TRANSACTIONS ON DATA PRIVACY 11 (2018) 279–297

Differentially Private Verification of Regression Predictions from Synthetic Data

Haoyang Yu*, **Jerome P. Reiter****

*Department of Statistical Science, Box 90251, Duke University, Durham, NC 27708, USA.

**Department of Statistical Science, Box 90251, Duke University, Durham, NC 27708, USA.

E-mail: haoyang.yu@duke.edu, jreiter@duke.edu

Received 12 June 2018; received in revised form 4 August 2018; accepted 6 August 2018

Abstract. One approach for releasing public use files is to make synthetic data, i.e., data simulated from statistical models estimated on the confidential data. Given access only to synthetic data, users cannot tell whether the synthetic data have been constructed in ways that provide sufficient accuracy for their particular purposes. To enable users to make such assessments, data providers also can allow users to request verification measures. These are summary statistics reflecting comparisons of the results of analysis based on the synthetic and confidential data. We present three verification measures that satisfy differential privacy for assessing the quality of linear regression models. We use simulation studies to illustrate the verification measures.

Keywords. Confidentiality; Kolmogorov; Regression; Simulated.

1 Introduction

When national statistics agencies and other data stewards—henceforth all called agencies—seek to release microdata files to the public, they are ethically and often legally obligated to protect the privacy and confidentiality of data subjects' identities and sensitive attributes. In the era of big data, agencies may need to alter large fractions of values to protect privacy and confidentiality. Modifying only a small number of values, e.g., only perturbing variables deemed quasi-identifiers, may not suffice when intruders have access to dozens or hundreds of variables on data subjects, for example, from private sector sources or administrative databases, that can be used for record linkage attacks.

In such settings, one approach is to release fully synthetic data (Rubin, 1993; Fienberg, 1994; Reiter, 2002, 2005a,b, 2009; Raghunathan *et al.*, 2003; Reiter and Raghunathan, 2007; Drechsler and Reiter, 2010; Drechsler, 2011). Here, the agency replaces every value on the file with draws from statistical models designed to preserve important relationships in the confidential data. When the models adequately describe the joint distribution of the confidential data, the synthetic data can preserve important associations in the confidential data. And, they can carry low disclosure risks, as the released data do not correspond to actual records. It is nonsensical for intruders to link synthetic records to external files, which reduces the risks from record linkage attacks that have broken many anonymization strategies (e.g., Sweeney, 1997, 2013; Narayanan and Shmatikov, 2008; Parry and Chase,

2011). Because of these potential benefits, the U.S. Census Bureau uses synthetic data as a dissemination strategy for several major data products, including the Survey of Income and Program Participation (SIPP) (Abowd *et al.*, 2006), the American Community Survey group quarters data (Hawala, 2008), the OnTheMap origin-destination data (OTM) (Machanavajjhala *et al.*, 2008), and the Longitudinal Business Database (LBD) (Kinney *et al.*, 2011). Other examples of synthetic data applications have appeared in the literature (e.g., Kennickell, 1997; Abowd and Woodcock, 2001, 2004; Little *et al.*, 2004; Graham and Penny, 2005; An and Little, 2007; Drechsler *et al.*, 2008a,b; Graham *et al.*, 2009; Slavkovic and Lee, 2010).

Synthetic data have a critical weakness. Users of synthetic data cannot determine if and how much their analysis results have been impacted by the synthesis process. Inevitably, the accuracy of some analyses deteriorates significantly because of imperfect data generation models. It is arguably essential that agencies develop ways to provide feedback to users about the quality of inferences from synthetic data—or data generated by any other disclosure protection method—for specific estimands.

Verification servers (Reiter *et al.*, 2009; Karr and Reiter, 2014) offer a way to provide that feedback. The basic idea is as follows. The analyst, who has access only to the synthetic data, submits a query to the verification server for the results of a statistical model; for example, the coefficients in a regression or the mean of some variable in a subpopulation. The server, which has both the confidential and synthetic data, performs the analysis on both data sources. From the results, the server calculates a measure of how similar one result is to the other. The server returns the value of the verification measure to the analyst. With such feedback, analysts can avoid publishing—in the broad sense—results with poor quality, and be confident about results with good quality (Reiter and Drechsler, 2010).

Verification measures, however, leak information about the confidential data. Clever intruders could use this information for disclosure attacks, as shown by Reiter *et al.* (2009) and McClure and Reiter (2012). Thus, agencies need verification measures that allow them to control the information leakage about the confidential data. Chen *et al.* (2016) propose that verification measures be designed to satisfy differential privacy (Dwork, 2006). In particular, they present methods for generating differentially private plots of residuals versus predicted values in linear regression, thereby helping users assess the reasonableness of the assumptions of a posited model when applied on the confidential data. Following on this work, Barrientos *et al.* (2018) present differentially private verification measures that allow users to assess whether specific regression coefficients exceed user-defined thresholds. We are not aware of other differentially private verification measures.

In this article, we present three differentially private (DP) verification algorithms aimed at helping users assess the accuracy of predictions based on linear regression results from synthetic data. The first algorithm, which we call DP prediction tolerance intervals, determines a DP estimate of the number of predicted values that fall inside user-specified intervals. The second algorithm, which we call DP prediction histograms, provides a DP graphical summary of how well the predicted values correspond to the assumptions of the linear model. The third algorithm, which we call DP prediction Kolmogorov-Smirnov tests, quantifies the distance between the empirical distributions of predicted values and the true values of the dependent variable. The algorithms are tuned to the assumptions of the linear model, but the general ideas can be adapted for other models.

The remainder of this article is organized as follows. In Section 2, we briefly review differential privacy and the Laplace Mechanism, which we use to design the DP verification algorithms. In Section 3, we present the three DP prediction verification algorithms. In Section 4, we illustrate the performance of the algorithms with simulation studies. In Section 5, we illustrate the methods using data from the U.S. Current Population Survey. In Section

6, we conclude with suggestions for implementation and discuss future research.

2 Review of Differential Privacy

Let $\mathcal{A}$ be an algorithm that takes as input a database D and outputs some quantity S, i.e., $\mathcal{A}(D) = S$. Let D' be a neighboring database of D. For this article, as done for most differentially private algorithms in the literature, we define neighboring databases as different in one row and identical for all other rows (Dwork and Roth, 2014).

Definition 1 (ϵ-differential privacy)**.** An algorithm $\mathcal{A}$ satisfies ϵ-differential privacy if for any pair of neighboring datasets (D, D'), and any output $S \in range(\mathcal{A})$, $Pr(\mathcal{A}(D) = S) \leq e^{\epsilon} \cdot Pr[\mathcal{A}(D') = S]$.

The ϵ, known as the privacy budget, controls the degree of privacy by limiting how well intruders can distinguish any dataset from neighboring datasets given the output. The smaller the value of ϵ, the more privacy is protected.

Differential privacy satisfies two key properties that we utilize in making DP verification measures. The first is sequential composition. Suppose $\mathcal{A}_1(D)$ and $\mathcal{A}_2(D)$ are ϵ_1-DP and ϵ_2-DP algorithms. Releasing both $\mathcal{A}_1(D)$ and $\mathcal{A}_2(D)$ satisfies $(\epsilon_1 + \epsilon_2)$-differential privacy. The second is post-processing composition. Suppose $\mathcal{A}_1(D)$ is an ϵ_1-DP algorithm. For any algorithm $\mathcal{A}_3(\cdot)$, releasing $\mathcal{A}_3(\mathcal{A}_1(D))$ still is ϵ_1-DP.

A common method for ensuring ϵ-differential privacy is the Laplace Mechanism (LM). For any function $f : D \to \mathbb{R}^d$, the global sensitivity $\mathcal{G}_{\mathcal{S}}(f)$ is defined to be the maximum L_1 distance in the outputs of f computed on any two neighboring datasets D and D', that is, $\mathcal{G}_{\mathcal{S}}(f) = \max_{(D,D')} ||f(D) - f(D')||_1$. For example, when f represents a counting query we generally have $\mathcal{G}_{\mathcal{S}}(f) = 1$. The LM sets $\mathcal{A}(D) = f(D)+\delta$, where δ is a vector of independent random variables drawn from a Laplace distribution with the probability density function $p(x) = \frac{1}{2(\mathcal{G}_{\mathcal{S}}(f)/\epsilon)} \exp(-|x|/(\mathcal{G}_{\mathcal{S}}(f)/\epsilon))$. See Dwork and Roth (2014) for examples of other algorithms that satisfy differential privacy.

3 Verification Methods for Predictions

We describe the verification methods for the context of linear regression. In particular, the user wants to estimate a regression of some continuous outcome y on a set of p explanatory variables, $x = (x_1, \ldots, x_p)$, using the standard linear model. For any individual i, the linear regression model is

$$y_i = \alpha + \sum_{j=1}^{p} x_{ij}\beta_j + \tau_i, \quad \tau_i \sim N(0, \sigma^2), \tag{1}$$

where α is an intercept, $\beta = (\beta_j, \ldots, \beta_p)$ are regression coefficients, and $N(0, \sigma^2)$ represents a normal distribution with mean 0 and variance σ^2. We define $\mu(x) = \alpha + \sum_{j=1}^{p} x_{ij}\beta_j$, the true regression mean at x. To simplify notation, we shall drop the (x) from the notation when the dependence is clear.

Given the confidential data D, we assume that the user would estimate the model parameters using maximum likelihood estimation, which is equivalent to ordinary least squares estimation. Given estimates of the parameters from this model fit, which we denote $\hat{\Theta} = (\hat{\alpha}, \hat{\beta}_1, \ldots, \hat{\beta}_p, \hat{\sigma}^2)$, the user can predict the value of y for any x using $\hat{\mu} = \hat{\alpha} + \sum_{j=1}^{p} x_j\hat{\beta}_j$.

We note that D may include more variables than x and y, but we use only those variables for the regression.

In our context, the user does not have access to D and hence cannot compute $\hat{\Theta}$. Instead, the user has access to a synthetic dataset $\tilde{D}$, generated by the agency. Using $\tilde{D}$, the user estimates some regression of y on x, ignoring any other variables. For now, we assume that the user seeks to estimate the same regression as in (1). Let $\tilde{\Theta} = (\tilde{\alpha}, \tilde{\beta}_1, \ldots, \tilde{\beta}_p, \tilde{\sigma}^2)$ be the resulting estimate of Θ. With $\tilde{\Theta}$, the user can make predictions of y for any x by using $\tilde{\mu} = \tilde{\alpha} + \sum_{j=1}^{p} x_j \tilde{\beta}_j$.

A key question for the user is: are the regression predictions from the synthetic data similar to what she would get from the confidential data? More broadly, does the regression offer reasonably accurate predictions? The three DP verification measures are designed to address these questions in different ways, as we now describe.

3.1 DP prediction tolerance intervals

The DP prediction tolerance interval is the most straightforward of the three intervals to describe. To begin, for all $i \in D$, the user supplies a tolerance interval $[l_i, u_i]$; we discuss interval choices momentarily. For each $i \in D$, let $I_i = 1$ when $y_i \in [l_i, u_i]$ and let $I_i = 0$ otherwise. We then compute $P(D) = \sum_{i \in D} I_i/n$, i.e., the fraction of cases where the true outcome variables are inside the user's tolerance intervals. We cannot release $P(D)$ since it depends on D. Instead, we add noise to this fraction drawn from the Laplace Mechanism with the appropriate global sensitivity. We label this algorithm as $\mathcal{P}$ and the noisy output as $\mathcal{P}(D)$, suppressing dependence on the method for interval construction to simplify notation.

The usefulness of $\mathcal{P}$ is determined by the way the user constructs the intervals. We seek intervals with the following characteristics. First, we want $I_i = 1$ when the synthetic data regression offers predictions within a satisfactory distance from y_i, where the user determines what is satisfactory. Second, we want to specify intervals so that $\mathcal{P}$ has low and computable global sensitivity.

To meet these desiderata, we allow users only to specify functions for making limits that are passed to the verification server, which makes the intervals without ever reporting their values to the user. We use $\tilde{\mu}_i = \tilde{\alpha} + \sum_{j=1}^{p} x_{ij}\tilde{\beta}_j$ as midpoints of the intervals. We use the actual x_i in $\tilde{\mu}_i$ to ensure any potential differences in $\tilde{\mu}_i$ from y_i are due to using a model based on $\tilde{D}$ and not based on using a different x_i. Using these midpoints, we consider two types of intervals. With additive intervals, the user sends each interval width, say Δ_i, to the verification server which computes each $[l_i, u_i] = [\tilde{\mu}_i - \Delta_i, \tilde{\mu}_i + \Delta_i]$. With multiplicative intervals, the user tells the verification server proportionality constants (a_i, b_i), and the server computes $[l_i, u_i] = [a_i\tilde{\mu}_i, b_i\tilde{\mu}_i]$. We expect users to make $a_i = a$ and $b_i = b$ for all i in practice. We emphasize that the server computes the limits and the corresponding values of I_i, but never allows users to see the values of the limits; it only reports $\mathcal{P}(D)$.

With multiplicative intervals, if each (a_i, b_i) is specified independently from any other $(a_{i'}, b_{i'})$, the global sensitivity of $\mathcal{P}$ (given $\tilde{D}$ is already released) equals $1/n$ since each I_i is independent of other $I_{i'}$. This makes for a straightforward and accurate implementation of the Laplace Mechanism.

To make a correspondingly simple additive interval, users could select a constant $\Delta_i = \Delta$ for all i. When users do not have such a Δ in mind, one approach is to set Δ_i by approximating prediction confidence intervals at each x_i. To define these limits, let X and $\tilde{X}$ represent the matrix of the predictors from D used to estimate $\hat{\Theta}$ and from $\tilde{D}$ used to estimate $\tilde{\Theta}$,

respectively. From linear regression theory, the usual prediction confidence interval for a new outcome y at some value x is $\hat{\mu} \pm t_{1-\alpha/2} SE(y|x)$, where $t_{1-\alpha/2}$ is the appropriate critical value from the t-distribution with $n-(p+1)$ degrees of freedom and $SE(y|x,D) = \sqrt{\hat{\sigma}^2(1+x^T(X^TX)^{-1}x)}$. We do not use $SE(y|x)$, however, because doing so would make the values of I_i dependent, since changing one (x_i, y_i) affects both $\hat{\sigma}^2$ and X^TX. We therefore use an approximate standard error, $SE(y|x,\tilde{D}) = \sqrt{\tilde{\sigma}^2(1+x^T(\tilde{X}^T\tilde{X})^{-1}x)}$. Since $\tilde{\sigma}^2$ and $\tilde{X}^T$ come from $\tilde{D}$, which is already released, each I_i is independent and the global sensitivity is $1/n$.

Values of $P(D)$ are generally small in two cases. First, $\tilde{D}$ may be a poor representation of D, in that it fails to capture the regression relationship adequately. In this case, the value of $\tilde{\beta}$ may be quite different from $\hat{\beta}$, so that $\tilde{\mu}_i$ can be quite far from y_i. Second, the regression model used to make predictions may predict y_i poorly, even in the confidential data. In this case, $P(D)$ will be small even if $\tilde{D}$ is drawn from the exact same distribution as D. Thus, users should view $\mathcal{P}$ as an omnibus measure of the quality of predictions made from the synthetic data.

3.2 DP prediction histograms

The DP prediction tolerance intervals offer a one-number summary of the quality of the predictions from the regression. However, users may be interested in finer details, for example, to visualize the distribution of the differences between y_i and $\tilde{\mu}_i$. The DP histogram facilitates such investigations. We begin with a description of the prediction histogram for the linear model without privacy considerations. We use the notation $\Phi(w)$ to represent the cumulative probability of the argument w under the standard normal distribution.

When the assumptions in (1) are valid, so that y is normally distributed around the true μ with true variance σ^2, the values of the cumulative probabilities, $F(y|x,\mu,\sigma^2) = \Phi((y-\mu)/\sigma)$, are uniformly distributed. Thus, given unbiased estimates $\hat{\mu}$ and $\hat{\sigma}^2$, with sufficient sample size we would expect the empirical cumulative probabilities, $F(y|x,\hat{\mu},\hat{\sigma}^2) = \Phi((y-\hat{\mu})/\hat{\sigma})$ also to be approximately uniformly distributed. When the linear regression assumptions are not reasonable, we expect to see deviations from uniformity in the empirical cumulative probabilities.

We adopt this method for use in checking the quality of regression predictions from synthetic data. Specifically, for each $x_i \in D$, the verification server computes $\tilde{\mu}_i$ as in Section 3.1. Using the estimated regression variance $\tilde{\sigma}^2$ computed with the regression estimated on $\tilde{D}$, the verification server computes $F(y_i|x_i,\tilde{D}) = \Phi((y_i-\tilde{\mu}_i)/\tilde{\sigma})$. For all $i \in D$, let $B_{i1} = 1$ if $0 < p(y_i|x_i,\tilde{D}) \leq 0.1$, and $B_{i1} = 0$ otherwise. Similarly, for $j = 2,\ldots,10$, let $B_{ij} = 1$ if $.1(j-1) < p(y_i|x_i,\tilde{D}) \leq .1(j)$, and $B_{ij} = 0$ otherwise. For $j = 1,\ldots,10$, let $M_j(D) = \sum_{i\in D} B_{ij}$ be the number of observations with $p(y_i|x_i,\tilde{D})$ in the bin corresponding to the jth decile of the unit interval. The counts $M_1(D),\ldots,M_{10}(D)$ represent a histogram with ten equal-sized bins. Adding or removing only one element from D affects the count in only one of the bins, so that global sensitivity (given $\tilde{D}$ is already released) is bounded by 1. We use the Laplace Mechanism to add noise to each $M_j(D)$, resulting in noisy counts $(\mathcal{M}_1(D),\ldots,\mathcal{M}_{10}(D))$ that are released to the user. Intuitively, when $\tilde{D}$ faithfully represents D and the linear regression assumptions are reasonable, the histogram of noisy counts should be flat. As with the DP prediction tolerance intervals, either non-representative synthesis or poorly specified predictive models for y could result in deviations from approximate uniformity.

3.3 DP prediction Kolmogorov-Smirnov test

The Kolmogorov-Smirnov (KS) statistic often is used to compare two univariate distributions. The basic idea is as follows. Define the empirical cumulative distribution function for a sample $Y = (y_1, \ldots, y_n)$ as $F_Y(y) = \sum_{i=1}^{n} I_i(-\infty < y_i \leq y)/n$, where y is some arbitrary value, $I_i(\cdot) = 1$ when the condition inside the parentheses is true, and $I_i(\cdot) = 0$ otherwise. For two samples, say Y_1 and Y_2, the KS statistic is the maximum difference in the empirical distributions functions,

$$KS(Y_1, Y_2) = sup_y |F_{Y_1}(y) - F_{Y_2}(y)|, \tag{2}$$

where F_{Y_1} and F_{Y_2} are empirical distribution functions for Y_1 and Y_2, respectively. The test statistic follows a known reference distribution, which can be used to test the null hypothesis that Y_1 and Y_2 come from the same distribution. A large value of $KS(Y_1, Y_2)$, corresponding to a small reference p-value, suggests that Y_1 and Y_2 come from different distributions.

We adapt the KS statistic to compare the set of y_i and a set of plausible predicted values of y_i based on $\tilde{\Theta}$. We let Y_1 in (2) be the observed values $Y = (y_1, \ldots, y_n)$ from D. We generate a plausible Y_2 as follows. For each $x_i \in D$, we sample a predicted value $\tilde{y}_i$ from a $N(\tilde{\mu}_i, \tilde{\sigma}^2)$, where $\tilde{\mu}_i$ and $\tilde{\sigma}^2$ are computed from $\tilde{\Theta}$, i.e., the parameters of the regression model of interest estimated with $\tilde{D}$. Let $\tilde{Y} = \{\tilde{y}_i : i = 1, \ldots, n\}$ be the n generated values, which we set as Y_2 in (2). We sample from the normal distribution so as to create data that accord with the assumptions in (1). If we instead use only $\tilde{\mu}_i$, the resulting univariate distribution automatically would have smaller variance than Y_1 and hence not be immediately comparable. We add noise to $KS(Y, \tilde{Y})$ using a Laplace Mechanism with global sensitivity $2/n$.

We now explain where $2/n$ comes from. First, we note that the global sensitivity of $KS(Y_1, Y_2)$ is $1/n$ when continuous-measured Y_1 and Y_2 are both $n \times 1$ vectors. In the most extreme case, the additional observation in the neighboring database is larger (or smaller) than all values in Y, in which case the maximum difference in the empirical distribution function will change by at most $1/n$. Suppose that, given a set of $(Y, \tilde{Y})$, we change one row in Y and the corresponding row in $\tilde{Y}$. Let $(Y', \tilde{Y}')$ be the data with the one changed row. Then, we have

$$\begin{aligned}
KS(Y', \tilde{Y}') &= sup_y |F_{Y'}(y) - F_{\tilde{Y}'}(y)| \\
&= sup_y |F_{Y'}(y) - F_Y(y) + F_Y(y) - F_{\tilde{Y}}(y) + F_{\tilde{Y}}(y) - F_{\tilde{Y}'}(y)| \\
&\leq sup_y |F_{Y'}(y) - F_Y(y)| + sup_y |F_Y(y) - F_{\tilde{Y}}(y)| + sup_y |F_{\tilde{Y}}(y) - F_{\tilde{Y}'}(y)| \\
&\leq sup_y |F_{Y'}(y) - F_Y(y)| + sup_y |F_{\tilde{Y}}(y) - F_{\tilde{Y}'}(y)| + KS(Y, \tilde{Y}) \\
&\leq 2/n + KS(Y, \tilde{Y})
\end{aligned}$$

Thus, $|KS(Y', \tilde{Y}') - KS(Y, \tilde{Y})| \leq 2/n$.

The noisy KS statistic, which we call $\mathcal{KS}(Y, \tilde{Y})$, is not easily interpreted on its own. We require the reference distribution for $\mathcal{KS}(Y, \tilde{Y})$ to find reference p-values; we now present a Monte Carlo approximation of this reference distribution. First, we create h equally spaced points between 0 and 1, where h is large, and compute the cumulative probability at each point according to the reference distribution for the KS statistic without any privacy considerations. In the simulations, we set $h = 1000$ for computational convenience; analysts can use larger h for finer approximations. Second, we use differences in the cumulative

probabilities for successive points to approximate the probability density function at each point. Third, we randomly sample a value from the h points using the approximate probability density function. We adopt this approximate sampling strategy because, as far as we are aware, convenient simulation routines for the Kolmogorov distribution do not exist (Marsaglia *et al.*, 2003). Fourth, we add to the sampled value an independent draw from the Laplace distribution with parameters that correspond to the desired LM. This is one sample from the Monte Carlo approximation to the reference distribution. Finally, we repeat the third and fourth steps thousands of times to come up with the approximate reference distribution. We note that the Kolmogorov distribution depends on the sample size n, so we must consider n not sensitive to release. If this is not the case, analysts could spend some privacy budget to add noise to n, which for large sample size should not meaningfully alter the reference distribution.

To compute the p-value for a particular $\mathcal{KS}(Y, \tilde{Y})$, we determine the percentage of samples from the reference distribution that exceed the value of $\mathcal{KS}(Y, \tilde{Y})$.

4 Simulation Studies

We illustrate the performance of the three algorithms using simulation studies of several scenarios. For each scenario, we create D from an underlying true model, which we call the authentic dataset generator. We call the model used to generate $\tilde{D}$ the synthetic dataset generator. We consider scenarios where the two generators are the same and where the two differ. We define the prediction model as the regression model the user estimates on $\tilde{D}$. In practice, the prediction model may differ from the models used in the true and synthetic data generators. We include this possibility in the scenarios.

Conceptually, the agency could generate $\tilde{D}$ in multiple ways. It could use a model-based approach, like those used to generate the synthetic SIPP and synthetic LBD, or use a differentially private synthesizer, like the one used to generate the synthetic data for OTM or by Zhao *et al.* (2015). We are not aware of any data synthesizers for multivariate data with continuous (unbounded) variables that offer low error for small values of ϵ. Therefore, to keep focus on the performance of the DP verification measures, we use model-based synthetic data generators akin to those used to release existing synthetic data products. We note that agencies releasing $\tilde{D}$ that do not satisfy differential privacy cannot use sequential composition properties to claim that the combined release of $\tilde{D}$ and the verification measures is DP. In this case, agencies should interpret the privacy guarantee as bounding the additional leakage of information by providing verification measures, given that $\tilde{D}$ is already released.

We construct each D to comprise $n = 1000$ individuals with three variables, one labeled y that we treat as a response for the prediction models and two labeled (x_1, x_2) that we treat as potential explanatory variables for the prediction models. In each D and for $i = 1, \ldots, n$, we generate (x_{i1}, x_{i2}) from draws of independent normal distributions with variances equal to one. As the bivariate means of these distributions, we use each integer pair in $\{(0,0), (1,1), (2,2), \ldots, (9,9)\}$, using each mean 100 times. This makes the explanatory variables take on a wide range of values. We generate y_i from various models, depending on the scenario. To generate the $\tilde{D}$ corresponding to D, we randomly sample new values $(\tilde{x}_{i1}, \tilde{x}_{i2})$, where $i = 1, \ldots, n$, using the same mixture of normal distribution approach used to generate the variables in D. We sample from the true distribution for (x_{i1}, x_{i2}) rather than an estimate of it computed from D so that the evaluations focus on

the ability of the measures to diagnose the quality of the regression models, as opposed to the quality of the method used to generate the explanatory variables. We generate new $\tilde{y}_i$ using predictive distributions estimated from different models depending on the scenario. Once we generate $\tilde{D}$, we act like a user and estimate the prediction model from $\tilde{D}$.

For all DP verification algorithms, we use $\epsilon = 1$. For the DP prediction tolerance interval, we set each $[l_i, u_i] = [\tilde{\mu}_i - t_{1-\alpha/2}SE(y|x, \tilde{D}), \tilde{\mu}_i + t_{1-\alpha/2}SE(y|x, \tilde{D})]$, where $\alpha = .05$.

To investigate the behavior of the verification measures, we consider five scenarios. The first three include best case scenarios where the authentic data and synthetic data generators use the same models. The next two include scenarios where the agency's synthetic data generator does not match the authentic data generator. We also vary the scenarios to use linear or quadratic regressions. We run each simulation scenario at least 50 times. The overall patterns conclusions are quite similar across repetitions. Therefore, to simplify the presentation, we only show results from one simulation per scenario.

4.1 Same authentic and synthetic data generators

We consider three scenarios where the authentic and synthetic data generators are the same. In the first, we use the same linear regression specification for the generators and the prediction model. In the second, we use the linear regression specifications for the generators but use different prediction models. In the third, we use regressions with quadratic functions of X for the generators and different prediction models.

4.1.1 Linear generators and ideal prediction model

The authentic dataset generator uses $y_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \tau_i$, where $\tau_i \sim N(0, \sigma^2)$ with $\sigma^2 = 1$ and $\alpha = \beta_1 = \beta_2 = 1$. From any D, we estimate the parameters using ordinary least squares to obtain $(\hat{\alpha}, \hat{\beta}_1, \hat{\beta}_2, \hat{\sigma}^2)$. The synthetic data generator then uses $\tilde{y}_i = \hat{\alpha} + \hat{\beta}_1 \tilde{x}_{i1} + \hat{\beta}_2 \tilde{x}_i^2 + \phi_i$, where ϕ_i is randomly drawn from $N(0, \hat{\sigma}^2)$. We note that Raghunathan *et al.* (2003) recommend sampling values of the regression parameters from their posterior distributions rather than using the ordinary least squares estimates. However, for the sample size used here practically this makes little difference in the generation of $\tilde{D}$. Finally, we estimate the same linear regression on $\tilde{D}$ as the prediction model.

The value of $\mathcal{P}(D) \approx .952$, which suggests that the model predictions tend to be within acceptable tolerance as defined by $\mathcal{P}$. This is as expected, since we based $[l_i, u_i]$ on the formula for 95% prediction intervals. Figure 1 displays the results from the DP prediction histogram. The shape is relatively flat, as would be expected since the prediction model and synthetic data generator have the same specification as the authentic data generator. The value of $\mathcal{KS}(D) \approx .011$, corresponding to a p-value of around 1. These also suggest that the predictions from $\tilde{D}$ are reliable, in that the estimated model based on $\tilde{D}$ can generate predictions that look like the authentic Y.

4.1.2 Linear generators and mis-specified prediction models

We use the linear model in 4.1.1 for the authentic and synthetic data generators. However, now we mimic two users who estimate mis-specified prediction models. Specifically, User 1 estimates a linear model using only x_1 as a predictor, that is, she excludes x_2 from the prediction model. User 2 fits a model with quadratic terms rather than linear terms, that is, he regresses y on x_1^2 and x_2^2.

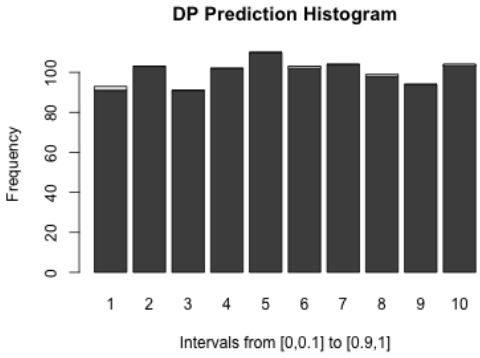

Figure 1: DP prediction histogram in scenario from Section 4.1.1 with matching linear authentic and synthetic data generators, and a matching linear prediction model. The flat shape indicates that the prediction model accurately describes the authentic data generation mechanism.

For User 1, the value of $\mathcal{P}(D) \approx .957$. Apparently, the intervals are wide enough that the mis-specification is not deemed problematic according to the criterion in $\mathcal{P}$. As evident in Figure 2, the DP prediction histogram for User 1 is relatively flat. The value of $\mathcal{KS}(D)$ for User 1 is around $.028$, corresponding to a p-value around $.835$. Thus, the $\mathcal{KS}$ algorithm also suggests that the predictions from $\tilde{D}$ are reliable, in that the estimated model based on $\tilde{D}$ can generate predictions that look like the authentic Y.

For User 2, the value of $\mathcal{P}(D) \approx .952$, which has approximately the same value and, hence, interpretation as the $\mathcal{P}(D)$ for User 1. The DP prediction histogram differentiates the two model fits more effectively. As evident in Figure 2, the DP prediction histogram for User 2 is bumpier than the one for User 1. Evidently, the mis-specification of User 2 harms prediction quality more than the mis-specification of User 1. Overall, the histogram suggests that the regression predictions from $\tilde{D}$ have a distribution that is dissimilar to the distribution of Y from D, in that the regression predictions from $\tilde{D}$ tend to be further away from the actual y_i values more frequently than would be expected if User 2's model was accurate. The $\mathcal{KS}(D)$ clearly reveals a lack of fit, with $\mathcal{KS}(D) \sim .091$ corresponding to a p-value of $.0005$.

Overall, the DP histogram and $\mathcal{KS}$ verification measures detect the dramatic model mis-specification of User 2, but they do not detect the apparently less harmful model mis-specification of User 1. The two sets of results also suggest that the prediction model of User 1 is preferred to that of User 2, because the assumptions for User 1 appear to fit the data more accurately. This illustrates how the verification measures can be used to compare the quality of predictions from different models.

4.1.3 Quadratic generators and mis-specified prediction models

The authentic dataset generator uses $y_i = \alpha + \beta_1 x_{i1}^2 + \beta_2 x_{i2}^2 + \tau_i$, where $\tau_i \sim N(0, \sigma^2)$ with $\sigma^2 = 1$ and $\alpha = \beta_1 = \beta_2 = 1$. From any D, we estimate the parameters using ordinary least squares to obtain $(\hat{\alpha}, \hat{\beta}_1, \hat{\beta}_2, \hat{\sigma}^2)$. The synthetic data generator uses $\tilde{y}_i = \hat{\alpha} + \hat{\beta}_1 \tilde{x}_{i1}^2 + \hat{\beta}_2 \tilde{x}_{i2}^2 + \phi_i$, where ϕ_i is a draw from $N(0, \hat{\sigma}^2)$. User 1 estimates a linear model using only x_1^2 as a

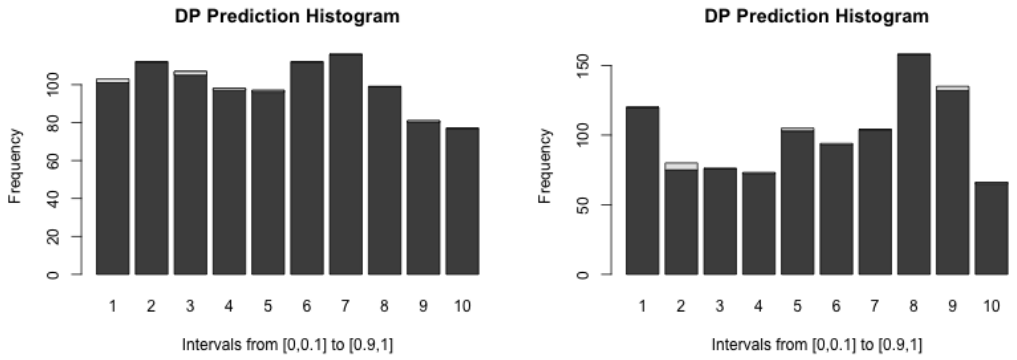

Figure 2: DP prediction histogram in scenario Section 4.1.2 with matching linear authentic and synthetic data generators, and two mis-specified prediction models. The left panel corresponds to User 1, who does not include an important term in the prediction model. The right panel corresponds to User 2, who includes quadratic terms rather than linear terms in the prediction model. The model for User 1 is preferred over the model for User 2.

predictor, that is, she excludes x_2^2 from the prediction model. User 2 fits a model with linear terms rather than quadratic terms, that is, he regresses y on x_1 and x_2 only.

For both users, the value of $\mathcal{P}(D)$ is slightly lower than .95, at around .930 for User 1 and .942 for User 2. For both users, the predictions still tend to be inside the prediction tolerance intervals, despite the model mis-specifications. As evident in Figure 3, the DP prediction histograms for both users are far from uniformly distributed, indicating the lack of model fit. The histogram for User 1 has most of its mass close to .5, suggesting that its predictions tend to be closer to the actual Y values than expected under the linear regression model, where distance is measured in terms of $\tilde{\sigma}$. Despite the model mis-specification, User 1 may be comfortable using the predictions from her model since it tends to offer accurate predictions. The histogram for User 2 has most of its mass far from 0.5, suggesting that its predictions tend to be further from the actual Y values than expected under the linear regression model. This shape suggests that User 2 not be comfortable using his model. The value of $\mathcal{KS}(D)$ for the prediction model of User 1 is around .109, which corresponds to a p-value reported as 0. The value of $\mathcal{KS}(D)$ for the prediction model of User 2 is around .177, corresponding to a p-value also reported as 0. Both values of $\mathcal{KS}(D)$ indicate the lack of model accuracy. As with the DP prediction histograms, the $\mathcal{KS}$ measure suggests that users prefer the prediction model for User 1 over the model for User 2.

4.2 Different authentic and synthetic data generators

We consider two scenarios where the authentic and synthetic data generators differ. In both scenarios, we evaluate the performance of the verification measures for two users' prediction models.

4.2.1 Quadratic authentic generator and linear synthetic generator

We generate D using the authentic data generator in 4.1.3, which uses squared terms of x_1 and x_2 to create y. As the synthesis model, however, we use only a linear regression.

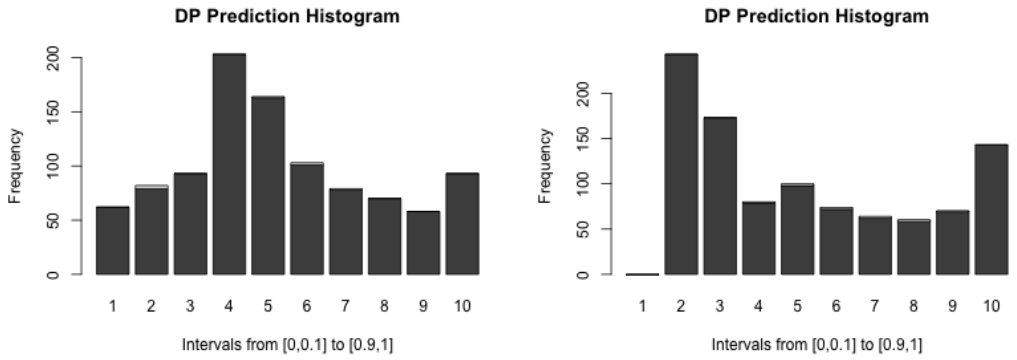

Figure 3: DP prediction histogram in scenario from Section 4.1.3 with matching quadratic authentic and synthetic data generators, and two mis-specified prediction models. The left panel corresponds to User 1, who does not include an important term in the prediction model. The right panel corresponds to User 2, who includes linear terms rather than quadratic terms in the prediction model. The model for User 1 is preferred over the model for User 2.

That is, from any D, we estimate the parameters of a linear model of y on (x_1, x_2) using ordinary least squares to obtain $(\hat{\alpha}, \hat{\beta}_1, \hat{\beta}_2, \hat{\sigma}^2)$. The synthetic data generator then uses $\tilde{y}_i = \hat{\alpha} + \hat{\beta}_1 \tilde{x}_{i1} + \hat{\beta}_2 \tilde{x}_{i2} + \phi_i$, where $\phi_i \sim N(0, \hat{\sigma}^2)$. User 1 regresses y on x_1^2 and x_2^2 as the prediction model; this happens to match the authentic data generator. User 2 regresses y on x_1 and x_2 as the prediction model; this happens to match the synthetic data generator.

For User 1, the value of $\mathcal{P}(D)$ is close to 1.0, which indicates that nearly all the predictions are within the tolerance band defined by $\mathcal{P}$. From Figure 4, the DP prediction histogram for User 1 is tightly clustered around .5. This suggests that the values of $\tilde{\mu}_i$ from the prediction model tend not to be too far from the corresponding y_i, at least according to the distribution implied by the prediction model. The lack of uniformity also highlights a lack of fit. Given that the prediction model matches the authentic data generation model for User 1, the lack of fit derives from the mismatch in the synthetic data generator. The value of $\mathcal{KS}(D) \approx .118$, which corresponds to a p-value reported as 0. The $\mathcal{KS}$ is sensitive enough to detect the mis-specification arising from the inaccurate synthetic data generator.

For User 2, the value of $\mathcal{P}(D) \approx .958$. The DP prediction histogram has high frequencies at the edge deciles of [0,.1] and [.9,1]. This indicates that many predicted values are far from their corresponding y_i, a feature suggesting that the user should not feel comfortable using the prediction model. The value of $\mathcal{KS}(D)$ confirms this, as it is around .167 corresponding to a p-value reported as 0. The two sets of measures suggest that the prediction model for User 1 be preferred over the model for User 2.

4.2.2 Linear authentic generator and polynomial synthetic generator

In the model-based synthetic data literature, agencies are advised to use richly specified synthesis models, in the sense that they should err on the side of including unimportant variables as predictors rather than excluding important variables (Reiter, 2005a). To mimic this advice, we generate the authentic dataset from a linear regression model and synthetic data from a polynomial regression model. Specifically, we generate authentic data from

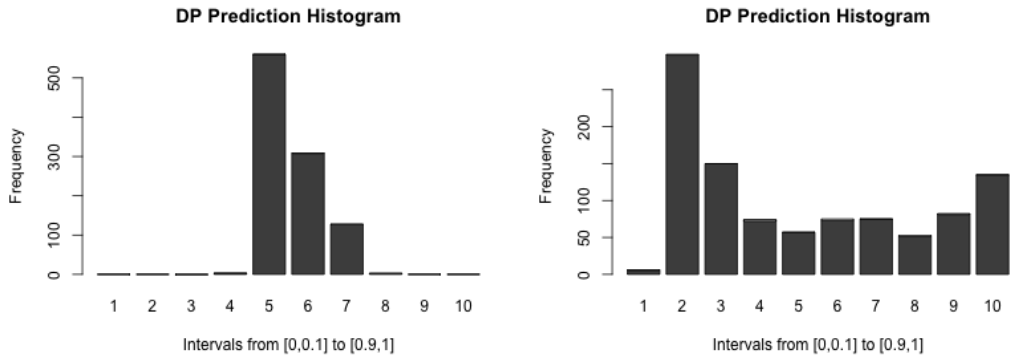

Figure 4: DP prediction histogram in scenario from Section 4.2.1 with quadratic authentic and linear synthetic data generators, and two prediction models. The left panel corresponds to User 1, whose prediction model happens to match the authentic data generator. The right panel corresponds to User 2, whose prediction model happens to match the synthetic data generator. The model for User 1 is preferred over the model for User 2.

the same model as in Section 4.1.1. To make synthetic data, we regress y on (x_1, x_1^2, x_2, x_2^2) using D, and use ordinary least squares to obtain $(\hat{\alpha}, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \hat{\beta}_4, \hat{\sigma}^2)$. The synthetic data generator uses $\tilde{y}_i = \hat{\alpha} + \hat{\beta}_1\tilde{x}_{i1} + \hat{\beta}_2\tilde{x}_{i1}^2 + \hat{\beta}_3\tilde{x}_{i2} + \hat{\beta}_4\tilde{x}_{i2}^2 + \phi_i$, where $\phi_i \sim N(0, \hat{\sigma}^2)$. User 1 regresses y on x_1 and x_2 as the prediction model; this happens to match the authentic data generator and is a submodel of the synthetic data generator. User 2 regresses y on x_1^2 and x_2^2 as the prediction model; this does not match either data generator.

For User 1, the value of $\mathcal{P}(D) \approx .948$; the DP prediction histogram in Figure 5 is flat; and, the value of $\mathcal{KS}(\mathcal{D}) \approx .027$ with a p-value of .86. From these results, we infer that the model of User 1 yields reliable predictions. This is not surprising, since the prediction model and synthetic data generator model include the authentic data generator model as submodels. Apparently, including irrelevant predictors in the synthesis model does not harm the quality of the predictions for User 1 substantially.

For User 2, the value of $\mathcal{P}(D)$ is still high at around .949. However, the DP prediction histogram in Figure 5 is rather bumpy, indicating some lack of fit. The value of $\mathcal{KS}(\mathcal{D}) \approx .084$, corresponding to a p-value of .002. Having a good synthesizer does not ameliorate problems that can arise from having a poor prediction model.

4.3 Findings from simulation results

Generally speaking, the DP prediction tolerance intervals return values around 95% across all simulation scenarios. In these scenarios, $SE(y|x, \tilde{D})$ is large even with a poorly fitting model, generating wide $[l_i, u_i]$ that usually include y_i. As a result, in these scenarios and with this additive Δ, $\mathcal{P}(D)$ is not useful for differentiating the quality of fit of competing models, nor for determining whether or not the assumptions of a prediction model are reasonable. This finding emphasizes the importance of selecting intervals that reflect user's specific desiderata (which we did not do for these artificial data simulation studies).

The DP prediction histogram allows users to visualize differences in the distributions of the actual values and the predicted values based on the synthetic data model. In these

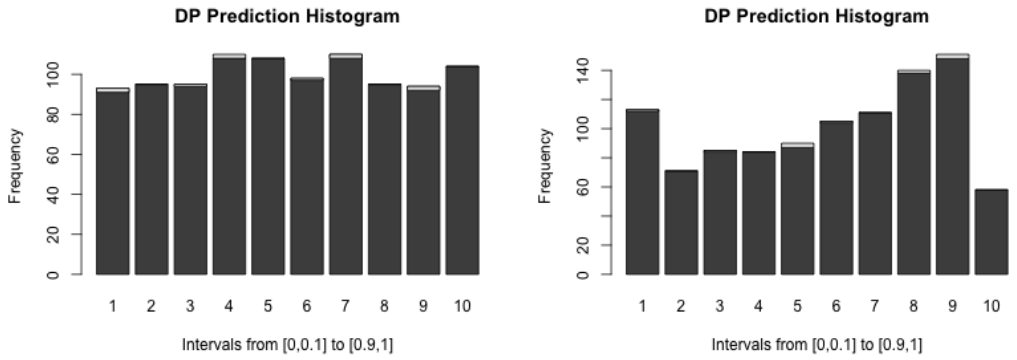

Figure 5: DP prediction histogram in scenario Section 4.2.2 with linear authentic and polynomial synthetic data generators, and two prediction models. The left panel corresponds to User 1, whose prediction model happens to match the authentic data generator and is a submodel of the synthetic data generator. The right panel corresponds to User 2, whose prediction model does not match either data generator. The model for User 1 is preferred over the model for User 2.

simulations, it reveals model mis-specifications both for the prediction model and the synthetic data generator. When the generators differ substantially, or when the assumptions in the prediction model are unreliable, the DP prediction histogram shows a non-uniform pattern.

The DP prediction KS test is very sensitive to model mis-specifications. In these simulations, it offers small p-values in all cases where the synthetic data generator or prediction model do not include the authentic data generator as a submodel. The DP prediction KS test appears to be especially useful as a one number summary for comparing different prediction models, with models corresponding to higher p-values (lower values of $\mathcal{KS}(\mathcal{D})$) being preferred.

5 Empirical Illustration

We now illustrate the DP prediction verification methods on genuine data. Specifically, we use a subset of the March 2000 Current Population Survey (CPS) public use file comprising $n = 49436$ heads of households with non-negative incomes and the variables described in Table 1. These data have been used in previous research (e.g., Reiter, 2005a,c). For simplicity, we generate a partially synthetic dataset (Little, 1993; Reiter, 2003), in which we synthesize only the income variable and leave all other variables at their fixed values.

To make synthetic values of income, we regress the cube root of income on all the variables in Table 1. In the mean function of the regression, we use linear and quadratic terms for each of the numerical variables, and indicator variables for each of the categorical variables. We also examined models that use the logarithm of income and the original scale of income as the dependent variable, but the cube root transformation offers the best fit based on residual plot diagnostics. After estimating the regression on the n records, we find the resulting estimated coefficients $(\hat{\alpha}, \hat{\beta})$ and estimated standard deviation $\hat{\sigma}$. For $i = 1, \ldots, n$, we compute $\hat{\mu}_i = \hat{\alpha} + x_i\hat{\beta}$, where x_i includes the values for all the explanatory variables

Variable	Type	Range
Sex	Categorical	Male, Female
Race	Categorical	White, Black, American Indian, Asian
Marital Status	Categorical	7 categories, coded 1-7
Highest attained education level	Categorical	16 categories, coded 31-48
Ages (years)	Numerical	15-90
Child Support Payments ($)	Numerical	0 - 23917.0
Social Security Payment($)	Numerical	0 - 50000.0
Household Property Taxes ($)	Numerical	0 - 99997.0
Household Income ($)	Numerical	1 - 768742.0

Table 1: Description of the variables used in the empirical study

used in the mean function of the regression. Using the estimated regression variance $\hat{\sigma}^2$, for each individual i we generate the synthetic income $\tilde{y}_i = (\hat{\mu}_i + \tilde{\tau}_i)^3$, where $\tilde{\tau}_i$ is a random draw from $N(0, \hat{\sigma}^2)$. The partially synthetic data $\tilde{D}$ comprise each record's $\tilde{y}_i$ and explanatory variables.

Acting now as the user of $\tilde{D}$, we consider two possible regression models for income on x. In the first we regress $\tilde{y}_i$ on x_i, and in the second we regress $\tilde{y}_i^{1/3}$ on x_i. We compute the maximum likelihood estimates for each model, which we use in the DP prediction verification measures. For the first model, we set $\tilde{\mu}_i = \tilde{\alpha} + x_i\tilde{\beta}$. For the second model, we set $\tilde{\mu}_i = (\tilde{\alpha} + x_i\tilde{\beta})^3$, so as to transform back to the original scale of income. For the DP prediction tolerance intervals, we use multiplicative bounds so that the intervals are of the form $[a\tilde{\mu}_i, b\tilde{\mu}_i]$. For illustration, we consider 10% and 20% tolerance levels, so that $(a, b) \in \{(.9, 1.1), (.8, 1.2)\}$. These are demanding levels of prediction accuracy that are hard to satisfy with income models, but our goal is to illustrate the methodology rather than evaluate the specific model. We note that users trivially could compute $\mathcal{P}(D)$ with other tolerance bounds. We also examine additive intervals based on the 95% prediction confidence interval described in Section 3.1 and used in Section 4. We show results for $\epsilon = 1$ and $\epsilon = .1$.

Table 2 displays the results of $\mathcal{P}(D)$ for all three intervals and both values of ϵ. We also include the value of $P(D)$, i.e., adding no noise to the percentage, as a baseline for comparison. For either multiplicative tolerance level, the fractions of intervals that cover the actual y_i are similar for both models, giving no compelling reason to prefer one model over the other. However, if we demand prediction accuracy with tolerance at either multiplicative intervals, we might conclude that the predictions based on the synthetic data are not sufficiently accurate and, therefore, not place much faith in predictions gleaned from $\tilde{D}$ for these models. As in Section 4, using additive bounds based on 95% prediction confidence intervals results in near 95% inclusion rates for both models. We note that conclusions are similar across both values of ϵ.

For the DP prediction histograms, we use the scales of the dependent variable when computing the probabilities. Thus, for the regression with $y^{1/3}$ as the outcome, we compare values of $y^{1/3}$ to $(\tilde{\alpha} + x_i\tilde{\beta})$. Figures 6 and 7 displays results for the DP prediction histograms. The loss in interpretability going from $\epsilon = 1$ to $\epsilon = .1$ is minimal. The plots reveal a lack of fit of the predictions for both models. The predictions from the synthetic data regressions tend to be closer to their corresponding actual Y values than we would expect under the model assumptions. Users desiring stringent adherence to the model assumptions should

Regression	Prediction Interval	No noise	$\epsilon = 1$	$\epsilon = .1$
y on x	$(a, b) = (.9, 1.1)$	.1282	.1282	.1284
	$(a, b) = (.8, 1.2)$	.2580	.2580	.2579
	Additive	.9576	.9576	.9573
$y^{1/3}$ on x	$(a, b) = (.9, 1.1)$	.1363	.1363	.1365
	$(a, b) = (.8, 1.2)$	.2707	.2707	.2712
	Additive	.9439	.9439	.9440

Table 2: Results of DP prediction tolerance interval for empirical illustration with CPS data using regression of y_i on x_i and regression of $y_i^{1/3}$ on x_i. The column labeled "No noise" refers to the output of the algorithm without adding any noise for privacy purposes.

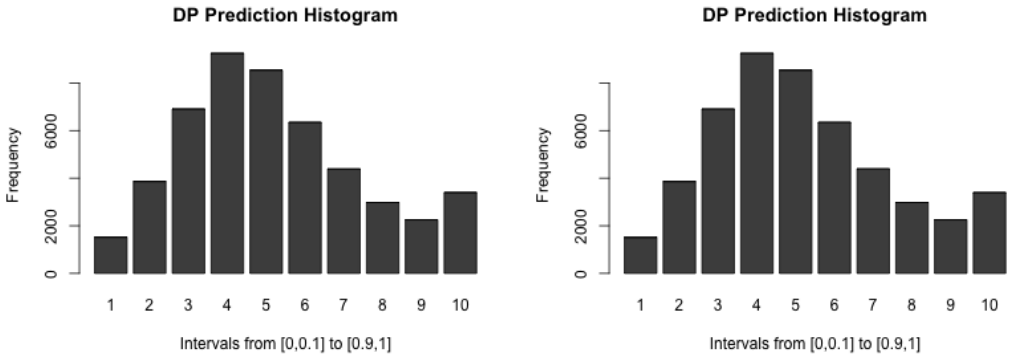

Figure 6: Results of DP prediction histogram for empirical illustration with CPS data from the regression of y_i on x_i. Left panel shows results for $\epsilon = 1$ and right panel shows results for $\epsilon = .1$

deem the resulting models inadequate. On the other hand, the deviations from uniformity are not drastic compared to those seen for the extremely poor fitting models in Section 4.1.3 and 4.2.1. Users willing to accept deviations from model assumptions, but still requiring predictions that are reasonably close to what one would expect under the model, may well find these models useful for prediction.

For the DP prediction Kolmogorov-Smirnoff tests, we work with predicted values of Y on the original scale. Thus, for the regression with $y^{1/3}$ as the outcome, we compute the statistic with plausible values of y rather than $y^{1/3}$. For the regression of y_i on x_i, the value of $\mathcal{KS}(Y, \tilde{Y}) \approx .1368$ when $\epsilon = 1$ and $\mathcal{KS}(Y, \tilde{Y}) \approx .1390$ when $\epsilon = .1$. For the regression of $y_i^{1/3}$ on x_i, the value of $\mathcal{KS}(Y, \tilde{Y}) \approx .1228$ when $\epsilon = 1$ and $\mathcal{KS}(Y, \tilde{Y}) \approx .1230$ when $\epsilon = .1$. Both correspond to p-values near 0, indicating that there are differences in the distributions of Y and $\tilde{Y}$. The smaller value of $\mathcal{KS}(Y, \tilde{Y})$ for the regression using $y^{1/3}$ suggests that it could be preferred to the regression using y.

Given that we still see lack of fit when using the same model for predictions that we used to generate $\tilde{D}$, the lack of fit stems from mismatches between the model used to generate the synthetic data and the true data generation mechanism. Investigations of the true data indicate that the normality assumption is not ideal for this regression. Apparently, more flexible synthetic data generators are needed to create more realistic synthetic data.

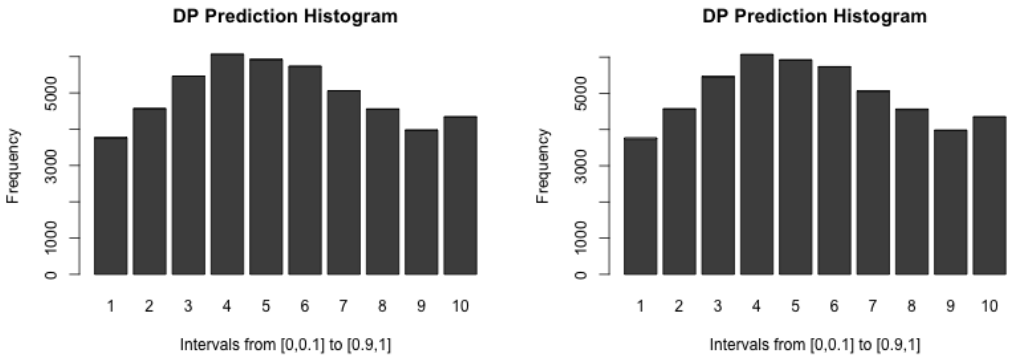

Figure 7: Results of DP prediction histogram for empirical illustration with CPS data from the regression of $y_i^{1/3}$ on x_i. Left panel shows results for $\epsilon = 1$ and right panel shows results for $\epsilon = .1$

6 Conclusion

The verification measures here are specific to linear regressions. However, they can be modified for other regression models with continuous outcomes. In particular, the DP prediction tolerance interval can be used easily with multiplicative intervals, as one simply applies the proportionality constants to the predicted values. The DP prediction histogram can be based on empirical cumulative probabilities derived from the assumptions of the model. For example, if the model assumes gamma distributed errors, then we compute the empirical cumulative probabilities using the gamma distribution and the predicted values. We can use a similar strategy to do the DP prediction KS test, in that we generate plausible predictions from the model assumptions and estimated parameters.

For regressions with binary outcomes, the DP verification measures presented here are not appropriate. Instead, as a measure of predictive accuracy, we recommend using the approach in Chen *et al.* (2016) to compute a DP receiver-operator characteristic curve, from which one can compute a DP area under the curve. For visualizations of prediction accuracy, we recommend using DP plots of binned residuals as described in Chen *et al.* (2018).

Acknowledgments

This work was supported by grants SES 1131897 and ACI 1443014 from the U.S. National Science Foundation.

References

Abowd, J., Stinson, M., and Benedetto, G. (2006). Final report to the Social Security Administration on the SIPP/SSA/IRS Public Use File Project. Tech. rep., U.S. Census Bureau Longitudinal Employer-Household Dynamics Program. Available at `http://www.census.gov/sipp/synth_data.html`.

Abowd, J. M. and Woodcock, S. D. (2001). Disclosure limitation in longitudinal linked

data. In P. Doyle, J. Lane, L. Zayatz, and J. Theeuwes, eds., *Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies*, 215–277. Amsterdam: North-Holland.

Abowd, J. M. and Woodcock, S. D. (2004). Multiply-imputing confidential characteristics and file links in longitudinal linked data. In J. Domingo-Ferrer and V. Torra, eds., *Privacy in Statistical Databases*, 290–297. New York: Springer-Verlag.

An, D. and Little, R. (2007). Multiple imputation: an alternative to top coding for statistical disclosure control. *Journal of the Royal Statistical Society, Series A* **170**, 923–940.

Barrientos, A. F., Bolton, A., Balmat, T., Reiter, J. P., de Figueiredo, J. M., Machanavajjhala, A., Chen, Y., Kneifel, C., and DeLong, M. (2018). Providing access to confidential research data through synthesis and verification: An application to data on employees of the U.S. federal government. *Annals of Applied Statistics* **12**, 1124–1156.

Chen, Y. Barrientos, A. F., Machanavajjhala, A., and Reiter, J. P. (2018). Is my model any good: Differentially private regression diagnostics. *Knowledge and Information Systems* **54**, 33–64.

Chen, Y., Machanavajjhala, A., Reiter, J. P., and Barrientos, A. F. (2016). Differentially private regression diagnostics. In *Proceedings of the IEEE International Conference on Data Mining*, 81–90.

Drechsler, J. (2011). *Synthetic Datasets for Statistical Disclosure Control*. New York: Springer-Verlag.

Drechsler, J., Bender, S., and Rässler, S. (2008a). Comparing fully and partially synthetic datasets for statistical disclosure control in the German IAB Establishment Panel. *Transactions on Data Privacy* **1**, 105–130.

Drechsler, J., Dundler, A., Bender, S., Rässler, S., and Zwick, T. (2008b). A new approach for disclosure control in the IAB Establishment Panel–Multiple imputation for a better data access. *Advances in Statistical Analysis* **92**, 439 – 458.

Drechsler, J. and Reiter, J. P. (2010). Sampling with synthesis: A new approach for releasing public use census microdata. *Journal of the American Statistical Association* **105**, 1347–1357.

Dwork, C. (2006). Differential privacy. In *33rd International Colloquium on Automata, Languages, and Programming, part II*, 1–12. Berlin: Springer.

Dwork, C. and Roth, A. (2014). *The Algorithmic Foundations of Differential Privacy*. Now Publishers.

Fienberg, S. E. (1994). A radical proposal for the provision of micro-data samples and the preservation of confidentiality. Tech. rep., Department of Statistics, Carnegie-Mellon University.

Graham, P. and Penny, R. (2005). Multiply imputed synthetic data files. Tech. rep., University of Otago, http://www.uoc.otago.ac.nz/departments/pubhealth/pgrahpub.htm.

Graham, P., Young, J., and Penny, R. (2009). Multiply imputed synthetic data: Evaluation of hierarchical Bayesian imputation models. *Journal of Official Statistics* **25**, 245–268.

Hawala, S. (2008). Producing partially synthetic data to avoid disclosure. In *Proceedings of the Joint Statistical Meetings*. Alexandria, VA: American Statistical Association.

Karr, A. F. and Reiter, J. P. (2014). Using statistics to protect privacy. In J. Lane, V. Stodden, S. Bender, and H. Nissenbaum, eds., *Privacy, Big Data, and the Public Good: Frameworks for Engagement*, 276–295. Cambridge University Press.

Kennickell, A. B. (1997). Multiple imputation and disclosure protection: The case of the 1995 Survey of Consumer Finances. In W. Alvey and B. Jamerson, eds., *Record Linkage Techniques, 1997*, 248–267. Washington, D.C.: National Academy Press.

Kinney, S. K., Reiter, J. P., Reznek, A. P., Miranda, J., Jarmin, R. S., and Abowd, J. M. (2011). Towards unrestricted public use business microdata: The synthetic Longitudinal Business Database. *International Statistical Review* **79**, 363–384.

Little, R. J. A. (1993). Statistical analysis of masked data. *Journal of Official Statistics* **9**, 407–426.

Little, R. J. A., Liu, F., and Raghunathan, T. E. (2004). Statistical disclosure techniques based on multiple imputation. In A. Gelman and X. L. Meng, eds., *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives*, 141–152. New York: John Wiley & Sons.

Machanavajjhala, A., Kifer, D., Abowd, J., Gehrke, J., and Vilhuber, L. (2008). Privacy: Theory meets practice on the map. In *IEEE 24th International Conference on Data Engineering*, 277–286.

Marsaglia, G., Tsang, W., and Wang, J. (2003). Evaluating Kolmogorov's distribution. *Journal of Statistical Software* **8:18**, 1–4.

McClure, D. and Reiter, J. P. (2012). Towards providing automated feedback on the quality of inferences from synthetic datasets. *Journal of Privacy and Confidentiality* **4:1**, Article 8.

Narayanan, A. and Shmatikov, V. (2008). Robust de-anonymization of large sparse datasets. In *Proceedings of the IEEE Symposium on Security and Privacy*, 111–125.

Parry, M. and Chase, J. (2011). Harvard researchers accused of breaching students' privacy. *Chronicle* **1**, 30.

Raghunathan, T. E., Reiter, J. P., and Rubin, D. B. (2003). Multiple imputation for statistical disclosure limitation. *Journal of Official Statistics* **19**, 1–16.

Reiter, J. P. (2002). Satisfying disclosure restrictions with synthetic data sets. *Journal of Official Statistics* **18**, 531–544.

Reiter, J. P. (2003). Inference for partially synthetic, public use microdata sets. *Survey Methodology* **29**, 181–189.

Reiter, J. P. (2005a). Releasing multiply-imputed, synthetic public use microdata: An illustration and empirical study. *Journal of the Royal Statistical Society, Series A* **168**, 185–205.

Reiter, J. P. (2005b). Significance tests for multi-component estimands from multiply-imputed, synthetic microdata. *Journal of Statistical Planning and Inference* **131**, 365–377.

Reiter, J. P. (2005c). Using CART to generate partially synthetic, public use microdata. *Journal of Official Statistics* **21**, 441–462.

Reiter, J. P. (2009). Using multiple imputation to integrate and disseminate confidential microdata. *International Statistical Review* **77**, 179–195.

Reiter, J. P. and Drechsler, J. (2010). Releasing multiply-imputed, synthetic data generated in two stages to protect confidentiality. *Statistica Sinica* **20**, 405–422.

Reiter, J. P., Oganian, A., and Karr, A. F. (2009). Verification servers: Enabling analysts to assess the quality of inferences from public use data. *Computational Statistics and Data Analysis* **53**, 1475–1482.

Reiter, J. P. and Raghunathan, T. E. (2007). The multiple adaptations of multiple imputation. *Journal of the American Statistical Association* **102**, 1462–1471.

Rubin, D. B. (1993). Discussion: Statistical disclosure limitation. *Journal of Official Statistics* **9**, 462–468.

Slavkovic, A. B. and Lee, J. (2010). Synthetic two-way contingency tables that preserve conditional frequencies. *Statistical Methodology* **7**, 225–239.

Sweeney, L. (1997). Weaving technology and policy together to maintain confidentiality. *The Journal of Law, Medicine & Ethics* **25**, 2-3, 98–110.

Sweeney, L. (2013). Matching known patients to health records in washington state data. Tech. rep. Data Privacy Lab, Harvard University.

Zhao, Y., Wang, X., Jiang, X., Ohno-Machado, L., and Tang, H. (2015). Choosing blindly but wisely: differentially private solicitation of DNA datasets for disease marker discovery. *Journal of the American Medical Informatics Association* **22**, 100-108.