Jackknife Variance Estimator for Datasets Containing Multiply Imputed Outcome Variables Under Uncongeniality: A Monte Carlo Simulation Study

Table of contents

Title Page {.unlisted ,.unnumbered}

A Thesis

Presented to the Department of Mathematics and Statistics

Hal Marcus College of Science and Engineering

and

The Kugelman Honors Program

of

The University of West Florida

In partial fulfillment of the requirements for graduation as a Kugelman Honors Scholar

Ihsan E. Buker

November, 2022

Acknowledgments {.unlisted ,.unnumbered}

I would like to thank my cat.

1 Abstract

Missing data is an issue ubiquitous in many fields of science. Today, multiple imputation (MI) is one of the most commonly utilized approaches to provide valid statistical inferences in the presence of missing data. Briefly, MI fills the missing cells in the original dataset by generating a series of plausible values based on an imputation model and, thereafter, creates multiple complete versions of the original dataset. Subsequently, the analysis model is applied to each imputed dataset, and the parameters of interest are pooled to accurately reflect the loss of information caused by the missing observations. Accompanying MI, however, is the issue of uncongeniality, which, imprecisely, occurs when the imputation model and the analysis model make different assumptions about the data. Not long after the conception of MI, Rubin's accompanying set of rules to pool parameter estimates from the multiply imputed datasets was shown to produce biased point estimates under uncongeniality, which led to under-coverage of confidence intervals for anti-conservative estimates of variance or over-coverage for conservative estimates. In response, certain combinations of MI and resampling methods have been proposed as robust variance estimators under uncongeniality; however, their main drawback, to this day, has been their associated computational cost. Moreover, bootstrapping, one of the most commonly utilized resampling methods alongside MI to obtain proper variance estimates, has its basis in asymptotic theory. As such, in small samples frequently encountered in biological studies, the need for a computationally efficient variance estimator with statistically desirable properties remains.

In response, a jackknife variance estimator for multiply imputed outcome variables under uncongeniality for small sample sizes is proposed, which provides asymptotically unbiased point estimates with appropriate confidence interval coverage under uncongeniality. The performance of the proposed jackknife variance estimator is investigated using a Monte Carlo simulation study and compared to other methods in the literature. Accordingly, the recommendation to replace Rubin's rules as the de facto standard in variance estimation with resampling-based robust variance estimators is made, particularly in light of the modern computational power statistical practitioners have at their disposal. Finally, an implementation of the proposed jackknife variance estimator in R is provided.

1.1 Keywords

Multiple imputation, uncongeniality, model misspecification, jackknife resampling

2 Background

Missing data is a discipline-agnostic issue commonly encountered by statistical practitioners. Given that many statistical procedures require data to be complete (i.e., in the form of an $n \times m$ matrix), the appropriate course of action to be taken in the presence of missing data has long been investigated by statisticians. Today, multiple imputation is accepted as the gold standard in missing data analysis, thanks to the work of Donald Rubin (Buuren 2012). In 1977, Rubin proposed using multiple completed versions of the dataset with missing observations, applying the complete-data procedure of interest, and pooling the estimates to draw valid inferences (Buuren 2012). The main advantage of multiple imputation, as opposed to single imputation, which had been used by researchers since the early 1970s, is its ability to properly estimate the variance caused by missing observations (Buuren 2012). The emphasis placed on variance and uncertainty by Rubin was a departure from the status quo of the time, which was to fill in the missing observation with the most probable value and to proceed with complete case analyses as if the observation had not been missing, to begin with (Buuren 2012). This approach, however, fails to incorporate the loss of information caused by missing observations into the estimation of parameters, resulting in the underestimation of variance (Rubin 1978).

Like all revolutionary ideas, multiple imputation received harsh criticism following its conception. Perhaps the most notable of the objections came from Fay in 1992, who demonstrated through counterexamples that multiple imputation produced biased covariance estimates (Jonathan W. Bartlett and Hughes 2020). Fay added that the need for unison between the imputation and analysis model made multiple imputation a poor general-purpose tool, particularly in instances where the imputer and analyst are different individuals (Fay 1992; Buuren 2012). Fay's arguments led to the conceptualization of congeniality[1] between the imputation and analysis model, which was later accepted to be a requirement to obtain valid inferences from multiple imputation using Rubin's pooling rules (hereafter, Rubin's rules) (Buuren 2012; Meng 1994). Although Fay's work initially criticized biases introduced to the covariance matrix following multiple imputation, a similar phenomenon of biased estimations were observed with variance under uncongeniality (Fay 1992; Meng 1994; Xie and Meng 2016).

Some of the earliest works demonstrating Rubin's variance estimator to be biased under uncongeniality were from Wang and Robins in 1998, who also proposed an alternate variance estimator in the same paper (Buuren 2012). The variance estimator proposed by Wang and Robins requires the calculation of several quantities, which are not easily accessible to the average statistical practitioner (Jonathan W. Bartlett and Hughes 2020). The challenging

[1] Please see the appendix for a detailed overview of congeniality.

construction of the variance estimator proposed by Wang and Robins has resulted in it receiving little-to-no attention in applied settings (Jonathan W. Bartlett and Hughes 2020). In an attempt to create a more user-friendly variance estimator in instances of suspected uncongeniality, researchers have proposed combining resampling methods with multiple imputation. Of the two main resampling methods, bootstrap has received more attention from multiple imputation researchers compared to jackknife resampling, which has mostly been investigated under single hot-deck imputation. Although particular combinations of bootstrap and multiple imputation have been demonstrated to create asymptotically unbiased estimates of variance, the associated computational cost makes this an active area of research (Jonathan W. Bartlett and Hughes 2020). Most recently, von Hippel has proposed a bootstrap variance estimator which addresses the issue of computational cost; however, it has been demonstrated to create confidence intervals that are slightly wider compared to traditional bootstrap and multiple imputation combinations (Jonathan W. Bartlett and Hughes 2020). Given the lower computational cost associated with jackknife resampling, as well as desirable properties demonstrated under single imputation, such as being unbiased in certain scenarios, it is an attractive alternative that should be considered as a variance estimator of multiply imputed data under uncongeniality (Chen and Shao 2001; Rao and Shao 1992). More importantly, however, is the advantages jackknife resampling has over bootstrap resampling in small sample sizes, which are frequently encountered in datasets associated with biological studies.

2.1 Bootstrap Resampling

Let q be the set of observations $(z_1, z_2, z_3, \dots, z_n)$ from the population Q such that $z_i \ \forall \ i \in \{1, 2, 3, \dots, n\}$ is an i.i.d sample from Q. Moreover, let θ be some parameter of interest, with the unbiased estimator $\hat{\theta}$, which is a statistic computed from $F(q)$. Finally, let G_θ be the sampling distribution of $F(q)$. The non-parametric bootstrap, as proposed by Efron, lets q define Q such that the set of observations $(z_1, z_2, z_3, \dots, z_n)$ appears with equal proportion in the infinitely large population. From there, the set of samples $q_1^*, q_2^*, q_3^*, \dots, q_m^*$ as $m \to \infty$ are generated by sampling n observations with replacement from q, and the statistic of interest $\hat{\theta}$ is calculated by applying $F(q_i) \ \forall \ i \in \{1, 2, 3, \dots, m\}$, which results in $\hat{\theta^*}_1, \hat{\theta^*}_2, \hat{\theta^*}_3, \dots, \hat{\theta^*}_m$.

Finally, the point estimate is obtained by

$$\hat{\theta} = m^{-1} \sum_{i=1}^{m} \hat{\theta^*}_i$$

and the variance is obtained by

$$\text{var}(\hat{\theta}) = m^{-1} \sum_{i=1}^{m} \left(\hat{\theta^*}_i - \hat{\theta} \right)^2$$

Since Efron's proposal of the non-parametric bootstrap, statisticians have widely utilized it thanks to its ease of implementation and the rapidly increasing computational power available to statistical practitioners (LaFontaine 2021). However, the properties of bootstrap resampling have their basis in the asymptotic theory, which holds in large sample sizes (Wang 1998; Mammen 1992). The minimum sample size required to utilize bootstrap resampling and obtain asymptotically unbiased estimates is highly context-dependent; in certain situations, a minimum sample size of $n = 200$ has been suggested, with other authors suggesting between $n = 100$ to $n = 500$ (Anderson and Gerbing 1984; Bentler and Chou 1987; Jackson 2001). Given that in many biological studies, the minimum sample size required for asymptotically unbiased estimates may not be achieved, jackknife resampling, a method that predates the bootstrap, may be considered (Faber and Fonseca 2014).

2.2 Jackknife Resampling

2.2.1 Leave-One-Out Jackknife

Let q be the set of observations $(z_1, z_2, z_3, \dots, z_n)$ from the population Q such that $z_i \ \forall \ i \in \{1, 2, 3, \dots, n\}$ is an i.i.d sample from Q. Moreover, let θ be some parameter of interest, with the unbiased estimator $\widehat{\theta}$, which is a statistic computed from $F(q)$. Finally, let G_θ be the sampling distribution of $F(q)$. The jackknife, as proposed by Quenouille and expanded on by Tukey, creates n leave-one-out subsamples from q such that $q_{-1} = (z_2, z_3, z_4, \dots, z_n)$, $q_{-2} = (z_1, z_3, z_4, \dots, z_n), \dots, q_{-n} = (z_1, z_2, z_3, \dots, z_{n-1})$ and $|q_{-i}| = n - 1 \ \forall i \in \{1, 2, 3, \dots, n\}$. Thereafter, the statistic of interest $\widehat{\theta}$ is calculated by applying $F(q_i) \ \forall \ i \in \{1, 2, 3, \dots, n\}$, which results in $\widehat{\theta^*}_1, \widehat{\theta^*}_2, \widehat{\theta^*}_3, \dots, \widehat{\theta^*}_m$.

Finally, the point estimate is obtained by

$$\widehat{\theta} = n^{-1} \sum_{i=1}^{n} \widehat{\theta^*}_i$$

and the variance is obtained by

$$\text{var}(\widehat{\theta}) = n^{-1} \sum_{i=1}^{n} \left(\widehat{\theta^*}_i - \widehat{\theta} \right)^2$$

2.2.2 Delete-*d* Jackknife

The delete-d jackknife may be seen as a generalized version of the traditional jackknife (hereinafter, jackknife). Although in many situations, the jackknife provides a computationally efficient means to estimate the variance of an estimator, for non-smooth statistics, such as the

percentiles of the data, the jackknife fails, as the statistic varies significantly between any two subsamples (Chen and Shao 2001). In such cases, the delete-d jackknife provides an alternative estimator that can provide asymptotically unbiased estimates for non-smooth statistics (Rao and Shao 1992). Similarly, let q be the set of observations $(z_1, z_2, z_3, \dots, z_n)$ from the population Q such that $z_i \ \forall \ i \in \{1, 2, 3, \dots, n\}$ is an i.i.d sample from Q. Moreover, let θ be some parameter of interest, with the unbiased estimator $\widehat{\theta}$, which is a statistic computed from $F(q)$. Finally, let G_θ be the sampling distribution of $F(q)$. The delete-d jackknife, creates $\binom{n}{d}$ subsamples of q such that $|q_i| = n - d$ Thereafter, the statistic of interest $\hat{\theta}$ is calculated by applying $F(q_i) \ \forall \ i \in \{1, 2, 3, \dots, \binom{n}{d}\}$, which results in $\widehat{\theta^*}_1, \widehat{\theta^*}_2, \widehat{\theta^*}_3, \dots, \widehat{\theta^*}_{\binom{n}{d}}$. In many cases, however, $\binom{n}{d}$ is a large value that yields this approach computationally unfeasible. In such instances, certain guidelines, as discussed in following chapters, may be employed to determine a value of subsamples that still provide proper estimates.

Finally, the point estimate is obtained by

$$\hat{\theta} = \binom{n}{d}^{-1} \sum_{i=1}^{\binom{n}{d}} \hat{\theta^*}_i$$

and the variance is obtained by

$$\text{var}(\hat{\theta}) = \binom{n}{d}^{-1} \sum_{i=1}^{\binom{n}{d}} \left(\hat{\theta^*}_i - \hat{\theta}\right)^2$$

2.3 Comparison of Jackknife and Bootstrap Resampling

Given their comparable construction and application, the jackknife and bootstrap make similar assumptions regarding the data. Most notably, both assume that the function used to estimate the parameter is smooth. Formally, a smooth function is defined as one that has continuous derivatives over some domain, with the minimum number of derivatives required to be considered smooth varying per the question at hand (Weisstein, n.d.). From a statistical perspective, if the function used to estimate the parameter is smooth on some domain (a, b), $[F(q_i) - F(q_z)] \to 0$ as $|q_i - q_z| \to 0$. Meaning, among a set of conceivable, non-identical samples from the population, minor differences between possible samples will only result in minor differences between the statistic estimated (Chen and Shao 2001). Due to its deterministic nature, the jackknife tends to perform poorly in estimating non-smooth statistics (Wicklin 2017). However, its deterministic nature also yields the jackknife superior to bootstrapping in smaller datasets. To combat the issue regarding non-smooth statistics, a generalized jackknife resampling scheme, called the *drop-d-jackknife* has been proposed, which was utilized in this estimator (Chen and Shao 2001).

2.4 Multiple Imputation

Multiple imputation is a missing data management method based on both Bayesian and frequentist inference proposed by Donald Rubin in the late 1970s (Jonathan W. Bartlett and Hughes 2020; Buuren 2012). Before Rubin's proposal, statisticians had been utilizing various single-imputation methods and proceeding with the analyses of interest as if data had not been missing. However, Rubin noted that single-imputation methods could not accurately capture the uncertainty caused by the missing observations (Buuren 2012). In response, he proposed imputing any given datum with a series of plausible values from its posterior distribution, thus creating several complete versions of the observed dataset and applying the complete data analysis procedure to each generated dataset. He proposed a series of rules that could derive point and variance estimates that would adequately reflect the uncertainty caused by the missing observations.

Rubin's idea to utilize multiple imputation was ridiculed at first, not only due to its drastically different interpretation of uncertainty but because of how unfeasible it was at the time (Buuren 2012). Rubin's method would require statisticians to come up with an imputation model that would allow them to draw values from the posterior distribution. After that, they would have to draw several values and repeat the analysis multiple times with the numerous complete datasets. The preceding workflow was challenging to implement in an era of low computational power and expensive digital storage (Buuren 2012). As such, Rubin's ideas did not receive immediate acceptance. However, since the late 1990s, with the increased access to computers and user-friendly statistical packages capable of implementing complex procedures, multiple imputations have been adopted and heavily researched, leading to various modified algorithms being utilized under different conditions to obtain valid inferences in the presence of missing data (Buuren 2012). At this time, one of the most notable challenges remaining with multiple imputation is the concept of congeniality. Congeniality may be thought of as the imputation model and the analysis model making compatible assumptions regarding the data. In the early 1990s, Fay and Meng demonstrated that congeniality was required to obtain valid inferences from multiple imputation (Jonathan W. Bartlett and Hughes 2020; Jonathan W. Bartlett 2021).

2.4.1 Current State of Multiple Imputation

Today, with the advent of public databases, the imputer and the analyst may no longer be the same person. Even in the absence of such cases, the imputation and analysis model may still be uncongenial if there does not exist a unifying Bayesian model which embeds the imputer's imputation model and the analyst's complete data procedure (Jonathan W. Bartlett 2021). As such, researchers have begun to develop approaches that combine resampling methods with multiple imputation to obtain valid inferences even under uncongeniality. Of the currently proposed methods, there exist two main limitations: a) the increased computational cost brought on by resampling methods alongside multiple imputation, and b) inference in smaller

sample sizes. At this time, nearly all mainstream approaches proposed by researchers utilize bootstrap resampling and multiple imputation to obtain valid inferences under uncongeniality; however, as discussed above, bootstrap resampling requires a certain sample size to provide proper estimates. As a viable alternative to be utilized in instances where there is a small sample size, a jackknife variance estimator is proposed.

3 Methods

For the proposed Monte Carlo simulation, $N = 10{,}000$ datasets will be generated with the following characteristics: A response variable, Y, with $p_{miss} = 0.3$ proportion of missing observations, where the mechanism of missingness is missing at random (MAR), and an $n \times q$ matrix of fully observed covariates, where $n = 50$ is the sample size, and $q = 3$ is the number of covariates.

Formally

$$\begin{bmatrix} V1 \\ V2 \\ V3 \end{bmatrix} \sim N \left(\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 & 0.5 & 0.5 \\ 0.5 & 1 & 0.5 \\ 0.5 & 0.5 & 1 \end{bmatrix} \right)$$

With

$$\beta_{V_1} = 2;\ \beta_{V_2} = 5;\ \beta_{V_3} = 8$$

The outcome variable was simulated such that:

$$Y = V_1 + V_2 + V_3 + \epsilon$$

Where

$$\epsilon \sim N(\mu = 0, \sigma \propto V_2)$$

Thus, data were simulated with heteroskedastic errors, which yields the imputation and analysis models congenial, yet misspecified.

Formally, the analysis model of interest was

$$\hat{Y} \sim \hat{\beta_{V_1}} + \hat{\beta_{V_2}} + \hat{\beta_{V_3}}$$

And the imputation model was

$$\hat{Y_{\text{mis}}} \sim \hat{\beta_{V_1}} + \hat{\beta_{V_2}} + \hat{\beta_{V_3}}$$

Where the imputation method of choice was predictive mean matching (PMM).

All generated datasets will be analysed using three approaches: * Bootstrap then multiply impute: The observed dataset with missing observations will initially be bootstrapped $n = 200$ times. Thereafter, each of the bootstrap samples will be imputed $m = 2$ times, with a maximum of $maxit = 5$ iterations. The mean of the bootstrap estimates will serve as the final point estimate, and a 95% confidence interval will be generated through the percentile method, where the $\alpha/2^{th}$ and $1-(\alpha/2)^{th}$ percentiles will be the lower and upper bounds, respectively. The R package, *bootImpute* will be utilized for this process.

- Multiply impute then use Rubin's rules: The observed dataset with missing observations will be imputed $m = 10$ times, with a maximum of $maxit = 5$ iterations. The point estimate, as well as the confidence interval will be obtained through the following rules proposed by Donald Rubin. The *mice* package will be utilized for this process.

$$\bar{\theta} = \frac{1}{m}\left(\sum_{i=1}^{m}\theta_i\right) \tag{3.1}$$

$$V_{within} = \frac{1}{m}\sum_{i=1}^{m} SE_i^2 \tag{3.2}$$

$$V_{between} = \frac{\sum_{i=1}^{m}(\theta_i - \bar{\theta})^2}{m-1} \tag{3.3}$$

$$V_{total} = V_W + V_B + \frac{V_B}{m} \tag{3.4}$$

$$\bar{\theta} \pm t_{df,1-\alpha/2} * \sqrt{V_{total}} \tag{3.5}$$

- Jackknife then multiply impute: Altough a more detailed overview of the proposed jackknife estimator is provided in Chapter 4, briefly, the observed dataset will be jackknifed to obtain $j = 200$ samples, each of which will be imputed $m = 2$ times, with a maximum of $maxit = 5$ iterations. The mean of the jackknife estimates will serve as the final point estimate, and a 95% confidence interval will be generated through the percentile method, where the $\alpha/2^{th}$ and $1-(\alpha/2)^{th}$ percentiles will be the lower and upper bounds, respectively.

Thereafter, the methods examined will be compared concerning their point estimates, confidence intervals, and computational expense.

All analyses will be conducted using R 4.2.1 (Funny-Looking Kid).

4 The Proposed Jackknife Estimator

A pseudocode overview of the jackknife estimator proposed may be seen in the following figure.

Briefly, the algorithm begins by obtaining j jackknife subsamples from the observed dataset with missing observations. Thereafter, each of the j subsamples are imputed m times, resulting in a total of $j \times m$ complete datasets. Subsequently, the analysis model of interest to estimate θ is applied to each of the completed datasets to produce $\hat{\theta^*}_1, \hat{\theta^*}_2, \hat{\theta^*}_3, \ldots, \hat{\theta^*}_{\binom{n}{d}}$. The point estimate then becomes the mean of the previously produced n pseudo-estimates, with the confidence interval the $\alpha/2^{th}$ and $1-\alpha/2^{th}$, respectively.

As part of the algorithm, researchers must choose values d and j, which will be context-dependent quantities. Ideally, a d value which satisfies $\frac{\sqrt{n}}{d} \to 0$ will provide asymptotically unbiased estimates even for non-smooth statistics (Chen and Shao 2001; Shao and Wu 1989). Rewriting the foregoing condition for d

$$\frac{\sqrt{n}}{d} \to 0 \tag{4.1}$$

$$\implies d >> \sqrt{n} \tag{4.2}$$

$$\text{Since } d < n \tag{4.3}$$

$$\implies n > d >> \sqrt{n} \tag{4.4}$$

It is evident that d should take on some value between n and $\sqrt{n}$ with d being closer to n, particularly for non-smooth statistics. At any rate, $j = \binom{n}{d}$ will likely be a value that is not computationally feasible to obtain. As such, the number of subsamples required, j, can be limited to yield the estimator more accessible. The choice of j will be a multifaceted decision, where, if possible, greater values are preferred. Ideally, a small pilot study may be performed with a range of j values to determine values of j for which estimates begin to converge.

Although not as widely applicable, researchers may consider utilizing a delete-one jackknife, as discussed in Section 2.2.1 Given the stochastic nature of multiple imputation, especially in instances where a high proportion of missingness is present, the pseudo-estimates may vary widely between any two given jackknife subsamples, similar to what would be observed in the case of percentiles. As such, the delete-one jackknife approach is not recommended for general use but could be considered in samples with low missingness proportion.

Algorithm 1: Algorithm for Jackknife Variance Estimator

$df \leftarrow n \times p$ matrix with missing observations.

df_m^i is the m^{th} imputed dataset.

df^j is the j^{th} jackknife sample with d observation dropped.

In many cases $\binom{n}{d}$ is a large value, which makes obtaining all possible jackknife subsamples infeasible. As such, an arbitrarily large number of subsamples are generated, denoted by j.

Imputations $\leftarrow$ **for** *1* **to** m **do**

  Create j jackknife subsamples from the observed dataset, resulting in j subsamples of size $n - d$, and impute each subsample m times, resulting in $m \times j$ complete datasets.

end

Point estimates $\leftarrow$ **for** *1* **to** $j \times m$ **do**

  Apply analysis model to **Imputations** to obtain a vector of length $j \times m$ containing estimates.

end

$$\hat{\theta}_{jack} = \frac{1}{j \times m} \sum_{i=1}^{j \times m} \hat{\theta}_i \tag{1}$$

Finally, the confidence interval will take on the form

$$(\hat{\theta}^{\alpha/2}, \hat{\theta}^{1-(\alpha/2)}) \tag{2}$$

In general, $m < 5$ while $\sqrt{n} << d < n$ is sufficient. However, for statistics sensitive to subtle changes between the different jackknife subsamples (such as the various percentiles), values for m, d, and j should be adjusted.

Figure 4.1: A pseudocode depiction of the proposed estimator.

5 Results

Per the Methods section, the performance of the proposed jackknife estimator was compared to two leading methods in the literature, Rubin's rules following multiple imputation and Bootstrap resampling prior to multiple imputation. The methods were compared concerning the coverage probabilities they generated, the widths of their respective confidence intervals, their computational expense, and the bias of their point estimators.

5.1 Point Estimates

All methods, perhaps with the exception of Rubin's rules, produced reasonable point estimates with minimal bias. Rubin's rules resulted in slightly anticonservative point estimates with greater standard deviation, indicative of a statistically inefficient estimator with high variability. This finding is unsurprising given the literature on Rubin's rules and its performance under uncongeniality.

In contrast, it is noted that both resampling methods examined provide nearly unbiased point estimates with smaller standard deviations compared to Rubin's rules. Again, given the literature on uncongeniality, this finding was unsurprising; however, given the lack of both empirical and theoretical justification for the bootstrap approach in small sample sizes, the desirable properties noted are worthy of further examination. Between all three methods, nevertheless, it is noted that the proposed jackknife estimator produced estimates with the smallest amount of bias, as well as the smallest standard deviation, an observation perhaps better observed in Section 5.1.2, where the distribution of the biases of the point estimates is compared. From the kernel density plots presented, it is evident that Rubin's rules slightly underestimate the parameter. Compared to Rubin's rules, both resampling approaches provide point estimates that are nearly unbiased; however, it is noted that the bootstrap estimates, despite being centered near zero, are slightly more dispersed compared to the jackknife estimates, which present as a narrow distribution centered at zero. This observation is justified by the values provided in Section 5.1.1 which demonstrate that the jackknife point estimates have a smaller standard deviation than the bootstrap approach and Rubin's rules.

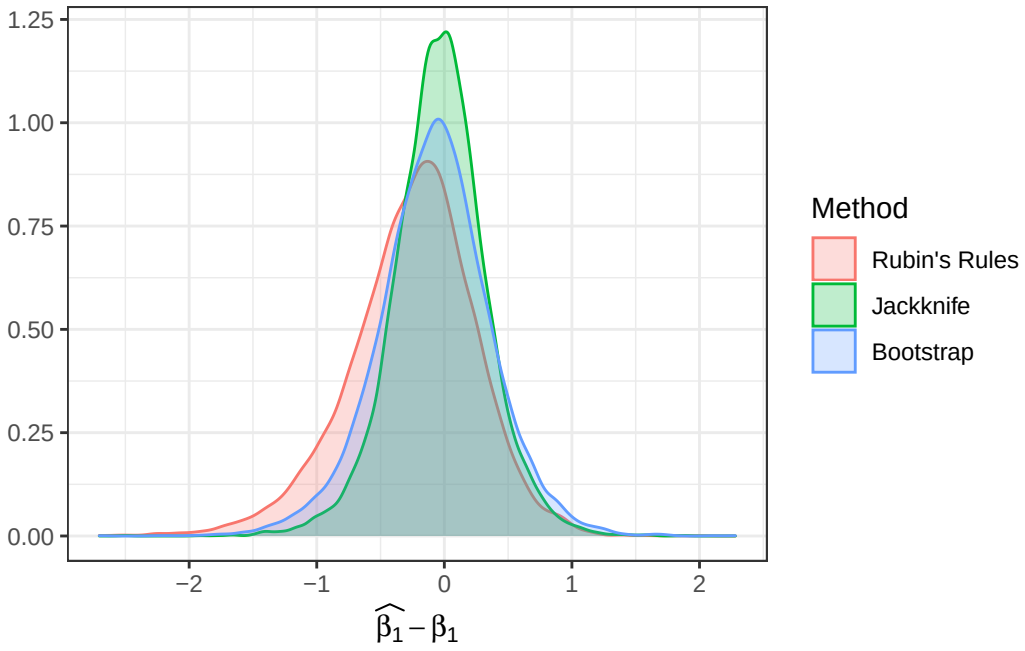
From a point estimation perspective, the statistically desirable properties of the jackknife estimator, alongside its theoretical and empirical justification when used with small sample sizes, yield it a desirable estimator.

5.1.1 Summary of Point Estimation Properties

```
# A tibble: 3 x 4
  Method        `Median Point Estimates ` `Mean Point Estimates` SD of Point E~1
  <chr>                             <dbl>                  <dbl>           <dbl>
1 Jackknife                          1.97                   1.97           0.365
2 Bootstrap                          1.93                   1.92           0.444
3 Rubin's Rules                      1.79                   1.75           0.496
# ... with abbreviated variable name 1: `SD of Point Estimates`
```

5.1.2 Distribution of Point Estimator Bias

```
point_dist
```



5.2 Confidence Intervals

Over-coverage of confidence intervals is noted across all methods; however, particulary with the jackknife and bootstrap approaches, such over-coverage is only slightly over the nominal, as such, they are likely not of concern and can be explained, in part, by the Monte Carlo standard error for the true coverage proabability of a 95% confidence interval. Among the three

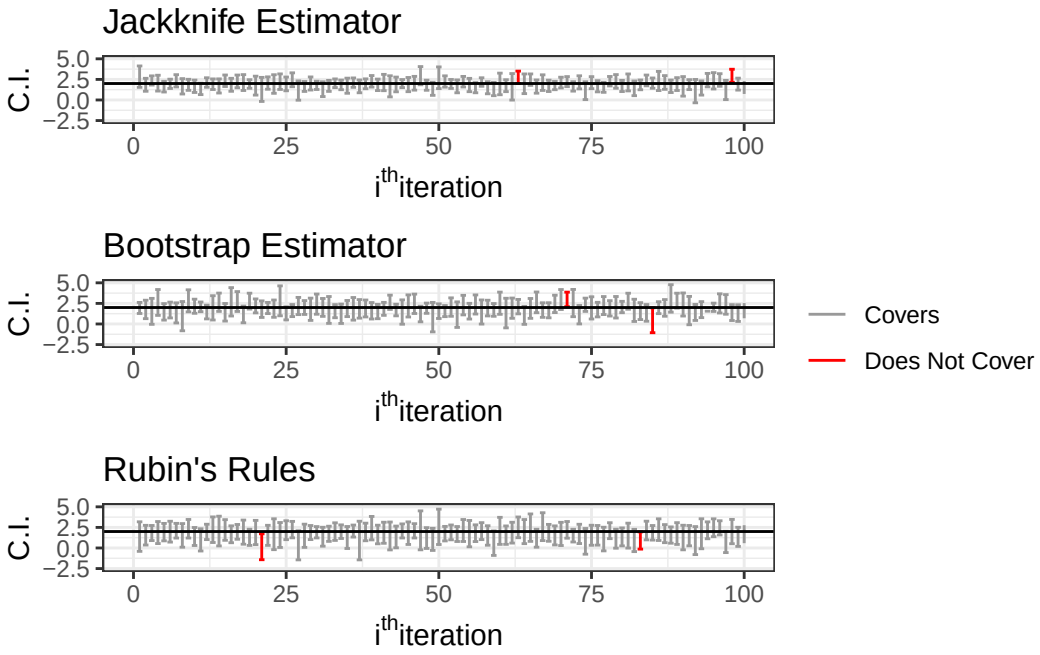
methods noted, Rubin's rules, by a significant margin, deviates from the nominal coverage, indicative of an overly conservative estimator. An argument could be made, particulary in the case of biological studies that an overly conservative estimator is safer than one that is anti-conservative; however, the statistical inefficieny created by conservative estimators can be of concern, particulary in instances where small sample sizes are present or the test being utilized already has low statistical power. Comparing these methods concerning confidence interval width, it is noted that both resampling approaches provide narrower confidence intervals, with the jackknife approach providing the narrowest confidence intervals by a wide margin. In instances where nominal, or near-nominal coverage is reached, narrower confidence intervals are indicative of more efficient estimators. Given the near-nominal coverage noted with the jackknife estimator, combined with the narrow confidence intervals it produces, its superity to the two other methods may be inferred.

A visual overview of the coverage probabilities may be noted in Section 5.2.1, where zipper plots of the methods are presented utilizing a simple random sample of 100 observations from the Monte Carlo simulation results of all methods examined. Given the small number of subsamples examined for visual clarity, the plots may not be indicative of the larger results presented above; however, they provide an appreciation for the meaning of coverage probabilities.

```
# A tibble: 3 x 5
  Method        `Coverage Probability` `Median C.I. Width` Mean C.I. W~1 SD of~2
  <chr>                          <dbl>               <dbl>         <dbl>   <dbl>
1 Jackknife                       97.7                1.61          1.68   0.520
2 Bootstrap                       97.7                2.16          2.23   0.642
3 Rubin's Rules                   98.5                2.46          2.54   0.723
# ... with abbreviated variable names 1: `Mean C.I. Width`,
#   2: `SD of C.I. Width`
```

5.2.1 Zipper Plots for Confidence Interval Coverage

```
ggarrange(jackk, boot, rubin, ncol = 1, nrow = 3, common.legend = TRUE, legend="right")
```



5.3 Performance Benchmark Results

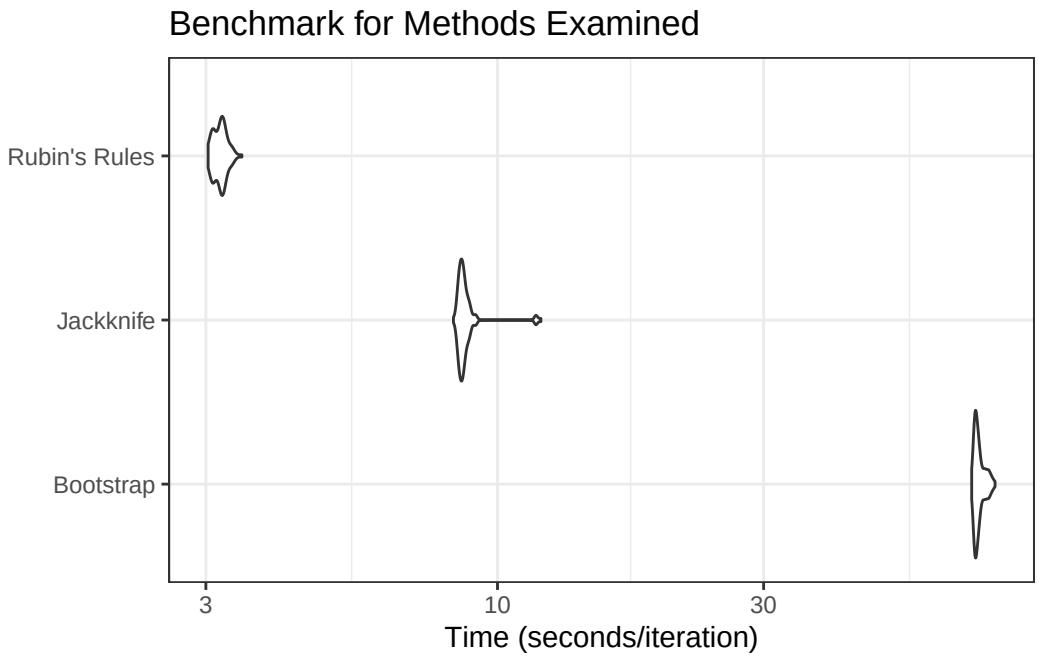
Finally, the three methods are compared concerning their computational expenses. Comparing the two approaches, which require further resampling, it is noted that the bootstrap approach takes nearly ten times longer than the jackknife approach per iteration. Unsurprisingly, Rubin's rules, which do not rely on any further resampling besides the one performed during multiple imputation, was the fastest approach. However, given its biased estimates under uncongeniality or misspecification, the computational advantage it brings to the table adds very little value.

Despite generating the same number of subsamples ($n = 200$), albeit in contrasting manners, with the same number of imputations ($m = 2$) and iterations ($maxit = 5$), it is surprising that the bootstrap approach took nearly ten times longer per iterations to provide estimates. Regardless, the significantly reduced computational expense of the jackknife estimator yields it superior to the bootstrap approach under this particular scenario.

```
# A tibble: 3 x 4
  Method        `Mean Time (seconds/iteration)` Median Time (seconds/i~1 SD of~2
  <chr>                                   <dbl>                     <dbl>   <dbl>
1 Jackknife                                8.80                      8.66  0.632
2 Bootstrap                               73.0                       72.5   1.62
3 Rubin's Rules                            3.18                      3.19  0.0913
# ... with abbreviated variable names 1: `Median Time (seconds/iteration)`,
```

```
#   2: `SD of Time (seconds/iteration)`
```

```
benchmark_plot
```



6 Discussion

6.1 Conclusion

In this paper, a jackknife estimator for multiply imputed outcome variables under the concern of uncongeniality was presented. The proposed estimator was compared to two alternative approaches in the literature employing a Monte Carlo simulation study, where all methods were evaluated on the bias of their point estimates, the width and coverage of their confidence intervals, and computational time. All procedures were found to slightly over-cover, suggesting conservative variance estimates, with Rubin's rules resulting in the broadest confidence intervals with significant over-coverage compared to the nominal level. In contrast, the two resampling-based approaches examined resulted in a substantial decline in confidence interval width, with the proposed jackknife estimator providing the narrowest confidence intervals by a wide margin while still attaining near-nominal coverage.

Unsurprisingly, Rubin's rules were the least computationally costly approach among the ones examined; however, given its downward-biased point estimates and wide confidence intervals, particularly in instances where uncongeniality is a concern, resampling-based robust methods should be preferred. Among the two resampling-based methods examined, the bootstrap approach took nearly ten times longer per iteration, which was a surprising observation, as both the jackknife and bootstrap methods utilized the same number of imputations, iterations, and subsamples. From a computational perspective, there is no evident reason why the bootstrap method should take nearly ten times longer than the proposed jackknife estimator. It is possible that the R package utilized for the bootstrap approach, *bootImpute*, is not optimized with speed concerns in mind. However, neither was the jackknife estimator. In both instances, the most time-consuming aspect of the processes was the imputation of the subsamples generated, either via bootstrap resampling or jackknife resampling, done with the same R package, *mice*, and the same parameters, number of iterations, and imputations. Given that the foregoing steps took place under identical conditions, perhaps a different aspect of the two approaches could explain the discrepancy noted.

Nevertheless, given the superior performance of the jackknife estimator noted, alongside its computational efficiency, the recommendation to replace Rubin's rules as the de facto standard in small, multiply imputed datasets with our proposed estimator is made.

6.2 Future Directions

Perhaps the most significant issue noted with the jackknife estimator was a slightly higher coverage probability compared to the nominal. Although the Monte Carlo error may partially explain this observation, alternative confidence interval construction approaches could be considered to yield more appropriate coverage. Namely, it is possible that a confidence interval could be generated by examining various values calculated based on Rubin's rules, such as the fraction of missing information, relative efficiency, and relative increase in variance. The aforestated statistics could be used to appropriately model the uncertainty in the imputation process, which, in turn, could be used to generate confidence intervals in a semi-parametric manner.

Otherwise, although the proposed estimator was efficient even with a small number of subsamples and imputations, alternative confidence interval constructions could allow one to utilize even fewer subsamples while still attaining nominal coverage, which may make the proposed method even more computationally feasible.

References

Anderson, James C., and David W. Gerbing. 1984. "The Effect of Sampling Error on Convergence, Improper Solutions, and Goodness-of-Fit Indices for Maximum Likelihood Confirmatory Factor Analysis." *Psychometrika* 49 (2): 155–73. https://doi.org/10.1007/BF02294170.

Bartlett, Jonathan W. 2021. "Reference-Based Multiple Imputation—What Is the Right Variance and How to Estimate It." *Statistics in Biopharmaceutical Research*, November, 1–9. https://doi.org/10.1080/19466315.2021.1983455.

Bartlett, Jonathan W, and Rachael A Hughes. 2020. "Bootstrap Inference for Multiple Imputation Under Uncongeniality and Misspecification." *Statistical Methods in Medical Research* 29 (12): 3533–46. https://doi.org/10.1177/0962280220932189.

Bentler, P. M., and Chih-Ping Chou. 1987. "Practical Issues in Structural Modeling." *Sociological Methods & Research* 16 (1): 78–117. https://doi.org/10.1177/0049124187016001004.

Buuren, Stef van. 2012. *Flexible Imputation of Missing Data*. Chapman & Hall/CRC Interdisciplinary Statistics Series. Boca Raton, FL: CRC Press.

Chen, Jiahua, and Jun Shao. 2001. "Jackknife Variance Estimation for Nearest-Neighbor Imputation." *Journal of the American Statistical Association* 96 (453): 260–69. https://doi.org/10.1198/016214501750332839.

Faber, Jorge, and Lilian Martins Fonseca. 2014. "How Sample Size Influences Research Outcomes." *Dental Press Journal of Orthodontics* 19 (4): 27–29. https://doi.org/10.1590/2176-9451.19.4.027-029.ebo.

Fay, Robert E. 1992. "When Are Inferences from Multiple Imputation Valid?."

Jackson, Dennis L. 2001. "Sample Size and Number of Parameter Estimates in Maximum Likelihood Confirmatory Factor Analysis: A Monte Carlo Investigation." *Structural Equation Modeling: A Multidisciplinary Journal* 8 (2): 205–23. https://doi.org/10.1207/S15328007SEM0802_3.

LaFontaine, Denise. 2021. "The History of Bootstrapping: Tracing the Development of Resampling with Replacement." *The Mathematics Enthusiast* 18 (1–2): 78–99. https://doi.org/10.54870/1551-3440.1515.

Mammen, Enno. 1992. *When Does Bootstrap Work?* Vol. 77. Lecture Notes in Statistics. New York, NY: Springer New York. https://doi.org/10.1007/978-1-4612-2950-6.

Meng, Xiao-Li. 1994. "Multiple-Imputation Inferences with Uncongenial Sources of Input." *Statistical Science* 9 (4): 538–58. https://doi.org/10.1214/ss/1177010269.

Rao, J. N. K., and J. Shao. 1992. "Jackknife Variance Estimation with Survey Data Under Hot Deck Imputation." *Biometrika* 79 (4): 811–22. https://doi.org/10.2307/2337236.

Rubin, Donald B. 1978. "Multiple Imputations in Sample Surveys-a Phenomenological

Bayesian Approach to Nonresponse." In *Proceedings of the Survey Research Methods Section of the American Statistical Association*, 1:20–34. American Statistical Association Alexandria, VA, USA.

Shao, Jun, and C. F. J. Wu. 1989. "A General Theory for Jackknife Variance Estimation." *The Annals of Statistics* 17 (3): 1176–97. https://doi.org/10.1214/aos/1176347263.

Wang, N. 1998. "Large-Sample Theory for Parametric Multiple Imputation Procedures." *Biometrika* 85 (4): 935–48. https://doi.org/10.1093/biomet/85.4.935.

Weisstein, Eric W. n.d. "Smooth Function." Text. https://mathworld.wolfram.com/.

Wicklin, Rick. 2017. "Jackknife Estimates in SAS." *SAS Blog*. https://blogs.sas.com/content/iml/2017/06/21/jackknife-estimate-standard-error-sas.html.

Xie, Xianchao, and Xiao-Li Meng. 2016. "Dissecting Multiple Imputation from a Multi-Phase Inference Perspective: What Happens When God's, Imputer's and Analyst's Models Are Uncongenial?" *Statistica Sinica*. https://doi.org/10.5705/ss.2014.067.

Appendix

Notation

- N is the total number of units in the finite population being targeted.
- X is an $N \times q$ matrix of fully observed covariates.
- Y is an $N \times p$ matrix of partially observed outcome variables.
- R is an $N \times p$ matrix of response indicators (i.e., $R_{ij} = 1$ if the response on Y_{ij} is obtained and $R_{ij} = 0$ otherwise.)
- Q is an unknown quantity of interest to the analyst.
- $Z_c = \{X, Y_{inc}\}$ is the complete data.
- $Z_o = \{X, Y_{obs}, R_{inc}\}$ is the incomplete (i.e., observed) data.
- The analyst's complete-data procedure is summarized by $\mathcal{P}_{com} = [\widehat{Q}(X, Y_{inc}), U(X, Y_{inc})]$, where $\widehat{Q}(X, Y_{inc})$ is an estimator of Q with associated variance $U(X, Y_{inc}.)$
- R is not a part of $\mathcal{P}_{com}$, as the missing at random assumption implies that the response behavior itself carries no information about Q.

Formal Definition of Congeniality

In short, one may define congeniality as the imputer and analyst making different assumptions regarding the data. The following two-part formal definition of uncongeniality was proposed by Meng in 1994, and will be utilized in our research. Meeting the assumptions set forth in the following two definitions qualifies the imputation model as being congenial to the analysis model, or vice versa.

Definition 1

Let E_f and V_f denote posterior mean and variance with respect to f, respectively. A Bayesian model f is said to be congenial to the analysis procedure $\mathcal{P} \equiv \{\mathcal{P}_{obs}; \mathcal{P}_{com}\}$ for given Z_o if the following hold:

- The posterior mean and variance of θ under f given the incomplete data are asymptotically the same as the estimate and variance from the analyst's incomplete-data procedure $\mathcal{P}_{obs}$, that is,
$$[\hat{\theta}(Z_o), U(Z_o)] \simeq [E_f[\theta|Z_o], V_f[\theta|Z_o]] \tag{6.1}$$

- The posterior mean and variance of θ under f given the complete data are asymptotically the same as the estimate and variance from the analyst's complete-data procedure $\mathcal{P}_{com}$, that is,
$$[\hat{\theta}(Z_c), U(Z_c)] \simeq [E_f[\theta|Z_c], V_f[\theta|Z_c]] \tag{6.2}$$

  for any possible $Y_{inc} = (Y_{obs}, Y_{miss})$ with Y_{obs} conditioned upon.

If the foregoing conditions are met, f is said to be second-moment congenial to $\mathcal{P}$.

Definition 2

The analysis procedure $\mathcal{P}$ is said to be congenial to the imputation model $g(Y_{miss}|Z_o, A)$ where A represents possible additional data the imputer has access to, if one can find an f such that (**i**) f is congenial to $\mathcal{P}$ and (**ii**) the posterior predictive density for Y_{miss} derived under f is identical to the imputation model $f(Y_{miss}|Z_o) = g(Y_{miss}|Z_o, A)\ \forall\ Y_{miss}$.

Raw Data

```
# A tibble: 10,000 x 13
   true_var UB_boot LB_boot point_es~1 point~2 LB_rub UB_rub    UB    LB point~3
      <dbl>   <dbl>   <dbl>      <dbl>   <dbl>  <dbl>  <dbl> <dbl> <dbl>   <dbl>
 1        2    2.76   0.971       2.04    2.13  1.07    3.20  2.62 1.54     2.19
 2        2    2.92   1.27        2.14    2.08  1.08    3.07  2.80 1.71     2.25
 3        2    2.90   0.937       1.94    2.10  1.09    3.11  2.76 1.30     2.03
 4        2    2.31   0.978       1.67    1.46  0.767   2.15  2.15 1.34     1.74
 5        2    3.13   1.08        2.03    2.12  1.30    2.94  2.68 1.01     1.87
 6        2    3.06   1.15        2.15    1.84  0.646   3.04  2.87 1.61     2.11
 7        2    2.68   1.23        1.95    1.80  0.728   2.87  2.60 1.46     1.98
 8        2    2.70  -0.184       1.52    1.33  0.132   2.53  2.52 0.418    1.54
 9        2    4.03   1.58        2.76    2.19  0.699   3.69  3.81 1.37     2.60
10        2    3.04   0.521       1.82    1.27 -0.252   2.78  2.90 0.998    1.89
# ... with 9,990 more rows, 3 more variables: rubin_width <dbl>,
#   jackknife_width <dbl>, boot_width <dbl>, and abbreviated variable names
#   1: point_estimate_boot, 2: point_estimate_rub, 3: point_estimate
```

Code

Please see our publicly available GitHub repo for the simulation code. Otherwise, the code utilized through the paper has been included below.

```
# Aggregated code here for organization, make it available in an appendix.
library(tidyverse)
library(ggpubr)

combined_results <- read_csv("final_sim_data_agg.csv") %>%
  select(-"...1")

benchmark_res <- read.csv("benchmark_detailed_res.csv") %>%
  select(-X) %>%
  group_by(expr) %>%
  summarise("Mean Time (seconds/iteration)" = mean(time)/1e6/100,
            "Median Time (seconds/iteration)" = median(time)/1e6/100,
            "SD of Time (seconds/iteration)" = sd(time)/1e6/100) %>%
  rename("Method" = expr)

benchmark_res_ord <- benchmark_res[c(2,1,3),]

jackknife_coverage <-
  round((sum(combined_results$true_var < combined_results$UB &
               combined_results$true_var > combined_results$LB) / nrow(combined_results))*

rubin_coverage <-
  round((sum(combined_results$true_var < combined_results$UB_rub &
               combined_results$true_var > combined_results$LB_rub) / nrow(combined_result

boot_coverage <-
  round((sum(combined_results$true_var < combined_results$UB_boot &
               combined_results$true_var > combined_results$LB_boot) / nrow(combined_resul

ci_sum <- tibble(
  "Method" = c("Jackknife", "Bootstrap", "Rubin's Rules"),

  "Coverage Probability" = c(jackknife_coverage, rubin_coverage, boot_coverage),

  "Median C.I. Width" = c(median(combined_results$jackknife_width), median(combined_result
```

```
  "Mean C.I. Width" = c(mean(combined_results$jackknife_width), mean(combined_results$boot

  "SD of C.I. Width" = c(sd(combined_results$jackknife_width), sd(combined_results$boot_wi
)

point_estimate_sum <- tibble(
  "Method" = c("Jackknife", "Bootstrap", "Rubin's Rules"),

  "Median Point Estimates " = c(median(combined_results$point_estimate), median(combined_r

  "Mean Point Estimates" = c(mean(combined_results$point_estimate), mean(combined_results$

  "SD of Point Estimates" = c(sd(combined_results$point_estimate), sd(combined_results$poi
)

set.seed(234)
jackk <- combined_results %>%
  sample_n(1e2, replace = FALSE) %>%
  mutate(covers = ifelse(UB > true_var & true_var > LB, "Covers", "Does Not Cover")) %>%
  ggplot(., aes(x = 1:100)) +
  geom_errorbar(aes(ymin = LB, ymax = UB, color = covers)) +
  theme_bw() +
  #coord_flip() +
  labs(
    x = latex2exp::TeX("$i^{th} iteration"),
    y = "C.I."
  ) +
  geom_hline(yintercept = combined_results$true_var) +
  scale_color_manual(name = NULL, values=c("#999999", "#FF0000")) +
  ggtitle("Jackknife Estimator") +
  xlim(c(0,100)) +
  ylim(c(-2.5,5))

set.seed(24123)
boot <- combined_results %>%
  sample_n(1e2, replace = FALSE) %>%
  mutate(covers = ifelse(UB_boot > true_var & true_var > LB_boot, "Covers", "Does Not Cove
  ggplot(., aes(x = 1:100)) +
  geom_errorbar(aes(ymin = LB_boot, ymax = UB_boot, color = covers)) +
  theme_bw() +
  #coord_flip() +
```

```
  labs(
    x = latex2exp::TeX("$i^{th} iteration"),
    y = "C.I."
  ) +
  geom_hline(yintercept = combined_results$true_var) +
  scale_color_manual(name = NULL, values=c("#999999", "#FF0000")) +
  ggtitle("Bootstrap Estimator") +
  xlim(c(0,100)) +
  ylim(c(-2.5,5))

set.seed(234)
rubin <- combined_results %>%
  sample_n(1e2, replace = FALSE) %>%
  mutate(covers = ifelse(UB_rub > true_var & true_var > LB_rub, "Covers", "Does Not Cover"
  ggplot(., aes(x = 1:100)) +
  geom_errorbar(aes(ymin = LB_rub, ymax = UB_rub, color = covers)) +
  theme_bw() +
  #coord_flip() +
  labs(
    x = latex2exp::TeX("$i^{th} iteration"),
    y = "C.I."
  ) +
  geom_hline(yintercept = combined_results$true_var) +
  scale_color_manual(name = NULL, values=c("#999999", "#FF0000")) +
  ggtitle("Rubin's Rules") +
  xlim(c(0,100)) +
  ylim(c(-2.5,5))

point_dist <- reshape2::melt(combined_results[c("point_estimate_rub", "point_estimate", "p
  ggplot(data = .,
         aes(x = value - 2, fill = variable, color = variable)) +
  geom_density(aes(y = ..density..), alpha = 0.25) +
  theme(axis.title.y = element_blank(),
        panel.spacing=unit(1.5,"lines")) +
  theme_bw() +
  theme(axis.title.y = element_blank(),
        panel.spacing=unit(1.5,"lines"),
        strip.text = element_text(
          size = 9)) +
  labs(
    x = latex2exp::TeX("$\\widehat{\\beta_1} - \\beta_1")
```

```
  ) +
  scale_fill_discrete(name = "Method", labels = c("Rubin's Rules", "Jackknife", "Bootstrap
  scale_color_discrete(name = "Method", labels = c("Rubin's Rules", "Jackknife", "Bootstra

benchmark_plot <- read.csv("benchmark_detailed_res.csv") %>%
  select(-X) %>%
  ggplot(., aes(x = expr, y = time/1e6/100)) +
  geom_violin() +
  scale_y_log10() +
  theme_bw() +
  ggtitle("Benchmark for Methods Examined") +
  coord_flip() +
  labs(
    y = "Time (seconds/iteration)"
  ) +
  theme(axis.title.y = element_blank())
```