

Microdata Simulation for Confidentiality of Tax Returns Using Quantile Regression and Hot Deck

Jennifer Hockett¹, Michael D. Larsen¹
Department of Statistics, Iowa State University¹

Abstract

Government agencies simultaneously maintain confidentiality of records and disseminate useful microdata. Iowa's Legislative Services Agency (LSA) submits requests for tax revenue predictions based on proposed policy changes to Iowa's Department of Revenue (IDR), which estimates revenue based on individual income tax returns. IDR cannot by law release the data themselves. We study combining quantile regression, hot deck imputation, and additional confidentiality preserving methods so that IDR can produce a synthetic microdata file for release to LSA. Releasing a synthetic microdata file that does not threaten confidentiality would help both LSA and IDR be more efficient. Several versions of microdata can be imputed to assess uncertainty. In addition to describing the methodology for producing a synthetic data set, measures of disclosure risk to evaluate confidentiality protection and data utility to study the suitability of synthetic data for statistical inference are considered. Methods are studied using simulated data with known properties. The proposed methodology could extend to any number of large establishment survey databases.

Keywords: Data sharing; Disclosure control; Disclosure limitation; Rank swapping; Statistical data protection; Synthetic data

1. Introduction

This paper describes research to develop a tool to allow an organization to release microdata to researchers while continuing to protect the confidentiality of identities and attributes on individual records. Research into the problem of confidentiality protection was motivated by an application at the Iowa Department of Revenue. Potential applications of the methodology are wide ranging and include data from establishment surveys and the American Community Survey of the U.S. Census Bureau. We propose to use a combination of simulating values of certain variables using quantile regression techniques developed by Koenker (2005) and swapping values of remaining variables using hot deck imputation techniques described in Little and Rubin (2002) to provide a "synthetic" data set with low disclosure risk and high

inferential utility. We consider measuring the risk of identity disclosure using probabilities of identifying a record proposed by Duncan and Lambert (1989) and extended by Reiter (2005). We compare the resulting disclosure risk of a data set obtained by implementing our methods versus current methods. We propose to measure data utility for inferential purposes based on marginal empirical distribution function plots and regression estimates obtained from the original data versus the synthetic data. Section 2 describes the motivating application at Iowa's Department of Revenue and summarizes some literature on confidentiality protection. Section 3 presents the methodology for producing synthetic data sets. Sections 4 and 5 discuss measuring disclosure risk and assessing inferential data utility, respectively. Final remarks and comments on future work are given in Section 6.

2. Confidentiality Protection

The initial motivation for research in the area of confidentiality protection was originally brought to my professors at Iowa State University by the Tax Research and Program Analysis Section of the Iowa Department of Revenue (IDR). Their hope was that we could help them to produce a method for generating synthetic or artificial tax return records for use in the Iowa Legislative Services Agency (LSA) tax calculators for determining tax revenue and burden. The end goal has been to provide a tool that will allow IDR to produce a dataset for release to LSA enabling Iowa legislators to explore various changes to the Iowa tax code by assessing the revenue implications of those potential changes without requiring direct IDR involvement with each calculation. Preliminary results, which are described in part in the next section, based on tax return data from IDR and data simulated using quantile regression have shown promising results.

We have since extended the scope of this project to include the development of new methods and procedures for producing confidential microdata as surrogate data for a variety of large databases. Our larger goal is to enhance the availability and usability of microdata in economic, behavioral, and social sciences. Such microdata could arise from surveys of

establishments, households, or individuals. Current work involves a data set from the U.S. Census Bureau based on its American Community Survey (ACS). We anticipate that methods developed using the IDR and ACS databases will be applicable to numerous federal and state data sources and many large-scale behavioral and social surveys.

The issue of confidentiality protection is also referred to in the literature as disclosure control, disclosure prevention or avoidance, and inference control. Books such as those by Willenborg and de Waal (1996, 2001) Domingo-Ferrer (1999, 2002), Domingo-Ferrer and Franconi (2006), Domingo-Ferrer and Torra (2004), and Doyle et al. (2001), a special issue of the *Journal of Official Statistics* (1998), and several technical reports at the National Institute of Statistical Sciences (NISS) (<http://www.niss.org>) have considered these issues and methods for making data available to researchers while preserving confidentiality. The most common existing methods include categorizing continuous data, top-coding sensitive values, swapping records, and noise addition (Reiter 2005a).

The idea for creating a full artificial set of data for public release comes from Rubin (1993). A number of authors have considered methods for implementing the proposal to create such artificial data sets. They include Reiter (2002, 2003, 2005b), Abowd and Woodcock (2004), Raghunathan, Reiter and Rubin (2003), and Little and Liu (2002, 2003). In the next section, we describe our new proposed method for generating a synthetic database.

3 Proposed Method

3.1 Quantile Regression Simulation

The method we propose to simulate a synthetic database is to use quantile regression to estimate quantiles of the conditional distributions of variables such as ages, wage income, or other sensitive or identifying variables, given a collection of demographic variables. The estimated quantiles can in turn be used to simulate values for ages and wage income conditional on demographic variables. Implementing quantile regression using the tax return data from IDR, we found that the proposed method can be used effectively to reproduce both the marginal and conditional distributions of ages and wage income.

3.1.1 Methods

Quantile regression is a method used to fit a curve or surface to a quantile of the response variable at various combinations of the independent variables. Whereas

least squares regression fits a curve or surface to the mean of the response via minimizing the sum of squared errors, quantile regression fits a curve or surface to the quantile of the response via minimizing the (weighted) sum of absolute errors (Koenker and Hallock 2001, Koenker 2005). When the quantile corresponding to $\tau \in (0,1)$ is of interest, the sum of the “tilted absolute value function”, denoted $\rho_\tau(y_i - \hat{y}_i)$, is minimized. That is, if $\xi(x_i, \beta_\tau)$ is any generalized linear function of predictors x_i and parameter vector β_τ , then the full statement of the quantile regression problem is to minimize $\sum \rho_\tau(y_i - \xi(x_i, \beta_\tau))$ over choices of β_τ . This minimization can be formulated so that it becomes a linear programming problem (Koenker and Hallock 2001). Parameter estimation can be performed in R using the ‘rq’ function of the ‘quantreg’ package. An introduction to using functions from the ‘quantreg’ package can be found in Koenker (2004).

3.1.2 Results

In the application for the Iowa Department of Revenue, we consider using quantile regression methods to simulate income tax return data by computing a predicted response for randomly selected quantiles of dependent variables age and wages conditional on key independent variables corresponding to the number of dependents, state filing status, and county reported on an individual income tax return. First, age is generated. Then, conditional on simulated age and the other independent variables, values of wages are generated. Using quantile regression we estimate parameters in this model for each of several quantiles in a sequence between 0 and 1. At each of the τ^{th} quantiles, given values of the independent predictors, an estimated response can be computed. For each record, we randomly select τ^* from a uniform distribution. Then we compute the estimated response by interpolating between the estimated responses corresponding to τ quantiles directly above and below τ^* in the set.

In particular, we simulate age conditional on the demographic predictors mentioned above. The distribution of age is skewed right and has a minimum at zero. After considering several models, we determine that the following conditional quantile function, with age conditioned on groups corresponding to the number of dependents and state filing status, is the best model in terms of producing

simulated, or synthetic ages with distribution most like the distribution of ages in the data:

$$Q_{age}(\tau | x) = \xi(d, fs, c, \beta_\tau) + F_u^{-1}(\tau),$$

where d , fs , and c stand for dependents, filing status, and county, respectively.

As a preliminary check we assess how well we did based on the plotted empirical cumulative distribution functions of both synthetic ages and ages in the data in Figure 1. As can be seen, the simulated and empirical quantiles match very precisely. Checks were also performed conditional on values of the independent variables (dependents, filing status, and county). Results are very promising (Huckett 2006).

In addition to ages, we simulate a sensitive variable in the tax return data set, wages, using the same quantile regression techniques. Again, after considering several models we find that the best model conditions wages on simulated age and the demographic predictors number of dependents, state filing status, and county group. The model is:

$$Q_{wage}(\tau | x) = \xi(age, d, fs, c, \beta_\tau) + F_u^{-1}(\tau).$$

Since the distribution of wages is so extremely skew right in the upper tail, in addition to quantile regression techniques used for simulating ages, we supplement simulation for wages with sampling from a shifted exponential distribution in the upper tail of the distribution.

The distribution of simulated wages is smooth up to roughly the 0.995 quantile at which point the distribution looks more like a step function. In order to attempt to improve the distribution in the upper tail, i.e., make the distribution smooth, we investigate using random draws from an exponential distribution to simulate wages in the upper tail. For randomly selected quantiles 0.995 and above, simulated wages are drawn from an exponential distribution with mean proportional to the distance between the 0.995th and the highest $[100(n-1)/n]^{th}$ quantile.

The resulting distribution is quite close to the actual distribution; see Figure 2 for illustration. Although it underestimates wages, the method offers improvement from using interpolation between quantiles alone (Huckett 2006). We recognize this method still underestimates for very high wages but are quite pleased and accept it for the time being. An interesting topic for future research might be to consider how

close we actually want to come to simulating the very high values with respect to disclosure risk.

3.2 Hot Deck Imputation and Rank Swap

At the initial stages of this study with IDR, we found that performing quantile regression on the entire database of individual income tax returns for any given year to be computationally intensive, if not impossible. At that time, we proposed performing hot deck imputation on remaining sensitive variables as a method to impute other variable values from similar records. We also performed additional perturbation using rank swapping. The result is one synthetic return with data values for key demographic predictors, synthetic values for wage income and ages via quantile regression techniques, and imputed values for additional variables. Since then, through work at the U.S. Census Bureau using a remote server and learning to use Linux and batch R code, the computational effort seems less daunting and it may be possible to synthesize most, if not all variables in a data set. Work to determine if this is realistic remains to be done. Imputation and rank swapping are described in the following paragraphs.

Hot deck imputation is performed using the following notions for imputing missing data, as described in Little and Rubin (2002). Consider a data set consisting of variables found in the IDR database: key demographic variables, ages, wages, and federal tax variables. Consider now, stacking these on another data set with the same values for key demographic variables, the synthetic values for ages and wages obtained using quantile regression methods described above, and missing values for all of federal tax. Records are classified by the key demographic variables and within each classification category we compare donors and candidates where donors are the records in the data and candidates are the records in the synthetic set that need to have federal tax values imputed.

Consider imputation within one classification category at a time. We compare donors and candidates at two levels. At the first level, we compute distances between donors and candidates based on wages only. The (squared) distance used here is

$$d^2(i, j) = (x_{i,w} - x_{j,w})^T S_{xx}^{-1} (x_{i,w} - x_{j,w}),$$

where $x_{i,w}$ and $x_{j,w}$ are values of wages in the donors (original data) and the candidate (synthetic) record, respectively. The estimated variance of the donor wages within the classification category is S_{xx} .

We randomly select 20 donors from the donors with the smallest distances to candidate $x_{j,w}$. At the second level of donor-candidate comparison, we consider (squared) Mahalanobis distance based on both wages and age,

$$d^2(i, j) = (x_i - x_j)^T S_{xx}^{-1} (x_i - x_j),$$

where x_i is a vector of wages and age values for a potential donor, x_j is a vector of wages and ages for the candidate record, and S_{xx} is the estimated variance-covariance matrix of wages and age values within the classification category. The record corresponding to the entry in x_i with the smallest Mahalanobis distance $d(i, j)$ becomes the donor and its value of federal tax is chosen for the candidate return.

An additional step is taken to further perturb the data. We compute the sample rank, r , of the federal tax in the candidate return and then draw a random rank r^* from the Uniform distribution centered at r , uniform[$r - 10, r + 10$]. We then impute the value of federal tax from the record that contains federal tax with rank r^* . These steps are repeated for all candidates x_j in each classification category and for each category in the synthetic set.

The result is a set of $2n$ records, n of which contain the original data, n of which contain the synthetic data. We then “unstack” this set, so to speak, and keep only the n records containing the synthetic data. We applied this method to federal tax variables in the IDR data set and obtained fairly similar empirical cumulative distribution functions, though some improvement in the lower tail may be desired. Figures 3 through 6 present results for total adjustments, deductions, federal tax, and credits on the log scale. The task of more fully assessing conditional distributions remains to be completed.

4. Measuring Disclosure Risk

To evaluate and demonstrate that the proposed quantile regression simulation combined with hot deck imputation method produces synthetic data with improved confidentiality protection over current data disclosure methods, we propose measuring disclosure risk based on a framework developed by Duncan and Lambert (1989) and Reiter (2005a) with extensions to

quantile regression inference. This framework involves assessing threats using the probability of identification of individual records in a released database. Probabilities of identification vary depending on assumptions of an intruder’s behavior and prior knowledge. Variables are separated into sets depending on whether we can assume the intruder has knowledge about an individual’s values prior to data release—and further, if the intruder knows these values exactly or only relative to other individuals’ values—or if the intruder only gains knowledge of the values after data are released. Those variables that the intruder has knowledge of prior to data release are further separated into sets depending on which variables are released with or without any disclosure control applied. Using this framework, the probability of a record being identified given the intruder’s knowledge and the released values for their “target” can be evaluated by splitting the probability into components based on the sets of variables that the intruder does or does not have knowledge about prior to obtaining the released data. We propose to compare historical disclosure limitation methods and quantile regression methods by comparing the probability of identification for some specific records. We anticipate that disclosure risk is substantially less for synthetic data.

Methods based on record linkage and re-identification of records after application of disclosure control methods are described in, for example, Winkler (2004a, 2004b). Future work could consider record linkage based measure of disclosure risk.

5. Quantifying Data Utility

Thus far our assessment of data utility has consisted of visually examining the empirical cumulative distribution functions of variables from the data and from the synthetic sets. We propose to implement more sophisticated measures of data utility, such as methods outlined in numerous papers in the literature by Karr and others at the National Institute of Statistical Sciences. They propose performing the same analysis on each data set (original and synthetic) and comparing results. For example, if one computes a regression and its standard errors for a model that is likely to be of interest to a user separately on the reported and on the synthetic data, then one can compare the resulting estimates, their standard errors, R^2 values, etc. Other utility measures may also be considered.

6. Concluding Remarks

Preliminary work on using quantile regression techniques to create synthetic data for release have shown promise. Much work to justify releasing sensitive data to the public still remains. Planned research to implement and assess disclosure risk measures is currently underway and will be followed by implementation and assessment of disclosure risk and data utility measures as well. Applications of this work in the near future will include databases from Iowa's Department of Revenue and data from the U.S. Census Bureau's American Community Survey.

Acknowledgements

We would like to acknowledge the Iowa Department of Revenue and Iowa's Legislative Services Agency for bringing this problem to the Statistics Department at Iowa State University. We have been intrigued by it and see much future work that lies ahead in order to fully address the confidentiality problem. Dr. Stephen Vardeman at ISU has contributed to the formulation of ideas. We would like to acknowledge the National Institute of Statistical Sciences for funding some of the work done in this research. The U.S. Census Bureau continues to fund this research in part through a gracious Dissertation Fellowship Award. Thanks also to the employees of the U.S. Census Bureau for their openness and encouragement during the summer of 2007.

References

- 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*. New York: Springer.
- Domingo-Ferrer, J. (ed.) (1999), *Statistical Data Protection*. Luxembourg: Office for Official Publications of the European Communities. <http://vneumann.etse.urv.es/sdp98/index.html>
- Domingo-Ferrer, J. (ed.) (2002), *Inference Control in Statistical Databases*. New York: Springer. Lecture Notes in Computer Science 2316.
- Domingo-Ferrer, J. and Franconi, L. (eds.) (2006), *Privacy in Statistical Databases*. New York: Springer. Lecture Notes in Computer Science 4302.
- Domingo-Ferrer, J. and Torra, V. (eds.) (2004), *Privacy in Statistical Databases*. New York: Springer. Lecture Notes in Computer Science 3050.
- Doyle, P., Lane, J.I., Theeuwes, J.J.M, and Zayatz, L.V. (eds.) (2001), *Confidentiality, Disclosure, and Data Access*. Amsterdam: Elsevier Science.
- Duncan, G.T. and Lambert, D. (1989), "The Risk of Disclosure for Microdata," *Journal of Business and Economic Statistics*, 7, 207-217.
- Huckett, J. (2006), "A study of a proposal to simulate tax records for the State of Iowa," Technical report to Iowa Department of Revenue.
- Journal of Official Statistics*. (1998), Special issue on Disclosure Limitation Methods for Protecting the Confidentiality of Statistical Data, 14, 4.
- Koenker, R. (2004), "Quantile Regression in R: A Vignette," <http://www.econ.uiuc.edu/~roger/research/rq/vig.pdf>.
- Koenker, R. (2005), *Quantile Regression*, Econometric Society Monograph Series, Cambridge University Press.
- Koenker, R. and Hallock, K. F. (2001), "Quantile Regression," *Journal of Economic Perspectives*, 15, 4, 143-156.
- Little, R. J. A. and Liu, F. (2002), "Selective Multiple Imputation of Keys for Statistical Disclosure Control in Microdata," *American Statistical Association, Proceedings of the Section on Survey Research Methods*, CD-ROM.
- Little, R. J. A. and Liu, F. (2003), "Comparison of SMIKe with Data-Swapping and PRAM for Statistical Disclosure Control of Simulated Microdata," *American Statistical Association, Proceedings of the Section on Survey Research Methods*, CD-ROM.
- Little, R.J.A. and Rubin, D.B. (2002), *Statistical Analysis with Missing Data*. Wiley, Hoboken, New Jersey.
- National Institute of Statistical Sciences. Technical reports. <http://www.niss.org/technicalreports.html>.
- Raghunathan, T.E., Reiter, J. P., and Rubin, D.R. (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, 531-543.
- Reiter, J.P. (2003), "Inference for Partially Synthetic, Public Use Data Sets," *Survey Methodology*, 29, 181-189.
- Reiter, J.P. (2005a), "Estimating Risks of Identification Disclosure in Microdata," *Journal of the American Statistical Association*, 100, 472, 1103-1113.
- Reiter, J.P. (2005b), "Releasing Multiply Imputed, Synthetic Public-Use Microdata: An Illustration and Empirical Study," *Journal of the Royal Statistical Society, A*, 168, 185-205.

- Rubin, D.B. (1993), Discussion: “Satisfying Confidentiality Constraints through the Use of Synthetic Multiply-Imputed Microdata,” *Journal of Official Statistics*, 9, 2, 461-468.
- Willenborg, L.C.R.J. and deWaal, T. (1996), *Statistical Disclosure Control in Practice*. Springer-Verlag, New York.
- Willenborg, L.C.R.J. and deWaal, T. (2001), *Elements of Statistical Disclosure Control*. Springer-Verlag, New York.

- Winkler, W. E. (2004a), “Masking and Re-identification Methods for Public-Use Microdata: Overview and Research Problems,” in (J. Domingo-Ferrer and V. Torra, eds.) *Privacy in Statistical Databases*. Springer: New York, 231-247.
- Winkler, W. E. (2004b), “Re-identification Methods for Masked Microdata,” in (J. Domingo-Ferrer and V. Torra, eds.) *Privacy in Statistical Databases*. Springer: New York, 216-230.

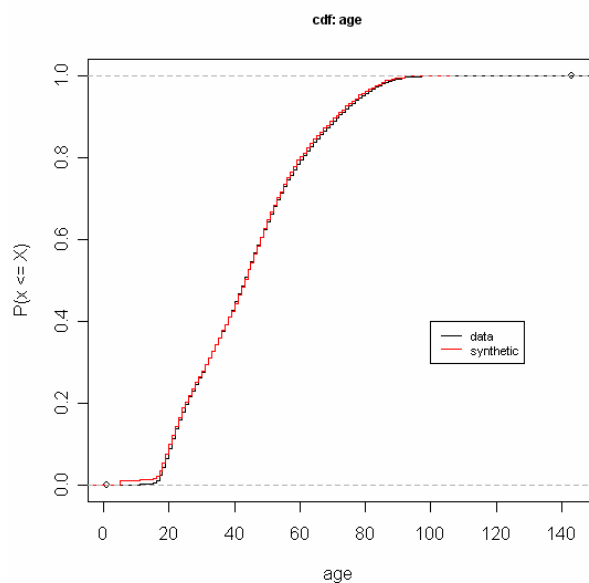


Figure 1: Empirical distribution functions of synthetic ages and ages in data.

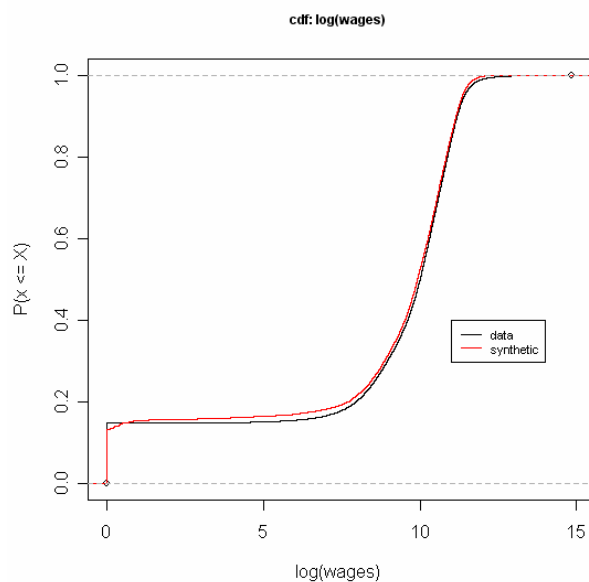


Figure 2: Empirical distribution of synthetic and reported log wages income.

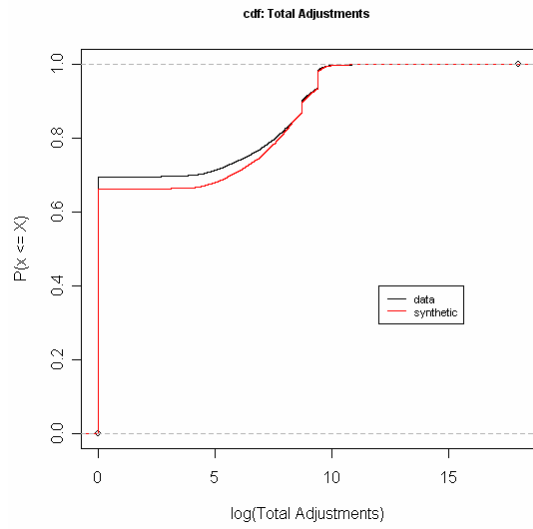


Figure 3: Empirical distribution of synthetic and reported log total adjustments.

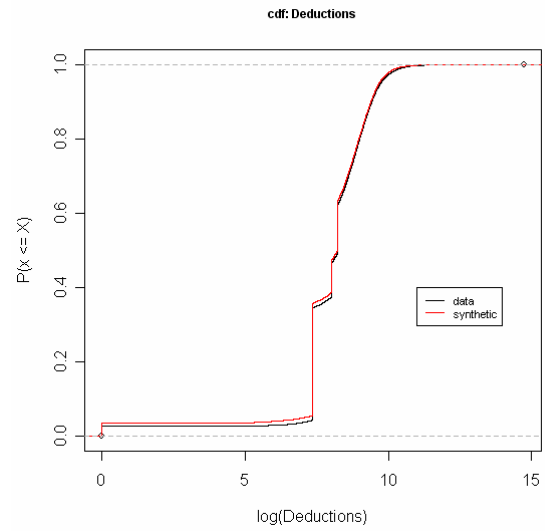


Figure 5: Empirical distribution of synthetic and reported log deductions.

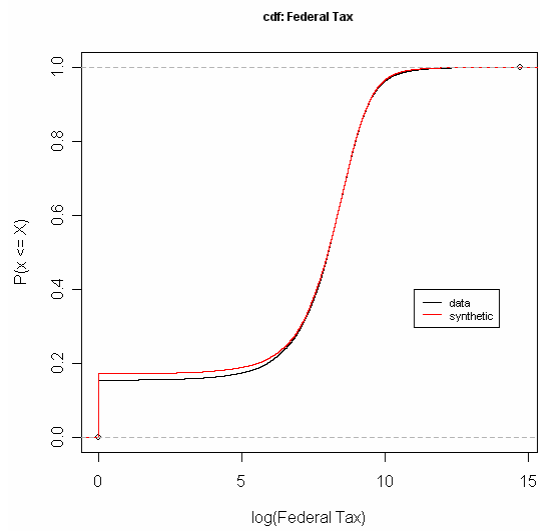


Figure 4: Empirical distribution of synthetic and reported log federal tax.

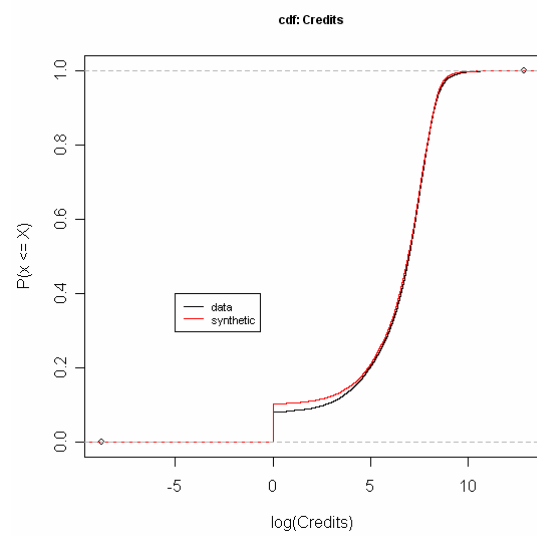


Figure 6: Empirical distribution of synthetic and reported log credits.