TITLE

Consistent Estimation with Deliberate Measurement Error to Protect Subject Confidentiality: An Application to Perturbed Location Data

AUTHORS

David Canning
Department of Global Health and Population
Harvard T.H. Chan School of Public Health
665 Huntington Ave, SPH 1, 1211
Boston, MA, 02115, USA
Tel: +1-617-432-6336
E-mail: dcanning@hsph.harvard.edu

Mahesh Karra (Corresponding Author)
Frederick S. Pardee School of Global Studies
Boston University
152 Bay State Rd, Room G04C
Boston, MA 02215
Tel: +1-703-969-1183
E-mail: mvkarra@bu.edu

ABSTRACT

Measurement error is often deliberately added to data in order to protect the confidentiality of human subjects. For example, variables such as birth date and district of residence are often coarsened, and sensitive survey questions may be asked in a way that deliberately induces noise to encourage truthful reporting. We focus on measurement error in spatial data, where a perturbation vector consisting of a random distance at a random angle is added to geographic coordinates. The use of perturbed data as explanatory variables in a regression model will generally make naïve estimates of all parameters biased and inconsistent.

We develop a general method for unbiased and consistent estimation for cases in which an explanatory variable is deliberately reported with error. Our method replaces the mismeasured variable with the expectation of the true variable and relies on knowledge of both the form of the measurement error and the underlying distribution of the true variable. We show how our method can be applied to several examples and conduct a Monte Carlo simulation exercise on an artificial dataset to show how replacing mismeasured distances to an exposure with expected distances using numerical integration over all possible true locations yields unbiased and consistent parameter estimates.

KEY WORDS

Measurement error; Data privacy; Demographic and Health Surveys (DHS); Numerical integration; Monte Carlo simulation

ABBREVIATIONS

DHS: Demographic and Health Survey

OLS: Ordinary Least Squares

1. INTRODUCTION

It is well known that measurement error in an explanatory variable in a regression usually biases the estimate of its coefficient towards zero (attenuation bias) and can render the coefficient estimates on all variables inconsistent. Usually, measurement error is inadvertent and its structure is unknown. Various approaches have been proposed to dealing with such measurement error, including regression calibration (Hardin et al. 2003; Spiegelman et al. 1997) and maximum likelihood methods (Rabe-Hesketh et al. 2003a, b,p.). In some instances, however, measurement error is deliberately added to reported data in order to prevent respondents from being identified and to protect their confidentiality. For example, variables such as a respondent's age are often coarsened such that the year and month of birth, but not the day, are reported. Similarly, a dataset may report a household's state of residence but not a smaller sub-unit (e.g. district, township, village), which together with other information in the dataset may identify the respondent. In some cases the measurement error may be built into the way the data is collected. The randomized response technique, which generates a random response by the respondent, has been used for questions on sensitive topics, such as sexual activity, where people may not want to reveal the information to the interviewer. For example, a respondent can be asked to throw a die privately and respond "no" on a one, and "yes" on a six, but answer the question truthfully on other numbers. Previous studies have proposed how to analyze such a randomized response as the dependent variable in a regression (Blair et al. 2015); our method shows how to such mismeasured responses as explanatory variables.

The motivating example that we examine is the case of adding perturbations to spatial location data in the Demographic and Health Surveys (DHS), which are nationally representative cross-sectional surveys that cover a range of health topics (USAID and ICF Macro International 2014). To protect the identity of interviewed households, the DHS masks the precise location data that are collected as part of the survey (Burgert et al. 2013; Perez-Heydrich et al. 2013). Reported location coordinates in the DHS are perturbed by adding a randomly generated distance, at a randomly generated angle, to the true location. This perturbed location data is frequently used by researchers to generate distance measures, for example, by measuring the minimum distance from the displaced respondent location to the known locations of health facilities or to other locations such as roads, rivers, schools, or markets. In such cases, distances that are generated from the perturbed location data will be measured with error.

Monte Carlo simulation studies of the effect of using perturbed location rather than actual locations to measure distance to an exposure have confirmed a large downward bias in parameter estimates (Arbia et al. 2015), and similar results have been found in a study where the actual location data is known and results using the actual locations are compared to those when using perturbed locations (Elkies et al. 2015). Adjusting for this bias may be important for our understanding of some important mechanisms. For example, some health studies that used perturbed location data frequently found that measured distance to the nearest health facility was not significantly associated with child mortality (Lohela et al. 2012), while other studies that used actual distances find significant effects (Schoeps et al. 2011; Karra et al. 2016).

While the structure of the measurement error that is added to location data is known, it is very complex, thereby making it difficult to recover good estimates of the actual location and distances to exposures. For example, while it can be shown that the expected measurement error in distances has positive mean and is bounded when measuring the distance to a fixed point (Elkies et al. 2015), when measuring distance to the nearest of a set of facilities, it is possible that the "nearest" facility may change when noise is added to the respondent location. If we assume that both the distributions of actual distances, and of the measurement errors in distances that are induced by using perturbed locations are independent and normally distributed, then we can apply the regression calibration method (Warren et al. 2016). However, the assumption of normality is generally not true (Arbia et al. 2015; Elkies et al. 2015).

Rather than deal with a specific case, we develop a theory that allows unbiased and consistent estimation of any linear regression model in which we have a known distribution of measurement error in an explanatory variable. In particular, we assume that the structure of the measurement error (the algorithm or process that was used to generate the error) is known to the researcher. We begin by showing that we can consistently estimate the linear regression so long as we can first obtain the expected value of the true variable of interest given the perturbed data for each observation. We show that we can construct this expected value by integrating over all the possible actual values of the true data, weighted by the conditional probability of the data values given the observed perturbed data and other observed variables in the dataset. These conditional probabilities can be derived provided that we know both how the measurement error is generated and the distribution of the true variable for observations in the dataset. Essentially, we replace our perturbed variables with new variables that contain only Berkson errors that allow consistent estimation in linear models (Fuller 2009).

Our method is related to the regression calibration approach; we show that regression calibration is equivalent to our approach when the underlying distributions of the true variable and the measurement error are both independent and normally distributed. However, our approach imposes no restrictions on the form of the underlying true distribution of the variable or of the measurement error. While we obtain unbiased and consistent estimates, we show our parameter estimates are less precise (i.e. have larger standard errors) than those estimates that could be obtained if we knew the true values of the variable in the survey.

In some sample cases that we examine the integration needed to calculate the required expected values may be straightforward and the value of the integral can be can be calculated analytically. In complex cases, however, the integration may be difficult, but we show that the value of the integral can be approximated arbitrarily closely by numerical methods.

Our approach requires an independent source for obtaining the underlying true distribution of the misreported data. For example, if we want to link a random sample of individuals to exposures at the district level, but the data only reports a cruder measure such as state of residence, we require independent information on the population numbers in each district in the state to find the conditional probably that a particular individual resides in a particular district. Similarly, for perturbed location data, we require a population density map, to give the unconditional probability an individual resides at a particular location.

Given that we know the form of the measurement error it may be possible to invert the distribution of observed perturbed data to generate the underlying distribution of the true data, which would remove the need for an external source for this underlying distribution. We show that the distributions of the true and perturbed observed variables are linked by a non-homogenous Fredholm integral equation of the first kind (Polyanin and Manzhirov 2008). The problem of solving such equations for the unknown true distribution, given a perturbed observed distribution, has been extensively studied (Hansen 1992). In general, however, this "inverse" problem is not well posed – we cannot guarantee existence or uniqueness of a solution. This is obvious for the case where the data is coarsened; if we have information on an individual's state of residence but not her district, any possible distribution of the state's residents across the state's districts is compatible with the observed state level data, and there will be multiple solutions to the integral equation linking the individual's actual district of residence to her observed state of residence. In this case, external data on the distribution of the population across districts is required for our method. However, in some simple cases, for example the randomized response technique for a binary variable, we show that inversion is possible, and the underlying distribution of the underlying variable can be constructed from the observed perturbed data.

In Section 2, we derive our theory that allows for consistent estimation with induced measurement error in one explanatory variable. We also describe a numerical method for calculating the expected value of the explanatory variable for complex cases where the integral cannot be calculated analytically. In Section 3, we extend our results to the case where several variables are measured with error and the model to be estimated has interactive effects. This is important for our motivating example of location data, where the two dimensional coordinates of the location are perturbed. In Section 4, we provide a number of examples that illustrate how our method can be used. We begin with two examples of coarsening, for discrete and continuous underlying variables, where the variable is reported only at a higher level of aggregation. We then show how out method can be applied when subjects provide randomized responses. Our fourth example is the case in which a normally distributed noise term is added to a normally distributed underlying variable, and we show that our method gives results that are the same as using the regression calibration approach. In these four simple examples, we solve the integral explicitly to derive the expected value of the true variable given the reported data. In Section 5, we apply our approach to the more complex case where the explanatory variable of interest is distance from the individual's location to the nearest facility. In this case, the integral required to calculate the expected distance given the perturbed location data cannot be calculated analytically and we approximate it through numerical integration over all possible true locations for the individual. We conduct a Monte Carlo simulation study to show that our method overcomes the bias in the estimates that results from directly using perturbed data.

2. THEORY

We begin with a simple case in which the variable that is measured with error has a true value that is a real number x drawn from the set X. Suppose we wish to estimate a relationship of the following form:

$$y_i = \alpha + \beta g(x_i) + \gamma z_i + \varepsilon_i \tag{1}$$

where y_i is the outcome, $g(x_i)$ is a known function g of the variable x_i, z_i is a covariate (generalization to many covariates is straightforward), and ε_i is a random error term with mean zero and that is independent of both x_i and z_i. In the data, x_i is not observed, but we do observe m_i, which is a perturbation of x_i, where we assume that the probability density function of the error generation process, denoted $p(m \mid x)$ is known.

It is well known that simply replacing x_i with m_i will typically make the results estimates of Equation 1 inconsistent and biased. However, suppose we can calculate the expected value

$$E\left[g(x_i) \mid m_i, z_i\right] = \int_X g(x)\, p(x \mid m_i, z_i)\, dx$$

Now set

$$u_i = g(x_i) - E\left[g(x_i) \mid m_i, z_i\right]$$

Here, u_i is the gap between the true value $g\left(x_i\right)$ and our calculated expected value. By the law of iterated expectations, we have

$$
\begin{gathered}
E\left[u_i \middle| E\left[g\left(x_i\right) | m_i, \mathrm{z}_i\right]\right] \\
=E\left[g\left(x_i\right) \middle| E\left[g\left(x_i\right) | m_i, \mathrm{z}_i\right]\right]-E\left[E\left[g\left(x_i\right) | m_i, \mathrm{z}_i\right] \middle| E\left[g\left(x_i\right) | m_i, \mathrm{z}_i\right]\right] \\
=E\left[g\left(x_i\right) | m_i, \mathrm{z}_i\right]-E\left[g\left(x_i\right) | m_i, \mathrm{z}_i\right]=0
\end{gathered}
$$

So u_i is an error term with expected value zero for all values of $E\left[g\left(x_i\right) | m_i, \mathrm{z}_i\right]$ and so is uncorrelated with $E\left[g\left(x_i\right) | m_i, \mathrm{z}_i\right]$.

Again by the Law of Iterated Expectations and the fact that $g\left(x_i\right)=E\left[g\left(x_i\right) | m_i, \mathrm{z}_i\right]+u_i$, we have

$$
E\left[g\left(x_i\right) | m_i, z_i\right]=E\left[\left(E\left[g\left(x_i\right) | m_i, z_i\right]+u_i\right) | m_i, z_i\right]=E\left[g\left(x_i\right) | m_i, z_i\right]+E\left[u_i | m_i, z_i\right]
$$

Hence, we have $E\left[u_i | m_i, z_i\right]=0$ for all m_i, z_i, and we see that u_i is an error term with expected value zero for all values of m_i, z_i and is therefore uncorrelated with both x_i and z_i.

Since $g\left(x_i\right)=E\left[g\left(x_i\right) | m_i, \mathrm{z}_i\right]+u_i$, we can rewrite the estimating equation as

$$
y_i=\alpha+\beta E\left[g\left(x_i\right) | m_i, z_i\right]+\gamma z_i+v_{i,} \quad v_i=\left(\beta u_i+\varepsilon_i\right) \tag{2}
$$

where the error term $v_i=\beta u_i+\varepsilon_i$ is mean zero and uncorrelated with either of the explanatory variables, $E\left[g\left(x_i\right) | m_i, z_i\right]$ or z_i.

It follows that by replacing the unknown $g\left(x_i\right)$ in the regression with $E\left[g\left(x_i\right) | m_i, z_i\right]$, we can estimate Equation 2 using the standard Ordinary Least Squares (OLS) methods to obtain consistent and unbiased estimates of α , β, and γ . Moreover, since all of the classical assumptions of OLS are satisfied following the correction, the standard errors of the parameter estimates will also be correct. While we have unbiased and consistent estimates, the standard deviation of the error term v_i , using our calculated expected values of the explanatory variable, will be larger than the standard deviation of the error term ε_i when estimating using the true values. This will make our estimates less precise, with higher standard errors, than if we used the true explanatory variable.

However, these results depend crucially on the fact that the underlying relationship that we want to estimate is linear.

In order to calculate $E\left[g\left(x_i\right)|\, m_i, z_i\right]$, we note that by using Bayes rule, the term can be written as

$$E\left[g\left(x_i\right)|\, m_i, z_i\right] = \int_X g\left(x\right) p\left(x\,|\, m_i, z_i\right) dx = \int_X g\left(x\right) \frac{p\left(m_i\,|\, x\right) p_{z_i}\left(x\right)}{\int_X p\left(m_i\,|\, x\right) p_{z_i}\left(x\right) dx} dx$$

Without loss of generality, we assume that the error generation process for m_i depends only on x_i and not the value of z_i. We can calculate this expectation, provided that we know the mechanism that was used to induce the error structure, $p\left(m_i\,|\, x\right)$, and we know the underlying probability density function of the true values x given z_i, given by $p_{z_i}\left(x\right)$. In order to calculate this expectation, we need to know the process generating the perturbed measure m_i given the true value x_i and also the true underlying distribution of (x, z) so we can calculate the marginal distribution $p_{z_i}\left(x\right)$ of x given a particular observed value z_i .

It is important to condition the expectation on covariates when they are present since they may be correlated with the unobserved variable x_i and may also contain information on it over and above what is present in the perturbed value m_i. In what follows, we assume for notational simplicity that we estimate a model without covariates so that we have

$$E\left[g\left(x_i\right)|\,m_i\right]=\int_X g\left(x\right)p\left(x\,|\,m_i\right)dx=\int_X g\left(x\right)\frac{p\left(m_i\,|\,x\right)p\left(x\right)}{\int_X p\left(m_i\,|\,x\right)p\left(x\right)dx}dx$$

While this can be calculated in principle, it may be difficult to calculate analytically if the functions are complex. For complex cases, suppose we divide the range of x, given by the interval $\left[x_{\min},x_{\max}\right]$, into an evenly spaced grid with grid with $S+1$ points and S intervals, at x_s for $s=0,\ldots,S$, where $x_0=x_{\min}$ and $x_S=x_{\max}$. Let the mesh of the grid be denoted as

$$h=\left|x_{s+1}-x_s\right|=\frac{\left(x_{\max}-x_{\min}\right)}{S}$$

Then, we have that for all $\varepsilon>0$, there exists a $h_0>0$ such that for $h<h_0$, we have

$$\left|\int_X g\left(x\right)\frac{p\left(m_i\,|\,x\right)p\left(x\right)}{\int_X p\left(m_i\,|\,x\right)p\left(x\right)dx}dx-\sum_{s=0}^{S-1}g\left(x_s\right)\frac{p\left(m_i\,|\,x_s\right)p\left(x_s\right)h}{\sum_{s=0}^{S-1}p\left(m_i\,|\,x_s\right)p\left(x_s\right)h}\right|<\varepsilon$$

by the definition of the Riemann integral, provided the functions $g\left(x\right)$, $p\left(m_i\,|\,x\right)$, and $p\left(x\right)$ are continuous almost everywhere, that is, the set of points at which there are discontinuities are of Lebesgue measure zero. For example, functions with any finite set of discontinuities are continuous almost everywhere (Rudin 1976). If the range of x is infinite, then we can replace the summation with fixed limits $\left[x_{\min},x_{\max}\right]$ and we can sum over a wider range as S increases, i.e. $\left[x_{\min}\left(S\right),x_{\max}\left(S\right)\right]$. Provided that this range goes to infinity, while the mesh converges to zero, then as S increases, we will obtain the same result. It therefore follows that we can approximate the continuous integral arbitrarily closely through numerical integration by taking a large number of grid points S, with sufficiently small mesh h, over all possible values of x.

In some cases, we may know the underlying distribution of the true values of x, given by $p(x)$, from external data. For example, if the sample is a random draw from an underlying population, and we know the population density $p(x)$ for the underlying population, then this will also be the underlying distribution of x in our sample. The distribution $p(x)$ should have the property that given the mechanism for producing the measurement error, the induced distribution of measured outcomes matches the empirical distribution $q(m)$ of the data, that is:

$$q(m) = \int_X p(m \mid x)\, p(x)\, dx \tag{3}$$

This is useful for checking the validity of the underlying distribution $p(x)$ if it is based on external information. In the absence of external information, we can, in principle, construct an estimate of the underlying distribution $p(x)$ based on a solution to Equation 3. Let $Z(m,x) = p(m \mid x)$, then

$$q(m) = \int_X Z(m,x)\, p(x)\, dx \tag{4}$$

Equation 4 is a non-homogenous Fredholm integral equation of the first kind with kernel $Z(m,x)$ (Polyanin and Manzhirov 2008). We can think of the integral in Equation 4 as a linear operator T from the Hilbert space of all square integrable functions on the interval $[x_{\min}, x_{\max}]$ into itself, given by $L^2[x_{\min}, x_{\max}]$. That is $q = T(p)$ where $p, q \in L^2[x_{\min}, x_{\max}]$. The problem of finding the inverse of the equation, that is the underlying distribution $p = T^{-1}(q)$, given the kernel $Z(m,x)$ and some observations from the distribution $q(m)$ has been extensively studied (Hansen 1992). Inverse problems of this kind are called "well posed" if a solution exists, is unique, and varies continuously in q. Unfortunately, in many cases, non-homogenous Fredholm integral equation of the first kind are not well posed in the sense that they do not have a unique solution. For example, in our examples section below, we consider the case where location data for a respondent is coarsened to protect confidentially by reporting only a larger administrative unit, such as state of residence and not a smaller administrative unit, like district, while the exposure to a covariate that affects the outcome for the respondent is measured at the district level. In this case, any probability distribution over the districts that adds up to the correct observed total for the state will be compatible with the observed data, and the solution of the Fredholm integral equation for the distribution of the true data will not be unique. Given the possibility of lack of uniqueness we propose the use of external data to determine the underlying true distribution on the mismeasured variable.

3. MULTIPLE VARIABLES WITH INDUCED MEASUREMENT ERROR AND INTERACTIVE EFFECTS

In some cases there may be several variable measured with error and interactive effects. This will have particular relevance in our application to perturbed location data, which is two dimensional. To make things explicit, suppose that we have two explanatory variables and an interactive effect

$$y_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \gamma g\left(x_{i1}, x_{i2}\right) + \varepsilon_i$$

where both $\left(x_{i1}, x_{i2}\right)$ are measured with error and where we observe only the mismeasured variables $\left(m_{i1}, m_{i2}\right)$.

Let $p\left(\left(m_{i1},m_{i2}\right)|\left(x_1,x_2\right)\right)$ be the known distribution of the mismeasured observation given the true pair $\left(x_{i1},x_{i2}\right)$. Note that we allow for the possibility that the each mismeasured variable depends on both true variables. For each observed pair $\left(m_{i1},m_{i2}\right)$, we wish to calculate

$$E\left[x_{i1}\,|\left(m_{i1},m_{i2}\right)\right]=\int_{X_1}\int_{X_2} x_1 p\left(\left(x_1,x_2\right)|\left(m_{i1},m_{i2}\right)\right)dx_2dx_1$$

$$=\int_{X_1}\int_{X_2} x_1\frac{p\left(\left(m_{i1},m_{i2}\right)|\left(x_1,x_2\right)\right)p\left(x_1,x_2\right)}{\int_{X_1}\int_{X_2} p\left(\left(m_{i1},m_{i2}\right)|\left(x_1,x_2\right)\right)p\left(x_1,x_2\right)dx_2dx_1}dx_2dx_1$$

Similarly

$$E\left[x_{i2}\,|\left(m_{i1},m_{i2}\right)\right]=\int_{X_2}\int_{X_1} x_2 p\left(\left(x_1,x_2\right)|\left(m_{i1},m_{i2}\right)\right)dx_1dx_2$$

$$=\int_{X_2}\int_{X_1} x_2\frac{p\left(\left(m_{i1},m_{i2}\right)|\left(x_1,x_2\right)\right)p\left(x_1,x_2\right)}{\int_{X_2}\int_{X_1} p\left(\left(m_{i1},m_{i2}\right)|\left(x_1,x_2\right)\right)p\left(x_1,x_2\right)dx_1dx_2}dx_1dx_2$$

Finally, we calculate the interactive effect

$$E\left[g\left(x_1,x_2\right)|\left(m_{i1},m_{i2}\right)\right]=\int_{X_2}\int_{X_1} g\left(x_1,x_2\right)p\left(\left(x_1,x_2\right)|\left(m_{i1},m_{i2}\right)\right)dx_1dx_2$$

$$=\int_{X_2}\int_{X_1} g\left(x_1,x_2\right)\frac{p\left(\left(m_{i1},m_{i2}\right)|\left(x_1,x_2\right)\right)p\left(x_1,x_2\right)}{\int_{X_2}\int_{X_1} p\left(\left(m_{i1},m_{i2}\right)|\left(x_1,x_2\right)\right)p\left(x_1,x_2\right)dx_1dx_2}dx_1dx_2$$

Note that we need to know the underlying joint distribution $p\left(x_1,x_2\right)$ to calculate all three of these expectations. The mismeasured value of one variable may contain information on the value of the other unless the underlying joint distribution of the two variables is independent. Given these expectations, we can then estimate

$$y_i=\alpha+\beta_1E\left[x_1\,|\left(m_{i2},m_{i2}\right)\right]+\beta_2E\left[x_2\,|\left(m_{i2},m_{i2}\right)\right]+\gamma E\left[g\left(x_1,x_2\right)|\left(m_{i2},m_{i2}\right)\right]+v_i$$

where, as before, each expectation is uncorrelated with v_i, and this equation can then be estimated by standard methods.

4. SIMPLE EXAMPLES

We now highlight several cases where our method may be used to correct for various types of measurement error that are induced in the exposure variable.

4.1. Measuring Exposure With Coarsened Discrete Variables

Suppose the model which we wish to estimate includes exposure variables measured at a low geographical level of aggregation, such as districts, while the survey data only has location measured at a higher level of aggregation, such as states. We would like to run the regression with the true exposure value for each individual based on their district of residence. Let d_{ki} be a set of k district dummies variables, one for each district, each of which takes the value 1 if the individual i is in district k and 0 otherwise. The survey data, however, only contains a state-level indicator m_i which takes different values for different states. Let our functional form for the exposure of individual i be

$$x_i = \sum_k x(k) d_{ki}$$

where $x(k)$ is the exposure level in district k. We assume that we have prior data on the probability that an individual in the sample lives in district k, given by $p(k)$. If the sample is a random sample of the whole population, then this probability will simply be the proportion of the total population that lives in district k. For more complex sample designs, $p(k)$ will depend on the structure of stratification and on the sample weights. Now, it is straightforward to show that the expected value of x_i given the state of residence measures m_i is

$$E(x_i|m_i) = \sum_k x(k)p(d_k|m_i) = \sum_k x(k)\frac{p(m_i|d_k)p(d_k)}{\sum_k p(m_i|d_k)p(d_k)}$$

Now, note that $p(d_{ki}) = p(k)$. Also, $p(m_i|d_k) = 1$ if district k is in state m_i and zero otherwise. Let $K(m_i) = \{k : p(m_i|d_k) = 1\}$. This is the set of districts that make up the state m_i . Therefore

$$E(x_i|m_i) = \frac{\sum_{k \in K(m_i)} x(k)p(k)}{\sum_{k \in K(m_i)} p(k)}$$

where $p(k)$ is the prior probability that an individual lives in district k. Our method is therefore equivalent to constructing the regressor as the population weighted exposure of people in the state using the relative district population sizes as weights. This approach is precisely what a researcher would have done intuitively to measure the average exposure level of an individual in the state.

While this method is equivalent to our approach in the simple case where we seek the expected exposure, it will also apply to more complex cases.

4.2 A Continuous Variable Reported in Intervals

In some cases, continuous variables may be reported as a discrete set of intervals. For example, rather than report household income directly, a respondent may be presented with a set of household income intervals and asked to choose the household income interval to which they belong. This discretization of the continuous variable may serve to protect respondent confidentiality or may increase response rates to a sensitive question. Let x_i be the true value of the continuous variable which has known distribution $p(x)$ while $m_i \in \{1, 2, ..., K\}$ is the income interval that is reported by respondent i. Let

x_k be the lower bound of income interval k and let x_{K+1} be the upper bound on the last income interval K. Note that it is possible that $x_1 = -\infty$ and $x_{K+1} = \infty$. Then, we have

$$E\left[g(x_i) \mid m_i = k\right] = \int_X g(x) p(x \mid m_i = k) dx = \int_X g(x) \frac{p(m_i = k \mid x) p(x)}{\int_X p(m_i = k \mid x) p(x) dx} dx$$

$$= \int_{x_k}^{x_{k+1}} g(x) \frac{p(x)}{\int_{x_k}^{x_{k+1}} p(x) dx} dx$$

If $g(x) = x$ and $p(x)$ is the uniform distribution, and each interval is bounded, this reduces to

$$\int_{x_k}^{x_{k+1}} \frac{x}{(x_{k+1} - x_k)} dx = \frac{x_{k+1} + x_k}{2}$$

which is the average value of x in the observed interval. However, in more complex cases, where $g(x)$ is nonlinear, or $p(x)$ is not uniform, an explicit calculation of the integral is required through numerical integration over the empirical distribution of the continuous variable.

4.3 Randomized Responses

Suppose a question has a binary response that takes on a value of 0 or 1. The true value of the variable for respondent i is $x_i \in \{0,1\}$. However, in order to help elicit a truthful response, the respondent is asked to give a random response under the following conditions: with probability q, she answers 0, with probability q she answers 1, and with probability $1-2q$ she answers truthfully. For example, if the respondent privately throws a die and answers 0 on a score of one, 1 on a score of six, and truthfully otherwise, we have $q = {}^1\!/_6$. This method of responding to the question implies that the interviewer cannot infer exactly the true answer given the randomized response. Let m_i be the random response that is provided by the respondent. We then have:

$$E(x_i|m_i) = p(x_i = 1|m_i) = \frac{p(m_i|x_i = 1)p(x_i = 1)}{p(m_i|x_i = 1)p(x_i = 1) + p(m_i|x_i = 0)p(x_i = 0)}$$

$$E(x_i|m_i = 1) = p(x_i = 1|m_i = 1) = \frac{p(m_i = 1|x_i = 1)p(x_i = 1)}{p(m_i = 1|x_i = 1)p(x_i = 1) + p(m_i = 1|x_i = 0)p(x_i = 0)}$$

$$= \frac{(1-q)p(x_i = 1)}{(1-q)p(x_i = 1) + qp(x_i = 0)}$$

$$E(x_i|m_i = 0) = p(x_i = 1|m_i = 0) = \frac{p(m_i = 0|x_i = 1)p(x_i = 1)}{p(m_i = 0|x_i = 1)p(x_i = 1) + p(m_i = 0|x_i = 0)p(x_i = 0)}$$

$$= \frac{qp(x_i = 1)}{qp(x_i = 1) + (1-q)p(x_i = 0)}$$

Hence, we can calculate the expected value of the true binary response if we know the underlying prevalence of the true response $p(x_i = 1)$ in the data. However, in this simple case, the Fredholm integral equation is invertible. Given this underlying unknown prevalence is the same for each individual, we have:

$$p(m = 1) = q + (1-2q)p(x = 1)$$

And hence

$$p(x = 1) = \frac{p(m = 1) - q}{(1-2q)}$$

Let $m^* = \frac{1}{N}\sum_{i=1}^{N} m_i$ be the mean of the observed randomized response. In large samples, this mean will converge to $p(m = 1)$ in the population, and we can use this to calculate

$$p(x = 1) = \frac{m^* - q}{(1-2q)}, p(x = 0) = 1 - \frac{m^* - q}{(1-2q)}$$

Note that, by construction, $q \le m^* \le 1 - q$, which implies that we can calculate the mean of the true underlying response given the mean of the randomized response. In particular, we have:

$$E(x_i \left| m_i = 1\right.) = \frac{(1-q)\dfrac{m*-q}{(1-2q)}}{(1-q)\dfrac{m*-q}{(1-2q)} + q\left(1-\dfrac{m*-q}{(1-2q)}\right)} = \frac{(1-q)(m*-q)}{(1-q)(m*-q)+q(1-q-m*)}$$

$$E(x_i \left| m_i = 0\right.) = \frac{q\dfrac{m*-q}{(1-2q)}}{q\dfrac{m*-q}{(1-2q)} + (1-q)\left(1-\dfrac{m*-q}{(1-2q)}\right)} = \frac{q(m*-q)}{q(m*-q)+(1-q)(1-q-m*)}$$

By replacing the observed randomized response with these expectations of the true underlying variable, we can then unbiasedly estimate the effect of the hidden true variable in a regression.

4.4 Normally Distributed Additive Measurement Errors and Underlying Variables

Suppose x is known to be normally distributed with mean x^* and variance σ_x^2, and the measured value is $m_i = x_i + u_i$, where u_i is a normally distributed error term with mean 0 and variance σ_u^2, then we have

$$p(x) = \frac{1}{\sqrt{2\pi\sigma_x^2}} \exp\left(\frac{\left(x-x^*\right)^2}{2\sigma_x^2}\right)$$

and

$$p(m_i \mid x) = p(u_i = m_i - x) = \frac{1}{\sqrt{2\pi\sigma_u^2}} \exp\left(\frac{\left(m_i - x\right)^2}{2\sigma_u^2}\right)$$

It therefore follows that

$$E\left[x \mid m_i\right] = \int_X xp\left(x \mid m_i\right)dx = \int_X x \frac{p\left(m_i \mid x\right)p\left(x\right)}{\int_X p\left(m_i \mid x\right)p\left(x\right)dx}dx$$

$$=\int_X x\frac{\frac{1}{\sqrt{2\pi\sigma_u^2}}\exp\left(\frac{\left(m_i-x\right)^2}{2\sigma_u^2}\right)\frac{1}{\sqrt{2\pi\sigma_x^2}}\exp\left(\frac{\left(x-x^*\right)^2}{2\sigma_x^2}\right)}{\frac{1}{\sqrt{2\pi\sigma_u^2}}\exp\left(\frac{\left(m_i-x\right)^2}{2\sigma_u^2}\right)\int_X\frac{1}{\sqrt{2\pi\sigma_x^2}}\exp\left(\frac{\left(x-x^*\right)^2}{2\sigma_x^2}\right)dx}dx$$

$$=\frac{\sigma_x}{\sigma_x+\sigma_u}m_i+\frac{\sigma_u}{\sigma_x+\sigma_u}x^*$$

So if we know the distribution of the error σ_u^2 and the distribution of x, and hence x^* and σ_x^2, we can calculate $E\left[x\,|\,m_i\right]$ and replace this term in the estimating equation, Equation 1.

This has a very close relationship with the regression calibration approach for normally distributed variables and measurement errors (Fuller 2009). Regression calibration assumes we have the perturbed normally distributed data m_i but also a validation data set that has both true values of the variable and perturbed values. Regression calibration uses the validation dataset to estimate σ_u as the standard deviation of the normally distributed residuals from a simple linear regression of m on x. The method then uses the fact that, for normally distributed variables and for $m_i = x_i + u_i$, we have $\sigma_m^2 = \sigma_x^2 + \sigma_u^2$, and hence we have $\sigma_x^2 = \sigma_m^2 - \sigma_u^2$ to estimate the variance of x σ_x^2 from the variance of the measured observations σ_m^2 and the estimate of the variance of the errors σ_u^2. In our approach, we have an advantage in that we know σ_u^2 as the variance of the error that is deliberately added to protect confidentiality, and so we do not need an external validation dataset. If we know the underlying distribution of x is normal, we can calculate it mean and variance from the mean and variance of the observed data and the error mechanism.

5. PERTURBED LOCATION DATA AND DISTANCE TO EXPOSURE: A NUMERICAL INTEGRATION EXAMPLE

The examples above allow for the explicit calculation of the integral or sum to give the expected value of the underlying true variable given the observed perturbed variable. However, this explicit calculation is not always possible, and we must therefore rely on numerical integration methods to estimate the expected value of the true exposure variable. We now examine the case of perturbed location data, an example where numerical integration has to be used.

Suppose we have true location data for an individual i given by a pair of coordinates (x_{i1}, x_{i2}). These coordinates will exactly identify the location of the individual. To protect respondent confidentiality, suppose the data collection team perturbs the coordinates by randomly displacing the coordinates (x_{i1}, x_{i2}) by a random angle that is uniformly distributed over 0 to 2π radians and by a random distance that is uniformly chosen between 0 and d kilometers at this angle. The resulting displaced coordinates for this respondent are given by (m_{i1}, m_{i2}), which are then shared with the researcher. It is easy to see that the probability density function of the perturbed data give the true data is:

$$p\left((m_{i1}, m_{i2}) \mid (x_{i1}, x_{i2})\right) = \left\{ \begin{array}{cc} 0 & \sqrt{(m_{i1}-x_{i1})^2 + (m_{i2}-x_{i2})^2} > d \\ \dfrac{1}{d \cdot 2\pi\sqrt{(m_{i1}-x_{i1})^2 + (m_{i2}-x_{i2})^2}} & \sqrt{(m_{i1}-x_{i1})^2 + (m_{i2}-x_{i2})^2} \le d \end{array} \right\}$$

Now, suppose we have a set of W exposures located at locations $r_w = (r_{w1}, r_{w2})$ for $w = 1, \ldots, W$. These may represent W health facilities or other locations that may affect the outcome of interest. The function g is given by the Euclidean distance from the individual to the nearest exposure unit (e.g. the nearest health facility), which is defined as follows:

$$g\left(x_{i1},x_{i2}\right)=\min_{w} d\left[\left(x_{i1},x_{i2}\right),\left(r_{w1},r_{w2}\right)\right]=\min_{w}\sqrt{\left(x_{i1}-r_{w1}\right)^{2}+\left(x_{i2}-r_{w2}\right)^{2}}$$

We then wish to estimate the relationship between the true minimum distance to a facility and some outcome y_i given by:

$$y_i=\alpha+\beta\cdot g\left(x_{i1},x_{i2}\right)+\varepsilon_i$$

A naïve estimation of the equation above would use the reported location data $\left(m_{i1},m_{i2}\right)$ and run the following regression:

$$y_i=\alpha_m+\beta_m g\left(m_{i1},\mathrm{m}_{i2}\right)+u_i$$

where the function $g\left(m_{i1},\mathrm{m}_{i2}\right)$ is the minimum distance from a reported location $\left(m_{i1},m_{i2}\right)$ to a facility. Estimates based on this approach are likely to be biased.

Using our method for each observed individual's location $\left(m_{i1},m_{i2}\right)$, we wish to calculate

$$E\left[g\left(x_{i1},x_{i2}\right)|\left(m_{i1},m_{i2}\right)\right]=\int_{X_2}\int_{X_1} g\left(x_1,x_2\right)p\left(\left(x_1,x_2\right)|\left(m_{i1},m_{i2}\right)\right)dx_1dx_2$$

$$=\int_{X_2}\int_{X_1} g\left(x_1,x_2\right)\frac{p\left(\left(m_{i1},m_{i2}\right)|\left(x_1,x_2\right)\right)p\left(x_1,x_2\right)}{\int_{X_2}\int_{X_1} p\left(\left(m_{i1},m_{i2}\right)|\left(x_1,x_2\right)\right)p\left(x_1,x_2\right)dx_1dx_2}dx_1dx_2$$

What we do is take each possible location for the individual and calculate the minimum distance from this location to a facility. We that calculate the expected minimum distance based on the probability that each location is the true location given the observed perturbed location.

Exact calculation of this integral is not possible. However, we can approximate this expectation using our numerical integration method by taking a grid of points $\left(x'_{j1},x'_{k2}\right)$, for $j=0,\ldots,S$ and

$k = 0,\dots,S$ that covers the complete (x_{i1}, x_{i2}) space. Let values of each coordinate lie in the range $\left[0, X\right]$, then we have $\left\{x'_{01} = 0,\dots,x'_{S1} = X\right\};\left\{x'_{02} = 0,\dots,x'_{S2} = X\right\}$. With equal spacing h on each axis, that covers all possible locations of the individual, and we can calculate

$$E\left[g\left(x_1, x_2\right)|\left(m_{i1}, m_{i2}\right)\right] \approx \sum_{j=0}^{S-1}\sum_{k=0}^{S-1} g\left(x'_{j1}, x'_{k2}\right)\frac{p\left(\left(m_{i1}, m_{i2}\right)|\left(x'_{j1}, x'_{k2}\right)\right)p\left(\left(x'_{j1}, x'_{k2}\right)\right)h^2}{\sum_{j=0}^{S-1}\sum_{k=0}^{S-1} p\left(\left(m_{i1}, m_{i2}\right)|\left(x'_{j1}, x'_{k2}\right)\right)p\left(\left(x'_{j1}, x'_{k2}\right)\right)h^2}$$

Given this expectation for each individual we can calculate the effect of the expected minimum distance on the outcome given by:

$$y_i = \alpha_c + \beta_c E\left[g\left(x_{i1}, x_{i2}\right)|\left(m_{i1}, m_{i2}\right)\right] + v_i$$

Our theory implies that this estimator will be consistent and will recover the same value of β as if we used the actual location of each individual. We test this in a Monte Carlo simulation. We first generate a 100 x 100 grid space over which 100 health facilities and 1,000 respondents are located. In particular, we generate 100 facilities at locations $r_w = \left(r_{w1}, r_{w2}\right)$ for $w = 1,\dots,100$ where each coordinate is randomly and uniformly distribution on the range $\left(0,100\right)$, that is $r_{w1}, r_{w2} \sim U\left(0,100\right)$, and we similarly generate 1,000 respondents to be at their true locations $x_i = \left(x_{i1}, x_{i2}\right)$, where $x_{i1}, x_{i2} \sim U\left(0,100\right)$.

We calculate the minimum distance from each facility to the true respondent location, and we then generate an outcome variable y_i, which we define by the following relationship:

$$y_i = 1 + 1 \cdot g\left(x_{i1}, x_{i2}\right) + \varepsilon_i \tag{5}$$

where $\varepsilon_i \sim N(0,1)$ is a randomly drawn error term. That is we generate the data on the assumption that the true values of the parameters are $\alpha = 1, \beta = 1$.

We then simulate perturbed location coordinates to their new location $m_i = (m_{i1}, m_{i2})$. We displace the respondent coordinates by a random distance d that is uniform on the interval $[0,5]$ and by random angle that is uniform on $[0,2\pi]$. This displacement algorithm implies that the probability density function of the perturbed data give the true data is:

$$p((m_{i1}, m_{i2}) | (x_{i1}, x_{i2})) = \begin{cases} 0 & \sqrt{(m_{i1} - x_{i1})^2 + (m_{i2} - x_{i2})^2} > 5 \\ \dfrac{1}{10\pi\sqrt{(m_{i1} - x_{i1})^2 + (m_{i2} - x_{i2})^2}} & \sqrt{(m_{i1} - x_{i1})^2 + (m_{i2} - x_{i2})^2} \le 5 \end{cases}$$

Since our clusters are uniformly distributed across the 100 x 100 space, our expectation simplifies to

$$E\left[g(x_1, x_2) | (m_{i1}, m_{i2})\right] \approx \sum_{j=0}^{S-1}\sum_{k=0}^{S-1} g(x'_{j1}, x'_{k2}) \frac{p((m_{i1}, m_{i2})|(x'_{j1}, x'_{k2}))}{\sum_{j=0}^{S-1}\sum_{k=0}^{S-1} p((m_{i1}, m_{i2})|(x'_{j1}, x'_{k2}))}$$

Our simulation approach has a number of steps. In each iteration, we undertake the following:

1. We randomly draw 100 facility locations $r_w = (r_{w1}, r_{w2})$ where $r_{w1}, r_{w2} \sim U(0,100)$.
2. We randomly draw 1000 individual locations $x_i = (x_{i1}, x_{i2})$, where $x_{i1}, x_{i2} \sim U(0,100)$.
3. We calculate the true minimum distances

$$g(x_{i1}, x_{i2}) = \min_w d\left[(x_{i1}, x_{i2}), (r_{w1}, r_{w2})\right] = \min_w \sqrt{(x_{i1} - r_{w1})^2 + (x_{i2} - r_{w2})^2}$$

4. We draw a random error term ε_i with $\varepsilon_i \sim N(0,1)$ and generate $y_i = \alpha + \beta \cdot g(x_{i1}, x_{i2}) + \varepsilon_i$ with $\alpha = 1, \beta = 1$.

5. We estimate the relationship $y_i = \hat{\alpha}_x + \hat{\beta}_x g(x_{i1}, x_{i2}) + \hat{\varepsilon}_i$, using the true minimum distance data, by OLS to give the estimated effect $\hat{\beta}_x$.

6. We generate a displacement vector and then perturb each respondent x_i to a new location m_i using a randomly generated distance that is uniform on the interval $[0,5]$ at a randomly generated angle in radians that is random and uniformly distributed on $[0, 2\pi]$.

7. We generate the minimum distance measure for each perturbed individual location

$$g(m_{i1}, \mathrm{m}_{i2}) = \min_{w} d\left[(m_{i1}, \mathrm{m}_{i2}), (r_{w1}, r_{w2})\right] = \min_{w} \sqrt{(m_{i1} - r_{w1})^2 + (m_{i2} - r_{w2})^2}$$

8. We estimate the relationship $y_i = \hat{\alpha}_m + \hat{\beta}_m g(m_{i1}, \mathrm{m}_{i2}) + \hat{u}_i$, using the perturbed minimum distance data, by OLS to give the estimated effect $\hat{\beta}_m$.

9. We then calculate the expectation of the true minimum distance given by

$$E\left[g(x_1, x_2) \mid (m_{i1}, m_{i2})\right] \approx \sum_{j=0}^{S-1} \sum_{k=0}^{S-1} g(x_{j1}, x_{k2}) \frac{p\left((m_{i1}, m_{i2}) \mid (x_{j1}, x_{k2})\right)}{\sum_{j=0}^{S-1} \sum_{k=0}^{S-1} p\left((m_{i1}, m_{i2}) \mid (x_{j1}, x_{k2})\right)}$$

by numerical integration.

10. We estimate the relationship $y_i = \hat{\alpha}_c + \hat{\beta}_c E\left[g(x_{i1}, x_{i2}) \mid (m_{i1}, m_{i2})\right] + \hat{v}_i$ using the corrected expectation of the distance to nearest facility.

11. We save the estimates. $\hat{\beta}_x, \hat{\beta}_m$ and $\hat{\beta}_c$for this iteration.

We repeat Steps 1-11 to generate empirical distributions for the parameter estimates $\hat{\beta}_x, \hat{\beta}_m$ and $\hat{\beta}_c$, for 1,000 iterations.

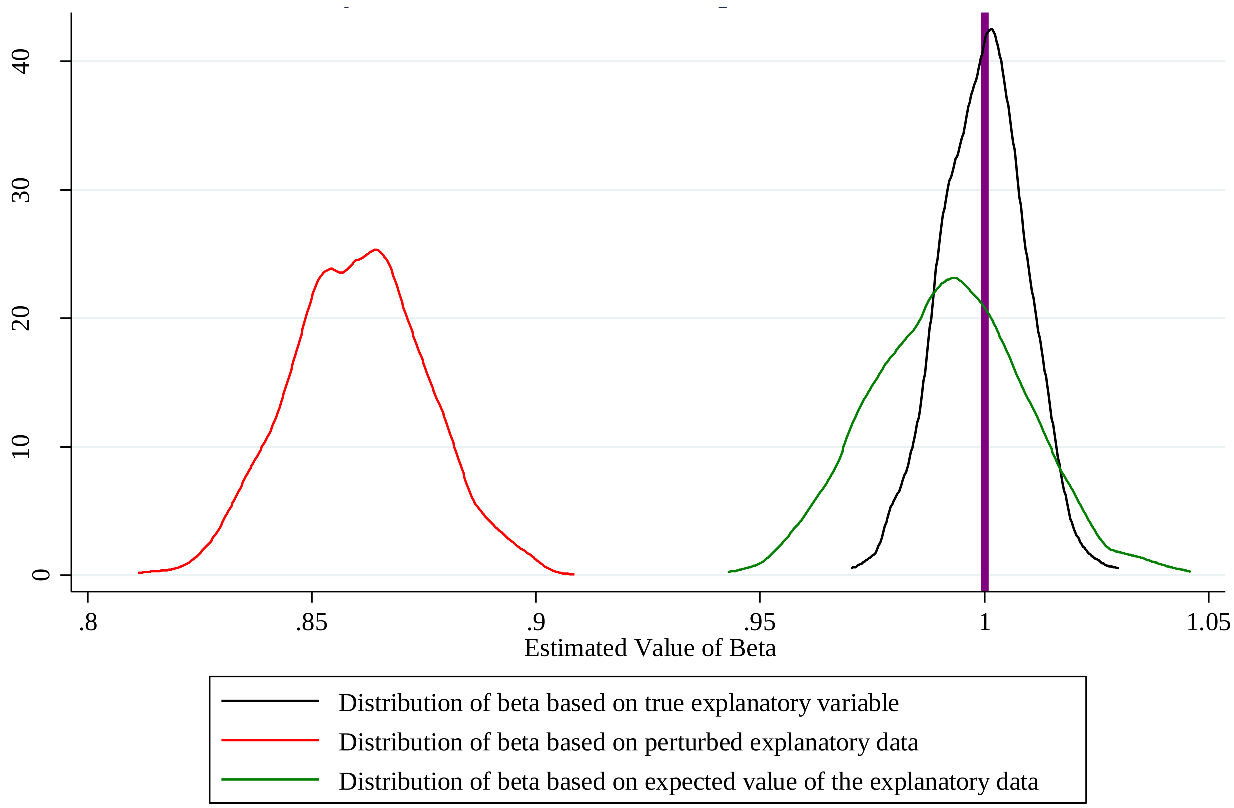
Table 1 presents summary statistics from the simulation exercise set out in steps 1-11, in which we run 1,000 iterations over a grid mesh length of $h = 1$ (a 100 x 100 mesh space) to generate empirical distributions for the parameter estimates.

Table 1: Summary Statistics from the Monte Carlo Simulation, 1,000 Iterations

	Mean	**SD**	**Minimum**	**Maximum**
$\hat{\beta}_x$	0.9997	0.0094	0.9703	1.0301
$\hat{\alpha}_x$	1.0004	0.0587	0.8193	1.1965
$\hat{\beta}_m$	0.8604	0.0151	0.8112	0.9085
$\hat{\alpha}_m$	1.7238	0.0951	1.4458	2.0546
$\hat{\beta}_c$	0.9920	0.0170	0.9427	1.0460
$\hat{\alpha}_c$	1.0524	0.0945	0.7785	1.3634
N	**1,000**			

As expected, the mean of the estimates $\hat{\alpha}_x, \hat{\beta}_x$ based on the true minimum distance data are very close to the true values of $\alpha = 1, \beta = 1$. The standard deviation of the estimates is small and the range is small and fairly symmetrical around the true values. However, when we run regressions using the perturbed location data to construct minimum distances, the estimate $\hat{\beta}_m$ is biased downwards. The mean of the estimate is well below the true value of $\beta = 1$; in fact, in all 1,000 simulations the estimated value is below one. The estimate of the intercept $\hat{\alpha}_m$ is correspondingly biased upwards. When we use the expected value of the minimum distance given the perturbed location data as a regressor, the mean estimates using this corrected data, $\hat{\alpha}_c, \hat{\beta}_c$, are again close to the true values of $\alpha = 1, \beta = 1$. However, the standard deviations and the ranges of the estimates are somewhat larger than what are observed for the true minimum distance data, which are expected given the greater noise in the regression error when using the corrected explanatory variable. Figure 1 shows the empirical distributions of the estimated values of β for our three different estimators using 1,000 iterations.

Figure 1: Empirical Distributions of Parameter Estimates for β, 1,000 Iterations



Note: The vertical purple line indicates the true parameter value, $\beta = 1$.

In practice, applying our method to actual data may be somewhat more difficult. Firstly, the method that is used to generate the perturbed data may the more complex than what we have employed. In the data reported by the DHS, for example, urban locations are perturbed by up to 2 kilometers at a random angle, while rural locations are perturbed by up to 5 kilometers, and a further randomly selected 1 percent of rural locations are perturbed by up to 10 kilometers. Moreover, if the displaced location lands outside the geographical area that is being surveyed, the DHS takes a new draw of the displacement. While this perturbation process is more complex than we have employed, it is feasible to generate the probability distribution of the perturbed data for each underlying possible location in a grid and implement our method. For the purposes of demonstrating our method, we have assumed a uniform underlying distribution of the population; however, a more realistic application would require us to use a population density map to generate the underlying distribution of the population.

6. CONCLUSIONS

In this study, we propose a general method for consistent inference under circumstances where an independent variable is deliberately measured with error. Our method is based on calculating the expected value of the true variable given the information we have, including the perturbed data that is reported. Our method relies on knowing how the measurement error is constructed and usually requires on knowing the underlying distribution of the underlying variable. We provide several examples of applications of the method; however, the method is applicable for any form of deliberately induced measurement error. We also conduct a Monte Carlo simulation of our method on an artificial dataset to show that replacing exposure data based on perturbed location with the expected exposure yields improved estimates in reasonable sample sizes.

REFERENCES

Arbia, G., Espa, G., and Giuliani, D. (2015), "Measurement Errors Arising When Using Distances in Microeconometric Modelling and the Individuals' Position Is Geo-Masked for Confidentiality," 3, 709–718. https://doi.org/10.3390/econometrics3040709.

Blair, G., Imai, K., and Zhou, Y.-Y. (2015), "Design and Analysis of the Randomized Response Technique," 110, 1304–1319. https://doi.org/10.1080/01621459.2015.1050028.

Burgert, C. R., Colston, J., Roy, T., and Zachary, B. (2013), *Geographic Displacement Procedure and Georeferenced Data Release Policy for the Demographic and Health Surveys*, DHS Spatial Analysis Reports No. 7, Calverton, MD: ICF International.

Elkies, N., Fink, G., and Bärnighausen, T. (2015), "'Scrambling' geo-referenced data to protect privacy induces bias in distance estimation," 1–16. https://doi.org/10.1007/s11111-014-0225-0.

Fuller, W. A. (2009), *Measurement Error Models*, New York, NY: Wiley and Sons.

Hansen, P. C. (1992), "Numerical tools for analysis and solution of Fredholm integral equations of the first kind," 8, 849. https://doi.org/10.1088/0266-5611/8/6/005.

Hardin, J. W., Schmiediche, H., and Carroll, R. J. (2003), "The regression-calibration method for fitting generalized linear models with additive measurement error," 3, 361–372.

Karra, M., Fink, G., and Canning, D. (2016), "Facility distance and child mortality: a multi-country study of health facility access, service utilization, and child health outcomes," dyw062. https://doi.org/10.1093/ije/dyw062.

Lohela, T. J., Campbell, O. M. R., and Gabrysch, S. (2012), "Distance to Care, Facility Delivery and Early Neonatal Mortality in Malawi and Zambia," 7, e52110. https://doi.org/10.1371/journal.pone.0052110.

Perez-Heydrich, C., Bragg-Gresham, J. L., Burgert, C. R., and Emch, M. E. (2013), *Guidelines on the Use of DHS GPS Data*, Spatial Analysis Reports No. 8, Calverton, MD: ICF International, pp. 626–627.

Polyanin, A. D., and Manzhirov, A. V. (2008), *Handbook of Integral Equations*, Boca Raton, FL: Chapman and Hall / CRC.

Rabe-Hesketh, S., Pickles, A., and Skrondal, A. (2003a), "Correcting for covariate measurement error in logistic regression using nonparametric maximum likelihood estimation," 3, 215–232.

Rabe-Hesketh, S., Skrondal, A., Pickles, A., and others (2003b), "Maximum likelihood estimation of generalized linear models with covariate measurement error," 3, 386–411.

Rudin, W. (1976), *Principles of Mathematical Analysis*, New York, NY: McGraw-Hill.

Schoeps, A., Gabrysch, S., Niamba, L., Sié, A., and Becher, H. (2011), "The Effect of Distance to Health-Care Facilities on Childhood Mortality in Rural Burkina Faso," 173, 492–498. https://doi.org/10.1093/aje/kwq386.

Spiegelman, D., McDermott, A., and Rosner, B. (1997), "Regression calibration method for correcting measurement-error bias in nutritional epidemiology," 65, 1179S–1186S.

USAID, and ICF Macro International (2014), "The DHS Program," Available at http://dhsprogram.com/.

Warren, J. L., Perez-Heydrich, C., Burgert, C. R., and Emch, M. E. (2016), "Influence of Demographic and Health Survey Point Displacements on Distance-Based Analyses," 4, 155–173. https://doi.org/10.1007/s40980-015-0014-0.