RESEARCH REPORT SERIES
(Statistics #2018-02)

Inference for Multivariate Regression Model based on Synthetic Data generated using Plug-in Sampling

Ricardo Moura
CMA, Faculty of Sciences and Technology, Nova University of Lisbon

Martin Klein
Center for Statistical Research and Methodology, U.S. Census Bureau

John Zylstra
Department of Mathematics and Statistics,
University of Maryland, Baltimore County

Carlos Coelho
CMA and Mathematics Department, Faculty of Sciences and Technology,
Nova University of Lisbon

Bimal Sinha
Department of Mathematics and Statistics,
University of Maryland, Baltimore County
and Center for Statistical Research and Methodology, U.S. Census Bureau

Center for Statistical Research and Methodology
Research and Methodology Directorate
U.S. Census Bureau
Washington, D.C. 20233

Report Issued: February 5, 2018 (Updated: March 9, 2021)

Disclaimer: This report is released to inform interested parties of research and to encourage discussion. The views expressed are those of the authors and not those of the U.S. Census Bureau.

Inference for Multivariate Regression Model based on Synthetic Data generated using Plug-in Sampling

Ricardo Moura
Center for Mathematics and Applications (CMA/UNL),
NOVA School of Science and Technology, NOVA University of Lisbon
Portuguese Navy Research Center (CINAV)
and Naval Academy, Alfeite, Almada
Martin Klein*
Division of Biometrics VIII, Office of Biostatistics,
Office of Translational Sciences, Center for Drug Evaluation and Research,
U.S. Food and Drug Administration,
Silver Spring, Maryland
John Zylstra
Department of Mathematics and Statistics,
University of Baltimore, Baltimore County (UMBC)
Carlos A. Coelho
Center for Mathematics and Applications (CMA/UNL)
and Mathematics Department,
NOVA School of Science and Technology, NOVA University of Lisbon
Bimal Sinha*
Department of Mathematics and Statistics,
University of Baltimore, Baltimore County (UMBC)
and Center for Statistical Research and Methodology (CSRM), U.S. Census Bureau

*__Disclaimer:__ This article reflects the views of the authors and should not be construed to represent the views and/or policies of neither the U.S. Food and Drug Administration or the U.S. Census Bureau

Abstract

In this paper, the authors derive the likelihood-based exact inference for *singly* and *multiply* imputed synthetic data in the context of a multivariate regression model. The synthetic data are generated via the Plug-in Sampling method, where the unknown parameters in the model are set equal to the observed values of their point estimators based on the original data, and synthetic data are drawn from this estimated version of the model. Simulation studies are carried out in order to confirm the theoretical results. In case multiple synthetic datasets are permissible, the authors provide an *exact* test procedure and compare their results with the asymptotic results of Reiter (2005a). The authors provide *exact* test procedures, which in case multiple synthetic datasets are permissible, are compared with the asymptotic results of Reiter (2005a). An application using 2000 U.S. Current Population Survey public use data is discussed. Furthermore, properties of the proposed methodology are evaluated in scenarios where some of the conditions that were used to derive the methodology do not hold, namely for nonnormal and discrete distributed random variables, cases in which the inferential procedures developed still show very good performances.

Key Words: Data Confidentiality, Finite sample analysis, Maximum likelihood estimators, Multivariate Regression, Partially Synthetic Data, Pivotal quantities, Plug-in Sampling, Statistical Disclosure Control.

1 Introduction

Methods of statistical disclosure control are used to achieve the competing goals of publishing statistical outputs from surveys, while protecting the survey respondents' confidential data from disclosure. Statistical disclosure control methods include data swapping, perturbation with randomly added or multiplied noise, and the release of synthetic data. The use of synthetic datasets has gained considerable popularity and importance in recent times (Klein et al., 2013). In this paper, we investigate some inferential aspects of statistical analysis based on synthetic data for situations when either a single or multiple synthetic datasets based on the original data are created as substitute for publication and analysis. Little (1993) and Rubin (1993) first advocated the use of synthetic data for statistical disclosure control, using the framework of multiple imputation (Rubin, 1987). Rubin (1993) argued that synthetic data so created do not correspond to any actual sampling unit, thus preserving the confidentiality of the respondents. Inferential methods for fully synthetic data were developed by Raghunathan et al. (2003). Reiter (2005a) presented an illustration and empirical study of fully synthetic data and Reiter and Raghunathan (2007) provided an overview of multiple imputation techniques, including its use in statistical disclosure control. Reiter (2003) presented methods for drawing inference for partially synthetic data. This is exactly the context of our paper.

There are two main methods one can use to generate synthetic data: Posterior Predictive Sampling and Plug-in Sampling (Reiter and Kinney, 2012), and statistical methods of data analysis can be developed for both methods.

Although most inferential methods for synthetic data are based on multiple imputation, Klein and Sinha (2015a,b,c, 2016)) in a series of recent papers developed exact parametric inferential methods based on singly imputed synthetic data for several probability models, including the multiple linear regression model where the sole response variable is taken as sensitive, thus requiring protection, while the covariates are treated as non-sensitive. There are cases where singly imputed synthetic data have been released (Hawala, 2008; Kinney et al., 2011, 2014), and therefore procedures for valid data analysis for this case are desirable.

Our main objective in this paper is to extend this scenario to the case of a multivariate linear regression model where there are multiple sensitive responses following a multivariate normal distribution with means modeled as linear combinations of multiple non-sensitive covariates. Based on the fitted multivariate linear regression model, we synthesize the sensitive responses based on the Plug-in Sampling method, and develop exact inferential data analysis procedures for both single and multiple imputation.

A brief description of the Plug-in Sampling method, which will be used throughout the paper, follows. Suppose that $\mathbf{Y} = (\mathbf{y_1}, ..., \mathbf{y_n})$ are the original data which are jointly distributed according to the probability density function (pdf) $f_{\boldsymbol{\theta}}(\mathbf{Y})$, where $\boldsymbol{\theta}$ is the unknown (scalar, vector or matrix) parameter. We start by taking the value of a point estimator $\hat{\boldsymbol{\theta}}(\mathbf{Y})$ of $\boldsymbol{\theta}$, and plug it into the joint pdf of $\mathbf{Y}$. The resulting pdf, with the unknown $\boldsymbol{\theta}$ replaced by the observed value of the point estimator $\hat{\boldsymbol{\theta}}(\mathbf{Y})$, is denoted by $f_{\hat{\boldsymbol{\theta}}}$. The singly imputed synthetic data, denoted by $\mathbf{V}$, are then generated by drawing $\mathbf{V} = (\mathbf{v_1}, ..., \mathbf{v_n})$ from the joint pdf $f_{\hat{\boldsymbol{\theta}}}(\mathbf{Y})$. In case of multiple imputation, this procedure is independently repeated M times to generate M synthetic datasets.

In terms of the multivariate linear regression model, in our context, we consider several *sensitive* response variables $y_j, j = 1, ..., m$, originating the vector of response variables $\mathbf{y} = (y_1, ..., y_m)'$, and a set of p non-*sensitive* predictors $\mathbf{x} = (x_1, ..., x_p)'$. We assume that $\mathbf{y}|\mathbf{x} \sim N_m(\mathbf{B}'\mathbf{x}, \boldsymbol{\Sigma})$, with $\mathbf{B}$ and $\boldsymbol{\Sigma}$ unknown. We write $\mathbf{Y} = (\mathbf{y_1}, ..., \mathbf{y_n})$ with $\mathbf{y_i} = (y_{1i}, ..., y_{mi})'$ and $\mathbf{X} = (\mathbf{x_1}, ..., \mathbf{x_n})$ with $\mathbf{x_i} = (x_{1i}, ..., x_{pi})'$. We also assume that $rank(\mathbf{X} : p \times n) = p < n$ and $n \geq m + p$. We are thus considering the following multivariate regression model

$$\mathbf{Y}_{m\times n} = \mathbf{B}'_{m\times p}\mathbf{X}_{p\times n} + \mathbb{E}_{m\times n} \tag{1}$$

where $\mathbb{E}_{m\times n}$ is distributed as $N_{mn}(\mathbf{0}, \mathbf{I_n} \otimes \boldsymbol{\Sigma})$. It is well known that, based on the original data, $\hat{\mathbf{B}} = (\mathbf{X}\mathbf{X}')^{-1}\mathbf{X}\mathbf{Y}'$ is the MLE and the UMVUE of $\mathbf{B}$, distributed as $N_{pm}(\mathbf{B}, \boldsymbol{\Sigma} \otimes (\mathbf{X}\mathbf{X}')^{-1})$, independent of $\hat{\boldsymbol{\Sigma}} = \frac{1}{n}(\mathbf{Y} - \hat{\mathbf{B}}'\mathbf{X})(\mathbf{Y} - \hat{\mathbf{B}}'\mathbf{X})'$ which is the MLE of $\boldsymbol{\Sigma}$, with $n\hat{\boldsymbol{\Sigma}} \sim W_m(\boldsymbol{\Sigma}, n - p)$, and therefore $\mathbf{S} = \frac{n\hat{\boldsymbol{\Sigma}}}{n-p}$ will be the unbiased estimator of $\boldsymbol{\Sigma}$.

There are several tests for $\mathbf{B}$, based on the original data, in the literature (Anderson, 2003). In this paper, the authors will develop two new procedures to be used with synthetic data to draw inference for $\mathbf{B}$, and also for $\mathbf{C} = \mathbf{AB}$ and $\boldsymbol{\Delta} = \mathbf{ABD}$ where $\mathbf{A}$ is a $k \times p$ matrix with $rank(\mathbf{A}) = k \leq p$ and $k \geq m$, and $\mathbf{D}$ is an $m \times r$ matrix with $rank(\mathbf{D}) = r \leq k$.

The organization of the paper is as follows. In Section 2, based on singly and multiply imputed synthetic data generated via Plug-in Sampling, we develop two exact inference procedures for the matrix of regression coefficients $\mathbf{B}$. These will be based on pivot statistics which are different from the classical test statistics for $\mathbf{B}$ under this model (see Anderson (2003)). These classical statistics are shown to be non-pivotal in the case of imputed synthetic data generated via Plug-in Sampling. The new exact inferential procedures are compared with Reiter's asymptotic methodology for multiple imputation synthetic data (Reiter, 2005a). In Section 3, we present some simulation results in order to check the accuracy of the theoretically derived results for the singly imputed and multiply imputed synthetic data, comparing the latter with the results obtained using an adaptation of Reiter's methodology. We also define the *radius* (distance between the center and the edge) of the confidence sets for the matrix of regression coefficients $\mathbf{B}$, both for the original data, as well as for the singly and the multiply imputed synthetic data. The Plug-in Sampling method offers smaller *radius* of the confidence sets than the Posterior Predictive Sampling (PPS) method and also gives estimates of the parameters closer to the ones obtained from the original data, despite giving slightly higher levels of disclosure risk (Moura, 2016). Section 4 presents data analyses under the proposed methods for singly and multiply imputed synthetic data in the context of public use data from the 2000 U.S. Current Population Survey (CPS), and the results are compared with those obtained from the original data. In Section 5, using the CPS data, the authors present an evaluation of the level of protection of the released synthetic datasets by comparing single and multiple imputation scenarios. Some concluding remarks are added in Section 6. Proofs of the theorems, corollaries, and other technical derivations appear in Appendices A and B.

We conclude this section with an observation regarding the existence of *sufficient statistics*. Suppose the original data are $\mathbf{Y} \sim f_{\boldsymbol{\theta}}$, and the synthetic data $\mathscr{V} = (\mathbf{V}_1, \ldots, \mathbf{V}_M)$ are generated such that $\mathbf{V}_1, \ldots, \mathbf{V}_M | \mathbf{Y}$ are *iid* from $f_{\hat{\boldsymbol{\theta}}}$. Suppose that $\mathbf{T}(\mathbf{Y})$ is a sufficient statistic for $\boldsymbol{\theta}$ based on the original data. Then the pdf of the synthetic data $\mathscr{V} = (\mathbf{V}_1, \ldots, \mathbf{V}_M)$ is

$$\int \Big\{\prod_{j=1}^{M} f_{\hat{\boldsymbol{\theta}}(\mathbf{Y})}(\mathbf{V}_j)\Big\} f_{\boldsymbol{\theta}}(\mathbf{Y})d\mathbf{Y} = \int \Big\{\prod_{j=1}^{M} g_{\hat{\boldsymbol{\theta}}(\mathbf{Y})}\left(\mathbf{T}(\mathbf{V}_j)\right) h(\mathbf{V}_j)\Big\} f_{\boldsymbol{\theta}}(\mathbf{Y})d\mathbf{Y}$$
$$= \Big\{\prod_{j=1}^{M} h(\mathbf{V}_j)\Big\} \int \Big\{\prod_{j=1}^{M} g_{\hat{\boldsymbol{\theta}}(\mathbf{Y})}\left(\mathbf{T}(\mathbf{V}_j)\right)\Big\} f_{\boldsymbol{\theta}}(\mathbf{Y})d\mathbf{Y},$$

which implies the following result.

Lemma 1.1. *Suppose that when the original data* $\mathbf{Y}$ *are observed,* $\mathbf{T}(\mathbf{Y})$ *is a sufficient statistic for* $\boldsymbol{\theta}$*. Then when the synthetic data* $\mathscr{V} = (\mathbf{V}_1, \ldots, \mathbf{V}_M)$ *are made available,* $(\mathbf{T}(\mathbf{V}_1), \ldots, \mathbf{T}(\mathbf{V}_M))$ *is jointly sufficient for* $\boldsymbol{\theta}$*. Furthermore, if* $M = 1$*, the sufficient statistic is simply* $\mathbf{T}(\mathbf{V}_1)$*, and if* $M > 1$*, then* $\sum_{j=1}^{M} \mathbf{T}(\mathbf{V}_j)$ *is sufficient if* $f_{\boldsymbol{\theta}}(\mathbf{Y}) = h(\mathbf{Y})\psi(\boldsymbol{\theta})\exp\{\gamma(\boldsymbol{\theta})'\mathbf{T}(\mathbf{Y})\}$*, i.e., if* $f_{\boldsymbol{\theta}}(\mathbf{Y})$ *belongs to the exponential family.*

2 Analysis under Single and Multiple Imputation

In this section, a likelihood-based approach for analysis of synthetic data generated from a multivariate regression model is presented for the Plug-in Sampling method. First, we provide two new and exact inferential procedures based on the likelihood principle for single and multiple imputation synthetic data (for $M = 1$, the single imputation case, both procedures concur) and then work out an adaptation of Reiter's method for our setup.

Consider the multivariate linear regression model (1) with $\mathbf{Y}$, $\mathbf{X}$, $\mathbf{B}$, $\boldsymbol{\Sigma}$, $\hat{\mathbf{B}}$ and $\mathbf{S}$ defined in that same context.

The synthetic data consist of M (≥ 1) synthetic versions of $\mathbf{Y}$ generated based on the Plug-in method as described below. To consider the single imputation case one only has to take $M = 1$. From the original data $(y_{i1}, ..., y_{im}, x_{1i}, ..., x_{pi}), i = 1, ..., n$, after estimating $\mathbf{B}$ and $\boldsymbol{\Sigma}$ by $\hat{\mathbf{B}}$ and $\mathbf{S}$, respectively, we generate the synthetic data, denoted as $\mathbf{V}_j = (\mathbf{v}_1^{(j)}, ..., \mathbf{v}_n^{(j)}), j = 1, \ldots, M$, where $\mathbf{v_i^{(j)}} = (v_{1i}^{(j)}, ..., v_{mi}^{(j)})'$, $i = 1, \ldots, n$, are independently distributed as

$$\mathbf{v}_i^{(j)}|(\hat{\mathbf{B}},\mathbf{S}) \sim N_m(\hat{\mathbf{B}}'\mathbf{x_i},\mathbf{S}). \tag{2}$$

Our goal is to draw inference on $\mathbf{B}$ based on the partial synthetic data $(v_{1i}^{(j)},...,v_{mi}^{(j)},x_{1i},...,x_{pi})$, for $i=1,...,n$ and $j=1,\dots,M$.

2.1 A First New Procedure based on the mean synthetic covariances

Towards the aforementioned objective of drawing inference on $\mathbf{B}$, based on the partial synthetic data, let $\mathbf{B}_j^* = (\mathbf{X}\mathbf{X}')^{-1}\mathbf{X}\mathbf{V}_j'$ and $\mathbf{S}_j^* = \frac{1}{n-p}(\mathbf{V}_j - \mathbf{B}_j^{*\prime}\mathbf{X})(\mathbf{V}_j - \mathbf{B}_j^{*\prime}\mathbf{X})'$ be the estimators of $\mathbf{B}$ and $\mathbf{\Sigma}$ based on $\mathbf{V}_j$. Conditionally on $(\hat{\mathbf{B}},\mathbf{S})$, for any $j=1,...,M$, $\mathbf{B}_j^*$ is independent of $\mathbf{S}_j^*$ and $\{(\mathbf{B}_1^*,\mathbf{S}_1^*),\dots,(\mathbf{B}_M^*,\mathbf{S}_M^*)\}$, by Lemma 1.1, are jointly sufficient estimators for $\mathbf{B}$ and $\mathbf{\Sigma}$. Let us also define

$$\overline{\mathbf{B}}_M^* = \frac{1}{M}\sum_{j=1}^{M}\mathbf{B}_j^* \quad \text{and} \quad \overline{\mathbf{S}}_M^* = \frac{1}{M}\sum_{j=1}^{M}\mathbf{S}_j^*, \tag{3}$$

which are mutually independent, conditionally on $\hat{\mathbf{B}}$ and $\mathbf{S}$. The main *inferential* results we derive are, for $p \geq m$ and $n \geq m+p$,

1. $\overline{\mathbf{B}}_M^*$ is an unbiased estimator for $\mathbf{B}$, with $Var(\overline{\mathbf{B}}_M^*) = \frac{M+1}{M}\mathbf{\Sigma}\otimes(\mathbf{X}\mathbf{X}')^{-1}$ (see Appendix B.1),
2. an unbiased estimator of $\mathbf{\Sigma}$ is $\overline{\mathbf{S}}_M^*$ (see Appendix B.1),
3. we prove in Corollary 2.2 (see below) that

$$T_M = \frac{\left|(\overline{\mathbf{B}}_M^* - \mathbf{B})'(\mathbf{X}\mathbf{X}')(\overline{\mathbf{B}}_M^* - \mathbf{B})\right|}{\left|(n-p)\overline{\mathbf{S}}_M^*\right|} \tag{4}$$

is a pivotal quantity and, for $\mathbf{W} \sim W_m(\mathbf{I}_m, n-p)$ and $F_{p,l} \sim \mathcal{F}_{p-l+1,M(n-p)-l+1}$ (where $\mathcal{F}_{m,n}$ denotes an F distribution with m and n degrees of freedom),

$$T_M|\mathbf{W} \overset{st}{\sim} \left\{ \prod_{l=1}^{m} \frac{p-l+1}{M(n-p)-l+1} F_{p,l} \right\} |M(n-p)\mathbf{W}^{-1} + \mathbf{I}_m|,$$

where $\overset{st}{\sim}$ means 'stochastically equivalent to',

4. if one wants to test the significance of a set of regression coefficients or more generally of a linear combination of these regression coefficients, $\mathbf{AB} = \mathbf{C}$ where $\mathbf{A}$ is a $k \times p$ matrix with $rank(\mathbf{A}) = k \leq p$ and $k \geq m$, one may define

$$T_{M,\mathbf{C}} = \frac{|(\mathbf{A}\overline{\mathbf{B}}_M^* - \mathbf{C})'(\mathbf{A}(\mathbf{XX}')^{-\mathbf{1}}\mathbf{A}')^{-1}(\mathbf{A}\overline{\mathbf{B}}_M^* - \mathbf{C})|}{|(n-p)\overline{\mathbf{S}}_M^*|}$$

and proceed by noting that, for $\mathbf{W} \sim W_m(\mathbf{I}_m, n-p)$ and $F_{k,l} \sim \mathcal{F}_{k-l+1,M(n-p)-l+1}$,

$$T_{M,\mathbf{C}}|\mathbf{W} \overset{st}{\sim} \left\{ \prod_{l=1}^{m} \frac{k-l+1}{M(n-p)-l+1} F_{k,l} \right\} |M(n-p)\mathbf{W}^{-1} + \mathbf{I}_m|; \qquad (5)$$

then, we may build

(i) *a test for the parameter matrix* $\mathbf{C}$*:* in order to test $H_0 : \mathbf{C} = \mathbf{C}_0$ versus $H_1 : \mathbf{C} \neq \mathbf{C}_0$ at a given γ level, we reject H_0 whenever the computed value of $T_{M,\mathbf{C}_0}$ exceeds its $1-\gamma$ quantile; in particular a test for $\mathbf{B} = \mathbf{B}_0$ follows upon taking $\mathbf{A} = \mathbf{I}_p$,

(ii) *a confidence set for* $\mathbf{C}$*:* a $(1-\gamma)$-level confidence set for $\mathbf{C}$ is given by

$$\Delta(\mathbf{C}) = \{\mathbf{C} : T_{M,\mathbf{C}} \leq \delta_{M,k,m,p,n;\gamma}\}, \qquad (6)$$

where $\delta_{M,k,m,p,n;\gamma}$ is the $1-\gamma$ quantile of $T_{M,\mathbf{C}}$ (the value of $\delta_{M,k,m,p,n;\gamma}$ can be obtained by simulating the distribution of $T_{M,\mathbf{C}}$, by first generating $\mathbf{W} \sim W_m(\mathbf{I}_m, n-p)$ and then generating $T_{M,\mathbf{C}}|\mathbf{W}$ from (5)),

5. to infer about $\mathbf{ABD} = \boldsymbol{\Delta}$ where $\mathbf{A}$ is a $k \times p$ matrix and $\mathbf{D}$ is a $m \times r$ with $r \leq k$, we start from its natural point estimator $\boldsymbol{\Delta}_M^* = \mathbf{A}\overline{\mathbf{B}}_M^*\mathbf{D}$ and propose to use the pivotal

quantity (see Corollary 2.3)

$$T_{M,\boldsymbol{\Delta}} = \frac{|(\boldsymbol{\Delta}^*_M - \boldsymbol{\Delta})' \left(\mathbf{A}(\mathbf{X}\mathbf{X}')^{-1}\mathbf{A}'\right) (\boldsymbol{\Delta}^*_M - \boldsymbol{\Delta})|}{|(n-p)\mathbf{D}'\overline{\mathbf{S}}^*_M\mathbf{D}|} \tag{7}$$

whose distribution is obtained from the relation

$$T_{M,\boldsymbol{\Delta}}|\mathbf{W}^* \overset{st}{\sim} \left\{\prod_{l=1}^{r} \frac{k-l+1}{M(n-p)-l+1} F_{k,l}\right\} \left|M(n-p)\mathbf{W}^{*-1} + \mathbf{I}_r\right|$$

where $F_{k,l} \sim \mathcal{F}_{k-l+1,M(n-p)-l+1}$ and $\mathbf{W}^* \sim W_r(\mathbf{I}_r, n-p)$, all independently; taking $r = 1$ and $k = 1$, and making $A : 1 \times p$ a matrix of zeros except for $\mathbf{A}_{1,g} = 1$, and $\mathbf{D} : m \times 1$ a matrix of zeros except for $\mathbf{D}_{h,1} = 1$, for $g = 1, \ldots, p$ and $h = 1, \ldots, m$ we may observe that

$$T_{M,\boldsymbol{\Delta}} = T_{M,\mathbf{B}_{(g,h)}} = \frac{|(\overline{\mathbf{B}}^*_{M(g,h)} - \mathbf{B}_{(g,h)})' \left(\mathbf{A}(\mathbf{X}\mathbf{X}')^{-1}\mathbf{A}\right) (\overline{\mathbf{B}}^*_{M(g,h)} - \mathbf{B}_{(g,h)})|}{|(n-p)\mathbf{D}'\overline{\mathbf{S}}^*_M\mathbf{D}|}$$

thus concluding that the $(1-\alpha)$ confidence interval for $\mathbf{B}_{(g,h)}$ will be given by

$$\overline{\mathbf{B}}^*_{M(g,h)} \pm \sqrt{q^*_{M,1-\alpha}(n-p)\overline{\mathbf{S}}^*_{M(h,h)}(\mathbf{X}\mathbf{X}')^{-1}_{g,g}},$$

with $q^*_{M,1-\alpha}$ being the value of the $1-\alpha$ cut-off point of the distribution of $T_{M,\boldsymbol{\Delta}}$, noting that $\mathbf{D}'\mathbf{S}^*\mathbf{D} = \mathbf{S}^*_{(h,h)}$ and $\mathbf{A}(\mathbf{X}\mathbf{X}')^{-1}\mathbf{A}' = (\mathbf{X}\mathbf{X}')^{-1}_{(g,g)}$ (for details in the proof of this result see Section S3 of Part II of the supplementary material).

Results in 1-5 are derived based on the following Theorem and Corollaries, whose proofs are provided in Appendix A.

Theorem 2.1. *The joint pdf of* $(\overline{\mathbf{B}}^*_M, \overline{\mathbf{S}}^*_M)$ *defined in (3) is proportional to*

$$\int \exp\left\{-\frac{1}{2} tr\left[(\boldsymbol{\Sigma} + \frac{1}{M}\mathbf{S})^{-1}(\overline{\mathbf{B}}^*_M - \mathbf{B})'(\mathbf{X}\mathbf{X}')(\overline{\mathbf{B}}^*_M - \mathbf{B}) + M(n-p)\mathbf{S}^{-1}\overline{\mathbf{S}}^*_M + (n-p)\boldsymbol{\Sigma}^{-\mathbf{1}}\mathbf{S}\right]\right\}$$
$$\times |\overline{\mathbf{S}}^*_M|^{\frac{M(n-p)-m-1}{2}} \times \frac{|\mathbf{S}|^{-\frac{M(n-p)-n+2p+m+1}{2}}}{|\boldsymbol{\Sigma}|^{\frac{n}{2}}} \times |\boldsymbol{\Sigma}^{-\mathbf{1}} + M\mathbf{S}^{-1}|^{-p/2}\, d\mathbf{S},$$

from which we can infer that, conditional on $\mathbf{S}$, $\overline{\mathbf{B}}^*_M$ *and* $\overline{\mathbf{S}}^*_M$ *are independent, with* $\overline{\mathbf{B}}^*_M|\mathbf{S} \sim N_{pm}(\mathbf{B}, (\boldsymbol{\Sigma} + \frac{1}{M}\mathbf{S}) \otimes (\mathbf{X}\mathbf{X}')^{-1})$ *and* $M(n-p)\overline{\mathbf{S}}^*_M|\mathbf{S} \sim W_m(\mathbf{S}, M(n-p))$.

Corollary 2.2. *The distribution of T_M defined in (4) can be obtained from the decomposition*

$$T_M|\mathbf{W} \overset{st}{\sim} \left\{ \prod_{l=1}^{m} \frac{p-l+1}{M(n-p)-l+1} F_{p,l} \right\} \left| M(n-p)\mathbf{W}^{-1} + \mathbf{I}_m \right|$$

where $\mathbf{W} \sim W_m(\mathbf{I}_m, n-p)$ and $F_{p,l} \sim \mathcal{F}_{p-l+1,M(n-p)-l+1}$. This implies that T_M is a pivotal quantity, that is, its distribution does not depend on $\mathbf{\Sigma}$.

Corollary 2.3. *The distribution of $T_{M,\mathbf{\Delta}}$ defined in (7) can be obtained from the decomposition*

$$T_{M,\mathbf{\Delta}}|\mathbf{W}^* \overset{st}{\sim} \left\{ \prod_{l=1}^{r} \frac{k-l+1}{M(n-p)-l+1} F_{k,l} \right\} \left| M(n-p)\mathbf{W}^{*-1} + \mathbf{I}_r \right|$$

where $\mathbf{W} \sim W_r(\mathbf{I}_r, n-p)$ and $F_{k,l} \sim \mathcal{F}_{k-l+1,M(n-p)-l+1}$, thus implying that $T_{M,\mathbf{\Delta}}$ is a pivotal quantity.

We may refer that all the above results remain valid for $M = 1$, that is, the single imputation case, for which inferential procedures were not available in the literature.

Remark 2.1. *When $m = 1$ and $M = 1$, the statistic T_M in (4) reduces to the statistic T^2 used in (Klein and Sinha, 2015a) which has a distribution obtained from the fact that*

$$T^2|_{W=w} \sim \frac{p}{n-p}\left[1 + \frac{n-p}{w}\right]\mathcal{F}_{p,n-p} \quad \text{where} \quad f_W(w) = \frac{1}{2^{\frac{n-p}{2}}\Gamma\left(\frac{n-p}{2}\right)} e^{-\frac{w}{2}} w^{\frac{n-p}{2}-1}. \qquad \square$$

Remark 2.2. *One could think that for our context we could suggest the use of the following adaptations of the classical test criteria for the multivariate regression model (see Anderson (2003) for the classical test criteria)*

(a) $T^{(1)} = \frac{|\overline{\mathbf{S}}^*_M|}{|\overline{\mathbf{S}}^*_M + (\overline{\mathbf{B}}^*_M - \mathbf{B})'(\mathbf{X}\mathbf{X}')(\overline{\mathbf{B}}^*_M - \mathbf{B})|}$ *(Wilks' Lambda Criterion)*

(b) $T^{(2)} = tr\left[(\overline{\mathbf{B}}^*_M - \mathbf{B})'(\mathbf{X}\mathbf{X}')(\overline{\mathbf{B}}^*_M - \mathbf{B})(\overline{\mathbf{S}}^*_M)^{-1}\right]$ *(Pillai's Trace Criterion)*

(c) $T^{(3)} = tr\left((\overline{\mathbf{B}}^*_M - \mathbf{B})'(\mathbf{X}\mathbf{X}')(\overline{\mathbf{B}}^*_M - \mathbf{B})[(\overline{\mathbf{B}}^*_M - \mathbf{B})'(\mathbf{X}\mathbf{X}')(\overline{\mathbf{B}}^*_M - \mathbf{B}) + \overline{\mathbf{S}}^*_M]^{-1}\right)$
(Hotelling-Lawley Trace Criterion)

(d) $T^{(4)} = \lambda_1$ *where* λ_1 *denotes the largest eigenvalue of* $(\overline{\mathbf{B}}^*_M - \mathbf{B})'(\mathbf{X}\mathbf{X}')(\overline{\mathbf{B}}^*_M - \mathbf{B})(\overline{\mathbf{S}}^*_M)^{-1}$ *(Roy's Largest Root Criterion),*

but, unfortunately, these test statistics are non-pivotal statistics (see Appendix B.1), opposite to what happens with the statistics T_M, $T_{M,\mathbf{C}}$ *and* $T_{M,\boldsymbol{\Delta}}$.

2.2 A Second New Procedure based on a combination of mean and cross synthetic covariances

Noting that it will be possible to gather more information from the released synthetic data we propose, in this Subsection, another likelihood-based approach for exact inference about $\mathbf{B}$, based on a combination of mean and cross synthetic covariances. Let us start by recalling that $\mathbf{V}_j$ is a $m \times n$ matrix formed by the vectors $(\mathbf{v}_1^{(j)}, ..., \mathbf{v}_n^{(j)})$, generated assuming $\mathbf{v}_i^{(j)} | \hat{\mathbf{B}}, \mathbf{S} \sim N_m(\hat{\mathbf{B}}'\mathbf{x_i}, \mathbf{S}), i = 1, ..., n$. Supposing that we have access to the M imputations $\mathbf{V}_1, ..., \mathbf{V}_M$, with $\mathbf{V}_j = (\mathbf{v}_1^{(j)}, ..., \mathbf{v}_n^{(j)})$, $j = 1, ..., M$, and noting that, conditionally on $\hat{\mathbf{B}}$ and $\mathbf{S}$, $(\mathbf{v}_i^{(1)}, ..., \mathbf{v}_i^{(M)})$ is a random sample from $N_m(\hat{\mathbf{B}}'\mathbf{x_i}, \mathbf{S})$, for any $i = 1, ..., n$, let us consider $\overline{\mathbf{v}}_i = \frac{1}{M}\sum_{j=1}^M \mathbf{v}_i^{(j)}$ and $\mathbf{S}_{\mathbf{v}i} = \sum_{j=1}^M (\mathbf{v}_i^{(j)} - \overline{\mathbf{v}}_i)(\mathbf{v}_i^{(j)} - \overline{\mathbf{v}}_i)'$ which are the sufficient statistics for $\boldsymbol{\Sigma}$, based on the i-th vector of covariates. Defining $\mathbf{S}_\mathbf{v} = \sum_{i=1}^n \mathbf{S}_{\mathbf{v}i}$, we have $(\overline{\mathbf{v}}_1, ..., \overline{\mathbf{v}}_n, \mathbf{S}_\mathbf{v})$ as the joint sufficient statistics for $(\mathbf{B}, \boldsymbol{\Sigma})$. Conditionally on $\hat{\mathbf{B}}$ and $\mathbf{S}$, we have $\overline{\mathbf{v}}_i \sim N_m(\hat{\mathbf{B}}'\mathbf{x_i}, \frac{1}{M}\mathbf{S})$ and $\mathbf{S}_\mathbf{v} \sim W_m(\mathbf{S}, n(M-1))$ since $\mathbf{S}_{\mathbf{v}i} \sim W_m(\mathbf{S}, M-1)$.

From the M released synthetic data matrices $\mathbf{V}_j, j = 1, ..., M$, we may define $\overline{\mathbf{V}}_M = \frac{1}{M}\sum_{j=1}^M \mathbf{V}_j$ and then define for $\mathbf{B}$ the estimator

$$\overline{\mathbf{B}}^*_M = (\mathbf{X}\mathbf{X}')^{-\mathbf{1}}\mathbf{X}\overline{\mathbf{V}}'_M, \tag{8}$$

which ends up being the same estimator defined in Subsection 2.1.

We may obtain additional information about $\boldsymbol{\Sigma}$ from $\mathbf{S}_{mean} = (\overline{\mathbf{V}}_M - \overline{\mathbf{B}}^{*'}_M\mathbf{X})'(\overline{\mathbf{V}}_M - \overline{\mathbf{B}}^{*'}_M\mathbf{X})$, which can be combined with the previous estimator $\mathbf{S}_\mathbf{v}$ to obtain

$$\mathbf{S}_{comb} = \frac{\mathbf{S}_\mathbf{v} + M \times \mathbf{S}_{mean}}{Mn - p}. \tag{9}$$

Analogous to what was done in Subsection 2.1, one can derive the following inferential results, for $p \geq m$, and $n > m + p$,

1. an unbiased estimator of $\mathbf{\Sigma}$ is $\mathbf{S}_{comb}$ (see Appendix B.2),
2. we prove in Corollary 2.5 (see below) that
$$T_{comb} = \frac{\left|(\overline{\mathbf{B}}^*_M - \mathbf{B})'(\mathbf{X}\mathbf{X}')(\overline{\mathbf{B}}^*_M - \mathbf{B})\right|}{\left|(n - \frac{p}{M})\mathbf{S}_{comb}\right|}, \tag{10}$$
is a pivotal quantity and that, for $\mathbf{W} \sim W_m(\mathbf{I}_m, n-p)$ and $F^*_{p,l} \sim \mathcal{F}_{p-l+1, Mn-p-l+1}$,
$$T_{comb}|\mathbf{W} \overset{st}{\sim} \left\{\prod_{l=1}^{m} \frac{p-l+1}{Mn-p-l+1} F^*_{p,l}\right\} \left|M(n-p)\mathbf{W}^{-1} + \mathbf{I}_m\right|,$$
3. if one wants to test the significance of a set of regression coefficients or more generally, a linear combination of $\mathbf{B}$, namely, $\mathbf{A}\mathbf{B} = \mathbf{C}$ where $\mathbf{A}$ is a $k \times p$ matrix with $rank(\mathbf{A}) = k \leq p$ and $k \geq m$, one may define
$$T_{comb,\mathbf{C}} = \frac{|(\mathbf{A}\overline{\mathbf{B}}^*_M - \mathbf{C})'(\mathbf{A}(\mathbf{X}\mathbf{X}')^{-\mathbf{1}}\mathbf{A}')^{-1}(\mathbf{A}\overline{\mathbf{B}}^*_M - \mathbf{C})|}{\left|(n - \frac{p}{M})\mathbf{S}_{comb}\right|}$$
and proceed by noting that, for $\mathbf{W} \sim W_m(\mathbf{I}_m, n-p)$ and $F^*_{k,l} \sim \mathcal{F}_{k-l+1, Mn-p-l+1}$,
$$T_{comb,\mathbf{C}}|\mathbf{W} \overset{st}{\sim} \left\{\prod_{l=1}^{m} \frac{k-l+1}{Mn-p-l+1} F^*_{k,l}\right\} \left|M(n-p)\mathbf{W}^{-1} + \mathbf{I}_m\right|; \tag{11}$$
then, we may build

   (i) *a test for the parameter matrix* $\mathbf{C}$*:* in order to test $H_0 : \mathbf{C} = \mathbf{C}_0$ versus $H_1 : \mathbf{C} \neq \mathbf{C}_0$, at a given γ level, we reject H_0 whenever the computed value of $T_{comb,\mathbf{C}_0}$ exceeds its $1-\gamma$ quantile; in particular a test for $\mathbf{B} = \mathbf{B_0}$ follows upon taking $\mathbf{A} = \mathbf{I_p}$,

   (ii) *a confidence set for* $\mathbf{C}$*:* a $(1-\gamma)$ level confidence set for $\mathbf{C}$ is given by
$$\Delta(\mathbf{C}) = \{\mathbf{C} : T_{comb,\mathbf{C}} \leq \omega_{M,k,m,p,n;\gamma}\}, \tag{12}$$
where $\omega_{M,k,m,p,n;\gamma}$ is the $1-\gamma$ quantile of $T_{comb,\mathbf{C}}$ (the value of $\omega_{M,k,m,p,n;\gamma}$ may be obtained by simulating the distribution of $T_{comb,\mathbf{C}}$, by first generating $\mathbf{W} \sim W_m(\mathbf{I}_m, n-p)$ and then generating $T_{comb,\mathbf{C}}|\mathbf{W}$ from (11)),

4. to infer about $\mathbf{ABD} = \boldsymbol{\Delta}$ where $\mathbf{A}$ is a $k \times p$ matrix and $\mathbf{D}$ is an $m \times r$ matrix with $r \leq k$, we start from its natural point estimator $\boldsymbol{\Delta}^*_M = \mathbf{A}\overline{\mathbf{B}}^*_M\mathbf{D}$ and propose to use pivotal quantity (see Corollary 2.6)

$$T_{comb,\boldsymbol{\Delta}} = \frac{|(\boldsymbol{\Delta}^*_M - \boldsymbol{\Delta})'\,(\mathbf{A}(\mathbf{XX}'\mathbf{A}')^{-1})\,(\boldsymbol{\Delta}^*_M - \boldsymbol{\Delta})|}{|(n-p)\mathbf{D}'\mathbf{S}_{comb}\mathbf{D}|} \tag{13}$$

whose distribution is given by

$$T_{comb,\boldsymbol{\Delta}}|\mathbf{W}^* \overset{st}{\sim} \left\{\prod_{l=1}^{r} \frac{k-l+1}{Mn-p-l+1} F^*_{k,l}\right\} \left|M(n-p)\mathbf{W}^{*-1} + \mathbf{I}_m\right|$$

where $F^*_{k,l} \sim \mathcal{F}_{k-l+1,Mn-p-l+1}$ and $\mathbf{W}^* \sim W_r(\mathbf{I}_r, n-p)$, all independently. Taking $r=1$ and $k=1$, and making $A : 1 \times p$ a matrix of zeros except for $\mathbf{A}_{1,g} = 1$, and $\mathbf{D} : m \times 1$ a matrix of zeros except for $\mathbf{D}_{h,1} = 1$, for $g = 1, \ldots, p$ and $h = 1, \ldots, m$ we may observe that the $(1-\alpha)$ confidence interval for $\boldsymbol{\Delta} = \mathbf{B}_{(g,h)}$ will be given by

$$\overline{\mathbf{B}}^*_{M(g,h)} \pm \sqrt{q^*_{comb,1-\alpha}(n-p)\mathbf{S}_{comb,(h,h)}(\mathbf{XX}')^{-1}_{g,g}}$$

(for details in the proof of this result see Section S3 of Part II of the supplementary material).

The above results are derived based on the observation that $\mathbf{S}_{mean}|\mathbf{S} \sim W_m(\frac{\mathbf{S}}{M}, n-p)$, and on the following Theorem and Corollaries, whose proofs are provided in Appendix A.

Theorem 2.4. *The joint pdf of* $(\overline{\mathbf{B}}^*_M, \mathbf{S}_{comb})$ *defined in (8) and (9) is proportional to*

$$\int \exp\left\{-\frac{1}{2}tr\left[(\boldsymbol{\Sigma} + \frac{1}{M}\mathbf{S})^{-1}(\overline{\mathbf{B}}^*_M - \mathbf{B})'(\mathbf{XX}')(\overline{\mathbf{B}}^*_M - \mathbf{B}) + (Mn-p)\mathbf{S}^{-1}\mathbf{S}_{comb} + (n-p)\boldsymbol{\Sigma}^{-1}\mathbf{S}\right]\right\}$$
$$\times |\mathbf{S}_{comb}|^{\frac{Mn-p-m-1}{2}} \times \frac{|\mathbf{S}|^{-\frac{Mn-p-n+2p+m+1}{2}}}{|\boldsymbol{\Sigma}|^{\frac{n}{2}}} \times |\boldsymbol{\Sigma}^{-1} + M\mathbf{S}^{-1}|^{-p/2}\, d\mathbf{S},$$

from which we can infer that, conditional on $\mathbf{S}$, $\overline{\mathbf{B}}^*_M$ *and* $\mathbf{S}_{comb}$ *are independent, with* $\overline{\mathbf{B}}^*_M|\mathbf{S} \sim N_{pm}(\mathbf{B}, (\boldsymbol{\Sigma} + \frac{1}{M}\mathbf{S}) \otimes (\mathbf{XX}')^{-1})$ *and* $(Mn-p)\mathbf{S}_{comb}|\mathbf{S} \sim W_m(\mathbf{S}, Mn-p)$.

Corollary 2.5. *The distribution of T_{comb} defined in (10) can be obtained from the decomposition*

$$T_{comb}|\mathbf{W} \overset{st}{\sim} \left\{ \prod_{l=1}^{m} \frac{p-l+1}{Mn-p-l+1} F^*_{p,l} \right\} \left| M(n-p)\mathbf{W}^{-1} + \mathbf{I}_m \right|$$

where $\mathbf{W} \sim W_m(\mathbf{I}_m, n-p)$ and $F^*_{p,l} \sim \mathcal{F}_{p-l+1, Mn-p-l+1}$, which implies that T_{comb} is a pivotal quantity, that is, its distribution does not depend on $\mathbf{\Sigma}$.

Corollary 2.6. *The distribution of $T_{comb,\mathbf{\Delta}}$ defined in (13) can be obtained from the decomposition*

$$T_{comb,\mathbf{\Delta}}|\mathbf{W} \overset{st}{\sim} \left\{ \prod_{l=1}^{r} \frac{k-l+1}{Mn-p-l+1} F^*_{k,l} \right\} \left| M(n-p)\mathbf{W}^{*-1} + \mathbf{I}_r \right|$$

where $\mathbf{W}^* \sim W_r(\mathbf{I}_r, n-p)$ and $F^*_{k,l} \sim \mathcal{F}_{k-l+1, Mn-p-l+1}$, implying that $T_{comb,\mathbf{\Delta}}$ is a pivotal quantity.

Remark 2.3. *It is the case that $Var(\overline{\mathbf{S}}^*_M) > Var(\mathbf{S}_{comb})$ for $M \geq 2$ (with equality for $M = 1$), and therefore the second new procedure is expected to outperform the first new procedure for $M \geq 2$. Anyway we still make both procedures available in the paper since the first procedure has an easier implementation, which the analyst may prefer to use given that for larger sample sizes there will be no big differences between the results from the two procedures, in terms of the radius, as it is shown in Section 3.*

The proof of this Remark may be seen in Appendix B.3.

2.3 Reiter's (2005) Methodology Under Multiple Imputation

Now we present an adaptation of Reiter (2005a) methodology for drawing inference on a vector valued parameter, based on multiply synthetic data, to draw inference on a matrix of regression coefficients. Although originally developed for synthetic data generated by

Posterior Predictive Sampling, Reiter and Kinney (2012) show that the methodology in Reiter (2005a) is also valid for synthetic data generated via Plug-in Sampling.

In order to be possible to use Reiter's (2005) methodology to estimate $\mathbf{B}$ from $\mathbf{V}_1, ..., \mathbf{V}_M$, the synthetic datasets defined at the beginning of Section 2, we consider $\mathbf{vec}(\mathbf{B}) = (\mathbf{B}_1' \ \mathbf{B}_2' \ ... \ \mathbf{B}_m')'$, a $pm \times 1$ parameter vector, where $\mathbf{B}_1', ..., \mathbf{B}_m'$ are the m columns of $\mathbf{B}$. Based on the original data, $\mathbf{vec}(\hat{\mathbf{B}})$ is an estimator of $\mathbf{vec}(\mathbf{B})$ and its covariance matrix estimator is $\mathbf{U} = \mathbf{S} \otimes (\mathbf{XX}')^{-1}$ a $pm \times pm$ matrix. Let $\mathbf{vec}(\mathbf{B}_j^*) = \mathbf{vec}((\mathbf{XX}')^{-\mathbf{1}}\mathbf{XV}_j')$ and $\mathbf{U}_j = \mathbf{S}_j^* \otimes (\mathbf{XX}')^{-1}$, where $\mathbf{S}_j^* = \frac{1}{n-p}(\mathbf{V}_j - \mathbf{B}_j^{*\prime}\mathbf{X})(\mathbf{V}_j - \mathbf{B}_j^{*\prime}\mathbf{X})'$, for $j = 1, \ldots, M$. Note that, conditionally on $\hat{\mathbf{B}}$ and $\mathbf{S}$, $\mathbf{vec}(\mathbf{B}_j^*)$ is an unbiased estimator of $\mathbf{vec}(\mathbf{B})$ and $\mathbf{U}_j$ is an unbiased estimator of its variance. Then the following estimators

$$\mathbf{vec}(\overline{\mathbf{B}}_M^*) = \frac{1}{M}\sum_{j=1}^{M}\mathbf{vec}(\mathbf{B}_j^*), \qquad \overline{\mathbf{U}}_M = \frac{1}{M}\sum_{j=1}^{M}\mathbf{U}_j,$$

$$\mathbf{b}_M = \frac{1}{M-1}\sum_{j=1}^{M}(\mathbf{vec}(\mathbf{B}_j^*) - \mathbf{vec}(\overline{\mathbf{B}}_M^*))(\mathbf{vec}(\mathbf{B}_j^*) - \mathbf{vec}(\overline{\mathbf{B}}_M^*))'$$

should be Reiter's estimators to be used to draw inference about $\mathbf{B}$, where $\mathbf{vec}(\overline{\mathbf{B}}_M^*)$ is an estimator for $\mathbf{vec}(\mathbf{B})$, its variance being estimated by $\frac{1}{M}\mathbf{b}_M + \overline{\mathbf{U}}_M$. Let us consider the statistic

$$T_{R,M} = \frac{(\mathbf{vec}(\overline{\mathbf{B}}_M^*) - \mathbf{vec}(\mathbf{B}))'(\overline{\mathbf{U}}_M)^{-1}(\mathbf{vec}(\overline{\mathbf{B}}_M^*) - \mathbf{vec}(\mathbf{B}))}{pm(1+r)}$$

where $r = \frac{tr(\mathbf{b_M}\overline{\mathbf{U}}_{\mathbf{M}}^{-\mathbf{1}})}{Mpm}$. The distribution of $T_{R,M}$ is approximated by an $\mathcal{F}_{pm,w(r)}$ distribution where $w(r) = 4 + [pm(M-1) - 4]\left[1 + 1/r - 2(rpm)^{-1}(M-1)^{-1}\right]^2$ (Reiter, 2005a).

3 Simulation Studies

In this section we present results of some simulations. The objectives of these simulations are (i) to show that the inferential methods used in Section 2 perform as we predicted for our proposed methodology for singly and multiply imputed synthetic data generated via

Plug-in Sampling, and (*ii*) to compare the accuracy of our proposed methodology with the accuracy of Reiter (2005a) methodology for multiply imputed partially synthetic data. All simulations were carried out using the software Mathematica®. To conduct the simulation, we take the population distribution as a multivariate normal distribution with expected value given by the right hand side of (1), with matrix of regression coefficients $\mathbf{B}$, and covariance matrix $\mathbf{\Sigma}$, for $m = 2$ and $p = 3$, given by

$$\mathbf{B} = \begin{pmatrix} 1 & 2 \\ 3 & 2 \\ 1 & 1 \end{pmatrix} \quad \text{and} \quad \mathbf{\Sigma} = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}.$$

The values $x_{1i}, x_{2i}, x_{3i}, i = 1, ..., n$, of the explanatory variables are generated as iid $N(0, 1)$ and held fixed for the entire simulation.

Based on a Monte Carlo simulation with 10^5 iterations, we compute estimates of the coverage probability (percentage of observed values of the statistics smaller than the respective theoretical cut-off points) of the following confidence regions (where in all cases, the level of the confidence region is set to 0.95):

1. for the two new procedures in Subsections 2.1 and 2.2, based on single and multiple synthetic data, the confidence sets for $\mathbf{B}$ and for $\mathbf{AB} = \mathbf{C}$, given by (6) and (12), are computed, with $\mathbf{A} = (\ \mathbf{0}_{2\times 1}|\ \mathbf{I}_2)$, using the methodology described in the two subsections referred above; for $M = 1, 2, 5, 10, 20$ synthetic datasets, the estimated average coverage probabilities of the confidence sets are shown in Table 1 under the columns $\mathbf{B}(1)$ and $\mathbf{AB}(1)$ for the new procedure in Subsect. 2.1, and under the columns $\mathbf{B}(2)$ and $\mathbf{AB}(2)$ for the new procedure in Subsect. 2.2; for $M = 1$ only one column is needed since the two procedures coincide, and Reiter's adapted procedure is not available for single imputed data;
2. the confidence set for $\mathbf{B}$ is obtained using the adapted methodology of Reiter (2005a) in Subsect. 2.3, for $M(> 1)$ synthetic datasets, and then for each of the cases $M =$

$2, 5, 10, 20$, the estimated coverage probabilities of the confidence sets are shown in Table 1 under the column **vec**(**B**).

Table 1: Estimated average coverage probabilities for **B** and **AB**

(a) Average coverage for **B**

	$M = 1$	$M = 2$			$M = 5$			$M = 10$			$M = 20$		
n		**vec**(**B**)	**B**(1)	**B**(2)	**vec**(**B**)	**B**(1)	**B**(2)	**vec**(**B**)	**B**(1)	**B**(2)	**vec**(**B**)	**B**(1)	**B**(2)
10	0.951	0.830	0.950	0.950	0.754	0.949	0.947	0.748	0.950	0.949	0.753	0.948	0.947
20	0.953	0.919	0.955	0.953	0.874	0.947	0.948	0.867	0.948	0.949	0.870	0.950	0.950
50	0.953	0.955	0.950	0.951	0.924	0.949	0.948	0.921	0.949	0.948	0.924	0.949	0.948
100	0.946	0.957	0.946	0.947	0.934	0.946	0.946	0.931	0.948	0.948	0.935	0.948	0.948
200	0.949	0.964	0.953	0.952	0.943	0.950	0.951	0.943	0.949	0.949	0.944	0.949	0.950

(b) Average coverage for **AB**

	$M = 1$	$M = 2$			$M = 5$			$M = 10$			$M = 20$		
n		**vec**	**AB**	**AB**	**vec**	**AB**	**AB**	**vec**	**AB**	**AB**	**vec**	**AB**	**AB**
		(**AB**)	(1)	(2)	(**AB**)	(1)	(2)	(**AB**)	(1)	(2)	(**AB**)	(1)	(2)
10	0.950	0.968	0.949	0.950	0.791	0.947	0.946	0.799	0.948	0.946	0.791	0.944	0.947
20	0.952	0.994	0.950	0.951	0.888	0.948	0.947	0.891	0.949	0.949	0.888	0.950	0.950
50	0.954	0.999	0.953	0.954	0.931	0.950	0.949	0.927	0.948	0.946	0.928	0.949	0.949
100	0.946	1.000	0.946	0.948	0.940	0.948	0.948	0.937	0.949	0.950	0.939	0.948	0.948
200	0.951	1.000	0.953	0.952	0.946	0.949	0.950	0.948	0.951	0.949	0.948	0.950	0.949

The results reported in Table 1 for sample sizes n=10, 20, 50, 100, 200, show that, based on singly imputed and multiply imputed synthetic data, the 0.95 confidence sets for **B** and **AB** have an estimated coverage probability approximately equal to 0.95, confirming that the confidence sets perform as predicted. Using the adapted methodology of Reiter (2005a) for multiply imputed partially synthetic data the estimated coverage probabilities fall short of the stipulated level of 0.95 for very small sample sizes, as expected, since this procedure is asymptotic in nature, but quickly attain the desired level even for moderate sample sizes for the cases where $M \geq 5$.

In Part I of the supplementary material we address cases where the response variables have non-normal distributions, namely when they have a multivariate t-distribution, a multivariate skew-normal distribution, a binomial distribution, a Poisson distribution and a distribution with a spike at zero. As it may be seen from Tables S1–S8 in the supplementary material, our procedures show, in general, for all these distributions values of estimated average coverage probabilities for $\mathbf{B}$ and $\mathbf{AB}$ (with $\mathbf{A} = (\ \mathbf{0}_{2\times 1}|\ \mathbf{I}_2)$), very close to the nominal value of 0.95, even for small sample sizes. We may note that in all cases the adaptation of Reiter (2005a) procedure gives somewhat similar results, at least for the larger sample sizes. Only for the case of distributions of response variables with a spike at zero, when testing for the matrix $\mathbf{B}$, the adapted Reiter procedure seems to present even lower average coverage probabilities than our procedures.

We may note that in the single imputation case ($M = 1$), the estimated average coverage probabilities have values that are slightly closer to the nominal value of 0.95, mainly when considering discrete distributions for the response variables. This may lead to the conclusion that the use/release of singly imputed datasets may be more adequate in these cases.

In order to measure the *radius* (distance between the center and the edge) of the confidence sets, we propose, for a level $0 < \gamma < 1$,

$$\Upsilon_M = \delta_{M,k,m,p,n;\gamma} \times |(n-p)\tilde{\mathbf{S}}_M|,$$

where $\delta_{M,k,m,p,n;\gamma}$ is the cut-off point and where we take $M = 1, 2, 5, 10, 20$, with $\tilde{\mathbf{S}}_M = \overline{\mathbf{S}}^*_M$ for the first new procedure, and $\tilde{\mathbf{S}}_M = (n - \frac{p}{M})/(n-p)\mathbf{S}_{comb}$ for the second new procedure, and making $k = p$.

In order to compare with the original data, we take $M = 0$, with $\tilde{\mathbf{S}}_0 = \mathbf{S}$, and the cut-off points $\delta_{0,k,m,p,n;\gamma}$ are obtained as the γ quantiles of the statistics

$$T_O = \frac{|(\hat{\mathbf{B}} - \mathbf{B})'(\mathbf{XX}')(\hat{\mathbf{B}} - \mathbf{B})|}{|(n-p)\mathbf{S}|} \overset{st}{\sim} \prod_{l=1}^{m} \frac{p-l+1}{n-p-l+1} F_{p,l}, \tag{14}$$

$$T_{O,\mathbf{C}} = \frac{|(\mathbf{A}\hat{\mathbf{B}} - \mathbf{C})'(\mathbf{A}(\mathbf{XX}')^{-1}\mathbf{A}')^{-1}(\mathbf{A}\hat{\mathbf{B}} - \mathbf{C})|}{|(n-p)\mathbf{S}|} \overset{st}{\sim} \prod_{l=1}^{m} \frac{k-l+1}{n-p-l+1} F_{k,l}, \tag{15}$$

where $F_{p,l} \sim \mathcal{F}_{p-l+1,n-p-l+1}$ and $F_{k,l} \sim \mathcal{F}_{k-l+1,n-p-l+1}$.

The expected value of this measure will be

$$E(\Upsilon_M) = \delta_{M,k,m,p,n;\gamma} \times \frac{(n-p)!}{(n-p-m)!} \times K_{M,p,n,m}|\mathbf{\Sigma}|$$

where $K_{0,p,n,m} = 1$ for the original data and, for $M \geq 1$,

$$K_{M,p,n,m} = \frac{1}{M^m(n-p)^m} \frac{(Mn-Mp)!}{(Mn-Mp-m)!}$$

for the procedure in Subsection 2.1, and

$$K_{M,p,n,m} = \frac{1}{M^m(n-p)^m} \frac{(Mn-p)!}{(Mn-p-m)!}$$

for the procedure in Subsection 2.2, where when $M = 1$ it will refer to the case of single imputed synthetic data. For more details about these expected values see Appendix B.4.

For $M = 0, 1, 2, 5, 10, 20$ and sample sizes $n = 10, 20, 50, 100, 200$, we present the average (avg) of simulated values of Υ_M, for 10^5 simulations, and its expected value (exp) $E(\Upsilon_M)$ for $\mathbf{B}$ in Table 2 and for $\mathbf{C} = \mathbf{AB}$ in Table 3.

Observing Tables 2 and 3, we conclude that as the number M of released synthetic datasets increases, Υ_M decreases and eventually coincides with the value of Υ_0, the value for the original data, indeed as expected, since as M increases, the amount of information about the original data released increases, leading us closer to the inference drawn from the original data. We also observe that the values of Υ_M, for $M > 1$, for both procedures become identical for larger sample sizes.

We may see that the Plug-in Sampling method offers *radius* that are much smaller than those obtained with the PPS method. For $M = 1$, that is, in the case of single imputation, the PPS method leads to *radius* which are approximately two and half times the *radius* obtained under Plug-in Sampling, what may be seen as an important advantage of the Plug-in Sampling method (Moura, 2016, Sec. 4.2.2).

Table 2: Average values of Υ_M and the values of $E(\Upsilon_M)$ for the confidence set for $\mathbf{B}$.

n	Original ($M = 0$)	$M = 1$		$M = 2$				$M = 5$				$M = 10$				$M = 20$			
		avg	exp	1st Approach		2nd Approach		1st Approach		2nd Approach		1st Approach		2nd Approach		1st Approach		2nd Approach	
				avg	exp	avg	exp	avg	exp	avg	exp	avg	exp	avg	exp	avg	exp	avg	exp
10	37.55	214.29	219.96	90.62	92.84	85.74	87.56	54.83	55.28	51.61	51.94	44.57	44.78	43.86	43.95	41.04	41.08	40.24	40.27
20	22.76	102.91	103.29	53.50	53.79	52.20	52.47	32.58	32.69	32.50	32.57	27.06	27.13	27.56	27.61	25.25	25.31	25.03	25.08
50	18.73	78.10	77.43	42.58	42.34	43.13	42.86	27.34	27.17	27.11	26.94	22.98	22.82	22.66	22.50	20.90	20.77	20.66	20.53
100	17.74	73.44	73.34	40.11	40.09	40.27	40.24	25.28	25.27	25.54	25.53	21.51	21.51	21.61	21.61	19.50	19.50	19.29	19.28
200	17.41	69.81	69.72	39.07	39.01	38.62	38.56	25.09	25.07	25.08	25.06	20.78	20.77	20.79	20.78	18.94	18.94	19.15	19.14

4 An Application Using the Current Population Survey Data

In this section we provide an application based on some real data and compare the inference based on the original data with the inference based on the synthetic data, according to the procedures developed in Section 2 and also the method of Reiter (2005a). The data are public use data from the 2000 U.S. Current Population Survey (CPS) March supplement, available online from https://www.census.gov/programs-surveys/cps.html. We will only focus on the household records. The full data has seventeen variables measured on 51,016 heads of households and it includes the variables age, race, sex and marital status as key identifiers and a mix of other categorical and numerical variables. For the vector $\mathbf{y}$ of response sensitive variables, we have selected two numerical variables, namely, *total household income* (I) and *household property tax* (PT). After deleting all entries where at least one of these variables are reported as 0, we were left with a sample size of 32,923.

Table 3: Average values of Υ_M and the values of $E(\Upsilon_M)$ for the confidence set for $\mathbf{AB}$.

n	Original ($M = 0$)	$M = 1$		$M = 2$				$M = 5$			
		avg	exp	1st Approach		2nd Approach		1st Approach		2nd Approach	
				avg	exp	avg	exp	avg	exp	avg	exp
10	13.36	72.57	74.49	32.62	33.42	30.61	31.26	19.56	19.72	19.00	19.12
20	8.66	39.00	39.15	20.12	20.23	20.17	20.27	12.71	12.75	12.45	12.48
50	7.44	31.23	30.97	16.67	16.58	16.79	16.69	10.70	10.63	10.55	10.48
100	7.14	28.30	28.27	15.89	15.88	16.12	16.11	10.07	10.07	10.10	10.10
200	6.92	27.76	27.72	15.80	15.78	15.56	15.54	9.72	9.72	9.85	9.84

n	$M = 10$				$M = 20$			
	1st Approach		2nd Approach		1st Approach		2nd Approach	
	avg	exp	avg	exp	avg	exp	avg	exp
10	16.07	16.15	15.61	15.64	14.44	14.45	14.49	14.50
20	10.38	10.41	10.41	10.43	9.55	9.58	9.54	9.56
50	8.94	8.88	8.76	8.70	8.25	8.20	8.10	8.04
100	8.42	8.43	8.63	8.63	7.68	7.68	7.77	7.77
200	8.40	8.40	8.16	8.16	7.64	7.63	7.56	7.56

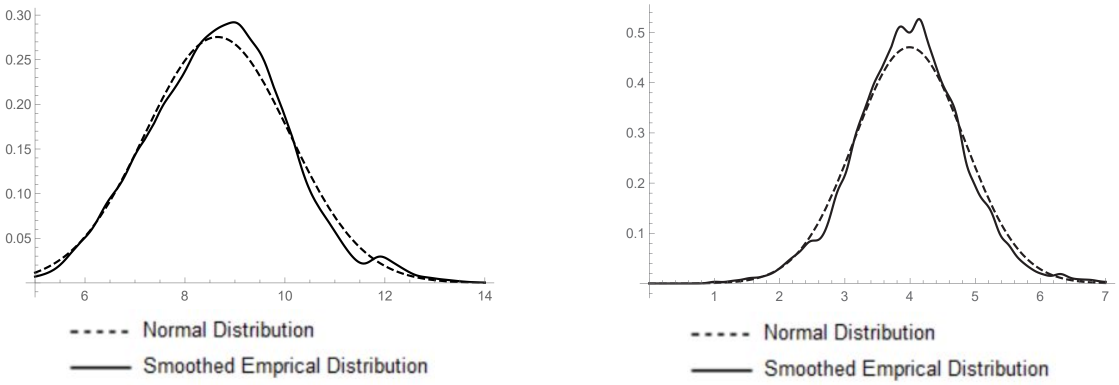
The example addressed below, using the proposed exact methods developed in Subsections 2.1 and 2.2, illustrates the capabilities of these methods. We will assume the normality of the fifth root of the two response variables I and PT. As we may observe in Figure 1, the marginal distribution of the transformed variables is approximately normal. Anyway, as it is shown by the results in the supplementary material, even if these variables would not be normally distributed, the procedures in Section 2 will still perform adequately.

We take the $n = 32,923$ households as a random sample, and I and PT as confidential variables. We will use the following set of covariates:

N: number of people in household;

L: number of people in the household who are less than 18 years old;

A: age for the head of household;



(a) household property tax

(b) total household income

Figure 1: Smothed Empirical distributions of response variables PT and I

E: education level for the head of the household(coded to take values 31-46);

M: marital status for the head of the household (coded to take values 1-7);

R: race of the head of the household (coded to take values 1-4);

S: sex of the head of the household (coded to take values 1,2).

For further details, namely on the coding of the variables, we refer to the Current Population Survey March 2000 technical documentation (available at http://www.census.gov/prod/techdoc/cps/cpsmar00.pdf) and to Klein and Sinha (2015a).

As such, in this application, $\mathbf{x}$, the vector of explanatory variables, is defined as

$$\begin{aligned}\mathbf{x} =& \Big(1, \mathrm{N}, \mathrm{L}, \mathrm{A}, \mathrm{I}(\mathrm{E}=32), ..., \mathrm{I}(\mathrm{E}=46), \mathrm{I}(\mathrm{M}=2), ..., \mathrm{I}(\mathrm{M}=7),\\ & \mathrm{I}(\mathrm{R}=2), ..., \mathrm{I}(\mathrm{R}=4), \mathrm{I}(\mathrm{S}=2)\Big)', \end{aligned} \tag{16}$$

where the indicator variables for the first code present in the sample for each variable is taken out in order to make the model matrix full rank, and where I(E = 32) is the indicator variable for E = 32, i.e. for individuals that have completed 1st, 2nd, 3rd or 4th grade, and so on. The model matrix $\mathbf{X} = (\mathbf{x}_1, \cdots, \mathbf{x}_n)$ has $p = 29$ rows and $n = 32,923$ columns, with rank equal to 29. Using the plug-in sampling method, we generate a single synthetic dataset. The realizations of the unbiased estimators $\mathbf{B}^*$ and $\mathbf{S}^*$ of $\mathbf{B}$ and $\boldsymbol{\Sigma}$, are respectively shown in Table 4 and in expression (17), along with the realizations of the original data estimators $\hat{\mathbf{B}}$ and $\mathbf{S}$. These estimates are respectively denoted by $\widetilde{\mathbf{B}}^*$, $\widetilde{\mathbf{S}}^*$, $\widetilde{\hat{\mathbf{B}}}$ and $\widetilde{\mathbf{S}}$, with

$$\widetilde{\mathbf{S}}^* = \begin{pmatrix} 0.6576 & 0.2090 \\ 0.2090 & 1.2905 \end{pmatrix}, \quad \widetilde{\mathbf{S}} = \begin{pmatrix} 0.6626 & 0.2130 \\ 0.2130 & 1.2898 \end{pmatrix}. \tag{17}$$

We see that the point estimates of $\hat{\mathbf{B}}$ based on the synthetic data and the original data tend to be in agreement. We also find that the two estimates of $\mathbf{\Sigma}$, $\widetilde{\mathbf{S}}$ and $\widetilde{\mathbf{S}}^*$, tend to have a general agreement.

As remarked in Moura (2016, Sec. 5.1) estimates obtained from Plug-in synthetic generated data seem to be more in agreement with the ones obtained from the original data than the ones obtained from synthetic data obtained from PPS.

We now present inferences on regression coefficients obtained by applying the methodology from Section 2 to analyze the singly imputed synthetic data and multiply synthetic data, considering $M = 2$ and $M = 5$. For this purpose, we use the statistics T, T_M, T_{comb} and $T_{R,M}$ defined in Section 2 and their empirical distributions (10^5 simulation size) to test the significance of the model, for $\gamma = 0.05$. For $M = 1$ the computed value of the statistic T was 4.96468, which is larger than the determined cut-off point for this case, $\delta_{2,32923,29;0.05} = 5.14914 \times 10^{-6}$, with a corresponding p-value approximately equal to 0, therefore, rejecting the non-significance of the model, that is, assuming that the explanatory variables in $\mathbf{x}$ have a significant role in determining the values of the response variables in $\mathbf{y}$. For $M = 2$ and $M = 5$, one finds similar p-values, with the computed values of T_M, for the first procedure, equal to 4.94839 and 5.06947, and the computed values of T_{comb}, for the second procedure, equal to 4.94420 and 5.06190. If we perform the same test on the original data, we obtain for T_O in (14) the computed value of 4.93432, that is also larger than the determined cut-off point 1.27984×10^{-6}, with a p-value approximately equal to 0, also rejecting the non-significance of the model. For Reiter's adapted procedure the p-values obtained were also approximately equal to zero both for $M = 2$ and $M = 5$.

Table 4: Estimates of the regression coefficients from the synthetic data and from the original data.

explanatory variable	*SyntheticData* ($\widetilde{\mathbf{B}}^*$)		*OriginalData* ($\widetilde{\widetilde{\mathbf{B}}}$)	
	I	PT	I	PT
Intercept	3.33024	7.56007	3.24037	7.52088
N	0.11642	0.51872	0.11527	0.52155
L	−0.08285	−0.53486	−0.08429	−0.53651
A	0.00087	−0.01440	0.00132	−0.01505
I(E=32)	−0.06863	−0.11734	0.08377	−0.07970
I(E=33)	−0.03361	−0.15427	0.03402	−0.03844
I(E=34)	0.05408	0.05322	0.13688	0.14479
I(E=35)	0.05304	0.11157	0.08928	0.23425
I(E=36)	0.11955	0.24970	0.14576	0.33577
I(E=37)	0.07023	0.23027	0.16636	0.29644
I(E=38)	0.20794	0.38589	0.21098	0.38226
I(E=39)	0.28300	0.72827	0.35955	0.79781
I(E=40)	0.36835	1.03459	0.44939	1.11411
I(E=41)	0.33921	1.06392	0.44562	1.10290
I(E=42)	0.49522	1.33937	0.57402	1.33862
I(E=43)	0.52201	1.59578	0.60579	1.67726
I(E=44)	0.76442	1.87793	0.88662	1.99260
I(E=45)	0.79513	2.36940	0.89977	2.50898
I(E=46)	0.81286	2.42916	0.91233	2.46191
I(M=2)	−0.29167	−0.18976	−0.20503	−0.10286
I(M=3)	−0.07052	−0.41459	−0.06588	−0.44057
I(M=4)	−0.03956	−0.47224	−0.05187	−0.46352
I(M=5)	−0.07136	−0.32840	−0.08825	−0.35516
I(M=6)	−0.03477	−0.63850	−0.06795	−0.66138
I(M=7)	−0.00992	−0.58561	−0.03941	−0.57123
I(R=2)	−0.09089	−0.14000	−0.07882	−0.12586
I(R=3)	−0.29051	−0.39652	−0.25237	−0.38992
I(R=4)	−0.07131	0.05753	−0.02879	0.05517
I(S=2)	0.02176	−0.10572	0.01507	−0.10844

In figure 2, one can see a histogram associated with the empirical distribution of T_M for $M = 1$ (10^5 simulation size).

We further considered the test of the null hypothesis $H_0 : \mathbf{AB} = \mathbf{0}$, using $\mathbf{A} = \left(\mathbf{0}_{2\times 3} \middle| \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} \middle| \mathbf{0}_{2\times 23} \right)$ and the statistics $T_{M,\mathbf{C}}$ and $T_{comb,\mathbf{C}}$ in (5) and (11), and also

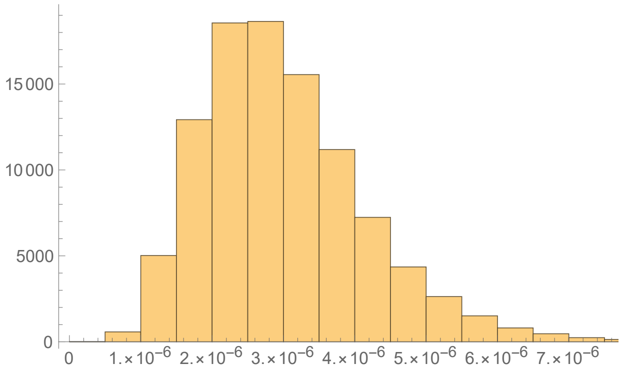


Figure 2: Histogram of the Empirical values of T_M for $M = 1$

the adapted procedure of Reiter (2005a). In the latter we replaced $\mathbf{vec}(\mathbf{B}_j^*)$ by $\mathbf{vec}(\mathbf{AB}_j^*)$, $\mathbf{vec}(\mathbf{B})$ by $\mathbf{vec}(\mathbf{AB})$ and took $\mathbf{U}_j = \mathbf{S}_j^* \otimes (\mathbf{A}(\mathbf{XX}')^{-1}\mathbf{A}')$ $(j = 1, \ldots, M)$. For $M = 1$ the computed value of $T_{M,\mathbf{C}}$ is $1.02081{\times}10^{-7}$ with a p-value of 0.00408, while for $M = 2$ we have a computed value of $T_{M,\mathbf{C}}$ equal to $2.07415{\times}10^{-7}$ and a computed value of $T_{comb,\mathbf{C}}$ of $2.07240{\times}10^{-7}$, to both of which corresponds a p-value of 0.00001. For $M = 5$ were obtained computed values of $3.47938{\times}10^{-7}$ and $3.47419{\times}10^{-7}$ respectively for $T_{M,\mathbf{C}}$ and $T_{comb,\mathbf{C}}$, with corresponding p-values of approximately 0. Reiter's adapted procedure gave p-values of 0.00016 for $M = 2$ and approximately zero for $M = 5$.

Also for this case, we may note that for all procedures the p-values are very close to zero as also is the p-value obtained from the original data, when using (15). As a consequence, it is interesting to observe that all p-values lead to the same conclusion, the rejection of the non-significance of the set of regression coefficients, and that the p-values obtained for $M\!=\!1$ are not very far from the ones obtained for $M\!=\!2$. Comparing the two multiple imputation procedures developed we observe that they present very similar p-values. Also, with the increase of the value of M the p-values get smaller, that is, closer to the p-values obtained with the original data, which although it may be seen as an advantage, it comes at the expense of a decrease in confidentiality.

Alternatively, it is possible to construct individual confidence intervals for all regression coefficients by using Result 5 in Section 2.1 and Result 4 in Section 2.2, based on Corollaries 2.3 and 2.6, and whose detailed proofs may be found in Section S3 of Part II of the supplementary material. In Subsections S3.1–S3.6 are shown the confidence intervals for all regression coefficients derived from the original data and the synthetic datasets for M=1, 2, 5. From these confidence intervals one may observe that for increasing values of M, the confidence intervals become smaller and smaller becoming closer and closer to the size of the one derived from the original data. This fact concurs with the study of the radius done in Section 3.

5 Privacy Protection of Singly Versus Multiply Imputed Synthetic Data

It is anticipated that singly imputed synthetic data will offer bigger protection than multiply imputed synthetic data. Nevertheless, one needs to evaluate this level of protection. In this section, we perform this evaluation using the CPS data referred to in the previous section. Let us consider $\mathbf{V}_j = (\mathbf{v}_1^{(j)}, ..., \mathbf{v}_n^{(j)})$, $j = 1, ..., M$, M synthetic datasets generated by the Plug-in Sampling method, where $\mathbf{v}_i^{(j)} = (v_{1i}^{(j)}, ..., v_{mi}^{(j)})', i = 1, ..., n$. Assume that after having access to the released synthetic data an "intruder" estimates the original values $\mathbf{y}_i = (y_{1i}, ..., y_{mi})'$ by $\hat{\mathbf{y}}_i = \frac{1}{M}\sum_{j=1}^{M} \mathbf{v}_i^{(j)}$. Then we propose the following three criteria as measures of the level of privacy protection

$$\Gamma_{1,\epsilon} = \frac{1}{mn}\sum_{l=1}^{m}\sum_{i=1}^{n} Pr\left[\left|\frac{\hat{y_{li}} - y_{li}}{y_{li}}\right| < \epsilon \,\middle|\, \mathbf{Y}\right];$$

$$\Gamma_{2,\epsilon} = \frac{1}{n}\sum_{i=1}^{n} Pr\left[\sqrt{\frac{1}{m}\sum_{l=1}^{m}\frac{(\hat{y_{li}} - y_{li})^2}{y_{li}^2}} < \epsilon \,\middle|\, \mathbf{Y}\right];$$

$$\Gamma_{3,\epsilon} = Pr\left[\frac{1}{mn}\sum_{l=1}^{m}\sum_{i=1}^{n}\left|\frac{\hat{y_{li}} - y_{li}}{y_{li}}\right| < \epsilon \,\middle|\, \mathbf{Y}\right].$$

Let us also consider from $\Gamma_{1,\epsilon}$ the following quantity, for $i = 1, ...n$ and $l = 1, .., m$,

$$D_{1,\epsilon} = Pr\left[\left|\frac{\hat{y_{li}} - y_{li}}{y_{li}}\right| < \epsilon \,\middle|\, \mathbf{Y}\right]$$

and from $\Gamma_{3,\epsilon}$ the

$$D_3 = \frac{1}{mn}\sum_{l=1}^{m}\sum_{i=1}^{n}\left|\frac{\hat{y_{li}} - y_{li}}{y_{li}}\right|.$$

We use Monte Carlo simultation with 10^4 iterations to estimate the above measures for each of the $n = 32,923$ households in the CPS data set.

In Table 5, we show the values of $\Gamma_{1,0.01}$ and $\Gamma_{2,0.01}$ and for $D_{1,0.01}$ its minimum, 1st quartile (Q_1), median, 3rd quartile (Q_3) and maximum. In Table 6, we show the values of $\Gamma_{3,0.01}$, $\Gamma_{3,0.1}$ and the minimum, Q_1, median, Q_3 and maximum of D_3. Looking at Table 5, we observe that the values of the measures increase as M increases, showing that the disclosure risk increases with the increase in the number of released synthetic data sets. We also observe that even for $M = 5$, the maximum value of $D_{1,0.01}$ is 0.3279, thus already indicating a substantial disclosure risk compared to 0.1491 from the singly imputed case. Likewise, we observe from Table 6 that if we set $\epsilon = 0.09$, we have $\Gamma_{3,\epsilon} = 0$ for $M = 1$ but $\Gamma_{3,\epsilon} = 0.1886$ for $M = 5$.

Table 5: Values of $\Gamma_{1,0.01}$, $\Gamma_{2,0.01}$ and a summary of the distribution of $D_{1,0.01}$.

M	$\Gamma_{1,0.01}$	$\Gamma_{2,0.01}$	$D_{1,0.01}$				
			Min	Q_1	Median	Q_3	Max
$M = 1$	0.0631	0.0006	0	0.0398	0.0552	0.0854	0.1491
$M = 2$	0.0754	0.0010	0	0.0331	0.0697	0.0954	0.2134
$M = 5$	0.0879	0.0018	0	0.0110	0.0792	0.1284	0.3279

Table 6: Values of $\Gamma_{3,0.01}$, $\Gamma_{3,0.09}$ and a summary of the distribution of D_3.

M	$\Gamma_{3,0.01}$	$\Gamma_{3,0.09}$	D_3				
			Min	Q_1	Median	Q_3	Max
$M = 1$	0	0	0.1050	0.1202	0.1233	0.1264	0.1379
$M = 2$	0	0	0.0948	0.1026	0.1051	0.1072	0.1159
$M = 5$	0	0.1886	0.0836	0.0905	0.0921	0.0937	0.1013

6 Concluding Remarks

The data analysis methodology of Reiter (2003), Reiter (2005a) and Raghunathan et al. (2003) are asymptotic in nature and can only be used when multiply imputed synthetic datasets are released. Moreover, their procedures were developed to draw inference only on scalar and vector parameters. In this paper, two exact likelihood-based solutions are offered for the case when multiple or single synthetic datasets are released and inference procedures are obtained for a matrix of regression coefficients under a Multivariate Linear Regression Model when synthetic data are generated via Plug-in Sampling. Furthermore, the authors provide an adaptation of Reiter (2005a) vector methodology to a matrix of parameters.

The second procedure proposed for multiple synthetic data presents slightly better performances than the first one for small sample sizes, and their performances are nearly the same for larger sample sizes. Nevertheless, the first procedure presents a simpler way of analyzing the synthetic datasets, thus being important to have access to both these procedures.

Although the singly imputed synthetic data will offer bigger protection than multiply imputed synthetic data, when releasing increasing numbers of multiple datasets, the confidence regions become smaller and smaller becoming closer and closer to the size of the ones derived from the original data.

Simulation studies show that the two new exact methodologies developed lead to confidence sets with the expected level of confidence, even for small sample sizes, both for single and multiple imputation.

Our simulations also reveal that as the number of synthetic datasets released increases, the inference derived from synthetic datasets comes closer to the one based on the original data, but of course at the expense of compromising privacy, namely by increasing the disclosure risk. Due to the increasing need to protect privacy, some entities already have decided to not release multiple imputation synthetic datasets, releasing only a single imputation dataset. The procedures developed in this work now allow the analysis of the data in the single imputation case, encouraging imputers to consider this scenario without having the concern about the feasibility of its analysis. We may note that one other advantage of the single imputation is that the estimated average coverage probabilities have values that are slightly closer to the nominal value of 0.95 than the ones obtained from multiple imputation, mainly when considering discrete distributions for the response variables, not forgetting that an analyst may find less confusing receiving a single dataset.

One of the advantages of using the Plug-in Sampling method is that it offers smaller *radius* (distance between the center and the edge) of the confidence sets than the Posterior Predictive Sampling (PPS) method, while also giving estimates of the parameters that are closer to the ones obtained from the original data, although at the expense of slightly higher levels of disclosure risk (Moura, 2016). Furthermore, the procedures developed, although based on model (1), which may seem to be a quite restrictive framework, allowed to develop inferential procedures with very good performances when data is generated by adaptations of the Plug-in method that generate non-normal or discrete response variables.

In the future it would be interesting to research how the procedures developed would perform on partial synthetic datasets generated by CART (Classification And Regression

Tree) methodology (Reiter, 2005b) and how similar techniques and procedures to the ones developed might be applied using LASSO (Least Absolute Shrinkage and Selection Operator) and other shrinkage and penalized regression methods.

Acknowledgments

This work is funded by national funds through FCT - Fundação para a Ciência e a Tecnologia, I.P., under the scope of projects UID/MAT/00297/2013 and UIDB/00297/2020 (Center for Mathematics and Applications).

Ricardo Moura's research was supported by a Fulbright Research Grant, and he sincerely thanks the faculty of the Department of Mathematics and Statistics at UMBC (University of Maryland, Baltimore County) for their support and encouragement. Part of Ricardo Moura's work was also completed during a 'Summer at Census' visit to the US Census Bureau for which he is thankful to Dr. Tommy Wright.

The authors also want to earnestly thank the constructive contributions of the Editors-in-Chief, Associate Editors and reviewers who dealt with the manuscript since its first submission.

References

Anderson, T. W. (2003). *An Introduction to Multivariate Statistical Analysis* (3rd ed.). Wiley.

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

Kinney, S. K., J. P. Reiter, and J. Miranda (2014). Synlbd 2.0: improving the synthetic longitudinal business database. *Statistical Journal of the IAOS 30*(2), 129–135.

Kinney, S. K., J. P. Reiter, A. P. Reznek, J. Miranda, R. S. Jarmin, and J. M. Abowd (2011). Towards unrestricted public use business microdata: The synthetic longitudinal business database. *International Statistical Review 79*(3), 362–384.

Klein, M., T. Mathew, and B. Sinha (2013). A comparison of statistical disclosure control methods: Multiple imputation versus noise multiplication. *U.S. Census Bureau Research Report Series*.

Klein, M. and B. Sinha (2013). Statistical analysis of noise multiplied data using multiple imputation. *Journal of Official Statistics 29*, 425–465.

Klein, M. and B. Sinha (2015a). Inference for singly imputed synthetic data based on posterior predictive sampling under multivariate normal and multiple linear regression models. *Sankhya B 77-B*, 293–311.

Klein, M. and B. Sinha (2015b). Likelihood-based finite sample inference for synthetic data based on exponential model. *Thailand Statistician 13*(1), 33–47.

Klein, M. and B. Sinha (2015c). Likelihood-based inference for singly and multiply imputed synthetic data under a normal model. *Statistics and Probability Letters 105*, 168–175.

Klein, M. and B. Sinha (2016). Likelihood based finite sample inference for singly imputed synthetic data under the multivariate normal and multiple linear regression models. *Journal of Privacy and Confidentiality 7*(1), 43–98.

Kollo, T. and D. Rosen (2005). *Advanced Multivariate Statistics with Matrices.* Springer.

Little, R. J. A. (1993). Statistical analysis of masked data. *Journal of Official Statistics 9*(2), 407.

Moura, R. (2016). *Likelihood-based Inference for Multivariate Regression Models using Synthetic Data*. Ph. D. thesis, NOVA University of Lisbon.

Muirhead, R. J. (2005). *Aspects of Multivariate Statistical Theory, 2nd ed.* John Wiley & Sons, Inc.

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

Reiter, J. P. (2003). Inference for partially synthetic public use microdata sets. *Survey Methodology 29*, 181–188.

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

Reiter, J. P. (2005b). Using cart to generate partially synthetic public use microdata. *Journal of Official Statistics 21*(3), 441.

Reiter, J. P. and S. K. Kinney (2012). Inferentially valid, partially synthetic data: Generating from posterior predictive distributions not necessary. *Journal of Official Statistics 28*, 583–590.

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

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. Wiley.

Rubin, D. B. (1993). Discussion: Statistical disclosure limitation. *Journal of Official Statistics 9*, 461–468.

Appendix A Proofs of Theorems and Corollaries

Some matrix identities and matrix calculations required in the proof of Theorem 2.1.

1. If the matrices A and B are positive-definite then

   (i) $\mathbf{A}^{-1} - \mathbf{A}^{-1}(\mathbf{A}^{-1} + \mathbf{B}^{-1})^{-1}\mathbf{A}^{-1} = \mathbf{A}^{-1}(\mathbf{A}^{-1} + \mathbf{B}^{-1})^{-1}\mathbf{B}^{-1}$ and

   (ii) $(\mathbf{A^{-1}} + \mathbf{B^{-1}})^{\mathbf{-1}} = \mathbf{A}(\mathbf{A}+\mathbf{B})^{\mathbf{-1}}\mathbf{B} = \mathbf{B}(\mathbf{A}+\mathbf{B})^{\mathbf{-1}}\mathbf{A}$.

2. Let $\mathbf{S}$ and $\mathbf{\Sigma}$ be symmetric, then

$$\begin{aligned}
&(\mathbf{C}-\mathbf{X})\mathbf{S^{-1}}(\mathbf{C}-\mathbf{X})' + (\mathbf{X}-\mathbf{D})\mathbf{\Sigma^{-1}}(\mathbf{X}-\mathbf{D})' \\
&= \left[\mathbf{X} - (\mathbf{CS^{-1}} + \mathbf{D\Sigma^{-1}})(\mathbf{S^{-1}} + \mathbf{\Sigma^{-1}})^{-1}\right](\mathbf{S^{-1}} + \mathbf{\Sigma^{-1}})\left[\mathbf{X} - (\mathbf{CS^{-1}} + \mathbf{D\Sigma^{-1}})(\mathbf{S^{-1}} + \mathbf{\Sigma^{-1}})^{-1}\right]' \\
&\qquad + \mathbf{CS^{-1}C'} + \mathbf{D\Sigma^{-1}D'} - (\mathbf{CS^{-1}} + \mathbf{D\Sigma^{-1}})(\mathbf{S^{-1}} + \mathbf{\Sigma^{-1}})^{-1}(\mathbf{CS^{-1}} + \mathbf{D\Sigma^{-1}})'.
\end{aligned}$$

3. Taking the last three terms of the previous sum, and using the identities from item 1, we have,

$$\begin{aligned}
&\mathbf{CS^{-1}C'} - \mathbf{CS^{-1}}(\mathbf{S^{-1}} + \mathbf{\Sigma^{-1}})^{-1}\mathbf{S^{-1}C'} + \mathbf{D\Sigma^{-1}D'} - \mathbf{D\Sigma^{-1}}(\mathbf{S^{-1}} + \mathbf{\Sigma^{-1}})^{-1}\mathbf{\Sigma^{-1}D'} \\
&\qquad - \mathbf{CS^{-1}}(\mathbf{S^{-1}} + \mathbf{\Sigma^{-1}})^{-1}\mathbf{\Sigma^{-1}D'} - \mathbf{D\Sigma^{-1}}(\mathbf{S^{-1}} + \mathbf{\Sigma^{-1}})^{-1}\mathbf{S^{-1}C'} \\
&= \mathbf{C}(\mathbf{S}+\mathbf{\Sigma})^{\mathbf{-1}}(\mathbf{C'} - \mathbf{D'}) + \mathbf{D}(\mathbf{S}+\mathbf{\Sigma})^{\mathbf{-1}}(\mathbf{D'} - \mathbf{C'}) = (\mathbf{C}-\mathbf{D})(\mathbf{S}+\mathbf{\Sigma})^{\mathbf{-1}}(\mathbf{C}-\mathbf{D})'.
\end{aligned}$$

Proof of Theorem 2.1. : From (2), given $(\hat{\mathbf{B}}, \mathbf{S})$, we have,

$$\overline{\mathbf{B}}^*_M | (\hat{\mathbf{B}}, \mathbf{S}) \sim N_{pm}(\hat{\mathbf{B}}, \frac{1}{M}\mathbf{S} \otimes (\mathbf{XX}')^{-1})$$

and for $\overline{\mathbf{S}}^*_M$ in (3), from the fact that $(n-p)\mathbf{S}^*_j | \mathbf{S} \sim W_m(\mathbf{S}, n-p)$, and these are independent for $j = 1, \ldots, M$,

$$M(n-p)\overline{\mathbf{S}}^*_M | \mathbf{S} \sim W_m(\mathbf{S}, M(n-p)). \tag{A.1}$$

Given the independence of $\overline{\mathbf{B}}^*_M$ and $\overline{\mathbf{S}}^*_M$, conditional on $(\hat{\mathbf{B}}, \mathbf{S})$, the conditional joint pdf of $(\overline{\mathbf{B}}^*_M, \overline{\mathbf{S}}^*_M)$ is proportional to

$$\exp\left\{-\frac{1}{2}tr\left(M\mathbf{S}^{-1}\left[(\overline{\mathbf{B}}^*_M - \hat{\mathbf{B}})'\mathbf{XX}'(\overline{\mathbf{B}}^*_M - \hat{\mathbf{B}}) + (n-p)\overline{\mathbf{S}}^*_M\right]\right)\right\} \times \frac{|\overline{\mathbf{S}}^*_M|^{\frac{M(n-p)-m-1}{2}}}{|\mathbf{S}|^{\frac{M(n-p)+p}{2}}}, \tag{A.2}$$

while, given the independence of $\hat{\mathbf{B}}$ and $\mathbf{S}$, defined in the Introduction after (1), the joint pdf of $(\hat{\mathbf{B}}, \mathbf{S})$ is proportional to

$$\exp\left\{-\frac{1}{2} tr\left(\boldsymbol{\Sigma}^{-1}\left[(\hat{\mathbf{B}}-\mathbf{B})'\mathbf{X}\mathbf{X}'(\hat{\mathbf{B}}-\mathbf{B})+(n-p)\mathbf{S}\right]\right)\right\}\frac{|\mathbf{S}|^{\frac{n-p-m-1}{2}}}{|\boldsymbol{\Sigma}|^{\frac{n}{2}}}. \qquad \text{(A.3)}$$

Therefore, we obtain the joint pdf of $(\overline{\mathbf{B}}^*_M, \overline{\mathbf{S}}^*_M, \hat{\mathbf{B}}, \mathbf{S})$ by multiplying the two joint pdf's (A.2) and (A.3).

Since

$$\begin{aligned} & tr\{M\mathbf{S}^{-1}(\overline{\mathbf{B}}^*_M-\hat{\mathbf{B}})'(\mathbf{X}\mathbf{X}')(\overline{\mathbf{B}}^*_M-\hat{\mathbf{B}})+\boldsymbol{\Sigma}^{-1}(\hat{\mathbf{B}}-\mathbf{B})'(\mathbf{X}\mathbf{X}')(\hat{\mathbf{B}}-\mathbf{B})\} \\ & \qquad = tr\{M(\overline{\mathbf{B}}^*_M-\hat{\mathbf{B}})\mathbf{S}^{-1}(\overline{\mathbf{B}}^*_M-\hat{\mathbf{B}})'(\mathbf{X}\mathbf{X}')+(\hat{\mathbf{B}}-\mathbf{B})\boldsymbol{\Sigma}^{-1}(\hat{\mathbf{B}}-\mathbf{B})'(\mathbf{X}\mathbf{X}')\}, \end{aligned}$$

where, from the identities in 1–3 above,

$$\begin{aligned} & M(\overline{\mathbf{B}}^*_M-\hat{\mathbf{B}})\mathbf{S}^{-1}(\overline{\mathbf{B}}^*_M-\hat{\mathbf{B}})'+(\hat{\mathbf{B}}-\mathbf{B})\boldsymbol{\Sigma}^{-1}(\hat{\mathbf{B}}-\mathbf{B})' = \\ & = \left[\hat{\mathbf{B}}-(M\overline{\mathbf{B}}^*_M\mathbf{S}^{-1}+\mathbf{B}\boldsymbol{\Sigma}^{-1})(M\mathbf{S}^{-1}+\boldsymbol{\Sigma}^{-1})^{-1}\right](M\mathbf{S}^{-1}+\boldsymbol{\Sigma}^{-1}) \\ & \qquad \left[\hat{\mathbf{B}}-(M\overline{\mathbf{B}}^*_M\mathbf{S}^{-1}+\mathbf{B}\boldsymbol{\Sigma}^{-1})(M\mathbf{S}^{-1}+\boldsymbol{\Sigma}^{-1})^{-1}\right]' \\ & \qquad + (\overline{\mathbf{B}}^*_M-\mathbf{B})(\frac{1}{M}\mathbf{S}+\boldsymbol{\Sigma})^{-1}(\overline{\mathbf{B}}^*_M-\mathbf{B})', \end{aligned}$$

integrating out $\hat{\mathbf{B}}$, we obtain the joint pdf of $(\overline{\mathbf{B}}^*_M, \overline{\mathbf{S}}^*_M, \mathbf{S})$ proportional to

$$\begin{aligned} & \exp\left\{-\frac{1}{2}tr\left[(\boldsymbol{\Sigma}+\frac{1}{M}\mathbf{S})^{-1}(\overline{\mathbf{B}}^*_M-\mathbf{B})'(\mathbf{X}\mathbf{X}')(\overline{\mathbf{B}}^*_M-\mathbf{B})+M(n-p)\mathbf{S}^{-1}\overline{\mathbf{S}}^*_M+(n-p)\boldsymbol{\Sigma}^{-1}\mathbf{S}\right]\right\} \\ & \qquad \times |\overline{\mathbf{S}}^*_M|^{\frac{M(n-p)-m-1}{2}} \times \frac{|\mathbf{S}|^{-\frac{M(n-p)-n+2p+m+1}{2}}}{|\boldsymbol{\Sigma}|^{\frac{n}{2}}} \times |\boldsymbol{\Sigma}^{-1}+M\mathbf{S}^{-1}|^{-p/2}, \end{aligned} \qquad \text{(A.4)}$$

and integrating out $\mathbf{S}$, we obtain the result in the body of the Theorem for the joint pdf of $\overline{\mathbf{B}}^*_M$ and $\overline{\mathbf{S}}^*_M$.

In (A.4), $\overline{\mathbf{S}}^*_M$ and $\overline{\mathbf{B}}^*_M$ are separable, showing that they are independent, with, $\overline{\mathbf{B}}^*_M|\mathbf{S} \sim N_{pm}(\mathbf{B}, (\boldsymbol{\Sigma}+\frac{1}{M}\mathbf{S})\otimes(\mathbf{X}\mathbf{X}')^{-1})$ and $M(n-p)\overline{\mathbf{S}}^*_M|\mathbf{S} \sim W_m(\mathbf{S}, M(n-p))$. □

Proof of Corollary 2.2. :

From the result in Theorem 2.1, $(\overline{\mathbf{B}}^*_M - \mathbf{B})'|\mathbf{S} \sim N(\mathbf{0}, (\mathbf{X}\mathbf{X}')^{-1} \otimes (\mathbf{\Sigma} + \frac{1}{M}\mathbf{S}))$, and by Theorem 2.4.1 in Kollo and Rosen (2005) we have that, for $p \geq m$,

$$(\overline{\mathbf{B}}^*_M - \mathbf{B})'(\mathbf{X}\mathbf{X}')(\overline{\mathbf{B}}^*_M - \mathbf{B})|\mathbf{S} \sim W_m(\mathbf{\Sigma} + \frac{1}{M}\mathbf{S}, p).$$

Therefore, for

$$\mathbf{H} = (\mathbf{\Sigma} + \frac{1}{M}\mathbf{S})^{-\frac{1}{2}}(\overline{\mathbf{B}}^*_M - \mathbf{B})'(\mathbf{X}\mathbf{X}')(\overline{\mathbf{B}}^*_M - \mathbf{B})(\mathbf{\Sigma} + \frac{1}{M}\mathbf{S})^{-\frac{1}{2}}$$

and

$$\mathbf{G} = M(n-p)\mathbf{S}^{-\frac{1}{2}}\overline{\mathbf{S}}^*_M\mathbf{S}^{-\frac{1}{2}},$$

from Theorem 2.4.2 in Kollo and Rosen (2005) and Subsection 7.3.3 in Anderson (2003), from the distribution of $\overline{\mathbf{S}}^*_M$, we have $\mathbf{H}|\mathbf{S} \sim W_m(\mathbf{I}_m, p)$ and $\mathbf{G}|\mathbf{S} \sim W_m(\mathbf{I}_m, M(n-p))$, that, given the conditional independence of $\overline{\mathbf{B}}^*_M|\mathbf{S}$ and $\overline{\mathbf{S}}^*_M|\mathbf{S}$, are two independent random variables.

Since we may write

$$T_M|\mathbf{S} = \frac{|(\overline{\mathbf{B}}^*_M - \mathbf{B})'(\mathbf{X}\mathbf{X}')(\overline{\mathbf{B}}^*_M - \mathbf{B})|}{|(n-p)\overline{\mathbf{S}}^*_M|}|\mathbf{S} = \frac{|M\mathbf{\Sigma} + \mathbf{S}|}{|\mathbf{S}|} \times \frac{|\mathbf{H}|}{|\mathbf{G}|}|\mathbf{S},$$

where, given $\mathbf{S}$, $|\mathbf{G}| \sim \prod_{l=1}^m \chi^2_{M(n-p)-l+1}$ and $|\mathbf{H}| \sim \prod_{l=1}^m \chi^2_{p-l+1}$, with the chi-square random variables in each product independent, we end up with a product of independent F-distributions. So, conditionally on $\mathbf{S}$, we have

$$T_M|\mathbf{S} \overset{st}{\sim} \prod_{l=1}^{m}\left[\frac{p-l+1}{M(n-p)-l+1}F_{p,l}\right] \times |\mathbf{S}^{-1}(M\mathbf{\Sigma} + \mathbf{S})|,$$

where $F_{p,l} \sim \mathcal{F}_{p-l+1, M(n-p)-l+1}$.

Note that $(n-p)\mathbf{S} \sim W_m(\mathbf{\Sigma}, n-p)$, thus implying $\frac{1}{n-p}\mathbf{S}^{-1} \sim W_m^{-1}((n-p)\mathbf{\Sigma^{-1}}, n-p)$, or $\frac{1}{n-p}\mathbf{\Sigma}^{1/2}\mathbf{S}^{-1}\mathbf{\Sigma}^{1/2} \sim W_m^{-1}(\mathbf{I}_m, n-p)$, which shows that the distribution of $|\mathbf{S}^{-1}(M\mathbf{\Sigma} + \mathbf{S})| = |M\mathbf{\Sigma}^{1/2}\mathbf{S}^{-1}\mathbf{\Sigma}^{1/2} + \mathbf{I}_m|$ does not depend on $\mathbf{\Sigma}$, concluding the proof. □

Proof of Corollary 2.3. The proof is identical to the proof of Corollary 2.2 replacing, conditional on $\hat{\mathbf{B}}$ and $\mathbf{S}$, the joint pdf of $(\overline{\mathbf{B}}^*_M, \overline{\mathbf{S}}^*_M)$ by the joint pdf of $(\mathbf{\Delta}^*_M, \mathbf{D}'\overline{\mathbf{S}}^*_M\mathbf{D})$, where $\mathbf{\Delta}^*_M = \mathbf{A}\overline{\mathbf{B}}^*_M\mathbf{D}$, noting that we have

$$\mathbf{\Delta}^*_M|\mathbf{S} \sim N_{kr}\left(\mathbf{\Delta}, \mathbf{D}'\left(\mathbf{\Sigma} + \frac{1}{M}\mathbf{S}\right)\mathbf{D} \otimes \mathbf{A}(\mathbf{X}\mathbf{X}')^{-1}\mathbf{A}'\right)$$

and

$$M(n-p)\mathbf{D}'\overline{\mathbf{S}}^*_M\mathbf{D}|_{\mathbf{S}} \sim W_m(\mathbf{D}'\mathbf{S}\mathbf{D}, M(n-p)).$$

□

Proof of Theorem 2.4. The proof is identical to the proof of Theorem 2.1, replacing, conditional on $\hat{\mathbf{B}}$ and $\mathbf{S}$, the joint pdf of $(\overline{\mathbf{B}}^*_M, \overline{\mathbf{S}}^*_M)$ by the joint pdf of $(\overline{\mathbf{B}}^*_M, \mathbf{S}_{comb})$, noting that (see Moura (2016, Sec. 2.3.2))

$$(Mn-p)\mathbf{S}_{comb}|\mathbf{S} \sim W_m(\mathbf{S}, Mn-p). \tag{A.5}$$

□

Proof of Corollary 2.5. The proof is identical to the proof of Corollary 2.2, replacing, conditional on $\mathbf{S}$, $\overline{\mathbf{S}}^*_M$ by $\mathbf{S}_{comb}$ and the corresponding degrees of freedom, $M(n-p)$ by $Mn-p$, and taking into account that from Theorem 2.4 we have that $(Mn-p)\mathbf{S}_{comb}|\mathbf{S} \sim W_m(\mathbf{S}, Mn-p)$ is independent of $\overline{\mathbf{B}}^*_M|\mathbf{S}$. □

Proof of Corollary 2.6. The proof is identical to the proof of Corollary 2.2, replacing, conditional on $\hat{\mathbf{B}}$ and $\mathbf{S}$, the joint pdf of $(\overline{\mathbf{B}}^*_M, \overline{\mathbf{S}}^*_M)$ by the joint pdf of $(\mathbf{\Delta}^*_M, \mathbf{D}'\mathbf{S}_{comb}\mathbf{D})$, where $\mathbf{\Delta}^*_M = \mathbf{A}\overline{\mathbf{B}}^*_M\mathbf{D}$, noting that we have

$$\mathbf{\Delta}^*_M|\mathbf{S} \sim N_{kr}\left[\mathbf{\Delta}, \mathbf{D}'\left(\mathbf{\Sigma} + \frac{1}{M}\mathbf{S}\right)\mathbf{D} \otimes \mathbf{A}(\mathbf{X}\mathbf{X}')^{-1}\mathbf{A}'\right]$$

and

$$(Mn-p)\mathbf{D}'\overline{\mathbf{S}}^*_M\mathbf{D}|\mathbf{S} \sim W_m(\mathbf{D}'\mathbf{S}\mathbf{D}, Mn-p).$$

□

Appendix B

B.1 Details on Results in Subsection 2.1 and Remark 2.2

In this Appendix, some details about the derivations of Results 1 and 2 in Subsection 2.1 are provided.

Details of Result 1:

$$E(\overline{\mathbf{B}}^*_M) = (\mathbf{X}\mathbf{X}')^{-\mathbf{1}}\mathbf{X}\frac{1}{M}\sum_{j=1}^{M} E(\mathbf{V'}_j) = (\mathbf{X}\mathbf{X}')^{-\mathbf{1}}\mathbf{X}\frac{1}{M}\sum_{i=1}^{M} E(\mathbf{X}'\hat{\mathbf{B}}) = \mathbf{B}$$

and

$$Var(\overline{\mathbf{B}}^*_{\mathbf{M}}) = E(Var_{\overline{\mathbf{B}}^*_{\mathbf{M}}|\hat{\mathbf{B}},\mathbf{S}}(\overline{\mathbf{B}}^*_{\mathbf{M}}|\hat{\mathbf{B}},\mathbf{S})) + Var(E_{\overline{\mathbf{B}}^*_{\mathbf{M}}|\hat{\mathbf{B}},\mathbf{S}}(\overline{\mathbf{B}}^*_{\mathbf{M}}|\hat{\mathbf{B}},\mathbf{S})) = \frac{M+1}{M}\mathbf{\Sigma} \otimes (\mathbf{X}\mathbf{X}')^{-\mathbf{1}}.$$

Details of Result 2: Noting that $M(n-p)\overline{\mathbf{S}}^*_M|\mathbf{S} \sim W_m(\mathbf{S}, M(n-p))$ and that $(n-p)\mathbf{S} \sim W_m(\mathbf{\Sigma}, n-p)$ then, immediately, $E(\overline{\mathbf{S}}^*_M|\mathbf{S}) = E(\mathbf{S}) = \mathbf{\Sigma}$.

Details of Remark 2.2: Let us consider $\mathbf{H}$ and $\mathbf{G}$ as we did in Appendix A for the particular case of the single imputation case, i.e. for $M = 1$, and let us consider $\mathbf{B}^* = \overline{\mathbf{B}}^*_1$ and $\mathbf{S}^* = \overline{\mathbf{S}}^*_1$, since the generalization for the multiple case is straightforward. We will begin to decompose all the four statistics in order to make them assume the same kind of form and then prove why all of them are non-pivotal. The first statistic is

$$T^{(1)} = \frac{|\mathbf{S}^*|}{|\mathbf{S}^* + (\mathbf{B}^* - \mathbf{B})'(\mathbf{X}\mathbf{X}')(\mathbf{B}^* - \mathbf{B})|}$$

that we can decompose as

$$\begin{aligned} T^{(1)} &= \frac{|\mathbf{S}||(n-p)\mathbf{S}^{-1/2}\mathbf{S}^*\mathbf{S}^{-1/2}|}{(n-p)^m|\mathbf{S}^* + (\mathbf{\Sigma}+\mathbf{S})^{1/2}(\mathbf{\Sigma}+\mathbf{S})^{-1/2}(\mathbf{B}^*-\mathbf{B})'(\mathbf{X}\mathbf{X}')(\mathbf{B}^*-\mathbf{B})(\mathbf{\Sigma}+\mathbf{S})^{-1/2}(\mathbf{\Sigma}+\mathbf{S})^{1/2}|} \\ &= \frac{|\mathbf{G}|}{|\mathbf{G} + (n-p)\mathbf{S}^{-1/2}(\mathbf{\Sigma}+\mathbf{S})^{1/2}\mathbf{H}(\mathbf{\Sigma}+\mathbf{S})^{1/2}\mathbf{S}^{-1/2}|}. \end{aligned}$$

Now let us consider the following statistics

$$T^{(2)} = (n-p)tr\left[\mathbf{H} \times (\mathbf{\Sigma}+\mathbf{S})^{1/2}\mathbf{S}^{-1/2} \times \mathbf{G}^{-1} \times \mathbf{S}^{-1/2}(\mathbf{\Sigma}+\mathbf{S})^{1/2}\right],$$

$$T^{(3)} = tr\{\mathbf{H} \times [\mathbf{H} + (\mathbf{\Sigma}+\mathbf{S})^{-1/2}\mathbf{S}^{1/2} \times (n-p)\mathbf{G} \times \mathbf{S}^{1/2}(\mathbf{\Sigma}+\mathbf{S})^{-1/2}]^{-1}\}$$

and $T^{(4)} = \lambda_1$ where λ_1 denotes the largest eigenvalue of

$$(n-p)\mathbf{H} \times (\mathbf{\Sigma}+\mathbf{S})^{1/2}\mathbf{S}^{-1/2} \times \mathbf{G}^{-1} \times \mathbf{S}^{-1/2}(\mathbf{\Sigma}+\mathbf{S})^{1/2}.$$

From $T^{(1)}$ we can observe that a term of the denominator is

$$\mathbf{S}^{-\frac{1}{2}}(\mathbf{\Sigma}+\mathbf{S})^{\frac{1}{2}}\mathbf{H}(\mathbf{\Sigma}+\mathbf{S})^{\frac{1}{2}}\mathbf{S}^{-\frac{1}{2}}|_{\mathbf{S}} \sim W_m(\mathbf{S}^{-\frac{1}{2}}(\mathbf{\Sigma}+\mathbf{S})\mathbf{S}^{-\frac{1}{2}}, p) \equiv W_m((\mathbf{S}^{-\frac{1}{2}}\mathbf{\Sigma}\mathbf{S}^{-\frac{1}{2}}+\mathbf{I}), p),$$

and in the other statistics there are similar terms. We can also observe that all of the terms involve a product similar to $\mathbf{S}^{-\frac{1}{2}}(\mathbf{\Sigma}+\mathbf{S})^{\frac{1}{2}}$ that cannot be simplified the same way we could do when using the determinant as in the statistic T_M used in this paper.

Thus, in order to prove that these statistics are dependent on $\mathbf{\Sigma}$, we can see the empirical distributions of $T^{(1)}$, $T^{(2)}$, $T^{(3)}$ and $T^{(4)}$ when we consider a simple case where $m=2$, $p=3$, $n=100$ and $\Sigma = \left(\begin{smallmatrix} 1 & \rho \\ \rho & 1 \end{smallmatrix}\right)$ with $\rho = \{0.2, 0.4, 0.6, 0.8\}$ for a simulation size of 10^4, in Figure 3.

After making the above simulations we can observe from its distributions and cut-off points ($\gamma = 0.05$) that these four statistics are non-pivotal.

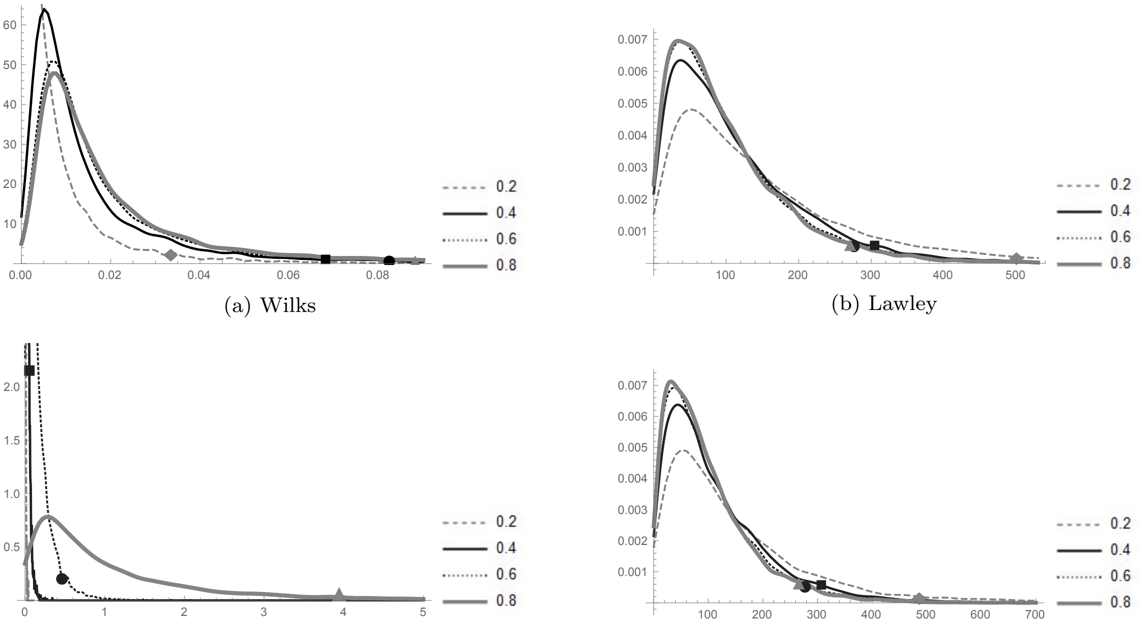
B.2 Details on Result 1 in Subsection 2.2

Noting that $(Mn-p)\mathbf{S}_{comb}|_{\mathbf{S}} \sim W_m(\mathbf{S}, Mn-p)$ and that $(n-p)\mathbf{S} \sim W_m(\mathbf{\Sigma}, n-p)$ then, immediately, $E(\mathbf{S}_{comb}) = E(\mathbf{S}) = \mathbf{\Sigma}$.

B.3 Proof of Remark 2.3

We may write

$$\text{Var}(\overline{\mathbf{S}}^*_M) = E\left[\text{Var}(\overline{\mathbf{S}}^*_M|\mathbf{S})\right] + \text{Var}\left[E(\overline{\mathbf{S}}^*_M|\mathbf{S})\right]$$



(a) Wilks

(b) Lawley

(c) Pillai

(d) Roy

Figure 3: Empirical distributions and cut-off points (γ=0.05) of $T^{(1)}$, $T^{(2)}$, $T^{(3)}$ and $T^{(4)}$ for $\rho = \{0.2, 0.4, 0.6, 0.8\}$.

and

$$\mathrm{Var}(\mathbf{S}_{comb}) = E\left[\mathrm{Var}(\mathbf{S}_{comb}|\mathbf{S})\right] + \mathrm{Var}\left[E(\mathbf{S}_{comb}|\mathbf{S})\right],$$

where, given the conditional distributions of $\overline{\mathbf{S}}^*_M$ and $\mathbf{S}_{comb}$, given $\mathbf{S}$, in (A.1) and (A.5) in Appendix A, we have

$$\mathrm{Var}\left[E(\overline{\mathbf{S}}^*_M|\mathbf{S})\right] = \mathrm{Var}\left[E(\mathbf{S}_{comb}|\mathbf{S})\right] = \mathrm{Var}(\mathbf{S})$$

and as such, to prove that $\mathrm{Var}(\overline{\mathbf{S}}^*_M) \geq \mathrm{Var}(\mathbf{S}_{comb})$ we only need to prove that

$$E\left[\mathrm{Var}(\overline{\mathbf{S}}^*_M|\mathbf{S})\right] \geq E\left[\mathrm{Var}(\mathbf{S}_{comb}|\mathbf{S})\right],$$

or, equivalently, that

$$\frac{E\left[\mathrm{Var}(\overline{\mathbf{S}}^*_M|\mathbf{S})\right]}{E\left[\mathrm{Var}(\mathbf{S}_{comb}|\mathbf{S})\right]} = \frac{\frac{1}{M^2(n-p)^2}E\left[\mathrm{Var}(M(n-p)\overline{\mathbf{S}}^*_M|\mathbf{S})\right]}{\frac{1}{(Mn-p)^2}E\left[\mathrm{Var}((Mn-p)\mathbf{S}_{comb}|\mathbf{S})\right]} \geq 1\,. \tag{B.1}$$

We know that if $\mathbf{A} \sim W_p(\mathbf{\Sigma}, \nu)$ then $\text{Var}(\mathbf{vec}(\mathbf{A})) = \nu(I_{p^2} + \mathbf{K})(\mathbf{\Sigma} \otimes \mathbf{\Sigma})$ where $\mathbf{K}$ is a commutation matrix of order $p^2{\times}p^2$ (Muirhead, 2005, pg. 90), so that, given the distributions of $M(n-p)\overline{\mathbf{S}}^*_M|\mathbf{S}$ and $(Mn-p)\mathbf{S}_{comb}|\mathbf{S}$ in (A.1) and (A.5), the two expected values on the right hand side of (B.1) have a common factor. On the numerator this factor appears multiplied by $M(n-p)$, while on the denominator it appears multiplied by $Mn-p$, so that we may write

$$\frac{E\left[\text{Var}(\overline{\mathbf{S}}^*_M|\mathbf{S})\right]}{E\left[\text{Var}(\mathbf{S}_{comb}|\mathbf{S})\right]} = \frac{\frac{1}{M(n-p)}}{\frac{1}{Mn-p}} = \frac{Mn-p}{M(n-p)}$$

which will be larger than 1 for $M \geq 2$ and equal to 1 for $M = 1$.

B.4 Details on Results in Section 3

Lastly, we provide some details about the derivations of the results in Section 3.

Details of Expect Values in Section 3: Recall that $(n-p)\mathbf{S} \sim W_m(\mathbf{\Sigma}, n-p)$, thus implying that $E(|(n-p)\mathbf{S}|) = |\mathbf{\Sigma}|E(\prod_{i=1}^m \chi^2_{n-p-i+1}) = \frac{(n-p)!}{(n-p-m)!}|\mathbf{\Sigma}|$, since $\prod_{i=1}^m \chi^2_{n-p-i+1}$ is a product of independent χ^2 variables. Also recalling that, conditionally on $\mathbf{S}$, we have $M(n-p)\overline{\mathbf{S}}^*{}_M \sim W_m(\mathbf{S}, M(n-p))$ and $(Mn-p)\mathbf{S}_{comb} \sim W_m(\mathbf{S}, Mn-p)$, thus implying that, conditionally on $\mathbf{S}$,

$$E(|(n-p)\overline{\mathbf{S}}^*_M|) = \frac{1}{M^m(n-p)^m} \times \frac{(Mn-Mp)!}{(Mn-Mp-m)!} \times |(n-p)\mathbf{S}|$$

and

$$E(|(n-p/M)\mathbf{S}_{comb}|) = \frac{1}{M^m(n-p)^m} \times \frac{(Mn-p)!}{(Mn-p-m)!} \times |(n-p)\mathbf{S}|.$$

Combining the result of $E(|(n-p)\mathbf{S}|)$ with each of the synthetic expected values, conditionally on $\mathbf{S}$, we end up with the expression for $E(\Upsilon_M)$ found in Section 3.

Supplementary material for “Inference for Multivariate Regression Model based on Synthetic Data generated using Plug-in Sampling”

Ricardo Moura
Center for Mathematics and Applications (CMA/UNL),
NOVA School of Science and Technology, NOVA University of Lisbon
Portuguese Navy Research Center (CINAV)
and Naval Academy, Alfeite, Almada
Martin Klein*
Division of Biometrics VIII, Office of Biostatistics,
Office of Translational Sciences, Center for Drug Evaluation and Research,
U.S. Food and Drug Administration,
Silver Spring, Maryland
John Zylstra
Department of Mathematics and Statistics,
University of Baltimore, Baltimore County (UMBC)
Carlos A. Coelho
Center for Mathematics and Applications (CMA/UNL)
and Mathematics Department,
NOVA School of Science and Technology, NOVA University of Lisbon
Bimal Sinha*
Department of Mathematics and Statistics,
University of Baltimore, Baltimore County (UMBC)
and Center for Statistical Research and Methodology (CSRM), U.S. Census Bureau

*__Disclaimer:__ This article reflects the views of the authors and should not be construed to represent the views and/or policies of neither the U.S. Food and Drug Administration or the U.S. Census Bureau

Part I – Analysis Under Non-Ideal Conditions

S1 Non-normal error distributions - continuous variables

In this section we briefly discuss the issue of robustness of our proposed synthetic data analysis methods when errors are non-normal, that is, we still assume model (1), with the error term $\mathbb{E}$ not normally distributed. In the sequel we consider two types of deviations from normality: t-type (keeping symmetry) and skew-normal type.

S1.1 t-Distribution

Regarding a multivariate t-type distribution, we generate $\mathbf{y}_i : m \times 1$ as

$$\mathbf{y}_i = \mathbf{B}'\mathbf{x}_i + \boldsymbol{\Sigma}^{1/2}\mathbf{t}_i\sqrt{\frac{\nu-2}{\nu}}, \text{ for } i = 1, 2, \ldots, n \tag{S.1}$$

where $\mathbf{t}_i = \mathbf{z}_i\sqrt{\frac{\nu}{\eta}}$ and $\mathbf{z}_i$'s are iid with each component distributed as $N(0,1)$ independent of $\eta \sim \chi^2_\nu$. This results in the $\mathbf{y}_i$'s being independent multivariate t-distributed vectors. In Table S1 we display the results of a simulation study under a similar scenario as in Section 3, except that the original data are now generated from the regression model (S.1) whose error term has a multivariate t-distribution. We observe in Table S1, that the average coverage probability of the confidence regions of our proposed procedures is approximately equal to the nominal value of 0.95. Moreover, we observe that for sufficiently large n the coverage probability of the procedures of Reiter (2005) is also approximately equal to 0.95 for the cases where $M \geq 5$.

S1.2 Skew-Normal Distribution

Under the skew-normal distribution (Azzalini, 1985; Henze, 1986), we generate $\mathbf{y}_i$ as

$$\mathbf{y}_i = \mathbf{B}'\mathbf{x}_i + \mathbf{\Sigma}^{1/2} z_i \tag{S.2}$$

where the m components of z_i are iid with each component distributed as $(s - \mu_s)/\sigma_s$. Here $s \sim$ Skew-normal$(0, 1, \lambda)$, that is, a skew-normal distribution with location 0, scale 1 and shape equal to λ, with $\mu_s = E(s) = \sqrt{\frac{2}{\pi}} \left[\frac{\lambda}{\sqrt{1+\lambda^2}}\right]$ and $\sigma_s = \sqrt{1 - \frac{2}{\pi}\left(\frac{\lambda^2}{1+\lambda^2}\right)}$. Under this data generation scheme, $\mathbf{y}_i$'s will have a skew-normal distribution. The parameter λ represents the extent of deviation from symmetry. In Table S2 are displayed the results of a simulation study under a similar scenario as in Section 3, except that the original data are now generated from regression model (S.2) whose error term has the skew-normal distribution. We may observe in Table S2 that the coverage probability of the confidence regions of our proposed procedures is again approximately equal to the nominal value of 0.95. We may also observe that again for sufficiently large values of n the coverage probability of the procedures of Reiter (2005) is again approximately equal to 0.95, for $M \geq 5$.

S2 Non-normal error distributions - discrete and other non-continuous variables

In this section we propose to show that our exact inference methods in Section 2 still perform well when the original variables are discrete or random variables with a spike at zero.

The following cases will be considered, regarding the distribution for the sensitive original variables: Binomial distribution, Poisson distribution, and a distribution with a spike at zero.

In all simulations we will consider that the original data is composed of two sensitive variables ($m = 2$) which will be independent of the non-sensitive variables. We will consider $p = 3$ non-sensitive variables, one of which will have all elements equal to 1, and the other will be generated as iid $N(0, 1)$, and held fixed for the entire simulation. The inclusion of a non-sensitive variable with all elements equal to 1 has to be made in this case since the sensitive variables will have expected values which are non-null.

For each case we generate $m = 2$ variables with the given discrete or spike at zero distribution, which we then suppose that are sensitive variables. From these variables we will synthesize m variables using the Plug-in sampling method and a multivariate regression model similar to model (1).

Let us denote by $w_{hi}^{(j)}$ the i-th value of the generated h-th synthetic variable in the j-th partially synthetic dataset ($i = 1, \dots, n; h = 1, \dots, m; j = 1, \dots, M$).

Then, for each case where we use a discrete distribution for the "original" sensitive variables we will follow two different approaches: (i) taking as the random sample of our synthetic variables, either the closest integer value to the value $w_{hi}^{(j)}$, that is, taking, for $h = 1, \dots, m$, $i = 1, \dots, n$ and $j = 1, \dots, M$,

$$y_{hi}^{(j)} = \begin{cases} \left\lfloor w_{hi}^{(j)} + 1/2 \right\rfloor , & \text{if } w_{hi}^{(j)} \geq 0 \\ 0\,, & \text{if } w_{hi}^{(j)} < 0 \end{cases} \tag{S.3}$$

where $\lfloor x \rfloor$ denotes the floor of the value x, as the i-th value for our h-th synthetic variable in the j-th imputed dataset, or (ii) we will simply take

$$y_{hi}^{(j)} = w_{hi}^{(j)}\,, \quad h = 1, \dots, m; i = 1, \dots, n; j = 1, \dots, M\,, \tag{S.4}$$

obtaining in this case a non-integer value, which anyway has been obtained from the original sensitive variables with a discrete distribution. We will call 'approach 1' the situation where we adopt (S.3) and 'approach 2' the situation where we adopt (S.4).

In the following Subsections, based on Monte Carlo simulations with 10^4 iterations, we evaluate the average coverage probability of our procedures and Reiter's adapted procedure when the released synthetic data is created using the two approaches referred above. Similarly to what was done in Section 3, we obtain confidence regions for $\mathbf{B}$ and for $\mathbf{AB}=\mathbf{C}$, respectively with $\mathbf{A}=\mathbf{I}_3$ and $\mathbf{A}=(\mathbf{0}_{2\times1}|\mathbf{I}_2)$, setting the level of the confidence region as 0.95.

The simulations carried out were done in a similar manner to the one used in Section 3 of the paper, with $m=2$ and $p=3$, with the difference that we now use a $\mathbf{B}$ parameter matrix of the form

$$\mathbf{B}=\begin{pmatrix}\mu_1 & \mu_2\\ 0 & 0\\ 0 & 0\end{pmatrix}$$

where $\mu_h=E(\mathbf{y}_h)$ $(h=1,2)$, since now the expected value of the discrete random variables is different from zero. We also used zeros for the other parameters in $\mathbf{B}$ since we choose to model the sensitive variables only through their mean values. The matrix $\mathbf{\Sigma}$ used was

$$\mathbf{\Sigma}=\begin{pmatrix}\sigma_1^2 & 0\\ 0 & \sigma_2^2\end{pmatrix}$$

since the two original sensitive variables were generated independently.

We understand that if the average coverage probability shows adequate values in these cases this is a good indication that similar behavior would be found for all cases.

S2.1 Binomial distribution

We assume that the original response variables are binomial distributed. Thus, we generate $\mathbf{Y}=(\mathbf{y}_1,\mathbf{y}_2)$ by independently generating $\mathbf{y}_1$ and $\mathbf{y}_2$ from Binomial distributions, as

$$y_{h,i}\sim B(n^*,p_h)\,,\qquad h=1,2;\; i=1,\ldots,n$$

with $p_h\in(0,1)$ and $n^*\in\mathbb{N}$.

In Tables S3 and S4, which correspond to approaches 1 and 2, considering different values of n, n^* and p_h ($h = 1, 2$), we may observe that the average coverage probability still presents values that are close to the nominal value of 0.95, and it may be important to notice that it is for the single imputation case that the results seem to be closer to 0.95.

S2.2 Poisson distribution

Now let us consider that the two response variables are originally Poisson distributed, by independently generating $\mathbf{y}_1$ and $\mathbf{y}_2$ from Poisson distributions, as

$$y_{h,i} \sim P(\lambda_h)$$

with $\lambda_h, \in \mathbb{R}^+$, for $i = 1, \dots, n$ and $h = 1, 2$.

In Tables S5 and S6, corresponding respectively to approaches 1 and 2, for different values of n and λ_h, we may observe that the average coverage probability still presents values that are close to the nominal value of 0.95, and that one more time it is noticeably the single imputation case that seems to present the best results.

S2.3 Distribution with a spike at zero

The idea is to simulate original data where the response variables have a percentage of values all equal zero and the remaining percentage has some random continuous distribution with values different from zero. With that objective in mind, we will consider that the sensitive variables are the product of a Bernoulli distributed variable and a Normal variable. We start by independently generating $\mathbf{y}_h$ as

$$y_{h,i} \sim Bernoulli(p_h) \times N(\mu_h, \sigma_h^2)$$

with $p_h \in (0, 1)$, $\mu_h \in \mathbb{R}$ and $\sigma_h^2 \in \mathbb{R}^+$, for $i = 1, \dots, n$ and $h = 1, 2$. By generating the original data this way it is expected to obtain, for each variable, an average of $n \times p_h$ values that are equal to zero.

To generate the Plug-in synthetic version ignoring the original format one as only to proceed as in Section 2. In order to create a synthetic dataset where the variable format maintains the original format, we propose the following procedure of generating the synthetic data:

1. Create two additional original variables $\mathbf{y}_1^\dagger$ and $\mathbf{y}_2^\dagger$ where $y_{h,i}^\dagger = 0$ if $y_{h,i} = 0$ and $y_{h,i}^\dagger = 1$ if $y_{h,i} \neq 0$
2. Consider the original data $\mathbf{Y} = (\mathbf{y}_1, \mathbf{y}_2)$ replaced by $\mathbf{Y}^\dagger = (\mathbf{y}_1, \mathbf{y}_1^\dagger, \mathbf{y}_2, \mathbf{y}_2^\dagger)$
3. Generate $\mathbf{V}_j^\dagger = (\mathbf{v}_1^{(j)}, \mathbf{v}_1^{\dagger(j)}, \mathbf{v}_2^{(j)}, \mathbf{v}_2^{\dagger(j)})$, proceeding as in Section 2
4. Consider the new version of the synthetic data $\mathbf{V}_j = (\mathbf{w}_1^{(j)}, \mathbf{w}_2^{(j)})$ where

$$\mathbf{w}_h^{(j)} = \mathbf{v}_h^{(j)} \mathbf{v}_h^{\dagger(j)},$$

   for $h = 1, 2$.

In Tables S7 and S8 we may observe that the average coverage probabilities still present values that are quite close to the nominal value of 0.95, for different values of n, p_h, μ_h and σ_h.

References

Azzalini, A. (1985). A class of distributions which includes the normal ones. *Scandinavian journal of statistics 12*(2), 171–178.

Henze, N. (1986). A probabilistic representation of the'skew-normal'distribution. *Scandinavian journal of statistics 13*(4), 271–275.

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

Table S1: Average coverage probabilities for **B** and **AB** when error distribution is multivariate t-type (where **vec**(**B**) and **vec**(**AB**) stand for the adapted Reiter procedure and **B**(1), **AB**(1) and **B**(2), **AB**(2) stand respectively for the first and second new inferential procedures)

(a) Average coverage for **B**

n	ν	$M = 1$	$M = 2$			$M = 5$			$M = 10$			$M = 20$		
			vec (**B**)	**B** (1)	**B** (2)	**vec** (**B**)	**B** (1)	**B** (2)	**vec** (**B**)	**B** (1)	**B** (2)	**vec** (**B**)	**B** (1)	**B** (2)
10	3	0.944	0.822	0.944	0.945	0.742	0.946	0.944	0.735	0.944	0.944	0.739	0.943	0.943
	5	0.946	0.828	0.944	0.947	0.743	0.948	0.946	0.735	0.945	0.946	0.742	0.945	0.946
	10	0.951	0.830	0.951	0.949	0.753	0.953	0.950	0.751	0.950	0.950	0.750	0.951	0.951
	15	0.950	0.828	0.951	0.950	0.751	0.950	0.949	0.744	0.949	0.950	0.747	0.950	0.948
	20	0.950	0.827	0.947	0.948	0.752	0.948	0.948	0.743	0.948	0.947	0.747	0.948	0.948
50	3	0.948	0.960	0.951	0.952	0.931	0.952	0.951	0.930	0.954	0.953	0.932	0.953	0.952
	5	0.948	0.953	0.947	0.948	0.927	0.951	0.950	0.926	0.950	0.949	0.929	0.951	0.950
	10	0.951	0.952	0.949	0.950	0.925	0.951	0.950	0.921	0.949	0.948	0.924	0.950	0.948
	15	0.950	0.950	0.949	0.950	0.928	0.950	0.949	0.924	0.952	0.951	0.924	0.950	0.949
	20	0.949	0.956	0.952	0.953	0.929	0.955	0.955	0.924	0.952	0.950	0.929	0.952	0.951
200	3	0.950	0.966	0.948	0.947	0.943	0.951	0.951	0.940	0.949	0.949	0.944	0.948	0.949
	5	0.954	0.965	0.950	0.948	0.943	0.946	0.946	0.939	0.948	0.948	0.943	0.946	0.947
	10	0.952	0.963	0.950	0.948	0.947	0.948	0.948	0.942	0.947	0.947	0.942	0.947	0.948
	15	0.950	0.963	0.950	0.949	0.947	0.951	0.951	0.942	0.950	0.950	0.944	0.947	0.948
	20	0.948	0.964	0.951	0.950	0.944	0.951	0.952	0.939	0.950	0.950	0.942	0.949	0.950

(b) Average coverage for **AB**

n	ν	$M = 1$	$M = 2$			$M = 5$			$M = 10$			$M = 20$		
			vec (**AB**)	**AB** (1)	**AB** (2)	**vec** (**AB**)	**AB** (1)	**AB** (2)	**vec** (**AB**)	**AB** (1)	**AB** (2)	**vec** (**AB**)	**AB** (1)	**AB** (2)
10	3	0.943	0.956	0.944	0.944	0.777	0.945	0.944	0.776	0.944	0.944	0.772	0.943	0.943
	5	0.944	0.958	0.946	0.946	0.778	0.945	0.945	0.781	0.944	0.944	0.776	0.942	0.943
	10	0.949	0.965	0.948	0.949	0.790	0.951	0.952	0.797	0.950	0.949	0.787	0.949	0.950
	15	0.952	0.967	0.951	0.951	0.798	0.953	0.953	0.799	0.950	0.950	0.789	0.951	0.951
	20	0.949	0.962	0.951	0.949	0.791	0.949	0.948	0.794	0.948	0.948	0.787	0.948	0.949
50	3	0.948	0.999	0.950	0.950	0.933	0.952	0.952	0.935	0.953	0.952	0.933	0.954	0.952
	5	0.949	0.999	0.950	0.951	0.934	0.952	0.951	0.932	0.952	0.950	0.933	0.953	0.952
	10	0.952	0.999	0.952	0.952	0.930	0.953	0.951	0.928	0.951	0.950	0.927	0.949	0.947
	15	0.955	0.999	0.955	0.955	0.935	0.953	0.952	0.935	0.954	0.952	0.932	0.951	0.951
	20	0.948	0.999	0.948	0.948	0.935	0.952	0.951	0.934	0.950	0.949	0.934	0.951	0.950
200	3	0.950	1.000	0.953	0.952	0.943	0.951	0.952	0.944	0.954	0.951	0.943	0.952	0.951
	5	0.957	1.000	0.953	0.952	0.949	0.950	0.951	0.947	0.953	0.951	0.946	0.952	0.952
	10	0.952	1.000	0.955	0.954	0.944	0.948	0.948	0.942	0.949	0.948	0.942	0.948	0.947
	15	0.952	1.000	0.952	0.952	0.945	0.952	0.953	0.946	0.952	0.949	0.944	0.951	0.950
	20	0.948	1.000	0.953	0.952	0.947	0.952	0.953	0.947	0.954	0.951	0.946	0.953	0.952

Table S2: Average coverage probabilities for **B** and **AB** when error distribution is skew-normal (where **vec**(**B**) and **vec**(**AB**) stand for the adapted Reiter procedure and **B**(1), **AB**(1) and **B**(2), **AB**(2) stand respectively for the first and second new inferential procedures)

(a) Average coverage for **B**

n	ν	$M = 1$	$M = 2$			$M = 5$			$M = 10$			$M = 20$		
			vec (**B**)	**B** (1)	**B** (2)	**vec** (**B**)	**B** (1)	**B** (2)	**vec** (**B**)	**B** (1)	**B** (2)	**vec** (**B**)	**B** (1)	**B** (2)
10	10	0.944	0.817	0.942	0.941	0.738	0.944	0.942	0.731	0.942	0.941	0.734	0.942	0.940
	-10	0.949	0.819	0.942	0.942	0.736	0.944	0.942	0.728	0.945	0.944	0.729	0.944	0.943
	100	0.944	0.815	0.942	0.941	0.739	0.943	0.942	0.730	0.941	0.943	0.731	0.942	0.942
	-100	0.950	0.816	0.944	0.944	0.735	0.944	0.942	0.726	0.943	0.942	0.730	0.944	0.942
50	10	0.942	0.947	0.941	0.944	0.917	0.945	0.945	0.913	0.941	0.944	0.917	0.946	0.945
	-10	0.955	0.954	0.950	0.951	0.922	0.946	0.945	0.913	0.942	0.945	0.918	0.944	0.943
	100	0.943	0.948	0.942	0.944	0.918	0.946	0.945	0.914	0.944	0.946	0.916	0.944	0.944
	-100	0.953	0.953	0.951	0.952	0.923	0.946	0.945	0.913	0.943	0.945	0.917	0.944	0.943
200	10	0.952	0.971	0.953	0.952	0.950	0.957	0.956	0.943	0.953	0.953	0.943	0.949	0.950
	-10	0.948	0.964	0.947	0.946	0.947	0.949	0.949	0.943	0.953	0.952	0.948	0.953	0.954
	100	0.953	0.969	0.953	0.952	0.947	0.956	0.956	0.945	0.953	0.953	0.941	0.950	0.952
	-100	0.950	0.963	0.950	0.948	0.948	0.949	0.949	0.946	0.950	0.950	0.947	0.953	0.954

(b) Average coverage for **AB**

n	ν	$M = 1$	$M = 2$			$M = 5$			$M = 10$			$M = 20$		
			vec (**AB**)	**AB** (1)	**AB** (2)	**vec** (**AB**)	**AB** (1)	**AB** (2)	**vec** (**AB**)	**AB** (1)	**AB** (2)	**vec** (**AB**)	**AB** (1)	**AB** (2)
10	10	0.950	0.964	0.950	0.950	0.798	0.950	0.950	0.796	0.949	0.948	0.793	0.950	0.950
	-10	0.953	0.965	0.952	0.950	0.792	0.950	0.950	0.798	0.951	0.950	0.791	0.950	0.950
	100	0.948	0.967	0.950	0.950	0.795	0.950	0.950	0.799	0.952	0.950	0.791	0.950	0.951
	-100	0.953	0.965	0.952	0.951	0.791	0.951	0.950	0.797	0.950	0.950	0.789	0.948	0.949
50	10	0.946	0.999	0.944	0.945	0.927	0.946	0.944	0.928	0.949	0.947	0.927	0.950	0.948
	-10	0.950	0.999	0.950	0.950	0.928	0.948	0.948	0.926	0.950	0.948	0.924	0.948	0.947
	100	0.945	0.999	0.946	0.946	0.926	0.947	0.946	0.928	0.950	0.948	0.926	0.951	0.949
	-100	0.951	1.000	0.953	0.953	0.928	0.947	0.946	0.926	0.948	0.947	0.924	0.948	0.947
200	10	0.950	1.000	0.955	0.954	0.949	0.953	0.954	0.947	0.951	0.950	0.946	0.956	0.955
	-10	0.951	1.000	0.951	0.951	0.944	0.946	0.947	0.947	0.951	0.949	0.945	0.950	0.950
	100	0.953	1.000	0.955	0.954	0.949	0.951	0.953	0.947	0.953	0.951	0.949	0.954	0.952
	-100	0.951	1.000	0.952	0.950	0.943	0.948	0.948	0.944	0.952	0.950	0.946	0.951	0.950

Table S3: Average coverage probabilities for **B** and **AB** when the original response variables are Binomial and the Plug-in response variables are left as continuous, for $p_1 = 0.3$ and $p_2 = 0.7$ (where **vec**(**B**) and **vec**(**AB**) stand for the adapted Reiter procedure and **B**(1), **AB**(1) and **B**(2), **AB**(2) stand respectively for the first and second new inferential procedures)

(a) Average coverage for **B**

			$M = 2$			$M = 5$			$M = 10$			$M = 20$		
n	n^*	$M = 1$	**vec**	**B**	**B**	**vec**	**B**	**B**	**vec**	**B**	**B**	**vec**	**B**	**B**
			(**B**)	(1)	(2)	(**B**)	(1)	(2)	(**B**)	(1)	(2)	(**B**)	(1)	(2)
20	4	0.948	0.935	0.952	0.951	0.870	0.951	0.952	0.871	0.958	0.958	0.878	0.962	0.962
	10	0.953	0.926	0.944	0.942	0.855	0.946	0.944	0.853	0.941	0.942	0.858	0.941	0.939
100	4	0.953	0.958	0.951	0.952	0.94	0.949	0.951	0.931	0.954	0.954	0.930	0.947	0.945
	10	0.954	0.955	0.953	0.953	0.927	0.942	0.943	0.921	0.933	0.934	0.928	0.932	0.931
200	4	0.943	0.959	0.946	0.946	0.933	0.950	0.950	0.934	0.946	0.945	0.938	0.951	0.951
	10	0.949	0.963	0.948	0.948	0.938	0.950	0.949	0.940	0.942	0.942	0.94	0.948	0.949

(b) Average coverage for **AB**

			$M = 2$			$M = 5$			$M = 10$			$M = 20$		
n	n^*	$M = 1$	**vec**	**AB**	**AB**	**vec**	**AB**	**AB**	**vec**	**AB**	**AB**	**vec**	**AB**	**AB**
			(**AB**)	(1)	(2)	(**AB**)	(1)	(2)	(**AB**)	(1)	(2)	(**AB**)	(1)	(2)
20	4	0.951	0.994	0.944	0.948	0.885	0.951	0.949	0.882	0.949	0.948	0.879	0.946	0.946
	10	0.950	0.991	0.960	0.962	0.883	0.944	0.944	0.884	0.945	0.946	0.875	0.942	0.941
100	4	0.946	1.000	0.944	0.945	0.929	0.953	0.954	0.934	0.953	0.953	0.931	0.952	0.953
	10	0.945	1.000	0.943	0.946	0.940	0.931	0.932	0.944	0.938	0.941	0.937	0.939	0.939
200	4	0.949	0.999	0.943	0.943	0.945	0.942	0.945	0.939	0.942	0.939	0.941	0.938	0.938
	10	0.946	1.000	0.955	0.953	0.939	0.940	0.940	0.946	0.945	0.942	0.943	0.948	0.947

Table S4: Average coverage probabilities for **B** and **AB** when the original response variables are Binomial and the Plug-in response variables have this same type, for $p_1 = 0.3$ and $p_2 = 0.7$ (where **vec**(**B**) and **vec**(**AB**) stand for the adapted Reiter procedure and **B**(1), **AB**(1) and **B**(2), **AB**(2) stand respectively for the first and second new inferential procedures)

(a) Average coverage for **B**

n	n^*	$M = 1$	$M = 2$			$M = 5$			$M = 10$			$M = 20$		
			vec	**B**	**B**	**vec**	**B**	**B**	**vec**	**B**	**B**	**vec**	**B**	**B**
			(**B**)	(1)	(2)	(**B**)	(1)	(2)	(**B**)	(1)	(2)	(**B**)	(1)	(2)
20	4	0.959	0.946	0.964	0.964	0.917	0.969	0.969	0.914	0.969	0.970	0.917	0.970	0.970
	10	0.954	0.933	0.954	0.953	0.893	0.955	0.956	0.888	0.956	0.957	0.892	0.957	0.957
100	4	0.956	0.970	0.958	0.959	0.955	0.961	0.962	0.956	0.965	0.965	0.958	0.963	0.963
	10	0.955	0.968	0.958	0.958	0.950	0.956	0.957	0.949	0.955	0.955	0.952	0.958	0.957
200	4	0.956	0.971	0.962	0.962	0.956	0.961	0.961	0.956	0.958	0.958	0.960	0.962	0.962
	10	0.952	0.968	0.953	0.952	0.953	0.958	0.958	0.953	0.958	0.958	0.957	0.958	0.959

(b) Average coverage for **AB**

n	n^*	$M = 1$	$M = 2$			$M = 5$			$M = 10$			$M = 20$		
			vec	**AB**	**AB**	**vec**	**AB**	**AB**	**vec**	**AB**	**AB**	**vec**	**AB**	**AB**
			(**AB**)	(1)	(2)	(**AB**)	(1)	(2)	(**AB**)	(1)	(2)	(**AB**)	(1)	(2)
20	4	0.962	0.998	0.964	0.965	0.933	0.969	0.969	0.934	0.966	0.967	0.933	0.968	0.968
	10	0.954	0.996	0.953	0.954	0.905	0.955	0.954	0.910	0.955	0.955	0.906	0.958	0.958
100	4	0.955	1.000	0.962	0.963	0.957	0.962	0.962	0.962	0.964	0.966	0.963	0.967	0.967
	10	0.952	1.000	0.954	0.956	0.949	0.954	0.955	0.950	0.956	0.957	0.948	0.953	0.954
200	4	0.958	1.000	0.963	0.962	0.964	0.964	0.964	0.965	0.966	0.964	0.966	0.966	0.965
	10	0.954	1.000	0.958	0.956	0.953	0.952	0.954	0.952	0.956	0.954	0.953	0.956	0.955

Table S5: Average coverage probabilities for **B** and **AB** when the original response variables are Poisson and the Plug-in response variables are left as continuous (where **vec**(**B**) and **vec**(**AB**) stand for the adapted Reiter procedure and **B**(1), **AB**(1) and **B**(2), **AB**(2) stand respectively for the first and second new inferential procedures)

(a) Average coverage for **B**

n	λ_h	$M = 1$	$M = 2$			$M = 5$			$M = 10$			$M = 20$		
			vec (**B**)	**B** (1)	**B** (2)	**vec** (**B**)	**B** (1)	**B** (2)	**vec** (**B**)	**B** (1)	**B** (2)	**vec** (**B**)	**B** (1)	**B** (2)
20	$\lambda_1 = 2$ $\lambda_2 = 3$	0.949	0.918	0.951	0.950	0.871	0.946	0.946	0.867	0.947	0.950	0.872	0.949	0.948
	$\lambda_1 = 7$ $\lambda_2 = 8$	0.950	0.920	0.948	0.948	0.870	0.946	0.947	0.864	0.947	0.949	0.869	0.949	0.949
100	$\lambda_1 = 2$ $\lambda_2 = 3$	0.946	0.959	0.947	0.950	0.939	0.944	0.946	0.930	0.949	0.949	0.930	0.948	0.947
	$\lambda_1 = 7$ $\lambda_2 = 8$	0.949	0.963	0.944	0.945	0.929	0.945	0.946	0.938	0.943	0.943	0.940	0.946	0.946
200	$\lambda_1 = 2$ $\lambda_2 = 3$	0.953	0.960	0.959	0.959	0.945	0.947	0.947	0.946	0.950	0.950	0.950	0.955	0.957
	$\lambda_1 = 7$ $\lambda_2 = 8$	0.950	0.960	0.947	0.946	0.945	0.954	0.954	0.940	0.943	0.943	0.937	0.940	0.941

(b) Average coverage for **AB**

n	λ_h	$M = 1$	$M = 2$			$M = 5$			$M = 10$			$M = 20$		
			vec (**AB**)	**AB** (1)	**AB** (2)	**vec** (**AB**)	**AB** (1)	**AB** (2)	**vec** (**AB**)	**AB** (1)	**AB** (2)	**vec** (**AB**)	**AB** (1)	**AB** (2)
50	$\lambda_1 = 2$ $\lambda_2 = 3$	0.949	0.996	0.952	0.953	0.889	0.951	0.950	0.892	0.951	0.951	0.891	0.952	0.952
	$\lambda_1 = 7$ $\lambda_2 = 8$	0.948	0.996	0.946	0.946	0.889	0.947	0.947	0.891	0.949	0.948	0.888	0.948	0.948
100	$\lambda_1 = 2$ $\lambda_2 = 3$	0.945	1.000	0.948	0.952	0.947	0.951	0.951	0.939	0.946	0.946	0.941	0.943	0.943
	$\lambda_1 = 7$ $\lambda_2 = 8$	0.942	0.999	0.943	0.942	0.937	0.948	0.948	0.94	0.944	0.945	0.941	0.951	0.952
200	$\lambda_1 = 2$ $\lambda_2 = 3$	0.958	1.000	0.955	0.955	0.944	0.947	0.947	0.951	0.958	0.9657	0.950	0.955	0.955
	$\lambda_1 = 7$ $\lambda_2 = 8$	0.944	1.000	0.950	0.947	0.935	0.942	0.943	0.938	0.945	0.945	0.938	0.944	0.945

Table S6: Average coverage probabilities for **B** and **AB** when the original response variables are Poisson and the Plug-in response variables have this same type (where **vec**(**B**) and **vec**(**AB**) stand for the adapted Reiter procedure and **B**(1), **AB**(1) and **B**(2), **AB**(2) stand respectively for the first and second new inferential procedures)

(a) Average coverage for **B**

n	λ_h	$M = 1$	$M = 2$			$M = 5$			$M = 10$			$M = 20$		
			vec	**B**	**B**	**vec**	**B**	**B**	**vec**	**B**	**B**	**vec**	**B**	**B**
			(**B**)	(1)	(2)	(**B**)	(1)	(2)	(**B**)	(1)	(2)	(**B**)	(1)	(2)
20	$\lambda_1 = 2$ $\lambda_2 = 3$	0.954	0.926	0.958	0.956	0.887	0.954	0.955	0.883	0.956	0.957	0.890	0.959	0.958
	$\lambda_1 = 7$ $\lambda_2 = 8$	0.953	0.924	0.952	0.952	0.876	0.948	0.948	0.869	0.950	0.952	0.874	0.952	0.953
100	$\lambda_1 = 2$ $\lambda_2 = 3$	0.945	0.962	0.950	0.951	0.948	0.949	0.950	0.934	0.954	0.955	0.940	0.949	0.949
	$\lambda_1 = 7$ $\lambda_2 = 8$	0.950	0.965	0.945	0.946	0.930	0.948	0.948	0.937	0.945	0.945	0.942	0.946	0.947
200	$\lambda_1 = 2$ $\lambda_2 = 3$	0.953	0.960	0.950	0.948	0.940	0.957	0.956	0.952	0.950	0.950	0.949	0.953	0.953
	$\lambda_1 = 7$ $\lambda_2 = 8$	0.947	0.961	0.942	0.941	0.944	0.950	0.950	0.941	0.948	0.947	0.944	0.947	0.948

(b) Average coverage for **AB**

n	λ_h	$M = 1$	$M = 2$			$M = 5$			$M = 10$			$M = 20$		
			vec	**AB**	**AB**	**vec**	**AB**	**AB**	**vec**	**AB**	**AB**	**vec**	**AB**	**AB**
			(**AB**)	(1)	(2)	(**AB**)	(1)	(2)	(**AB**)	(1)	(2)	(**AB**)	(1)	(2)
20	$\lambda_1 = 2$ $\lambda_2 = 3$	0.952	0.997	0.958	0.958	0.906	0.958	0.957	0.912	0.958	0.959	0.910	0.960	0.960
	$\lambda_1 = 7$ $\lambda_2 = 8$	0.950	0.996	0.947	0.947	0.894	0.949	0.947	0.896	0.950	0.950	0.890	0.951	0.950
100	$\lambda_1 = 2$ $\lambda_2 = 3$	0.948	0.999	0.953	0.953	0.954	0.958	0.958	0.951	0.952	0.954	0.950	0.950	0.951
	$\lambda_1 = 7$ $\lambda_2 = 8$	0.946	0.999	0.947	0.947	0.939	0.951	0.951	0.941	0.949	0.949	0.943	0.952	0.953
200	$\lambda_1 = 2$ $\lambda_2 = 3$	0.955	1.000	0.958	0.958	0.956	0.953	0.954	0.959	0.960	0.957	0.957	0.956	0.956
	$\lambda_1 = 7$ $\lambda_2 = 8$	0.947	1.000	0.949	0.948	0.942	0.945	0.946	0.943	0.949	0.947	0.942	0.950	0.949

Table S7: Average coverage probabilities for **B** and **AB** when the original response variables have a spike at zero and the Plug-in response variables are left as continuous, for $\mu_1 = 20, \sigma_1 = 1, \mu_2 = 150$ and $\sigma_2 = 50$ (where $\mathbf{vec}(\mathbf{B})$ and $\mathbf{vec}(\mathbf{AB})$ stand for the adapted Reiter procedure and $\mathbf{B}(1)$, $\mathbf{AB}(1)$ and $\mathbf{B}(2)$, $\mathbf{AB}(2)$ stand respectively for the first and second new inferential procedures)

(a) Average coverage for **B**

n	p_h	$M = 1$	$M = 2$			$M = 5$			$M = 10$			$M = 20$		
			vec (**B**)	**B** (1)	**B** (2)	**vec** (**B**)	**B** (1)	**B** (2)	**vec** (**B**)	**B** (1)	**B** (2)	**vec** (**B**)	**B** (1)	**B** (2)
50	$p_1 = 0.3$ $p_2 = 0.7$	0.958	0.945	0.943	0.944	0.932	0.960	0.957	0.924	0.956	0.958	0.922	0.954	0.954
	$p_1 = 0.4$ $p_2 = 0.6$	0.956	0.959	0.950	0.948	0.948	0.958	0.957	0.936	0.958	0.959	0.930	0.957	0.957
100	$p_1 = 0.3$ $p_2 = 0.7$	0.945	0.949	0.943	0.942	0.931	0.940	0.941	0.924	0.947	0.947	0.929	0.947	0.945
	$p_1 = 0.4$ $p_2 = 0.6$	0.947	0.945	0.946	0.946	0.935	0.938	0.940	0.939	0.946	0.946	0.930	0.941	0.940
200	$p_1 = 0.3$ $p_2 = 0.7$	0.950	0.963	0.954	0.951	0.933	0.950	0.951	0.945	0.948	0.948	0.946	0.953	0.953
	$p_1 = 0.4$ $p_2 = 0.6$	0.947	0.962	0.955	0.953	0.948	0.955	0.957	0.946	0.958	0.958	0.950	0.957	0.957

(b) Average coverage for **AB**

n	p_h	$M = 1$	$M = 2$			$M = 5$			$M = 10$			$M = 20$		
			vec (**AB**)	**AB** (1)	**AB** (2)	**vec** (**AB**)	**AB** (1)	**AB** (2)	**vec** (**AB**)	**AB** (1)	**AB** (2)	**vec** (**AB**)	**AB** (1)	**AB** (2)
50	$p_1 = 0.3$ $p_2 = 0.7$	0.946	0.999	0.956	0.956	0.947	0.963	0.96	0.941	0.963	0.959	0.926	0.957	0.956
	$p_1 = 0.4$ $p_2 = 0.6$	0.952	0.999	0.948	0.947	0.947	0.949	0.948	0.934	0.946	0.946	0.932	0.945	0.943
100	$p_1 = 0.3$ $p_2 = 0.7$	0.941	0.999	0.946	0.947	0.937	0.939	0.939	0.940	0.945	0.947	0.937	0.947	0.948
	$p_1 = 0.4$ $p_2 = 0.6$	0.946	0.999	0.944	0.944	0.936	0.945	0.945	0.930	0.938	0.938	0.933	0.943	0.945
200	$p_1 = 0.3$ $p_2 = 0.7$	0.942	0.999	0.943	0.943	0.941	0.944	0.944	0.953	0.945	0.946	0.953	0.948	0.947
	$p_1 = 0.4$ $p_2 = 0.6$	0.944	0.999	0.952	0.952	0.944	0.951	0.953	0.949	0.954	0.951	0.950	0.955	0.955

Table S8: Average coverage probabilities for **B** and **AB** when the original response variables have a spike at zero and the Plug-in response variables have this same type, for $\mu_1 = 20, \sigma_1 = 1, \mu_2 = 150$ and $\sigma_2 = 50$ (where **vec**(**B**) and **vec**(**AB**) stand for the adapted Reiter procedure and **B**(1), **AB**(1) and **B**(2), **AB**(2) stand respectively for the first and second new inferential procedures)

(a) Average coverage for **B**

n	p_h	$M = 1$	$M = 2$			$M = 5$			$M = 10$			$M = 20$		
			vec (**B**)	**B** (1)	**B** (2)	**vec** (**B**)	**B** (1)	**B** (2)	**vec** (**B**)	**B** (1)	**B** (2)	**vec** (**B**)	**B** (1)	**B** (2)
50	$p_1 = 0.3$ $p_2 = 0.7$	0.940	0.936	0.940	0.942	0.899	0.944	0.942	0.899	0.943	0.945	0.903	0.944	0.942
	$p_1 = 0.4$ $p_2 = 0.6$	0.944	0.949	0.948	0.949	0.918	0.952	0.951	0.916	0.950	0.952	0.920	0.950	0.950
100	$p_1 = 0.3$ $p_2 = 0.7$	0.926	0.928	0.922	0.922	0.874	0.915	0.916	0.864	0.920	0.921	0.867	0.922	0.919
	$p_1 = 0.4$ $p_2 = 0.6$	0.932	0.935	0.927	0.928	0.893	0.923	0.924	0.887	0.926	0.926	0.89	0.929	0.928
200	$p_1 = 0.3$ $p_2 = 0.7$	0.903	0.894	0.890	0.884	0.795	0.879	0.880	0.784	0.874	0.876	0.775	0.871	0.874
	$p_1 = 0.4$ $p_2 = 0.6$	0.913	0.906	0.902	0.900	0.822	0.878	0.877	0.804	0.868	0.868	0.803	0.865	0.865

(b) Average coverage for **AB**

n	p_h	$M = 1$	$M = 2$			$M = 5$			$M = 10$			$M = 20$		
			vec (**AB**)	**AB** (1)	**AB** (2)	**vec** (**AB**)	**AB** (1)	**AB** (2)	**vec** (**AB**)	**AB** (1)	**AB** (2)	**vec** (**AB**)	**AB** (1)	**AB** (2)
50	$p_1 = 0.3$ $p_2 = 0.7$	0.966	1.000	0.970	0.971	0.970	0.975	0.974	0.974	0.979	0.978	0.974	0.981	0.980
	$p_1 = 0.4$ $p_2 = 0.6$	0.964	1.000	0.971	0.971	0.967	0.974	0.974	0.970	0.976	0.975	0.971	0.977	0.976
100	$p_1 = 0.3$ $p_2 = 0.7$	0.963	1.000	0.971	0.972	0.976	0.974	0.973	0.980	0.976	0.977	0.981	0.976	0.976
	$p_1 = 0.4$ $p_2 = 0.6$	0.963	1.000	0.968	0.969	0.976	0.973	0.973	0.980	0.974	0.975	0.98	0.976	0.977
200	$p_1 = 0.3$ $p_2 = 0.7$	0.966	1.000	0.974	0.973	0.976	0.976	0.978	0.980	0.974	0.973	0.979	0.977	0.977
	$p_1 = 0.4$ $p_2 = 0.6$	0.965	1.000	0.973	0.973	0.973	0.973	0.974	0.975	0.978	0.978	0.971	0.980	0.980

Part II – Individual confidence intervals for Regression coefficients

S3 Detailed proofs of Result 5 in Subsection 2.1 and Result 4 in Subsection 2.2

To infer about $\mathbf{ABD} = \boldsymbol{\Delta} : k \times r$ where $\mathbf{A} : k \times p$, $\mathbf{B} : p \times m$, $\mathbf{D} : m \times r$ with $r \leq k$, we start from its natural point estimator

$$\boldsymbol{\Delta}_M^* = \mathbf{A}\overline{\mathbf{B}}_M^*\mathbf{D},$$

and propose to use the statistic

$$T_{M,\boldsymbol{\Delta}} = \frac{|(\boldsymbol{\Delta}_M^* - \boldsymbol{\Delta})'[\mathbf{A}(\mathbf{XX}')^{-1}\mathbf{A}']^{-1}(\boldsymbol{\Delta}_M^* - \boldsymbol{\Delta})|}{|(n-p)\mathbf{D}'\overline{\mathbf{S}}_M^*\mathbf{D}|}.$$

A more detailed proof of the distribution of $T_{M,\boldsymbol{\Delta}}$, may be obtained by writing

$$T_{M,\boldsymbol{\Delta}} = T_{\boldsymbol{\Delta}}^{(1)} \times T_{\boldsymbol{\Delta}}^{(2)} = \frac{|(\boldsymbol{\Delta}_M^* - \boldsymbol{\Delta})'[\mathbf{A}(\mathbf{XX}')^{-1}\mathbf{A}']^{-1}(\boldsymbol{\Delta}_M^* - \boldsymbol{\Delta})|}{|\mathbf{D}'(\boldsymbol{\Sigma} + \frac{1}{M}\mathbf{S})\mathbf{D}|} \times \frac{|\mathbf{D}'(\boldsymbol{\Sigma} + \frac{1}{M}\mathbf{S})\mathbf{D}|}{|(n-p)\mathbf{D}'\overline{\mathbf{S}}_M^*\mathbf{D}|}.$$

Recalling that we have

$$\overline{\mathbf{B}}_M^*|\mathbf{S} \sim N_{pm}\left(\mathbf{B}, \left(\boldsymbol{\Sigma} + \frac{1}{M}\mathbf{S}\right) \otimes (\mathbf{XX}')^{-1}\right),$$

we obtain

$$\boldsymbol{\Delta}_M^*|\mathbf{S} \sim N_{kr}\left(\boldsymbol{\Delta}, \mathbf{D}'\left(\boldsymbol{\Sigma} + \frac{1}{M}\mathbf{S}\right)\mathbf{D} \otimes \mathbf{A}(\mathbf{XX}')^{-1}\mathbf{A}'\right).$$

Analogous to what was done in Corollary 2.2, we may conclude that

$$(\boldsymbol{\Delta}_M^* - \boldsymbol{\Delta})'[\mathbf{A}(\mathbf{XX}')^{-1}\mathbf{A}']^{-1}(\boldsymbol{\Delta}_M^* - \boldsymbol{\Delta})|\mathbf{S} \sim W_r\left(\mathbf{D}'\left(\boldsymbol{\Sigma} + \frac{1}{M}\mathbf{S}\right)\mathbf{D}, k\right)$$

and hence

$$T_{\mathbf{\Delta}}^{(1)} = \frac{|(\mathbf{\Delta}_M^* - \mathbf{\Delta})'[\mathbf{A}(\mathbf{X}\mathbf{X}')^{-1}\mathbf{A}']^{-1}(\mathbf{\Delta}_M^* - \mathbf{\Delta})|}{|\mathbf{D}'(\mathbf{\Sigma} + \frac{1}{M}\mathbf{S})\mathbf{D}|}|\mathbf{S} \sim \prod_{l=1}^{r} \chi^2_{k-l+1},$$

which will be independent of $\mathbf{S}$.

Let us write $T_{\mathbf{\Delta}}^{(2)}$ as

$$T_{\mathbf{\Delta}}^{(2)} = \frac{|\mathbf{D}'\mathbf{S}\mathbf{D}|}{|M(n-p)\mathbf{D}'\overline{\mathbf{S}}_M^*\mathbf{D}|} \times \frac{|\mathbf{D}'(M\mathbf{\Sigma} + \mathbf{S})\mathbf{D}|}{|\mathbf{D}'\mathbf{S}\mathbf{D}|}.$$

Recalling that $M(n-p)\overline{\mathbf{S}}_M^*|\mathbf{S} \sim W_m(\mathbf{S}, M(n-p))$, we may obtain

$$|M(n-p)\mathbf{D}'\overline{\mathbf{S}}_M^*\mathbf{D}|/|\mathbf{D}'\mathbf{S}\mathbf{D}||\mathbf{S} \sim \prod_{l=1}^{r} \chi^2_{M(n-p)-l+1},$$

which will also be independent of $\mathbf{S}$.

Lastly, it is easy to show by standard arguments that

$$\frac{|\mathbf{D}'(M\mathbf{\Sigma} + \mathbf{S})\mathbf{D}|}{|\mathbf{D}'\mathbf{S}\mathbf{D}|} \sim \frac{|\mathbf{W}^* + M(n-p)\mathbf{I}_r|}{|\mathbf{W}^*|} \sim |\mathbf{I}_r + M(n-p)\mathbf{W}^{*-1}|,$$

where $\mathbf{W}^* \sim W_r(\mathbf{I}_r, n-p)$, which is independent of $\mathbf{S}$.

Combining the above terms, we conclude that

$$T_{M,\mathbf{\Delta}} \overset{st}{\sim} \left\{ \prod_{l=1}^{r} \frac{k-l+1}{M(n-p)-l+1} F_{k,l} \right\} |M(n-p)\mathbf{W}^{*-1} + \mathbf{I}_r|$$

where $F_{k,l} \sim \mathcal{F}_{k-l+1, M(n-p)-l+1}$.

Taking $r = 1$ and $k = 1$, and making $\mathbf{A} : 1 \times p$ a matrix of zeros except for $\mathbf{A}_{1,g} = 1$, and $\mathbf{D} : m \times 1$ a matrix of zeros except for $\mathbf{D}_{h,1} = 1$, for $g = 1, ..., p$ and $h = 1, ..., p$ we may observe that

$$T_{M,\mathbf{\Delta}} = T_{M,\mathbf{B}_{(g,h)}} = \frac{(\overline{\mathbf{B}}^*_{M(g,h)} - \mathbf{B}_{(g,h)})[\mathbf{A}(\mathbf{X}\mathbf{X}')^{-1}\mathbf{A}']^{-1}(\overline{\mathbf{B}}^*_{M(g,h)} - \mathbf{B}_{(g,h)})}{(n-p)\mathbf{D}'\overline{\mathbf{S}}_M^*\mathbf{D}}$$

therefore concluding that the $(1-\alpha)$ confidence interval for $\mathbf{B}_{(g,h)}$ will be given by

$$\overline{\mathbf{B}}^*_{M(g,h)} \pm \sqrt{q^*_{M,1-\alpha}(n-p)\mathbf{D}'\overline{\mathbf{S}}^*_M\mathbf{D}\mathbf{A}(\mathbf{X}\mathbf{X}')^{-1}\mathbf{A}'}$$

where in fact, $\mathbf{D}'\overline{\mathbf{S}}^*_M\mathbf{D} = \overline{\mathbf{S}}^*_{M,(h,h)}$ and $\mathbf{A}(\mathbf{X}\mathbf{X}')^{-1}\mathbf{A}' = (\mathbf{X}\mathbf{X}')^{-1}_{(g,g)}$, with $q^*_{M,1-\alpha}$ being the value of the $1-\alpha$ cut-off point of the distribution of $T_{M,\mathbf{\Delta}}$, for $g = 1,..,p$ and $h = 1,...,m$.

Analogously, if we consider the following statistic

$$T_{comb,\mathbf{\Delta}} = \frac{|(\mathbf{\Delta}^*_M - \mathbf{\Delta})'[\mathbf{A}(\mathbf{X}\mathbf{X}')^{-1}\mathbf{A}']^{-1}(\mathbf{\Delta}^*_M - \mathbf{\Delta})|}{|(n-\frac{p}{M})\mathbf{D}'\mathbf{S}^*_{comb}\mathbf{D}|}$$

associated to second procedure, we have

$$T_{comb,\mathbf{\Delta}} \overset{st}{\sim} \left\{\prod_{l=1}^{r} \frac{k-i+1}{Mn-p-l+1}F_{k,l}\right\} |M(n-p)\mathbf{W}^{*-1} + \mathbf{I}_r|,$$

where $\mathbf{W}^* \sim W_r(\mathbf{I}_r, n-p)$ and $F_{k,l} \sim \mathcal{F}_{k-l+1,Mn-p-l+1}$.

Therefore, we will have the $(1-\alpha)$ confidence interval for $\mathbf{B}_{(g,h)}$ given by, for $T_{comb,\mathbf{\Delta}}$,

$$\mathbf{B}_{(g,h)} = \mathbf{B}^*_{comb(g,h)} \pm \sqrt{q^*_{comb,1-\alpha}\left(n-\frac{p}{M}\right)\mathbf{S}^*_{comb(h,h)}(\mathbf{X}\mathbf{X}')^{-1}_{(g,g)}}.$$

In the next subsections we present the confidence intervals for each of the individual coefficients in $\mathbf{B}$ based on the original data, the single synthetic dataset (M=1) and the multiple synthetic datasets (M=2 and M=5) where min$\mathbf{B}_1$ and min$\mathbf{B}_2$ are the lower bounds, and max$\mathbf{B}_1$ and max$\mathbf{B}_2$ are the upper bounds of the confidence intervals for each of the coefficients in $\mathbf{B}_1$ and $\mathbf{B}_2$, which are the vectors of coefficients respectively for the first and second sensitive variables in Section 4, which are the *total household income* (I) and the *property tax* (PT), relative to the explanatory variables in expression (16) in Section 4, that is, the 28 explanatory variables or covariates (the 3 continuous explanatory variables, and the indicator variables for the other 4 categorical explanatory variables) and the intercept, and where $\tilde{\tilde{\mathbf{B}}}_1$ and $\tilde{\tilde{\mathbf{B}}}_2$ are the estimates of the same coefficients.

For the original data we use the statistic

$$T_{O,\boldsymbol{\Delta}} = \frac{|(\mathbf{A}\hat{\mathbf{B}}\mathbf{D} - \boldsymbol{\Delta})'(\mathbf{A}(\mathbf{X}\mathbf{X}')^{-1}\mathbf{A}')^{-1}(\mathbf{A}\hat{\mathbf{B}}\mathbf{D} - \boldsymbol{\Delta})|}{|(n-p)\mathbf{D}'\mathbf{S}\mathbf{D}|} \overset{st}{\sim} \prod_{l=1}^{r} \frac{k-l+1}{n-p-l+1} F_{k,l},$$

where $F_{k,l} \sim \mathcal{F}_{k-l+1,n-p-l+1}$, in order to compare the results obtained from the original data with the ones obtained from the synthetic data.

S3.1 Individual regression coefficients confidence intervals for the original data

$$\left(\begin{array}{cccc}
\mathrm{min}\mathbf{B}_1 & \mathrm{max}\mathbf{B}_1 & \mathrm{min}\mathbf{B}_2 & \mathrm{max}\mathbf{B}_2 \\
3.0276 & 3.45314 & 7.22404 & 7.81773 \\
0.101763 & 0.128775 & 0.50271 & 0.540397 \\
-0.100943 & -0.0676293 & -0.55975 & -0.513273 \\
0.000592196 & 0.00205233 & -0.0160693 & -0.0140322 \\
-0.140275 & 0.307819 & -0.392277 & 0.232881 \\
-0.181448 & 0.249481 & -0.339044 & 0.262165 \\
-0.0719615 & 0.345724 & -0.146577 & 0.436157 \\
-0.124093 & 0.302649 & -0.0634389 & 0.53193 \\
-0.065135 & 0.356647 & 0.0415413 & 0.629991 \\
-0.0445564 & 0.377282 & 0.00218244 & 0.59071 \\
-0.0126063 & 0.434573 & 0.0703166 & 0.694198 \\
0.155254 & 0.563845 & 0.512783 & 1.08283 \\
0.244623 & 0.654163 & 0.828421 & 1.39979 \\
0.237687 & 0.653551 & 0.812801 & 1.39299 \\
0.365105 & 0.782933 & 1.04715 & 1.63009 \\
0.400888 & 0.810689 & 1.39139 & 1.96312 \\
0.680126 & 1.09312 & 1.7045 & 2.28069 \\
0.686246 & 1.1133 & 2.21108 & 2.80688 \\
0.696918 & 1.12775 & 2.16138 & 2.76245 \\
-0.432626 & 0.0225628 & -0.420386 & 0.214671 \\
-0.152141 & 0.020387 & -0.560916 & -0.320213 \\
-0.0876056 & -0.0161402 & -0.513375 & -0.41367 \\
-0.118923 & -0.057579 & -0.397948 & -0.312365 \\
-0.143437 & 0.0075315 & -0.766689 & -0.556065 \\
-0.0741748 & -0.00465169 & -0.619728 & -0.522733 \\
-0.113329 & -0.0443148 & -0.174004 & -0.0777186 \\
-0.341987 & -0.162744 & -0.514953 & -0.264881 \\
-0.08713 & 0.0295535 & -0.0262251 & 0.136566 \\
-0.00459359 & 0.0347371 & -0.135873 & -0.0810009
\end{array}\right), \left(\begin{array}{cc}
\tilde{\tilde{\mathbf{B}}}_1 & \tilde{\tilde{\mathbf{B}}}_2 \\
3.24037 & 7.52088 \\
0.115269 & 0.521554 \\
-0.0842859 & -0.536512 \\
0.00132226 & -0.0150507 \\
0.0837721 & -0.0796979 \\
0.0340165 & -0.0384396 \\
0.136881 & 0.14479 \\
0.0892781 & 0.234245 \\
0.145756 & 0.335766 \\
0.166363 & 0.296446 \\
0.210983 & 0.382257 \\
0.35955 & 0.797806 \\
0.449393 & 1.11411 \\
0.445619 & 1.1029 \\
0.574019 & 1.33862 \\
0.605789 & 1.67726 \\
0.886624 & 1.9926 \\
0.899771 & 2.50898 \\
0.912332 & 2.46191 \\
-0.205032 & -0.102857 \\
-0.0658772 & -0.440565 \\
-0.0518729 & -0.463522 \\
-0.0882509 & -0.355157 \\
-0.0679526 & -0.661377 \\
-0.0394132 & -0.571231 \\
-0.0788219 & -0.125861 \\
-0.252366 & -0.389917 \\
-0.0287882 & 0.0551704 \\
0.0150718 & -0.108437
\end{array}\right)$$

S3.2 Individual regression coefficients confidence intervals for the single synthetic data

$$\left(\begin{array}{cccc}
\min\mathbf{B}_1 & \max\mathbf{B}_1 & \min\mathbf{B}_2 & \max\mathbf{B}_2 \\
3.02817 & 3.6323 & 7.13692 & 7.98322 \\
0.0972511 & 0.1356 & 0.491859 & 0.545581 \\
-0.106496 & -0.059202 & -0.567982 & -0.50173 \\
-0.000170873 & 0.00190206 & -0.0158549 & -0.012951 \\
-0.386709 & 0.249444 & -0.562915 & 0.328241 \\
-0.339504 & 0.272278 & -0.58278 & 0.274238 \\
-0.242414 & 0.350569 & -0.362125 & 0.468556 \\
-0.249879 & 0.355959 & -0.312773 & 0.535917 \\
-0.179851 & 0.418947 & -0.16971 & 0.669118 \\
-0.229205 & 0.369672 & -0.189196 & 0.649742 \\
-0.109485 & 0.525368 & -0.0587802 & 0.830556 \\
-0.0070404 & 0.57303 & 0.321972 & 1.13457 \\
0.0776444 & 0.659063 & 0.627344 & 1.44183 \\
0.0440082 & 0.634403 & 0.650395 & 1.47745 \\
0.198631 & 0.791815 & 0.923888 & 1.75485 \\
0.231117 & 0.812905 & 1.18828 & 2.00328 \\
0.471254 & 1.05758 & 1.46725 & 2.28861 \\
0.491991 & 1.09827 & 1.94475 & 2.79406 \\
0.507036 & 1.11868 & 2.00075 & 2.85757 \\
-0.614778 & 0.0314465 & -0.642398 & 0.262868 \\
-0.192984 & 0.0519521 & -0.586147 & -0.243027 \\
-0.090293 & 0.0111654 & -0.54331 & -0.401182 \\
-0.114903 & -0.027814 & -0.389401 & -0.267403 \\
-0.141936 & 0.0723915 & -0.788623 & -0.488382 \\
-0.0592764 & 0.0394245 & -0.654749 & -0.516484 \\
-0.139882 & -0.041904 & -0.208622 & -0.0713688 \\
-0.417745 & -0.163276 & -0.574755 & -0.218281 \\
-0.154133 & 0.0115208 & -0.0584952 & 0.173561 \\
-0.0061631 & 0.0496741 & -0.144825 & -0.066605
\end{array}\right),
\left(\begin{array}{cc}
\tilde{\hat{\mathbf{B}}}_1 & \tilde{\hat{\mathbf{B}}}_2 \\
3.33024 & 7.56007 \\
0.116426 & 0.51872 \\
-0.0828492 & -0.534856 \\
0.000865592 & -0.0144029 \\
-0.0686325 & -0.117337 \\
-0.0336128 & -0.154271 \\
0.0540774 & 0.0532158 \\
0.0530399 & 0.111572 \\
0.119548 & 0.249704 \\
0.0702339 & 0.230273 \\
0.207941 & 0.385888 \\
0.282995 & 0.728269 \\
0.368354 & 1.03459 \\
0.339206 & 1.06392 \\
0.495223 & 1.33937 \\
0.522011 & 1.59578 \\
0.764416 & 1.87793 \\
0.795129 & 2.3694 \\
0.812856 & 2.42916 \\
-0.291666 & -0.189765 \\
-0.0705159 & -0.414587 \\
-0.0395638 & -0.472246 \\
-0.0713583 & -0.328402 \\
-0.0347721 & -0.638503 \\
-0.00992597 & -0.585616 \\
-0.0908932 & -0.139996 \\
-0.290511 & -0.396518 \\
-0.0713061 & 0.0575331 \\
0.0217555 & -0.105715
\end{array}\right)$$

S3.3 Individual regression coefficients confidence intervals for the multiple M=2 synthetic data (1st procedure)

$$\left(\begin{array}{cccc}
\min\mathbf{B}_1 & \max\mathbf{B}_1 & \min\mathbf{B}_2 & \max\mathbf{B}_2 \\
2.96109 & 3.48503 & 7.13854 & 7.87099 \\
0.0930352 & 0.126294 & 0.495816 & 0.542311 \\
-0.0994441 & -0.0584271 & -0.560381 & -0.503041 \\
0.000358738 & 0.00215653 & -0.0158025 & -0.0132893 \\
-0.202296 & 0.34942 & -0.489447 & 0.28183 \\
-0.190543 & 0.340038 & -0.455128 & 0.286603 \\
-0.0662091 & 0.448067 & -0.205374 & 0.513563 \\
-0.124584 & 0.400842 & -0.181933 & 0.552591 \\
-0.0390392 & 0.480281 & -0.0560079 & 0.66998 \\
-0.100476 & 0.418913 & -0.0854874 & 0.640597 \\
0.0225152 & 0.573105 & -0.0228145 & 0.746887 \\
0.138073 & 0.641151 & 0.436303 & 1.13959 \\
0.228236 & 0.732483 & 0.742801 & 1.44772 \\
0.205758 & 0.71779 & 0.772304 & 1.4881 \\
0.346188 & 0.860639 & 0.994694 & 1.71388 \\
0.38443 & 0.888998 & 1.30263 & 2.008 \\
0.646802 & 1.1553 & 1.61589 & 2.32676 \\
0.64891 & 1.17472 & 2.08452 & 2.81958 \\
0.671446 & 1.2019 & 2.11805 & 2.8596 \\
-0.552546 & 0.00790641 & -0.710236 & 0.0732532 \\
-0.184376 & 0.0280502 & -0.557329 & -0.260367 \\
-0.116768 & -0.0287761 & -0.541284 & -0.418275 \\
-0.132051 & -0.0565214 & -0.376457 & -0.270869 \\
-0.135489 & 0.0503911 & -0.835584 & -0.575731 \\
-0.0697217 & 0.0158787 & -0.622451 & -0.502785 \\
-0.121066 & -0.0360918 & -0.194518 & -0.0757282 \\
-0.374165 & -0.153471 & -0.544733 & -0.236212 \\
-0.143336 & 0.000330931 & -0.050877 & 0.149963 \\
-0.0027869 & 0.0456391 & -0.142886 & -0.0751888
\end{array}\right),
\left(\begin{array}{cc}
\tilde{\tilde{\mathbf{B}}}_1 & \tilde{\tilde{\mathbf{B}}}_2 \\
3.22306 & 7.50477 \\
0.109665 & 0.519064 \\
-0.0789356 & -0.531711 \\
0.00125763 & -0.0145459 \\
0.0735619 & -0.103808 \\
0.0747474 & -0.0842629 \\
0.190929 & 0.154095 \\
0.138129 & 0.185329 \\
0.220621 & 0.306986 \\
0.159219 & 0.277555 \\
0.29781 & 0.362036 \\
0.389612 & 0.787945 \\
0.48036 & 1.09526 \\
0.461774 & 1.1302 \\
0.603414 & 1.35429 \\
0.636714 & 1.65531 \\
0.901054 & 1.97132 \\
0.911813 & 2.45205 \\
0.936675 & 2.48883 \\
-0.27232 & -0.318491 \\
-0.0781627 & -0.408848 \\
-0.072772 & -0.479779 \\
-0.0942861 & -0.323663 \\
-0.0425487 & -0.705658 \\
-0.0269215 & -0.562618 \\
-0.0785787 & -0.135123 \\
-0.263818 & -0.390472 \\
-0.0715024 & 0.0495431 \\
0.0214261 & -0.109038
\end{array}\right)$$

S3.4 Individual regression coefficients confidence intervals for the multiple M=2 synthetic data (2nd procedure)

$$\left(\begin{array}{cccc}
\mathrm{min}\mathbf{B}_1 & \mathrm{max}\mathbf{B}_1 & \mathrm{min}\mathbf{B}_2 & \mathrm{max}\mathbf{B}_2 \\
2.96171 & 3.48441 & 7.13942 & 7.87011 \\
0.0930745 & 0.126255 & 0.495872 & 0.542255 \\
-0.0993956 & -0.0584755 & -0.560312 & -0.50311 \\
0.000360862 & 0.00215441 & -0.0157995 & -0.0132923 \\
-0.201644 & 0.348768 & -0.488519 & 0.280903 \\
-0.189916 & 0.339411 & -0.454237 & 0.285711 \\
-0.0656015 & 0.447459 & -0.20451 & 0.512699 \\
-0.123963 & 0.400221 & -0.18105 & 0.551708 \\
-0.0384256 & 0.479667 & -0.055135 & 0.669107 \\
-0.099862 & 0.418299 & -0.0846144 & 0.639724 \\
0.0231657 & 0.572454 & -0.0218891 & 0.745962 \\
0.138667 & 0.640557 & 0.437149 & 1.13874 \\
0.228832 & 0.731888 & 0.743649 & 1.44687 \\
0.206363 & 0.717185 & 0.773165 & 1.48724 \\
0.346796 & 0.860032 & 0.995559 & 1.71301 \\
0.385026 & 0.888402 & 1.30348 & 2.00715 \\
0.647403 & 1.1547 & 1.61674 & 2.3259 \\
0.649532 & 1.17409 & 2.08541 & 2.81869 \\
0.672073 & 1.20128 & 2.11894 & 2.85871 \\
-0.551883 & 0.00724418 & -0.709294 & 0.0723112 \\
-0.184125 & 0.0277992 & -0.556972 & -0.260724 \\
-0.116664 & -0.02888 & -0.541136 & -0.418423 \\
-0.131962 & -0.0566106 & -0.37633 & -0.270996 \\
-0.135269 & 0.0501715 & -0.835271 & -0.576044 \\
-0.0696206 & 0.0157776 & -0.622307 & -0.502929 \\
-0.120965 & -0.0361922 & -0.194375 & -0.0758711 \\
-0.373904 & -0.153732 & -0.544362 & -0.236583 \\
-0.143166 & 0.000161176 & -0.0506355 & 0.149722 \\
-0.00272968 & 0.0455819 & -0.142805 & -0.0752702
\end{array}\right),
\left(\begin{array}{cc}
\tilde{\hat{\mathbf{B}}}_1 & \tilde{\hat{\mathbf{B}}}_2 \\
3.22306 & 7.50477 \\
0.109665 & 0.519064 \\
-0.0789356 & -0.531711 \\
0.00125763 & -0.0145459 \\
0.0735619 & -0.103808 \\
0.0747474 & -0.0842629 \\
0.190929 & 0.154095 \\
0.138129 & 0.185329 \\
0.220621 & 0.306986 \\
0.159219 & 0.277555 \\
0.29781 & 0.362036 \\
0.389612 & 0.787945 \\
0.48036 & 1.09526 \\
0.461774 & 1.1302 \\
0.603414 & 1.35429 \\
0.636714 & 1.65531 \\
0.901054 & 1.97132 \\
0.911813 & 2.45205 \\
0.936675 & 2.48883 \\
-0.27232 & -0.318491 \\
-0.0781627 & -0.408848 \\
-0.072772 & -0.479779 \\
-0.0942861 & -0.323663 \\
-0.0425487 & -0.705658 \\
-0.0269215 & -0.562618 \\
-0.0785787 & -0.135123 \\
-0.263818 & -0.390472 \\
-0.0715024 & 0.0495431 \\
0.0214261 & -0.109038
\end{array}\right)$$

S3.5 Individual regression coefficients confidence intervals for the multiple M=5 synthetic data (1st procedure)

$$\left(\begin{array}{cccc}
\min\mathbf{B}_1 & \max\mathbf{B}_1 & \min\mathbf{B}_2 & \max\mathbf{B}_2 \\
2.93058 & 3.39889 & 7.17972 & 7.83333 \\
0.0972081 & 0.126935 & 0.498923 & 0.540413 \\
-0.0968526 & -0.0601912 & -0.559855 & -0.508687 \\
0.000700045 & 0.00230693 & -0.0159042 & -0.0136615 \\
-0.108209 & 0.38492 & -0.423626 & 0.264626 \\
-0.147025 & 0.327213 & -0.381087 & 0.2808 \\
-0.0115419 & 0.448123 & -0.160998 & 0.480549 \\
-0.0618431 & 0.407787 & -0.0978739 & 0.557582 \\
-0.0053142 & 0.458858 & 0.00350566 & 0.651344 \\
-0.0116139 & 0.45262 & -0.0418757 & 0.606048 \\
0.0441611 & 0.536283 & 0.0454834 & 0.73233 \\
0.204335 & 0.653991 & 0.48795 & 1.11553 \\
0.29911 & 0.749811 & 0.803562 & 1.4326 \\
0.286226 & 0.743884 & 0.800312 & 1.43906 \\
0.410814 & 0.870635 & 1.02998 & 1.67174 \\
0.451483 & 0.90247 & 1.3587 & 1.98813 \\
0.724857 & 1.17936 & 1.67296 & 2.3073 \\
0.73176 & 1.20173 & 2.18023 & 2.83616 \\
0.727331 & 1.20146 & 2.13127 & 2.79301 \\
-0.489139 & 0.0117984 & -0.561753 & 0.137396 \\
-0.168175 & 0.0216926 & -0.529872 & -0.264877 \\
-0.0989116 & -0.0202637 & -0.548395 & -0.438628 \\
-0.121477 & -0.0539678 & -0.401783 & -0.307562 \\
-0.145168 & 0.0209728 & -0.824159 & -0.592279 \\
-0.066873 & 0.00963738 & -0.614174 & -0.507389 \\
-0.115578 & -0.0396281 & -0.163004 & -0.0570012 \\
-0.344065 & -0.146807 & -0.568873 & -0.293563 \\
-0.116207 & 0.0122035 & -0.0401409 & 0.13908 \\
-0.00421745 & 0.0390661 & -0.1346 & -0.0741894
\end{array}\right), \left(\begin{array}{cc}
\tilde{\tilde{\mathbf{B}}}_1 & \tilde{\tilde{\mathbf{B}}}_2 \\
3.16474 & 7.50652 \\
0.112072 & 0.519668 \\
-0.0785219 & -0.534271 \\
0.00150349 & -0.0147829 \\
0.138355 & -0.0794999 \\
0.0900943 & -0.0501434 \\
0.21829 & 0.159775 \\
0.172972 & 0.229854 \\
0.226772 & 0.327425 \\
0.220503 & 0.282086 \\
0.290222 & 0.388906 \\
0.429163 & 0.801739 \\
0.524461 & 1.11808 \\
0.515055 & 1.11969 \\
0.640725 & 1.35086 \\
0.676976 & 1.67342 \\
0.952109 & 1.99013 \\
0.966745 & 2.5082 \\
0.964395 & 2.46214 \\
-0.23867 & -0.212178 \\
-0.0732414 & -0.397375 \\
-0.0595876 & -0.493511 \\
-0.0877223 & -0.354672 \\
-0.0620976 & -0.708219 \\
-0.0286178 & -0.560782 \\
-0.0776033 & -0.110003 \\
-0.245436 & -0.431218 \\
-0.0520018 & 0.0494693 \\
0.0174243 & -0.104394
\end{array}\right)$$

S3.6 Individual regression coefficients confidence intervals for the multiple M=5 synthetic data (2nd procedure))

$$
\left(\begin{array}{cccc}
\min\mathbf{B}_1 & \max\mathbf{B}_1 & \min\mathbf{B}_2 & \max\mathbf{B}_2 \\
2.93001 & 3.39946 & 7.17893 & 7.83412 \\
0.0971719 & 0.126972 & 0.498873 & 0.540463 \\
-0.0968972 & -0.0601466 & -0.559917 & -0.508625 \\
0.000698089 & 0.00230888 & -0.0159069 & -0.0136588 \\
-0.108809 & 0.38552 & -0.424461 & 0.265461 \\
-0.147602 & 0.327791 & -0.38189 & 0.281603 \\
-0.0121014 & 0.448682 & -0.161776 & 0.481327 \\
-0.0624147 & 0.408359 & -0.0986693 & 0.558377 \\
-0.00587915 & 0.459423 & 0.00271958 & 0.65213 \\
-0.0121789 & 0.453185 & -0.0426619 & 0.606834 \\
0.0435621 & 0.536882 & 0.0446499 & 0.733163 \\
0.203788 & 0.654538 & 0.487188 & 1.11629 \\
0.298562 & 0.75036 & 0.802799 & 1.43336 \\
0.285669 & 0.744441 & 0.799537 & 1.43983 \\
0.410254 & 0.871195 & 1.0292 & 1.67252 \\
0.450934 & 0.903019 & 1.35793 & 1.9889 \\
0.724303 & 1.17991 & 1.67219 & 2.30807 \\
0.731188 & 1.2023 & 2.17944 & 2.83696 \\
0.726754 & 1.20204 & 2.13047 & 2.79381 \\
-0.489748 & 0.0124081 & -0.562601 & 0.138245 \\
-0.168407 & 0.0219237 & -0.530194 & -0.264555 \\
-0.0990073 & -0.0201679 & -0.548528 & -0.438494 \\
-0.121559 & -0.0538856 & -0.401897 & -0.307447 \\
-0.14537 & 0.021175 & -0.82444 & -0.591997 \\
-0.0669662 & 0.00973051 & -0.614303 & -0.50726 \\
-0.115671 & -0.0395357 & -0.163132 & -0.0568725 \\
-0.344305 & -0.146567 & -0.569207 & -0.293229 \\
-0.116363 & 0.0123598 & -0.0403584 & 0.139297 \\
-0.00427013 & 0.0391188 & -0.134673 & -0.0741161
\end{array}\right),
\left(\begin{array}{cc}
\tilde{\tilde{\mathbf{B}}}_1 & \tilde{\tilde{\mathbf{B}}}_2 \\
3.16474 & 7.50652 \\
0.112072 & 0.519668 \\
-0.0785219 & -0.534271 \\
0.00150349 & -0.0147829 \\
0.138355 & -0.0794999 \\
0.0900943 & -0.0501434 \\
0.21829 & 0.159775 \\
0.172972 & 0.229854 \\
0.226772 & 0.327425 \\
0.220503 & 0.282086 \\
0.290222 & 0.388906 \\
0.429163 & 0.801739 \\
0.524461 & 1.11808 \\
0.515055 & 1.11969 \\
0.640725 & 1.35086 \\
0.676976 & 1.67342 \\
0.952109 & 1.99013 \\
0.966745 & 2.5082 \\
0.964395 & 2.46214 \\
-0.23867 & -0.212178 \\
-0.0732414 & -0.397375 \\
-0.0595876 & -0.493511 \\
-0.0877223 & -0.354672 \\
-0.0620976 & -0.708219 \\
-0.0286178 & -0.560782 \\
-0.0776033 & -0.110003 \\
-0.245436 & -0.431218 \\
-0.0520018 & 0.0494693 \\
0.0174243 & -0.104394
\end{array}\right)
$$