Finite Sample Inference for Multiply Imputed Synthetic Data under a Multiple Linear Regression Model

Martin D. Klein[a,*], **John Zylstra**[b], **Bimal K. Sinha**[c]

Abstract

In this paper we develop finite sample inference based on multiply imputed synthetic data generated under the multiple linear regression model. We consider two methods of generating the synthetic data, namely, posterior predictive sampling and plug-in sampling. Simulation results are presented to confirm that the proposed methodology performs as the theory predicts, and to numerically compare the proposed methodology with the current state of the art procedures for analyzing multiply imputed partially synthetic data.

Keywords: Partially synthetic data; Pivotal quantity; Plug-in sampling; Posterior predictive sampling; Statistical disclosure control

1 Introduction

Statistical disclosure control (SDC) methodology aims to suitably modify a dataset prior to its release so that the modified dataset does not disclose confidential information about the individual

[a] Martin D. Klein, Center for Statistical Research and Methodology, U.S. Census Bureau, 4600 Silver Hill Road, Washington, DC 20233, U.S.A., Email: martin.klein@census.gov

[b] John Zylstra, Department of Mathematics and Statistics, University of Maryland, Baltimore County, 1000 Hilltop Circle, Baltimore, MD 21250, U.S.A., Email: zylstra1@umbc.edu

[c] Bimal K. Sinha, Center for Statistical Research and Methodology, U.S. Census Bureau, 4600 Silver Hill Road, Washington, DC 20233, U.S.A., and Department of Mathematics and Statistics, University of Maryland, Baltimore County, 1000 Hilltop Circle, Baltimore, MD 21250, U.S.A., Email: sinha@umbc.edu

*Corresponding author.

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

units that contributed their information to the dataset (for example, survey respondents). At the same time, it is also a goal that a dataset that has been modified using SDC methodology would still be useful for drawing inference on the relevant population. The release of synthetic data is a form of SDC methodology where (all or part of) the real data are not released, but are instead used to create synthetic data which are released. Generally the synthetic data literature refers to two types of synthetic data: *fully* and *partially* synthetic data. Fully synthetic data were proposed by Rubin[19] and methodology for drawing valid inference from such data was developed by Raghunathan, Reiter, and Rubin[12]. Partially synthetic data were proposed by Little[9] and methodology for drawing valid inference from these data was developed by Reiter[13]. Raghunathan, Reiter, and Rubin[12] and Reiter[13] developed methodologies for scalar valued estimands, while Reiter[14] extended these procedures for vector values estimands. Drechsler[3] provides a detailed account of both partially and fully synthetic data methodology. Both the methodology of Raghunathan, Reiter, and Rubin[12] and that of Reiter[13] are general in the sense that they can be applied under a variety of models and for a variety of parameters, and these methodologies provide inference that is approximately valid if the sample size is sufficiently large. Fully and partially synthetic data approaches both utilize concepts of multiple imputation for missing data as developed by Rubin[18], and therefore, both approaches call for releasing a total of $m > 1$ multiply imputed synthetic datasets. Examples of major data sources where partially synthetic data products have been produced include the Survey of Income and Program Participation (Abowd, Stinson, and Benedetto[1], Benedetto, Stinson, and Abowd[2]), the American Community Survey Group Quarters data (Hawala[4], Rodríguez[17]), OnTheMap data displaying where workers live and where they work (Machanavajjhala et al.[10]), and the Longitudinal Business Database (Kinney, Reiter, and Miranda[5], Kinney et al.[6]).

In this paper we focus on a specific synthetic data problem, namely, synthetic data under the multiple linear regression model. This synthetic data problem fits into the framework of partially synthetic data, and hence the methodology of Reiter[13] can be used to obtain approximately valid inference if the sample size is sufficiently large and the number of multiply imputed synthetic datasets available is $m > 1$. However, given the specific structure in this problem, we shall instead exploit the model structure to derive finite sample inference for the unknown regression coefficients. While the methodology we derive is specific to the problem at hand, it yields exact inference for both large and small samples using the $m \geq 1$ multiply imputed synthetic datasets that are available. (Of course, if $m = 1$ then we have only a singly imputed synthetic dataset, but the proposed method will still provide valid inference.)

Throughout, let $\boldsymbol{y} = (y_1, \ldots, y_n)'$ be the $n \times 1$ vector of sensitive response variables, and suppose $\boldsymbol{y} \sim N_n(\boldsymbol{X}'\boldsymbol{\beta}, \sigma^2 \boldsymbol{I}_n)$, where $\boldsymbol{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n]$ is a $p \times n$ dimensional matrix of fixed and non-sensitive predictor variables with rank$(\boldsymbol{X}) = p < n$, and the unknown parameters are $\boldsymbol{\beta} \in \mathbb{R}^p$ and $\sigma^2 > 0$. Note that the original data are $(\boldsymbol{y}, \boldsymbol{X})$, and based on the original data, $\boldsymbol{b} = (\boldsymbol{X}\boldsymbol{X}')^{-1}\boldsymbol{X}\boldsymbol{y}$ is the maximum likelihood estimator (MLE) and uniformly minimum variance unbiased estimator (UMVUE) of $\boldsymbol{\beta}$, distributed as $N_p(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}\boldsymbol{X}')^{-1})$, independent of $\hat{\sigma}^2 = \text{RSS}/(n-p)$ which is the UMVUE of σ^2 where $\text{RSS} = (\boldsymbol{y} - \boldsymbol{X}'\boldsymbol{b})'(\boldsymbol{y} - \boldsymbol{X}'\boldsymbol{b}) \sim \sigma^2 \chi^2_{n-p}$.

Since $\boldsymbol{y}$ is sensitive and hence cannot be released, instead it is replaced with $m \geq 1$ multiply imputed synthetic copies which are released. We consider two ways of generating the $m \geq 1$ synthetic copies of $\boldsymbol{y}$, namely, *posterior predictive sampling* and *plug-in sampling.*

Posterior Predictive Sampling. Assume a prior $\pi(\boldsymbol{\beta}, \sigma^2)$ for $(\boldsymbol{\beta}, \sigma^2)$, then the posterior distribution of $(\boldsymbol{\beta}, \sigma^2)$ given $\boldsymbol{y}$ is derived and used to draw m independent replications $\{(\boldsymbol{\beta}^*_j, \sigma^{*2}_j), j = 1, \ldots, m\}$ (known as posterior draws). Next, for each such posterior draw of $(\boldsymbol{\beta}, \sigma^2)$, a corresponding replicate of $\boldsymbol{y}$ is generated. Thus the synthetic data $\{\boldsymbol{z}_j = (z_{j1}, \ldots, z_{jn})', j = 1, \ldots, m\}$ are generated by drawing $\boldsymbol{z}_j$ from the $N_n(\boldsymbol{X}'\boldsymbol{\beta}^*_j, \sigma^{*2}_j \boldsymbol{I}_n)$ distribution, independently, for $j = 1, \ldots, m$. The data $(\boldsymbol{z}_1, \boldsymbol{X}), \ldots, (\boldsymbol{z}_m, \boldsymbol{X})$ are then released to the public. For the scenario described here, the usual practice for drawing inference on the unknown parameters from the synthetic data, assuming $m > 1$, is based on the methods of Reiter[13] for multiply imputed partially synthetic data. In the specific case of $m = 1$, likelihood based methods for drawing inference in this scenario were derived by Klein and Sinha[7].

Plug-in Sampling. An alternative way to generate synthetic data is to take the observed values of $\boldsymbol{b}$ and $\text{RSS}/(n-p)$, the point estimators of the unknown parameters $\boldsymbol{\beta}$ and σ^2, plug them into the distribution of $\boldsymbol{y}$, and use the resulting distribution to generate synthetic data. Thus the synthetic data $\{\boldsymbol{z}_j = (z_{j1}, \ldots, z_{jn})', j = 1, \ldots, m\}$ are generated by drawing each $\boldsymbol{z}_j$ independently from the $N_n\left(\boldsymbol{X}'\boldsymbol{b}, \frac{\text{RSS}}{n-p}\boldsymbol{I}_n\right)$ distribution. The data $(\boldsymbol{z}_1, \boldsymbol{X}), \ldots, (\boldsymbol{z}_m, \boldsymbol{X})$ are then released to the public. As discussed by Reiter and Kinney[15], in this scenario the procedures of Reiter[13] for drawing inference based on $m > 1$ multiply imputed partially synthetic datasets appear to remain valid. In the specific case of $m = 1$, likelihood based methods for drawing inference in this scenario were derived by Klein and Sinha[8].

As mentioned above, in the case of singly imputed synthetic data ($m = 1$), finite sample procedures for drawing inference have appeared in Klein and Sinha[7,8] for posterior predictive sampling and plug-in sampling. The results derived in the present paper extend those finite sample procedures for $m > 1$ multiply imputed synthetic datasets. The organization of the rest of this paper is as follows. In Section 2.1, we derive finite sample inference based on synthetic data generated using posterior predictive sampling. Here we use a general form of the prior $\pi(\boldsymbol{\beta}, \sigma^2)$, involving a hyperparameter α. In Section 2.2, we carry out finite sample inference based on synthetic data generated using the plug-in sampling method. In Section 3 we review the inference procedures of Reiter[13,14] for multiply imputed partially synthetic data. In Section 4 we present results of some simulation studies, and Section 5 presents some concluding remarks. Appendix A contains proofs of theorems and results that appear in this paper. Appendix B contains some details of the simulation studies presented in Section 4.

2 Methodology

In this section we derive the finite sample inference for multiply imputed synthetic data, first under posterior predictive sampling in Section 2.1, and then for plug-in sampling in Section 2.2.

2.1 Posterior Predictive Sampling

To generate synthetic data $\boldsymbol{z}_1 = (z_{11}, \ldots, z_{1n})', \ldots, \boldsymbol{z}_m = (z_{m1}, \ldots, z_{mn})'$ under *posterior predictive sampling* we start from a joint prior distribution $\pi(\boldsymbol{\beta}, \sigma^2) \propto \frac{1}{\sigma^\alpha}$ for $\boldsymbol{\beta} \in \mathbb{R}^p$ and $\sigma^2 > 0$. The posterior distribution of $(\boldsymbol{\beta}, \sigma^2)$, given $\boldsymbol{y}$, has the representation:

$$\boldsymbol{\beta} \,\big|\, \sigma^2, \boldsymbol{y} \;\sim\; N_p\left[\boldsymbol{b}, \sigma^2(\boldsymbol{X}\boldsymbol{X}')^{-1}\right], \qquad \frac{\text{RSS}}{\sigma^2}\,\bigg|\, \boldsymbol{y} \;\sim\; \chi^2_{n+\alpha-p-2}. \tag{1}$$

We assume throughout that $n + \alpha > p + 4$. The synthetic data are generated by repeating Steps 1 and 2 below independently for each $j = 1, \ldots, m$.

Step 1. Draw $(\boldsymbol{\beta}^*_j, \sigma^{2*}_j)$ from the posterior distribution (1).

Step 2. Draw $\boldsymbol{z}_j = (z_{j1}, \ldots, z_{jn})' \sim N_n(\boldsymbol{X}'\boldsymbol{\beta}^*_j, \sigma^{2*}_j \boldsymbol{I}_n)$.

The released synthetic data are $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_m$ along with the matrix of predictor variables $\boldsymbol{X}$. The inferential procedures presented below are derived from the frequentist perspective where $\boldsymbol{\beta}$ and σ^2 are fixed but unknown quantities. In view of the sampling mechanism (Steps 1 and 2 above), it follows that from the frequentist perspective, the joint distribution of $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_m$, $\boldsymbol{\beta}^*_1, \ldots, \boldsymbol{\beta}^*_m$, $\sigma^{2*}_1, \ldots, \sigma^{2*}_m$, and $\boldsymbol{y}$ has the following hierarchical structure:

$$\begin{aligned}
\boldsymbol{z}_j \,\big|\, \boldsymbol{\beta}^*_1, \ldots, \boldsymbol{\beta}^*_m, \sigma^{2*}_1, \ldots, \sigma^{2*}_m, \boldsymbol{y} \;&\sim\; N_n(\boldsymbol{X}'\boldsymbol{\beta}^*_j,\, \sigma^{2*}_j \boldsymbol{I}_n), \text{ independently for } j = 1, \ldots, m, \\
\boldsymbol{\beta}^*_j \,\big|\, \sigma^{2*}_1, \ldots, \sigma^{2*}_m, \boldsymbol{y} \;&\sim\; N_p(\boldsymbol{b},\, \sigma^{2*}_j(\boldsymbol{X}\boldsymbol{X}')^{-1}), \text{ independently for } j = 1, \ldots, m, \\
\sigma^{2*}_1, \ldots, \sigma^{2*}_m \,\big|\, \boldsymbol{y} \;&\overset{iid}{\sim}\; \frac{\text{RSS}}{\chi^2_{n+\alpha-p-2}}, \quad \boldsymbol{y} \;\sim\; N_n(\boldsymbol{X}'\boldsymbol{\beta},\, \sigma^2 \boldsymbol{I}_n).
\end{aligned} \tag{2}$$

Therefore the joint probability density function (pdf) of $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_m$, $\boldsymbol{\beta}^*_1, \ldots, \boldsymbol{\beta}^*_m$, $\sigma^{2*}_1, \ldots, \sigma^{2*}_m$, $\boldsymbol{y}$ is $f_{\boldsymbol{\beta},\sigma^2}(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_m, \boldsymbol{\beta}^*_1, \ldots, \boldsymbol{\beta}^*_m, \sigma^{2*}_1, \ldots, \sigma^{2*}_m, \boldsymbol{y})$

$$\begin{aligned}
&= \prod_{j=1}^m (2\pi\sigma^{2*}_j)^{-n/2} \exp\left[-\frac{1}{2\sigma^{2*}_j}(\boldsymbol{z}_j - \boldsymbol{X}'\boldsymbol{\beta}^*_j)'(\boldsymbol{z}_j - \boldsymbol{X}'\boldsymbol{\beta}^*_j)\right] \\
&\times \prod_{j=1}^m (2\pi\sigma^{2*}_j)^{-p/2} |\boldsymbol{X}\boldsymbol{X}'|^{1/2} \exp\left[-\frac{1}{2\sigma^{2*}_j}(\boldsymbol{\beta}^*_j - \boldsymbol{b})'(\boldsymbol{X}\boldsymbol{X}')(\boldsymbol{\beta}^*_j - \boldsymbol{b})\right] \\
&\times \prod_{j=1}^m \frac{(\text{RSS}/2)^{(n+\alpha-p-2)/2}}{\Gamma\left(\frac{n+\alpha-p-2}{2}\right)} (\sigma^{2*}_j)^{-(n+\alpha-p-2)/2-1} \exp\left[-\frac{\text{RSS}}{2\sigma^{2*}_j}\right] \\
&\times (2\pi\sigma^2)^{-n/2} \exp\left[-\frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{X}'\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}'\boldsymbol{\beta})\right],
\end{aligned}$$

and hence the marginal pdf of $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_m$ is

$$f_{\boldsymbol{\beta},\sigma^2}(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_m) = \int \cdots \int f_{\boldsymbol{\beta},\sigma^2}(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_m, \boldsymbol{\beta}_1^*, \ldots, \boldsymbol{\beta}_m^*, \sigma_1^{2*}, \ldots, \sigma_m^{2*}, \boldsymbol{y}) d\boldsymbol{\beta}_1^* \cdots d\boldsymbol{\beta}_m^* d\sigma_1^{2*} \cdots d\sigma_m^{2*} d\boldsymbol{y}.$$

The inferential results presented below are based on the marginal distribution of $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_m$, and we will utilize the hierarchical structure (2) to derive the results.

For exact inference based on the released synthetic data, we define $\boldsymbol{b}_j^* = (\boldsymbol{X}\boldsymbol{X}')^{-1}\boldsymbol{X}\boldsymbol{z}_j$ and $\text{RSS}_j^* = (\boldsymbol{z}_j - \boldsymbol{X}'\boldsymbol{b}_j^*)'(\boldsymbol{z}_j - \boldsymbol{X}'\boldsymbol{b}_j^*)$, for $j = 1, \ldots, m$. It can be shown that $(\boldsymbol{b}_1^*, \text{RSS}_1^*), \ldots, (\boldsymbol{b}_m^*, \text{RSS}_m^*)$ are jointly sufficient for $\boldsymbol{\beta}$ and σ^2. Define $\overline{\boldsymbol{b}^*} = \frac{1}{m}\sum_{j=1}^m \boldsymbol{b}_j^*$ and $\widetilde{\text{RSS}^*} = \sum_{j=1}^m \text{RSS}_j^*$. Also let $\overline{\boldsymbol{\beta}^*} = \frac{1}{m}\sum_{j=1}^m \boldsymbol{\beta}_j^*$. Then $E(\overline{\boldsymbol{b}^*}) = E[E(\overline{\boldsymbol{b}^*}|\boldsymbol{\beta}_1^*, \ldots, \boldsymbol{\beta}_m^*, \sigma_1^{2*}, \ldots, \sigma_m^{2*})] = E(\overline{\boldsymbol{\beta}^*}) = E[E(\overline{\boldsymbol{\beta}^*}|\boldsymbol{y})] = E(\boldsymbol{b}) = \boldsymbol{\beta}$. Hence $\overline{\boldsymbol{b}^*}$ is an unbiased estimator of $\boldsymbol{\beta}$. The variance of $\overline{\boldsymbol{b}^*}$ is

$$\text{Var}(\overline{\boldsymbol{b}^*}) = \left[1 + \frac{2(n-p)}{m(n+\alpha-p-4)}\right] \sigma^2 \left(\boldsymbol{X}\boldsymbol{X}'\right)^{-1}, \tag{3}$$

where proof of (3) appears in Appendix A. Also,

$$E(\widetilde{\text{RSS}^*}) = \sigma^2 \frac{m(n-p)^2}{n+\alpha-p-4}, \tag{4}$$

implying that $\frac{(\widetilde{\text{RSS}^*})(n+\alpha-p-4)}{m(n-p)^2}$ is an unbiased estimator of σ^2. Proof of (4) appears in Appendix A. To construct an exact test and confidence set for $\boldsymbol{\beta}$, define

$$T^2 = (\overline{\boldsymbol{b}^*} - \boldsymbol{\beta})(\boldsymbol{X}\boldsymbol{X}')(\overline{\boldsymbol{b}^*} - \boldsymbol{\beta})/\widetilde{\text{RSS}^*}.$$

Then we have the following distributional result about T^2, whose proof is in Appendix A.

Theorem 1 The distribution of T^2 can be represented as $T^2 \sim T_1 \times T_2$ where T_1 and T_2 are independently distributed. Furthermore,

$$T_1 \sim \chi_p^2, \text{ and } T_2 \sim \frac{1 + 2m^{-2}A\sum_{j=1}^m (1/B_j)}{A\sum_{j=1}^m (C_j/B_j)},$$

where $A, B_1, \ldots, B_m, C_1, \ldots, C_m$ are independently distributed with $A \sim \chi_{n-p}^2$, $B_j \sim \chi_{n+\alpha-p-2}^2$, and $C_j \sim \chi^2_{n-p}$ for $j = 1, \ldots, m$.

A $(1-\gamma)$ confidence region for $\boldsymbol{\beta}$ can be obtained as follows. For given values of γ, p, m, n, α, let $d_{\gamma,p,m,n,\alpha}$ satisfy $1-\gamma = P(T^2 \leq d_{\gamma,p,m,n,\alpha})$. Notice that it is straightforward to simulate from the distribution of T^2 using the representation of the distribution given in Theorem 1, and therefore, one can compute $d_{\gamma,p,m,n,\alpha}$ using Monte Carlo simulation. Then a $(1-\gamma)$ confidence region for $\boldsymbol{\beta}$ based on T^2 is

$$\left\{ \boldsymbol{\beta} : \frac{(\overline{\boldsymbol{b}^*} - \boldsymbol{\beta})'(\boldsymbol{X}\boldsymbol{X}')(\overline{\boldsymbol{b}^*} - \boldsymbol{\beta})}{\widetilde{\text{RSS}^*}} \leq d_{\gamma,p,m,n,\alpha} \right\}, \tag{5}$$

with its volume given by $\frac{\pi^{p/2}}{\Gamma(\frac{p}{2}+1)}(d_{\gamma,p,m,n,\alpha})^{p/2} \left|\boldsymbol{X}\boldsymbol{X}'\right|^{-1/2} (\widetilde{\text{RSS}^*})^{p/2}$. To compute the expected volume, it can be shown that $E[(\widetilde{\text{RSS}^*})^{p/2}] = \sigma^p E[(\chi^2_{n-p})^{\frac{p}{2}}] E[\sum_{j=1}^m \frac{\chi^2_{n-p;j}}{\chi^2_{n+\alpha-p-2;j}}]^{p/2}$, where all the χ^2 variables are independent.

Let $\boldsymbol{A}$ be a $k \times p$ dimensional matrix with $\text{rank}(\boldsymbol{A}) = k < p$. Inference about $\boldsymbol{\eta} = \boldsymbol{A}\boldsymbol{\beta}$ easily follows upon noting that, based on the existing arguments presented in the proof of Theorem 1,

$$\begin{aligned} \overline{\boldsymbol{b}^*}|\Delta \sim N_p\left(\boldsymbol{\beta}, (\sigma^2 + 2\Delta)\left(\boldsymbol{X}\boldsymbol{X}'\right)^{-1}\right) &\implies \boldsymbol{A}\overline{\boldsymbol{b}^*}|\Delta \sim N_k\left(\boldsymbol{\eta}, (\sigma^2+2\Delta)\boldsymbol{A}\left(\boldsymbol{X}\boldsymbol{X}'\right)^{-1}\boldsymbol{A}'\right) \\ &\implies \frac{(\boldsymbol{A}\overline{\boldsymbol{b}^*} - \boldsymbol{\eta})'[\boldsymbol{A}(\boldsymbol{X}\boldsymbol{X}')^{-1}\boldsymbol{A}']^{-1}(\boldsymbol{A}\overline{\boldsymbol{b}^*} - \boldsymbol{\eta})}{\sigma^2 + 2\Delta} \sim \chi^2_k, \end{aligned}$$

where $\Delta = \sum_{j=1}^m \frac{\sigma_j^{2*}}{m^2}$. Defining $T^2_{\boldsymbol{\eta}} = (\boldsymbol{A}\overline{\boldsymbol{b}^*} - \boldsymbol{\eta})'[\boldsymbol{A}(\boldsymbol{X}\boldsymbol{X}')^{-1}\boldsymbol{A}']^{-1}(\boldsymbol{A}\overline{\boldsymbol{b}^*} - \boldsymbol{\eta})/\widetilde{\text{RSS}^*}$, it then follows that $T^2_{\boldsymbol{\eta}}$ is distributed as the product of $T_{1,k}$ and T_2, where $T_{1,k} \sim \chi^2_k$, the distribution of T_2 is defined in Theorem 1, and $T_{1,k}$ and T_2 are independent. By simulating the distribution of $T^2_{\boldsymbol{\eta}}$ we can compute the value $\delta_{k,\gamma,p,m,n,\alpha}$ satisfying $1-\gamma = P(T^2_{\boldsymbol{\eta}} \leq \delta_{k,\gamma,p,m,n,\alpha})$, and obtain a $(1-\gamma)$ confidence region for η as

$$\left\{ \boldsymbol{\eta} : \frac{(\boldsymbol{A}\overline{\boldsymbol{b}^*} - \boldsymbol{\eta})'[\boldsymbol{A}(\boldsymbol{X}\boldsymbol{X}')^{-1}\boldsymbol{A}']^{-1}(\boldsymbol{A}\overline{\boldsymbol{b}^*} - \boldsymbol{\eta})}{\widetilde{\text{RSS}^*}} \leq \delta_{k,\gamma,p,m,n,\alpha} \right\}. \tag{6}$$

In particular, taking $\boldsymbol{A}$ to be a $1 \times p$ dimensional vector having a 1 in column i, and 0 in all other columns, we see that inference about a single regression coefficient β_i can be based on $t_i^2 = \frac{(\overline{b_i^*} - \beta_i)^2}{D_{ii}\widetilde{\text{RSS}^*}}$ where D_{ii} is the ith diagonal element of $(\boldsymbol{X}\boldsymbol{X}')^{-1}$. From the preceding discussion it is clear that the distribution of t_i^2 is that of the product of two independent random variables, $T_{1,1}$ and T_2, where

$T_{1,1} \sim \chi_1^2$ and the distribution of T_2 is defined in Theorem 1. The resulting $(1-\gamma)$ confidence interval for β_i is

$$\left[\overline{b_i^*} - \left(D_{ii} \times \widetilde{\mathrm{RSS}^*} \times \delta_{1,\gamma,p,m,n,\alpha}\right)^{1/2},\ \overline{b_i^*} + \left(D_{ii} \times \widetilde{\mathrm{RSS}^*} \times \delta_{1,\gamma,p,m,n,\alpha}\right)^{1/2}\right]. \tag{7}$$

2.2 Plug-in Sampling

To generate synthetic data $\boldsymbol{z}_1 = (z_{11}, \ldots, z_{1n})', \ldots, \boldsymbol{z}_m = (z_{m1}, \ldots, z_{mn})'$ under *plug-in* sampling, we start from the point estimates $\boldsymbol{b}$ and $\mathrm{RSS}/(n-p)$, of $\boldsymbol{\beta}$ and σ^2, respectively. The synthetic data are obtained by drawing $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_m$ as *iid* from $N_n\left(\boldsymbol{X}'\boldsymbol{b}, \frac{\mathrm{RSS}}{n-p}\boldsymbol{I}_n\right)$. Equivalently, the synthetic data are obtained by drawing $z_{ji} \sim N(\boldsymbol{x}_i'\boldsymbol{b}, \frac{\mathrm{RSS}}{n-p})$, independently for $i = 1, \ldots, n$ and $j = 1, \ldots, m$. We will now proceed to derive inferential procedures from the frequentist perspective where $\boldsymbol{\beta}$ and σ^2 are fixed but unknown quantities. In view of the sampling mechanism, it follows that the joint distribution of $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_m$ and $\boldsymbol{y}$ has the following hierarchical structure:

$$\begin{aligned} \boldsymbol{z}_1, \ldots, \boldsymbol{z}_m \,\big|\, \boldsymbol{y} &\overset{iid}{\sim} N_n(\boldsymbol{X}'\boldsymbol{b}, \frac{\mathrm{RSS}}{n-p}\boldsymbol{I}_n) \quad \left[\text{equivalently}, z_{ji} \overset{\text{independent}}{\sim} N\left(\boldsymbol{x}_i'\boldsymbol{b}, \frac{\mathrm{RSS}}{n-p}\right)\right], \\ \boldsymbol{y} &\sim N_n(\boldsymbol{X}'\boldsymbol{\beta}, \sigma^2\boldsymbol{I}_n). \end{aligned} \tag{8}$$

Therefore the joint probability density function (pdf) of $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_m$, and $\boldsymbol{y}$ is

$$\begin{aligned} f_{\boldsymbol{\beta},\sigma^2}(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_m, \boldsymbol{y}) = \prod_{j=1}^{m} \left(2\pi\frac{\mathrm{RSS}}{n-p}\right)^{-n/2} &\exp\left[-\frac{1}{2\frac{\mathrm{RSS}}{n-p}}(\boldsymbol{z}_j - \boldsymbol{X}'\boldsymbol{b})'(\boldsymbol{z}_j - \boldsymbol{X}'\boldsymbol{b})\right] \\ \times (2\pi\sigma^2)^{-n/2} &\exp\left[-\frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{X}'\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}'\boldsymbol{\beta})\right], \end{aligned}$$

and hence the marginal pdf of $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_m$ is $f_{\boldsymbol{\beta},\sigma^2}(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_m) = \int f_{\boldsymbol{\beta},\sigma^2}(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_m, \boldsymbol{y})d\boldsymbol{y}$. The inferential results presented below are based on the marginal distribution of $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_m$, and we will utilize the hierarchical structure (8) to derive the results.

We now provide an exact inference procedure for $\boldsymbol{\beta}$ based on $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_m$. Recall that each $\boldsymbol{z}_j$ is the $n \times 1$ vector $(z_{j1}, \ldots, z_{jn})'$ and, conditionally given $\boldsymbol{b}$ and RSS, $z_{ji} \sim N(\boldsymbol{x}_i'\boldsymbol{b}, \frac{\mathrm{RSS}}{n-p})$. Therefore, conditional on $\boldsymbol{b}$ and RSS, $(z_{1i}, \ldots, z_{mi})$ is a random sample from $N(\boldsymbol{x}_i'\boldsymbol{b}, \frac{\mathrm{RSS}}{n-p})$ for arbitrary but fixed $i = 1, \ldots, n$. Let $\bar{z}_i = \frac{1}{m}\sum_{j=1}^{m} z_{ji}$, $S_{zi}^2 = \sum_{j=1}^{m}(z_{ji} - \bar{z}_i)^2$, and $S_z^2 = \sum_{i=1}^{n} S_{zi}^2$. If $m > 1$, then it follows that, conditional on $\boldsymbol{b}$ and RSS,

$$S_z^2 \sim \left[\frac{\text{RSS}}{(n-p)}\right]\chi^2_{n(m-1)}, \quad \bar{z}_i \sim N\left[\boldsymbol{x}_i'\boldsymbol{b}, \frac{\text{RSS}}{m(n-p)}\right], \ i = 1, \ldots, n,$$

with these terms being (conditionally) independent. If $m = 1$, then the situation reduces to $\bar{z}_i = z_{1i}$ and $S_{zi}^2 = 0$ for $i = 1, \ldots, n$, and hence $S_z^2 = 0$.

Let $\bar{\boldsymbol{z}} = (\bar{z}_1, \ldots, \bar{z}_n)'$ and $\boldsymbol{b}_j^* = (\boldsymbol{X}\boldsymbol{X}')^{-1}\boldsymbol{X}\boldsymbol{z}_j$. We define $\overline{\boldsymbol{b}^*} = (\boldsymbol{X}\boldsymbol{X}')^{-1}\boldsymbol{X}\bar{\boldsymbol{z}} = \frac{1}{m}\sum_{j=1}^{m}\boldsymbol{b}_j^*$ and $S^2_{\text{comb}} = S_z^2 + m(\bar{\boldsymbol{z}} - \boldsymbol{X}'\overline{\boldsymbol{b}^*})'(\bar{\boldsymbol{z}} - \boldsymbol{X}'\overline{\boldsymbol{b}^*})$, and note that, conditionally given $\boldsymbol{b}$ and RSS,

$$\overline{\boldsymbol{b}^*} \sim N_p\left[\boldsymbol{b}, \frac{\text{RSS}}{m(n-p)}(\boldsymbol{X}\boldsymbol{X}')^{-1}\right], \quad S^2_{\text{comb}} \sim \left[\frac{\text{RSS}}{(n-p)}\right]\chi^2_{n(m-1)+n-p}.$$

Then $E(\overline{\boldsymbol{b}^*}) = E[E(\overline{\boldsymbol{b}^*}|\boldsymbol{y})] = E(\boldsymbol{b}) = \boldsymbol{\beta}$. Thus $\overline{\boldsymbol{b}^*}$ is an unbiased estimator of $\boldsymbol{\beta}$ whose variance is given by

$$\begin{aligned}\text{Var}(\overline{\boldsymbol{b}^*}) &= E[\text{Var}(\overline{\boldsymbol{b}^*}|\boldsymbol{y})] + \text{Var}[E(\overline{\boldsymbol{b}^*}|\boldsymbol{y})] = E\left[\frac{\text{RSS}}{m(n-p)}(\boldsymbol{X}\boldsymbol{X}')^{-1}\right] + \text{Var}(\boldsymbol{b}) \\ &= \frac{\sigma^2}{m}(\boldsymbol{X}\boldsymbol{X}')^{-1} + \sigma^2(\boldsymbol{X}\boldsymbol{X}')^{-1} = \sigma^2\left(1 + \frac{1}{m}\right)(\boldsymbol{X}\boldsymbol{X}')^{-1}.\end{aligned}$$

Also note that $E(S^2_{\text{comb}}) = E[E(S^2_{\text{comb}}|\boldsymbol{y})] = E[\frac{\text{RSS}}{n-p}(nm-p)] = \sigma^2(nm-p)$; hence $S^2_{\text{comb}}/(mn-p)$ is an unbiased estimator of σ^2. To construct an exact test and confidence set for $\boldsymbol{\beta}$, define

$$T^2_{\text{comb}} = (\overline{\boldsymbol{b}^*} - \boldsymbol{\beta})'(\boldsymbol{X}\boldsymbol{X}')(\overline{\boldsymbol{b}^*} - \boldsymbol{\beta})/S^2_{\text{comb}}.$$

We have the following distributional result about T^2_{comb}, whose proof is in Appendix A.

Theorem 2 The distribution of T^2_{comb} can be represented as follows:

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

To obtain a $(1-\gamma)$ confidence set for $\boldsymbol{\beta}$, for given values of γ, p, m, n, let $\ell_{\gamma,p,m,n}$ be such that

$1-\gamma = P(T^2_{\text{comb}} \leq \ell_{\gamma,p,m,n})$. The value $\ell_{\gamma,p,m,n}$ can be computed using Monte Carlo simulation, by using Theorem 2 to simulate from the distribution of T^2_{comb}. A $(1-\gamma)$ level confidence set for $\boldsymbol{\beta}$ based on T^2 is

$$\left\{\boldsymbol{\beta} : \frac{(\overline{\boldsymbol{b}^*} - \boldsymbol{\beta})'(\boldsymbol{X}\boldsymbol{X}')(\overline{\boldsymbol{b}^*} - \boldsymbol{\beta})}{S^2_{\text{comb}}} \leq \ell_{\gamma,p,m,n}\right\}, \tag{9}$$

and its volume is $\frac{\pi^{p/2}}{\Gamma(\frac{p}{2}+1)}(\ell_{\gamma,p,m,n})^{p/2}\,|\boldsymbol{X}\boldsymbol{X}'|^{-1/2}\,(S^2_{\text{comb}})^{p/2}$. The expected volume can be computed using $E[(S^2_{\text{comb}})^{p/2}] = E[(\text{RSS})^{p/2}]E[(\chi^2_{mn-p})^{p/2}]/(n-p)^{p/2} = \sigma^p E[(\chi^2_{n-p})^{p/2}]E[(\chi^2_{mn-p})^{p/2}]/(n-p)^{p/2}$.

Let $\boldsymbol{A}$ be a $k \times p$ dimensional matrix with $\text{rank}(\boldsymbol{A}) = k < p$. Inference about $\boldsymbol{\eta} = \boldsymbol{A}\boldsymbol{\beta}$ easily follows upon noting that, conditional on $\boldsymbol{b}$ and RSS,

$$\boldsymbol{A}\overline{\boldsymbol{b}^*} \sim N_k\left(\boldsymbol{A}\boldsymbol{b}, \frac{\text{RSS}}{m(n-p)}\boldsymbol{A}(\boldsymbol{X}\boldsymbol{X}')^{-1}\boldsymbol{A}'\right), \quad S^2_{\text{comb}} \sim \frac{\text{RSS}}{n-p}\chi^2_{mn-p},$$

along with their (conditional) independence. Thus, by arguments entirely analogous to those presented in the proof of Theorem 2, we get

$$T^2_{\text{comb};\boldsymbol{\eta}}|\psi \sim \frac{k}{m(mn-p)}\left[1 + \frac{m(n-p)}{\psi}\right]F_{k,mn-p},$$

where $T^2_{\text{comb};\boldsymbol{\eta}} = (\boldsymbol{A}\overline{\boldsymbol{b}^*} - \boldsymbol{\eta})'[\boldsymbol{A}(\boldsymbol{X}\boldsymbol{X}')^{-1}\boldsymbol{A}']^{-1}(\boldsymbol{A}\overline{\boldsymbol{b}^*} - \boldsymbol{\eta})/S^2_{\text{comb}}$, and the distribution of ψ is the same is an Theorem 2. Thus, by simulating the distribution of $T^2_{\text{comb};\boldsymbol{\eta}}$, we can compute the constant $\lambda_{k,\gamma,p,m,n}$ satisfying $1-\gamma = P(T^2_{\text{comb};\boldsymbol{\eta}} \leq \lambda_{k,\gamma,p,m,n})$, and obtain a $(1-\gamma)$ confidence region for $\boldsymbol{\eta}$ as

$$\left\{\boldsymbol{\eta} : \frac{(\boldsymbol{A}\overline{\boldsymbol{b}^*} - \boldsymbol{\eta})'[\boldsymbol{A}(\boldsymbol{X}\boldsymbol{X}')^{-1}\boldsymbol{A}']^{-1}(\boldsymbol{A}\overline{\boldsymbol{b}^*} - \boldsymbol{\eta})}{S^2_{\text{comb}}} \leq \lambda_{k,\gamma,p,m,n}\right\}. \tag{10}$$

In particular, taking $\boldsymbol{A}$ to be a $1 \times p$ dimensional vector having a 1 in column i, and 0 in all other columns, we see that inference about a single regression coefficient β_i can be based on $t_i^2 = \frac{(\overline{b_i^*} - \beta_i)^2}{D_{ii}S^2_{\text{comb}}}$, where D_{ii} is the ith diagonal element of $(\boldsymbol{X}\boldsymbol{X}')^{-1}$. Obviously, the distribution of t_i^2 is the same as that of $T^2_{\text{comb};\boldsymbol{\eta}}$ with $k=1$, and the resulting $(1-\gamma)$ confidence interval for β_i is

$$\left[\overline{b_i^*} - \left(D_{ii} \times S^2_{\text{comb}} \times \lambda_{1,\gamma,p,m,n}\right)^{1/2},\ \overline{b_i^*} + \left(D_{ii} \times S^2_{\text{comb}} \times \lambda_{1,\gamma,p,m,n}\right)^{1/2}\right]. \tag{11}$$

3 Review of Standard Methodology for Multiply Imputed Partially Synthetic Data

In this section we review the state of the art methodology for drawing approximately valid inference based on multiply imputed partially synthetic data. This methodology was developed by Reiter [13] for a scalar-valued parameter of interest, and extended by Reiter [14] to a vector-valued parameter of interest. Reiter [13,14] assumes that the partially synthetic data are generated using posterior predictive sampling, however, Reiter and Kinney [15] indicate that the methodology still yields valid inference when partially synthetic data are generated using plug-in sampling. We shall now summarize the application of these procedures under our linear regression scenario, however, we should re-iterate that these procedures are quite general; these yield approximately valid inference under a variety of models and for a variety of estimands. This methodology requires the availability of multiply imputed synthetic datasets, it cannot be applied if only a singly imputed synthetic dataset is released.

In our notation, the synthetic data are $(\boldsymbol{z}_1, \boldsymbol{X}), \ldots, (\boldsymbol{z}_m, \boldsymbol{X})$ where $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_m$ are the synthetic copies of $\boldsymbol{y}$, generated either using posterior predictive sampling as in Section 2.1, or plug-in sampling as in Section 2.2. Let $\boldsymbol{Q} = \boldsymbol{Q}(\boldsymbol{\beta}, \sigma^2)$ be the parameter of interest, which is a function of $\boldsymbol{\beta}$ and σ^2.

Inference for a Scalar-Valued Parameter. Suppose the parameter of interest $Q = Q(\boldsymbol{\beta}, \sigma^2)$ is a scalar. Let $q = q(\boldsymbol{y}, \boldsymbol{X})$ be a point estimator of Q based on the original data, and let $u = u(\boldsymbol{y}, \boldsymbol{X})$ be an estimator of the variance of q, also based on the original data. Let $q_j = q(\boldsymbol{z}_j, \boldsymbol{X})$ and $u_j = u(\boldsymbol{z}_j, \boldsymbol{X})$ be the values of q and u, respectively, when computed on the jth synthetic dataset. Define

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

Then $\bar{q}_m$ is an estimate of Q, the variance of $\bar{q}_m$ is approximated by $T_m = \mathcal{B}_m/m + \bar{u}_m$, and the distribution of $(\bar{q}_m - Q)/\sqrt{T_m}$ is approximated by a t distribution with $\nu = (m-1)\left[1 + \frac{\bar{u}_m}{\mathcal{B}_m/m}\right]^2$ degrees of freedom. The quantity $(\bar{q}_m - Q)/\sqrt{T_m}$ is used, along with its approximate t distribution, to obtain a confidence interval and significance test for Q.

Inference for a Vector-Valued Parameter. Suppose the parameter of interest $\boldsymbol{Q} = \boldsymbol{Q}(\boldsymbol{\beta}, \sigma^2)$ is a $k \times 1$ dimensional vector. Let $\boldsymbol{q} = \boldsymbol{q}(\boldsymbol{y}, \boldsymbol{X})$ be a point estimator of $\boldsymbol{Q}$ based on the original data, and let $\boldsymbol{u} = \boldsymbol{u}(\boldsymbol{y}, \boldsymbol{X})$ be an estimator of the covariance matrix of $\boldsymbol{q}$, also based on the original data. Let $\boldsymbol{q}_j = \boldsymbol{q}(\boldsymbol{z}_j, \boldsymbol{X})$ and $\boldsymbol{u}_j = \boldsymbol{u}(\boldsymbol{z}_j, \boldsymbol{X})$ be the values of $\boldsymbol{q}$ and $\boldsymbol{u}$, respectively, when computed on the jth synthetic dataset. Define

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

Then $\bar{\boldsymbol{q}}_m$ is an estimate of $\boldsymbol{Q}$, and the covariance matrix of $\bar{\boldsymbol{q}}_m$ is approximated by $\boldsymbol{T}_m = \boldsymbol{\mathcal{B}}_m/m + \bar{\boldsymbol{u}}_m$. Define $\boldsymbol{\mathcal{S}}_{\boldsymbol{m}} = (\bar{\boldsymbol{q}}_m - \boldsymbol{Q})'(\bar{\boldsymbol{u}}_m)^{-1}(\bar{\boldsymbol{q}}_m - \boldsymbol{Q})/[k(1+r)]$ where $r = \mathrm{tr}(\boldsymbol{\mathcal{B}}_m \bar{\boldsymbol{u}}_m^{-1})/(mk)$. The distribution of $\boldsymbol{\mathcal{S}}_m$ 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 $\boldsymbol{\mathcal{S}}_m$ is used, along with its approximate F distribution, to obtain a confidence region and significance test for $\boldsymbol{Q}$. Alternative methods of inference based on the log-likelihood ratio test statistic from the m synthetic datasets are also developed by Reiter[14].

4 Simulation Studies

In this section we present simulation results in order to demonstrate that the methodology developed in Section 2 performs as our theory predicts, and to compare the proposed methodology with the state of the art methodology which we reviewed in Section 3. All simulations were performed using the statistical computing software R (R Core Team[11]). To perform the simulations, we define $\boldsymbol{X}$, $\boldsymbol{\beta}$ and σ^2 as explained in Appendix B. Note that this simulation model was also used by Klein and Sinha[8] to perform simulation studies to evaluate finite sample methodology for singly imputed synthetic data generated under plug-in sampling. To conduct the simulations for a given sample size n, we generate $\boldsymbol{X}$ one time as described in Appendix B, and then hold it fixed from one iteration to the next. Using Monte Carlo simulation with 10^6 iterations, we compute an estimate of the coverage probability and expected volume or expected length (as appropriate) of the following confidence regions for $\boldsymbol{\beta}$ and the following confidence intervals for β_2.

1. The confidence region (5) for $\boldsymbol{\beta}$ under posterior predictive sampling.

2. The confidence region (9) for $\boldsymbol{\beta}$ under plug-in sampling.

3. The confidence region for $\boldsymbol{\beta}$ obtained using the methodology of Reiter[14] as reviewed in Section 3, where we take $\boldsymbol{Q} = \boldsymbol{\beta}$, $\boldsymbol{q}(\boldsymbol{y}, \boldsymbol{X}) = (\boldsymbol{X}\boldsymbol{X}')^{-1}\boldsymbol{X}\boldsymbol{y} = \boldsymbol{b}$, and $\boldsymbol{u}(\boldsymbol{y}, \boldsymbol{X}) = \frac{\text{RSS}}{n-p}(\boldsymbol{X}\boldsymbol{X}')^{-1}$.

4. The confidence interval (7) for β_2 under posterior predictive sampling.

5. The confidence interval (11) for β_2 under plug-in sampling.

6. The confidence interval for β_2 obtained using the methodology of Reiter[13] as reviewed in Section 3, where we take $Q = \beta_2$, $q(\boldsymbol{y}, \boldsymbol{X}) = b_2$, $u(\boldsymbol{y}, \boldsymbol{X}) = \frac{\text{RSS}}{n-p} D_{22}$.

For the sake of comparison, we also compute a Monte Carlo estimate of the coverage probability and expected volume or expected length of the standard confidence region for $\boldsymbol{\beta}$ and standard confidence interval for β_2, based on the original data, which are obtained from the standard results (Rencher and Schaalje[16]):

$$\frac{(\boldsymbol{b} - \boldsymbol{\beta})'(\boldsymbol{X}\boldsymbol{X}')(\boldsymbol{b} - \boldsymbol{\beta})}{\left(\frac{\text{RSS}}{n-p}\right) p} \sim F_{p,n-p} \quad \text{and} \quad \frac{b_i - \beta_i}{\sqrt{D_{ii}\left(\frac{\text{RSS}}{n-p}\right)}} \sim t_{n-p}.$$

The simulation results related to $\boldsymbol{\beta}$ are displayed in Tables 1 and 2, where Table 1 gives the results under posterior predictive sampling, while Table 2 gives the results under plug-in sampling. Simulation results related to β_2 are displayed in Tables 3 and 4, where Table 3 give the results under posterior predictive sampling, while Table 4 gives the results under plug-in sampling. In Tables 1 - 4, results for the original data confidence regions/intervals are displayed under the heading *Original Data*, results for the confidence regions/intervals derived in Section 2 are displayed under the heading *Finite Sample Analysis*, and results for the confidence regions/intervals reviewed in Section 3 are displayed under the heading *Combination Formula Based Analysis*. Also in Tables 1 - 4 the Monte Carlo estimates of coverage probability appear under the heading *cvg* (coverage). We have plotted the Monte Carlo estimates of coverage probability under the combination formula based analysis in Figure 1, where the estimates from Tables 1 - 4 are plotted in Figure 1(a) - 1(d), respectively. In Figure 1 we have not plotted the Monte Carlo estimates of coverage probability under our proposed finite sample analysis because these values are approximately on the horizontal line at 0.95 in all simulation scenarios considered (as predicted from the theory in Section 2). In Tables 1 and 2, the columns labeled *rel vol* (relative volume) give the Monte Carlo estimate of the expected volume of the confidence region divided by the Monte Carlo estimate of the expected volume of the corresponding original data confidence region. Similarly, in Tables 3 and 4, the columns labeled *rel len* (relative length) give the Monte Carlo estimate of the expected length of the confidence interval divided by the Monte Carlo estimate of the expected length of the corresponding original data confidence interval. These tables show results for $n = 50, 100, 200, 500, 1000, 2000, 3000$ and $m = 2, 5, 10$; for the finite sample analysis we also show results for $m = 1$, but the combination formula based analysis cannot be applied in the $m = 1$ case. In all cases the nominal level of the confidence region/interval is set at 0.95.

The following is a summary of the results of Tables 1 - 4 and Figure 1.

1. We observe that for all values of n and all values of m that we consider, the finite sample confidence regions/intervals of Section 2 have coverage equal to the nominal value of 0.95. This finding is true for both the methodology of Section 2.1 under posterior predictive sampling, and the methodology of Section 2.2 under plug-in sampling. Thus the proposed methodology performs as the theory predicts.

2. As expected, the combination formula based analysis performs well as long as the sample size is sufficiently large and the number of imputations is not too small. We observe that if $m = 2$, then the combination formula based analysis yields confidence regions/intervals whose coverage is not quite equal to the nominal value of 0.95, and this statement is true for both large and small values of n. Similarly, we observe that if the sample size n is too small, then the combination formula based analysis yields confidence regions/intervals whose coverage is not quite equal to the nominal value of 0.95, and this is true for all values of m considered in the simulation study. However, for the cases $m = 5$ and $m = 10$, the combination formula based analysis yields confidence regions/intervals whose coverage appears to converge to the nominal value of 0.95 as the sample size increases.

Table 1: Inference for $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_{10})'$ when synthetic data are generated via posterior predictive sampling

	Original Data		*Synthetic Data*													
			Finite Sample Analysis								*Combination Formula Based Analysis*					
			$m=1$		$m=2$		$m=5$		$m=10$		$m=2$		$m=5$		$m=10$	
n	*cvg*	*rel vol*	*cvg*	*rel vol*	*cvg*	*rel vol*	*cvg*	*rel vol*	*cvg*	*rel vol*	*cvg*	*rel vol*	*cvg*	*rel vol*	*cvg*	*rel vol*
50	0.950	1.000	0.950	854.938	0.950	55.024	0.950	6.408	0.950	2.658	0.942	149.133	0.921	3.729	0.917	1.467
100	0.950	1.000	0.950	424.439	0.950	40.628	0.950	5.763	0.950	2.562	0.951	180.062	0.937	4.842	0.935	1.977
200	0.950	1.000	0.950	315.993	0.950	35.833	0.950	5.536	0.949	2.502	0.954	196.294	0.945	5.413	0.943	2.247
500	0.950	1.000	0.950	268.660	0.950	33.526	0.951	5.469	0.950	2.480	0.956	205.303	0.948	5.767	0.947	2.413
1000	0.950	1.000	0.950	255.270	0.950	32.832	0.949	5.368	0.950	2.504	0.956	208.201	0.949	5.882	0.949	2.470
2000	0.950	1.000	0.950	248.627	0.950	32.374	0.950	5.371	0.950	2.492	0.957	210.878	0.950	5.943	0.949	2.498
3000	0.950	1.000	0.950	247.252	0.950	32.133	0.950	5.382	0.950	2.496	0.956	211.037	0.950	5.969	0.950	2.507

Table 2: Inference for $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_{10})'$ when synthetic data are generated via plug-in sampling

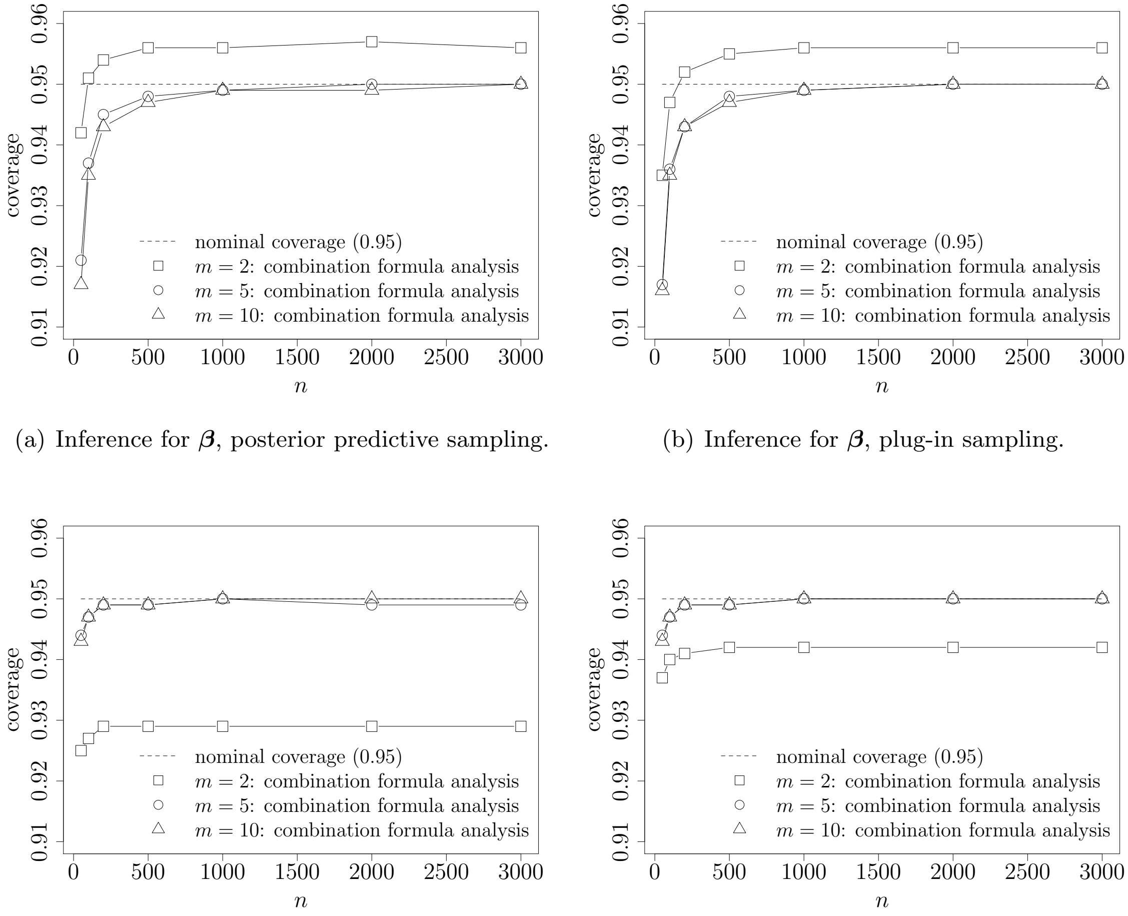
	Original Data		*Synthetic Data*													
			Finite Sample Analysis								*Combination Formula Based Analysis*					
			$m=1$		$m=2$		$m=5$		$m=10$		$m=2$		$m=5$		$m=10$	
n	*cvg*	*rel vol*	*cvg*	*rel vol*	*cvg*	*rel vol*	*cvg*	*rel vol*	*cvg*	*rel vol*	*cvg*	*rel vol*	*cvg*	*rel vol*	*cvg*	*rel vol*
50	0.950	1.000	0.950	62.709	0.949	9.270	0.950	2.592	0.950	1.621	0.935	11.942	0.917	1.469	0.916	0.896
100	0.950	1.000	0.950	43.602	0.950	8.463	0.950	2.547	0.950	1.620	0.947	16.377	0.936	2.005	0.935	1.240
200	0.950	1.000	0.950	37.152	0.950	8.062	0.950	2.526	0.950	1.610	0.952	18.800	0.943	2.291	0.943	1.426
500	0.950	1.000	0.950	33.965	0.951	7.781	0.950	2.504	0.950	1.610	0.955	20.273	0.948	2.465	0.947	1.540
1000	0.950	1.000	0.950	32.934	0.950	7.676	0.950	2.493	0.950	1.610	0.956	20.838	0.949	2.524	0.949	1.579
2000	0.950	1.000	0.950	32.427	0.950	7.642	0.950	2.492	0.950	1.612	0.956	21.025	0.950	2.553	0.950	1.598
3000	0.950	1.000	0.950	32.220	0.949	7.603	0.950	2.495	0.950	1.609	0.956	21.058	0.950	2.564	0.950	1.605

Table 3: Inference for β_2 when synthetic data are generated via posterior predictive sampling

	Original Data		*Synthetic Data*													
			Finite Sample Analysis								*Combination Formula Based Analysis*					
			$m = 1$		$m = 2$		$m = 5$		$m = 10$		$m = 2$		$m = 5$		$m = 10$	
n	*cvg*	*rel len*	*cvg*	*rel len*	*cvg*	*rel len*	*cvg*	*rel len*	*cvg*	*rel len*	*cvg*	*rel len*	*cvg*	*rel len*	*cvg*	*rel len*
50	0.950	1.000	0.950	1.743	0.950	1.414	0.950	1.179	0.950	1.093	0.925	2.127	0.944	1.173	0.943	1.065
100	0.950	1.000	0.949	1.735	0.950	1.416	0.950	1.183	0.950	1.095	0.927	2.161	0.947	1.195	0.947	1.084
200	0.950	1.000	0.950	1.733	0.950	1.415	0.950	1.185	0.950	1.095	0.929	2.177	0.949	1.204	0.949	1.092
500	0.950	1.000	0.949	1.731	0.950	1.414	0.950	1.185	0.950	1.095	0.929	2.183	0.949	1.209	0.949	1.096
1000	0.950	1.000	0.950	1.731	0.950	1.414	0.951	1.185	0.950	1.095	0.929	2.184	0.950	1.210	0.950	1.098
2000	0.950	1.000	0.950	1.732	0.950	1.414	0.950	1.185	0.950	1.096	0.929	2.188	0.949	1.211	0.950	1.099
3000	0.950	1.000	0.950	1.730	0.950	1.414	0.949	1.182	0.950	1.096	0.929	2.187	0.949	1.212	0.950	1.099

Table 4: Inference for β_2 when synthetic data are generated via plug-in sampling

	Original Data		*Synthetic Data*													
			Finite Sample Analysis								*Combination Formula Based Analysis*					
			$m = 1$		$m = 2$		$m = 5$		$m = 10$		$m = 2$		$m = 5$		$m = 10$	
n	*cvg*	*rel len*	*cvg*	*rel len*	*cvg*	*rel len*	*cvg*	*rel len*	*cvg*	*rel len*	*cvg*	*rel len*	*cvg*	*rel len*	*cvg*	*rel len*
50	0.950	1.000	0.950	1.423	0.950	1.223	0.950	1.095	0.950	1.047	0.937	1.480	0.944	1.071	0.943	1.018
100	0.950	1.000	0.950	1.419	0.950	1.226	0.950	1.094	0.950	1.048	0.940	1.500	0.947	1.090	0.947	1.036
200	0.950	1.000	0.950	1.416	0.950	1.226	0.950	1.094	0.950	1.049	0.941	1.510	0.949	1.098	0.949	1.043
500	0.950	1.000	0.949	1.414	0.950	1.226	0.950	1.095	0.950	1.050	0.942	1.517	0.949	1.103	0.949	1.047
1000	0.950	1.000	0.950	1.414	0.950	1.225	0.950	1.095	0.950	1.049	0.942	1.517	0.950	1.104	0.950	1.049
2000	0.950	1.000	0.950	1.414	0.950	1.225	0.950	1.095	0.950	1.049	0.942	1.517	0.950	1.105	0.950	1.049
3000	0.950	1.000	0.950	1.414	0.950	1.226	0.950	1.097	0.950	1.050	0.942	1.517	0.950	1.105	0.950	1.049



(a) Inference for $\boldsymbol{\beta}$, posterior predictive sampling.

(b) Inference for $\boldsymbol{\beta}$, plug-in sampling.

(c) Inference for β_2, posterior predictive sampling.

(d) Inference for β_2, plug-in sampling.

Figure 1: Monte Carlo estimates of confidence region/interval coverage.

3. In Tables 1 and 2 and Figures 1(a) and 1(b), we observe that for the combination formula based analysis, for each n the coverage decreases when m increases from 2 to 5. In particular, we note that for $m = 2$, the coverage exceeds the nominal value of 0.95 for most values of n. Thus the observed decrease in coverage when m increases from 2 to 5 may occur because $m = 2$ may be too small to invoke the approximations used to justify the combination formula based analysis. Also, in Table 2 we observe that for the combination formula based analysis, when $m = 10$ and $n = 50$, the relative volume is less than 1. This feature may occur because the sample size $n = 50$ in this scenario is again too small to invoke the approximations used to justify the combination formula based analysis, as evident from the fact that the corresponding coverage is less than the nominal value of 0.95. For larger values of n, where the coverage is approximately 0.95, the relative volume exceeds 1.

4. In cases where the sample size and number of imputations is sufficiently large so that the combination formula based analysis yields confidence regions/intervals having coverage approximately equal to the nominal value of 0.95, we observe in the tables that the finite sample analysis can offer a slight improvement in terms of reduced expected volume or expected length of the confidence region/interval. However, as n and m increase, this improvement diminishes, and when n is sufficiently large, and m is not too small, the two methods appear to have similar performance.

5 Concluding Remarks

In this paper we have developed model based inference for multiply imputed synthetic data under the multiple linear regression model when synthetic data are generated via posterior predictive sampling, and when synthetic data are generated via plug-in sampling. The proposed methodology is derived from finite sample theory, and hence provides valid inference for both large and small values of the sample size n, and for any value of $m \geq 1$, where m is the number of released synthetic datasets.

Acknowledgments

We thank Dr. Tommy Wright, Dr. William Winkler, and Dr. Eric Slud of the U.S. Census Bureau for encouragement; and we thank the reviewers and Editor for providing helpful comments that enhanced the quality of the manuscript.

Appendices

A Proofs

Proof of (3). We work under the setup of Section 2.1. Let $\mathfrak{D}^* = (\boldsymbol{\beta}_1^*, \ldots, \boldsymbol{\beta}_m^*, \sigma_1^{2*}, \ldots, \sigma_m^{2*})$. We have $\mathrm{Var}(\overline{\boldsymbol{b}^*}) = E[\mathrm{Var}(\overline{\boldsymbol{b}^*}|\mathfrak{D}^*)] + \mathrm{Var}[E(\overline{\boldsymbol{b}^*}|\mathfrak{D}^*)] = E\left[(\boldsymbol{XX'})^{-1}\, m^{-2} \sum_{j=1}^{m} \sigma_j^{2*}\right] + \mathrm{Var}(\overline{\boldsymbol{\beta}^*}) = \sigma^2(\boldsymbol{XX'})^{-1}\frac{n-p}{m(n+\alpha-p-4)} + \mathrm{Var}(\overline{\boldsymbol{\beta}^*})$ where we have used the fact that $E(\sigma_j^{2*}) = E[E(\sigma_j^{2*}|\boldsymbol{y})] = E[\frac{\mathrm{RSS}}{n+\alpha-p-4}] = \frac{\sigma^2(n-p)}{n+\alpha-p-4}$. Noting that $\mathrm{Var}(\boldsymbol{\beta}_j^*|\boldsymbol{y}) = E[\mathrm{Var}(\boldsymbol{\beta}_j^*|\sigma_j^{2*}, \boldsymbol{y})|\boldsymbol{y}] + \mathrm{Var}[E(\boldsymbol{\beta}_j^*|\sigma_j^{2*}, \boldsymbol{y})|\boldsymbol{y}] = E[\sigma_j^{2*}(\boldsymbol{XX'})^{-1}|\boldsymbol{y}] + \mathrm{Var}[\boldsymbol{b}|\boldsymbol{y}] = (\boldsymbol{XX'})^{-1}\frac{\mathrm{RSS}}{n+\alpha-p-4}$, and $E(\boldsymbol{\beta}_j^*|\boldsymbol{y}) = E[E(\boldsymbol{\beta}_j^*|\sigma_j^{2*}, \boldsymbol{y})|\boldsymbol{y}] = E[\boldsymbol{b}|\boldsymbol{y}] = \boldsymbol{b}$, we obtain, $\mathrm{Var}(\overline{\boldsymbol{\beta}^*}) = E[\mathrm{Var}(\overline{\boldsymbol{\beta}^*}|\boldsymbol{y})] + \mathrm{Var}[E(\overline{\boldsymbol{\beta}^*}|\boldsymbol{y})] = E[(\boldsymbol{XX'})^{-1}\frac{\mathrm{RSS}}{m(n+\alpha-p-4)}] + \mathrm{Var}(\boldsymbol{b}) = \sigma^2(\boldsymbol{XX'})^{-1}\frac{n-p}{m(n+\alpha-p-4)} + \sigma^2(\boldsymbol{XX'})^{-1}$. Therefore, $\mathrm{Var}(\overline{\boldsymbol{b}^*}) = \sigma^2(\boldsymbol{XX'})^{-1}\frac{n-p}{m(n+\alpha-p-4)} + \mathrm{Var}(\overline{\boldsymbol{\beta}^*}) = 2\sigma^2(\boldsymbol{XX'})^{-1}\frac{n-p}{m(n+\alpha-p-4)} + \sigma^2(\boldsymbol{XX'})^{-1} = \sigma^2(\boldsymbol{XX'})^{-1}[1 + \frac{2(n-p)}{m(n+\alpha-p-4)}]$. □

Proof of (4). Working under the setup of Section 2.1, we have $E(\mathrm{RSS}_j^*) = E[E(\mathrm{RSS}_j^*|\boldsymbol{\beta}_j^*, \sigma_j^{2*})] = E[(n-p)\sigma_j^{2*}] = (n-p)E[E(\sigma_j^{2*}|\boldsymbol{y})] = (n-p)E[\frac{\mathrm{RSS}}{n+\alpha-p-4}] = \sigma^2\frac{(n-p)^2}{n+\alpha-p-4}$. Hence $E(\widetilde{\mathrm{RSS}^*}) = \sum_{j=1}^{m} E(\mathrm{RSS}_j^*) = \sigma^2\frac{m(n-p)^2}{n+\alpha-p-4}$. □

Proof of Theorem 1. We define $\Delta = \sum_{j=1}^{m}\frac{\sigma_j^{2*}}{m^2}$ and express T^2 as $T_1 \times T_2$ where $T_1 = \frac{(\overline{\boldsymbol{b}^*}-\boldsymbol{\beta})(\boldsymbol{XX'})(\overline{\boldsymbol{b}^*}-\boldsymbol{\beta})}{\sigma^2+2\Delta}$ and $T_2 = \frac{\sigma^2+2\Delta}{\widetilde{\mathrm{RSS}^*}}$. To derive the distribution of T^2, we show that T_1 and T_2 are independent, and establish their marginal distributions as stated in the theorem. Towards this end, we begin with some facts regarding the distributions of the random variables involved.

1. We have $\boldsymbol{b}_j^*|(\boldsymbol{\beta}_1^*, \sigma_1^{2*}, \ldots, \boldsymbol{\beta}_m^*, \sigma_m^{2*}) \sim N_p[\boldsymbol{\beta}_j^*, (\boldsymbol{XX'})^{-1}\sigma_j^{2*}]$, independently for $j = 1, \ldots, m$, and therefore, $\overline{\boldsymbol{b}^*}|(\boldsymbol{\beta}_1^*, \sigma_1^{2*}, \ldots, \boldsymbol{\beta}_m^*, \sigma_m^{2*}) \sim N_p\left[\overline{\boldsymbol{\beta}^*}, (\boldsymbol{XX'})^{-1}\Delta\right]$.

2. We have $\boldsymbol{\beta}_j^* | (\sigma_j^{2*}, \boldsymbol{b}, \mathrm{RSS}) \sim N_p[\boldsymbol{b}, \sigma_j^{2*}(\boldsymbol{XX'})^{-1}]$, independently for $j = 1, \ldots, m$, and therefore, $\overline{\boldsymbol{\beta}^*} | (\sigma_1^{2*}, \ldots, \sigma_m^{2*}, \boldsymbol{b}, \mathrm{RSS}) \sim N_p\left[\boldsymbol{b}, (\boldsymbol{XX'})^{-1}\Delta\right]$.

3. It follows by multiplication and use of the identity

$$\begin{aligned}
&(\overline{\boldsymbol{\beta}^*} - \boldsymbol{b})'(\boldsymbol{XX'})(\overline{\boldsymbol{\beta}^*} - \boldsymbol{b}) + (\overline{\boldsymbol{b}^*} - \overline{\boldsymbol{\beta}^*})'(\boldsymbol{XX'})(\overline{\boldsymbol{b}^*} - \overline{\boldsymbol{\beta}^*}) \\
&= 2\left(\overline{\boldsymbol{\beta}^*} - \frac{\boldsymbol{b} + \overline{\boldsymbol{b}^*}}{2}\right)'(\boldsymbol{XX'})\left(\overline{\boldsymbol{\beta}^*} - \frac{\boldsymbol{b} + \overline{\boldsymbol{b}^*}}{2}\right) + \frac{1}{2}(\overline{\boldsymbol{b}^*} - \boldsymbol{b})'(\boldsymbol{XX'})(\overline{\boldsymbol{b}^*} - \boldsymbol{b}),
\end{aligned}$$

that the conditional joint pdf of $\overline{\boldsymbol{b}^*}$ and $\overline{\boldsymbol{\beta}^*}$, given $\Delta, \boldsymbol{b}, \mathrm{RSS}$, is

$$\begin{aligned}
&f(\overline{\boldsymbol{b}^*}, \overline{\boldsymbol{\beta}^*} | \Delta, \boldsymbol{b}) \\
&\propto \exp\left\{-\frac{1}{2}\left[\frac{2}{\Delta}\left(\overline{\boldsymbol{\beta}^*} - \frac{\boldsymbol{b} + \overline{\boldsymbol{b}^*}}{2}\right)'(\boldsymbol{XX'})\left(\overline{\boldsymbol{\beta}^*} - \frac{\boldsymbol{b} + \overline{\boldsymbol{b}^*}}{2}\right) + \frac{1}{2\Delta}(\overline{\boldsymbol{b}^*} - \boldsymbol{b})'(\boldsymbol{XX'})(\overline{\boldsymbol{b}^*} - \boldsymbol{b})\right]\right\}.
\end{aligned}$$

4. Integrating out $\overline{\boldsymbol{\beta}^*}$, we get

$$f(\overline{\boldsymbol{b}^*} | \Delta, \boldsymbol{b}, \mathrm{RSS}) \propto \exp\left\{-\frac{1}{2}\left[\frac{1}{2\Delta}(\overline{\boldsymbol{b}^*} - \boldsymbol{b})'(\boldsymbol{XX'})(\overline{\boldsymbol{b}^*} - \boldsymbol{b})\right]\right\}.$$

5. Using $\boldsymbol{b} \sim N_p(\boldsymbol{\beta}, \sigma^2(\boldsymbol{XX'})^{-1})$, we multiply $f(\overline{\boldsymbol{b}^*} | \Delta, \boldsymbol{b}, \mathrm{RSS})$ with $f(\boldsymbol{b})$ and use the identity below to integrate out $\boldsymbol{b}$:

$$\begin{aligned}
&\frac{(\overline{\boldsymbol{b}^*} - \boldsymbol{b})'(\boldsymbol{XX'})(\overline{\boldsymbol{b}^*} - \boldsymbol{b})}{2\Delta} + \frac{(\boldsymbol{b} - \boldsymbol{\beta})'(\boldsymbol{XX'})(\boldsymbol{b} - \boldsymbol{\beta})}{\sigma^2} \\
&= \left(\frac{1}{\sigma^2} + \frac{1}{2\Delta}\right)\left[\boldsymbol{b} - \frac{\left(\frac{\boldsymbol{\beta}}{\sigma^2} + \frac{\overline{\boldsymbol{b}^*}}{2\Delta}\right)}{\left(\frac{1}{\sigma^2} + \frac{1}{2\Delta}\right)}\right]'(\boldsymbol{XX'})\left[\boldsymbol{b} - \frac{\left(\frac{\boldsymbol{\beta}}{\sigma^2} + \frac{\overline{\boldsymbol{b}^*}}{2\Delta}\right)}{\left(\frac{1}{\sigma^2} + \frac{1}{2\Delta}\right)}\right] + \frac{(\overline{\boldsymbol{b}^*} - \boldsymbol{\beta})'(\boldsymbol{XX'})(\overline{\boldsymbol{b}^*} - \boldsymbol{\beta})}{\sigma^2 + 2\Delta}.
\end{aligned}$$

6. This readily yields the required conditional density of $\overline{\boldsymbol{b}^*}$, given Δ as

$$f(\overline{\boldsymbol{b}^*} | \Delta) \propto \exp\left\{-\frac{1}{2}\frac{(\overline{\boldsymbol{b}^*} - \boldsymbol{\beta})'(\boldsymbol{XX'})(\overline{\boldsymbol{b}^*} - \boldsymbol{\beta})}{\sigma^2 + 2\Delta}\right\}.$$

7. It then follows that
$$\frac{(\overline{\boldsymbol{b}^*}-\boldsymbol{\beta})'(\boldsymbol{X}\boldsymbol{X}')(\overline{\boldsymbol{b}^*}-\boldsymbol{\beta})}{\sigma^2+2\Delta}\Big|\Delta \sim \chi^2_p,$$
and hence unconditionally as well. Obviously, T_1 is independent of Δ, and hence that of T_2 because $\boldsymbol{b}_j^*$'s are independent of RSS_j^*'s, conditionally given $\boldsymbol{b}$ and RSS.

To determine the distribution of $T_2 = \frac{\sigma^2+2\Delta}{\sum_{j=1}^m \text{RSS}_j^*}$, note that this expression involves the random variables $\{(\text{RSS}_j^*, \sigma_j^{2*}), j=1,\ldots,m\}$, and their distributions in turn depend on RSS. We proceed in a very logical fashion by taking up the above random variables one group at a time in the natural order they appear, namely, first $\{\text{RSS}_j^*, j=1,\ldots,m\}$, then $\{\sigma_j^{2*}, j=1,\ldots,m\}$, and at the end RSS. We again observe the following facts.

1. For $j=1,\ldots,m$,
$$\frac{\text{RSS}_j^*}{\sigma_j^{2*}}\Big|\sigma_1^{2*},\ldots,\sigma_m^{2*},\text{RSS} \sim \chi^2_{n-p},$$
and all the χ^2 variables are independent, and obviously they are also independent of σ_j^{2*}'s. Let us denote them by C_j, $j=1,\ldots,m$. This fact is used in the denominator of T_2.

2. We next note that
$$\frac{\text{RSS}}{\sigma_j^{2*}}\Big|\text{RSS} \sim \chi^2_{n+\alpha-p-2},\ j=1,\ldots,m,$$
and, conditionally given RSS, these χ^2 variables are independent, and obviously they are also independent of RSS. We denote them by B_j, $j=1,\ldots,m$.

3. Using the above two steps, T_2 can be expressed as
$$\frac{\sigma^2+2\Delta}{\sum_{j=1}^m \text{RSS}_j^*}\Big|\text{RSS} \overset{d}{=} \frac{\sigma^2+2m^{-2}\text{RSS}(\sum_{j=1}^m \frac{1}{B_j})}{\text{RSS}\sum_{j=1}^m \frac{C_j}{B_j}}.$$

4. Finally, using the fact that $\text{RSS}/\sigma^2 \sim \chi^2_{n-p}$, and noting this last χ^2, denoted as A, is independent of all the previous χ^2 variables, we get
$$T_2 \overset{d}{=} \frac{1+2m^{-2}A(\sum_{j=1}^m \frac{1}{B_j})}{A\sum_{j=1}^m \frac{C_j}{B_j}}.$$

This completes the proof. □

Proof of Theorem 2. The proof is based on the following steps.

1. Given $(\boldsymbol{b}, \mathrm{RSS})$, the conditional joint pdf of $(\overline{\boldsymbol{b}^*}, S^2_{\text{comb}})$ is given by

$$f(\boldsymbol{b}^*, \mathrm{RSS}^*|\boldsymbol{b}, \mathrm{RSS}) \propto e^{-\frac{n-p}{2}\left[m\frac{(\overline{\boldsymbol{b}^*}-\boldsymbol{b})'(\boldsymbol{X}\boldsymbol{X}')(\boldsymbol{b}^*-\boldsymbol{b})+S^2_{\text{comb}}}{\mathrm{RSS}}\right]} \times \frac{(S^2_{\text{comb}})^{\frac{nm-p}{2}-1}}{\mathrm{RSS}^{nm/2}}.$$

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

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

Combining the above, we get the joint pdf of $(\overline{\boldsymbol{b}^*}, S^2_{\text{comb}}, \boldsymbol{b}, \mathrm{RSS})$ which we use to sequentially integrate out $\boldsymbol{b}$ and RSS. Writing $\widetilde{\mathrm{RSS}} = \mathrm{RSS}/(n-p)$, since

$$\begin{aligned}
&\frac{m(\overline{\boldsymbol{b}^*}-\boldsymbol{b})'(\boldsymbol{X}\boldsymbol{X}')(\overline{\boldsymbol{b}^*}-\boldsymbol{b})}{\widetilde{\mathrm{RSS}}} + \frac{(\boldsymbol{b}-\boldsymbol{\beta})'(\boldsymbol{X}\boldsymbol{X}')(\boldsymbol{b}-\boldsymbol{\beta})}{\sigma^2} \\
&= \left(\frac{1}{\sigma^2} + \frac{m}{\widetilde{\mathrm{RSS}}}\right)\left[\boldsymbol{b} - \frac{(\frac{\boldsymbol{\beta}}{\sigma^2} + \frac{m\overline{\boldsymbol{b}^*}}{\widetilde{\mathrm{RSS}}})}{(\frac{1}{\sigma^2} + \frac{m}{\widetilde{\mathrm{RSS}}})}\right]'(\boldsymbol{X}\boldsymbol{X}')\left[\boldsymbol{b} - \frac{(\frac{\boldsymbol{\beta}}{\sigma^2} + \frac{m\overline{\boldsymbol{b}^*}}{\widetilde{\mathrm{RSS}}})}{(\frac{1}{\sigma^2} + \frac{m}{\widetilde{\mathrm{RSS}}})}\right] + \frac{(\overline{\boldsymbol{b}^*}-\boldsymbol{\beta})'(\boldsymbol{X}\boldsymbol{X}')(\overline{\boldsymbol{b}^*}-\boldsymbol{\beta})}{(\sigma^2 + (1/m)\widetilde{\mathrm{RSS}})},
\end{aligned}$$

integrating out $\boldsymbol{b}$, we get the joint pdf of $(\overline{\boldsymbol{b}^*}, S^2_{\text{comb}}, \mathrm{RSS})$ as

$$\begin{aligned}
&f_{\boldsymbol{\beta},\sigma^2}(\overline{\boldsymbol{b}^*}, S^2_{\text{comb}}, \mathrm{RSS}) \\
&\propto e^{-\frac{1}{2}\left[\frac{(\boldsymbol{b}^*-\boldsymbol{\beta})'(\boldsymbol{X}\boldsymbol{X}')(\boldsymbol{b}^*-\boldsymbol{\beta})}{\sigma^2+(1/m)\widetilde{\mathrm{RSS}}}+\frac{S^2_{\text{comb}}}{\widetilde{\mathrm{RSS}}}+\frac{\mathrm{RSS}}{\sigma^2}\right]} \times \frac{(S^2_{\text{comb}})^{\frac{nm-p}{2}-1}}{(\mathrm{RSS})^{nm/2}} \times \frac{(\mathrm{RSS})^{-\frac{p+2}{2}}}{\sigma^n} \times \left[\frac{1}{\sigma^2} + \frac{m}{\widetilde{\mathrm{RSS}}}\right]^{-p/2}.
\end{aligned}$$

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

$$\begin{aligned}
&f_{\boldsymbol{\beta},\sigma^2}(\overline{\boldsymbol{b}^*}, S^2_{\text{comb}}, \psi) \\
&\propto e^{-\frac{1}{2}\left[\frac{(\overline{\boldsymbol{b}^*}-\boldsymbol{\beta})'(\boldsymbol{X}\boldsymbol{X}')(\overline{\boldsymbol{b}^*}-\boldsymbol{\beta})}{\sigma^2(1+\frac{\psi}{m(n-p)})}+\frac{(n-p)S^2_{\text{comb}}}{\sigma^2\psi}+\psi\right]} \times (S^2_{\text{comb}})^{\frac{nm-p}{2}-1} \times (\psi)^{-\frac{p+2}{2}} \times \left[1 + \frac{m(n-p)}{\psi}\right]^{-p/2}.
\end{aligned}$$

To complete the proof, note that, conditionally given ψ,

$$\frac{(\overline{\boldsymbol{b}^*} - \boldsymbol{\beta})'(\boldsymbol{X}\boldsymbol{X}')(\overline{\boldsymbol{b}^*} - \boldsymbol{\beta})}{\sigma^2(1 + \frac{\psi}{m(n-p)})} \sim \chi^2_p, \quad \frac{(n-p)S^2_{\text{comb}}}{\sigma^2 \psi} \sim \chi^2_{nm-p}, \text{ independent of } \overline{\boldsymbol{b}^*},$$

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,

$$T^2_{\text{comb}} \left[\frac{\psi}{m(n-p) + \psi} \right] \sim \frac{p}{m(mn-p)} F_{p,mn-p}.$$

This completes the proof. □

B Details of the Simulation Model

To perform the simulation studies of Section 4, we take

$$\boldsymbol{X} = \begin{pmatrix} 1 & 1 & \dots & 1 \\ x_{11} & x_{12} & \dots & x_{1n} \\ x_{21} & x_{22} & \dots & x_{2n} \\ x_{31} & x_{32} & \dots & x_{3n} \\ x_{41} & x_{42} & \dots & x_{4n} \\ I(x_{51}=2) & I(x_{52}=2) & \dots & I(x_{5n}=2) \\ I(x_{51}=3) & I(x_{52}=3) & \dots & I(x_{5n}=3) \\ I(x_{51}=4) & I(x_{52}=4) & \dots & I(x_{5n}=4) \\ I(x_{51}=5) & I(x_{52}=5) & \dots & I(x_{5n}=5) \\ I(x_{51}=6) & I(x_{52}=6) & \dots & I(x_{5n}=6) \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \\ \beta_6 \\ \beta_7 \\ \beta_8 \\ \beta_9 \\ \beta_{10} \end{pmatrix} = \begin{pmatrix} 10 \\ 2 \\ 2 \\ -3 \\ -1 \\ -2 \\ 1 \\ 2 \\ 2 \\ 4 \end{pmatrix}, \quad \sigma^2 = 1,$$

where $I(\cdot)$ is the indicator function and the variables appearing in $\boldsymbol{X}$ are generated independently for $i = 1, \dots, n$ according to

$$x_{1i} \sim N(1,1),\ \log x_{2i} \sim N(0,1),\ x_{3i} \sim \text{Exponential}(\text{mean} = 1),\ x_{4i} \sim \text{Poisson}(1),$$
$$P(x_{5i} = 1) = P(x_{5i} = 3) = P(x_{5i} = 4) = P(x_{5i} = 5) = 0.2, P(x_{5i} = 2) = P(x_{5i} = 6) = 0.1.$$

References

[1] Abowd JM, Stinson M, Benedetto G. Final Report to the Social Security Administration on the SIPP/SSA/IRS Public Use File Project. Technical Report. 2006; URL `http://www2.vrdc.cornell.edu/news/wp-content/uploads/2007/11/ssafinal.pdf`.

[2] Benedetto G, Stinson MH, Abowd JM. The Creation and Use of the SIPP Synthetic Beta. Technical Report. 2013; URL `http://www.census.gov/content/dam/Census/programs-surveys/sipp/methodology/SSBdescribe_nontechnical.pdf`.

[3] Drechsler J. Synthetic Datasets for Statistical Disclosure Control. New York: Springer; 2011.

[4] Hawala S. Producing Partially Synthetic Data to Avoid Disclosure. Proceedings of the Joint Statistical Meetings. American Statistical Association. 2008; 1345-1350.

[5] Kinney SK, Reiter JP, Miranda J. SynLBD 2.0: Improving the Synthetic Longitudinal Business Database. Statistical Journal of the International Association for Official Statistics. 2014; 30(2): 129-135.

[6] Kinney SK, Reiter JP, Reznek AP, Miranda J, Jarmin RS, Abowd JM. Towards Unrestricted Public Use Business Microdata: The Synthetic Longitudinal Business Database, International Statistical Review. 2011; 79(3): 362-384.

[7] Klein M, Sinha B. Inference for Singly Imputed Synthetic Data Based on Posterior Predictive Sampling under Multivariate Normal and Multiple Linear Regression Models. Sankhyā: The Indian Journal of Statistics. 2015; 77-B(2): 293-311.

[8] Klein M, Sinha B. Likelihood Based Finite Sample Inference for Singly Imputed Synthetic Data Under the Multivariate Normal and Multiple Linear Regression Models. Journal of Privacy and Confidentiality. 2016; 7(1): 43-98.

[9] Little RJA. Statistical Analysis of Masked Data. Journal of Official Statistics. 1993; 9(2): 407-426.

[10] Machanavajjhala A, Kifer D, Abowd J, Gehrke J, Vilhuber L. Privacy: Theory Meets Practice on the Map. IEEE 24th International Conference on Data Engineering. 2008; 277-286.

[11] R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing: Vienna, Austria; 2013; URL `http://www.R-project.org/`.

[12] Raghunathan TE, Reiter JP, Rubin DB. Multiple Imputation for Statistical Disclosure Limitation. Journal of Official Statistics. 2003; 19(1): 1-16.

[13] Reiter JP. Inference for Partially Synthetic, Public Use Microdata Sets. Survey Methodology. 2003; 29(2): 181-188.

[14] Reiter JP. Significance Tests for Multi-Component Estimands From Multiply Imputed, Synthetic Microdata. Journal of Statistical Planning and Inference. 2005; 131(2): 365-377.

[15] Reiter JP, Kinney SK. Inferentially Valid, Partially Synthetic Data: Generating from Posterior Predictive Distributions not Necessary. Journal of Official Statistics. 2012; 28(4): 583-590.

[16] Rencher AC, Schaalje GB. Linear Models in Statistics, Second Edition, New Jersey: John Wiley & Sons; 2008.

[17] Rodríguez R. Synthetic Data Disclosure Control for American Community Survey Group Quarters. Proceedings of the Joint Statistical Meetings. American Statistical Association. 2007; 1439-1450.

[18] Rubin DB. Multiple Imputation for Nonresponse in Surveys. New Jersey: John Wiley & Sons; 1987.

[19] Rubin DB. Discussion: Statistical Disclosure Limitation. Journal of Official Statistics. 1993; 9(2): 461-468.