Journal of Privacy and Confidentiality (2015–2016) **7**, Number 1, 43–98

Likelihood Based Finite Sample Inference for Singly Imputed Synthetic Data Under the Multivariate Normal and Multiple Linear Regression Models

Martin Klein[*] and Bimal Sinha[†]

Abstract

In this paper we develop likelihood-based finite sample inference based on singly imputed partially synthetic data, when the original data follow either a multivariate normal or a multiple linear regression model. We assume that the synthetic data are generated by using the plug-in sampling method, where unknown parameters in the data model are set equal to observed values of their point estimators based on the original data, and synthetic data are drawn from this estimated version of the model. Empirical studies are presented to show that the proposed methods do indeed perform as the theory predicts, and to compare the proposed methods for singly imputed synthetic data with the combining rules that are used to analyze multiply imputed partially synthetic data. Some theoretical comparisons between singly and multiply imputed partially synthetic data inference are also provided. A data analysis example and disclosure risk evaluation of singly and multiply imputed partially synthetic data is presented based on public use data from the Current Population Survey. We discuss the specific conditions under which the proposed methodology will yield valid inference, and evaluate the performance of the methodology when certain conditions do not hold. We outline some ways to extend the proposed methodology for certain scenarios where the required set of conditions do not hold.

Keywords: Partially synthetic data, Pivotal quantity, Plug-in sampling, Statistical disclosure control, Unbiased estimator.

[*] Center for Statistical Research and Methodology, U.S. Census Bureau, Washington, DC, USA,mailto:martin.klein@census.gov.

[†] Center for Disclosure Avoidance Research, U.S. Census Bureau, Washington, DC 20233, USA, and Department of Mathematics and Statistics, University of Maryland Baltimore County, Baltimore, MD, USA,mailto:sinha@umbc.edu.

Disclaimer: This article is released to inform interested parties of ongoing research and to encourage discussion of work in progress. The views expressed are those of the authors and not necessarily those of the U.S. Census Bureau.

1 Introduction

Statistical agencies are often faced with two conflicting objectives: (1) collect and publish useful datasets for designing public policies and building scientific theories, and (2) protect confidentiality of survey respondents which is essential to uphold public trust, leading to better response rates and data accuracy. The synthetic data approach aims to satisfy these two objectives, and some statistical agencies now release synthetic data products.

Generally, there are two types of synthetic data discussed in the literature: *fully synthetic data* and *partially synthetic data*, and methodology for drawing inference based on synthetic data has been developed using concepts of multiple imputation (Rubin, 1987). In fully synthetic data methodology, all units in the population not selected in the sample are treated as missing, and are multiply imputed based on the information from sampled units, to create multiple synthetic populations. A sample is then drawn from each synthetic population, and these samples are released to the public. This approach was suggested by Rubin (1993), and methods for drawing inference based on the synthetic data generated using this approach were developed by Raghunathan et al. (2003). In the partially synthetic data approach, the released data is comprised of only the originally sampled units, but any responses deemed to be confidential are replaced by multiple imputations. For any particular variable, the responses could be deemed as confidential for some or all respondents. This approach was suggested by Little (1993), and methods for drawing inference based on synthetic data under this approach were developed by Reiter (2003). We refer to the monograph by Drechsler (2011) for a detailed and general discussion on synthetic data methodology.

There are several examples where partially synthetic data products have been produced based on major data sources. Some examples in the United States include the Survey of Income and Program Participation (Abowd et al., 2006; Benedetto et al., 2013), the American Community Survey Group Quarters data (Hawala, 2008), OnTheMap data displaying where workers live and where they work (Machanavajjhala et al., 2008), and the Longitudinal Business Database (Kinney et al., 2011; Kinney et al., 2014). To obtain valid inference on population quantities using synthetic data, the current practice requires multiple synthetic datasets to be released, but there are cases where it is indeed desirable to release only a single partially synthetic dataset. For example, the Synthetic Longitudinal Business Database, accessible through the VirtualRDC at Cornell University, is a partially synthetic version of the U.S. Census Bureau's Longitudinal Business Database (LBD). As discussed in Kinney et al. (2011) and Kinney et al. (2014), the decision was made to release only a single version of the LBD in the synthetic file, instead of multiple copies, to avoid the perception of high disclosure risk. Similarly, in the application of partially synthetic data to American Community Survey Group Quarters data presented by Hawala (2008), only a single synthetic dataset is released because of the concern that releasing multiple synthetic copies may increase disclosure risk.

The primary purpose of this paper is to develop new likelihood-based procedures for drawing inference based on a singly imputed partially synthetic dataset in some particular scenarios. Moreover, since the synthetic datasets are generated based on the assumption of an underlying probability model for the observed data, it is also natural to explore exact inference procedures based on the likelihood of the released synthetic data. This is precisely what is accomplished in this paper for two useful probability models: multivariate normal and multiple linear regression. In the former model we assume that all observed variables are sensitive in nature and hence require protection. In the latter model, on the other hand, it is assumed that there is only one sensitive response requiring protection, and the response depends on a set of non-sensitive non-stochastic predictors.

We now explain the basic mechanism for generating synthetic data that is assumed throughout this paper. Let $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ be the original confidential data, which are jointly distributed according to the probability density function (pdf) $f_{[\theta]}(\boldsymbol{X})$, where $\boldsymbol{\theta}$ is the unknown (scalar or vector) parameter. To generate partially synthetic data, let $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}(\boldsymbol{X})$ be the observed value of a point estimator of $\boldsymbol{\theta}$, and we plug it into the joint pdf of $\boldsymbol{X}$. The resulting pdf, with the unknown $\boldsymbol{\theta}$ replaced by the observed value $\hat{\boldsymbol{\theta}}(\boldsymbol{X})$ of the point estimator, is denoted by $f_{\hat{\boldsymbol{\theta}}}$. The singly imputed synthetic data, denoted by $\boldsymbol{Y}$, are then generated by drawing $\boldsymbol{Y}$ from the joint pdf $f_{\hat{\boldsymbol{\theta}}}$. Notice that here the synthetic data $\boldsymbol{Y}$ are not generated from posterior predictive sampling under a Bayesian framework as in Reiter (2003), instead they are drawn via plug-in sampling, where we plug a point estimate for $\boldsymbol{\theta}$ into the original model $f_{\boldsymbol{\theta}}(\boldsymbol{X})$, and sample from the resulting distribution. Reiter and Kinney (2012) show that for partially synthetic data, it appears to be unnecessary to sample from a posterior predictive distribution in order to use the inferential procedures of Reiter (2003), and one can instead use plug-in sampling. Indeed this is also the focus of this paper, namely, to concentrate only on the plug-in sampling method for our chosen probability models.

The organization of the paper is as follows. In Section 2 we review the currently available methodology for drawing inference from multiply imputed partially synthetic data. Based on singly imputed synthetic data generated via plug-in sampling, we develop inference for the multivariate normal mean vector and dispersion matrix in Section 3, and in Section 4 we develop inference for the parameters of a multiple linear regression model. Section 5 presents simulation results for assessing the performance of the derived procedures, and comparing their performance with procedures for multiply imputed synthetic data. Subsection 6.1 presents data analysis results under the proposed methods for singly imputed partially synthetic data for a linear regression model in the context of public use data from the 2000 U.S. Current Population Survey, and the results are compared with those obtained under multiply imputed partially synthetic data. Subsection 6.2 presents a disclosure risk evaluation of singly and multiply imputed partially synthetic data in the context of the 2000 U.S. Current Population Survey data. Subsection 7.1 discusses some of the practical conditions under which the proposed methodology will yield valid inference. Subsection 7.2 studies properties of the methodology under some realistic scenarios where the conditions may be violated, and provides some comparisons with multiple imputation based methodology. Subsection 7.3 outlines some ways of extending the proposed methodology. We conclude the paper in Section 8 with a discussion of advantages, disadvantages, and possible extensions of the proposed methods. Proofs of the theorems, and other technical derivations appear in Appendices 1, 2, and 3.

We conclude this section with an observation regarding the existence of *sufficient statistics* in the context of the synthetic data $\boldsymbol{Y}$ generated as above. Suppose based on the original data $\boldsymbol{X}$, $\boldsymbol{T}(\boldsymbol{X})$ is a sufficient statistic for the unknown parameter $\boldsymbol{\theta}$ in the original model $f_{\boldsymbol{\theta}}(\boldsymbol{X})$. Then we can write $f_{\boldsymbol{\theta}}(\boldsymbol{X}) = h(\boldsymbol{X}) g_{\boldsymbol{\theta}}[\boldsymbol{T}(\boldsymbol{X})]$, and the pdf of the synthetic data $\boldsymbol{Y}$ is

$$\int f_{\hat{\boldsymbol{\theta}}(\boldsymbol{X})}(\boldsymbol{Y}) f_{\boldsymbol{\theta}}(\boldsymbol{X}) d\boldsymbol{X} = \int g_{\hat{\boldsymbol{\theta}}(\boldsymbol{X})}[\boldsymbol{T}(\boldsymbol{Y})] h(\boldsymbol{Y}) f_{\boldsymbol{\theta}}(\boldsymbol{X}) d\boldsymbol{X} = h(\boldsymbol{Y}) \int g_{\hat{\boldsymbol{\theta}}(\boldsymbol{X})}[\boldsymbol{T}(\boldsymbol{Y})] f_{\boldsymbol{\theta}}(\boldsymbol{X}) d\boldsymbol{X}. \tag{1}$$

Equation (1) implies the following result, which we use in the sequel.

Lemma 1.1. Suppose that when the original data $\boldsymbol{X}$ are observed, $\boldsymbol{T}(\boldsymbol{X})$ is a sufficient statistic for the unknown parameter $\boldsymbol{\theta}$ in the original model $f_{\boldsymbol{\theta}}(\boldsymbol{X})$. Then when the synthetic data $\boldsymbol{Y}$ are observed, $\boldsymbol{T}(\boldsymbol{Y})$ is a sufficient statistic for $\boldsymbol{\theta}$.

2 Review of Methodology for Multiply Imputed Partially Synthetic Data

In this section we briefly review the state of the art methodology for drawing inference based on multiply imputed partially synthetic data, as developed by Reiter (2003) for a scalar parameter of interest, and extended by Reiter (2005b) for a vector valued parameter of interest. We shall explain these procedures under our specific setting of model based partially synthetic data generated via plug-in sampling. As we discussed in Section 1, the methodology presented in this section was originally developed by Reiter (2003; 2005b) for synthetic data generated by posterior predictive sampling, but Reiter and Kinney (2012) indicate that the procedures are still valid when synthetic data are generated via plug-in sampling.

As in Section 1, let $\boldsymbol{X}$ be the originally observed confidential data, jointly distributed according to the pdf $f_{\boldsymbol{\theta}}(\boldsymbol{X})$, where $\boldsymbol{\theta}$ is the unknown parameter. Let $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}(\boldsymbol{X})$ be the observed value of a point estimator of $\boldsymbol{\theta}$. Then $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_m$, $m > 1$ partially synthetic datasets, are obtained by drawing each $\boldsymbol{Y}_j$ independently and identically, conditional on $\boldsymbol{X}$, such that

$$\boldsymbol{Y}_j \sim f_{\hat{\boldsymbol{\theta}}}, \text{ for } j = 1, \ldots, m. \tag{2}$$

Inference for a Scalar Valued Parameter. We now explain the methodology of Reiter (2003), which is used to draw inference on $Q = Q(\boldsymbol{\theta})$, a scalar parameter of interest. Let $q = q(\boldsymbol{X})$ be an estimator of Q based on the original data $\boldsymbol{X}$, and let $u = u(\boldsymbol{X})$ be an estimator of the variance of q, also computed on the original data $\boldsymbol{X}$. Let $q_j = q(\boldsymbol{Y}_j)$ and $u_j = u(\boldsymbol{Y}_j)$, be the values of q and u, respectively, when computed on the jth synthetic dataset $\boldsymbol{Y}_j$. The following quantities are used to draw inference on

Q:

$$\bar{q}_m = \frac{1}{m}\sum_{j=1}^{m} q_j, \qquad b_m = \frac{1}{m-1}\sum_{j=1}^{m}(q_j - \bar{q}_m)^2, \qquad \bar{u}_m = \frac{1}{m}\sum_{j=1}^{m} u_j.$$

Then $\bar{q}_m$ is an estimate of Q and the variance of $\bar{q}_m$ is estimated by $T_m = b_m/m + \bar{u}_m$. The distribution of $(\bar{q}_m - Q)/\sqrt{T_m}$ is approximated by a t_v distribution where $v = (m-1)\left[1 + \frac{\bar{u}_m}{(b_m/m)}\right]^2$. Thus the quantity $(\bar{q}_m - Q)/\sqrt{T_m}$, with its approximating t distribution, can be used to obtain tests of significance for Q, and a $(1-\gamma)$ confidence interval for Q.

Inference for a Vector Valued Parameter. The methodology of Reiter (2005b), which is used to draw inference on $\boldsymbol{Q} = \boldsymbol{Q}(\boldsymbol{\theta})$, a $k \times 1$ dimensional parameter of interest, can now be explained. Let $\boldsymbol{q} = \boldsymbol{q}(\boldsymbol{X})$ be a $k \times 1$ dimensional estimator of $\boldsymbol{Q}$ based on the original data $\boldsymbol{X}$, and let $\boldsymbol{u} = \boldsymbol{u}(\boldsymbol{X})$ be a $k \times k$ dimensional estimator of the covariance matrix of $\boldsymbol{q}$, also computed on the original data $\boldsymbol{X}$. Let $\boldsymbol{q}_j = \boldsymbol{q}(\boldsymbol{Y}_j)$ and $\boldsymbol{u}_j = \boldsymbol{u}(\boldsymbol{Y}_j)$, be the values of $\boldsymbol{q}$ and $\boldsymbol{u}$, respectively, when computed on the jth synthetic dataset $\boldsymbol{Y}_j$. The following quantities are used to draw inference on $\boldsymbol{Q}$:

$$\bar{\boldsymbol{q}}_m = \frac{1}{m}\sum_{j=1}^{m} \boldsymbol{q}_j, \qquad \boldsymbol{b}_m = \frac{1}{m-1}\sum_{j=1}^{m}(\boldsymbol{q}_j - \bar{\boldsymbol{q}}_m)(\boldsymbol{q}_j - \bar{\boldsymbol{q}}_m)', \qquad \bar{\boldsymbol{u}}_m = \frac{1}{m}\sum_{j=1}^{m} \boldsymbol{u}_j.$$

Then $\bar{\boldsymbol{q}}_m$ is an estimate of $\boldsymbol{Q}$ and the covariance matrix of $\bar{\boldsymbol{q}}_m$ is estimated by $\boldsymbol{T}_m = \boldsymbol{b}_m/m + \bar{\boldsymbol{u}}_m$. Define the quantity

$$S_m = \frac{(\bar{\boldsymbol{q}}_m - \boldsymbol{Q})'(\bar{\boldsymbol{u}}_m)^{-1}(\bar{\boldsymbol{q}}_m - \boldsymbol{Q})}{k(1+r)}$$

where $r = m^{-1}\mathrm{tr}(\boldsymbol{b}_m\bar{\boldsymbol{u}}_m^{-1})/k$. The distribution of S_m is approximated by an $F_{k,w(r)}$ distribution where

$$w(r) = 4 + (t-4)\left[1 + \frac{(1-\frac{2}{t})}{r}\right]^2$$

and $t = k(m-1)$. Thus the quantity S_m, with its approximating F distribution, can be used to obtain tests of significance for $\boldsymbol{Q}$, and a $(1-\gamma)$ confidence ellipsoid for $\boldsymbol{Q}$. Alternative methods of inference based on log-likelihood ratio test statistics from m synthetic datasets are also developed by Reiter (2005b).

3 Methodology Under a Multivariate Normal Distribution

In this section we present the likelihood-based approach for analysis of singly imputed synthetic data generated from a multivariate normal population with both mean vector and dispersion matrix unknown under the plug-in sampling method. Assume the original confidential data are

$$\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) \sim \text{ independent and identically distributed } (iid) \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \quad (3)$$

where $n > p$, and define $\hat{\boldsymbol{\mu}} = \bar{\boldsymbol{x}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i$ (sample mean) and $\hat{\boldsymbol{\Sigma}} = \mathscr{S}_x/(n-1)$ where $\mathscr{S}_x = \boldsymbol{W} = \sum_{i=1}^{n}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})'$ is the sample Wishart matrix. Obviously, $(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}})$ are jointly sufficient for $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ when the original data are observed.

The singly imputed synthetic data, denoted by $\boldsymbol{Y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)$, are obtained by drawing

$$\boldsymbol{Y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n) \,\big|\, \boldsymbol{X} \sim iid \sim N_p\left(\bar{\boldsymbol{x}},\ \frac{\mathscr{S}_x}{n-1}\right). \quad (4)$$

Define $\bar{\boldsymbol{y}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{y}_i$ (sample mean based on $\boldsymbol{Y}$) and $\mathscr{S}_y = \sum_{i=1}^{n}(\boldsymbol{y}_i - \bar{\boldsymbol{y}})(\boldsymbol{y}_i - \bar{\boldsymbol{y}})'$ (sample Wishart matrix based on $\boldsymbol{Y}$). It follows from Lemma 1.1 that $\bar{\boldsymbol{y}}$ and $\mathscr{S}_y$ are jointly sufficient for $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. The following *fundamental* theorem, whose proof appears in Appendix 1, can be used to derive the inferential results presented in this section.

Theorem 3.1. The distribution of $T^2 = n(\bar{\boldsymbol{y}} - \boldsymbol{\mu})'\mathscr{S}_y^{-1}(\bar{\boldsymbol{y}} - \boldsymbol{\mu})$ has the representation: $T^2 = T_1 \times T_2$ where $T_1 \sim \frac{1}{\chi^2_{n-p}}$, independent of T_2, and the conditional distribution of T_2, given a Wishart matrix $\boldsymbol{W}^*$, is $\sum_{i=1}^{p} \lambda_i \chi^2_{1i}$ where χ^2_{1i} are independent χ^2 variables each with 1 degree of freedom and $\lambda_1, \ldots, \lambda_p$ are the roots of $|(n-1)\boldsymbol{I}_p + (1-\lambda)\boldsymbol{W}^*| = 0$ and $\boldsymbol{W}^* \sim \text{Wishart}_p(\boldsymbol{I}_p, n-1)$.

Below are the main *inferential* results related to $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ in this case. Appendix 2.1 contains some further details about the derivation of these above results.

Result 3.1. The maximum likelihood estimator (MLE) of $\boldsymbol{\mu}$ is $\bar{\boldsymbol{y}}$, which is unbiased for $\boldsymbol{\mu}$, with $\text{Var}(\bar{\boldsymbol{y}}) = 2\boldsymbol{\Sigma}/n$.

Result 3.2. An unbiased estimator of $\boldsymbol{\Sigma}$ is $\mathscr{S}_y/(n-1)$.

Result 3.3. Define $T^2 = n(\bar{\boldsymbol{y}} - \boldsymbol{\mu})'\mathscr{S}_y^{-1}(\bar{\boldsymbol{y}} - \boldsymbol{\mu})$. Theorem 3.1 shows that T^2 is a pivotal quantity, and therefore a $(1-\gamma)$ confidence ellipsoid $\boldsymbol{\mu}$ based on T^2 is given by

$$\Delta_{\text{MVN}}(\boldsymbol{\mu}) = \left\{\boldsymbol{\mu} : n(\boldsymbol{\mu} - \bar{\boldsymbol{y}})'\mathscr{S}_y^{-1}(\boldsymbol{\mu} - \bar{\boldsymbol{y}}) \le c_{n,p,\gamma}\right\}, \quad (5)$$

where $c_{n,p,\gamma}$ is the $(1-\gamma)$percentile from the distribution of T^2. From Theorem 3.1, it follows that the cut-off point $c_{n,p,\gamma}$ can be obtained by simulating the distribution of T^2 as follows:

1. Generate $\lambda_1, \ldots, \lambda_p$, the roots of $|(n-1)\boldsymbol{I}_p + (1-\lambda)\boldsymbol{W}^*| = 0$ where $\boldsymbol{W}^* \sim \text{Wishart}_p(\boldsymbol{I}_p, n-1)$.

2. Generate $T_2 {=} \sum_{i=1}^{p} \lambda_i \chi^2_{1i}$ where χ^2_{1i} are independent χ^2 variables each with 1 degree of freedom.

3. Generate $T_1 \sim \frac{1}{\chi^2_{n-p}}$, independent of T_2.

4. Finally, compute $T^2 = T_1 \times T_2$.

The *volume* of the confidence ellipsoid $\Delta_{\mathrm{MVN}}(\boldsymbol{\mu})$ is given by

$$V_{\boldsymbol{\mu}}(\boldsymbol{Y}) = \frac{\pi^{p/2}}{n^{p/2}\Gamma\left(\frac{p}{2}+1\right)}(c_{n,p,\gamma})^{p/2}|\mathscr{S}_y|^{\frac{1}{2}}.$$

Since $E\left(|\mathscr{S}_y|^{1/2}\right) = \frac{\mathscr{C}^2_{n,p}}{(n-1)^{p/2}}|\boldsymbol{\Sigma}|^{\frac{1}{2}}$ with $\mathscr{C}_{n,p} = \prod_{i=1}^{p}\left[2^{1/2}\Gamma\left(\frac{n-i+1}{2}\right)\Big/\Gamma\left(\frac{n-i}{2}\right)\right]$, the expected volume is obtained as

$$E\left[V_{\boldsymbol{\mu}}(\boldsymbol{Y})\right] = \frac{\pi^{p/2}}{n^{p/2}\Gamma\left(\frac{p}{2}+1\right)}(c_{n,p,\gamma})^{p/2}\frac{\mathscr{C}^2_{n,p}}{(n-1)^{p/2}}|\boldsymbol{\Sigma}|^{\frac{1}{2}}. \tag{6}$$

Remark 3.1. Suppose that $m \geq 1$ synthetic datasets are generated by repeating the sampling in (4) a total of m times, independently. Let $(\boldsymbol{y}_{1j}, \ldots, \boldsymbol{y}_{nj})$ denote the jth synthetic dataset for $j = 1, \ldots, m$. Applying the combination formulas of Reiter (2003; 2005b), as outlined in Section 2, one would use $\hat{\boldsymbol{\mu}}_m = \sum_{j=1}^{m} \bar{\boldsymbol{y}}_j/m$ to estimate $\boldsymbol{\mu}$, where $\bar{\boldsymbol{y}}_j = \sum_{i=1}^{n} \boldsymbol{y}_{ij}/n$. The estimate $\hat{\boldsymbol{\mu}}_m$ is unbiased for $\boldsymbol{\mu}$ with $\mathrm{Var}(\hat{\boldsymbol{\mu}}_m) = \mathrm{Var}[E(\sum_{j=1}^{m} \bar{\boldsymbol{y}}_j/m|\bar{\boldsymbol{x}}, \mathscr{S}_x)] + E[\mathrm{Var}(\sum_{j=1}^{m} \bar{\boldsymbol{y}}_j/m|\bar{\boldsymbol{x}}, \mathscr{S}_x)]$. Since $E(\bar{\boldsymbol{y}}_j|\bar{\boldsymbol{x}}, \mathscr{S}_x) = \bar{\boldsymbol{x}}$, $j = 1, \ldots, m$, the first term is $\boldsymbol{\Sigma}/n$. Since, conditionally given $(\bar{\boldsymbol{x}}, \mathscr{S}_x)$, $\bar{\boldsymbol{y}}_j$ are *iid* with conditional covariance matrix equal to $\mathscr{S}_{\boldsymbol{x}}/n(n-1)$, and $E(\mathscr{S}_{\boldsymbol{x}}/(n-1)) = \boldsymbol{\Sigma}$, the second term is $\boldsymbol{\Sigma}/(mn)$, resulting in the final expression $\mathrm{Var}(\hat{\boldsymbol{\mu}}_m) = (1+\frac{1}{m})\frac{\boldsymbol{\Sigma}}{n}$. Obviously, when $m = 1$ we get $\mathrm{Var}(\hat{\boldsymbol{\mu}}_m) = 2\boldsymbol{\Sigma}/n$, which agrees with Result 3.1, and in general, the expression $\mathrm{Var}(\hat{\boldsymbol{\mu}}_m) = (1+\frac{1}{m})\frac{\boldsymbol{\Sigma}}{n}$ shows how the variance of $\hat{\boldsymbol{\mu}}_m$ decreases as the number of imputations m increases.

Remark 3.2. Continuing with the scenario of Remark 3.1, suppose now that $m > 1$ synthetic datasets are generated. Applying the methodology of Reiter (2003; 2005b) outlined in Section 2, one would estimate the covariance matrix of the multiple imputation estimator $\hat{\boldsymbol{\mu}}_m$ by $\boldsymbol{T}_m = \boldsymbol{b}_m/m + \bar{\boldsymbol{u}}_m$, where $\boldsymbol{b}_m = \frac{1}{m-1}\sum_{j=1}^{m}(\boldsymbol{y}_j - \hat{\boldsymbol{\mu}}_m)(\boldsymbol{y}_j - \hat{\boldsymbol{\mu}}_m)'$, $\bar{\boldsymbol{u}}_m = \sum_{j=1}^{m} \boldsymbol{u}_j/m$, $\boldsymbol{u}_j = \mathscr{S}_{yj}/(n(n-1))$, and $\mathscr{S}_{yj} = \sum_{i=1}^{n}(\boldsymbol{y}_{ij} - \bar{\boldsymbol{y}}_j)(\boldsymbol{y}_{ij} - \bar{\boldsymbol{y}}_j)'$. We will show that $E(\bar{\boldsymbol{u}}_m) = E(\boldsymbol{b}_m) = \boldsymbol{\Sigma}/n$, and hence $E(\boldsymbol{T}_m) = (1+\frac{1}{m})\frac{\boldsymbol{\Sigma}}{n}$. In other words, under multiple synthetic data, $\boldsymbol{T}_m$ provides an unbiased estimate of $\mathrm{Var}(\hat{\boldsymbol{\mu}}_m)$. Since $\boldsymbol{u}_j = \mathscr{S}_{yj}/(n(n-1))$, for $j = 1, \ldots, m$, we get $E(\bar{\boldsymbol{u}}_m) = \boldsymbol{\Sigma}/n$. We next show that $E(\boldsymbol{b}_m) = \boldsymbol{\Sigma}/n$, thus proving the assertion. Without any loss of generality, we can assume $\boldsymbol{\mu} = \boldsymbol{0}$. Since $E(\bar{\boldsymbol{y}}_j\bar{\boldsymbol{y}}_j') = E(\bar{\boldsymbol{x}}\bar{\boldsymbol{x}}' + \frac{\mathscr{S}_x}{n(n-1)}) = \frac{2\boldsymbol{\Sigma}}{n}$ for each j, and $E(\hat{\boldsymbol{\mu}}_m\hat{\boldsymbol{\mu}}_m') = E(\bar{\boldsymbol{x}}\bar{\boldsymbol{x}}' + \frac{\mathscr{S}_{\boldsymbol{x}}}{mn(n-1)}) = (\frac{1+m}{m})\frac{\boldsymbol{\Sigma}}{n}$, we readily get $E[(m-1)\boldsymbol{b}_m] = [2m\frac{\boldsymbol{\Sigma}}{n} - (m+1)\frac{\boldsymbol{\Sigma}}{n}] = (m-1)\boldsymbol{\Sigma}/n$. Hence the result.

4 Methodology Under a Multiple Linear Regression Model

In this section we turn to the case of a standard *multiple linear regression* model involving a *sensitive response* variable y and a $p \times 1$ dimensional vector of *non-sensitive predictors* $\boldsymbol{x}$. We assume that

$$y_1, \ldots, y_n \text{ are independent such that } y_i \sim N(\boldsymbol{x}_i'\boldsymbol{\beta}, \sigma^2), \tag{7}$$

where $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ are fixed, and $\boldsymbol{\beta}$ and σ^2 are both unknown. Thus the original data consist of $\{(y_i, \boldsymbol{x}_i) : i = 1, \ldots, n\}$. We define $\boldsymbol{y} = (y_1, \ldots, y_n)'$ as the $n \times 1$ dimensional vector of response variables, and $\boldsymbol{X} = [\boldsymbol{x}_1 \ \cdots \ \boldsymbol{x}_n]$ as the $p \times n$ dimensional matrix of predictor variables, and we assume that $\text{rank}(\boldsymbol{X}) = p < n$. Based on the original data, $\boldsymbol{b} = (\boldsymbol{X}\boldsymbol{X}')^{-1}\boldsymbol{X}\boldsymbol{y}$ is the MLE and uniformly minimum variance unbiased estimator (UMVUE) of $\boldsymbol{\beta}$, and $\hat{\sigma}^2 = \text{RSS}/(n-p)$ is the UMVUE of σ^2 where $\text{RSS} = (\boldsymbol{y} - \boldsymbol{X}'\boldsymbol{b})'(\boldsymbol{y} - \boldsymbol{X}'\boldsymbol{b})$. Furthermore, $\boldsymbol{b}$ and RSS are independently distributed such that $\boldsymbol{b} \sim N_p[\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}\boldsymbol{X}')^{-1}]$ and $\text{RSS} \sim \sigma^2\chi^2_{n-p}$. When the original data are observed, $\boldsymbol{b}$ and RSS are jointly sufficient for $\boldsymbol{\beta}$ and σ^2.

The singly imputed synthetic data in this case consist of a single synthetic version of $\boldsymbol{y} = (y_1, \ldots, y_n)'$, which is denoted as $\boldsymbol{v} = (v_1, \ldots, v_n)'$, and obtained by drawing

$$v_1, \ldots, v_n \text{ independently such that } v_i \sim N\left(\boldsymbol{x}_i'\boldsymbol{b},\, \frac{\text{RSS}}{n-p}\right). \tag{8}$$

Thus the released data will be of the form $\{(v_i, \boldsymbol{x}_i) : i = 1, \ldots, n\}$, and our goal is to discuss inference on $\boldsymbol{\beta}$ and σ^2 based on this released data. Towards this end, analogous to $\boldsymbol{b}$ and RSS, we define $\boldsymbol{b}^* = (\boldsymbol{X}\boldsymbol{X}')^{-1}\boldsymbol{X}\boldsymbol{v}$ and $\text{RSS}^* = (\boldsymbol{v} - \boldsymbol{X}'\boldsymbol{b}^*)'(\boldsymbol{v} - \boldsymbol{X}'\boldsymbol{b}^*)$, which, by Lemma 1.1, are jointly sufficient for $(\boldsymbol{\beta}, \sigma^2)$ when $\{(v_i, \boldsymbol{x}_i) : i = 1, \ldots, n\}$ are observed.

The inferential results presented in this section can be derived based on the following three *fundamental* theorems whose proofs are deferred to Appendix 1.

Theorem 4.1. The joint pdf of $(\boldsymbol{b}^*, \text{RSS}^*)$ is given by

$$f_{\boldsymbol{\beta},\sigma^2}(\boldsymbol{b}^*, \text{RSS}^*)$$
$$\propto \int_0^\infty e^{-\frac{1}{2}\left[\frac{(\boldsymbol{b}^*-\boldsymbol{\beta})'(\boldsymbol{X}\boldsymbol{X}')(\boldsymbol{b}^*-\boldsymbol{\beta})}{\sigma^2(1+\frac{\psi}{n-p})} + \frac{(n-p)\text{RSS}^*}{\sigma^2\psi} + \psi\right]} \times \frac{(\text{RSS}^*)^{\frac{n-p}{2}-1}}{(\sigma^2)^{\frac{n-p}{2}}} \frac{(\psi)^{-\frac{p+2}{2}}}{\sigma^p} \left[1 + \frac{n-p}{\psi}\right]^{-p/2} d\psi.$$

Theorem 4.2. The pdf of $V = \dfrac{\text{RSS}^*}{\sigma^2}$ is given by

$$f_{n,p}(v) = K_{n,p} \int_0^\infty e^{-\frac{1}{2}[\frac{(n-p)v}{\psi} + \psi]} v^{\frac{n-p}{2}-1} \psi^{-1} d\psi$$

where $[K_{n,p}]^{-1} = [\Gamma(\frac{n-p}{2})]^2[2^{n-p}(n-p)^{\frac{n-p}{2}}]$.

Theorem 4.3. The distribution of $T^2 = (\boldsymbol{b}^* - \boldsymbol{\beta})'(\boldsymbol{X}\boldsymbol{X}')(\boldsymbol{b}^* - \boldsymbol{\beta})/\text{RSS}^*$ can be represented as follows:

$$T^2|\psi \sim \left[\frac{p}{n-p}\right]\left[1 + \frac{n-p}{\psi}\right] F_{p,n-p} \quad \text{and} \quad \psi \sim \chi^2_{n-p}.$$

Here are the main *inferential* results related to $\boldsymbol{\beta}$ and σ^2 in this case. Appendix 2.2 contains some further details about the derivations of these results.

Result 4.1. The MLE of $\boldsymbol{\beta}$ is $\boldsymbol{b}^* = (\boldsymbol{X}\boldsymbol{X}')^{-1}\boldsymbol{X}\boldsymbol{y}$, which is unbiased for $\boldsymbol{\beta}$, with $\mathrm{Var}(\boldsymbol{b}^*) = 2\sigma^2(\boldsymbol{X}\boldsymbol{X}')^{-1}$.

Result 4.2. An unbiased estimator of σ^2 is $\mathrm{RSS}^*/(n-p)$, and therefore an unbiased estimator of $\mathrm{Var}(\boldsymbol{b}^*)$ is $\widehat{\mathrm{Var}}(\boldsymbol{b}^*) = 2\left(\frac{\mathrm{RSS}^*}{n-p}\right)(\boldsymbol{X}\boldsymbol{X}')^{-1}$.

Result 4.3. The MLE of σ^2 is $(n-p)\mathrm{RSS}^*/\Delta_{\max}$ where $\Delta_{\max}$ is the value of Δ that maximizes

$$\Delta^{n/2} \times \int_0^\infty e^{-\frac{1}{2}\left[\psi + \frac{\Delta}{\psi}\right]} \psi^{-\frac{p+2}{2}} \left(1 + \frac{n-p}{\psi}\right)^{-p/2} d\psi.$$

Result 4.4. A $(1-\gamma)$ level confidence interval for σ^2 based on $V = \mathrm{RSS}^*/\sigma^2$ is

$$\left[\frac{\mathrm{RSS}^*}{b_{n,p;\gamma}}, \frac{\mathrm{RSS}^*}{a_{n,p;\gamma}}\right], \tag{9}$$

where $a_{n,p;\gamma}$ and $b_{n,p;\gamma}$ are any two constants that satisfy $1-\gamma = \Pr(a_{n,p;\gamma} \leq V \leq b_{n,p;\gamma})$. The equal-tail confidence interval is obtained by taking $a_{n,p;\gamma}$ and $b_{n,p;\gamma}$ as the $\gamma/2$ and $1-\gamma/2$ quantiles, respectively, of V, which can be readily computed using Monte Carlo simulation. The pdf of V is given by Theorem 4.2, and to simulate from this distribution, one can simply draw $\psi \sim \chi^2_{n-p}$ and $V|\psi \sim \frac{\psi\chi^2_{n-p}}{n-p}$.

The shortest length $(1-\gamma)$ level confidence interval for σ^2 based on V on the other hand can be obtained by choosing $a_{n,p;\gamma}$ and $b_{n,p;\gamma}$ to satisfy

$$\int_{a_{n,p;\gamma}}^{b_{n,p;\gamma}} f_{n,p}(v)dv = 1-\gamma \quad \text{and} \quad a^2_{n,p;\gamma} f_{n,p}(a_{n,p;\gamma}) = b^2_{n,p;\gamma} f_{n,p}(b_{n,p;\gamma}),$$

where $f_{n,p}(v)$ the pdf of V (given in Theorem 4.2). A method to compute the constants $a_{n,p;\gamma}$ and $b_{n,p;\gamma}$ is explained in Remark 4.1 below.

The expected length of the above confidence interval for σ^2 is equal to

$$(n-p)\sigma^2 \left[\frac{1}{a_{n,p;\gamma}} - \frac{1}{b_{n,p;\gamma}}\right]. \tag{10}$$

Result 4.5. Define $T^2 = (\boldsymbol{b}^* - \boldsymbol{\beta})'(\boldsymbol{X}\boldsymbol{X}')(\boldsymbol{b}^* - \boldsymbol{\beta})/\mathrm{RSS}^*$. Theorem 4.3 shows that T^2 is a pivotal quantity, and therefore a confidence ellipsoid for $\boldsymbol{\beta}$ based on T^2 is given by

$$\Delta_{\mathrm{MLR}}(\boldsymbol{\beta}) = \left\{\boldsymbol{\beta} : T^2 \leq d_{n,p;\gamma}\right\} \tag{11}$$

where $d_{n,p;\gamma}$ satisfies $1-\gamma = \Pr(T^2 \leq d_{n,p;\gamma})$. The cut-off point $d_{n,p;\gamma}$ can be readily obtained by simulating the distribution of T^2 as follows.

1. Generate $\psi \sim \chi^2_{n-p}$.

2. Generate $T^2|\psi \sim \left[\frac{p}{n-p}\right]\left[1 + \frac{n-p}{\psi}\right] F_{p,n-p}$.

The volume of the confidence ellipsoid $\Delta_{\mathrm{MLR}}(\boldsymbol{\beta})$ is given by

$$V_{\boldsymbol{\beta}}(\boldsymbol{v}, \boldsymbol{X}) = \frac{\pi^{p/2}}{\Gamma\left(\frac{p}{2}+1\right)} (d_{n,p;\gamma})^{p/2} \left|\boldsymbol{X}\boldsymbol{X}'\right|^{-1/2} (\mathrm{RSS}^*)^{p/2},$$

and since $E\left[(\mathrm{RSS}^*)^{p/2}\right] = \sigma^p (n-p)^{-p/2} \left\{E\left[\left(\chi^2_{n-p}\right)^{p/2}\right]\right\}^2$, the expected volume is

$$E\left[V_{\boldsymbol{\beta}}(\boldsymbol{v}, \boldsymbol{X})\right] = \frac{\pi^{p/2}}{\Gamma\left(\frac{p}{2}+1\right)} (d_{n,p;\gamma})^{p/2} \left|\boldsymbol{X}\boldsymbol{X}'\right|^{-1/2} \sigma^p \frac{\left\{E\left[\left(\chi^2_{n-p}\right)^{p/2}\right]\right\}^2}{(n-p)^{p/2}}, \tag{12}$$

where $E\left[\left(\chi^2_{n-p}\right)^{p/2}\right] = 2^{p/2}\Gamma\left(\frac{n}{2}\right)/\Gamma\left(\frac{n-p}{2}\right)$.

Result 4.6. If one is interested in the significance of a single regression coefficient or more generally in the significance of a linear combination of $\boldsymbol{\beta}$, namely, $\boldsymbol{A}\boldsymbol{\beta} = \boldsymbol{\eta}$ where $\boldsymbol{A}$ is a $k \times p$ dimensional matrix with $\mathrm{rank}(\boldsymbol{A}) = k < p$, we define $T^2_{\boldsymbol{\eta}} = (\boldsymbol{A}\boldsymbol{b}^* - \boldsymbol{\eta})' \left\{\boldsymbol{A}(\boldsymbol{X}\boldsymbol{X}')^{-1}\boldsymbol{A}'\right\}^{-1} (\boldsymbol{A}\boldsymbol{b}^* - \boldsymbol{\eta})/\mathrm{RSS}^*$, and proceed by noting that

$$T^2_{\boldsymbol{\eta}}|\psi \sim \left[\frac{k}{n-p}\right]\left[1+\frac{n-p}{\psi}\right] F_{k,n-p} \quad \text{and} \quad \psi \sim \chi^2_{n-p}.$$

Test for the significance of $\boldsymbol{\eta}$. For testing $H_0 : \boldsymbol{\eta} = \boldsymbol{\eta}_0$ versus $H_1 : \boldsymbol{\eta} \neq \boldsymbol{\eta}_0$ at level γ, we reject H_0 whenever $T^2_{\boldsymbol{\eta}_0}$ exceeds $\delta_{k,n,p;\gamma}$ where $\delta_{k,n,p;\gamma}$ satisfies $1-\gamma = \Pr(T^2_{\boldsymbol{\eta}_0} \leq \delta_{k,n,p;\gamma})$ when H_0 is true.

Confidence ellipsoid for $\boldsymbol{\eta}$. A $(1-\gamma)$ level confidence ellipsoid for $\boldsymbol{\eta}$ is given by

$$\Delta_{\mathrm{MLR}}(\boldsymbol{\eta}) = \left\{\boldsymbol{\eta} : T^2_{\boldsymbol{\eta}} \leq \delta_{k,n,p;\gamma}\right\}. \tag{13}$$

The constant $\delta_{k,n,p;\gamma}$ above is obtained by simulating the distribution of $T^2_{\boldsymbol{\eta}}$ directly from the above representation, namely, by (i) generating $\psi \sim \chi^2_{n-p}$, and then (ii) generating $T^2_{\boldsymbol{\eta}}|\psi \sim \left[\frac{k}{n-p}\right] F_{k,n-p} \left[1+\frac{n-p}{\psi}\right]$. The volume of the confidence ellipsoid $\Delta_{\mathrm{MLR}}(\boldsymbol{\eta})$ is

$$V_{\boldsymbol{\eta}}(\boldsymbol{v}, \boldsymbol{X}) = \frac{\pi^{k/2}}{\Gamma\left(\frac{k}{2}+1\right)} (\delta_{k,n,p;\gamma})^{k/2} \left|\boldsymbol{A}(\boldsymbol{X}\boldsymbol{X}')^{-1}\boldsymbol{A}'\right|^{1/2} (\mathrm{RSS}^*)^{k/2},$$

and since $E\left[(\mathrm{RSS}^*)^{k/2}\right] = \sigma^k (n-p)^{-k/2} \left\{E\left[\left(\chi^2_{n-p}\right)^{k/2}\right]\right\}^2$, the expected volume is

$$E\left[V_{\boldsymbol{\eta}}(\boldsymbol{v}, \boldsymbol{X})\right] = \frac{\pi^{k/2}}{\Gamma\left(\frac{k}{2}+1\right)} (\delta_{k,n,p;\gamma})^{k/2} \left|\boldsymbol{A}(\boldsymbol{X}\boldsymbol{X}')^{-1}\boldsymbol{A}'\right|^{1/2} \sigma^k \frac{\left\{E\left[\left(\chi^2_{n-p}\right)^{k/2}\right]\right\}^2}{(n-p)^{k/2}},$$

where $E\left[\left(\chi^2_{n-p}\right)^{k/2}\right] = 2^{k/2}\Gamma\left(\frac{n+k-p}{2}\right)/\Gamma\left(\frac{n-p}{2}\right)$.

Inference for a single regression coefficient. Writing $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$, inference for a single regression coefficient β_i is readily obtained by taking $\boldsymbol{A}$ as the $1 \times p$ dimensional vector having a 1 in column i, and 0 in all other columns, in which case $\boldsymbol{A\beta} = \beta_i$. For a given value β_{i0}, one can test $H_0 : \beta_i = \beta_{i0}$ versus $H_1 : \beta_i \neq \beta_{i0}$ at level γ by rejecting H_0 whenever $(b_i^* - \beta_{i0})^2/(D_{ii} \times \mathrm{RSS}^*) > \delta_{1,n,p;\gamma}$, where D_{ii} is the (i,i)th entry of the matrix $(\boldsymbol{XX}')^{-1}$, and b_i^* is the ith entry of the vector $\boldsymbol{b}^*$. Furthermore, a $(1-\gamma)$ level confidence interval for β_i is given by

$$\left[b_i^* - \sqrt{D_{ii} \times \mathrm{RSS}^*}\sqrt{\delta_{1,n,p;\gamma}}\,,\ \ b_i^* + \sqrt{D_{ii} \times \mathrm{RSS}^*}\sqrt{\delta_{1,n,p;\gamma}} \right]. \tag{14}$$

The length of the above confidence interval for β_i is

$$V_{\beta_i}(\boldsymbol{v}, \boldsymbol{X}) = 2\sqrt{D_{ii} \times \mathrm{RSS}^*}\sqrt{\delta_{1,n,p;\gamma}},$$

and the expected length is

$$E\left[V_{\beta_i}(\boldsymbol{v}, \boldsymbol{X})\right] = 2(D_{ii})^{1/2}(\delta_{1,n,p;\gamma})^{1/2}\frac{\sigma}{(n-p)^{1/2}}\left[\frac{2^{1/2}\Gamma\left(\frac{n-p+1}{2}\right)}{\Gamma\left(\frac{n-p}{2}\right)}\right]^2. \tag{15}$$

Remark 4.1. To compute the constants $a_{n,p;\gamma}$ and $b_{n,p;\gamma}$ that give the shortest version of the $1-\gamma$ level confidence interval in (9), note that the pdf of V in Theorem 4.2 is a continuous mixture of the form $f_{n,p}(v) = \int_0^\infty f(\psi)f(v|\psi)d\psi$, where $f(\psi)$ is the pdf of $\psi \sim \chi^2_{n-p}$ and $f(v|\psi)$ is the conditional pdf of $V|\psi \sim \frac{\psi\chi^2_{n-p}}{n-p}$. Recall that we need $a^2_{n,p;\gamma}f_{n,p}(a_{n,p;\gamma}) - b^2_{n,p;\gamma}f_{n,p}(b_{n,p;\gamma}) = 0$ and $\Pr(a_{n,p;\gamma} \leq V \leq b_{n,p;\gamma}) = 1-\gamma$. Since

$$a^2_{n,p;\gamma}f_{n,p}(a_{n,p;\gamma}) - b^2_{n,p;\gamma}f_{n,p}(b_{n,p;\gamma}) = \int_0^\infty \left\{a^2_{n,p;\gamma}f(a_{n,p;\gamma}|\psi) - b^2_{n,p;\gamma}f(b_{n,p;\gamma}|\psi)\right\} f(\psi)d\psi,$$

it follows that a Monte Carlo estimator of $a^2_{n,p;\gamma}f_{n,p}(a_{n,p;\gamma}) - b^2_{n,p;\gamma}f_{n,p}(b_{n,p;\gamma})$ is

$$\frac{1}{m}\sum_{j=1}^m \left\{a^2_{n,p;\gamma}f(a_{n,p;\gamma}|\psi_j) - b^2_{n,p;\gamma}f(b_{n,p;\gamma}|\psi_j)\right\}, \tag{16}$$

where $\psi_j \sim iid \sim \chi^2_{n-p}$. To compute $\Pr(a_{n,p;\gamma} \leq V \leq b_{n,p;\gamma})$, note that

$$\begin{aligned}\Pr(a_{n,p;\gamma} \leq V \leq b_{n,p;\gamma}) &= E\left[\Pr\left\{a_{n,p;\gamma} \leq \frac{\psi\chi^2_{n-p}}{n-p} \leq b_{n,p;\gamma}\,\Big|\,\psi\right\}\right] \\ &= E\left[F_{\chi^2_{n-p}}\left(\frac{(n-p)b_{n,p;\gamma}}{\psi}\right) - F_{\chi^2_{n-p}}\left(\frac{(n-p)a_{n,p;\gamma}}{\psi}\right)\right],\end{aligned}$$

where $F_{\chi^2_{n-p}}(\cdot)$ is the cumulative distribution function (cdf) of the χ^2_{n-p} distribution. Hence we get a Monte Carlo estimator of $\Pr(a_{n,p;\gamma} \leq V \leq b_{n,p;\gamma})$ as

$$\frac{1}{m}\sum_{j=1}^m \left[F_{\chi^2_{n-p}}\left(\frac{(n-p)b_{n,p;\gamma}}{\psi_j}\right) - F_{\chi^2_{n-p}}\left(\frac{(n-p)a_{n,p;\gamma}}{\psi_j}\right)\right], \tag{17}$$

which can be viewed as a Rao-Blackwellized version of the simple Monte Carlo estimator

$$\frac{1}{m}\sum_{j=1}^{m} I(a_{n,p;\gamma} \leq V_j \leq b_{n,p;\gamma}),$$

where $V_j \sim iid \sim f_{n,p}(v)$. We can therefore compute $a_{n,p;\gamma}$ and $b_{n,p;\gamma}$ as follows:

(I) Fix a value $a_{n,p;\gamma}$ and solve for $b_{n,p;\gamma}$ satisfying $a^2_{n,p;\gamma} f_{n,p;\gamma}(a_{n,p;\gamma}) - b^2_{n,p;\gamma} f_{n,p}(b_{n,p;\gamma}) = 0$ using the Monte Carlo estimator.

(II) Evaluate $\Pr(a_{n,p;\gamma} \leq V \leq b_{n,p;\gamma})$ using the Rao-Blackwellized Monte Carlo estimator given displayed in (17).

(III) Repeat steps (I) and (II) over a grid of values for $a_{n,p;\gamma}$, choose the values of $a_{n,p;\gamma}$ and $b_{n,p;\gamma}$ that yield $\Pr(a_{n,p;\gamma} \leq V \leq b_{n,p;\gamma})$ closest to $(1-\gamma)$.

Remark 4.2. Suppose that $m \geq 1$ synthetic datasets are generated by repeating the sampling in (8) a total of m times, independently. Let $\{(v_{ij}, \boldsymbol{x}_i) : i = 1, \ldots, n\}$ denote the jth synthetic dataset for $j = 1, \ldots, m$. Applying the methodology of Reiter (2003; 2005b) outlined in Section 2, one would use $\hat{\boldsymbol{\beta}}_m = \sum_{j=1}^{m} \boldsymbol{b}_j^*/m$ to estimate $\boldsymbol{\beta}$, where $\boldsymbol{b}_j^* = (\boldsymbol{X}\boldsymbol{X}')^{-1}\boldsymbol{X}\boldsymbol{v}_j$, $\boldsymbol{v}_j = (v_{1j}, \ldots, v_{nj})'$. The estimate $\hat{\boldsymbol{\beta}}_m$ is unbiased for $\boldsymbol{\beta}$ with $\text{Var}(\hat{\boldsymbol{\beta}}_m) = \text{Var}[E(\sum_{j=1}^{m} \boldsymbol{b}_j^*/m|\boldsymbol{b}, \text{RSS})] + E[\text{Var}(\sum_{j=1}^{m} \boldsymbol{b}_j^*/m|\boldsymbol{b}, \text{RSS})]$. Since $E(\boldsymbol{b}_j^*|\boldsymbol{b}, \text{RSS}) = \boldsymbol{b}$, $j = 1, \cdots, m$, the first term is $\text{Var}(\boldsymbol{b}) = \sigma^2(\boldsymbol{X}\boldsymbol{X}')^{-1}$. Since, conditionally given $(\boldsymbol{b}, \text{RSS})$, $\boldsymbol{b}_j^*$ are iid with conditional variance equal to $(\text{RSS}/(n-p))(\boldsymbol{X}\boldsymbol{X}')^{-1}$, and $E(\text{RSS}/(n-p)) = \sigma^2$, the second term is $\sigma^2(\boldsymbol{X}\boldsymbol{X}')^{-1}/m$, resulting in the final expression $\text{Var}(\hat{\boldsymbol{\beta}}_m) = (1+\frac{1}{m})\sigma^2(\boldsymbol{X}\boldsymbol{X}')^{-1}$. Obviously, when $m = 1$ we get $\text{Var}(\hat{\boldsymbol{\beta}}_m) = 2\sigma^2(\boldsymbol{X}\boldsymbol{X}')^{-1}$, which agrees with Result 4.1, and in general, the expression $\text{Var}(\hat{\boldsymbol{\beta}}_m) = (1+\frac{1}{m})\sigma^2(\boldsymbol{X}\boldsymbol{X}')^{-1}$ shows how the variance of $\hat{\boldsymbol{\beta}}_m$ decreases as the number of imputations m increases.

Remark 4.3. Continuing with the scenario of Remark 4.2, suppose now that $m > 1$ synthetic datasets are released. Applying the methodology of Reiter (2003; 2005b) outlined in Section 2, one would estimate the covariance matrix of the multiple imputation estimator $\hat{\boldsymbol{\beta}}_m$ by $\boldsymbol{T}_m = \boldsymbol{b}_m/m + \bar{\boldsymbol{u}}_m$ where $\boldsymbol{b}_m = \frac{1}{m-1}\sum_{j=1}^{m}(\boldsymbol{b}_j^* - \hat{\boldsymbol{\beta}}_m)(\boldsymbol{b}_j^* - \hat{\boldsymbol{\beta}}_m)'$, $\bar{\boldsymbol{u}}_m = \sum_{j=1}^{m} \boldsymbol{u}_j/m$, $\boldsymbol{u}_j = (\text{RSS}_j^*/(n-p))(\boldsymbol{X}\boldsymbol{X}')^{-1}$, and $\text{RSS}_j^* = (\boldsymbol{v}_j - \boldsymbol{X}'\boldsymbol{b}_j^*)'(\boldsymbol{v}_j - \boldsymbol{X}'\boldsymbol{b}_j^*)$. We will show that $E(\bar{\boldsymbol{u}}_m) = E(\boldsymbol{b}_m) = \sigma^2(\boldsymbol{X}\boldsymbol{X}')^{-1}$, and hence $E(\boldsymbol{T}_m) = (1+\frac{1}{m})\sigma^2(\boldsymbol{X}\boldsymbol{X}')^{-1}$. In other words, under multiple synthetic data, $\boldsymbol{T}_m$ provides an unbiased estimate of $\text{Var}(\hat{\boldsymbol{\beta}}_m)$. Since $\boldsymbol{u}_j = (\text{RSS}_j^*/(n-p))(\boldsymbol{X}\boldsymbol{X}')^{-1}$, for $j = 1, \ldots, m$, we get $E(\bar{\boldsymbol{u}}_m) = \sigma^2(\boldsymbol{X}\boldsymbol{X}')^{-1}$. We next show that $E(\boldsymbol{b}_m) = \sigma^2(\boldsymbol{X}\boldsymbol{X}')^{-1}$, thus proving the assertion. Without any loss of generality, we can assume $\boldsymbol{\beta} = \boldsymbol{0}$. Since $E[\boldsymbol{b}_j^*(\boldsymbol{b}_j^*)'] = E[\boldsymbol{b}\boldsymbol{b}' + (\text{RSS}/(n-p))(\boldsymbol{X}\boldsymbol{X}')^{-1}] = 2\sigma^2(\boldsymbol{X}\boldsymbol{X}')^{-1}$ for each j, and $E[\hat{\boldsymbol{\beta}}_m\hat{\boldsymbol{\beta}}_m'] = E[\boldsymbol{b}\boldsymbol{b}' + (\boldsymbol{X}\boldsymbol{X}')^{-1}\text{RSS}/(m(n-p))] = \sigma^2[\frac{1+m}{m}](\boldsymbol{X}\boldsymbol{X}')^{-1}$, we readily get $E[(m-1)\boldsymbol{b}_m] = 2m\sigma^2(\boldsymbol{X}\boldsymbol{X}')^{-1} - (m+1)\sigma^2(\boldsymbol{X}\boldsymbol{X}')^{-1} = (m-1)\sigma^2(\boldsymbol{X}\boldsymbol{X}')^{-1}$. Hence the result.

5 Simulation Studies

In this section we report simulation results. The primary purposes of these simulation studies are (1) to demonstrate that the inferential methods developed in Sections 3 and 4 perform as our theory predicts; (2) to compare the accuracy of inference of our proposed methodology for singly imputed partially synthetic data with accuracy of inference of the methodology of Reiter (2003; 2005b) for multiply imputed partially synthetic data, in both cases using plug-in sampling to generate synthetic data; and (3) to compare the accuracy of inference of our proposed methodology for singly imputed partially synthetic data generated via plug-in sampling with accuracy of inference of standard methods applied on the original data. All simulation results were obtained using the statistical computing software R (R Development Core Team, 2013).

5.1 Multivariate Normal

In this subsection we present a simulation study designed to evaluate the performance of the methodology developed in Section 3, and thus we work under the notations of Section 3. To conduct the simulation, the population distribution is taken to be the multivariate normal model (3) with

$$p = 10, \;\; \boldsymbol{\mu} = 0.1 \times \begin{pmatrix} 1 & 2 & \dots & 10 \end{pmatrix}', \;\; \boldsymbol{\Sigma} = 0.25\boldsymbol{I}_p + 0.75\boldsymbol{J}_p, \tag{18}$$

where $\boldsymbol{I}_p$ is the $p \times p$ dimensional identity matrix and $\boldsymbol{J}_p$ is the $p \times p$ matrix of 1's. Based on Monte Carlo simulation with 10^6 iterations, we compute an estimate of the coverage probability (*avg cvg*) and an estimate of the expected volume (*avg vol*) for the following confidence ellipsoids for $\boldsymbol{\mu}$, where in all cases, the nominal level of the confidence ellipsoid is set at 0.95.

(a) The confidence ellipsoid for $\boldsymbol{\mu}$ given by (5), as developed in Section 3, based on a single synthetic dataset that is generated as in (4). The results are displayed in Table 1, under the headings *Synthetic Data* and $m = 1$. In general, m refers to the number of synthetic datasets.

(b) The confidence ellipsoid for $\boldsymbol{\mu}$ presented in Section 2, based on $m > 1$ synthetic datasets that are obtained by repeating the generation in (4) a total of $m > 1$ times. When applying the methods of Section 2 to get a confidence ellipsoid for $\boldsymbol{\mu}$, the vector valued parameter of interest is obviously $\boldsymbol{Q} = \boldsymbol{\mu}$, and we take $\boldsymbol{q} = \bar{\boldsymbol{x}}$ and $\boldsymbol{u} = \mathscr{S}_x/[n(n-1)]$. These results are displayed in Table 1 for the cases $m = 5$ and $m = 10$.

The simulation results are reported in Table 1 for the cases when the sample size n equals 1000, 2000, and 4000. While the methodology presented here for analyzing singly imputed synthetic data is valid even for small sample sizes, we have only considered larger sample sizes in the simulations because larger sample sizes are more realistic in applications. Table 1 also displays (under the headings *Synthetic Data*, $m = 1$ and *exp vol*) the numerical value of the theoretical expected volume of (5), which is computed using formula (6). Furthermore, for comparison sake, the table displays the expected volume (under the headings *Original Data* and *exp vol*) of the confidence ellipsoid for $\boldsymbol{\mu}$ that is obtained from the original data using the well known result (Hotelling, 1931; Anderson, 2003; Muirhead, 2005):

$$\frac{n-p}{p(n-1)}\left[n(\bar{\boldsymbol{x}}-\boldsymbol{\mu})'\left(\frac{\mathscr{S}_x}{n-1}\right)^{-1}(\bar{\boldsymbol{x}}-\boldsymbol{\mu})\right]\sim F_{p,n-p}. \tag{19}$$

The following is a summary of the main findings of the simulation study.

1. The results in Table 1, under the headings *Synthetic Data* and $m = 1$, show that the *avg cvg* of the nominal 0.95 confidence ellipsoid (5) is approximately equal to 0.95, for all values of n considered. Furthermore, *exp vol*, the numerical value of the theoretical expected volume (6), is approximately equal to *avg vol*, the Monte Carlo estimate of the expected volume. Thus the confidence ellipsoid performs exactly as predicted by the theory developed in Section 3.

2. In the cases of $m = 5$ or $m = 10$ multiply imputed synthetic datasets, we see in Table 1 that *avg cvg* is approximately equal to the nominal level for the sample sizes considered. This observation is in agreement with Reiter and Kinney (2012), who found that the inferential methodology for multiply imputed partially synthetic data, which was derived by Reiter (2003; 2005b) under posterior predictive sampling, remains valid under plug-in sampling.

3. In Table 1, when comparing *avg vol* based on singly imputed synthetic data ($m = 1$) with *avg vol* based on multiply imputed synthetic data ($m = 5$ and $m = 10$), we find that *avg vol* under multiply imputed synthetic data is considerably less than *avg vol* under singly imputed synthetic data. Thus for larger sample sizes, multiply imputed synthetic data tend to yield smaller confidence regions as compared with singly imputed synthetic data. We also observe that the mean volume of the confidence ellipsoid based on synthetic datasets decreases when m increases, and gets closer to the mean volume of the confidence ellipsoid based on the original data. Recall that in Remark 3.1 we showed that if $m \geq 1$ synthetic datasets are released, then $\hat{\boldsymbol{\mu}}_m$, the estimator of $\boldsymbol{\mu}$ based on the synthetic data, has $\mathrm{Var}(\hat{\boldsymbol{\mu}}_m) = (1+\frac{1}{m})\frac{\boldsymbol{\Sigma}}{n}$. Thus for $m = 1$, 5, and 10, the variance of $\hat{\boldsymbol{\mu}}_m$ equals $2\boldsymbol{\Sigma}/n$, $1.2\boldsymbol{\Sigma}/n$, and $1.1\boldsymbol{\Sigma}/n$, respectively. These expressions give a clear indication of why the expected volume of the confidence region can be considerably less under multiple imputation in comparison with single imputation. Furthermore, when m is large, the variance of $\hat{\boldsymbol{\mu}}_m$ is approximately equal to $\frac{\boldsymbol{\Sigma}}{n}$, which is the variance of the sample mean based on the original data.

5.2 Linear Regression

In this subsection we present a simulation study designed to evaluate the performance of the methodology developed in Section 4, and thus we work under the notations of

Table 1: Inference for the multivariate normal mean vector $\boldsymbol{\mu}$ when $p = 10$.

	Original Data	*Synthetic Data*						
		$m = 1$			$m = 5$		$m = 10$	
	exp	*avg*	*avg*	*exp*	*avg*	*avg*	*avg*	*avg*
n	*vol*	*cvg*	*vol*	*vol*	*cvg*	*vol*	*cvg*	*vol*
1000	2.986E-11	0.950	9.688E-10	9.688E-10	0.946	7.165E-11	0.946	4.489E-11
2000	9.117E-13	0.949	2.900E-11	2.900E-11	0.948	2.270E-12	0.948	1.422E-12
4000	2.817E-14	0.951	9.062E-13	9.062E-13	0.949	7.142E-14	0.949	4.475E-14

Section 4. To conduct the simulation, the population distribution is taken to be the linear regression model (7) with

$$p = 10, \qquad \boldsymbol{x}_i = \begin{pmatrix} 1 \\ x_{1i} \\ x_{2i} \\ x_{3i} \\ x_{4i} \\ I(x_{5i} = 2) \\ I(x_{5i} = 3) \\ I(x_{5i} = 4) \\ I(x_{5i} = 5) \\ I(x_{5i} = 6) \end{pmatrix}, \qquad \boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \\ \beta_6 \\ \beta_7 \\ \beta_8 \\ \beta_9 \\ \beta_{10} \end{pmatrix} = \begin{pmatrix} 10 \\ 2 \\ 2 \\ -3 \\ -1 \\ -2 \\ 1 \\ 2 \\ 2 \\ 4 \end{pmatrix}, \qquad \sigma^2 = 1. \tag{20}$$

The regressor variables in $\boldsymbol{x}_i$ are generated one time at the beginning of the simulation, and then held fixed from one iteration to the next. We generate the regressor variables (all independently) as follows:

$$x_{1i} \sim N(1, 1), \qquad \log x_{2i} \sim N(0, 1), \qquad x_{3i} \sim \text{Exponential}(\text{mean} = 1),$$

$$x_{4i} \sim \text{Poisson}(1), \qquad x_{5i} = \begin{cases} 1 \text{ with probability } 0.2 \\ 2 \text{ with probability } 0.1 \\ 3 \text{ with probability } 0.2 \\ 4 \text{ with probability } 0.2 \\ 5 \text{ with probability } 0.2 \\ 6 \text{ with probability } 0.1 \end{cases}$$

Based on Monte Carlo simulation with 10^6 iterations, we compute an estimate of the coverage probability and we compute an estimate of the expected volume or length (as appropriate) of the following confidence regions, where in all cases, the nominal level of the confidence region is set at 0.95.

(a) The confidence ellipsoid for $\boldsymbol{\beta}$ given by (11), the confidence interval for β_2 given by

(14), and the shortest length confidence interval for σ^2 given by (9). Each of these confidence regions is based on a single synthetic dataset that is generated as in (8). The estimated coverage probability (*avg cvg*) and estimated expected volume (*avg vol*) of the confidence ellipsoid for $\boldsymbol{\beta}$ are shown in Table 2; the estimated coverage probability (*avg cvg*) and estimated expected length (*avg len*) of the confidence interval for β_2 are shown in Table 3; and the estimated coverage probability (*avg cvg*) and estimated expected length (*avg len*) of the confidence interval for σ^2 are shown in Table 4. Because we have just a single synthetic dataset, these results are shown in each of the Tables 2 - 4 under the heading $m = 1$.

(b) The confidence ellipsoid for $\boldsymbol{\beta}$, confidence interval for β_2, and confidence interval for σ^2 obtained using the methodology of Section 2 for $m > 1$ synthetic datasets. The m synthetic datasets are obtained by repeating the generation in (8) a total of m times. In applying the methods of Section 2 to get a confidence ellipsoid for $\boldsymbol{\beta}$, we take $\boldsymbol{Q} = \boldsymbol{\beta}$, $\boldsymbol{q} = \boldsymbol{b}$, $\boldsymbol{u} = \left(\frac{\text{RSS}}{n-p}\right)(\boldsymbol{X}\boldsymbol{X}')^{-1}$; in applying these methods to get a confidence interval for β_2, we take $Q = \beta_2$, $q = b_2$, $u = \left(\frac{\text{RSS}}{n-p}\right) D_{22}$; and in applying these methods to get a confidence interval for σ^2, we take $Q = \sigma^2$, $q = \frac{\text{RSS}}{n-p}$, $u = 2\left(\frac{\text{RSS}}{n-p}\right)^2/(n-p)$. For each of the cases $m = 5$ and $m = 10$, the estimated coverage probability (*avg cvg*) and estimated expected volume (*avg vol*) of the confidence ellipsoid for $\boldsymbol{\beta}$ are shown in Table 2; the estimated coverage probability (*avg cvg*) and estimated expected length (*avg len*) of the confidence interval for β_2 are shown in Table 3; and the estimated coverage probability (*avg cvg*) and estimated expected length (*avg len*) of the confidence interval for σ^2 are shown in Table 4.

Table 2: Inference for the vector of regression parameters $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_{10})'$

	Original Data	*Synthetic Data*						
		$m = 1$			$m = 5$		$m = 10$	
	exp	*avg*	*avg*	*exp*	*avg*	*avg*	*avg*	*avg*
n	*vol*	*cvg*	*vol*	*vol*	*cvg*	*vol*	*cvg*	*vol*
1000	6.506E-07	0.950	2.132E-05	2.132E-05	0.949	1.643E-06	0.948	1.027E-06
2000	1.607E-08	0.951	5.243E-07	5.242E-07	0.950	4.101E-08	0.949	2.567E-08
4000	5.894E-10	0.950	1.910E-08	1.909E-08	0.950	1.514E-09	0.949	9.478E-10

The simulation results are reported in Tables 2, 3, and 4, for the cases when the sample size n equals 1000, 2000, and 4000. In addition, Table 2 displays the numerical value of the theoretical expected volume of (11), computed using formula (12); Table 3 displays the numerical value of the theoretical expected length of (14), computed using formula (15); and Table 4 displays the numerical value of the theoretical expected length of the shortest length confidence interval (9), computed using formula (10). For the sake of comparison, Table 2 displays the expected volume (under the headings *Original*

Table 3: Inference for the scalar regression parameter β_2.

	Original Data	*Synthetic Data*						
		$m = 1$			$m = 5$		$m = 10$	
	exp	*avg*	*avg*	*exp*	*avg*	*avg*	*avg*	*avg*
n	*len*	*cvg*	*len*	*len*	*cvg*	*len*	*cvg*	*len*
1000	0.125	0.950	0.177	0.177	0.950	0.138	0.949	0.131
2000	0.086	0.950	0.122	0.122	0.950	0.095	0.950	0.090
4000	0.062	0.950	0.087	0.087	0.950	0.068	0.950	0.065

Table 4: Inference for the residual variance σ^2.

	Original Data	*Synthetic Data*						
		$m = 1$			$m = 5$		$m = 10$	
	exp	*avg*	*avg*	*exp*	*avg*	*avg*	*avg*	*avg*
n	*len*	*cvg*	*len*	*len*	*cvg*	*len*	*cvg*	*len*
1000	0.177	0.948	0.248	0.248	0.949	0.195	0.949	0.185
2000	0.124	0.952	0.177	0.177	0.950	0.137	0.950	0.131
4000	0.088	0.949	0.124	0.124	0.950	0.097	0.950	0.092

Data and *exp vol*) of the confidence ellipsoid for $\boldsymbol{\mu}$ obtained from the original data; Table 3 displays the expected length (under the headings *Original Data* and *exp len*) of the confidence interval for β_2 obtained from the original data; and Table 4 displays the expected length (under the headings *Original Data* and *exp len*) of the confidence interval for σ^2 obtained from the original data. These original data confidence regions are obtained from the following standard results (Rencher and Schaalje, 2008):

$$\frac{(\boldsymbol{b}-\boldsymbol{\beta})'(\boldsymbol{X}\boldsymbol{X}')(\boldsymbol{b}-\boldsymbol{\beta})}{\left(\frac{RSS}{n-p}\right)p} \sim F_{p,n-p}, \qquad \frac{b_i-\beta_i}{\sqrt{D_{ii}\left(\frac{\mathrm{RSS}}{n-p}\right)}} \sim t_{n-p}, \qquad \frac{\mathrm{RSS}}{\sigma^2} \sim \chi^2_{n-p}. \tag{21}$$

In order to get a fair comparison, for the confidence interval for σ^2 based on the original data, we use the shortest length 0.95 confidence interval based on the pivot $\frac{\mathrm{RSS}}{\sigma^2}$ (as opposed to, for example, the equal-tail interval). The shortest length 0.95 confidence interval based on this pivot is, of course, $\left[\frac{\mathrm{RSS}}{b}, \frac{\mathrm{RSS}}{a}\right]$, where a and b satisfy $F_{\chi^2_{n-p}}(b) - F_{\chi^2_{n-p}}(a) = 0.95$ and $b^2 f_{\chi^2_{n-p}}(b) = a^2 f_{\chi^2_{n-p}}(a)$, and $F_{\chi^2_{n-p}}(\cdot)$ and $f_{\chi^2_{n-p}}(\cdot)$ denote the cdf and pdf, respectively, of the χ^2_{n-p} distribution (Casella and Berger, 2001).

The findings of this simulation study are essentially analogous to those reported in Section 5.1 for the multivariate normal model, and are stated below.

1. The results in Tables 2, 3, and 4, under the headings *Synthetic Data* and $m = 1$, show that, based on singly imputed synthetic data, the 0.95 confidence ellipsoid for

$\boldsymbol{\beta}$, 0.95 confidence interval for β_2, and the 0.95 confidence interval for σ^2, each have *avg cvg* approximately equal to 0.95. Furthermore, each of these confidence regions has *exp vol* or *exp len* approximately equal to *avg vol* or *avg len*. Thus the simulation confirms that the confidence regions (11), (14), and (9), perform as predicted by the theory in Section 4.

2. In the cases of $m = 5$ or $m = 10$ multiply imputed synthetic datasets, we see in Tables 2, 3, and 4 that *avg cvg* is approximately equal to the nominal level. Thus, as in Section 5.1, this observation is in agreement with Reiter and Kinney (2012), who found that the inferential methodology for multiply imputed partially synthetic data, which was derived by Reiter (2003; 2005b) under posterior predictive sampling, remains valid under plug-in sampling.

3. Comparing *avg vol* or *avg len* based on singly imputed synthetic data ($m = 1$) with *avg vol* or *avg len* based on multiply imputed synthetic data ($m = 5$ and $m = 10$) in Tables 2, 3, and 4, we find that multiply imputed synthetic data tends to yield smaller confidence sets as compared with singly imputed synthetic data. We also see that the mean volume or length of the confidence regions based on synthetic datasets decreases when m increases, and gets closer to the mean volume or length of the confidence regions based on the original data. Recall that in Remark 4.2 we showed that if $m \geq 1$ synthetic datasets are released, then $\hat{\boldsymbol{\beta}}_m$, the estimator of $\boldsymbol{\beta}$ based on the synthetic data, has $\mathrm{Var}(\hat{\boldsymbol{\beta}}_m) = (1 + \frac{1}{m})\sigma^2(\boldsymbol{X}\boldsymbol{X}')^{-1}$. Thus for $m = 1$, 5, and 10, the variance of $\hat{\boldsymbol{\beta}}_m$ equals $2\sigma^2(\boldsymbol{X}\boldsymbol{X}')^{-1}$, $1.2\sigma^2(\boldsymbol{X}\boldsymbol{X}')^{-1}$, and $1.1\sigma^2(\boldsymbol{X}\boldsymbol{X}')^{-1}$, respectively. These expressions give a clear indication of why the expected volume/length of the confidence sets can be considerably less under multiple imputation in comparison with single imputation. Furthermore, when m is large, the variance of $\hat{\boldsymbol{\beta}}_m$ is approximately equal to $\sigma^2(\boldsymbol{X}\boldsymbol{X}')^{-1}$, which is the variance of the least squares estimate of $\boldsymbol{\beta}$ based on the original data.

Remark 5.1. We conclude this section with the following remark. Suppose that a data analyst, in possession of a singly imputed synthetic dataset, were to simply analyze the synthetic data as if it were the original data. Such an analysis would, in general, obviously lead to invalid inference. To show what can happen, we conducted two simulation studies, each based on 10^6 iterations. In the first simulation study, we generated multivariate normal data under (18), and computed a Monte Carlo estimate of the coverage probability of the confidence ellipsoid that is obtained from (19), when the confidence ellipsoid is naively computed based on a singly imputed synthetic dataset, instead of the original data. In the second simulation study, we generated linear regression data under (20), and computed a Monte Carlo estimate of the coverage probability of each of the confidence sets that are obtained from (21), when each confidence set is naively computed based on a singly imputed synthetic dataset, instead of on the original data. In all cases, the nominal confidence level is set at 0.95. The results, which are displayed in Table 5, indicate that when original data confidence sets formulae are naively applied to singly imputed synthetic data, the resulting confidence sets have coverage probability well below the nominal level.

Table 5: Coverage of original data 0.95 confidence sets when applied to synthetic data.

	Parameter of Interest			
n	$\boldsymbol{\mu}$	$\boldsymbol{\beta}$	β_2	σ^2
1000	0.482	0.485	0.834	0.835
2000	0.482	0.484	0.835	0.834
4000	0.482	0.483	0.834	0.835

6 Empirical Evaluations Using Current Population Survey Data

In this section we present a real data application using public use data from the 2000 Current Population Survey (CPS) March Supplement. These data are available online from `http://www.census.gov.cps/`. In Subsection 6.1 we present inference on regression parameters obtained by applying methodology of Section 4 to analyze a singly imputed partially synthetic dataset, and we compare with the inference obtained by applying the methodology of Reiter (2003) to analyze multiply imputed partially synthetic data. In Subsection 6.2 we compare the disclosure risk of singly imputed partially synthetic data with that of multiply imputed partially synthetic data in the context of this CPS data example. These CPS data were previously used by Drechsler and Reiter (2010) and Reiter (2005a;c) for illustrating aspects of synthetic data methodology, and by Klein et al. (2014) for illustrating methodology of noise multiplication for statistical disclosure control. While the entire data file contains household, family, and individual records, we focus only on the household records, as did Drechsler and Reiter (2010), Reiter (2005a;c), and Klein et al. (2014). There are 51,016 household records, and 50,661 of those have positive household income. For the purpose of this illustration, we proceed as if the $n = 50,661$ households with positive income are a random sample, and as if household income is confidential for all households (in reality, these are public use data). Thus we treat these public use data as the original data in this illustration. In the notation of Section 4, we let the response variable y be the natural logarithm of household income. A number of covariates are available on the data file, and for the illustration presented here we use the same set of covariates as in Klein et al. (2014), namely,

P: household property tax,

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 the household,

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

M: martial status for the head of the household (coded to take values 1-7), R: race for the head of the household (coded to take values 1-4),

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

We refer to the Current Population Survey March 2000 technical documentation (available at `http://www.census.gov/prod/techdoc/cps/cpsmar00.pdf`) and Klein et al. (2014) for details. Therefore $\boldsymbol{x}$, the vector of regressor variables, is defined as

$$\boldsymbol{x} = \Big(1,\ \mathrm{P},\ \mathrm{N},\ \mathrm{L},\ \mathrm{A},\ I(\mathrm{E}{=}32), I(\mathrm{E}{=}33)\ldots, I(\mathrm{E}{=}46), \\ I(\mathrm{M}{=}2), I(\mathrm{M}{=}3), \ldots, I(\mathrm{M}{=}7),\ I(\mathrm{R}{=}2), I(\mathrm{R}{=}3), I(\mathrm{R}{=}4),\ I(\mathrm{S}{=}2)\Big)', \tag{22}$$

where $I(\mathrm{E}{=}32)$ is an indicator for E=32, $I(\mathrm{E}{=}33)$ is an indicator for E=33, etc.; and the model matrix $\boldsymbol{X} = [\boldsymbol{x}_1 \ \cdots \ \boldsymbol{x}_n]$ has $p = 30$ rows and $n = 50,661$ columns, and has rank equal to 30. This model is used throughout this section for generating synthetic data, and for performing the data analysis. The adjusted value of the coefficient of determination when fitting this model to the CPS data is 0.3629. We use R (R Development Core Team, 2013) for all computations reported in this section.

6.1 Data Analysis

We present inference on the unknown regression parameters based on (1) singly imputed synthetic data, (2) multiply imputed synthetic data, and (3) the original data. Under each of these scenarios, we compute point estimates of the parameters and individual 0.95 confidence intervals for each parameter. We generate a single ($m = 1$) synthetic dataset using (8), and we report the unbiased estimators $\boldsymbol{b}^*$ and $\mathrm{RSS}^*/(n-p)$ as defined in Section 4, as well as the individual confidence intervals for the regression coefficients and residual variance using the methods developed in Section 4. We also generate both $m = 5$ and $m = 10$ synthetic datasets by repeating (8) a total of m times, and we obtain point estimates and individual confidence intervals for the parameters using the methods reviewed in Section 2 (when applying these methods, for each β_i we take $Q = \beta_i$, $q = b_i$, $u = \left(\frac{\mathrm{RSS}}{n-p}\right) D_{ii}$, and for σ^2 we take $Q = \sigma^2$, $q = \frac{\mathrm{RSS}}{n-p}$, $u = 2\left(\frac{\mathrm{RSS}}{n-p}\right)^2/(n-p)$). For the sake of comparison, we also use the original data to compute the usual unbiased estimates ($\boldsymbol{b}$ and $\frac{\mathrm{RSS}}{n-p}$) and individual confidence intervals for each β_i and σ^2, which are obtained from (21). The data analysis results are displayed in Table 6, and the following is a summary of the main findings.

1. We see in Table 6 that the point estimates based on the original data, and based on $m = 1$, $m = 5$, and $m = 10$ synthetic datasets all tend to be in agreement.
2. By comparing the original data inference with synthetic data inference, as expected, we generally find that the synthetic data yield wider confidence intervals than the original data.

3. Comparing the singly imputed ($m = 1$) synthetic data inference with multiply imputed ($m = 5$ and $m = 10$) synthetic data inference, we find that singly imputed data yield wider confidence intervals than multiply imputed synthetic data. Furthermore, in general when comparing the inference for $m = 5$ synthetic datasets with the inference for $m = 10$ synthetic datasets, we see that $m = 5$ tends to a wider confidence interval in comparison with $m = 10$.

4. Suppose one tests $H_0 : \beta_i = 0$ versus $H_1 : \beta_i \neq 0$ at level 0.05 by rejecting H_0 if zero is not contained in the 0.95 confidence interval. Upon examining Table 6, we find that there are two times when the test based on the synthetic data yields a different conclusion than the test based on the original data. For the regression coefficient of $I(E = 34)$, we would not reject H_0 based on the original data, but would reject H_0 based on the $m = 10$ multiply imputed synthetic datasets; and for the regression coefficient of $I(M = 2)$, we would not reject H_0 based on the original data, but would reject H_0 based on the singly imputed synthetic data. In all other cases the test based on the synthetic data yields the same conclusion as the test based on the original data.

In summary, we find that the synthetic data estimates tend to agree with the original data estimates, and as m (the number of imputed synthetic datasets) increases, the length of the confidence interval decreases. Therefore, singly imputed synthetic data would tend to yield less efficient inference than multiply imputed data. Such a finding is expected, since multiply imputed synthetic data would appear to release more information than singly imputed synthetic data. This finding is also in agreement with the findings of our simulation studies, as discussed in Section 5. While singly imputed synthetic data appear to yield less efficient inference than multiply imputed synthetic data, one would expect that singly imputed synthetic data would provide an enhanced level of privacy protection in comparison with multiply imputed synthetic data. The next section precisely explores this issue.

6.2 Disclosure Risk Evaluation of Singly Versus Multiply Imputed Partially Synthetic Data

We have seen that singly imputed synthetic data appear to yield less efficient inference than multiply imputed synthetic data. However, one would also reasonably expect that singly imputed synthetic data would tend to yield an enhanced level of privacy protection in comparison with multiply imputed synthetic data. Therefore the purpose of this section is to evaluate the level of privacy protection offered by singly imputed synthetic data in comparison with multiply imputed synthetic data, in the context of the CPS data, and find out exactly what happens. In order to evaluate the disclosure risk, we use the following framework. Let $\{(v_{11}, \dots, v_{n1}), \dots, (v_{1m}, \dots, v_{nm})\}$ denote $m \geq 1$ synthetic versions of $(y_1, \dots, y_n)$, which are generated by repeating the sampling in (8) m times. Having observed the synthetic data $\{(v_{11}, \dots, v_{n1}), \dots, (v_{1m}, \dots, v_{nm})\}$, we assume that the intruder will estimate a confidential target value y_i as $\hat{y}_i = \frac{1}{m} \sum_{j=1}^{m} v_{ij}$.

Table 6: Analysis of CPS data (parameter estimates, 0.95 confidence intervals, and length of confidence interval).

	Original Data Inference			Synthetic Data Inference								
					$m = 1$			$m = 5$			$m = 10$	
Parameter	Est	CI	Len	Est	CI	Len	Est	CI	Len	Est	CI	Len
Intercept	9.53	(9.41, 9.65)	0.245	9.48	(9.31, 9.65)	0.346	9.57	(9.44, 9.70)	0.257	9.52	(9.39, 9.65)	0.254
P	0.00	(0.00, 0.00)	0.000	0.00	(0.00, 0.00)	0.000	0.00	(0.00, 0.00)	0.000	0.00	(0.00, 0.00)	0.000
N	0.36	(0.35, 0.37)	0.021	0.36	(0.35, 0.38)	0.029	0.36	(0.35, 0.37)	0.023	0.36	(0.35, 0.37)	0.022
NL	-0.37	(-0.38, -0.36)	0.025	-0.37	(-0.39, -0.36)	0.036	-0.37	(-0.38, -0.35)	0.027	-0.37	(-0.39, -0.36)	0.027
A	-0.00	(-0.01, -0.00)	0.001	-0.00	(-0.01, -0.00)	0.002	-0.00	(-0.01, -0.00)	0.001	-0.00	(-0.01, -0.00)	0.001
I(E=32)	-0.05	(-0.19, 0.08)	0.261	0.00	(-0.18, 0.19)	0.369	-0.09	(-0.24, 0.05)	0.286	-0.04	(-0.18, 0.10)	0.275
I(E=33)	0.02	(-0.10, 0.15)	0.246	0.03	(-0.14, 0.20)	0.348	-0.02	(-0.15, 0.11)	0.257	0.04	(-0.09, 0.17)	0.259
I(E=34)	0.10	(-0.02, 0.22)	0.238	0.15	(-0.02, 0.31)	0.336	0.07	(-0.05, 0.19)	0.239	0.13	(0.00, 0.25)	0.249
I(E=35)	0.15	(0.03, 0.27)	0.244	0.19	(0.02, 0.37)	0.345	0.14	(0.01, 0.26)	0.248	0.17	(0.04, 0.30)	0.255
I(E=36)	0.19	(0.07, 0.31)	0.241	0.21	(0.04, 0.38)	0.341	0.16	(0.04, 0.28)	0.244	0.20	(0.07, 0.33)	0.258
I(E=37)	0.21	(0.09, 0.33)	0.240	0.24	(0.07, 0.41)	0.340	0.19	(0.06, 0.31)	0.244	0.22	(0.10, 0.35)	0.250
I(E=38)	0.31	(0.18, 0.44)	0.261	0.42	(0.23, 0.60)	0.368	0.31	(0.18, 0.44)	0.265	0.34	(0.20, 0.47)	0.271
I(E=39)	0.56	(0.45, 0.68)	0.229	0.60	(0.44, 0.77)	0.324	0.53	(0.41, 0.64)	0.234	0.58	(0.46, 0.70)	0.241
I(E=40)	0.75	(0.64, 0.87)	0.230	0.78	(0.62, 0.95)	0.325	0.72	(0.60, 0.84)	0.233	0.77	(0.64, 0.89)	0.241
I(E=41)	0.80	(0.68, 0.91)	0.238	0.84	(0.67, 1.01)	0.336	0.77	(0.65, 0.89)	0.241	0.81	(0.68, 0.93)	0.253
I(E=42)	0.91	(0.79, 1.03)	0.240	0.93	(0.76, 1.10)	0.339	0.88	(0.76, 1.01)	0.245	0.93	(0.80, 1.05)	0.250
I(E=43)	1.10	(0.99, 1.22)	0.231	1.12	(0.96, 1.28)	0.326	1.07	(0.96, 1.19)	0.234	1.12	(1.00, 1.24)	0.241
I(E=44)	1.29	(1.17, 1.41)	0.235	1.33	(1.16, 1.49)	0.332	1.26	(1.14, 1.38)	0.239	1.30	(1.17, 1.42)	0.246
I(E=45)	1.50	(1.37, 1.63)	0.254	1.49	(1.31, 1.67)	0.358	1.47	(1.34, 1.60)	0.260	1.52	(1.39, 1.66)	0.265
I(E=46)	1.47	(1.34, 1.60)	0.259	1.47	(1.29, 1.65)	0.366	1.43	(1.30, 1.56)	0.263	1.49	(1.35, 1.62)	0.266
I(M=2)	-0.06	(-0.20, 0.09)	0.288	-0.25	(-0.45, -0.04)	0.407	-0.05	(-0.21, 0.10)	0.307	-0.07	(-0.23, 0.08)	0.310
I(M=3)	-0.30	(-0.36, -0.25)	0.115	-0.30	(-0.38, -0.21)	0.163	-0.33	(-0.39, -0.26)	0.126	-0.30	(-0.36, -0.24)	0.119
I(M=4)	-0.31	(-0.33, -0.28)	0.058	-0.30	(-0.34, -0.26)	0.082	-0.31	(-0.35, -0.28)	0.064	-0.30	(-0.33, -0.27)	0.061
I(M=5)	-0.21	(-0.23, -0.18)	0.047	-0.19	(-0.23, -0.16)	0.066	-0.21	(-0.24, -0.19)	0.050	-0.20	(-0.23, -0.18)	0.051
I(M=6)	-0.41	(-0.45, -0.36)	0.087	-0.39	(-0.45, -0.33)	0.123	-0.40	(-0.46, -0.35)	0.102	-0.41	(-0.45, -0.36)	0.091
I(M=7)	-0.41	(-0.43, -0.38)	0.046	-0.39	(-0.42, -0.36)	0.065	-0.41	(-0.44, -0.39)	0.047	-0.40	(-0.43, -0.38)	0.048
I(R=2)	-0.13	(-0.15, -0.11)	0.047	-0.12	(-0.15, -0.09)	0.066	-0.14	(-0.16, -0.11)	0.048	-0.13	(-0.16, -0.11)	0.050
I(R=3)	-0.24	(-0.30, -0.17)	0.130	-0.28	(-0.37, -0.19)	0.183	-0.28	(-0.36, -0.21)	0.148	-0.25	(-0.32, -0.18)	0.140
I(R=4)	-0.11	(-0.15, -0.07)	0.083	-0.14	(-0.19, -0.08)	0.117	-0.11	(-0.16, -0.06)	0.104	-0.10	(-0.14, -0.05)	0.087
I(S=2)	-0.12	(-0.13, -0.10)	0.030	-0.11	(-0.13, -0.09)	0.043	-0.12	(-0.13, -0.10)	0.033	-0.12	(-0.14, -0.11)	0.032
σ^2	0.63	(0.62, 0.63)	0.015	0.62	(0.61, 0.63)	0.022	0.63	(0.62, 0.64)	0.016	0.63	(0.62, 0.63)	0.016

Then we use the following criterion as a measure of the level of privacy protection:

$$p_{i,\epsilon} = \Pr\left\{ \left| \frac{\hat{y}_i - y_i}{y_i} \right| \leq \epsilon \,\middle|\, y_1, \ldots, y_n \right\}, \tag{23}$$

where $\epsilon > 0$. The above criterion was also used by Klein et al. (2014) to measure the level of privacy protection in the context of noise multiplication to protect extreme values, and it is similar to a criterion used by Lin and Wise (2012). Notice that if the probability (23) is small, then we would conclude that there is a high level of protection against disclosure; and if this probability is large, then we would conclude that there is a low level of protection against disclosure.

We use Monte Carlo simulation with 10^5 iterations to estimate $p_{i,0.01}$ for each of the $n = 50,661$ households in the CPS dataset; these results are summarized in Table 7. The table shows the minimum, maximum, and α-quantile (Q_α) for $\alpha = 0.1, 0.2, \ldots, 0.9$, of $p_{i,0.01}$ over $i = 1, \ldots, n$; the rows of the table correspond to the cases when m, the number of imputations, equals 1, 5, 10, 15, 20, 25, 50, and 100. Histograms of the $p_{i,0.01}$ values for $m = 1$, 5, 10, and 15 are shown in Figure 1, and the histograms for $m = 20$, 25, 50, and 100 are shown in Figure 2. A summary of the findings of the disclosure risk evaluation is as follows:

Looking at Table 7, we observe that $Q_{0.9}$ and the maximum of $p_{i,0.01}$ both increase as m increases. On the other hand, the minimum, $Q_{0.1}$, $Q_{0.2}$, $Q_{0.3}$, and $Q_{0.4}$ of $p_{i,0.01}$ tend to decrease (or remain equal) as m increases. To examine the situation further, from the histograms in Figure 1 and Figure 2, we observe that as m increases, the disclosure risk for many observations actually decreases ($p_{i,0.01}$ is small), but for some observations the disclosure risk increases substantially ($p_{i,0.01}$ is quite large). For instance, when $m = 100$, the maximum value of $p_{i,0.01}$ is 0.88, which indicates a high disclosure risk, but when $m = 1$, the maximum value of $p_{i,0.01}$ is 0.13 which indicates much less disclosure risk. To further explain this phenomenon, note that for fixed $y_1, \ldots, y_n$, by the Law of Large Numbers, we have

$$\hat{y}_i = \frac{1}{m} \sum_{j=1}^{m} v_{ij} \overset{a.s.}{\to} E(v_{ij} | y_1, \ldots, y_n) = \boldsymbol{x}_i' \boldsymbol{b} \ \text{ as } \ m \to \infty.$$

Therefore, for large m, we simply have $\hat{y}_i \approx \boldsymbol{x}_i' \boldsymbol{b}$ where $\boldsymbol{x}_i' \boldsymbol{b}$ is the fitted value from the regression of y_i on $\boldsymbol{x}_i$. This indicates that the values for which $p_{i,0.01}$ is large for large m, are precisely those values that fit the regression line very closely. Graphically, Figure 3(a) shows a plot of y_i versus the fitted value $\boldsymbol{x}_i' \boldsymbol{b}$ when $m = 100$ for those y_i having $p_{i,0.01} > 0.63$ (0.63 is the 0.9-quantile of $p_{i,0.01}$ values in this setting), while Figure 3(b) shows a plot of y_i versus the fitted value $\boldsymbol{x}_i' \boldsymbol{b}$ when $m = 100$ for those y_i having $p_{i,0.01} \leq 0.63$. In Figure 3(a) we clearly see that the large values of $p_{i,0.01}$ occur for y_i values that are close to their fitted values, and in Figure 3(b) we see that the small values of $p_{i,0.01}$ occur for y_i values that are not as close to their fitted values.

In summary, we can conclude that as m increases the disclosure risk for some observations increases. When $m = 1$, the maximum value of $p_{i,0.01}$ is only 0.13, and when

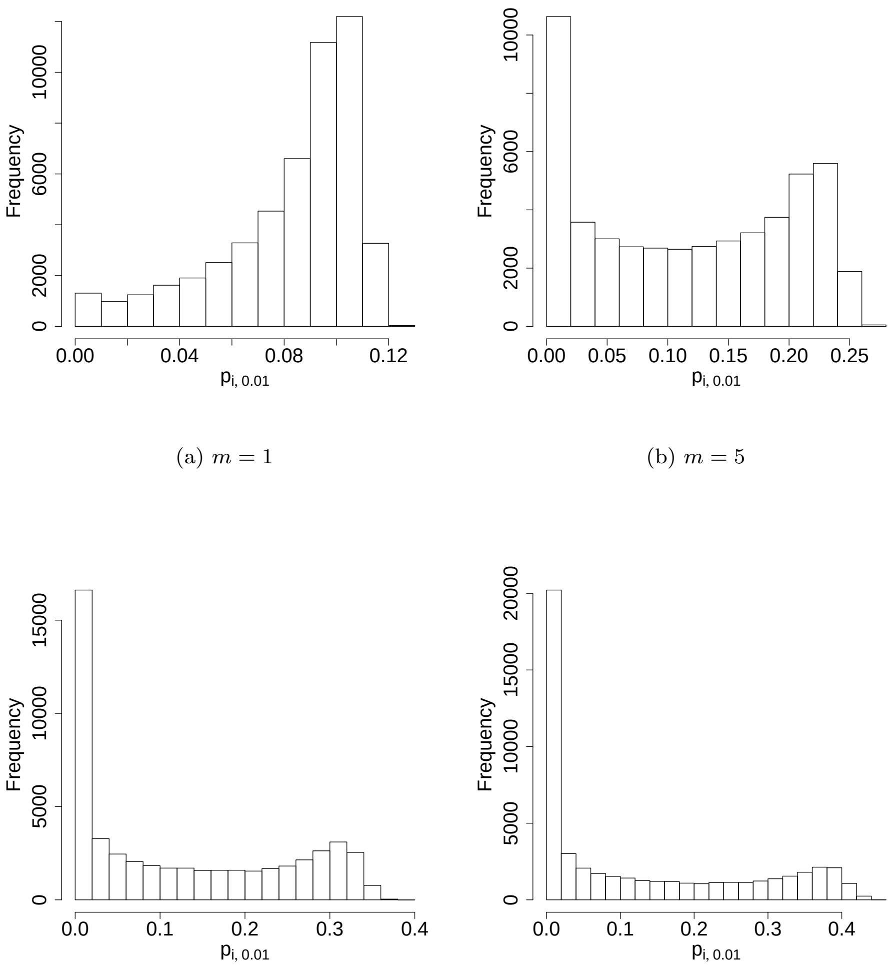
$m = 5$ we see that the maximum value is more than double at 0.28. When $m = 100$ the maximum value of $p_{i,0.01}$ increases substantially to 0.88. So if it is desired that $p_{i,0.01}$ be kept uniformly small over all i, then a single imputation would provide the most protection. Even a modest jump from $m = 1$ to $m = 5$ more than doubles the maximum value of $p_{i,0.01}$. The potential for disclosure risk to increase as m increases has also been discussed by Reiter and Mitra (2009), who used the probability that an intruder can identify a target individual, conditional on the released data, as the basis for evaluating disclosure risk.

Remark 6.1. For each i, we see that $p_{i,\epsilon}$ is increasing in ϵ. Therefore if a large value of ϵ is chosen, then the values of $\{p_{i,\epsilon} : i = 1, \dots, n\}$ would tend to be larger, as compared to if a small value of ϵ is chosen. Thus to ensure a large amount of protection against disclosure, as measured by (23), a statistical agency could evaluate the $p_{i,\epsilon}$ values using a large value of ϵ, and then choose a statistical disclosure control strategy that yields small values of $p_{i,\epsilon}$ for the chosen ϵ. Table 8 is similar to Table 7, but with ϵ increased to 0.05. As expected, comparing Tables 7 and 8, we observe that the values of $p_{i,0.05}$ tend to be larger than the values of $p_{i,0.01}$. The behavior of $p_{i,0.05}$ as m increases is generally inline with the above discussion for $p_{i,0.01}$.

Remark 6.2. Suppose that instead of releasing synthetic data, we release the observed value of the least squares estimate $\boldsymbol{b}$, which is computed on the original data. Then the fitted value $\hat{y}_i = \boldsymbol{x}_i'\boldsymbol{b}$ is a natural estimate of the intruder's target value y_i. In this case the criterion $p_{i,\epsilon}$ simply equals 0 or 1 for each i, because conditional on $y_1, \dots, y_n$, the quantity $\hat{y}_i = \boldsymbol{x}_i'\boldsymbol{b}$ is not random. In the CPS data example, we find that $p_{i,0.01}$ equals 0 for 43,801 of the y_i's (86%) and $p_{i,0.01}$ equals 1 for 6,860 of the y_i's (14%). We noted above that when multiply imputed synthetic data are released $\hat{y}_i = \frac{1}{m}\sum_{j=1}^{m} v_{ij} \overset{a.s.}{\to} \boldsymbol{x}_i'\boldsymbol{b}$ as $m \to \infty$ for fixed $y_1, \dots, y_n$. Therefore, whether the agency releases a large number of multiply imputed synthetic datasets, or the estimated parameter $\boldsymbol{b}$ based on the original data, the intruder's estimate of y_i would be approximately the same.

Table 7: Distribution of $p_{i,0.01} = \Pr\left\{\left|\frac{\hat{y}_i - y_i}{y_i}\right| \le 0.01 \,\middle|\, y_1, \dots, y_n\right\}$ in the CPS data example over all 50,661 y_i-values.

m	Min	$Q_{0.1}$	$Q_{0.2}$	$Q_{0.3}$	$Q_{0.4}$	$Q_{0.5}$	$Q_{0.6}$	$Q_{0.7}$	$Q_{0.8}$	$Q_{0.9}$	Max
1	0.00	0.04	0.06	0.08	0.08	0.09	0.10	0.10	0.10	0.11	0.13
5	0.00	0.00	0.02	0.05	0.08	0.12	0.16	0.19	0.21	0.23	0.28
10	0.00	0.00	0.00	0.01	0.04	0.09	0.15	0.21	0.27	0.31	0.38
15	0.00	0.00	0.00	0.00	0.02	0.06	0.13	0.21	0.30	0.36	0.45
20	0.00	0.00	0.00	0.00	0.01	0.04	0.10	0.20	0.32	0.41	0.51
25	0.00	0.00	0.00	0.00	0.00	0.03	0.08	0.19	0.33	0.44	0.57
50	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.12	0.32	0.55	0.73
100	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.05	0.27	0.63	0.88



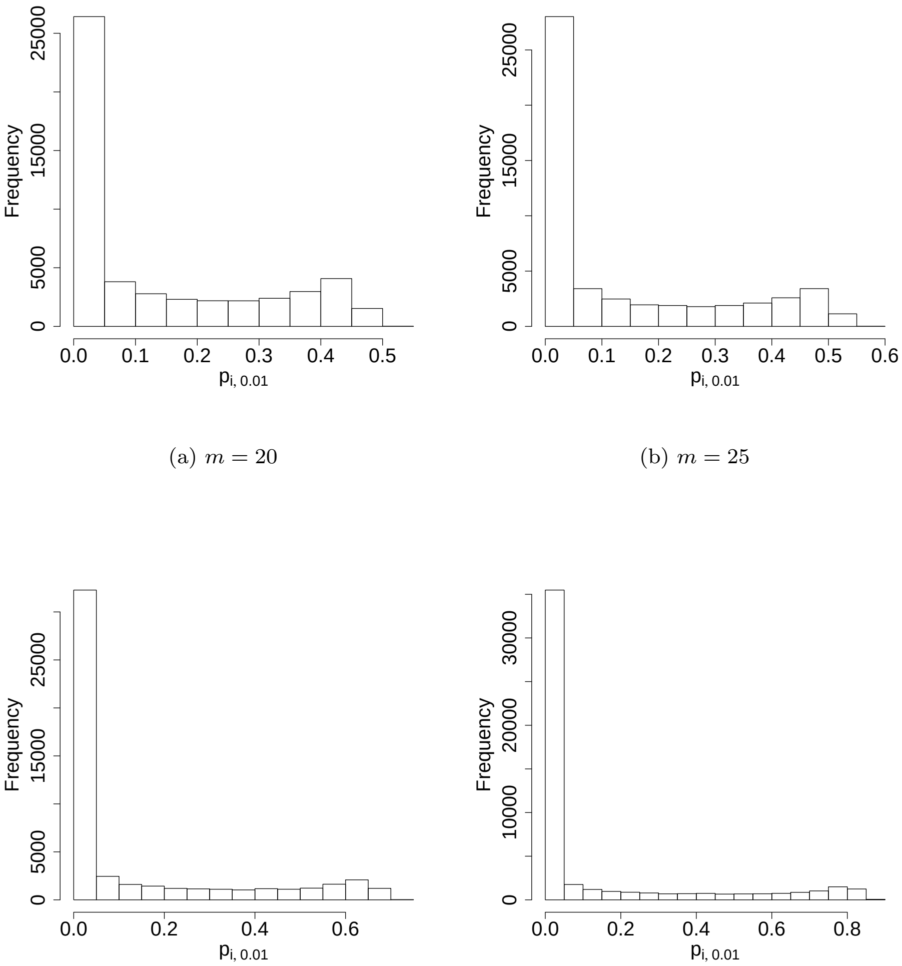
(a) $m = 1$

(b) $m = 5$

(c) $m = 10$

(d) $m = 15$

Figure 1: Histograms of $p_{i,0.1}$ values.



(a) $m = 20$

(b) $m = 25$

(c) $m = 50$

(d) $m = 100$

Figure 2: Histograms of $p_{i,0.1}$ values.

Table 8: Distribution of $p_{i,0.05} = \Pr\left\{\left|\frac{\hat{y}_i - y_i}{y_i}\right| \leq 0.05 \,\middle|\, y_1, \ldots, y_n\right\}$ in the CPS data example over all 50,661 y_i-values.

m	Min	$Q_{0.1}$	$Q_{0.2}$	$Q_{0.3}$	$Q_{0.4}$	$Q_{0.5}$	$Q_{0.6}$	$Q_{0.7}$	$Q_{0.8}$	$Q_{0.9}$	Max
1	0.00	0.21	0.31	0.37	0.41	0.44	0.46	0.47	0.49	0.50	0.57
5	0.00	0.05	0.21	0.37	0.51	0.62	0.71	0.78	0.82	0.85	0.92
10	0.00	0.01	0.12	0.32	0.51	0.68	0.79	0.88	0.93	0.95	0.99
15	0.00	0.00	0.08	0.28	0.52	0.71	0.84	0.92	0.96	0.98	1.00
20	0.00	0.00	0.05	0.25	0.52	0.74	0.88	0.95	0.98	0.99	1.00
25	0.00	0.00	0.03	0.23	0.52	0.77	0.90	0.97	0.99	1.00	1.00
50	0.00	0.00	0.00	0.15	0.53	0.85	0.97	1.00	1.00	1.00	1.00
100	0.00	0.00	0.00	0.07	0.55	0.93	1.00	1.00	1.00	1.00	1.00

7 Conditions for Valid Inference, Further Analysis, Extensions

In this section we summarize the key practical conditions under which the proposed methodology for analyzing singly imputed data will yield valid inference (Subsection 7.1); we provide some results to illustrate what can happen when certain conditions are violated (Subsection 7.2); and we indicate some extensions of the methodology that relax some of the conditions (Subsection 7.3). The discussion in this section focuses only on the case of the multiple linear regression model as discussed in Section 4.

7.1 Conditions for Valid Inference

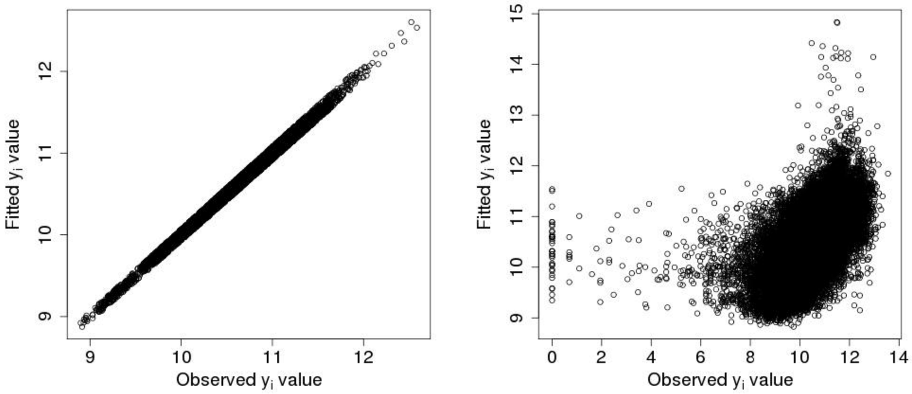
When generating and analyzing synthetic data, there are essentially three statistical models underlying the whole process:

1. *Data generating model* (DM): The true population model that generated the original data.

2. *Imputation model* (IM): The statistical agency's assumed model for the original data. Based on this model, the statistical agency determines a procedure for creating synthetic data.

3. *Analysis model* (AM): The data analyst's assumed model for the original data. Based on this model, the data analyst, who only has access to the synthetic data, determines an appropriate procedure for analyzing the synthetic data.

Section 4 explains the specific mathematical assumptions under which the methodology developed in that section is derived. These assumptions yield a set of conditions under which the proposed methodology will provide valid statistical inference. For the sake of clarity, below we list the necessary conditions in practical terms.



(a) y_i with $p_{i,0.01} > 0.63$ when $m = 100$

(b) y_i with $p_{i,0.01} \leq 0.63$ when $m = 100$

Figure 3: Plots of observed y_i values versus fitted values (defined as $\boldsymbol{x}_i'\boldsymbol{b}$) for large and small $p_{i,0.01}$ values when the number of imputations is $m = 100$.

C1. The DM is the multiple linear regression model (7).

C2. The IM is also (7), and based on this model, the statistical agency obtains the estimates $\boldsymbol{b}$ and RSS$/(n-p)$ for $\boldsymbol{\beta}$ and σ^2, respectively, and hence uses the fitted regression model to generate the synthetic data, as in (8).

C3. The variable y_i is synthesized for each $i = 1, \ldots, n$; none of the regressor variables $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ are synthesized.

C4. The original data $\{(y_i, \boldsymbol{x}_i) : i = 1, \ldots, n\}$ are fully available to the statistical agency, i.e., there are no missing data.

C5. The released data $\{(v_i, \boldsymbol{x}_i) : i = 1, \ldots, n\}$ are available to the data analyst; the data analyst correctly knows that v_i is a synthetic version of y_i for each $i = 1, \ldots, n$; and also knows that $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ are original (not synthetic) values.

C6. The AM is also (7), and the analyst will use the data $\{(v_i, \boldsymbol{x}_i) : i = 1, \ldots, n\}$ to draw inference on the parameters $\boldsymbol{\beta}$ (or a linear function of $\boldsymbol{\beta}$) and σ^2 that appear in (7).

If the conditions above hold, then the data analyst can apply the methods of Section 4 to draw valid inference on $\boldsymbol{\beta}$ and σ^2 using the released data $\{(v_i, \boldsymbol{x}_i) : i = 1, \ldots, n\}$. Obviously, in a real life scenario, these conditions can be violated in a number of ways. In subsections that follow, we consider some scenarios that could lead to a violation of one or more of these conditions. We evaluate the properties of our methodology when some conditions do not hold, and discuss their implications.

7.2 Scenarios where the DM, IM, and AM Differ

Imputer and/or Analyst Overfit or Underfit the Regression Model

One way the DM, IM, and AM can differ is by including either too many (overfitting), or not enough (underfitting) covariates in the linear regression model and the imputation and/or analysis models. To study this scenario, we assume for simplicity the case of two covariates. Let M_F and M_R denote the full and reduced models, respectively, which we define as follows:

$$M_F: \ y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + \epsilon_i, \ \epsilon_i \overset{iid}{\sim} N(0, \sigma^2), \ \text{for } i = 1, \ldots, n, \ \text{and } (\beta_1, \beta_2, \sigma^2) \text{ unknown;} \tag{24}$$

$$M_R: \ y_i = \beta_1 x_{1i} + \epsilon_i, \ \epsilon_i \overset{iid}{\sim} N(0, \sigma^2), \ \text{for } i = 1, \ldots, n, \ \text{and } (\beta_1, \sigma^2) \text{ unknown.} \tag{25}$$

Under both models M_F and M_R, the regressor variables are treated as fixed. For the purpose of this study, we restrict attention to the possibility of just the two models M_F and M_R. We evaluate the performance of our methodology under the following eight possible cases, where in each case we consider β_1 as the parameter of interest.

Case 1:	DM $= M_F$,	IM $= M_F$,	AM $= M_F$.
Case 2:	DM $= M_F$,	IM $= M_F$,	AM $= M_R$.
Case 3:	DM $= M_F$,	IM $= M_R$,	AM $= M_F$.
Case 4:	DM $= M_F$,	IM $= M_R$,	AM $= M_R$.
Case 5:	DM $= M_R$,	IM $= M_F$,	AM $= M_F$.
Case 6:	DM $= M_R$,	IM $= M_F$,	AM $= M_R$.
Case 7:	DM $= M_R$,	IM $= M_R$,	AM $= M_F$.
Case 8:	DM $= M_R$,	IM $= M_R$,	AM $= M_R$.

For comparison sake, in each case we also discuss what happens under multiple imputation. If DM $= M_F$ then (24) is the true model that generated the original data $\boldsymbol{y}$; and if DM $= M_R$ then (25) is the true model that generated $\boldsymbol{y}$.

Data Analysis Under Single Imputation. Under single imputation, the released data are $\mathscr{D} = \{(v_i, x_{1i}, x_{2i}) : i = 1, \ldots, n\}$. If IM $= M_F$, then the statistical agency generates the synthetic data as in (8) with $\boldsymbol{x}_i = \begin{pmatrix} x_{1i} \\ x_{2i} \end{pmatrix}$; and if IM $= M_R$, then the statistical agency generates the synthetic data again as in (8), but this time taking $\boldsymbol{x}_i = x_{1i}$. If AM $= M_F$, then the data user applies the results of Section 4, with $p = 2$, $\boldsymbol{x}_i = \begin{pmatrix} x_{1i} \\ x_{2i} \end{pmatrix}$, $\boldsymbol{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \end{pmatrix}$; and hence the data user estimates β_1 by b_1^* where $\boldsymbol{b}^* = \begin{pmatrix} b_1^* \\ b_2^* \end{pmatrix}$ is defined in Result 4.1; the data user estimates $\text{Var}(b_1^*)$ by the

$(1,1)$ element in the matrix $\widehat{\mathrm{Var}}(\boldsymbol{b}^*)$ which is defined in Result 4.2; and the data user computes a $(1-\gamma)$ level confidence interval for β_1 using (14). If AM $= M_R$, then the data user again applies the results of Section 4, but this time with $p = 1$, $\boldsymbol{x}_i = x_{1i}$, $\boldsymbol{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1n} \end{pmatrix}$; and hence the data user estimates β_1 by b_1^* where $\boldsymbol{b}^* = b_1^*$ is defined in Result 4.1; the data user estimates $\mathrm{Var}(b_1^*)$ by $\widehat{\mathrm{Var}}(b_1^*) = \widehat{\mathrm{Var}}(\boldsymbol{b}^*)$ which is defined in Result 4.2; and the data user computes a $(1-\gamma)$ level confidence interval for β_1 using (14) (which is equivalent to (11) in this case, since $p = 1$).

Data Analysis Under Multiple Imputation. Under multiple imputation, the released data are $\{\mathscr{D}_1, \ldots, \mathscr{D}_m\}$, where $m > 1$, and $\mathscr{D}_j = \{(v_{ij}, x_{1i}, x_{2i}) : i = 1, \ldots, n\}$ for $j = 1, \ldots, m$. If IM $= M_F$, then the statistical agency generates the synthetic data by repeating (8) independently m times, taking $\boldsymbol{x}_i = \begin{pmatrix} x_{1i} \\ x_{2i} \end{pmatrix}$; and if IM $= M_R$, then the statistical agency generates the synthetic data again by repeating (8) independently m times, but taking $\boldsymbol{x}_i = x_{1i}$. If AM $= M_F$, then the data user takes $p = 2$, $\boldsymbol{x}_i = \begin{pmatrix} x_{1i} \\ x_{2i} \end{pmatrix}$, $\boldsymbol{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \end{pmatrix}$, and then estimates β_1 using the methodology of Section 2 with $Q = \beta_1$, $q = b_1$, $u = \mathrm{RSS}/(n-p)$. If AM $= M_R$, then the data user again applies the results of Section 2 with $Q = \beta_1$, $q = b_1$, $u = \mathrm{RSS}/(n-p)$, but now $p = 1$, $\boldsymbol{x}_i = x_{1i}$, and $\boldsymbol{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1n} \end{pmatrix}$, i.e., b_1 and RSS are now the least squares estimate and residual sum of squares based on just one regressor variable.

Analysis of Single and Multiple Imputation Inference Under Cases 1-8. Table 9 provides some theoretical properties of the estimator b_1^* in Cases 1-8 under singly imputed synthetic data, including $E(b_1^*)$ as well as $\mathrm{Var}(b_1^*)$ and $E[\widehat{\mathrm{Var}}(b_1^*)]$ in cases where the bias of b_1^* is negligible. Proofs of these expression appear in Appendix 3. For the sake of comparison, Table 10 provides similar theoretical properties for the estimator $\bar{b}_{1,m}^*$ in Case 1-8 under multiply imputed synthetic data; we omit the proofs of these expressions because the derivations are similar to those that appear in Appendix 3 for $m = 1$. In Tables 9 and 10, under Cases 2, 3, and 4 we omit the expressions for the variance of the estimator of β_1, and the expected value of the estimator of this variance, because in these cases the estimator of β_1 has non-negligible bias, and this bias generally will cause inferences to be invalid.

To complement the theoretical results, we also used Monte Carlo simulation based on 10^6 iterations to estimate the bias, variance, and expected value of the estimated variance of b_1^* and $\bar{b}_{1,m}^*$. Using Monte Carlo simulation we also estimated the coverage of the nominal 0.95 confidence interval for β_1, and the expected length of this confidence interval, under both single and multiple imputation. For the simulation studies, we set $\sigma^2 = 4$, $n = 2000$, and for multiple imputation, the number of imputations was $m = 10$. We generated $(x_{11}, \ldots, x_{1n})$ and $(x_{21}, \ldots, x_{2n})$ all independently from the $N(12, 9)$ distribution, one time at the beginning of the simulation, and then held the x-values fixed across iterations of the simulation. We used the same x-values under each Case 1-8. For cases where the DM is M_F, we set $\beta_1 = 5$ and $\beta_2 = 4$; and for cases where the DM is M_R, we set $\beta_1 = 5$ ($\beta_2 = 0$ when the IM is M_R). The simulation results appear in Table 11.

Below we discuss the results of our theoretical and empirical evaluations that appear in Tables 9, 10, and 11. Notice that in Cases 1 and 8, the DM, IM, and AM all agree, and both Cases 1 and 8 fall under the assumptions of Section 4, with $\boldsymbol{X}$ appropriately defined. On the other hand, Cases 2-7 do not fall under the assumptions of Section 4, because in these cases, the DM, IM, and AM are not all the same.

Imputer incorrectly assumes β_2 equals 0. In Cases 3 and 4 the DM is M_F and the IM is M_R. In both of these cases the imputer has incorrectly assumed $\beta_2 = 0$. Because of this incorrect assumption, the estimator of β_2 is biased under both single and multiple imputation, as expected (Meng, 1994). It is seen from the tables that the bias is exactly the same for both single and multiple imputation, and it would vanish if $\sum_{i=1}^n x_{1i}x_{2i} = 0$. Table 11 indicates that the bias is substantial enough that the nominal 0.95 confidence interval has actual coverage of approximately 0 for both single and multiple imputation.

Imputer assumes more than the analyst. In Cases 3 and 7 the IM is M_R and the AM is M_F; thus the imputer assumes $\beta_2 = 0$ while the analyst does not. Above we discussed that there is bias in Case 3, so we will not discuss that case any further. In Case 7 the DM is M_R and therefore the imputer is correct in assuming $\beta_2 = 0$, and therefore the synthetic data incorporate this extra information. However, the analyst being unaware of the fact that $\beta_2 = 0$, fits the full model M_F to the synthetic data. Under single and multiple imputation, it is readily seen from Tables 9 and 10 that $E[\widehat{\text{Var}}(b_1^*)] > \text{Var}(b_1^*)$ and $E[\widehat{\text{Var}}(\bar{b}_{1,m}^*)] > \text{Var}(\bar{b}_{1,m}^*)$, i.e., the variance estimator is positively biased for the true variance under single and multiple imputation. The simulation results in Table 11 also confirm this statement. Notice that under single imputation $\text{Var}(b_1^*) = \sigma^2\left(\frac{\sum_{i=1}^n x_{2i}^2}{\Delta}\right) + \sigma^2\left(\frac{1}{\sum_{i=1}^n x_{1i}^2}\right)$, and under multiple imputation $\text{Var}(\bar{b}_{1,m}^*) = \sigma^2\left(\frac{1}{m}\right)\left(\frac{\sum_{i=1}^n x_{2i}^2}{\Delta}\right) + \sigma^2\left(\frac{1}{\sum_{i=1}^n x_{1i}^2}\right)$. Under multiple imputation, if the number of imputations m is large, then the first term in the variance becomes negligible, yielding

$$\text{Var}(\bar{b}_{1,m}^*) \approx \sigma^2\left(\frac{1}{\sum_{i=1}^n x_{1i}^2}\right) = \text{Var}\left(\frac{\sum_{i=1}^n x_{1i}y_i}{\sum_{i=1}^n x_{1i}^2}\right) \leq \text{Var}\left(\frac{\sum_{i=1}^n c_i y_i}{\Delta}\right) = \sigma^2\left(\frac{\sum_{i=1}^n x_{2i}^2}{\Delta}\right),$$

where $c_i = x_{1i}(\sum_{i=1}^n x_{2i}^2) - x_{2i}(\sum_{i=1}^n x_{1i}x_{2i})$ and $\Delta = (\sum_{i=1}^n x_{1i}^2)(\sum_{i=1}^n x_{2i}^2) - (\sum_{i=1}^n x_{1i}x_{2i})^2$. Because the AM is M_F, even if the data analysis had observed the original data, the analyst would estimate β_1 using $\frac{\sum_{i=1}^n c_i y_i}{\Delta}$, the least squares estimate under M_F (which is unbiased), instead of $\frac{\sum_{i=1}^n x_{1i}y_i}{\sum_{i=1}^n x_{1i}^2}$, the least squares estimator under the correct model M_R (which is the UMVUE). Thus the data analyst's estimate of β_1 based on multiply imputed synthetic data has smaller variance than the analyst's estimate under the original data, if m is large. One can also easily show that $\bar{b}_{1,\infty}^* \equiv \lim_{m\to\infty} \bar{b}_{1,m}^* = \frac{\sum_{i=1}^n x_{1i}y_i}{\sum_{i=1}^n x_{1i}^2}$, a.s., for fixed $\boldsymbol{y}$. Rubin (1996) discusses this phenomenon in the context of multiple imputation for missing data, and refers to the imputations in such a scenario as *strongly superefficient*. This phenomenon has also been discussed by Meng (1994) in the context of multiple imputation for missing data. The imputations here are referred to as strongly superefficient because while the data analyst is unaware that $\beta_2 = 0$, the imputer has incorporated this information into the imputations; as a result, for sufficiently large m, the variance of $\bar{b}^*_{1,m}$ is less than the variance of $\frac{\sum_{i=1}^n c_i y_i}{\Delta}$ (which is the estimate the data analyst would have used if given access to the original data). This phenomenon does not occur under single imputation, because

$$\mathrm{Var}(b_1^*) = \sigma^2 \left(\frac{\sum_{i=1}^n x_{2i}^2}{\Delta} \right) + \sigma^2 \left(\frac{1}{\sum_{i=1}^n x_{1i}^2} \right) > \sigma^2 \left(\frac{\sum_{i=1}^n x_{2i}^2}{\Delta} \right) = \mathrm{Var} \left(\frac{\sum_{i=1}^n c_i y_i}{\Delta} \right).$$

Under both single and multiple imputation, it is true that the variance estimator is positively biased for the true variance of the estimator of β_1; as a result, we observe in Table 11 that under both single and multiple imputation, the confidence interval for β_1 has true coverage well above the nominal level of 0.95.

Analyst assumes more than the imputer. In Cases 2 and 6 the IM is M_F and the AM is M_R; thus the analyst assumes $\beta_2 = 0$ while the imputer does not. In Case 2 the analyst is incorrect in making this assumption, because the DM is M_F, as a result the analyst's estimate is biased. The bias observed in Case 6 is expected because even if the original data were observed, it is well known that fitting an underspecified regression model generally leads to biased estimates (Rencher and Schaalje, 2008). In Case 2 the bias is the same under both single and multiple imputation, and it is large enough so that in the simulation results the nominal 0.95 level confidence interval has true coverage approximately equal to 0. In Case 6 the analyst is correct to assume $\beta_2 = 0$, because the DM is M_R, however, the imputer has used an overspecified model to create the synthetic data. In this case under single imputation we find that $\mathrm{Var}(b_1^*) = 2\sigma^2 \left(\frac{1}{\sum_{i=1}^n x_{1i}^2} \right)$, which is the same as the variance of b_1^* under Case 8, therefore overspecification of the IM has not inflated the variance. Similarly, under multiple imputation $\mathrm{Var}(\bar{b}^*_{1,m}) = \sigma^2 \left(1 + \frac{1}{m}\right) \left(\frac{1}{\sum_{i=1}^n x_{1i}^2} \right)$, which is the same as the variance of $\bar{b}^*_{1,m}$ under Case 8, again indicating that overspecification of the IM has not inflated the variance. We also notice that under both single and multiple imputation, the bias of the variance estimator is negligible for large sample sizes because it is of the order $O(n^{-1})$. The simulation results in Table 11 also indicate that in Case 6, overspecification of the IM, when the AM is correct, has not inflated the variance, and the confidence interval for β_1 still covers at the nominal rate of 0.95.

Analyst and imputer make the same assumptions. In Cases 1, 4, 5, and 8 the IM and AM are the same. In both Cases 1 and 8, the DM, IM and AM are all the same. Cases 1 and 8 fall under the assumptions of Section 4, and the results agree with the theory developed in that section. The comparison between single and model imputation in Cases 1 and 8 is inline with the comparison given in Remark 4.2, and in Subsection 5.2 and Section 6. In Case 4, the IM and AM are both M_R, while the DM is M_F; thus the IM and AM are both underspecified, causing the estimator of β_1

to be biased. We see in Tables 9, 10, and 11 that in Case 4 the bias is the same under single and multiple imputation, and it is substantial enough so that the nominal 0.95 level confidence interval has actual coverage approximately equal to 0. In Case 5 the IM and AM are both M_F, while the DM is M_R; thus both the imputer and analyst have overspecified the model. Under both single and multiple imputation, we find that the estimator of β_1 is unbiased in this case, and the estimated variance estimator is unbiased for the true variance. The simulation results also indicate that the coverage of the nominal 0.95 level confidence interval is in fact, approximately equal to 0.95. Comparing the variance of b_1^* in Case 5, with the variance of b_1^* in Case 8, we see that the variance is larger in Case 5; this inflation of the variance is caused by overspecification of both the IM and AM.

Summary. In the model disagreement framework described by Cases 1-8, single and multiple imputation tend to agree in the sense that single and multiple imputation both provide valid inference in Cases 1, 5, 6, 7, and 8, and invalid inference in Cases 2, 3, and 4 (due to the presence of bias). One notable difference between single and multiple imputation is that multiple imputation can offer superefficiency (Rubin, 1996) in Case 7, while single imputation cannot; however in Case 7, both single and multiple imputation yield a positively biased variance estimate, resulting in a confidence interval whose coverage is well above the nominal rate.

Table 9: Properties of the estimator of β_1 in Cases 1-8 of the DM, IM, and AM under singly imputed synthetic data.

Case	$E(b_1^*)$	$\text{Var}(b_1^*)$	$E[\widehat{\text{Var}}(b_1^*)]$
1	β_1	$2\sigma^2\left(\frac{\sum_{i=1}^n x_{2i}^2}{\Delta}\right)$	$2\sigma^2\left(\frac{\sum_{i=1}^n x_{2i}^2}{\Delta}\right)$
2	$\beta_1 + \beta_2\left(\frac{\sum_{i=1}^n x_{1i}x_{2i}}{\sum_{i=1}^n x_{1i}^2}\right)$	–	–
3	$\beta_1 + \beta_2\left(\frac{\sum_{i=1}^n x_{1i}x_{2i}}{\sum_{i=1}^n x_{1i}^2}\right)$	–	–
4	$\beta_1 + \beta_2\left(\frac{\sum_{i=1}^n x_{1i}x_{2i}}{\sum_{i=1}^n x_{1i}^2}\right)$	–	–
5	β_1	$2\sigma^2\left(\frac{\sum_{i=1}^n x_{2i}^2}{\Delta}\right)$	$2\sigma^2\left(\frac{\sum_{i=1}^n x_{2i}^2}{\Delta}\right)$
6	β_1	$2\sigma^2\left(\frac{1}{\sum_{i=1}^n x_{1i}^2}\right)$	$2\sigma^2\left(\frac{n}{n-1}\right)\left(\frac{1}{\sum_{i=1}^n x_{1i}^2}\right)$
7	β_1	$\sigma^2\left(\frac{\sum_{i=1}^n x_{2i}^2}{\Delta}\right) + \sigma^2\left(\frac{1}{\sum_{i=1}^n x_{1i}^2}\right)$	$2\sigma^2\left(\frac{\sum_{i=1}^n x_{2i}^2}{\Delta}\right)$
8	β_1	$2\sigma^2\left(\frac{1}{\sum_{i=1}^n x_{1i}^2}\right)$	$2\sigma^2\left(\frac{1}{\sum_{i=1}^n x_{1i}^2}\right)$

$\Delta = (\sum_{i=1}^n x_{1i}^2)(\sum_{i=1}^n x_{2i}^2) - (\sum_{i=1}^n x_{1i}x_{2i})^2$

Table 10: Properties of the estimator of β_1 in Cases 1-8 of the DM, IM, and AM under multiply imputed synthetic data.

Case	$E(\overline{b}^*_{1,m})$	$\mathrm{Var}(\overline{b}^*_{1,m})$	$E[\widehat{\mathrm{Var}}(\overline{b}^*_{1,m})]$
1	β_1	$\sigma^2\left(1+\frac{1}{m}\right)\left(\frac{\sum_{i=1}^n x_{2i}^2}{\Delta}\right)$	$\sigma^2\left(1+\frac{1}{m}\right)\left(\frac{\sum_{i=1}^n x_{2i}^2}{\Delta}\right)$
2	$\beta_1+\beta_2\left(\frac{\sum_{i=1}^n x_{1i}x_{2i}}{\sum_{i=1}^n x_{1i}^2}\right)$	–	–
3	$\beta_1+\beta_2\left(\frac{\sum_{i=1}^n x_{1i}x_{2i}}{\sum_{i=1}^n x_{1i}^2}\right)$	–	–
4	$\beta_1+\beta_2\left(\frac{\sum_{i=1}^n x_{1i}x_{2i}}{\sum_{i=1}^n x_{1i}^2}\right)$	–	–
5	β_1	$\sigma^2\left(1+\frac{1}{m}\right)\left(\frac{\sum_{i=1}^n x_{2i}^2}{\Delta}\right)$	$\sigma^2\left(1+\frac{1}{m}\right)\left(\frac{\sum_{i=1}^n x_{2i}^2}{\Delta}\right)$
6	β_1	$\sigma^2\left(1+\frac{1}{m}\right)\left(\frac{1}{\sum_{i=1}^n x_{1i}^2}\right)$	$\sigma^2\left(\frac{1}{m}+\frac{n}{n-1}\right)\left(\frac{1}{\sum_{i=1}^n x_{1i}^2}\right)$
7	β_1	$\sigma^2\left(\frac{1}{m}\right)\left(\frac{\sum_{i=1}^n x_{2i}^2}{\Delta}\right)+\sigma^2\left(\frac{1}{\sum_{i=1}^n x_{1i}^2}\right)$	$\sigma^2\left(1+\frac{1}{m}\right)\left(\frac{\sum_{i=1}^n x_{2i}^2}{\Delta}\right)$
8	β_1	$\sigma^2\left(1+\frac{1}{m}\right)\left(\frac{1}{\sum_{i=1}^n x_{1i}^2}\right)$	$\sigma^2\left(1+\frac{1}{m}\right)\left(\frac{1}{\sum_{i=1}^n x_{1i}^2}\right)$

$\Delta = (\sum_{i=1}^n x_{1i}^2)(\sum_{i=1}^n x_{2i}^2) - (\sum_{i=1}^n x_{1i}x_{2i})^2$

Analysis Model is the Regression of x on y

Another scenario where the IM and AM differ occurs if the data analyst chooses to fit a linear regression model where one of the x-variables is the response, and the y-variable is among the regressors. In order to study the effect that this scenario can have on the single imputation inference methodology of Section 4, and to compare with the multiple imputation methodology of Section 2, we consider the following scenario. Suppose that the original data are $\boldsymbol{y} = (y_1, \ldots, y_n)$ and $\boldsymbol{x} = (x_1, \ldots, x_n)$, where $\boldsymbol{y}$ is sensitive and $\boldsymbol{x}$ is not sensitive. Suppose the DM is

$$\begin{pmatrix} y_1 \\ x_1 \end{pmatrix}, \ldots, \begin{pmatrix} y_n \\ x_n \end{pmatrix} \sim iid \sim N_2\left[\begin{pmatrix} \mu_y \\ \mu_x \end{pmatrix}, \begin{pmatrix} \sigma_y^2 & \sigma_{xy} \\ \sigma_{xy} & \sigma_x^2 \end{pmatrix}\right]. \tag{26}$$

Under model (26), it follows that

$$y_1, \ldots, y_n \,|\, \boldsymbol{x} \text{ are independently distributed such that } y_i|\boldsymbol{x} \sim N(\beta_1 + \beta_2 x_i,\ \sigma_{y|x}^2); \tag{27}$$

$$x_1, \ldots, x_n \,|\, \boldsymbol{y} \text{ are independently distributed such that } x_i|\boldsymbol{y} \sim N(\omega_1 + \omega_2 y_i,\ \sigma_{x|y}^2); \tag{28}$$

Table 11: Simulation results in Cases 1-8 of the DM, IM, and AM under single and multiple imputation.

m	Case	Bias	Var	$\widehat{\text{Var}}$	Cvg	Len
1	1	1.946E-05	2.190E-04	2.188E-04	0.950	0.058
	2	3.728	2.560E-05	1.867E-03	0.000	0.170
	3	3.728	7.992E-03	1.597E-02	0.000	0.496
	4	3.728	9.448E-04	1.867E-03	0.000	0.170
	5	-4.206E-05	2.188E-04	2.188E-04	0.950	0.058
	6	5.718E-07	2.557E-05	2.560E-05	0.950	0.020
	7	-1.556E-06	1.222E-04	2.188E-04	0.991	0.058
	8	1.317E-06	2.569E-05	2.559E-05	0.950	0.020
10	1	-1.280E-05	1.205E-04	1.204E-04	0.949	0.043
	2	3.728	1.408E-05	9.349E-04	0.000	0.120
	3	3.728	8.115E-04	8.783E-03	0.000	0.368
	4	3.728	1.063E-04	1.027E-03	0.000	0.126
	5	-9.737E-06	1.203E-04	1.204E-04	0.950	0.043
	6	-1.869E-06	1.408E-05	1.408E-05	0.950	0.015
	7	-1.435E-06	2.375E-05	1.204E-04	1.000	0.043
	8	1.161E-06	1.410E-05	1.407E-05	0.950	0.015

where

$$\beta_1 = \mu_y - \frac{\sigma_{xy}}{\sigma_x^2}\mu_x, \qquad \beta_2 = \frac{\sigma_{xy}}{\sigma_x^2}, \qquad \sigma_{y|x}^2 = \sigma_y^2 - \frac{\sigma_{xy}^2}{\sigma_x^2},$$

$$\omega_1 = \mu_x - \frac{\sigma_{xy}}{\sigma_y^2}\mu_y, \qquad \omega_2 = \frac{\sigma_{xy}}{\sigma_y^2}, \qquad \sigma_{x|y}^2 = \sigma_x^2 - \frac{\sigma_{xy}^2}{\sigma_y^2}.$$

Because $\boldsymbol{y}$ is sensitive, and $\boldsymbol{x}$ is not sensitive, the IM is (27), and thus the imputer generates synthetic $\boldsymbol{y}$-data using (4) with $\boldsymbol{x}_i = (1, x_i)'$. Specifically, in the case of single imputation, the released synthetic data are $\mathscr{D} = \{(v_i, x_i) : i = 1, \ldots, n\}$ where $(v_1, \ldots, v_n)$ are generated as in (4). In the case of multiple imputation, the released synthetic data are $\{\mathscr{D}_1, \ldots, \mathscr{D}_m\}$, where $m > 1$, $\mathscr{D}_j = \{(v_{ij}, x_i) : i = 1, \ldots, n\}$, and the vectors $(v_{11}, \ldots, v_{n1}), \ldots, (v_{1m}, \ldots, v_{nm})$ are generated by repeating (4) m times (independently). Now suppose that the AM is (28), and the data analyst's goal is to estimate ω_2. In the sequel, we present an empirical study to investigate properties of the analyst's inference for ω_2.

Single Imputation. Suppose that the singly imputed synthetic data $\mathscr{D} = \{(v_i, x_i) : i = 1, \ldots, n\}$ are released, and the data analyst, whose goal is to estimate ω_2, uses the methodology developed in Section 4 to draw inference. Although these methods were not designed for this scenario, because here the regressor variable is synthetic, and the response is original, our goal here is to examine what can happen if these methods are used in this situation. Therefore, we assume that the data analyst uses the inferential procedures of Section 4, but with the roles of $\boldsymbol{x}$ and $\boldsymbol{v}$ reversed. Applying Result 4.1 in this scenario, the data analyst's estimate of ω_2 is

$$\omega_2^* = \frac{\sum_{i=1}^n x_i(v_i - \bar{v})}{\sum_{i=1}^n (v_i - \bar{v})^2},$$

where $\bar{v} = n^{-1}\sum_{i=1}^n v_i$. Based on Result 4.2, the data analyst's estimate of $\mathrm{Var}(\omega_2^*)$ is

$$\widehat{\mathrm{Var}}(\omega_2^*) = 2\left(\frac{\sum_{i=1}^n (x_i - \omega_1^* - \omega_2^* v_i)^2}{n-2}\right)\left(\frac{1}{\sum_{i=1}^n (v_i - \bar{v})^2}\right),$$

where $\omega_1^* = \bar{x} - \omega_2^*\bar{v}$ and $\bar{x} = n^{-1}\sum_{i=1}^n x_i$. Furthermore, based on (14), the data analyst's $(1-\gamma)$ level confidence interval for ω_2 is

$$\left[\omega_2^* - \sqrt{\frac{\sum_{i=1}^n (x_i - \omega_1^* - \omega_2^* v_i)^2}{\sum_{i=1}^n (v_i - \bar{v})^2}}\sqrt{\delta_{1,n,2;\gamma}},\ \omega_2^* + \sqrt{\frac{\sum_{i=1}^n (x_i - \omega_1^* - \omega_2^* v_i)^2}{\sum_{i=1}^n (v_i - \bar{v})^2}}\sqrt{\delta_{1,n,2;\gamma}}\right]. \tag{29}$$

To evaluate the properties of τ_2^*, $\widehat{\mathrm{Var}}(\omega_2^*)$, and the confidence interval (29), we used Monte Carlo simulation to approximate the bias of τ_2^*, variance of τ_2^*, expected value of $\widehat{\mathrm{Var}}(\omega_2^*)$, coverage of the nominal 0.95 confidence interval computed using (29), and expected length of this confidence interval. For the Monte Carlo simulation, we used 10^6 iterations, and we fixed the parameters as

$$\mu_y = 0.1,\ \mu_x = 0.2,\ \sigma_y^2 = 1,\ \sigma_x^2 = 4,\ \sigma_{xy} = 1.6, \tag{30}$$

which then yield

$$\beta_1 = 0.02,\ \beta_2 = 0.40,\ \omega_1 = 0.04,\ \omega_2 = 1.6,\ \sigma_{y|x}^2 = 0.36,\ \sigma_{x|y}^2 = 1.44.$$

Simulation results when the sample size is n = 10, 30, 50, 100, 500, 1000, 2000, 4000, are displayed in the top panel of Table 12, under the heading $m = 1$. Based on this table, we observe that in the simulation scenarios, the estimator τ_2^* is approximately unbiased for τ^2, and the bias is getting closer to 0 as n increases. We also observe that the estimator $\widehat{\mathrm{Var}}(\omega_2^*)$ is positively biased for $\mathrm{Var}(\omega_2^*)$, for all chosen values of n, and the confidence interval coverage is slightly larger than the nominal level of 0.95.

Multiple Imputation. Suppose that the multiply imputed synthetic data $\{\mathscr{D}_1, \ldots, \mathscr{D}_m\}$ are released, and the data analyst using the methodology reviewed in Section 2, to draw inference on ω_2. Let the original data estimator of ω_2 be the usual least squares estimator

$$q = \hat{\omega}_2 = \frac{\sum_{i=1}^n x_i(y_i - \bar{y})}{\sum_{i=1}^n (y_i - \bar{y})^2},$$

along with the standard variance estimator

$$u = \widehat{\mathrm{Var}}(\hat{\omega}_2) = \left(\frac{\sum_{i=1}^n (x_i - \hat{\omega}_1 - \hat{\omega}_2 y_i)^2}{n-2}\right)\left(\frac{1}{\sum_{i=1}^n (y_i - \bar{y})^2}\right)$$

where $\hat{\omega}_1 = \bar{x} - \hat{\omega}_2\bar{y}$ and $\bar{y} = n^{-1}\sum_{i=1}^{n} y_i$. Thus the data analyst would compute the analogous quantities q_j and u_j based on each synthetic dataset $\mathscr{D}_j$, and use the combination formulas mentioned in Section 2 to get the point estimator $\bar{q}_m$ and corresponding variance estimator T_m, as well as confidence interval for ω_2 based on the t-distribution approximation. To evaluate the properties of $\bar{q}_m$, T_m, and the t-based confidence interval in this scenario, we used Monte Carlo simulation. For the simulation, we used the same parameters as in (30), once again using 10^6 iterations. We set the number of imputations $m = 10$. The lower panel of Table 12, under the heading $m = 10$, shows the Monte Carlo estimates of the bias of $\bar{q}_m$, variance of $\bar{q}_m$, expected value of T_m, coverage of the nominal 0.95 confidence interval, and expected length of this confidence interval. As in the single imputation scenario, results are shown for the sample sizes $n = 10$, 30, 50, 100, 500, 1000, 2000, 4000. Based on the results in Table 12, we observe that in the simulation scenarios, $\bar{q}_m$ is approximately unbiased for τ_2 (as in the single imputation case). We also observe in the simulation scenarios that as long as n is reasonably large, T_m is approximately unbiased for the variance of $\bar{q}_m$, and the confidence interval has coverage close to the nominal value of 0.95.

Table 12: Simulation results under the scenario where the data analyst wants to estimate the coefficient ω_2 in the regression of x on y.

m	n	Bias	Var	$\widehat{\text{Var}}$	Cvg	Len
1	10	8.454E-03	3.711E-01	4.848E-01	0.965	3.065
	30	4.748E-03	8.556E-02	1.121E-01	0.971	1.353
	50	3.052E-03	4.824E-02	6.309E-02	0.973	1.002
	100	1.455E-03	2.311E-02	3.010E-02	0.973	0.685
	500	3.294E-04	4.464E-03	5.810E-03	0.974	0.299
	1000	1.940E-04	2.223E-03	2.893E-03	0.974	0.211
	2000	1.075E-04	1.111E-03	1.443E-03	0.974	0.149
	4000	1.035E-04	5.549E-04	7.208E-04	0.975	0.105
10	10	7.978E-03	2.350E-01	2.578E-01	0.922	1.847
	30	4.636E-03	5.782E-02	5.914E-02	0.942	0.935
	50	3.082E-03	3.277E-02	3.327E-02	0.946	0.707
	100	1.507E-03	1.577E-02	1.587E-02	0.948	0.491
	500	3.828E-04	3.056E-03	3.062E-03	0.950	0.217
	1000	2.142E-04	1.523E-03	1.524E-03	0.950	0.153
	2000	9.301E-05	7.627E-04	7.605E-04	0.950	0.108
	4000	9.481E-05	3.797E-04	3.798E-04	0.950	0.076

In summary, in the simulation scenario, we find that if the data analyst applies the results of Section 4, with the roles of $\boldsymbol{x}$ and $\boldsymbol{v}$ reversed, the point estimate of the target parameter τ_2 is approximately unbiased, but the estimated variance tends to be too large, and as a result, the confidence interval has coverage greater than the nominal level. In our numerical studies, the nominal 0.95 confidence interval tends to have actual coverage of about 0.97. On the other hand, in the simulation scenario, multiple imputation also gives an approximately unbiased point estimate, as long as the sample size is not very small, the variance estimate is also approximately unbiased, and the confidence interval coverage is approximately equal to the nominal value of 0.95. Table 12 also indicates that in the simulation scenario, multiple imputation tends to yield a shorter confidence interval than single imputation, as expected.

7.3 Extensions of the Methodology to Other Scenarios

Only Part of y is Sensitive

In Section 4 we have assumed that all the n observations $\boldsymbol{y} = (y_1, \ldots, y_n)$ in the multiple linear regression model are sensitive. Of course, this need not be the case, and quite generally we can partition $\boldsymbol{y}$ into two parts: $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$ of dimensions r and $(n-r)$, respectively, and assume that the first r observations $\boldsymbol{y}_1$ are sensitive, thus requiring privacy protection, and the remaining $(n-r)$ observations $\boldsymbol{y}_2$ are non-sensitive, and can remain unprotected. Let $\boldsymbol{X} = [\boldsymbol{X}_1\, \boldsymbol{X}_2]$ be the corresponding partitioning of the matrix $\boldsymbol{X}$, so that $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are of dimensions $p \times r$ and $p \times (n-r)$, respectively. The reasons for some of the y-values being sensitive can vary depending on the context. For example, for income data, large incomes (extreme values) may be sensitive. The sensitive nature of y may also depend on the (extreme) values of the corresponding covariates $\boldsymbol{x}$. We outline below a data analysis procedure when the latter situation holds, namely, the sensitivity of the first r values of $\boldsymbol{y}$ is due to the nature of the covariates, which makes r a *non-random* integer. We assume that our interest lies in drawing valid inference about the regression coefficients $\boldsymbol{\beta}$.

We propose to synthesize the r sensitive y-values $\boldsymbol{y}_1$ by applying the plug-in sampling method based on these r y-values, as discussed in Section 4. The synthetic version of $\boldsymbol{y}_1$ is $\boldsymbol{y}_1^* = (y_1^*, \cdots, y_r^*)$, where $y_1^*, \cdots, y_r^*$ are generated independently such that $y_i \sim N(\boldsymbol{x}_i'\hat{\boldsymbol{\beta}}_1, \hat{\sigma}_1^2)$, $i = 1, \ldots, r$, where $\hat{\boldsymbol{\beta}}_1 = (\boldsymbol{X}_1\boldsymbol{X}_1')^{-1}\boldsymbol{X}_1\boldsymbol{y}_1$ and $\hat{\sigma}_1^2 = (\boldsymbol{y}_1 - \boldsymbol{X}_1'\hat{\boldsymbol{\beta}}_1)'(\boldsymbol{y}_1 - \boldsymbol{X}_1'\hat{\boldsymbol{\beta}}_1)/(r-p)$. We assume that $r > p$ and $(n-r) > p$ so that we can draw valid inference about the p regression coefficients $\boldsymbol{\beta}$ separately for each data set. We suggest a data analysis method combining $\boldsymbol{y}_1^*$ with the unperturbed remaining y-values $\boldsymbol{y}_2$ along the lines of statistical meta-analysis (Hartung et al. , 2008). Based on the synthesized part $\boldsymbol{y}_1^*$, we proceed as in Section 4 and observe from Result 4.1 that $\hat{\boldsymbol{\beta}}_1^* = (\boldsymbol{X}_1\boldsymbol{X}_1')^{-1}\boldsymbol{X}_1\boldsymbol{y}_1^*$ is an unbiased estimate of $\boldsymbol{\beta}$ with $\mathrm{Var}(\hat{\boldsymbol{\beta}}_1^*) = 2\sigma^2(\boldsymbol{X}_1\boldsymbol{X}_1')^{-1}$. On the other hand, an unbiased estimate of $\boldsymbol{\beta}$ based on the unperturbed second part of the data, namely $\boldsymbol{y}_2$, is given by $\hat{\boldsymbol{\beta}}_2 = (\boldsymbol{X}_2\boldsymbol{X}_2')^{-1}\boldsymbol{X}_2\boldsymbol{y}_2$ with $\mathrm{Var}(\hat{\boldsymbol{\beta}}_2) = \sigma^2(\boldsymbol{X}_2\boldsymbol{X}_2')^{-1}$. In view of independence of the two estimates of $\boldsymbol{\beta}$, a combined unbiased estimate of $\boldsymbol{\beta}$ is given by

$$\hat{\boldsymbol{\beta}}_{\mathrm{comb}} = \left[\frac{1}{2}(\boldsymbol{X}_1\boldsymbol{X}_1') + (\boldsymbol{X}_2\boldsymbol{X}_2')\right]^{-1} \left[\frac{1}{2}\boldsymbol{X}_1\boldsymbol{y}_1^* + \boldsymbol{X}_2\boldsymbol{y}_2\right]$$

with

$$\mathrm{Var}(\hat{\boldsymbol{\beta}}_{\mathrm{comb}}) = \sigma^2 \left[\frac{1}{2}(\boldsymbol{X}_1\boldsymbol{X}_1') + (\boldsymbol{X}_2\boldsymbol{X}_2')\right]^{-1}.$$

To test hypotheses about $\boldsymbol{\beta}$ and to construct a confidence set for $\boldsymbol{\beta}$ based on the two data sets $\boldsymbol{y}_1^*$ and $\boldsymbol{y}_2$, we define

$$T^2_{\text{comb}} = \frac{(\hat{\boldsymbol{\beta}}_{\text{comb}} - \boldsymbol{\beta})' \left[\frac{1}{2}(\boldsymbol{X}_1\boldsymbol{X}_1') + (\boldsymbol{X}_2\boldsymbol{X}_2')\right] (\hat{\boldsymbol{\beta}}_{\text{comb}} - \boldsymbol{\beta})}{\hat{\sigma}^2_{\text{comb}}}$$

where $\hat{\sigma}^2_{\text{comb}} = (r-p)\text{RSS}_1^* + \text{RSS}_2$, and $\text{RSS}_1^* = (\boldsymbol{y}_1^* - \boldsymbol{X}_1'\hat{\boldsymbol{\beta}}_1^*)'(\boldsymbol{y}_1^* - \boldsymbol{X}_1'\hat{\boldsymbol{\beta}}_1^*)$ and $\text{RSS}_2 = (\boldsymbol{y}_2 - \boldsymbol{X}_2'\hat{\boldsymbol{\beta}}_2)'(\boldsymbol{y}_2 - \boldsymbol{X}_2'\hat{\boldsymbol{\beta}}_2)$ are the standard residual sum of squares based on the two data sets $\boldsymbol{y}_1^*$, $\boldsymbol{y}_2$, upon fitting multiple linear regression on their respective covariates. We demonstrate below that T^2_{comb} is a pivot. This is in the same spirit as in Theorem 4.3 in Section 4.

To derive the distribution of T^2_{comb}, note from Theorem 4.1 that, conditionally given ψ, $\hat{\boldsymbol{\beta}}_{\text{comb}} \sim N_p[\boldsymbol{\beta}, \sigma^2\boldsymbol{\Sigma}(\psi)]$ where

$$\boldsymbol{\Sigma}(\psi) = \left[\frac{1}{2}(\boldsymbol{X}_1\boldsymbol{X}_1') + (\boldsymbol{X}_2\boldsymbol{X}_2')\right]^{-1} \left[\frac{1}{2}(\boldsymbol{X}_1\boldsymbol{X}_1')\left(1 + \frac{\psi}{r-p}\right) + (\boldsymbol{X}_2\boldsymbol{X}_2')\right] \left[\frac{1}{2}(\boldsymbol{X}_1\boldsymbol{X}_1') + (\boldsymbol{X}_2\boldsymbol{X}_2')\right]^{-1}.$$

Hence, conditionally given ψ, the numerator of T^2_{comb} is distributed as $\sigma^2\sum_{i=1}^p \lambda_i\chi^2_{1i}$ where χ^2_{1i}'s are independent χ^2 variables each with 1 degree of freedom, and λ_i's are the roots of the matrix $\boldsymbol{A}(\psi)$ where

$$\boldsymbol{A}(\psi) = \left[\frac{1}{2}(\boldsymbol{X}_1\boldsymbol{X}_1')\left(1 + \frac{\psi}{r-p}\right) + (\boldsymbol{X}_2\boldsymbol{X}_2')\right] \left[\frac{1}{2}(\boldsymbol{X}_1\boldsymbol{X}_1') + (\boldsymbol{X}_2\boldsymbol{X}_2')\right]^{-1}.$$

The above roots can be expressed as $\lambda_i = (1 + \frac{\psi}{r-p})\lambda_{1i} + \lambda_{2i}$, where λ_{1i}'s and λ_{2i}'s are the roots of $\frac{1}{2}(\boldsymbol{X}_1\boldsymbol{X}_1')[\frac{1}{2}(\boldsymbol{X}_1\boldsymbol{X}_1') + (\boldsymbol{X}_2\boldsymbol{X}_2')]^{-1}$ and $(\boldsymbol{X}_2\boldsymbol{X}_2')[\frac{1}{2}(\boldsymbol{X}_1\boldsymbol{X}_1') + (\boldsymbol{X}_2\boldsymbol{X}_2')]^{-1}$, respectively.

On the other hand, again from Theorem 4.1, conditionally given ψ, $\frac{(r-p)\text{RSS}_1^*}{\sigma^2\psi} \sim \chi^2_{r-p}$, independent of $\text{RSS}_2 \sim \sigma^2\chi^2_{n-r-p}$. Hence, conditionally given ψ, $[(r-p)\text{RSS}_1^* + \text{RSS}_2] \sim \sigma^2[\psi\chi^2_{r-p} + \chi^2_{n-r-p}]$, and this variable is independent of $\hat{\boldsymbol{\beta}}_{\text{comb}}$. It then follows that the conditional distribution of T^2_{comb}, given ψ, can be expressed as $\frac{\sum_{i=1}^p \chi^2_{1i}[(1+\frac{\psi}{r-p})\lambda_{1i} + \lambda_{2i}]}{\psi\chi^2_{r-p} + \chi^2_{n-r-p}}$ where all the χ^2 variables are independent, and, using an argument similar to that used in Theorem 4.3, the marginal distribution of ψ is given by $f_{r,p}(\psi) \propto e^{-\psi/2}\psi^{\frac{r-p}{2}-1}$. That T^2_{comb} is a pivot is clear, and it is indeed possible to easily determine the cut-off points of T^2_{comb}. We omit the details.

Inference about $\boldsymbol{\eta} = \boldsymbol{A}\boldsymbol{\beta}$. Assume $\boldsymbol{A}$ is of dimension $k \times p$ with $\text{rank}(\boldsymbol{A}) = k < p$. From the preceding arguments, it follows that, conditionally given ψ, $\hat{\boldsymbol{\eta}}_{\text{comb}} = \boldsymbol{A}\hat{\boldsymbol{\beta}}_{\text{comb}} \sim N_k[\boldsymbol{\eta}, \sigma^2\boldsymbol{A}\Sigma(\psi)\boldsymbol{A}']$, and, marginally, $\text{Var}(\hat{\boldsymbol{\eta}}_{\text{comb}}) = \sigma^2\boldsymbol{A}[\frac{1}{2}(\boldsymbol{X}_1\boldsymbol{X}_1') + (\boldsymbol{X}_2\boldsymbol{X}_2')]^{-1}\boldsymbol{A}'$. Therefore, defining

$$T^2_{\boldsymbol{\eta},\text{comb}} = \frac{(\hat{\boldsymbol{\eta}}_{\text{comb}} - \boldsymbol{\eta})'\{\boldsymbol{A}[\frac{1}{2}(\boldsymbol{X}_1\boldsymbol{X}_1') + (\boldsymbol{X}_2\boldsymbol{X}_2')]^{-1}\boldsymbol{A}'\}^{-1}(\hat{\boldsymbol{\eta}}_{\text{comb}} - \boldsymbol{\eta})}{\hat{\sigma}^2_{\text{comb}}},$$

it follows from the previous arguments that the conditional distribution of $T^2_{\boldsymbol{\eta},\text{comb}}$, given ψ, can be expressed as $\frac{\sum_{i=1}^{k}\chi^2_{1i}[(1+\frac{\psi}{r-p})\lambda^*_{1i}+\lambda^*_{2i}]}{\psi\chi^2_{r-p}+\chi^2_{n-r-p}}$ where λ^*_{1i}'s and λ^*_{2i}'s are the roots of $\frac{1}{2}(\boldsymbol{X}_1\boldsymbol{X}_1')[\frac{1}{2}(\boldsymbol{X}_1\boldsymbol{X}_1')+(\boldsymbol{X}_2\boldsymbol{X}_2')]^{-1}\boldsymbol{A}'\{\boldsymbol{A}[\frac{1}{2}(\boldsymbol{X}_1\boldsymbol{X}_1')+(\boldsymbol{X}_2\boldsymbol{X}_2')]^{-1}\boldsymbol{A}'\}^{-1}\boldsymbol{A}[\frac{1}{2}(\boldsymbol{X}_1\boldsymbol{X}_1')+(\boldsymbol{X}_2\boldsymbol{X}_2')]^{-1}$ and $(\boldsymbol{X}_2\boldsymbol{X}_2')[\frac{1}{2}(\boldsymbol{X}_1\boldsymbol{X}_1')+(\boldsymbol{X}_2\boldsymbol{X}_2')]^{-1}\boldsymbol{A}'\{\boldsymbol{A}[\frac{1}{2}(\boldsymbol{X}_1\boldsymbol{X}_1')+(\boldsymbol{X}_2\boldsymbol{X}_2')]^{-1}\boldsymbol{A}'\}^{-1}\boldsymbol{A}[\frac{1}{2}(\boldsymbol{X}_1\boldsymbol{X}_1')+(\boldsymbol{X}_2\boldsymbol{X}_2')]^{-1}$, respectively. As noted before, here all the χ^2 variables are independent, and the marginal distribution of ψ is again given by $f_{r,p}(\psi) \propto e^{-\psi/2}\psi^{\frac{r-p}{2}-1}$. That $T^2_{\boldsymbol{\eta},\text{comb}}$ is a pivot is clear, and it is indeed possible to easily determine the cut-off points of $T^2_{\boldsymbol{\eta},\text{comb}}$ to draw inference about $\boldsymbol{\eta}$. We omit the details.

Response and Covariates are all Sensitive

Under the same linear regression model as in Section 4, we now assume that the covariates $\boldsymbol{x}$ are also sensitive along with the primary response variable y, so that the entire data set $\{(y_i,\boldsymbol{x}_i) : i = 1,\ldots,n\}$ needs to be privacy protected. We discuss exact inference about the regression coefficients $\boldsymbol{\beta}$ based on a singly imputed plug-in synthetic data under the assumption of a multivariate normal distribution of the $(p+1)$-dimensional vector $(y,\boldsymbol{x})$.

$$\text{Assume} \begin{pmatrix} y_1 \\ x_{11} \\ \vdots \\ x_{p1} \end{pmatrix}, \ldots, \begin{pmatrix} y_n \\ x_{1n} \\ \vdots \\ x_{pn} \end{pmatrix}, n > p+1, \text{ are } iid \text{ as } N_{p+1}\left[\boldsymbol{\mu} = \begin{pmatrix} \mu_y \\ \boldsymbol{\mu}_x \end{pmatrix}, \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_{yy} & \boldsymbol{\sigma}'_{y\boldsymbol{x}} \\ \boldsymbol{\sigma}_{y\boldsymbol{x}} & \boldsymbol{\Sigma}_{\boldsymbol{xx}} \end{pmatrix}\right].$$

Define $\hat{\mu}_y = \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$, $\hat{\boldsymbol{\mu}}_x = \bar{\boldsymbol{x}} = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i$, $\hat{\boldsymbol{\Sigma}} = \mathscr{S}/(n-1)$ where $\boldsymbol{x}_i = \begin{pmatrix} x_{1i} \\ \vdots \\ x_{pi} \end{pmatrix}$, $\mathscr{S} = \begin{pmatrix} \mathscr{S}_{yy} & \mathscr{S}'_{y\boldsymbol{x}} \\ \mathscr{S}_{y\boldsymbol{x}} & \mathscr{S}_{\boldsymbol{xx}} \end{pmatrix}$ and $\mathscr{S}_{yy} = \sum_{i=1}^{n}(y_i-\bar{y})^2$, $\mathscr{S}'_{y\boldsymbol{x}} = (\mathscr{S}_{y1},\ldots,\mathscr{S}_{yp})$ with $\mathscr{S}_{yj} = \sum_{i=1}^{n}(y_i-\bar{y})(x_{ji}-\bar{x}_j)$, $j = 1,\cdots,p$, and $\mathscr{S}_{\boldsymbol{xx}} = \sum_{i=1}^{n}(\boldsymbol{x}_i-\bar{\boldsymbol{x}})(\boldsymbol{x}_i-\bar{\boldsymbol{x}})'$ is the sample Wishart matrix based on the $\boldsymbol{x}$-data. Obviously, $(\hat{\boldsymbol{\mu}},\hat{\boldsymbol{\Sigma}})$ are jointly sufficient for $(\boldsymbol{\mu},\boldsymbol{\Sigma})$. The central parameter of interest in this context is $\boldsymbol{\beta}$, the $p \times 1$ dimensional vector of regression coefficients of y on $\boldsymbol{x}$, defined by

$$\boldsymbol{\beta} = \boldsymbol{\Sigma}^{-1}_{\boldsymbol{xx}}\boldsymbol{\sigma}_{y\boldsymbol{x}}.$$

The standard inference for $\boldsymbol{\beta}$ based on the original data $\{(y_1,\boldsymbol{x}_1),\ldots,(y_n,\boldsymbol{x}_n)\}$ uses the following well known facts (Anderson, 2003).

1. $\hat{\boldsymbol{\beta}} = \boldsymbol{b} = \mathscr{S}^{-1}_{\boldsymbol{xx}}\mathscr{S}_{\boldsymbol{yx}}$ is an unbiased estimate of $\boldsymbol{\beta}$.
2. $\boldsymbol{b}|\mathscr{S}_{\boldsymbol{xx}} \sim N_p[\boldsymbol{\beta},\sigma_{yy.\boldsymbol{x}}\mathscr{S}^{-1}_{\boldsymbol{xx}}]$ where $\sigma_{yy.\boldsymbol{x}} = \sigma_{yy} - \boldsymbol{\sigma}'_{y\boldsymbol{x}}\boldsymbol{\Sigma}_{\boldsymbol{xx}}\boldsymbol{\sigma}_{y\boldsymbol{x}}$.
3. $\mathscr{S}_{\boldsymbol{xx}} \sim W_p(\boldsymbol{\Sigma}_{\boldsymbol{xx}}, n-1)$
4. $\mathscr{S}_{yy.\boldsymbol{x}} = \mathscr{S}_{yy} - \mathscr{S}'_{y\boldsymbol{x}}\mathscr{S}^{-1}_{\boldsymbol{xx}}\mathscr{S}_{y\boldsymbol{x}} \sim \sigma_{yy.\boldsymbol{x}}\chi^2_{n-p-1}$, independent of $\boldsymbol{b}$ and $\mathscr{S}_{\boldsymbol{xx}}$.

Based on the above distributional properties, a confidence interval for $\boldsymbol{\beta}$ is readily obtained from the fact that

$$F = \frac{p}{n-p-1}\left[\frac{(\boldsymbol{b}-\boldsymbol{\beta})'\mathscr{S}_{\boldsymbol{xx}}(\boldsymbol{b}-\boldsymbol{\beta})}{\mathscr{S}_{yy.\boldsymbol{x}}}\right] \sim F_{p,n-p-1}.$$

Following the approach of Section 3, we now develop the likelihood-based approach for analysis of singly imputed synthetic data generated under the plug-in sampling method. Under this method, the synthetic data, denoted by $(\boldsymbol{u},\boldsymbol{v}) = \{(u_1,\boldsymbol{v}_1),\ldots(u_n,\boldsymbol{v}_n)\}$ are obtained by drawing *iid* samples from $N_{p+1}[(\bar{y},\bar{\boldsymbol{x}})',\mathscr{S}/(n-1)]$. Analogous to the statistics defined with the original data, let $\bar{u} = \frac{1}{n}\sum_{i=1}^{n} u_i$ (sample mean based on u-values), $031\boldsymbol{v} = (1\text{''}7016\bar{v}_1,\cdots,1\text{''}7016v_p)$ where $\bar{v}_i = \frac{1}{n}\sum_{j=1}^{n} v_{ij},\ i = 1,\cdots,p$, and $\tilde{\mathscr{S}} = \begin{pmatrix}\mathscr{S}_{uu} & \mathscr{S}'_{u\boldsymbol{v}}\\ \mathscr{S}_{u\boldsymbol{v}} & \mathscr{S}_{\boldsymbol{vv}}\end{pmatrix}$ where $\mathscr{S}_{uu} = \sum_{i=1}^{n}(u_i-\bar{u})^2$, $\mathscr{S}_{u\boldsymbol{v}} = (\mathscr{S}_{u1},\cdots,\mathscr{S}_{up})$ with $\mathscr{S}_{uj} = \sum_{i=1}^{n}(u_i-\bar{u})(v_{ji}-\bar{v}_j)$, and $\mathscr{S}_{\boldsymbol{vv}} = \sum_{i=1}^{n}(\boldsymbol{v}_i-\bar{\boldsymbol{v}})(\boldsymbol{v}_i-\bar{\boldsymbol{v}})'$.

It then follows from a general observation that $(\bar{u},\bar{\boldsymbol{v}})$ and $\tilde{\mathscr{S}}$ are jointly sufficient for $(\boldsymbol{\mu},\boldsymbol{\Sigma})$. To derive the main *inferential* results for $\boldsymbol{\beta}$, we define (analogous to $\boldsymbol{b}$)

$$\tilde{\boldsymbol{b}} = \mathscr{S}_{\boldsymbol{vv}}^{-1}\mathscr{S}_{u\boldsymbol{v}},\ \ \mathscr{S}_{uu.\boldsymbol{v}} = \mathscr{S}_{uu} - \mathscr{S}'_{u\boldsymbol{v}}\mathscr{S}_{\boldsymbol{vv}}^{-1}\mathscr{S}_{u\boldsymbol{v}}$$

and observe the following facts:

1. $\tilde{\boldsymbol{b}}|\mathscr{S},\mathscr{S}_{vv} \sim N_p[\boldsymbol{\beta},\mathscr{S}_{vv}^{-1}s_{yy.\boldsymbol{x}} + \mathscr{S}_{\boldsymbol{xx}}^{-1}\sigma_{yy.\boldsymbol{x}}]$, where $s_{yy.\boldsymbol{x}} = \mathscr{S}_{yy.\boldsymbol{x}}/(n-1)$.
2. $\mathscr{S}_{vv}|\mathscr{S} \sim W_p(\frac{\mathscr{S}_{xx}}{n-1}, n-1)$
3. $\mathscr{S}_{uu.\boldsymbol{v}}|\mathscr{S} \sim s_{yy.\boldsymbol{x}}\chi^2_{n-p-1}$, independent of $\tilde{\boldsymbol{b}},\mathscr{S}_{vv}$.

Facts 2 and 3 are obvious (Anderson, 2003), and a proof of the first fact appears at the end of this section.

To carry out inference for $\boldsymbol{\beta}$, we proceed as we normally do with the original multivariate data $\{(y_1,\boldsymbol{x}_1),\ldots,(y_n,\boldsymbol{x}_n)\}$, and propose

$$\tilde{T}^2_{\boldsymbol{\beta}} = \frac{(\tilde{\boldsymbol{b}}-\boldsymbol{\beta})'\mathscr{S}_{vv}(\tilde{\boldsymbol{b}}-\boldsymbol{\beta})}{\mathscr{S}_{uu.\boldsymbol{v}}}$$

and demonstrate that this is a *pivot*. We also provide enough computational details to easily simulate the distribution of $\tilde{T}^2_{\boldsymbol{\beta}}$.

Proof. Observe from the above that, conditionally given $\mathscr{S}$ and $\mathscr{S}_{vv}$,

$$(\tilde{\boldsymbol{b}}-\boldsymbol{\beta})'\mathscr{S}_{vv}(\tilde{\boldsymbol{b}}-\boldsymbol{\beta}) = \sum_{i=1}^{p} c_i\chi^2_{1i},$$

where χ^2_{1i}'s are independent chi-squared random variables each with 1 degree of freedom, and independent of $c_1,\cdots,,c_p$, which are the roots of $\mathscr{Q} = \mathscr{S}_{\boldsymbol{vv}}[s_{yy.\boldsymbol{x}}\mathscr{S}_{\boldsymbol{vv}}^{-1} + \sigma_{yy.\boldsymbol{x}}\mathscr{S}_{\boldsymbol{xx}}^{-1}] = s_{yy.\boldsymbol{x}}\boldsymbol{I}_p + \sigma_{yy.\boldsymbol{x}}\mathscr{S}_{\boldsymbol{vv}}\mathscr{S}_{\boldsymbol{xx}}^{-1}$.

We now argue as follows. Since $\mathscr{S}_{\boldsymbol{vv}}|\mathscr{S} \sim W_p(\frac{\mathscr{S}_{\boldsymbol{xx}}}{n-1}, n-1)$, the roots of $\mathscr{S}_{\boldsymbol{vv}}\mathscr{S}_{\boldsymbol{xx}}^{-1}$ are the same as the roots of $W_p(\frac{\boldsymbol{I}_p}{n-1}, n-1)$, and let us denote them by $\lambda_1, \cdots, \lambda_p$, implying $c_i = s_{yy.\boldsymbol{x}} + \sigma_{yy.\boldsymbol{x}}\lambda_i$, $i = 1, \cdots, p$. We next note that

$$\frac{c_i}{\mathscr{S}_{uu.\boldsymbol{v}}} = \frac{s_{yy.\boldsymbol{x}}}{\mathscr{S}_{uu.\boldsymbol{v}}} + \frac{\sigma_{yy.\boldsymbol{x}}}{\mathscr{S}_{yy.\boldsymbol{x}}}\frac{\mathscr{S}_{yy.\boldsymbol{x}}}{\mathscr{S}_{uu.\boldsymbol{v}}}\lambda_i.$$

We finally observe that (i) $\mathscr{S}_{uu.\boldsymbol{v}}/s_{yy.\boldsymbol{x}}|\mathscr{S}_{yy.\boldsymbol{x}} \sim \chi^2_{n-p-1}$, independent of $\mathscr{S}_{yy.\boldsymbol{x}}$, and (ii) $\mathscr{S}_{yy.\boldsymbol{x}}/\sigma_{yy.\boldsymbol{x}} \sim \chi^2_{n-p-1}$. Hence, writing $\tilde{T}^2_{\boldsymbol{\beta}} = \sum_{i=1}^p d_i\chi^2_{1i}$, we find that

$$\boldsymbol{d} = \frac{\boldsymbol{c}}{\mathscr{S}_{uu.\boldsymbol{v}}} \sim \frac{1}{\chi^2_{n-p-1}}\left[\boldsymbol{1} + \frac{(n-p-1)\boldsymbol{\lambda}}{\chi^2_{n-p-1}}\right] \implies \tilde{T}^2_{\boldsymbol{\beta}} = \frac{1}{\chi^2_{n-p-1}}\left[\chi^2_p + \frac{(n-p-1)\sum_{i=1}^p \lambda_i\chi^2_{1i}}{\chi^2_{n-p-1}}\right]$$

where all the chi-squares are independent. That $\tilde{T}^2_{\boldsymbol{\beta}}$ is a pivot follows directly.

Remark 7.1. Here are the steps to generate the values of $\tilde{T}^2_{\boldsymbol{\beta}}$.

1. Generate independent χ^2 variables each with 1 degree of freedom.
2. Generate roots λ^* of $W_p(\boldsymbol{I}_p, n-1)$.
3. Generate independently $A \sim \chi^2_p$, $B \sim \chi^2_{n-p-1}$, and $C \sim \chi^2_{n-p-1}$.
4. Generate values of $\tilde{T}^2_{\boldsymbol{\beta}}$ using $\tilde{T}^2_{\boldsymbol{\beta}} = \frac{1}{B}[A + (\frac{n-p-1}{n-1})\frac{\sum_{i=1}^p \lambda_i^*\chi^2_{1i}}{C}]$.

Remark 7.2. Inference about the residual variance $\sigma_{yy.\boldsymbol{x}}$ can be easily carried out by defining $V = \mathscr{S}_{uu.\boldsymbol{v}}/\sigma_{yy.\boldsymbol{x}}$ and noting that V is distributed as $[\frac{\mathscr{S}_{uu.\boldsymbol{v}}}{\mathscr{S}_{yy.\boldsymbol{x}}}][\frac{\mathscr{S}_{yy.\boldsymbol{x}}}{\sigma_{yy.\boldsymbol{x}}}] \sim \chi^2_{n-p-1} \times \chi^2_{n-p-1}/(n-1)$, and the two χ^2's are independent.

Proof of $\tilde{\boldsymbol{b}}|\mathscr{S}, \mathscr{S}_{vv} \sim N_p[\boldsymbol{\beta}, \mathscr{S}_{vv}^{-1}s_{yy.\boldsymbol{x}} + \mathscr{S}_{xx}^{-1}\sigma_{yy.\boldsymbol{x}}]$. The proof, which is analogous to that of Theorem 4.1, proceeds in several steps.

Step 1. The conditional joint pdf of $\tilde{\boldsymbol{b}}$, $\mathscr{S}_{vv}$ and $\mathscr{S}_{uu.\boldsymbol{v}}$, given $\mathscr{S}$, is the product of three terms, namely, $A \times B \times C$, where

A is the conditional normal pdf of $\tilde{\boldsymbol{b}}$, given by $\propto \exp[-\frac{1}{2}\{\frac{(\tilde{\boldsymbol{b}}-\boldsymbol{b})'\mathscr{S}_{vv}(\tilde{\boldsymbol{b}}-\boldsymbol{b})}{s_{yy.\boldsymbol{x}}}\}]\frac{|\mathscr{S}_{vv}|^{\frac{1}{2}}}{s_{yy.\boldsymbol{x}}^{\frac{p}{2}}}$;

B is the conditional Wishart pdf of $\mathscr{S}_{\boldsymbol{vv}}$, given by $\propto \exp[-\frac{n-1}{2}\mathrm{tr}(\mathscr{S}_{xx}^{-1}\mathscr{S}_{vv})]\frac{|\mathscr{S}_{vv}|^{\frac{n-p-2}{2}}}{|\mathscr{S}_{xx}|^{\frac{n-1}{2}}}$;

C is the conditional χ^2_{n-p-1} pdf of $\mathscr{S}_{uu.\boldsymbol{v}}$, given by $\propto \exp[-\frac{1}{2}\frac{\mathscr{S}_{uu.\boldsymbol{v}}}{s_{yy.\boldsymbol{x}}}]\frac{\mathscr{S}_{uu.\boldsymbol{v}}^{\frac{n-p-1}{2}-1}}{s_{yy.\boldsymbol{x}}^{\frac{n-p-1}{2}}}$.

Step 2. The joint pdf of $(\boldsymbol{b}, \mathscr{S}_{\boldsymbol{xx}}, S_{yy.\boldsymbol{x}})$ is the product of three terms, namely, $D \times E \times F$, where

D is the conditional normal pdf of $\boldsymbol{b}$, given $\mathscr{S}_{\boldsymbol{xx}}$, given by $\propto \exp[-\frac{1}{2}\frac{(\boldsymbol{b}-\boldsymbol{\beta})'\mathscr{S}_{\boldsymbol{xx}}(\boldsymbol{b}-\boldsymbol{\beta})}{\sigma_{yy.\boldsymbol{x}}}]|\mathscr{S}_{\boldsymbol{xx}}|^{\frac{p}{2}}$;

E is the marginal Wishart pdf of $\mathscr{S}_{\boldsymbol{xx}}$, given by $\propto \exp[-\frac{1}{2}\mathrm{tr}\left(\boldsymbol{\Sigma}_{\boldsymbol{xx}}^{-1}\mathscr{S}_{\boldsymbol{xx}}\right)]|\mathscr{S}_{\boldsymbol{xx}}|^{\frac{n-p-2}{2}}$;

F is the marginal χ^2_{n-p-1} pdf of $\mathscr{S}_{yy.\boldsymbol{x}}$, independent of $\boldsymbol{b}$ and $\mathscr{S}_{\boldsymbol{xx}}$, given by $\propto$ $\exp[-\frac{1}{2}\frac{\mathscr{S}_{yy.\boldsymbol{x}}}{\sigma_{yy.\boldsymbol{x}}}]\mathscr{S}_{yy.\boldsymbol{x}}^{\frac{n-p-1}{2}-1}$.

Step 3. Now proceeding as in the proof of Theorem 4.1, we combine the terms involving $\boldsymbol{b}$ and integrate it out to get the conditional pdf of $\tilde{\boldsymbol{b}}$, given $\mathscr{S}$ and $\mathscr{S}_{vv}$, as $N_p[\boldsymbol{\beta}, \mathscr{S}_{vv}^{-1}s_{yy.\boldsymbol{x}} + \mathscr{S}_{xx}^{-1}\sigma_{yy.\boldsymbol{x}}]$. This completes the proof.

8 Discussion

In this paper, we developed new likelihood-based methods for drawing valid inference based on a singly imputed partially synthetic dataset, generated via plug-in sampling, under a multivariate normal as well as a multiple linear regression model. In these two cases, namely, multivariate normal and multiple linear regression, the methodology presented here allows one to draw valid inference when only a single partially synthetic dataset is available. The simulation studies presented in Section 5 illustrate that these methods perform just as our theory predicts. It should be noted that the methodology developed here is model based, and thus it does not immediately generalize to cases that do not fall under the multivariate normal or multiple linear regression models. In other cases, such as when there are a mixture of continuous and categorical variables, it may very well be possible to derive analogous likelihood based methods for analyzing singly imputed partially synthetic data, and we hope to pursue this problem in future work.

Kinney et al. (2011) mention that a singly imputed version of the Synthetic LBD is released, as opposed to multiply imputed versions, due to concerns about disclosure risk. Similarly, Hawala (2008) mentions that for American Community Survey Group Quarters data, the released data are based on singly imputed synthetic data, not multiply imputed synthetic data, because of disclosure risk concerns. Intuitively, it would appear that as m, the number of synthetic datasets, increases, the disclosure risk would also increase. In Subsection 6.2 we confirmed this statement, through a disclosure risk evaluation in the context of Current Population Survey (CPS) data. On the other hand, intuitively it would also appear that as m increases, inference derived from the synthetic data would become more efficient. We confirmed this statement too through the simulation studies in Section 5 and the CPS data analysis in Subsection 6.1.

In deriving the methodology of Sections 3 and 4, we have made assumptions about the process that generated the original data, and about the mechanism used to create synthetic data. Indeed, these assumptions are used to derive the likelihood-based inference for singly imputed synthetic data. In Subsection 7.1 we discussed the practical implications of these assumptions (in the case of the linear regression model), which yield a set of conditions under which the proposed methodology is expected to yield valid inference. Subsection 7.2 explored the performance of our methodology when some of the conditions do not hold (i.e., scenarios where the imputer and/or data analyst overfit or underfit the regression model, and a scenario where the imputer's model is the regression of y on x, but the data analyst's model is the regression of x on y). We leave it as future work to develop likelihood-based methodology that provides valid inference for singly imputed synthetic data in some important scenarios that do not fall under the conditions of Subsection 7.1, such as the following: only some records of the variable y are synthesized (an approach is outlined in Subsection 7.3); the response variable and the regressor variables are all synthesized (an approach is outlined in Subsection 7.3); multiple y-variables are synthesized, while multiple x-variables are not; the original data are from a census, not a sample; and the original data contain missing observations. We should mention that in case of multiple imputation, methodology for handling the preceding scenarios is available. The multiple imputation methodology of Reiter (2003) covers partially synthetic data scenarios where all or some variables and/or records in the database are synthesized. Furthermore, An and Little (2007) present specific methodology for using multiple imputation when only large values of a variable are sensitive; Drechsler and Reiter (2010) present methodology for generating and analyzing synthetic data when the original data are from a census; and Reiter (2004) presents methodology for simultaneously handling missing data and synthetic data. A similar line of research for singly imputed synthetic data can be developed.

Acknowledgments

The authors thank three anonymous referees and the Editor for providing many helpful comments that enhanced the quality of the paper; Paul Massell for reviewing a previous version of the manuscript; and Laura McKenna, Joseph Schafer, Eric Slud, Yves Thibaudeau, William Winkler, and Tommy Wright for encouragement.

Appendices

1 Proofs of the Theorems

Proof of Theorem 3.1. Recall $\boldsymbol{W} = \mathscr{S}_x = (n-1)\hat{\boldsymbol{\Sigma}}$. The proof proceeds in several steps.

1. The conditional joint pdf of $(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)$, given $(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}})$, is
$$\propto \exp\left[-\frac{n(n-1)}{2}(\bar{\boldsymbol{y}} - \bar{\boldsymbol{x}})'\boldsymbol{W}^{-1}(\bar{\boldsymbol{y}} - \bar{\boldsymbol{x}}) + \frac{n-1}{2}\mathscr{S}_y\boldsymbol{W}^{-1}\right] \times |\boldsymbol{W}|^{-n/2}.$$

2. The joint pdf of $(\bar{\boldsymbol{x}}, \mathscr{S}_x)$ is
$$\propto \exp\left[-\frac{n}{2}(\bar{\boldsymbol{x}} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\bar{\boldsymbol{x}} - \boldsymbol{\mu}) + \frac{1}{2}\mathscr{S}_x\boldsymbol{\Sigma}^{-1}\right] \times |\boldsymbol{\Sigma}|^{-n/2} \times |\mathscr{S}_x|^{\frac{n-p-2}{2}}.$$

3. We now combine the terms involving $\bar{\boldsymbol{x}}$ from the two exponents as
$$\begin{aligned}
&(n-1)(\bar{\boldsymbol{y}} - \bar{\boldsymbol{x}})'\boldsymbol{W}^{-1}(\bar{\boldsymbol{y}} - \bar{\boldsymbol{x}}) + (\bar{\boldsymbol{x}} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\bar{\boldsymbol{x}} - \boldsymbol{\mu}) \\
&= \left\{\bar{\boldsymbol{x}} - [(n-1)\boldsymbol{W}^{-1} + \boldsymbol{\Sigma}^{-1}]^{-1}[(n-1)\boldsymbol{W}^{-1}\bar{\boldsymbol{y}} + \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}]\right\}' \left\{(n-1)\boldsymbol{W}^{-1} + \boldsymbol{\Sigma}^{-1}\right\} \\
&\times \left\{\bar{\boldsymbol{x}} - [(n-1)\boldsymbol{W}^{-1} + \boldsymbol{\Sigma}^{-1}]^{-1}[(n-1)\boldsymbol{W}^{-1}\bar{\boldsymbol{y}} + \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}]\right\} \\
&+ (n-1)\bar{\boldsymbol{y}}'\boldsymbol{W}^{-1}\bar{\boldsymbol{y}} + \boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} \\
&- \{(n-1)\boldsymbol{W}^{-1}\bar{\boldsymbol{y}} + \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\}'\{(n-1)\boldsymbol{W}^{-1} + \boldsymbol{\Sigma}^{-1}\}^{-1}\{(n-1)\boldsymbol{W}^{-1}\bar{\boldsymbol{y}} + \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\}.
\end{aligned}$$

4. Using the fact that for any two positive definite matrices $\boldsymbol{A}$ and $\boldsymbol{B}$, $\boldsymbol{A}^{-1} - \boldsymbol{A}^{-1}(\boldsymbol{A}^{-1} + \boldsymbol{B}^{-1})^{-1}\boldsymbol{A}^{-1} = \boldsymbol{A}^{-1}(\boldsymbol{A}^{-1} + \boldsymbol{B}^{-1})^{-1}\boldsymbol{B}^{-1}$, the last term in the above expression can be simplified as $(\bar{\boldsymbol{y}} - \boldsymbol{\mu})'(\boldsymbol{\Sigma} + \frac{\boldsymbol{W}}{n-1})^{-1}(\bar{\boldsymbol{y}} - \boldsymbol{\mu})$.

5. Now integrating out $\bar{\boldsymbol{x}}$, the conditional joint pdf of $(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)$, given $\hat{\boldsymbol{\Sigma}}$, is obtained as
$$\propto \exp\left[-\frac{n}{2}(\bar{\boldsymbol{y}} - \boldsymbol{\mu})'\left(\boldsymbol{\Sigma} + \frac{\boldsymbol{W}}{n-1}\right)^{-1}(\bar{\boldsymbol{y}} - \boldsymbol{\mu}) + \frac{n-1}{2}\mathscr{S}_y\boldsymbol{W}^{-1}\right].$$

6. The unconditional joint pdf of $(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)$, which is the required likelihood to carry out subsequent inference for $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, can then be derived by integrating out $\boldsymbol{W}$ from the joint pdf of $(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n, \boldsymbol{W})$.

We note from the above that the MLE of $\boldsymbol{\mu}$ is $\bar{\boldsymbol{y}}$, and, conditional on $\boldsymbol{W}$, $\bar{\boldsymbol{y}} \sim N[\boldsymbol{\mu}, \frac{\boldsymbol{\Sigma}}{n} + \frac{\boldsymbol{W}}{n(n-1)}]$ and $\mathscr{S}_y \sim \text{Wishart}_p(\frac{\boldsymbol{W}}{n-1}, n-1)$, independent of $\bar{\boldsymbol{y}}$. We are now in a position to derive the distribution of T^2. We first express T^2 as
$$T^2 = \left[\frac{(\bar{\boldsymbol{y}} - \boldsymbol{\mu})'\mathscr{S}_y^{-1}(\bar{\boldsymbol{y}} - \boldsymbol{\mu})}{(n-1)(\bar{\boldsymbol{y}} - \boldsymbol{\mu})'\boldsymbol{W}^{-1}(\bar{\boldsymbol{y}} - \boldsymbol{\mu})}\right] \times \left[n(n-1)(\bar{\boldsymbol{y}} - \boldsymbol{\mu})'\boldsymbol{W}^{-1}(\bar{\boldsymbol{y}} - \boldsymbol{\mu})\right] = T_1 \times T_2.$$

Using standard properties of Wishart distribution (Anderson, 2003; Kshirsagar, 1972), it follows that $T_1|\boldsymbol{W},\bar{\boldsymbol{y}} \sim \frac{1}{\chi^2_{n-p}}$, which is independent of both $\boldsymbol{W}$ and $\bar{\boldsymbol{y}}$. To derive the distribution of T_2, note that $[n(n-1)]^{1/2}\boldsymbol{W}^{-1/2}(\bar{\boldsymbol{y}}-\boldsymbol{\mu})|\boldsymbol{W} \sim N_p[\mathbf{0},\boldsymbol{I}_p+(n-1)\boldsymbol{W}^{-1/2}\boldsymbol{\Sigma}\boldsymbol{W}^{-1/2}]$. We use the fact that if $\boldsymbol{Z} \sim N_p(\mathbf{0},\boldsymbol{A})$, then $\boldsymbol{Z}'\boldsymbol{Z} \sim \sum_{i=1}^{p}\lambda_i\chi^2_{1i}$ where $\lambda_1,\ldots,\lambda_p$ are the eigenvalues of $\boldsymbol{A}$ and χ^2_{1i} are independent χ^2 variables each with 1 degree of freedom. In our case, since $\boldsymbol{A}=\boldsymbol{I}_p+(n-1)\boldsymbol{W}^{-1/2}\boldsymbol{\Sigma}\boldsymbol{W}^{-1/2}$, $\lambda_1,\ldots,\lambda_p$ are the roots of $|(n-1)\boldsymbol{I}_p+(1-\lambda)\boldsymbol{\Sigma}^{-1/2}\boldsymbol{W}\boldsymbol{\Sigma}^{-1/2}|=0$ which can be easily generated based on the fact that $\boldsymbol{\Sigma}^{-1/2}\boldsymbol{W}\boldsymbol{\Sigma}^{-1/2} \sim \text{Wishart}_p(\boldsymbol{I}_p,n-1)$. Taking $\boldsymbol{Z}=[n(n-1)]^{1/2}\boldsymbol{W}^{-1/2}(\bar{\boldsymbol{y}}-\boldsymbol{\mu})$, it finally follows that:

(a) the conditional distribution of T_2, given $\boldsymbol{W}$, is $\sum_{i=1}^{p}\lambda_i\chi^2_{1i}$ where $\lambda_1,\ldots,\lambda_p$ are the roots of $|(n-1)\boldsymbol{I}_p+(1-\lambda)\boldsymbol{W}^*|=0$ and $\boldsymbol{W}^* \sim \text{Wishart}_p(\boldsymbol{I}_p,n-1)$; and

(b) the unconditional distribution of T_2 is obtained by averaging over the joint distribution of the roots $\lambda_1,\ldots,\lambda_p$.

This completes the proof. □

Proof of Theorem 4.1. The proof is based on the following steps.

1. Given $(\boldsymbol{b},\text{RSS})$, the conditional joint pdf of $(\boldsymbol{b}^*,\text{RSS}^*)$ is given by

$$f(\boldsymbol{b}^*,\text{RSS}^*|\boldsymbol{b},\text{RSS}) \propto e^{-\frac{n-p}{2}\left[\frac{(\boldsymbol{b}^*-\boldsymbol{b})'(\mathbf{XX}')(\boldsymbol{b}^*-\boldsymbol{b})+\text{RSS}^*}{\text{RSS}}\right]} \times \frac{(\text{RSS}^*)^{\frac{n-p}{2}-1}}{\text{RSS}^{n/2}}.$$

2. The joint pdf of $(\boldsymbol{b},\text{RSS})$ is given by

$$f_{\boldsymbol{\beta},\sigma^2}(\boldsymbol{b},\text{RSS}) \propto e^{-\frac{1}{2}\left[\frac{(\boldsymbol{b}-\boldsymbol{\beta})'(\mathbf{XX}')(\boldsymbol{b}-\boldsymbol{\beta})}{\sigma^2}+\frac{\text{RSS}}{\sigma^2}\right]} \times \frac{(\text{RSS})^{\frac{n-p}{2}-1}}{\sigma^n}.$$

Combining the above, we get the joint pdf of $(\boldsymbol{b}^*,\text{RSS}^*,\boldsymbol{b},\text{RSS})$ which we use to sequentially integrate out $\boldsymbol{b}$ and RSS. Writing $\widetilde{\text{RSS}}=\text{RSS}/(n-p)$, since

$$\frac{(\boldsymbol{b}^*-\boldsymbol{b})'(\mathbf{XX}')(\boldsymbol{b}^*-\boldsymbol{b})}{\widetilde{\text{RSS}}}+\frac{(\boldsymbol{b}-\boldsymbol{\beta})'(\mathbf{XX}')(\boldsymbol{b}-\boldsymbol{\beta})}{\sigma^2}$$
$$=\left(\frac{1}{\sigma^2}+\frac{1}{\widetilde{\text{RSS}}}\right)\left[\boldsymbol{b}-\frac{(\frac{\boldsymbol{\beta}}{\sigma^2}+\frac{\boldsymbol{b}^*}{\widetilde{\text{RSS}}})}{(\frac{1}{\sigma^2}+\frac{1}{\widetilde{\text{RSS}}})}\right]'(\mathbf{XX}')\left[\boldsymbol{b}-\frac{(\frac{\boldsymbol{\beta}}{\sigma^2}+\frac{\boldsymbol{b}^*}{\widetilde{\text{RSS}}})}{(\frac{1}{\sigma^2}+\frac{1}{\widetilde{\text{RSS}}})}\right]+\frac{(\boldsymbol{b}^*-\boldsymbol{\beta})'(\mathbf{XX}')(\boldsymbol{b}^*-\boldsymbol{\beta})}{(\sigma^2+\widetilde{\text{RSS}})},$$

integrating out $\boldsymbol{b}$, we get the joint pdf of $(\boldsymbol{b}^*,\text{RSS}^*,\text{RSS})$ as

$$f_{\boldsymbol{\beta},\sigma^2}(\boldsymbol{b}^*,\text{RSS}^*,\text{RSS})$$
$$\propto e^{-\frac{1}{2}\left[\frac{(\boldsymbol{b}^*-\boldsymbol{\beta})'(\mathbf{XX}')(\boldsymbol{b}^*-\boldsymbol{\beta})}{\sigma^2+\widetilde{\text{RSS}}}+\frac{\text{RSS}^*}{\widetilde{\text{RSS}}}+\frac{\text{RSS}}{\sigma^2}\right]} \times \frac{(\text{RSS}^*)^{\frac{n-p}{2}-1}}{(\text{RSS})^{n/2}} \times \frac{(\text{RSS})^{-\frac{p+2}{2}}}{\sigma^n} \times \left[\frac{1}{\sigma^2}+\frac{1}{\widetilde{\text{RSS}}}\right]^{-p/2}.$$

Putting $\psi = \mathrm{RSS}/\sigma^2$, the joint pdf of $(\boldsymbol{b}^*, \mathrm{RSS}^*, \psi)$ simplifies as

$$f_{\boldsymbol{\beta},\sigma^2}(\boldsymbol{b}^*, \mathrm{RSS}^*, \psi)$$
$$\propto e^{-\frac{1}{2}\left[\frac{(\boldsymbol{b}^*-\boldsymbol{\beta})'(\mathbf{XX}')(\boldsymbol{b}^*-\boldsymbol{\beta})}{\sigma^2(1+\frac{\psi}{n-p})}+\frac{(n-p)\mathrm{RSS}^*}{\sigma^2\psi}+\psi\right]} \times \frac{(\mathrm{RSS}^*)^{\frac{n-p}{2}-1}}{(\sigma^2)^{\frac{n-p}{2}}} \times \frac{(\psi)^{-\frac{p+2}{2}}}{\sigma^p} \times \left[1+\frac{n-p}{\psi}\right]^{-p/2}.$$

Integrating out ψ, we get the desired result. □

Proof of Theorem 4.2. This is immediate from the joint pdf of $(\boldsymbol{b}^*, \mathrm{RSS}^*)$ upon integrating out $\boldsymbol{b}^*$ and making the transformation $V = \mathrm{RSS}^*/\sigma^2$. □

Proof of Theorem 4.3. From Theorem 4.1, it follows that, conditionally given ψ,

$$\frac{(\boldsymbol{b}^*-\boldsymbol{\beta})'(\mathbf{XX}')(\boldsymbol{b}^*-\boldsymbol{\beta})}{\sigma^2(1+\frac{\psi}{n-p})} \sim \chi^2_p, \quad \frac{(n-p)\mathrm{RSS}^*}{\sigma^2\psi} \sim \chi^2_{n-p}, \quad \textit{independent of } \boldsymbol{b}^*$$

and, marginally, $\psi \sim f_{n,p}(\psi) \propto e^{-\frac{\psi}{2}}(\psi)^{\frac{n-p}{2}-1}$. The result follows immediately upon noting that, conditionally given ψ,

$$\frac{(\boldsymbol{b}^*-\boldsymbol{\beta})'(\mathbf{XX}')(\boldsymbol{b}^*-\boldsymbol{\beta})\psi}{(n-p+\psi)\mathrm{RSS}^*} \sim \frac{p}{n-p}F_{p,n-p}.$$

This completes the proof. □

2 Derivations of the Results in Sections 3 and 4

Here we provide details about the derivations of the Results that appear in Sections 4 and 3.

2.1 Results in Section 3

Details of Result 3.1. Since $\boldsymbol{y}_1, \ldots \boldsymbol{y}_n$, conditional on $\boldsymbol{X}$, are *iid* from $N_p\left(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}\right)$, and $E(\hat{\boldsymbol{\mu}}) = E(\bar{\boldsymbol{x}}) = \boldsymbol{\mu}$, it follows from a standard conditional argument that $E(\bar{\boldsymbol{y}}) = \boldsymbol{\mu}$. Likewise,

$$\mathrm{Var}(\bar{\boldsymbol{y}}) = E[\mathrm{Var}(\bar{\boldsymbol{y}}|\bar{\boldsymbol{x}}, \mathscr{S}_x)] + \mathrm{Var}[E(\bar{\boldsymbol{y}}|\bar{\boldsymbol{x}}, \mathscr{S}_x)] = E\left[\frac{\mathscr{S}_x}{n(n-1)}\right] + \mathrm{Var}(\bar{\boldsymbol{x}}) = \frac{2\boldsymbol{\Sigma}}{n}.$$

We also note from step 5 of the proof of Theorem 3.1 that the MLE of $\boldsymbol{\mu}$ is $\bar{\boldsymbol{y}}$.

Details of Result 3.2. $E(\mathscr{S}_y/(n-1)) = E(\hat{\boldsymbol{\Sigma}}) = \boldsymbol{\Sigma}$.

Details of Result 3.3. The expression for the volume $V_{\boldsymbol{\mu}}(\boldsymbol{Y})$ follows from a general result that if $\mathscr{A}$ is a $p \times p$ dimensional positive definite matrix, $\boldsymbol{v} \in \mathbb{R}^p$, and $C > 0$, then the volume of the ellipsoid $\{\boldsymbol{x} \in \mathbb{R}^p : (\boldsymbol{x} - \boldsymbol{v})'\mathscr{A}(\boldsymbol{x} - \boldsymbol{v}) \leq C\}$ is equal to $[\pi^{p/2}/\Gamma\left(\frac{p}{2}+1\right)]C^{p/2}|\mathscr{A}|^{-1/2}$. The expression for the expected volume follows because $\mathscr{S}_y|\hat{\boldsymbol{\Sigma}} \sim \text{Wishart}_p(\hat{\boldsymbol{\Sigma}}, n-1)$, and therefore, conditional on $\hat{\boldsymbol{\Sigma}}$, $|\mathscr{S}_y|$ is distributed as $|\hat{\boldsymbol{\Sigma}}| \times \prod_{i=1}^p \chi^2_{n-i}$, where $\prod_{i=1}^p \chi^2_{n-i}$ is the product of p independent χ^2 random variables with the ith component having $n-i$ degrees of freedom (Anderson, 2003; Kshirsagar, 1972). Hence we get $E\left(|\mathscr{S}_y|^{1/2}\right) = \mathscr{C}_{n,p}E\left(|\hat{\boldsymbol{\Sigma}}|^{1/2}\right)$ where $\mathscr{C}_{n,p} = E[(\prod_{i=1}^p \chi^2_{n-i})^{1/2}] = \prod_{i=1}^p 2^{1/2}\Gamma(\frac{n-i+1}{2})/\Gamma(\frac{n-i}{2})$. Similarly, since $(n-1)\hat{\boldsymbol{\Sigma}} = \mathscr{S}_x \sim \text{Wishart}_p(\boldsymbol{\Sigma}, n-1)$, we get $E\left(|\mathscr{S}_x|^{1/2}\right) = \mathscr{C}_{n,p}|\boldsymbol{\Sigma}|^{1/2}$. Combining these two results, we get the expression for $E[V_{\boldsymbol{\mu}}(\boldsymbol{Y})]$. The result pertaining to the determination of the cut-off point $c_{n,p,\gamma}$ directly follows from Theorem 3.1.

2.2 Results in Section 4

Here we provide details about the derivations of the Results that appear in Section 4.

Details of Result 4.1. That the MLE of $\boldsymbol{\beta}$ is $\boldsymbol{b}^*$ directly follows from Theorem 4.1. Also, we have $E(\boldsymbol{b}^*) = E\left[E(\boldsymbol{b}^*|\boldsymbol{b}, \text{RSS})\right] = E(\boldsymbol{b}) = \boldsymbol{\beta}$. Furthermore, from a standard conditional argument, we get

$$\begin{aligned}\text{Var}(\boldsymbol{b}^*) &= E[\text{Var}(\boldsymbol{b}^*|\boldsymbol{b}, \text{RSS})] + \text{Var}[E(\boldsymbol{b}^*|\boldsymbol{b}, \text{RSS})]\\ &= E[(\boldsymbol{X}\boldsymbol{X}')^{-1}(\text{RSS}/(n-1))] + \text{Var}[\boldsymbol{b}]\\ &= 2\sigma^2(\boldsymbol{X}\boldsymbol{X}')^{-1}.\end{aligned}$$

Details of Result 4.2. $E(\text{RSS}^*) = E[E(\text{RSS}^*|\text{RSS})] = E(\text{RSS}) = (n-p)\sigma^2$.

Details of Result 4.3. Plugging in the MLE $\boldsymbol{b}^*$ of $\boldsymbol{b}$ in Theorem 4.1, we get the *restricted* likelihood of $\sigma^2|\text{RSS}^*$, whose maximization yields the MLE of σ^2. Defining $\Delta = (n-p)\text{RSS}^*/\sigma^2$, it amounts to maximizing the expression

$$(\Delta)^{n/2} \times \int_0^\infty e^{-\frac{1}{2}[\psi + \frac{\Delta}{\psi}]}(\psi)^{-\frac{p+2}{2}}(1 + \frac{n-p}{\psi})^{-p/2}d\psi$$

with respect to Δ, and hence the MLE of σ^2 is readily obtained.

Details of Result 4.4. The conditions for the constants $a_{n,p;\gamma}$ and $b_{n,p;\gamma}$ that yield the shortest confidence interval for σ^2 can be derived using the argument on page 444 of Casella and Berger (2001). The expected confidence interval length follows directly from Result 4.2.

Details of Result 4.5. From Details of Result 3.3 given in Appendix 2.1, we get the volume of the confidence ellipsoid $\Delta_{\text{MLR}}(\boldsymbol{\beta})$ as $V_{\boldsymbol{\beta}}(\boldsymbol{v}, \boldsymbol{X}) = \frac{\pi^{p/2}}{\Gamma(\frac{p}{2}+1)} d_{n,p;\gamma}^{p/2} |\boldsymbol{X}\boldsymbol{X}'|^{-1/2} (\text{RSS}^*)^{p/2}$. Since $\text{RSS}^*/(\text{RSS}/(n-p))|\text{RSS} \sim \chi^2_{n-p}$ and $\text{RSS}/\sigma^2 \sim \chi^2_{n-p}$, we get $E(\text{RSS}^*)^{p/2} = E[(\text{RSS})^{p/2}]/(n-p)^{p/2} = \sigma^p E\{(\chi^2_{n-p})^{p/2}\}/(n-p)^{p/2}$. Combining all the terms, we get the desired result.

Details of Result 4.6. From the proof of Theorem 4.1, it follows that, conditionally given ψ,

$$\frac{(\mathbf{A}\boldsymbol{b}^* - \boldsymbol{\eta})'[\mathbf{A}(\mathbf{X}\mathbf{X}')^{-1}\mathbf{A}']^{-1}(\mathbf{A}\boldsymbol{b}^* - \boldsymbol{\eta})}{\sigma^2(1 + \frac{\psi}{n-p})} \sim \chi^2_k, \quad \frac{(n-p)\text{RSS}^*}{\sigma^2\psi} \sim \chi^2_{n-p}, \text{ independent of } \boldsymbol{b}^*,$$

and, marginally, $\psi \sim f_{n,p}(\psi) \propto e^{-\frac{\psi}{2}}(\psi)^{\frac{n-p}{2}-1}$. Hence, one readily obtains

$$T^2_{\boldsymbol{\eta}}|\psi \sim \left[\frac{k}{n-p}\right]\left[1 + \frac{n-p}{\psi}\right] F_{k,n-p} \quad \text{and} \quad f_{n,p}(\psi) \propto e^{-\frac{\psi}{2}}(\psi)^{\frac{n-p}{2}-1}.$$

3 Proofs of the Expressions in Table 9

Here we provide proofs of the expressions appearing in Table 9. We show that in Cases 2, 3, 4, the data analyst's estimate of β_1 is biased. However, in Cases 5, 6, 7, such an estimate is unbiased and the estimated variance of the estimate of β_1 from the data analyst's point of view is unbiased in Case 5, approximately unbiased in Case 6 if the sample size is large, and is biased in Case 7. Recall that the $p \times n$ design matrix is denoted by $\boldsymbol{X}$ with row vectors as $\boldsymbol{x}_i'$'s.

Case 1: DM = M_F, IM = M_F, AM = M_F. In this case the assumptions of Section 4 hold where $\boldsymbol{X} = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1n} \\ x_{21} & x_{22} & \dots & x_{2n} \end{pmatrix}$. Therefore it follows from Results 4.1 and 4.2 that $E(b_1^*) = \beta_1$, $\text{Var}(b_1^*) = 2\sigma^2\left(\frac{\sum_{i=1}^n x_{2i}^2}{\Delta}\right)$, and $E[\widehat{\text{Var}}(b_1^*)] = 2\sigma^2\left(\frac{\sum_{i=1}^n x_{2i}^2}{\Delta}\right)$ where $\Delta = (\sum_{i=1}^n x_{1i}^2)(\sum_{i=1}^n x_{2i}^2) - (\sum_{i=1}^n x_{1i}x_{2i})^2$.

Case 2: DM = M_F, IM = M_F, AM = M_R. In this case the data analyst's estimate of β_1 is given by $b_1^* = [\sum_{i=1}^n v_i x_{1i}]/[\sum_{i=1}^n x_{1i}^2]$ where the v_i's are generated as independent normal variables with $v_i \sim N[x_{1i}b_1 + x_{2i}b_2, \text{RSS}/(n-2)]$ where b_1, b_2 and $\text{RSS}/(n-2)$ are the usual estimates of β_1, β_2 and σ^2 based on the original data $\boldsymbol{y}$. Since $E(v_i) = x_{1i}\beta_1 + x_{2i}\beta_2$, we readily get $E(b_1^*) = \beta_1 + \beta_2[(\sum_{i=1}^n x_{1i}x_{2i})/(\sum_{i=1}^n x_{1i}^2)]$.

Case 3: DM = M_F, IM = M_R, AM = M_F. Obviously, here the data analyst's estimate of β_1 (based on M_F) is given by $b_1^* = \sum_{i=1}^n v_i c_i/\Delta$, where $c_i = x_{1i}(\sum_{i=1}^n x_{2i}^2) - x_{2i}(\sum_{i=1}^n x_{1i}x_{2i})$ and $\Delta = (\sum_{i=1}^n x_{1i}^2)(\sum_{i=1}^n x_{2i}^2) - (\sum_{i=1}^n x_{1i}x_{2i})^2$. Recall that under the data imputation model M_R, $(v_1, \cdots, v_n)$ are generated as independent normal variables with $v_i \sim N[x_{1i}b_1, \mathrm{RSS}/(n-1)]$ where b_1 and $\mathrm{RSS}/(n-1)$ are the usual unbiased estimates of β_1 and σ^2 based on the original data $\boldsymbol{y}$ and M_R. Since $E(v_i) = x_{1i}E(b_1)$ and under IM (M_R), $E(b_1) = E[(\sum_{i=1}^n y_i x_{1i})/(\sum_{i=1}^n x_{1i}^2)] = \beta_1 + \beta_2[(\sum_{i=1}^n x_{1i}x_{2i})/(\sum_{i=1}^n x_{1i}^2)]$, using $\sum_{i=1}^n c_i x_{1i} = \Delta$, we get $E(b_1^*) = \beta_1 + \beta_2 \left(\frac{\sum_{i=1}^n x_{1i}x_{2i}}{\sum_{i=1}^n x_{1i}^2}\right)$.

Case 4: DM $= M_F$, IM $= M_R$, AM $= M_R$. In this case, as in Case 2, the data analyst's estimate of β_1 is given by $b_1^* = [\sum_{i=1}^n v_i x_{1i}]/[\sum_{i=1}^n x_{1i}^2]$ where the v_i's are generated (similar to Case 3) as independent normal variables with $v_i \sim N[x_{1i}b_1, \mathrm{RSS}/(n-1)]$ where b_1 and $\mathrm{RSS}/(n-1)$ are the usual unbiased estimates of β_1 and σ^2 based on the original data $\boldsymbol{y}$ and M_R. Since $E(v_i) = x_{1i}E(b_1)$ and under DM (M_F), $E(b_1) = E[(\sum_{i=1}^n y_i x_{1i})/(\sum_{i=1}^n x_{1i}^2)] = \beta_1 + \beta_2[(\sum_{i=1}^n x_{1i}x_{2i})/(\sum_{i=1}^n x_{1i}^2)]$, we get $E(b_1^*) = \beta_1 + \beta_2[(\sum_{i=1}^n x_{1i}x_{2i})/(\sum_{i=1}^n x_{1i}^2)]$.

Case 5: DM $= M_R$, IM $= M_F$, AM $= M_F$. Obviously, here the data analyst's estimate of β_1 is given by $b_1^* = \sum_{i=1}^n v_i c_i/\Delta$, where $c_i = x_{1i}(\sum_{i=1}^n x_{2i}^2) - x_{2i}(\sum_{i=1}^n x_{1i}x_{2i})$ and $\Delta = (\sum_{i=1}^n x_{1i}^2)(\sum_{i=1}^n x_{2i}^2) - (\sum_{i=1}^n x_{1i}x_{2i})^2$. Recall that under the imputation model M_F, $(v_1, \cdots, v_n)$ are generated as independent normal variables with $v_i \sim N[x_{1i}b_1 + x_{2i}b_2, \mathrm{RSS}/(n-2)]$ where b_1, b_2 and $\mathrm{RSS}/(n-2)$ are the usual estimates of β_1, β_2 and σ^2 based on the original data $\boldsymbol{y}$. This means $\boldsymbol{b} = [\boldsymbol{X}\boldsymbol{X}']^{-1}\begin{pmatrix}\boldsymbol{x}_1'\boldsymbol{y}\\ \boldsymbol{x}_2'\boldsymbol{y}\end{pmatrix}$ and $\mathrm{RSS} = \boldsymbol{y}'[\boldsymbol{I}_n - \boldsymbol{X}'(\boldsymbol{X}\boldsymbol{X}')^{-1}\boldsymbol{X}]\boldsymbol{y}$. Unbiasedness of b_1^* for β_1 easily follows because $\sum_{i=1}^n c_i x_{1i} = \Delta$ and $\sum_{i=1}^n c_i x_{2i} = 0$.

The true variance of this unbiased estimate b_1^* of β_1 under the DM (M_R) consists of two terms given by

$$\mathrm{Var}(b_1^*) = \frac{1}{\Delta^2}\left[E\left\{\mathrm{Var}\left(\sum_{i=1}^n v_i c_i \,\Big|\, \boldsymbol{y}\right)\right\} + \mathrm{Var}\left\{E\left(\sum_{i=1}^n v_i c_i \,\Big|\, \boldsymbol{y}\right)\right\}\right].$$

Again, using the fact that (a) $\sum_{i=1}^n c_i x_{1i} = \Delta$ and (b) $\sum_{i=1}^n c_i x_{2i} = 0$, the 2nd term simplifies to $\sigma^2(\sum_{i=1}^n x_{2i}^2)/\Delta$. Likewise, using the fact that $\boldsymbol{x}_1'[\boldsymbol{I}_n - \boldsymbol{X}'(\boldsymbol{X}\boldsymbol{X}')^{-1}\boldsymbol{X}] = \boldsymbol{0}$, the 1st term also simplifies to $\sigma^2(\sum_{i=1}^n x_{2i}^2)/\Delta$, resulting in $\mathrm{Var}(b_1^*) = 2\sigma^2(\sum_{i=1}^n x_{2i}^2)/\Delta$. On the other hand, data analyst's inference is based on using the same estimate b_1^* but with its variance computed as $2\tau^2(\sum_{i=1}^n c_i^2)/\Delta^2$. Then τ^2 is estimated by $\hat{\tau}^2 = \boldsymbol{v}'[\boldsymbol{I}_n - \boldsymbol{X}'(\boldsymbol{X}\boldsymbol{X}')^{-1}\boldsymbol{X}]\boldsymbol{v}/(n-2)$, resulting in the estimated variance as $\widehat{\mathrm{Var}}(b_1^*) = 2\hat{\tau}^2(\sum_{i=1}^n c_i^2)/\Delta^2$. We show below that $E[\hat{\tau}^2] = \sigma^2$, implying the fact that analyst's inference is valid in the sense of providing an unbiased estimate of the true variance.

Note that $(n-2)E[\hat{\tau}^2] = \mathrm{tr}([\boldsymbol{I}_n - \boldsymbol{X}'(\boldsymbol{X}\boldsymbol{X}')^{-1}\boldsymbol{X}]E(\boldsymbol{v}\boldsymbol{v}'))$. Since $\boldsymbol{x}_i'[\boldsymbol{I}_n - \boldsymbol{X}'(\boldsymbol{X}\boldsymbol{X}')^{-1}\boldsymbol{X}] = \boldsymbol{0}$, $i = 1, 2$, we get $E(\boldsymbol{v}\boldsymbol{v}') = \boldsymbol{I}_n E(\mathrm{RSS}/(n-2)) = \sigma^2\boldsymbol{I}_n$. Since $\mathrm{tr}[\boldsymbol{I}_n - \boldsymbol{X}'(\boldsymbol{X}\boldsymbol{X}')^{-1}\boldsymbol{X}] = n-2$, this proves the result.

Case 6: DM $= M_R$, IM $= M_F$, AM $= M_R$. In this case the data analyst's estimate of β_1 is given by $b_1^* = [\sum_{i=1}^n v_i x_{1i}]/[\sum_{i=1}^n x_{1i}^2]$ where the v_i's are generated in the same way as in Case 5. Obviously, $E(b_1^*) = \beta_1$ with $\text{Var}(b_1^*) = 2\sigma^2/[\sum_{i=1}^n x_{1i}^2]$. This is because

$$E[\sum_{i=1}^n v_i x_{1i}]/[\sum_{i=1}^n x_{1i}^2] = E(b_1) = \beta_1$$

and $E(b_2) = 0$ under DM (M_R) model. The expression for $\text{Var}(b_1^*)$ also follows directly from the standard conditional argument because $E[\text{RSS}/(n-2)] = \sigma^2$.

On the other hand, analyst's inference is based on using the same estimate b_1^* but with its variance computed as $[2\tau^2]/[\sum_{i=1}^n x_{1i}^2]$ where τ^2 is estimated by $\hat{\tau}^2 = \boldsymbol{v}'[\boldsymbol{I}_n - \boldsymbol{X}_1(\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1']\boldsymbol{v}/(n-1) = [\sum_{i=1}^n v_i^2 - \frac{(\sum_{i=1}^n v_1 x_{1i})^2}{\sum_{i=1}^n x_{1i}^2}]/(n-1)$ in view of data analyst's use of the reduced model M_R, resulting in the estimated variance as $\widehat{\text{Var}}(b_1^*)] = 2\hat{\tau}^2/[\sum_{i=1}^n x_{1i}^2]$. We show below that $E[\hat{\tau}^2] = \sigma^2(\frac{n}{n-1})$, implying the fact that analyst's inference is *almost* valid in the sense of providing an *almost* unbiased estimate of the true variance.

Note that $(n-1)E[\hat{\tau}^2] = \text{tr}[\boldsymbol{I}_n - \boldsymbol{X}_1(\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1']E[\boldsymbol{v}\boldsymbol{v}']$, and $E[\boldsymbol{v}\boldsymbol{v}'] = \boldsymbol{I}_n E[\text{RSS}/(n-2)] + E[(b_1\boldsymbol{X}_1 + b_2\boldsymbol{X}_2)(b_1\boldsymbol{X}_1 + b_2\boldsymbol{X}_2)'] = \boldsymbol{I}_n E[\text{RSS}/(n-2)] + E[b_2^2\boldsymbol{X}_2\boldsymbol{X}_2']$ since the rest of the terms are 0 under multiplication by $[\boldsymbol{I}_n - \boldsymbol{X}_1(\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1]$. The result now follows because $E[\text{RSS}/(n-2)] = \sigma^2$, $\boldsymbol{X}_2'[\boldsymbol{I}_n - \boldsymbol{X}_1(\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1']\boldsymbol{X}_2' = \frac{\Delta}{\sum_{i=1}^n x_{1i}^2}$, and $E(b_2^2) = \sigma^2(\sum_{i=1}^n x_{1i}^2/\Delta)$.

Case 7: DM $= M_R$, IM $= M_R$, AM $= M_F$. Obviously, here the data analyst's estimate of β_1 (based on M_F) is given by (as in Case 5) $b_1^* = \sum_{i=1}^n v_i c_i/\Delta$, where $c_i = x_{1i}(\sum_{i=1}^n x_{2i}^2) - x_{2i}(\sum_{i=1}^n x_{1i}x_{2i})$ and $\Delta = (\sum_{i=1}^n x_{1i}^2)(\sum_{i=1}^n x_{2i}^2) - (\sum_{i=1}^n x_{1i}x_{2i})^2$. Recall that under the data imputation model M_R, $(v_1, \cdots, v_n)$ are generated as independent normal variables with $v_i \sim N[x_{1i}b_1, \text{RSS}/(n-1)]$ where b_1 and $\text{RSS}/(n-1)$ are the usual unbiased estimates of β_1 and σ^2 based on the original data $\boldsymbol{y}$ and M_R. This means $b_1 = [\sum_{i=1}^n y_i x_{1i}]/[\sum_{i=1}^n x_{1i}^2]$ and $\text{RSS} = \boldsymbol{y}'[\boldsymbol{I}_n - \boldsymbol{X}_1'(\boldsymbol{X}_1\boldsymbol{X}_1')^{-1}\boldsymbol{X}_1]\boldsymbol{y} = [\sum_{i=1}^n y_i^2 - \frac{(\sum_{i=1}^n y_1 x_{1i})^2}{\sum_{i=1}^n x_{1i}^2}]$. Obviously, $E(b_1^*) = \beta_1$.

The true variance of this unbiased estimate b_1^* of β_1 under the DM model (M_R) consists of two terms given by

$$\text{Var}(b_1^*) = \frac{1}{\Delta^2}\left[E\left\{\text{Var}\left(\sum_{i=1}^n v_i c_i \,\middle|\, \boldsymbol{y}\right)\right\} + \text{Var}\left\{E\left(\sum_{i=1}^n v_i c_i \,\middle|\, \boldsymbol{y}\right)\right\}\right].$$

Using the fact that $(v_1, \ldots, v_n)$ are generated independently such that $v_i \sim N[x_{1i}b_1, \frac{\text{RSS}}{n-1}]$ and $\sum_{i=1}^n c_i x_{1i} = \Delta$, the 2nd term simplifies to $\sigma^2/\sum_{i=1}^n x_{1i}^2$. Likewise, using the fact that $\sum_{i=1}^n c_i^2 = (\sum_{i=1}^n x_{2i}^2)\Delta$ and $E(\text{RSS}/(n-1)) = \sigma^2$, the 1st term simplifies to $\sigma^2(\sum_{i=1}^n x_{2i}^2)/\Delta$, resulting in

$$\text{Var}(b_1^*) = \sigma^2\left(\frac{\sum_{i=1}^n x_{2i}^2}{\Delta}\right) + \sigma^2\left(\frac{1}{\sum_{i=1}^n x_{1i}^2}\right).$$

On the other hand, analyst's inference is based on using the same estimate b_1^* but with its variance computed as $2\tau^2(\sum_{i=1}^n c_i^2)/\Delta^2$, and τ^2 is estimated by $\hat{\tau}^2 = \boldsymbol{v}'[\boldsymbol{I}_n -$

$\boldsymbol{X}'(\boldsymbol{X}\boldsymbol{X}')^{-1}\boldsymbol{X}]\boldsymbol{v}/(n-2)$, resulting in the estimated variance as $\widehat{\mathrm{Var}}(b_1^*) = 2\hat{\tau}^2(\sum_{i=1}^n c_i^2)/\Delta^2 = 2(\sum_{i=1}^n x_{2i}^2)/\Delta$. We show below that $E[\hat{\tau}^2] = \sigma^2$, implying the fact that analyst's inference is *not* valid in this case in the sense of *not* providing an unbiased estimate of the true variance. Note that $(n-2)E[\hat{\tau}^2] = \mathrm{tr}([\boldsymbol{I}_n - \boldsymbol{X}'(\boldsymbol{X}\boldsymbol{X}')^{-1}\boldsymbol{X}]E(\boldsymbol{v}\boldsymbol{v}'))$. In view of $(v_1, \cdots, v_n)$ being generated as independent normal variables with $v_i \sim N[x_{1i}b_1, \mathrm{RSS}/(n-1)]$, using the fact that $\boldsymbol{x}_1'[\boldsymbol{I}_n - \boldsymbol{X}'(\boldsymbol{X}\boldsymbol{X}')^{-1}\boldsymbol{X}] = \boldsymbol{0}$, we get $E(\boldsymbol{v}\boldsymbol{v}') = \boldsymbol{I}_n E(\mathrm{RSS}/(n-1)) = \sigma^2\boldsymbol{I}_n$. Since $\mathrm{tr}[\boldsymbol{I}_n - \boldsymbol{X}'(\boldsymbol{X}\boldsymbol{X}')^{-1}\boldsymbol{X}] = n-2$, this proves the result.

Case 8: DM = M_R, IM = M_R, AM = M_R. In this case the assumptions of Section 4 hold where $\boldsymbol{X} = \begin{pmatrix} x_{11} & x_{12} & \ldots & x_{1n} \end{pmatrix}$. Therefore it follows from Results 4.1 and 4.2 that $E(b_1^*) = \beta_1$, $\mathrm{Var}(b_1^*) = 2\sigma^2\left(\frac{1}{\sum_{i=1}^n x_{1i}^2}\right)$, and $E[\widehat{\mathrm{Var}}(b_1^*)] = 2\sigma^2\left(\frac{1}{\sum_{i=1}^n x_{1i}^2}\right)$.

References

Abowd, J., Stinson, M., and Benedetto, G. (2006). Final Report to the Social Security Administration on the SIPP/SSA/IRS Public Use File Project. *Technical Report.* Available at `http://www2.vrdc.cornell.edu/news/wp-content/uploads/2007/11/ssafinal.pdf`.

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

Anderson, T.W. (2003). *An Introduction to Multivariate Statistical Analysis.* Third edition. Wiley.

Benedetto, G., Stinson, M.H., and Abowd, J.M. (2013). The Creation and Use of the SIPP Synthetic Beta, *Technical Report.* Available at `http://www.census.gov/content/dam/Census/programs-surveys/sipp/methodology/SSBdescribe_nontechnical.pdf`.

Casella, G., and Berger, R.L. (2001). *Statistical Inference.* Second edition. Duxbury.

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

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

Hartung, J., Knapp, G., and Sinha, B.K. (2008). *Statistical Meta-Analysis With Applications.* Wiley.

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

Hotelling, H. (1931). The generalization of student's ratio, *The Annals of Mathematical Statistics*, 2, 360-378.

Kinney, S.K., Reiter, J.P., and Miranda, J. (2014). SynLBD 2.0: Improving the synthetic longitudinal business database, *Statistical Journal of the International Association for Official Statistics*, 30, 129-135.

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

Klein, M., Mathew, T., and Sinha, B. (2014). Noise multiplication for statistical disclosure control of extreme values in log-normal regression samples, *Journal of Privacy and Confidentiality*, 6, 77-125.

Kshirsagar, A.M. (1972). *Multivariate Analysis.* Marcel Dekker.

Lin, Y.-X and Wise, P. (2012). Estimation of regression parameters from noise multiplied data. *Journal of Privacy and Confidentiality*, 4, 61-94.

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

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

Meng, X.L. (1994). Multiple-imputation inferences with uncongenial sources of input, *Statistical Science*, 9, 538-573.

Muirhead, R.J. (2005). *Aspects of Multivariate Statistical Theory.* Wiley.

R Development Core Team (2013). R: *A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. `http://www.R-project.org/`.

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

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

Reiter, J.P. (2004). Simultaneous use of multiple imputation for missing data and disclosure limitation, *Survey Methodology*, 30, 235-242.

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

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

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

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

Reiter, J.P., and Mitra, R. (2009). Estimating risks of identification disclosure in partially synthetic data, *Journal of Privacy and Confidentiality*, 1, 99-110.

Rencher, A.C., and Schaalje, G.B. (2008). *Linear Models in Statistics*. Second edition. Wiley.

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.

Rubin, D.B. (1996). Multiple imputation after 18+ years, *Journal of the American Statistical Association*, 91, 473-489.