*Brookings Papers on Economic Activity*, Spring 2015

**Online Technical Appendix** to **"Economic Analysis and Statistical Disclosure Limitation"**

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

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

**August 7, 2015**

# A Ignorable and Nonignorable 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). Little (1993) anticipated much of our analysis, including the use of hierarchical models that introduced SDL via generalized randomized response. We first define the economic process model that the econometrician is trying to learn. 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*.

## A.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 $\theta_p$, $P \times 1$, and two probability distributions: the *data model* $p_Y\left(Y \left| \theta_p \right.\right)$ and the *process-parameter prior distribution* $p_{\theta_p}\left(\theta_p\right)$.

The econometrician seeks to conduct estimation and inference concerning finite-population estimands, functions of $Y$ only, and super-population estimands, functions of the parameters $\theta_p$. 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.[9]

## A.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

[9]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 whereas econometricians tend to focus on parametric (or semi-parametric) modeling focused on $\theta_p$. 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.

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 $p_{R|Y}\left(R\left|Y,\theta_D\right.\right)$. $\theta_D$, 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 $p_{\theta_D|\theta_p}\left(\theta_D\left|\theta_p\right.\right)$ allows for potential dependence of the design on the process parameters. The complete-data likelihood function[10] is then

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

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 superpopulation.

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 SDL is

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

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 $(\theta_p, \theta_D)$, given the observed data, which we assemble from the pieces defined above as

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

In general, we focus interest on the posterior distribution of $\theta_p$ which, in the absence of

[10] 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

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

The data inclusion model is *ignorable* if

$$p_{\theta_P|Y^{(obs)}R}\left(\theta_p \left| Y^{(obs)}, R\right.\right) \equiv p_{\theta_P|Y^{(obs)}}\left(\theta_p \left| Y^{(obs)}\right.\right). \tag{A.6}$$

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.

## A.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* $p_{Z|Y,R}\left(Z\left|Y, R, \theta_S\right.\right)$ with *SDL-parameter* vector $\theta_S$. The *SDL-parameter prior distribution* is $p_{\theta_S|\theta_D\theta_p}\left(\theta_S\left|\theta_D, \theta_p\right.\right)$. The likelihood function for the published data is

$$\pounds_\theta^{(pub)}\left(\theta_p, \theta_D, \theta_S\left|Z, R\right.\right) = \int p_{Z|YR}\left(Z\left|Y, R, \theta_S\right.\right) p_{YR}\left(Y, R\left|\theta_p, \theta_D\right.\right) dY \tag{A.7}$$
$$= \int p_{Z|YR}\left(Z\left|Y, R, \theta_S\right.\right) p_{R|Y}\left(R\left|Y, \theta_D\right.\right) p_Y\left(Y\left|\theta_p\right.\right) dY$$

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

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

where $p_\theta\left(\theta\right) = p_{\theta_S|\theta_D\theta_p}\left(\theta_S\left|\theta_D, \theta_p\right.\right) p_{\theta_D|\theta_p}\left(\theta_D\left|\theta_p\right.\right) p_{\theta_p}\left(\theta_p\right)$. So that the posterior distribution of the process parameters is

$$p_{\theta_P|ZR}\left(\theta_p\left|Z, R\right.\right) = \int\int p_{\theta|ZR}\left(\theta_p, \theta_D, \theta_S\left|Z, R\right.\right) d\theta_D d\theta_S. \tag{A.8}$$

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

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

That is, the posterior distribution of the process parameters $\theta_p$ 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

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

for all $Y^{(obs)}$, $Z$, and $R$.

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

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

for all $Y^{(obs)}$ and $Z$. Equation (A.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.

## A.4 Implementing SDL-aware Data Analysis

Since equation (A.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 $p_{\theta_P|ZR}\left(\theta_p\,|Z,R\right)$ can be assembled. We will define an SDL method as *fully discoverable* if $p_{\theta_P|ZR}\left(\theta_p\,|Z,R\right)$ can be computed. If the SDL process is not fully discoverable, then we will consider some diagnostic methods that can be used to approximate $p_{\theta_P|ZR}\left(\theta_p\,|Z,R\right)$ or to detect failures of equation (A.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

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

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

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

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 implementing 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 $\theta_D$ 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,

$$\begin{aligned}p_{\theta_P|ZR}\left(\theta_p\left|Z,R\right.\right) &= p_{\theta_P|Z}\left(\theta_p\left|Z\right.\right)\\ &= \int p_{\theta_P|Y^{(obs)}}\left(\theta_p\left|Y^{(obs)}\right.\right)p_{Y^{(obs)}|Z}\left(Y^{(obs)}\left|Z\right.\right)dY^{(obs)}\end{aligned}\tag{A.12}$$

$$p_{\theta_p\theta_D|Y^{(obs)}R}\left(\theta_p,\theta_D\left|Y^{(obs)},R\right.\right)=p_{\theta_p|Y^{(obs)}}\left(\theta_p\left|Y^{(obs)}\right.\right)\tag{A.13}$$

$$p_{\theta_S|ZR\theta_p\theta_D}\left(\theta_S\left|Z,R,\theta_p,\theta_D\right.\right)=p_{\theta_S|Z\theta_p}\left(\theta_S\left|Z,\theta_p\right.\right)\tag{A.14}$$

and

$$p_{Y^{(obs)}|ZR\theta_p\theta_D\theta_S}\left(Y^{(obs)}\left|Z,R,\theta_p,\theta_D,\theta_S\right.\right)=p_{Y^{(obs)}|Z\theta_p\theta_S}\left(Y^{(obs)}\left|Z,\theta_p,\theta_S\right.\right).\tag{A.15}$$

Estimation and inference using the SDL-aware system described by equations (A.12)-(A.15) 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 many examples where the inclusion model's properties are well understood or provably ignorable.

## A.5 Using Conditional Probability Models to Discover Nonignorable SDL

Consider the data model in which $y_i$ contains $K$ variables with $y_{i1}$ binary and the remaining variables either continuous or discrete. Although the formal data model remains $p_{y|\theta_P}\left(Y^{(obs)}\,|\theta_P\right)$, interest focuses on estimation and inference for the conditional probabilities

$$\Pr\left[y_{i1}=1\,|y_{i2},\beta\right]$$

where $\beta$ is the process parameter vector of interest (linear probability model coefficients, logit coefficients, probit coefficients, etc.). The remaining process parameters are nuisance parameters in an analysis with access to $Y^{(obs)}$. We consider here the use of conditional probability models as diagnostic tools for discovering nonignorable SDL.

If the analyst is completely ignorant of the process generating $y_i$ the SDL is not discoverable unless its details are published by the agency or it is generated by a formal privacy model with public parameters. The intuitive notion that information in related data can be used to discover SDL properties lies at the heart of the Alexander et al. (2010) analysis of the 2000 Census and ACS Public-Use Microdata Samples (PUMS). In those examples $y_{i1}$ is the individual's sex and $y_{i2}$ is the individual's birth date (or age). They (implicitly) use an informative prior on $\beta$ based on the population summary files for the 2000 Census (which are based on all records, not just the PUMS records) and the published tabulations for the ACS (which are based on the full ACS sample, not just the records in the PUMS) to estimate the effects of SDL on analyses using the PUMS files. In their case, the informative prior distribution was sufficient to estimate $\beta$ accurately because $\beta$ was actually a finite-population estimand (the proportion of the age cohort that is in each sex for the U.S. population at a point in time). In addition, because they used a finite-population estimand where the variability of $\beta$ in the prior distribution was negligible, they could assess the probability that the differences were due to chance from the posterior variability in the PUMS files alone. In general, this won't be the case, but the intuition underlying their method is more broadly applicable for discovering nonignorable SDL.

Conditional probability models analyzed using SDL-aware procedures with informative priors can render the SDL discoverable in both our formal sense and the intuitive sense used by Alexander et al. To develop this point formally, we can no longer assume that the inclusion process, in this case the sampling model, is ignorable because this process contributes to the posterior distribution of the process parameters and to an informative prior distribution on those parameters, but not necessarily in the same manner. In addition, we will need to be precise in making assumptions about the SDL. SDL processes used in related data publications may share parameters, random noise, and conditioning variables. We will have to be formal about conditioning on or integrating out these SDL components in the informative prior as well as in posterior of interest. The payoff is that we can get probability models that provide a formal basis for what Alexander et al. did and are more generally applicable. In our empirical examples, we are careful to select published data files where the dependencies in the SDL have been documented by the suppliers.

# B Details of Estimating Population Proportions with Noise Infusion

Suppose the confidential data, $y_i$, contain $K$ variables with $y_{i1}$ binary and the remaining variables either continuous or discrete. We are interested in estimation and inference for the conditional probabilities $\Pr[y_{i1} = 1 | y_{i2}, \beta]$, where $\beta$ is the parameter of interest. The problem arises from using $\Pr[z_{i1} = 1 | z_{i2}, \beta, \theta_S]$ where the $z_i$ variables are the published versions of $y_i$ and $\theta_S$ are the parameters of the SDL.

To facilitate the exposition, consider just one outcome $z_{i1}$, which can be either zero or one. For example, the observed $z_{i1}$ could be an indicator that the respondent is male and the conditioning set, $z_{i2}$ could be age 65. With probability $\rho$, the published data come from the same conditioning set as in the confidential data; that is, $z_{i2} = y_{i2}$. For example, if the stratification is on age, then with probability $\rho$, the observed outcome comes from the true age category; that is, $z_{i1} = y_{i1}$ for $y_{i2} = 1$ [true age $= 65$]. With the complementary probability, the observed outcome is a binary random variable with expected value $\mu \neq \beta$, for example, the average value of proportion male over all age categories at risk to be changed by the SDL model.

Under these conditions and using $\mathrm{E}[z_{i1} = 1 | z_{i2}, \beta, \rho, \mu]$ the consistent estimator for the process parameter of interest, $\beta$, is

$$\hat{\beta} = \frac{\bar{z}_1 - (1 - \rho)\mu}{\rho} \tag{B.16}$$

where $\bar{z}_1$ is the estimated sample proportion of ones (i.e., males). The estimator for the conditional proportion of interest $\hat{\beta}$ is confounded by the two SDL parameters, except in the special case that $\rho = 1$, which implies that none of the published age data has been infused with noise. If all of observations have been subjected to this noise infusion, then $\hat{\beta}$ is undefined, and the expected value of $\bar{z}_1$ is just $\mu$. In the starkest possible terms, the estimator in equation (B.16) is hopelessly underidentified in the absence of information about $\rho$ and $\mu$.

If $\rho$ and $\mu$ are not known, they may still be discoverable if the analyst has access to estimates of conditional probabilities like $\beta$ from an alternative source. Here is an example based on the analysis in ADS. Comparing the sex proportions estimated from the Census 2000 PUMS to the published Census 2000 data, and treating the published Census 2000 estimates as the true values, we have

$$\mathrm{E}[\bar{z}_{j1} - \bar{y}_{j1} | \bar{y}_{j1}] = \rho_j \bar{y}_{j1} + (1 - \rho_j)\mu_j - \bar{y}_{j1} \tag{B.17}$$

for $j =$ ages $65, 66, 67, \ldots 89$.

The SDL process is still underidentified if we consider only a single outcome like sex, but there are quite a few other binary outcomes that could also be studied, conditional on age, for example, marital status, race and ethnicity. The differences between Census 2000 estimates of the proportion married at ages 65 and greater and their comparable Census 2000 PUMS estimates have exactly the same functional form as equation (B.17) with exactly the same SDL parameters. Since these proportions condition on the same

age variable, all of the other outcomes that also have an official Census 2000 published proportion can be used to estimate $\rho_j$ and $\mu_j$. The identifying assumptions are: (1) all proportions are all conditioned on the same noisy age variable, and (2) the noisy age variable can be reasonably modeled as randomized-response noise.

### B.1 History of the Census Bureau's Correction to Census 2000 and ACS PUMS Files

The original announcement that the PUMS files would not be corrected can be found in Census 2000 Public Use Microdata Sample Data Note 12 (October 2010) and the reversal in Data Note 13 (October 2010) http://www.census.gov/prod/cen2000/doc/pums.pdf. The original announcement that the ACS PUMS files would not be corrected can be found in Errata 47 (February 18, 2010) and 50 (December 18, 2009). The reversal is in Erratum 65 (January 25, 2012). See also User note 3. Cited documents http://www.census.gov/acs/www/data_documentation/errata/#Err47, http://www.census.gov/acs/www/data_documentation/errata/#Err50, http://www.census.gov/acs/www/data_documentation/errata/index.php#Err65, and http://www.census.gov/acs/www/data_documentation/user_notes/index.php#n03(cited March 19, 2015).

## C Details of Estimating Regression Models with SDL

### C.1 Bias due to SDL in the Dependent Variable

For the case in which SDL is applied to the dependent variable, our derivation of the bias formula is a direct extension of the analysis in Sections 1.1, 1.2, and 1.3 in the Appendix to Bollinger and Hirsch (2006). Our only modification is to the equation characterizing the distribution of imputed data. In the Bollinger and Hirsch Appendix, equation (1) states

$$f_I\left(y_i, z_i | x_i\right) = f_O\left(y_i | x_i\right) f_M(z_i | x_i),$$

where $f_I\left(y_i, z_i | x_i\right)$ is the joint distribution of $y$ and $z$ in the imputed data, $f_O\left(y_i | x_i\right)$ is the distribution of the dependent variable, $y$, given the matching variables, $x$, among the observed data, and $f_M(z_i | x_i)$ is the distribution of the regressors, $z$, conditional on $x$, among the missing data.

In our application, $y$ is sometimes missing, not because it was not reported, but because it is suppressed. That means the imputed values can be drawn from the distribution of the suppressed data. Formally, this just amounts to changing the above equation to

$$f_I\left(y_i, z_i | x_i\right) = f_M\left(y_i | x_i\right) f_M(z_i | x_i).$$

This change does not affect the remaining derivations in sections 1.1–1.3 of the Bollinger and Hirsch Appendix; the bias formula remains the same. This is just a change of interpretation. Specifically, whereas Bollinger and Hirsch must make an assumption on

the missing data process (namely, that the data are conditionally missing at random), we require no such assumption on the suppression process.

The SDL is ignorable for estimation and inference of $\beta$ if the solution to the least squares projection of $z_{1i}$ on $z_{2i}$ yields estimates consistent for the parameters of the true regression model: $\mathrm{E}[y_{i1}|y_{i2}] = \alpha + y_{i2}\beta$. The solution to the least squares projection is $\left(\hat{a}, \hat{b}\right) = \arg\min_{a,b} \mathrm{E}\left[(z_{i1} - a - z_{2i}b)^2\right]$.

We start with the case of a single right-hand side variable, where the intuition is simpler. The regressor $z_{2i} = y_{2i}$ and the conditioning variables $x_i$ are scalar. Allowing SDL to be conditional on the suppression indicator, we follow the derivations in the unpublished Appendix to Bollinger and Hirsch (2006) to obtain the following result:

$$\operatorname{plim} \hat{b} = \beta - (1-\rho)\,\mu\beta = (1 - (1-\rho)\,\mu)\beta. \tag{C.18}$$

The bias term on the right hand side depends on two factors: the share of suppressed observations, $(1-\rho)$, and the error from using $x_i$ to impute the suppressed value instead of $z_{2i}$, measured by $\mu$. The term $\mu$ may be derived as follows. First, compute the residual from predicting the regressor with the conditioning variables: $e_i = z_{2i} - E(z_{2i}|x_i, \gamma_i = 0)$. Now $\mu$ is the slope parameter from the regression $e_i = \ell + \mu z_{2i}$. That is, $\mu$ measures the signal from the regressor $z_{2i}$ left in $e_i$ after conditioning on $x_i$ and $\gamma_i = 0$.

The same result holds for the more general case in which $z_{2i}$ and $x_i$ are vectors. Now

$$\operatorname{plim} \hat{b} = \beta - (1-\rho)\,M\beta = (I - (1-\rho)\,M)\beta, \tag{C.19}$$

Formally, $M$ is derived analogously to $\mu$. First, measure the vector of residuals from the system $e_i = z_{2i} - E(z_{2i}|x_i, \gamma_i = 0)$. Then, $M$ is the parameter matrix from estimating of $e_i = L + M z_{2i}$.

The case of general $z_i$ is similar, but the derivations are complicated and provide little intuition for the applications under consideration here. They are available upon request.

## C.2 Bias Due to SDL in a Single Regressor

The case in which SDL is applied to a single regressor turns out to be identical to the case of SDL applied to the dependent variable. That this is so may be intuitive when the regression model includes only one regressor. It is less transparent in the case of multiple regressors, so we present the relevant derivations here. For ease of presentation, we use notation similar to Bollinger and Hirsch.

The data vector is $(y_i, z_i, t_i, x_i, R_i)$. $y_i$ is the dependent variable, $t_i$ is the scalar variable to which SDL is applied, $z_i$ is a vector of regressors that are not distorted, $x_i$ is a vector of conditioning variables used to impute replacements for $t$ when it is suppressed, and $R_i$ is a variable equal to 1 if $t_i$ is suppressed, and 0 otherwise. Define the population distributions $f_O(y_i, z_i, t_i, x_i|R_i = 0)$ for observations with no suppression, and $f_M(y_i, z_i, t_i, x_i|R_i = 1)$ for observations where $t$ was suppressed and imputed. Also, let $p = \Pr[R_i = 1]$.

We make the following assumptions on the data generating process:

- Only $t_i$ is suppressed;
- the matching variables depend only on $z_i$ and $t_i$, $x_i = h(z_i, t_i)$;
- the researcher has the correct model, and so $x$ cannot provide any additional information:
$$E[y_i|z_i, t_i, x_i] = E[y_i|z_i, t_i] = \alpha + z_i^T\beta + \gamma t_i$$
- when $R_i = 1$, the published value $t_i$ is sampled from the distribution $f_M(t|x_i)$.

The conditional distribution of the suppressed data is

$$f_I(y_i, z_i, t_i|x_i) = f_M(y_i, z_i|x_i) f_M(t_i|x_i).$$

It follows the distribution of the published data is

$$f_S(y_i, z_i, t_i|x_i) = (1-p) f_O(y_i, z_i, t_i|x_i) + p f_M(y_i, z_i|x_i) f_M(t_i|x_i).$$

After some algebraic transformations, and taking expectations with respect to $z_i$,$t_i$, and $x_i$, we get the key moment equation characterizing the conditional expectation of $y_i$ given $(z_i, t_i, x_i)$ in the published data:

$$\begin{aligned} E_S[y_i|z_i, t_i, x_i] &= (1-p) E_O[y_i|z_i, t_i, x_i] \frac{f_O(z_i, t_i, x_i)}{f_S(z_i, t_i, x_i)} \\ &\quad + p E_M[y_i|z_i, x_i] \frac{f_M(z_i, x_i) f_M(t_i|x_i)}{f_S(z_i, t_i, x_i)}. \end{aligned}$$

Using the definition of $E_O[y_i|z_i, t_i, x_i]$ and adding and subtracting $p\gamma t \frac{f_M(z_i,x_i) f_M(t_i|x_i)}{f_S(z_i,t_i,x_i)}$,

$$\begin{aligned} E_S[y_i|z_i, t_i, x_i] &= \alpha + z_i^T\beta + \gamma t_i \\ &\quad - p\gamma [t - E_M(t_i|x_i)] \frac{f_M(z_i, x_i) f_M(t_i|x_i)}{f_S(z_i, t_i, x_i)}. \end{aligned}$$

We now show that the least-squares solution will not be consistent for the parameters of interest and derive the bias correction. In the published data, the least-squares solution to the regression of $y_i$ on $z_i$ and $t_i$ is

$$\arg\min_{a,b,c} E_S\left[\left(E_s[y_i|z_i, t_i, x_i] - \left(a + z_i^T b + c t_i\right)\right)^2\right];$$

that is

$$\arg\min_{a,b,c} \int \left(E_s[y_i|z_i, t_i, x_i] - \left(a + z_i^T b + ct\right)\right)^2 f_s(z_i, t_i, x_i)\, dz_i dt_i dx_i.$$

The first-order conditions for a minimum are given by:

$$\begin{aligned} &\alpha + E_S\left(z_i^T\right)\beta + \gamma E_s(t_i) - \gamma p E_I(t_i - E_M(t_i|x_i)) \\ &\quad - \left(a + E_S\left(z_i^T\right) b + c E_s(t_i)\right) = 0 \end{aligned}$$

from differentiating with respect to $a$,

$$\begin{aligned}\alpha E_S\left(z_i\right)+E_S\left(z_i z_i^T\right)\beta+\gamma E_s\left(z_i t_i\right)-\gamma p E_I\left(z_i\left(t_i-E_M\left(t_i|x_i\right)\right)\right)\\ -\left(aE_S\left(z_i\right)+E_S\left(z_i z_i^T\right)b+cE_s\left(z_i t_i\right)\right)=0\end{aligned}$$

from differentiating with respect to $b$, and

$$\begin{aligned}\alpha E_S\left(t_i\right)+E_S\left(t_i z_i^T\right)\beta+\gamma E_s\left(t_i^2\right)-\gamma p E_I\left(t_i\left(t_i-E_M\left(t_i|x_i\right)\right)\right)\\ -\left(aE_S\left(t_i\right)+E_S\left(t_i z_i^T\right)b+cE_s\left(t_i^2\right)\right)=0\end{aligned}$$

from differentiating with respect to $c$.

In the case where $t_i$ is the only regressor ($z$ is identically zero), it is easy to show

$$c=\gamma\left\{1-p\left[E_I\left[t_i\left(t_i-E_M\left(t_i|x_i\right)\right)\right]-E_I\left[t-E_M\left(t_i|x_i\right)\right]E_s\left(t_i\right)\right]\right\}.$$

By inspection, this is identical to the formula for the case in which SDL is applied to the dependent variable.

# D Details of the RD Model

## D.1 Generalized Randomized Response SDL

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. 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 $\text{N}\left(0,\delta^2\right)$.

## D.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 $\rho_0$ and $\delta_0$, we have

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

and

$$\begin{aligned}\mathrm{E}\left[y_{i4}\left|z_{i3},\rho_0,\delta_0\right.\right] &= \mathrm{E}\left[1\left[y_{i3}\geq\tau\right]\left|z_{i3},\rho_0,\delta_0\right.\right] \\ &= \rho_0\,1\left[z_{i3}\geq\tau\right]+\left(1-\rho_0\right)\Phi\left(\frac{z_{i3}-\tau}{\delta_0}\right)\end{aligned} \tag{D.20}$$

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 $\theta_{RD}$ 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

$$\mathrm{E}\left[\theta_{RD}\left|Z,\rho_0,\delta_0\right.\right]=\mathrm{E}\left[\lim_{y_{i3}\downarrow\tau}\mathrm{E}\left[y_{i2}\left|y_{i3}=\tau\right.\right]\left|Z,\rho_0,\delta_0\right.\right] \tag{D.21}$$

$$-\mathrm{E}\left[\lim_{y_{i3}\uparrow\tau}\mathrm{E}\left[y_{i1}\left|y_{i3}=\tau\right.\right]\left|Z,\rho_0,\delta_0\right.\right] \tag{D.22}$$

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

and

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

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

$$(1-\rho_0)\left\{\mathrm{E}\left[\lim_{z_{i3}\downarrow\tau} f_2\left(z_{i3}-\varepsilon_i\right)|Z,\gamma_i=0,\delta_0\right]-\mathrm{E}\left[\lim_{z_{i3}\uparrow\tau} f_1\left(z_{i3}-\varepsilon_i\right)|Z,\gamma_i=0,\delta_0\right]\right\} \quad \text{(D.23)}$$

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 $\tau$–the continuous function $f_1\left(z_{i3}\right)$ transitions smoothly to $f_2\left(z_{i3}\right)$ over the support of $\varepsilon_i$. The SDL noise needn't be normal, but it must be drawn from a continuous distribution.

### D.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

$$\begin{aligned} h(z_i) &= \mathrm{E}\left[t_i\,|z_i,\rho_0,\delta_0\right] && \text{(D.24)}\\ &= \rho_0 p_{T|R}\left(t_i=1|z_i\right)+\left(1-\rho_0\right)\int p_{T|R}(t_i=1|r_i)p_{R|Z}(r_i|z_i)dr. && \text{(D.25)}\end{aligned}$$

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 $\tau$. 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.

# E Details for Tabular Methods

## E.1 Swapped Household Data

This part of our discussion of tabular data that applies only to tabulations based on household data. The tables produced from the decennial censuses and the American Community Survey are based on swapped input data. The effects of swapping can be assessed using the methods we discussed in Technical Appendix Section B. The condition for discoverable consequences of swapping, without the cooperation of the data provider, requires getting at least two tabulations that cover the same subpopulation, have known sampling variation, and use independent SDL models. The best general diagnostic we can derive is to perform simulations under the assumption that the contribution to the

posterior variance of a parameter of interest due to swapping is less than the contribution due to edit and imputation. If an agency were to state in its published documentation that this hypothesis was correct, then it might be worth unleashing the full posterior simulation technology.

## E.2 Custom Tabulations

An agency's officially tabulated estimates are those listed in the defined data products for the agency's publications. If the tabulation isn't listed in the defined data products, then an official estimate of that item is called a custom tabulation. All custom tabulations (also called special tabulations) are done sequentially, then released to the general public. The suppression rules applied to the official tabulations carry over to the first custom tabulation, and then to all successive custom tabulations. The effects are order dependent and cumulative. If an item was explicitly suppressed from any previous official or custom tabulation, then it will be suppressed for all future tabulations as well. This statement applies to both primary and complementary suppressions. Some agencies will not produce custom tabulations as a matter of policy. It is also worth pointing out that not all suppressions are due to SDL. There are also minimum data quality standards that can result in a suppression. These do not always cause additional complementary suppressions, but they do always cumulate. The data quality cannot be improved by calling it a custom tabulation.

## E.3 Directly Tabulating Published Microdata

After data collection has ended, the raw survey, census or administrative-record data are edited, imputed for missing data, weight-corrected, and subjected to SDL. Only then are the publication tables generated. The statistical agencies consider any released public-use microdata samples to also be publication tables. In general, if a researcher computes an estimate of a moment or quantile from the public-use microdata, then compares that estimate to its published equivalent in the tabular summaries, those two estimates will not agree exactly. Assuming that the correct selection criteria and weights were used, there remain three reasons why these calculations don't match. The first possible cause is differences in the computational formulas used. The second possible cause is sampling variability.

The third possible cause is SDL. The Census Bureau, for example, applies additional swapping and noise infusion to the publication-ready ACS records before selecting the PUMS. Public-use files produced by the Statistics of Income Division of the IRS are also subjected to extensive SDL beyond what is used for the tabular summaries.

By far, the most important SDL explanation for a discrepancy is that the equivalent tabular estimate is suppressed. A researcher can calculate some estimate of interest from the public-use microdata file, but the agency didn't release an official tabulation of the same item for several reasons. Some researchers may not consider suppressions in official tabulations to be a discrepancy with respect to estimates produced from the public-use microdata files, but there is an important sense in which they are. The microdata files are produced so that researchers can perform analyses that are not possible using the

aggregated tabulations. If there were no microdata files, as we will see is usually the case for the aggregated business data discussed in Section 5.2.2, then any research design would require a strategy for handling the missing data caused by the suppression.

Even when public-use microdata are available, suppression is still a problem. If a substantial proportion of the estimates a researcher computes from the microdata files correspond to suppressed official tabulations, it is a warning sign that the inputs to the researcher's statistical model may be of poor quality. Household tabular estimates are suppressed most often when the number of households in the cell is below the publication threshold. The statistical agency considers that item to be poorly estimated in the underlying confidential data. Furthermore, these are exactly the cells most likely to contain edit, imputation, and SDL-induced noise. A good research strategy is to consider pooling those estimates, for example by using a shrinkage estimator that averages a specific moment with a pooled estimate of the same moment.

## E.4 Tabular Regression Models with Nonignorable SDL

The noise infusion in QWI may be nonignorable. Univariate regression of a variable, say from another dataset, onto a QWI aggregate, provides a simple illustration. Suppose the part of the process model of interest is:

$$\mathrm{E}\left[Y_{(k)t} \left| W_{(k)t}\right.\right] = \alpha + \beta W_{(k)t} \tag{E.26}$$

where $W_{(k)t}$ is the quarterly payroll in county $k$ and $Y_{(k)t}$ is any outcome of interest collected from a different data source. $Y_{(k)t}$ can also be subject to SDL, but we will assume that it is statistically independent of the SDL applied to the QWI data. The published aggregate data are $[Y_{(k)t}, W^*_{(k)t}]$. The undistorted values, $W_{(k)t}$ are confidential.

The probability limit of the OLS estimator for $\beta$ based using the published data is

$$p\lim \hat{\beta}_{OLS} = \frac{\mathrm{Cov}\left[Y_{(k)t}, W_{(k)t}\right]}{\mathrm{Var}\left[\delta_j\right]\mathrm{E}\left[W^2_{(k)t}H^W_{(k)t}\right] + \mathrm{Var}\left[W_{(k)t}\right]} \tag{E.27}$$

The term $\mathrm{E}\left[W^2_{(k)t}H^W_{(k)t}\right]$ is the expected Herfindahl index for payroll within aggregate $k$, as derived in the Data Appendix E.5. The noise infusion is clearly nonignorable in this setting. Algebraic manipulation reveals the bias to be

$$p\lim \frac{\hat{\beta}_{OLS}}{\beta} = \frac{\mathrm{Var}\left[W_{(k)t}\right]}{\mathrm{Var}\left[\delta_j\right]\mathrm{E}\left[W^2_{(k)t}H^W_{(k)t}\right] + \mathrm{Var}\left[W_{(k)t}\right]}. \tag{E.28}$$

The bias factor lies between 0 and 1.

One option is to correct the bias analytically. If $\mathrm{Var}\left[\delta_j\right]$ is known, or can be estimated, the bias can be corrected directly. An unbiased estimate for $E[W_{(k)t}]^2$ is available from $E[W^*_{(k)t}]^2$ once $\mathrm{Var}\left[\delta_j\right]$ is known, after which it only remains to recover $\mathrm{Var}\left[W_{(k)t}\right]$ from the definition of $\mathrm{Var}\left[W^*_{(k)t}\right]$.

The second possibility is to find instruments. Any instrument, $Z_{(k)t}$, correlated with $W_{(k)t}$ and uncorrelated with the SDL noise infusion process will work, since

$$p\lim \hat{\beta}_{IV} = \frac{\text{Cov}\left[\alpha + \beta W_{(k)t} + \varepsilon_{(k)t}, Z_{(k)t}\right]}{\text{Cov}\left[W^*_{(k)t}, Z_{(k)t}\right]} = \frac{\beta \,\text{Cov}\left[W_{(k)t}, Z_{(k)t}\right]}{\text{Cov}\left[W_{(k)t}, Z_{(k)t}\right]} = \beta. \tag{E.29}$$

## E.5 Details of Estimating the Variance Contribution of SDL for the QWI

It is possible to recover the variance of the noise factor $\text{Var}\left[\delta_j\right]$, which is needed to correct directly for bias in the univariate and multivariate regression examples using the QWI. The noise in a magnitude estimate from a particular cell and the confidential magnitude value are independent by construction. By design, there is no bias:

$$\text{E}\left[W^*_{(k)t} - W_{(k)t} \,\middle|\, W_{(k)t}\right] = \text{E}\left[\sum_{j \in \Omega_{(k)t}} W_{jt}\left(\delta_j - 1\right) \,\middle|\, W_{(k)t}\right] = 0, \tag{E.30}$$

where the last equality results from the independence of $W_{jt}$ and $\delta_j$ for all $t$. This is a common feature of noise-infusion SDL. The designers eliminated the bias in published tabulations. However, this was accomplished by inflating the variance of the published aggregate. The exact formula for the variance in the difference between noisy and noise-free estimates of is

$$\text{V}\left[W^*_{(k)t} - W_{(k)t} \,\middle|\, W_{(k)t}\right] = \text{V}\left[\delta\right] \sum_{j \in \Omega_{(k)t}} W^2_{jt}. \tag{E.31}$$

Our leverage in this analysis comes from the fact that QWI and QCEW use identical frames (QCEW establishments). Hence, we can use $W^{(QCEW)}_{(k)t}$ as the noise-free estimate of $W_{(k)t}$, as long as it has not been suppressed too often.

Although the data come from a different administrative record system, the concepts underlying the CBP payroll variable are very similar to both the QWI and QCEW inputs. The SDL system used for CBP data is very similar to the one used for QWI but the random noise in CBP is independent of the random noise in QWI. The formulas for recovering both systems SDL parameters are in the Data Appendix Section H.

**Data Appendix for "Economic Analysis and Statistical Disclosure Limitation"**

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

**August 7, 2015**

# F Variables from the QWI, QCEW, and CBP

We work with several variables from the QWI data:

- $B_{jt} \equiv$ employment at establishment $j$ at the beginning of quarter $t$ (record-linkage definition; first calendar day of the quarter)
- $W_{jt} \equiv$ total quarterly payroll for all statutory employees during the quarter (state unemployment insurance system definition)

From the QCEW, we use the following variables, which are analogous to the employment and payroll variables reported in the QWI.

- $E_{jt}^{(QCEW,1)} \equiv$ month 1 employment (on the payroll for the pay period covering the $12^{th}$ day of the first month of the quarter)
- $E_{jt}^{(QCEW,3)} \equiv$ month 3 employment (on the payroll for the pay period covering the $12^{th}$ day of the third month of the quarter)
- $W_{jt}^{(QCEW)} \equiv$ total quarterly payroll for all statutory employees during the quarter (state unemployment insurance system definition)

From the CBP, we use the following variables, which are analogous to the employment and payroll variables in the QWI.

- $L_{jt} \equiv$ employment (on the payroll for the pay period covering the March $12^{th}$)
- $P_{jt} \equiv$ total first-quarter payroll for all statutory employees (Federal Insurance Contributions Act (FICA) definition)

# G Statistical Disclosure Limitation Methods

## G.1 QWI

The QWI SDL system is based on multiplicative input noise infusion applied to all variables used to compute tabular magnitude estimates. These include employment and

payroll, of course, but also hires, separations, job creations, job destructions, and similar statistics for stocks and flows based on stable employment definitions.

A random fuzz factor, $\delta_j$, is drawn for each establishment, $j$, from a double-ramp distribution with the following probability distribution function:

$$p(\delta_j) = \begin{cases} 0, \ \delta < 1-b \\ (1+b+\delta-2)/(b-a)^2, \ \delta \in [1-b, 1-a] \\ 0, \ \delta \in (1-a, 1+a) \\ (1+b-\delta)/(b-a)^2, \ \delta \in [1+a, 1+b] \\ 0, \ \delta > 1+b \end{cases} \tag{G.32}$$

and associated density

$$F(\delta_j) = \begin{cases} 0, \ \delta < 1-b \\ (\delta+b-1)^2/\left[2(b-a)^2\right], \ \delta \in [1-b, 1-a] \\ 0.5, \ \delta \in (1-a, 1+a) \\ 0.5 + \left[(b-a)^2 - (1+b-\delta)^2\right]/\left[2(b-a)^2\right], \ \delta \in [1+a, 1+b] \\ 1, \ \delta > 1+b \end{cases} \tag{G.33}$$

The values $0 < a < b < 1$ are parameters chosen such that each establishment's value of the statistic is distorted by a minimum of $100a$ percent and a maximum of $100b$ percent.[11] Thus $\delta_j$ has the following properties:

$$\mathrm{E}[\delta_j] = 1$$

and

$$\mathrm{V}[\delta_j] = a^2 + \frac{1}{6}(b-a)^2 + \frac{2}{3}a(b-a).$$

The probability distribution of $\delta_j$ is plotted in Figure H.1a on page 67 and the cumulative distribution is plotted in Figure H.1b on page 67 for the values $a = 0.05, b = 0.3$, which were chosen for illustrative purposes only. Note, the distribution of $\delta_j$ is independent of all other variables. The SDL system is implemented so that an establishment is assigned a value of $\delta_j$ at the time it first enters the database. The establishment retains the assigned value until it disappears from the data permanently.

## G.2 SDL for the QCEW

The QCEW data use a primary/complementary suppression system for SDL. The only public information about this system appears in Statistical Policy Working Paper 22:

"For example, the Quarterly Census of Employment and Wages (QCEW), a census of monthly employment and quarterly wage information from Unemployment Insurance filings, uses a threshold rule and the $p$ percent rule for calendar year (CY) 2002 data and beyond. Prior to CY 2002, QCEW used a threshold rule and a concentration rule of $(n, k)$. In a few cases, a two-step rule is used–an $(n, k)$ rule for a single establishment is followed

[11] The exact percentage distortions are Census Bureau confidential.

by an $(n, k)$ rule for two establishments." (Harris-Kojetin et al. (2005), page 47)

The BLS quantifies the amount of suppression with the following statement:

"The finest level of geographic detail is the county-industry level, as aggregates of establishments classified to varying degrees of industry detail. While the input data are coded with meaningful address locations, the data are generally unavailable at greater detail. The QCEW program is constrained by the need to protect the confidentiality of data provided by employers, and richer geographic detail would threaten that confidentiality. Even the county by industry data cited above is at the margin of being disclosable-approximately 60 percent of the most detailed level data are suppressed for confidentiality reasons." (http://www.bls.gov/cew/cewfaq.htm)

The only public detail of the complementary suppression algorithm is that it does not include table margins (unless the margin itself fails the primary suppression rule):

"However, published totals of higher-level aggregations, when disclosed, include the suppressed lower-level data." (http://www.bls.gov/cew/cewfaq.htm)

The public QCEW data are not rounded.

## G.3 SDL for County Business Patterns

CBP uses noise infusion that is similar in the cross-section to the method used by QWI. There are also primary suppressions when the number of establishments in a cell is deemed too small to allow publication and when the value of the cell was distorted by more than five percent. The official specification of the noise infusion system is quoted here from the CBP documentation.

"County Business Patterns continues to apply the Noise Infusion method of data protection that began in 2007. Noise infusion is a method of disclosure avoidance in which values for each establishment are perturbed prior to table creation by applying a random noise multiplier to the magnitude data (i.e., characteristics such as first-quarter payroll, annual payroll, and number of employees) for each company. Disclosure protection is accomplished in a manner that results in a relatively small change in the vast majority of cell values. Each published cell value has an associated noise flag, indicating the relative amount of distortion in the cell value resulting from the perturbation of the data for the contributors to the cell. The flag for 'low noise' (G) indicates the cell value was changed by less than 2 percent with the application of noise, and the flag for 'moderate noise' (H) indicates the value was changed by 2 percent or more but less than 5 percent. Cells that have been changed by 5 percent or more are suppressed from the published tables. Additionally, other cells in the table may be suppressed for additional protection from disclosure or because the quality of the data does not meet publication standards. Though some of these suppressed cells may be derived by subtraction, the results are not official and may differ substantially from the true estimate. The number of establishments in a particular tabulation cell is not considered a disclosure; therefore, this information may be released without the addition of protective noise." (http://www.census.gov/econ/cbp/methodology.htm, citing Evans et al. (1998)).

Tabular cell magnitudes for $L_{jt}$ and $P_{jt}$ in CBP are computed using a multiplicative fuzz factor from a ramp distribution as in equations (G.32) and (G.33) with confidential

parameters. The factor $\delta_{jt}^{(CBP)}$ is drawn fresh for each establishment every year. The same fuzz factor is applied to all values from an establishment. Census uses a "balancing" algorithm to reduce the amount of noise in a particular cell of CBP. In this context, balancing means that the conditional distribution of $\delta_{jt}^{(CBP)}$ is not independent of the values of $L_{jt}$ or $P_{jt}$ (depending upon which variable has been used to balance, which is not disclosed). CBP also uses non-standard establishment level rounding to ensure the protection of noise infusion for cells with a small number of small establishments. Employment size class distributions are tabulated from unfuzzed employment data.

The published CBP data on payroll are rounded to the nearest thousand dollars, which is also an SDL. Published employment is not further rounded from the record-level edits.

# H Discovery of SDL parameters in QWI data

It is the differences in data construction that give rise to our strategy for revealing features of the SDL applied in each source. Therefore, it is necessary to discuss in some detail how each data source constructs and reports aggregate summaries from the underlying microdata.

**QWI variable construction:** Aggregates are formed over a classification $k = 1, \ldots, K$ that partitions the universe of establishments $\Omega_t$ into $K$ mutually exclusive and exhaustive subsets $\Omega_{(k)t}$. These partitions usually have detailed geographic and industrial dimensions. For all three data sources, geography is coded using FIPS county codes. Industrial classifications are by NAICS sectors, sub-sectors, and industry groups.

The tabular magnitudes are computed by aggregating the values over the establishments in the partition $k$. In the QWI, total private employment in a state is benchmarked to the month-1 employment from QCEW using an establishment weight, $\omega_{jt}$. In the absence of SDL, the beginning-of-quarter employment in $k$ would be estimated by

$$B_{(k)t} = \sum_{j \in \Omega_{(k)t}} \omega_{jt} B_{jt}.$$

The QCEW does not use weights. The comparable employment magnitudes for months 1 and 3 are

$$E_{(k)t}^{(QCEW,1)} = \sum_{j \in \Omega_{(k)t}} E_{jt}^{(QCEW,1)}$$

and similarly for $E_{(k)t}^{(QCEW,3)}$.

**SDL through multiplicative noise infusion:** Published aggregates from the QWI are computed using the multiplicative noise factors $\delta_j$. Beginning-of-quarter employment is computed as

$$B_{(k)t}^{*} = \sum_{j \in \Omega_{(k)t}} \delta_j \omega_{jt} B_{jt},$$

where we have adopted the convention of tagging the post-SDL value with an asterisk.

Similarly, the unprotected and protected values of total payroll in QWI are computed as

$$W_{(k)t} = \sum_{j \in \Omega_{(k)t}} \omega_{jt} W_{jt}$$

and

$$W^*_{(k)t} = \sum_{j \in \Omega_{(k)t}} \delta_j \omega_{jt} W_{jt}.$$

Notice that the same weight and the same fuzz-factor are used to aggregate total payroll and beginning-of-quarter employment (and, in fact, for all of the QWI).

**QCEW variable construction:** The total payroll variable in the QCEW is computed as

$$W^{(QCEW)}_{(k)t} = \sum_{j \in \Omega_{(k)t}} W^{(QCEW)}_{jt}.$$

To implement the SDL system for the QCEW, order statistics for the employment and payroll variables from the establishments in $\Omega_{(k)t}$ are used to compute the $p$-percent primary suppression rule. The partition size, $\left|\Omega_{(k)t}\right|$, is used to compute cell size thresholds, when they are used for suppression. As noted above, the formulas for the complementary suppressions have not been published.

**Comparability of QWI and QCEW data:** If the only partition is geography at the state level, so $k$ indexes states, then the QWI benchmarking ensures that

$$B_{(k)t} = E^{(QCEW,1)}_{(k)t}. \tag{H.34}$$

We note for clarity that equation (H.34) holds only for beginning-of-quarter employment at the state level for all private employers, and not for any other variable or aggregation. In addition, the benchmarking is performed using the confidential inputs before SDL; hence, it does not hold exactly for the published values.

CBP data are not weighted. Hence, the published values are computed as

$$L^*_{(k)t} = \sum_{j \in \Omega_{(k)t}} \delta^{(CBP)}_{jt} L_{jt}$$

and

$$P^*_{(k)t} = \sum_{j \in \Omega_{(k)t}} \delta^{(CBP)}_{jt} P_{jt}.$$

$\left|\Omega_{(k)t}\right|$ is used to compute cell sizes for primary suppressions. The criteria for suppression based on "data quality" have not been published. There are no complementary suppressions. When computing the bins for the employment size distribution of establishments, $L_{jt}$ is used without fuzzing.[12]

[12]Some details of the CBP protections are taken from an non-confidential presentation by Richard Moore, 2010, available from the authors.

## H.1 Estimating the Variance Contribution of SDL for the QWI

In QWI, the variance in the difference between noisy (published) and noise-free estimates of start-of-quarter employment, $B_{(k)t}$, conditional on the noise-free estimates, is

$$\begin{aligned} \mathrm{V}\left[B^*_{(k)t} - B_{(k)t} \,\middle|\, B_{(k)t}\right] &= \mathrm{E}\left[\sum_{j\in\Omega_{(k)t}} \omega_{jt}\left(B_{jt}\left(\delta_j - 1\right)\right)^2 \,\middle|\, B_{(k)t}\right] \\ &= \mathrm{V}\left[\delta\right] \sum_{j\in\Omega_{(k)t}} \omega_{jt} B^2_{jt} \end{aligned} \tag{H.35}$$

with a similar formula for $W_{(k)t}$.

If information about the properties of the size distribution of employment are available for the classification represented by $\Omega$, then equation (H.35) can be re-expressed as

$$\mathrm{V}\left[B^*_{(k)t} - B_{(k)t} \,\middle|\, B_{(k)t}\right] = \mathrm{V}\left[\delta\right] B^2_{(k)t} H^{(B)}_{(k)t}$$

where $H^{(B)}_{(k)t}$ is the Herfindahl index of employment shares of establishments in category $k$. It is straightforward to derive an equivalent equation for the variance of the difference between the published and noise-free payroll totals, conditional on the noise-free level, which depends on $H^{(W)}_{(k)t}$, the Herfindahl index of payroll shares of establishments in category $k$. Dividing both sides of Equation (H.35) by the square of the noise-free estimate and taking positive square roots yields

$$\sqrt{\frac{\mathrm{V}\left[B^*_{(k)t} - B_{(k)t} \,\middle|\, B_{(k)t}\right]}{B^2_{(k)t}}} \equiv \mathrm{CV}\left[B^*_{(k)t} - B_{(k)t} \,\middle|\, B_{(k)t}\right] \tag{H.36}$$

$$= \sqrt{\mathrm{V}\left[\delta\right] \frac{\sum_{j\in\Omega_{(k)t}} \omega_{jt} B^2_{jt}}{B^2_{(k)t}}} \tag{H.37}$$

$$= \sqrt{\mathrm{V}\left[\delta\right] H^{(B)}_{(k)t}}$$

with a similar formula for $W^*_{(k)t}$.

### H.1.1 Empirical Specification

Taking logarithms of Equation (H.36) yields the estimating equation

$$\ln \mathrm{CV}\left[B^*_{(k)t} - B_{(k)t} \,\middle|\, B_{(k)t}\right] = \frac{1}{2}\ln \mathrm{V}\left[\delta\right] + \ln\sqrt{H^{(B)}_{(k)t}}. \tag{H.38}$$

The dependent variable is defined in terms of the noise-free estimates, which are confidential in the QWI system. Fortunately, we have access to noise-free variables from the

published QCEW and CBP data. We assume

$$\mathrm{CV}\left[B^*_{(k)t} - B_{(k)t} \middle| B_{(k)t}\right] = \mathrm{CV}\left[B^*_{(k)t} - E^{(QCEW,1)}_{(k)t} \middle| E^{(QCEW,1)}_{(k)t}\right]$$

and

$$\mathrm{CV}\left[W^*_{(k)t} - W_{(k)t} \middle| W_{(k)t}\right] = \mathrm{CV}\left[W^*_{(k)t} - W^{(QCEW)}_{(k)t} \middle| W^{(QCEW)}_{(k)t}\right].$$

Both assumptions are justified by the fact that the noise factors in the QWI are completely independent of the underlying data. Furthermore, all three sources use identical frames and identical input data sources to measure the same variables in the same manner. The concepts used to measure $B_{(k)t}$ in QWI were constructed to approximate as closely as possible first month employment in the QCEW. Furthermore, the QWI establishment weights force the private state-level aggregate $B_{(k)t}$ to match exactly its QCEW counterpart.

The key explanatory variable in Equation H.38 is based on the Herfindahl index over employment shares, $H^{(B)}_{(k)t}$. This can be computed directly when size class information about the distribution within $B_{(k)t}$ and $W_{(k)t}$ is available, as is the case in the CBP data. Alternatively, we model this term as a power law (Cobb-Douglas) function of the number of establishments used to form the cell $k$

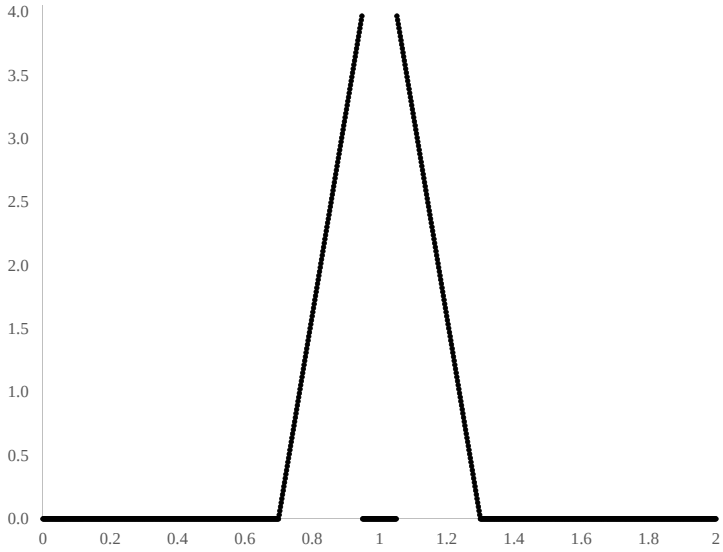
$$H^{(B)}_{(k)t} = \frac{\sum_{j\in\Omega_{(k)t}} \omega_{jt} B^2_{jt}}{B^2_{(k)t}} = \alpha_{(k)} N^{\beta_{(k)}}_{(k)t}, \tag{H.39}$$

where the scaling coefficient $\alpha_{(k)}$ is a potential confounder for the estimation of $\mathrm{V}\left[\delta\right]$.

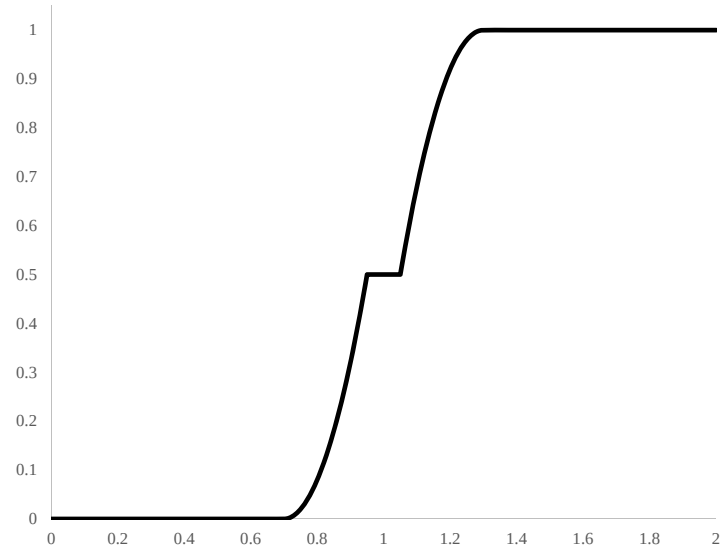
When no size-class information is available, substitution of (H.39) gives the estimating equation:

$$\ln \mathrm{CV}\left[B^*_{(k)t} - E^{(QCEW,1)}_{(k)t} \middle| E^{(QCEW,1)}_{(k)t}\right] = \frac{1}{2}\ln \mathrm{V}\left[\delta\right] + \ln \alpha_{(k)} + \frac{\beta_{(k)}}{2}\ln N_{(k)t}. \tag{H.40}$$

Equation (H.40) is a smooth function of the logarithm of the number of establishments used to form the table cell with estimates $B^*_{(k)t}$ and $W^*_{(k)t}$. Data on the number of establishments in each cell is reported in the QCEW. The derivation of the estimating equation for the coefficient of variation in payroll is identical.




(a) PDF




(b) CDF

Figure H.1: Distribution of Establishment Noise for the Quarterly Workforce Indicators