THE ADVANTAGES AND DISADVANTAGES OF STATISTICAL DISCLOSURE LIMITATION FOR PROGRAM EVALUATION

John M. Abowd
Department of Economics
Labor Dynamics Institute
Cornell University
john.abowd@cornell.edu

Ian M. Schmutte
Department of Economics
Terry College of Business
University of Georgia
schmutte@uga.edu

October 2015

CONFERENCE DRAFT – PLEASE REQUEST FINAL VERSION FOR CITATION

We acknowledge direct support from Alfred P. Sloan Foundation Grant G-2015-13903. Abowd acknowledges direct support from NSF Grants BCS-0941226, TC-1012593 and SES-1131848.

Abstract

This paper formalizes the manner in which statistical disclosure limitation (SDL) hinders empirical research in economics. We also highlight a hitherto unappreciated advantage of SDL, formal privacy models, and synthetic data systems: they can serve as a defense against model overfitting and false-discovery bias. More specifically, a synthetic data validation system can – and we argue should – be used in conjunction with systems in which researchers register their research design ahead of analysis. The key insight is that privacy-protected data can be used for model development while minimizing risk of model overfitting. To demonstrate these points, we develop a model in which the statistical agency collects data from a population, but publishes a version in which the data that have been intentionally distorted by some SDL process. We say the SDL process is *ignorable* if inferences based on the published data are indistinguishable from inferences based on the unprotected data. SDL is rarely ignorable. If the researcher has knowledge of the SDL model, she can conduct an SDL-aware analysis that explicitly corrects for the effects of SDL. If, as is often the case, if the SDL model is unknown, we describe circumstances under which SDL can still be learned.

1 Introduction

Most modern methods for evaluating economic models and public policy rely on data measured at high levels of disaggregation. The need of data providers to protect respondent confidentiality often involves the use of methods for statistical disclosure limitation that compromise the quality of published data. The features of the data empirical researchers value: details of age, place of residence, earnings, and so on, are precisely those that require the most protection.

In general, the standard approach of ignoring statistical disclosure limitation leads to inconsistent estimation and incorrect inference. Despite this, and in part because details of the methods used to protect the data are not published most empirical research does not directly confront the effects of statistical disclosure limitation. This paper endeavors to precisely describe the circumstances under which SDL can safely be ignored, and what to do when it cannot.

It turns out that the steps taken to protect privacy in published data can have another, more salubrious, consequence for empirical research. Formal privacy systems and synthetic data can act as a natural defense against overfitting and false-discovery bias. We develop this intuition with an application of our model to regression discontinuity analysis. Researchers can use synthetic, or otherwise privacy protected data to develop their research design. However, they can only estimate the model on the true data once. The use of synthetic data thus protects against false-discovery bias and can therefore serve as the basis of experimental (and quasiexperimental) registration systems.

We develop a formal model that allows us to define *ignorable* statistical disclosure limitation. In the model, the statistical agency collects data from a population, but published a version of the data that have been intentionally distorted. Intuitively, SDL is ignorable if inferences based on the published data are indistinguishable from inferences based on the unprotected data. For most questions of interest to economists, SDL is rarely ignorable.

Finally, we show that the intuition underlying the RD example is true more generally for differentially private data publication systems. These methods al-

locate researchers a 'privacy budget' which is expended each time the researcher wants to conduct a new analysis. Under such a system, the privacy budget acts as a constraint on the amount of specification searching that can be undertaken. Moreover, the distortions in the published data, while compromising inference, also serve as a natural barrier to overfitting.

2 Economic Modeling in the Presence of SDL

We formalize the role of SDL in economic analysis using the concept of ignorability. Our approach is a direct extension of the ignorability of missing data developed by Rubin (1976). We first define the economic process model that the econometrician is trying to learn about. We then define the inclusion process that determines which parts of the economic process are actually observed. This gives rise to the well-known concept of *ignorable missing data* or, equivalently, *ignorable inclusion*. Finally, we formally define the SDL model and define *ignorable statistical disclosure limitation*.

2.1 The Economic Process Model

We consider a population of N entities that is described by a *complete-data* matrix Y, $N \times K$, a *process-parameter* vector θ_p, $P \times 1$, and two probability distributions: the *data model* $p_Y\left(Y \left| \theta_p \right.\right)$ and the *process-parameter prior distribution* $p_{\theta_p}\left(\theta\right)$.

The econometrician seeks to conduct estimation and inference concerning finite-population estimands, functions of Y only, and super-population estimands, functions of the parameters θ_p. We distinguish between these two estimand types because the statistical agencies that collect and disseminate the data we are discussing in this paper consider themselves to be engaged in producing finite-population estimands whereas the economists who analyze these data are primarily conducting super-population estimation and inference.[1]

[1]Many SDL methods, as well as methods from the newer data-privacy literature in computer science, explicitly consider the properties of these methods for finite-population estimands

2.2 The Data Inclusion Model and Ignorable Inclusion

Next, we define the tools necessary to understand the properties of published (released) data from conventional surveys, censuses, and administrative record systems. The population *inclusion matrix*, R, $N \times K$, indicates that an entity i has data for the associated variable, $r_{ij} = 1$, or not, $r_{ij} = 0$. If you think that this is needlessly complex, remember that we have not said that N is known nor how the statistician came to observe any element of Y. That is the role of the *inclusion model*: the distribution of R given Y is $p_{R|Y}\left(R\,|Y, \theta_D\right)$. θ_D, is the *design* parameter vector, so named because it characterizes how Y is observed, or the design of the survey or experiment. The *design-parameter prior distribution* is $p_{\theta_D|\theta_p}\left(\theta_D\,|\theta_p\right)$ allows for potential dependence of the design on the process parameters. The complete-data likelihood function[2] is then

$$\pounds_\theta\left(\theta_p, \theta_D\,|Y, R\right) = p_Y\left(Y\,|\theta_p\right) p_{R|Y}\left(R\,|Y, \theta_D\right) = p_{YR}\left(Y, R\,|\theta_p, \theta_D\right). \tag{1}$$

The term "complete data" means that this likelihood function applies to estimation and inference on the process and design parameters given a realization of Y, R from the super-population.

The *observed data* matrix, in the absence of SDL, is $Y^{(obs)}$, $N \times P$, contains a data item in $y_{ij}^{(obs)}$, if and only if $r_{ij} = 1$. The complement to the observed data matrix, in the absence of SDL is $Y^{(mis)}$, which contains the unobserved data items corresponding to $r_{ij} = 0$. The observed data likelihood function, in the absence of

whereas econometricians tend to focus on parametric (or semi-parametric) modeling focused on θ_p. The concept of ignorability was invented to allow a clean characterization of how the data collection process affects both types of modeling. We are not trying to be overly philosophical, just to provide a direct link between the way the data collectors think about the methods they use and the way data analysts trained in economics and econometrics use those data.

[2]The Rubin formulation includes the notion of fully observed covariates–variables that are never missing in the population and never have to be collected. In a known-finite population, these consist of variables on the frames used for sampling. Since these variables are also subjected to SDL when the data are published, we include them in the population data matrix Y.

SDL is

$$\begin{aligned} \pounds_{\theta}^{(obs)}\left(\theta_p, \theta_D \left| Y^{(obs)}, R\right.\right) &= p_{Y^{(obs)}R}\left(Y^{(obs)}, R \left| \theta_p, \theta_D \right.\right) \quad (2) \\ &= \int p_{YR}\left(Y, R \left| \theta_p, \theta_D \right.\right) dY^{(mis)}. \quad (3) \end{aligned}$$

The term "observed data" derives from the application of these modeling concepts to sampling, experimental design, and unintentionally missing data (missing survey records or responses, unreported administrative records, etc.). In the standard analysis of ignorability (e.g., Gelman et al. 2013), the published data would be $Y^{(obs)}$. The notation may seem awkward for the application to SDL, but it seems better to us to use this conventional notation. Wherever the term $Y^{(obs)}$ occurs, think: the actual confidential data collected by the statistical agency.

Inference and estimation, in the absence of SDL, are based on the joint posterior distribution of (θ_p, θ_D), given the observed data, which we assemble from the pieces defined above as

$$\begin{aligned} p_{\theta_p\theta_D|Y^{(obs)}R}\left(\theta_p, \theta_D \left| Y^{(obs)}, R\right.\right) &\propto p_{\theta_D|\theta_p}\left(\theta_D \left| \theta_p \right.\right) p_{\theta_p}\left(\theta_p\right) p_{Y^{(obs)}R}\left(Y^{(obs)}, R \left| \theta_p, \theta_D \right.\right) \\ &= p_{\theta_D|\theta_p}\left(\theta_D \left| \theta_p \right.\right) p_{\theta_p}\left(\theta_p\right) \pounds_{\theta}^{(obs)}\left(\theta_p, \theta_D \left| Y^{(obs)}, R\right.\right) \quad (4) \end{aligned}$$

In general, we focus interest on the posterior distribution of θ_p which, in the absence of SDL, is

$$\begin{aligned} p_{\theta_P|Y^{(obs)}R}\left(\theta_p \left| Y^{(obs)}, R\right.\right) &= \int p_{\theta|Y^{(obs)}R}\left(\theta_p, \theta_D \left| Y^{(obs)}, R\right.\right) d\theta_D \quad (5) \\ &\propto \int\int p_Y\left(Y \left| \theta_p \right.\right) p_{R|Y}\left(R \left| Y, \theta_D \right.\right) p_{\theta_D|\theta_p}\left(\theta_D \left| \theta_p \right.\right) p_{\theta_p}\left(\theta_p\right) dY^{(mis)} d\theta_D \end{aligned}$$

The data inclusion model is *ignorable* if

$$p_{\theta_P|Y^{(obs)}R}\left(\theta_p \left| Y^{(obs)}, R\right.\right) = p_{\theta_P|Y^{(obs)}}\left(\theta_p \left| Y^{(obs)}\right.\right). \quad (6)$$

For reasons that will be clear shortly, we call this *ignorable inclusion* (or *ignorable sampling*, or *ignorable missing data*, if the context of the inclusion model is clear).

Our definition of ignorability is general enough to cover observational data, survey designs, experiments, and unintentional missing data models. It says that inference and estimation about the super-population parameters is ignorable if it does not depend on the unobserved data, $Y^{(mis)}$. It is not general enough to cover SDL because $Y^{(obs)}$ undergoes an additional transformation before being published.

2.3 The SDL Model and Ignorable SDL

We characterize the SDL probabilistically using the same tools as we have used for the data model, the inclusion model, and their parameters. The *published data* Z, $N \times K$, are generated by the *SDL model* $p_{Z|Y,R}\left(Z\,|Y,R,\theta_S\right)$ with *SDL-parameter* vector θ_S. The *SDL-parameter prior distribution* is $p_{\theta_S|\theta_D\theta_p}\left(\theta_S\,|\theta_D,\theta_p\right)$. The likelihood function for the published data is

$$\begin{aligned} \pounds_{\theta}^{(pub)}\left(\theta_p,\theta_D,\theta_S\,|Z,R\right) &= \int p_{Z|YR}\left(Z\,|Y,R,\theta_S\right) p_{YR}\left(Y,R\,|\theta_p,\theta_D\right) dY \\ &= \int p_{Z|YR}\left(Z\,|Y,R,\theta_S\right) p_{R|Y}\left(R\,|Y,\theta_D\right) p_{Y}\left(Y\,|\theta_p\right) dY \end{aligned} \tag{7}$$

Once again, estimation and inference are based on the posterior distribution of the process parameters, which is derived from the joint posterior distribution of the model, inclusion, and publication parameters given the published data and the inclusion matrix

$$\begin{aligned} p_{\theta|ZR}\left(\theta_p,\theta_D,\theta_S\,|Z,R\right) &\propto \int p_{Z|YR}\left(Z\,|Y,R,\theta_S\right) p_{YR}\left(Y,R\,|\theta_p,\theta_D\right) p_{\theta}\left(\theta\right) dY \\ &= p_{\theta}\left(\theta\right) \pounds_{\theta}^{(pub)}\left(\theta_p,\theta_D,\theta_S\,|Z,R\right) \end{aligned}$$

where $p_{\theta}\left(\theta\right) = p_{\theta_S|\theta_D\theta_p}\left(\theta_S\,|\theta_D,\theta_p\right) p_{\theta_D|\theta_p}\left(\theta_D\,|\theta_p\right) p_{\theta_p}\left(\theta_p\right)$. So that the posterior dis-

tribution of the process parameters is

$$\begin{aligned} p_{\theta_P|ZR}\left(\theta_p\left|Z,R\right.\right) &= \int\int p_{\theta|ZR}\left(\theta_p,\theta_D,\theta_S\left|Z,R\right.\right)d\theta_D d\theta_S \\ &\propto \int\int p_\theta\left(\theta\right)\pounds_\theta^{(pub)}\left(\theta_p,\theta_D,\theta_S\left|Z,R\right.\right)p_{\theta_D\theta_S}\left(\theta_D,\theta_S\right)d\theta_D d\theta_S \end{aligned} \tag{8}$$

The relation between equations (5) and (8) is

$$p_{\theta_P|ZR}\left(\theta_p\left|Z,R\right.\right) = \int p_{\theta_P|Y^{(obs)}R}\left(\theta_p\left|Y^{(obs)},R\right.\right)p_{Y^{(obs)}|ZR}\left(Y^{(obs)}\left|Z,R\right.\right)dY^{(obs)} \tag{9}$$

That is, the posterior distribution of the process parameters θ_p given the published data and inclusion matrix is the expectation of the posterior distribution of the process parameters given the observed data (the actual confidential data used by the agency) and inclusion matrix with the expectation taken over the posterior predictive distribution of the observed data given the published data and inclusion matrix. This formulation assumes that the agency also publishes R, which is not innocuous but we will usually be analyzing models in which we assume ignorable inclusion.

We define *ignorable statistical disclosure limitation* as

$$p_{\theta_P|Y^{(obs)}R}\left(\theta_p\left|Y^{(obs)}=Z,R\right.\right) = p_{\theta_P|ZR}\left(\theta_p\left|Z,R\right.\right) \tag{10}$$

The definition is subtle, so we repeat it in words. The SDL is ignorable if and only if analyzing the posterior distribution of the process parameters given the published data is equivalent to analyzing the posterior distribution of process parameters given the observed data and assuming that the published data are identical to the (confidential) observed data.

If the model possesses both ignorable inclusion and ignorable SDL then

$$p_{\theta_P|Y^{(obs)}}\left(\theta_p\left|Y^{(obs)}=Z\right.\right) = p_{\theta_P|Z}\left(\theta_p\left|Z\right.\right). \tag{11}$$

Equation (11) summarizes both the sampling (or inclusion) and SDL assumptions

that are embodied in any economic analysis that treats the published data as if they had been produced by an ignorable inclusion process without SDL; that is, without explicitly modeling the sample design and SDL.

2.4 Example: Ignorable and Nonignorable Coarsening

Heitjan and Rubin (1991) consider the problem of inference when the data contain reporting errors where, for instance, individuals round hours or earnings to salient, whole numbers. The same model is relevant to those types of microdata masking that aggregate attributes, including topcoding.

Assume X is a vector-valued random variable distributed according to $f(x|\theta_X)$, and the goal of research is to learn something about θ_X. Rather than observe X, the researcher observes $Y = M(X, G)$ where M is the microdata mask and G is a random variable that determines how the mask will be applied. In the topcoding example, the mask is a topcode, and G is a binary random variable, conditional on X, that indicates whether a particular data item is to be topcoded. The random variable G is never directly observed, and the effect of the mask on inference depends, in part, on whether the researcher can infer G from the published data. The distribution of G is parameterized by θ_G.

The coarsening is deterministic once the true data X, and the variable G are both known. The conditional distribution of the published data is degenerate with point mass on the coarsened data, Y:

$$p_{y|x,g}(y|x, g, \theta_X, \theta_G) = \begin{cases} 1, & if \quad y = M(x, g) \\ 0, & if \quad y \neq M(x, g) \end{cases} \tag{12}$$

Inference should be based on the likelihood for the published data given the coarsening rule:

$$L_C(\theta_X, \theta_G|y) = \int \int p_{y|x,g}(y|x, g, \theta_X, \theta_G) p_{g|x,\theta_G} dg p_x(x|\theta_X) dx \tag{13}$$

However, it is common to ignore the stochastic nature of the coarsening process,

estimating the model on the grouped data from

$$L_G(\theta_X|y) = \int p_{y|x,g,\theta_X} p_x(x|\theta_X)dx. \tag{14}$$

Heitjan and Rubin (1991) prove that if the data are coarsened at random, then inference based on (13) is equivalent to inference based on (14). That is, the coarsening process is formally ignorable. The data are coarsened at random if the probability $G = 1$ is independent of the value of y.

3 Implementing SDL-aware Data Analysis

Since equation (9) is an identity, it is, in principle, possible to do any data analysis using methods that account for the SDL. In practice, we must confront whether or not the SDL process is known, and if it is known, whether the components required to compute $p_{\theta_P|ZR}(\theta_p|Z,R)$ can be assembled. We will define an SDL method as *fully discoverable* if $p_{\theta_P|ZR}(\theta_p|Z,R)$ can be computed. If the SDL process is not fully discoverable, then we will consider some diagnostic methods that can be used to approximate $p_{\theta_P|ZR}(\theta_p|Z,R)$ or to detect failures of equation (10).

At the heart of the implementation is the computation of $p_{Y^{(obs)}|ZR}\left(Y^{(obs)}|Z,R\right)$, which is the posterior predictive distribution of the data that would have been published in the absence of SDL, given the published data and the inclusion matrix. In the absence of any ignorability assumptions the computations can be done using Markov Chain Monte Carlo sampling from the conditional distributions

$$p_{\theta_p\theta_D|Y^{(obs)}R}\left(\theta_p,\theta_D\left|Y^{(obs)},R\right.\right)$$

$$p_{\theta_S|ZR\theta_p\theta_D}\left(\theta_S|Z,R,\theta_p,\theta_D\right)$$

$$p_{Y^{(obs)}|ZR\theta_p\theta_D\theta_S}\left(Y^{(obs)}|Z,R,\theta_p,\theta_D,\theta_S\right)$$

starting from arbitrary initial values of $Y^{(obs)}$, and $(\theta_p,\theta_D,\theta_S)$.

In many ways, implementing SDL-aware data analysis is similar to imple-

menting ignorable and nonignorable missing data models. Since there are many excellent discussions of missing data issues and in order to focus our contribution more clearly, we consider next implementing SDL-aware analysis when the inclusion model is provably ignorable. A leading case is the inclusion model in which data are missing at random in the sense of Rubin (1987); then, inclusion model can be ignored because

$$p_{R|Y}\left(R\left|Y,\theta_D\right.\right) = p_{R|Y}\left(R\left|Y^{(obs)},\theta_D\right.\right)$$

and

$$p_{\theta}\left(\theta\right) = p_{\theta_S|\theta_p\theta_D}\left(\theta_S\left|\theta_p,\theta_D\right.\right) p_{\theta_D}\left(\theta_D\right) p_{\theta_p}\left(\theta_p\right)$$

To further simplify, simple random sampling implies that the inclusion model does not depend upon any unknown parameters nor on the population data; hence $p_{R|Y}\left(R\left|Y,\theta_D\right.\right) = p_R\left(R\right)$, which allows R and θ_D to be eliminated altogether from the analysis of the published data.

It is enlightening to study the SDL-aware data analysis equations under the assumption that the inclusion model is ignorable and known. Then,

$$\begin{aligned} p_{\theta_P|ZR}\left(\theta_p\left|Z,R\right.\right) &= p_{\theta_P|Z}\left(\theta_p\left|Z\right.\right) \\ &= \int p_{\theta_P|Y^{(obs)}}\left(\theta_p\left|Y^{(obs)}\right.\right) p_{Y^{(obs)}|Z}\left(Y^{(obs)}\left|Z\right.\right) dY^{(obs)} \end{aligned} \tag{15}$$

$$p_{\theta_p\theta_D|Y^{(obs)}R}\left(\theta_p,\theta_D\left|Y^{(obs)},R\right.\right) = p_{\theta_p|Y^{(obs)}}\left(\theta_p\left|Y^{(obs)}\right.\right) \tag{16}$$

$$p_{\theta_S|ZR\theta_p\theta_D}\left(\theta_S\left|Z,R,\theta_p,\theta_D\right.\right) = p_{\theta_S|Z\theta_p}\left(\theta_S\left|Z,\theta_p\right.\right) \tag{17}$$

and

$$p_{Y^{(obs)}|ZR\theta_p\theta_D\theta_S}\left(Y^{(obs)}\left|Z,R,\theta_p,\theta_D,\theta_S\right.\right) = p_{Y^{(obs)}|Z\theta_p\theta_S}\left(Y^{(obs)}\left|Z,\theta_p,\theta_S\right.\right) \tag{18}$$

Estimation and inference using the SDL-aware system described by equations (15)-(18) can be applied to many common SDL methods, including those introduced in the data-privacy literature in CS.

Although we largely limit our attention in this paper to SDL-aware analyses that assume that the inclusion model is known and ignorable, we do not mean to endorse these assumptions universally. In particular, we have chosen examples where the inclusion model's properties are well understood or provably ignorable for most of our examples below.

4 Application: Estimating Proportions with Randomized Response

Randomized response (Warner 1965) is a survey technique in which the respondent is presented with one of two questions that can both be answered "yes" or "no." The respondent is asked a sensitive question with a certain probability (e.g. "Have you ever committed a violent crime?"), and a non-sensitive question with complementary probability (e.g., "Is your birthday in December?"). The survey records only the binary answer (yes or no) and the contents of the question are destroyed. Data publication under randomized response is identical to the input noise infusion SDL procedure in which the published variable is either the actual response or an unrelated binary random variable with a known distribution. This form of SDL is provably private, but also provably non-ignorable, as we now formally demonstrate.

4.1 Formal Model of Randomized Response

Consider the SDL-aware analysis of randomized response assuming that the inclusion model is known and ignorable. The data model is

$$y_i \sim \text{Bin}\left(\theta, 1\right) \; i.i.d. \; i = 1, \ldots n.$$

The process-parameter prior distribution is

$$\theta \sim \text{Beta}\left(\alpha, \beta\right).$$

In the absence of SDL, the observed data likelihood function is

$$\mathcal{L}_{\theta}^{(obs)}\left(\theta_p \left| Y^{(obs)} \right.\right) = \prod_{i=1}^{n} \theta^{y_i} \left(1-\theta\right)^{(1-y_i)}.$$

The posterior distribution of the process parameter given $Y^{(obs)}$ is

$$\theta \sim \text{Beta}\left(\alpha + \sum y_i, \beta + n - \sum y_i\right).$$

The posterior mean is $\text{E}\left[\theta \left| Y^{(obs)} \right.\right] = \frac{\alpha+\sum y_i}{\alpha+\beta+n}$. The usual frequentist estimator just sets $\alpha = \beta = 0$. For large n the posterior is dominated by the likelihood component; hence, we won't be concerned with sensitivity analysis of the prior distribution here.

Now add the SDL model. Let γ_i be the random variable that controls which question is asked. Then,

$$\gamma_i \sim \text{Bin}\left(\rho, 1\right) \ i.i.d. \ i = 1, \ldots n,$$

where ρ is the SDL parameter controlling the probability that the question of interest is asked, event $\gamma_i = 1$. Let δ_i be the random variable that determines the correct answer to the innocuous question (or the input noise outcome). Then,

$$\delta_i \sim \text{Bin}\left(\mu, 1\right) \ i.i.d. \ i = 1, \ldots n,$$

where μ is the SDL parameter that controls the probability that the answer to the innocuous is "yes" (or the input noise infusion has $\delta_i = 1$).

The likelihood for the published data Z comes from from the equation

$$z_i = \gamma_i y_i + \left(1-\gamma_i\right)\delta_i$$

which implies

$$\begin{aligned}\mathcal{L}_{\theta}^{(pub)}\left(\theta_p,\theta_D,\theta_S\left|Z,R\right.\right) &= \mathcal{L}_{\theta}^{(pub)}\left(\theta,\rho,\mu\left|Z\right.\right)\\ &= \prod_{i=1}^{n}\left[\begin{array}{c}\left\{\left[\rho+\left(1-\rho\right)\mu\right]\theta+\left[\left(1-\rho\right)\mu\right]\left(1-\theta\right)\right\}^{z_i}\\ \left\{1-\left[\rho+\left(1-\rho\right)\mu\right]\theta+\left[\left(1-\rho\right)\mu\right]\left(1-\theta\right)\right\}^{\left(1-z_i\right)}\end{array}\right]\end{aligned} \tag{19}$$

where θ is the only process parameter and the SDL parameters are $\theta_S=(\rho,\mu)$.

To finish the specification, we need to put a prior distribution on the SDL parameters. Continuing the analogy to randomized response for the moment, we could assume that (ρ,μ) is known, say (ρ^0,μ^0). This is always the case when randomized response is used in the original survey context; however, as a general SDL method, only certain features might be public. For example, the data publisher might say: "A small percentage of the cases (less than 5%) was replaced with binomial noise where the binomial probability was equal to the estimated proportion of true yes answers in the confidential sample." Such a statement is not ridiculous. It corresponds approximately to the way the Survey of Consumer Finances applies SDL to certain variables. In this case, a reasonable prior puts $\mu=\theta$ and $\Pr\left[\rho>0.05\right]=0$. In the differential privacy context, (ρ^0,μ^0) are also known (and used to determine the level of differential privacy provided). Assuming the SDL parameters are known implies

$$\begin{aligned}p_{\theta|Z}\left(\theta,\rho,\mu\left|Z\right.\right) \propto & p_{\theta}\left(\theta\right)\mathcal{L}_{\theta}^{(pub)}\left(\theta\left|Z,\rho^0,\mu^0\right.\right)\\ = & \prod_{i=1}^{n}\left[\begin{array}{c}\left\{\left[\rho^0+\left(1-\rho^0\right)\mu^0\right]\theta+\left[\left(1-\rho^0\right)\mu^0\right]\left(1-\theta\right)\right\}^{z_i}\\ \left\{1-\left[\rho^0+\left(1-\rho^0\right)\mu^0\right]\theta+\left[\left(1-\rho^0\right)\mu^0\right]\left(1-\theta\right)\right\}^{\left(1-z_i\right)}\end{array}\right]\\ & \times\frac{\Gamma\left(\alpha+\beta\right)}{\Gamma\left(\alpha\right)\Gamma\left(\beta\right)}\theta^{\alpha}\left(1-\theta\right)^{\beta}\end{aligned} \tag{20}$$

Because the likelihood function is a mixture of binomials, exact Bayesian analysis of even this simple case requires MCMC sampling. If we use an improper prior, then we can perform estimation and inference on just the likelihood component.

In this case we have

$$p\lim \bar{z} = \left[\rho^0 + \left(1-\rho^0\right)\mu^0\right]\theta^0 + \left[\left(1-\rho^0\right)\mu^0\right]\left(1-\theta^0\right)$$

where $\bar{z}$ is the sample mean of z_i and θ^0 is the mean of the superpopulation. The implied estimator for θ is the usual randomized response MLE

$$\hat{\theta}_{MLE} = \frac{\bar{z} - \left(1-\rho^0\right)\mu^0}{\rho^0}.$$

If the only superpopulation estimand of interest is θ, then the MLE is probably adequate, since it is also the mode of posterior distribution when n is large.

It has been recognized for many years that the SDL is not ignorable in this situation. It's straightforward to confirm by examining

$$\begin{aligned}
p_{\theta|Y^{(obs)}}\left(\theta \left| Y^{(obs)} = Z\right.\right) &= \frac{\Gamma\left(\alpha+\beta+n\right)}{\Gamma\left(\alpha+\sum z_i\right)\Gamma\left(\beta+n-\sum z_i\right)}\theta^{\alpha+\sum z_i}\left(1-\theta\right)^{\beta+n-\sum z_i} \\
&\neq \prod_{i=1}^{n}\left[\begin{array}{c}\left\{\left[\rho^0+\left(1-\rho^0\right)\mu^0\right]\theta + \left[\left(1-\rho^0\right)\mu^0\right]\left(1-\theta\right)\right\}^{z_i} \\ \left\{1-\left[\rho^0+\left(1-\rho^0\right)\mu^0\right]\theta + \left[\left(1-\rho^0\right)\mu^0\right]\left(1-\theta\right)\right\}^{(1-z_i)}\end{array}\right] \\
&\quad \times \frac{\Gamma\left(\alpha+\beta\right)}{\Gamma\left(\alpha\right)\Gamma\left(\beta\right)}\theta^{\alpha}\left(1-\theta\right)^{\beta} \\
&= p_{\theta_P|Z}\left(\theta_p \left| Z\right.\right).
\end{aligned}$$

Although we did not need the full appartus of equations (15)-(18) to get reasonable data evidence about θ from the disclosure-limited publication data, that situation changes when we consider SDL-aware analysis of multivariate relationships when some of the variables have been subjected to nonignorable SDL with known parameters.

Consider the simplest multivariate analysis in which we can embed randomized response SDL. We continue to assume that the inclusion model is ignorable. The data model is

$$y_i \sim \text{Multinomial}\left(\theta_p, 1\right) \; i.i.d. \; i = 1, \ldots n$$

where the process parameter θ is a vector of K probabilities. The process-parameter prior distribution is

$$\theta \sim \text{Dirichlet}\,(\alpha)$$

where α is a vector of K parameters, usually called the prior sample sizes. Passing directly to the posterior distribution for θ we have

$$\theta \sim \text{Dirichlet}\,(\alpha + \sum y_i)$$

where the vector $\alpha + \sum y_i$ is usually called the posterior sample sizes.

Now apply randomized response SDL to each element of y_i using the process parameter vectors ρ and μ, each $K \times 1$ arranged conformably to θ_p. Marginally, each publication variable z_{ij} can be used for estimation and inference on θ_i because the marginal likelihood is in the form given in equation (19) and the posterior distribution of each process parameter θ_i is in the form given in equation (20). If the K-way table implied by cross-classifying all of the y_{ij} variables is not too cumbersome, then the multivariate generalization of equation (20) can be used for SDL-aware analysis. We are not going to analyze that case because economists generally don't use multiway contingency table models. Instead, we consider models that combine discrete and continuous variables in the form of linear probability models, logististic regression and probit models. These models are very similar to the regression models considered in the next two sections. We defer to those sections the analysis of regression-like implementations of these conditional probabilities.

5 Application to Regression Discontinuity Models

In our analysis of the effect of SDL on regression discontinuity designs, we consider the case in which the following model of SDL was applied to the running variable.

5.1 Generalized Randomized Response SDL

The published data are

$$\begin{aligned} \omega_i &= w_i^* \\ z_{i3} &\quad \text{sampled from } p_{Z_3|Y_3}\left(z_{i3}\left|y_{i3}, \theta_S\right.\right) \\ z_{i4} &= 1\left[z_{i3} \geq \tau\right] \end{aligned}$$

with $p_{Z_3|Y_3}\left(z_{i3}\left|y_{i3}, \theta_S\right.\right)$ given by the following mixture model, which is a generalization of randomized response. The randomization variable is $\gamma_i \sim \text{Bin}\left(\rho, 1\right)$. When $\gamma_i = 1$, $z_{i3} = y_{i3}$; otherwise $z_{i3} = y_{i3} + \varepsilon_i$ with $\varepsilon_i \sim \text{N}\left(0, \delta^2\right)$, (*i.e.,* additive noise infusion).

These assumptions imply

$$z_{i3} = \gamma_i y_{i3} + \left(1 - \gamma_i\right)\left(y_{i3} + \varepsilon_i\right),$$

$$z_{i4} = \begin{cases} 1\left[y_{i3} \geq \tau\right] \text{ if } \gamma_i = 1 \\ 1\left[y_{i3} + \varepsilon_i \geq \tau\right] \text{ if } \gamma_i = 0 \end{cases}$$

and

$$p_{Z_3Z_4|Y_3}\left(z_{i3}, z_{i4}\left|y_{i3}, \theta_S\right.\right) = \rho p_{Y_3Y_4}\left(Z_3, Z_4\left|\theta_p\right.\right) + \left(1 - \rho\right) p^*_{Y_3Y_4}\left(Z_3, Z_4\left|\theta_p, \delta^2\right.\right),$$

where $p^*_{Y_3Y_4}\left(Z_3, Z_4\left|\theta_p, \delta^2\right.\right)$ is the distribution function from the convolution of $p_{Y_3Y_4}\left(Y_3, Y_4\left|\theta_p\right.\right)$ and $\text{N}\left(0, \delta^2\right)$.

5.2 SDL Aware Analysis of the RD Model

Using the posterior predictive distribution for y_{i3} given z_{i3} and assuming that the SDL parameters are fixed at the known values ρ_0 and δ_0, we have

$$\text{E}\left[y_{i3}\left|z_{i3}, \rho_0, \delta_0\right.\right] = \text{E}\left[z_{i3} - \left(1 - \gamma_i\right)\varepsilon_i\left|z_{i3}, \rho_0, \delta_0\right.\right] = z_{i3}$$

and

$$\begin{aligned} \mathrm{E}\left[y_{i4}\left|z_{i3},\rho_0,\delta_0\right.\right] &= \mathrm{E}\left[1\left[y_{i3}\geq\tau\right]\left|z_{i3},\rho_0,\delta_0\right.\right] \\ &= \rho_0\,1\left[z_{i3}\geq\tau\right]+\left(1-\rho_0\right)\Phi\left(\frac{z_{i3}-\tau}{\delta_0}\right) \end{aligned} \tag{21}$$

where $\Phi\left(\right)$ is the standard normal cumulative distribution function. The SDL-aware analysis has converted the original sharp RD into a fuzzy RD. To complete the analysis we should use the posterior distribution of θ_{RD} given the published data Z and the SDL parameters, assumed known or with an informative prior given agency-provided data.

In the RD literature, functional form assumptions about $f_1\left(y_{i3}\right)$, $f_2\left(y_{i3}\right)$, and $\pounds_\theta^{(obs)}\left(\theta_p\left|Y^{(obs)}\right.\right)$ are minimized. Respecting this analysis style, without implying that it is the best way to analyze a finite sample of size n from a superpopulation with size N, we analyze a few posterior moments, making the assumption that those exist.

We want to estimate

$$\mathrm{E}\left[\theta_{RD}\left|Z,\rho_0,\delta_0\right.\right] = \mathrm{E}\left[\lim_{y_{i3}\downarrow\tau}\mathrm{E}\left[y_{i2}\left|y_{i3}=\tau\right.\right]\left|Z,\rho_0,\delta_0\right.\right] \tag{22}$$

$$-\mathrm{E}\left[\lim_{y_{i3}\uparrow\tau}\mathrm{E}\left[y_{i1}\left|y_{i3}=\tau\right.\right]\left|Z,\rho_0,\delta_0\right.\right] \tag{23}$$

$$\begin{aligned} &= \mathrm{E}\left[\lim_{y_{i3}\downarrow\tau}f_2\left(y_{i3}\right)\left|Z,\rho_0,\delta_0\right.\right]-\mathrm{E}\left[\lim_{y_{i3}\uparrow\tau}f_1\left(y_{i3}\right)\left|Z,\rho_0,\delta_0\right.\right] \\ &= \rho_0\left\{\begin{array}{c}\mathrm{E}\left[\lim_{z_{i3}\downarrow\tau}f_2\left(z_{i3}\right)\left|Z,\gamma_i=1,\delta_0\right.\right] \\ -\mathrm{E}\left[\lim_{z_{i3}\uparrow\tau}f_1\left(z_{i3}\right)\left|Z,\gamma_i=1,\delta_0\right.\right]\end{array}\right\} \\ &+ \left(1-\rho_0\right)\left\{\begin{array}{c}\mathrm{E}\left[\lim_{z_{i3}\downarrow\tau}f_2\left(z_{i3}-\varepsilon_i\right)\left|Z,\gamma_i=0,\delta_0\right.\right] \\ -\mathrm{E}\left[\lim_{z_{i3}\uparrow\tau}f_1\left(z_{i3}-\varepsilon_i\right)\left|Z,\gamma_i=0,\delta_0\right.\right]\end{array}\right\} \\ &= \rho_0\left(\lim_{z_{i3}\downarrow\tau}f_2\left(\tau\right)-\lim_{z_{i3}\uparrow\tau}f_1\left(\tau\right)\right) \end{aligned}$$

and

$$\begin{aligned}\rho_0 &= \lim_{z_{i3}\downarrow\tau}\left[\rho_0\,1\left[z_{i3}\geq\tau\right]+(1-\rho_0)\,\Phi\left(\frac{z_{i3}-\tau}{\delta_0}\right)\right]\\ &\quad-\lim_{z_{i3}\uparrow\tau}\left[\rho_0\,1\left[z_{i3}\geq\tau\right]+(1-\rho_0)\,\Phi\left(\frac{z_{i3}-\tau}{\delta_0}\right)\right]\end{aligned}$$

The regime where $\gamma_i = 1$ is a conventional RD. The existence of the regime $\gamma_i = 0$ converts the problem to a fuzzy RD where $\mathrm{E}\left[y_{i4}\left|z_{i3},\rho_0,\delta_0\right.\right] = g\left(z_{i3}\right)$ plays the role of the "compliance status" function. The term

$$(1-\rho_0)\left\{\mathrm{E}\left[\lim_{z_{i3}\downarrow\tau} f_2\left(z_{i3}-\varepsilon_i\right)|Z,\gamma_i=0,\delta_0\right]-\mathrm{E}\left[\lim_{z_{i3}\uparrow\tau} f_1\left(z_{i3}-\varepsilon_i\right)|Z,\gamma_i=0,\delta_0\right]\right\} \tag{24}$$

is zero because $\varepsilon_i \sim \mathrm{N}\left(0,\delta^2\right)$ implies that in the regime $\gamma_i = 0$, there is no point mass at $\varepsilon_i = 0$; hence there is no jump at τ–the continuous function $f_1\left(z_{i3}\right)$ transitions smoothly to $f_2\left(z_{i3}\right)$ over the support of ε_i. The SDL noise needn't be normal, but it must be drawn from a continuous distribution.

5.2.1 Implications of SDL in the Running Variable for other RD Models

If generalized random response SDL is applied to the running variable, then the SDL is ignorable for parameter estimation when the true RD design is fuzzy. The FRD compliance function, augmented with the contribution from SDL, becomes

$$h(z_i) = \mathrm{E}\left[t_i\left|z_i,\rho_0,\delta_0\right.\right] \tag{25}$$

$$= \rho_0 p_{T|R}\left(t_i=1|z_i\right)+(1-\rho_0)\int p_{T|R}(t_i=1|r_i)p_{R|Z}(r_i|z_i)dr. \tag{26}$$

It immediately follows

$$\lim_{z_i\downarrow\tau} h\left(z_i\right)-\lim_{z_i\uparrow\tau} h\left(z_i\right)=\rho_0\left[\lim_{z_i\downarrow\tau} p_{T|R}\left(t_i=1|z_i\right)-\lim_{z_{i1}\uparrow\tau} p_{T|R}\left(t_i=1|z_i\right)\right].$$

The second summand in the expression for $h(z_i)$ is zero. When the running variable is distorted with normally distributed noise, there is no point mass anywhere, and hence no discontinuity in the probability of treatment at τ. The claim that the SDL is ignorable for consistent estimation of the treatment effect in the fuzzy RD design follows. Imbens and Lemieux (2008) show that the IV estimator that uses the RD as an exclusion restriction is formally equivalent to the fuzzy RD estimator, so the SDL is also ignorable for consistent estimation in this case.

6 Conclusion

In Progress.

Bibliography

Heitjan, D. F. and Rubin, D. B. (1991). Ignorability and coarse data, *The Annals of Statistics* **19**(4): 2244–2253.

Imbens, G. W. and Lemieux, T. (2008). Regression discontinuity designs: A guide to practice, *Journal of Econometrics* **142**(2): 615–635.

Rubin, D. B. (1976). Inference and missing data, *Biometrika* **63**(3): 581–592.