Analysis of randomized trials with non-compliance and a binary outcome

James Robins[1] and Andrea Rotnitzky[2]
Harvard School of Public Health, Departments of [1]Epidemiology and [1;2]Biostatistics
and [2]Di Tella University, Department of Economics
655 Huntington Ave, Boston, USA
[1]robins@hsph.harvard.edu and [2]andrea@hsph.harvard.edu

In this presentation we summarize results in Robins and Rotnitzky (2001) concerning the problem of estimating the actual e®ect of treatment in randomized clinical trials with noncompliance and a dichotomous outcome. We allow both for the assigned and actual treatment to be continuous categorical or ordinal and for the possibility that the assigned treatment has a direct e®ect on the outcome through pathways other than actual treatment..Our proposed methodology is based on the logistic structural mean model of Robins (1997) and Robins et. al. (1999).

Special challenges arise for analyzing binary outcomes with logistic structural mean models. Speci¯cally, for continuous or count outcomes, unbiased estimating functions for the (structural) treatment e®ect of the additive and multiplicative structural mean models of Robins (1994) exist even in the presence of many continuous and discrete exogenous baseline covariates, provided the randomization probabilities (i.e. the conditional density of the instrument given the exogenous baseline covariates) are known. However, unless the causal sharp null hypothesis of no e®ect of either assigned or actual treatment holds, such unbiased estimating functions do not exist for the structural parameters of the logistic structural mean model. As a consequence, the best that can be hoped for are estimators, such as those proposed in this presentation, that are guaranteed both (i) to consistently estimate the (null) treatment e®ect when the null hypothesis of no e®ect of either assigned or actual treatment is true and, (ii) to have small bias when the true treatment e®ect is close but not equal to zero. In particular, one can use an estimator satisfying (i) in conjunction with its estimated standard error to construct a test that is an asymptotically distribution free ® ¡ level test both of the intention to treat null hypothesis that the outcome is conditionally independent of assigned treatment given baseline covariates and of the null hypothesis of no causal e®ect of either assigned or actual treatment. In the presence of noncompliance the possibility of constructing estimators that satisfy conditions (i) and (ii) is one of the most important bene¯ts to be realized from a randomized experiment. Because of the considerable cost associated with conducting a randomized experiment, we would regard as less than satisfactory any analytic approach that failed to satisfy (i) and (ii). Our work was stimulated by a recent paper on placebo controlled trials by Vansteelandt and Goetghebeur (2001) in which these authors, in conjunction with J Robins, proposed an estimation method that satis¯ed (i) and (ii). However, when some subjects in the placebo arm take the active treatment Vansteelandt and Goetghebeur modelling strategy can be internally inconsistent. For this reason, we prefer our proposed approach to that of Vansteelandt and Goetghebeur, in spite of their approach being less computationally demanding.

Under the exclusion assumption that assigned treatment has no direct e®ect on the outcome, the problem treated in this paper is equivalent to the problem of how to use an instrumental variable (i.e. assigned treatment) to estimate the e®ect of an endogenous (i.e. actual) treatment on a dichotomous response variable. This latter problem is a long standing problem in econometrics that was recently reviewed by Angrist (2001). When data on many continuous and discrete ex-

ogenous baseline covariates have been obtained, to the best of our knowledge, the only proposed econometric solutions are for the special case in which both assigned and actual treatment are themselves dichotomous, there are no de¯ers (i.e. there are no subjects such that they take treatment if and only if they are assigned to the control arm), the known randomization probabilities may depend on baseline covariates, and interest is in the treatment e®ect among those who comply with their assigned treatment (Abadie, 2001; Hirano, Imbens, Rubin and Zhou, 2000). For this special case, Abadie's estimator, in contrast to Hirano et. al's, satis¯es conditions (i) and (ii). In essence, the reason for the failure of Hirano's et al estimator is that it does not use knowledge of the randomization probabilities and thus, as shown by Robins and Ritov (1997), it cannot satisfy (i).

Often it is logistically or ethically impossible to conduct a randomized experiment. In that case one can sometimes ¯nd a variable that one is willing to consider an instrumental variable within levels of a vector of covariates. For example, one might regard the prescribed dose of a drug therapy to be an instrument for the dose actually taken within levels of the often high dimensional vector of prognostic factors used by physicians to determine the need for the therapy. Unlike in designed experiments, it would be typical for the probability of the instrument to depend in a complex fashion on such a high dimensional vector of covariates. Furthermore, unlike in a designed experiment, the probability of the instrument (prescribed therapy) conditional on the vector of prognostic factors would not be known and would have be estimated from the data using a dimension reducing, usually parametric, model. In such cases, no estimators satisfying (i) and (ii) will exist. The best that can be hoped for with a logistic structural mean model is, as argued in this presentation, to succeed in constructing an estimator that is so-called doubly robust or doubly protected under the null hypothesis of no e®ect of either prescribed or actual treatment. For additive or multiplicative structural mean models, estimators that are also doubly robust at non-null parameter values are available (Robins, 2000).

We now summarize the model assumptions and the inferential problem treated in this presentation. We assume that we observe n iid copies $O_i; i = 1; :::; n;$ of the vector $O = (L; Z; X; Y)$ where L is a vector recording pre-randomization covariates, Z records the randomly assigned treatment, X is the treatment actually received, and Y is the dichotomous indicator outcome of interest which we assume takes the values 0 or 1. Both Z and X can be discrete or continuous variables. Throughout, random variables are denoted with capital letters and their supports with the corresponding calligraphic letter.

To de¯ne treatment e®ects and our model we ¯rst de¯ne for each value $z \in \mathcal{Z}; x \in \mathcal{X}$ the potential outcome Y_{zx} (Neyman, 1923, Rubin, 1978, Robins 1986): This is de¯ned as the value of the dichotomous outcome of interest had, possible contrary to fact, Z been set to z and X set to x: Our semiparametric model, which throughout we denote by A; has observed data O and is de¯ned by the following restrictions: a) f (z j L) is known by design, b) $Z \quad Y_{zx} \mid L$; w.p.1 for all $z \in \mathcal{Z}; x \in \mathcal{X}$; c) $Y = Y_{zx}$ if $Z = z; X = x$; d)

$$\circ(l; z; x) = \circ(`; z; x; \tilde{A}_0) \tag{1}$$

where

$$\circ(z; x; l) = \log \left[\frac{P(Y_{zx} = 1 \mid Z = z; X = x; L = l)}{P(Y_{zx} = 0 \mid Z = z; X = x; L = l)} \right] - \left[\frac{P(Y_{z0} = 1 \mid Z = z; X = x; L = l)}{P(Y_{z0} = 0 \mid Z = z; X = x; L = l)} \right]$$

and $\tilde{A}_0$ is an unknown parameter vector of dimension p and $\circ(`; z; x; \tilde{A})$ is a known function

satisfying $\gamma(\ell, z, 0, \tilde{A}) = \gamma(\ell, z, x, 0) = 0$ and e) $b(z0 \mid \ell) = t(\ell) + t^{\natural}(\ell, z)$;with $t(\ell)$ a completely unknown function, $t^{\natural}(\ell, z)$ a known function satisfying $t(\ell, 0) = 0$ and $b(zx|d) \equiv \mathbb{C}^{-1}\{P(Y_{zx} = 1|D = d)\}$ where $\mathbb{C}(u) = e^u/(1 + e^u)$ and $\mathbb{C}^{-1}(u) = \log\{u/(1-u)\}$: Our goal is to estimate the function $\gamma(z, x, l)$ which measures the causal effect, on the logistic scale, of assigned treatment z and actual treatment x on the population assigned to z and treated with x and with baseline covariates equal to l: Randomization guarantees that Assumption b) holds. Assumption c) links the counterfactual data to the observed data. We make assumption d) in order to identify $\gamma(z, x, l)$: We do so because without modeling restrictions on $\gamma(l, z, x)$; this function is not generally identified from the observed data O: (See Robins,1994). Following Robins (1997) and Robins et al (1999) we call the model (1); a logistic structural mean model. Note that $\gamma(\ell, z, x, 0) = 0$ implies that if $\tilde{A}_0 = 0$ then the treatment $X = x$ has no effect on the outcome in the subpopulation with X observed to be x: Finally, assumption e) says that the direct effect of assigned treatment on the outcome when treatment actually taken is set to zero is known. Note that the special choice $t^{\natural}(\ell, z) = 0$ is equivalent to the assumption that $P(Y_{z0} = 1|L)$ is the same for all z and therefore is implied by (but does not imply) the exclusion assumption $Y_{zx} = Y_x$ w.p.1 for all $z \in \mathcal{Z}$ and $x \in \mathcal{X}$.

Though ideally we would like to make inference about $\tilde{A}_0$ under model A; unfortunately, due to the curse of dimensionality (Huber, 1985, Robins and Ritov, 1997) this is not feasible unless $\tilde{A}_0 = 0$ and $t^{\natural}(\ell, z) = 0$. Thus, in order to obtain well behaved estimators of $\tilde{A}_0$ when $\tilde{A}_0 \neq 0$ we propose estimating $\tilde{A}_0$ under a parametric model for the law of the observed data. However, we do not fit the model by maximum likelihood because the MLE of $\tilde{A}_0$ is not guaranteed to either (i) consistently estimate the (null) treatment effect $\tilde{A}_0 = 0$ when the sharp null hypothesis $Y_{zx} = Y$ w:p:1 of no effect of actual or assigned treatment is true or (ii) to have small bias when the true treatment effect $\tilde{A}_0$ is close but not equal to zero and assigned treatment has no direct effect on the outcome i.e. $t^{\natural}(l, z) = 0$.

We now describe estimating equations whose solutions satisfy (i) and (ii). Let $q(\ell, z, x) \equiv b(z0 \mid \ell, z, x) - b(z0 \mid \ell, z, x = 0)$ where $b(zx^{\emptyset}|l, z, x)$ denotes $\mathbb{C}^{-1}\{E(Y_{zx^{\emptyset}}|l, z, x)\}$ and $\mathbb{C}(u) := e^u/(1 + e^u)$: Robins and Rotnitzky (2001) showed that the observed data law can be parameterized in terms of $\gamma(\ell, z, x)$; $f(x \mid \ell, z)$; $t(\ell)$; and $q(\ell, z, x)$ because $b(z0 \mid \ell, z, 0)$ depends on $q(\ell, z, x)$; $t(\ell)$ and $f(x \mid \ell, z)$ since it is the unique solution to the integral equation

$$\mathbb{C}\{t(\ell) + t^{\natural}(\ell, z)\} = \int \mathbb{C}\{q(\ell, z, x) + v(\ell, z)\}\, dF(x \mid \ell, z) \tag{2}$$

Robins and Rotnitzky (2001) provide the explicit analytical solution of (2) when X is binary.

Our proposed estimator of $\tilde{A}$ is computed by the following two-stage estimation procedure. At stage 1, postulate working parametric models $f(x \mid \ell, z; \circledR)$; $t(\ell; !)$ and $q(\ell, z, x; \mu)$ for the unknowns $f(x \mid \ell, z)$; $t(\ell)$ and $q(\ell, z, x)$: For each fixed value of $\tilde{A}$; compute the estimator $\hat{\eta}(\tilde{A}) = (\hat{\circledR}(\tilde{A}), \hat{!}(\tilde{A}), \hat{\mu}(\tilde{A}))$ of $\eta = (\circledR, !, \mu)$; that maximize the parametric log-likelihood $\prod_i P(Y_i|L_i, Z_i, X_i, \tilde{A}, \eta)^{Y_i} f(X_i \mid L_i, Z_i, \circledR)$; or otherwise compute $\hat{\circledR}(\tilde{A}) = \hat{\circledR}$ by maximizing $\prod_i f(X_i \mid L_i, Z_i, \circledR)$ and then compute $\hat{\mu}(\tilde{A}), \hat{!}(\tilde{A})$ by maximizing the parametric conditional likelihood $\prod_i P(Y_i|L_i, Z_i, X_i, \tilde{A}, \mu, \hat{\circledR}, !)^{Y_i}$ over $\mu, !$ with $\hat{\circledR}$ and $\tilde{A}$ held fixed. At stage 2, compute the estimator $\hat{\tilde{A}}$ of $\tilde{A}$ that solves $\sum_i [h_{eff}(L_i, Z_i, \tilde{A}) - E\{h_{eff}(L_i, Z_i, \tilde{A}) \mid L_i\}]\, H_i(\tilde{A}, \hat{\eta}(\tilde{A})) = 0$; where

$$H(\tilde{A}, \eta) = \frac{E_\eta\{\mathbb{C}^{\emptyset}[M_1(\eta)] \mid L, Z\}\, \{Y - \mathbb{C}[M_2(\tilde{A}, \eta)]\}}{E_\eta\{\mathbb{C}^{\emptyset}[M_2(\tilde{A}, \eta)] \mid L, Z\}} + \mathbb{C}\{M_1(\eta)\} - \mathbb{C}\{M_3(\eta)\}$$

with $M_1(\psi) = q(L;Z{:}X;\mu)+v(L;Z;X;\psi)$; $M_2(\tilde{A};\psi) = \circ(L;Z;X;\tilde{A})+q(L;Z;X;\mu)+v(L;Z;X;\psi)$; $M_3(\psi) = t(L;!) + t^{\alpha}(L;Z)$ and $h_{eff}(L_i;Z_i;\tilde{A}) = c_{eff}(L_i;Z_i;\tilde{A};\psi(\tilde{A}))$ where

$$c_{eff}(L;Z;\tilde{A};\psi) = s(L;Z;\tilde{A};\psi)\,E\left\{\copyright^{\prime}[M_1(\psi)][@\circ(L;Z;X;\tilde{A})/@\tilde{A}]\,|L;Z\right\}_i$$
$$\frac{E\left\{\copyright\, s(L;Z;\tilde{A};\psi)\copyright^{\prime}[M_1(\psi)][@\circ(L;Z;X;\tilde{A})/@\tilde{A}]\,|L\right\}}{E\left\{s(L;Z;\tilde{A};\psi)\copyright^{\prime}[M_3(\psi)]\,|L\right\}} s(L;Z;\tilde{A};\psi)\copyright^{\prime}[M_3(\psi)]$$

with $s(L;Z;\tilde{A};\psi) = E\left\{\copyright^{\prime}[M_2(\psi)]\,|L;Z\right\}^{a} = \left\{\copyright\, E\left\{\copyright^{\prime}[M_1(\psi)]\,|L;Z\right\}^{\alpha a}\right\}^{2}$: The asymptotic variance of $\hat{A}$ is equal to the semiparametric variance bound for $\tilde{A}$ in model A under correct specification of the working models for $f(x\,|\,\grave{};z)$; $t(\grave{})$ and $q(\grave{};z;x)$:

REFERENCES

Abadie, A. (1999). Semiparametric instrumental variable estimation of treatment response models. Technical Report, MIT.

Angrist, J. (2001). Estimation of Limited-dependent Variable Models with Dummy Endogenous Regressors: Simple Strategies for Empirical Practice. J. of Business and Economic Statistics, 19, 2-16.

Hirano, K., Imbens, G., Rubin, D., and Zhou, A. (2000). Assessing the effect of an influenza vaccine in an encouragement design. Biostatistics, 1.1, 69-88.

Huber, P. (1985). Projection Pursuit. Annals of Statistics, 13, 435-475.

Neyman, J. (1923). On the application of probability theory to agricultural experiments: essay on principles, section 9. Translated in Statistical Science, 5, 565-480, 1990

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. Mathematical Modelling, 7:1393-1512.

Robins, J. M. (1994) Correcting for non-compliance in randomized trials using structural nested mean models, Communications in Statistics, Theory and Methods, 23, 2379-2412.

Robins, J.M. (1997). Causal Inference from Complex Longitudinal Data. Latent Variable Modeling and Applications to Causality. Lecture Notes in Statistics (120), M. Berkane, Editor. NY: Springer Verlag, pp. 69-117.

Robins, J. and Ritov, Y. (1997). Toward a curse of dimensionality appropriate (CODA) asymptotic theory for semiparametric models. Statistics in Medicine. 16, 285-319.

Robins, J.M. (1999). Marginal Structural models vs structural nested mean models as tools for causal inference. Statistical Models in Epidemiology: The environment and Clinical Trials, New York: Springer Verlag, 95-134.

Robins, J.M. (2000). Robust estimation in sequentially ignorable missing data and causal inference models. Proceedings of the American Statistical Association Section on Bayesian Statistical Science 1999, pp. 6-10.

Robins, J.M. and Rotnitzky, A. (2001). Estimation of treatment effects in randomized trials with noncompliance and a dichotomous outcome using logistic structural mean models. Submitted.

Rubin, D. (1978). Bayesian inference for causal models. Annals of Statistitcs, 6, 34-58.

Vansteelandt, S. and Goetghebeur, E. (2001) Causal inference with generalized structural mean models. To appear in the Journal of the Royal Statistical Society, series B.