Statistics and Its Interface Volume 15 (2022) 515–526

The more data, the better? Demystifying deletion-based methods in linear regression with missing data*

Tianchen Xu†, Kun Chen, and Gen Li

We compare two deletion-based methods for dealing with the problem of missing observations in linear regression analysis. One is the complete-case analysis (CC, or listwise deletion) that discards all incomplete observations and only uses common samples for ordinary least-squares estimation. The other is the available-case analysis (AC, or pairwise deletion) that utilizes all available data to estimate the covariance matrices and applies these matrices to construct the normal equation. We show that the estimates from both methods are asymptotically unbiased under missing completely at random (MCAR) and further compare their asymptotic variances in some typical situations. Surprisingly, using more data (i.e., AC) does not necessarily lead to better asymptotic efficiency in many scenarios. Missing patterns, covariance structure and true regression coefficient values all play a role in determining which is better. We further conduct simulation studies to corroborate the findings and demystify what has been missed or misinterpreted in the literature. Some detailed proofs and simulation results are available in the online supplemental materials.

AMS 2000 subject classifications: Primary 62D10, 62J05; secondary 62F12.

Keywords and phrases: Asymptotic variance, Available-case analysis, Complete-case analysis, Missing data.

1. INTRODUCTION

Missing data are very common in linear regression analysis. Dong and Peng (2013) described missing data as "a rule rather than an exception in quantitative research." For instance, longitudinal data may be incomplete due to unexpected dropout, and survey data may be incomplete due to refusal of respondents or wrong answers. Since inappropriate treatments on missing data can severely undermine the validity of inference and conclusion of a study, researchers have developed many methods to conquer this challenge.

Deletion-based methods usually involve complete-case analysis (CC, or listwise deletion) and available-case analysis (AC, or pairwise deletion). Peng et al. (2006) examined 569 papers with missing data published in 11 education journals from 1998 to 2004 and found that 552 (97%) employed deletion-based methods; Lang and Little (2018) reviewed 169 papers with missing data in *Prevention Science* from February 2013 to July 2015 and found that 62 (37%) studies used deletion-based methods. Especially recently, there is an increasing trend in applying AC method or its variants to high dimensional data such as block-missing multi-modality datasets where each subject has missing blocks from certain modality sources (Yu et al., 2020; Xue and Qu, 2020). CC method utilizes the complete dataset in which any incomplete rows are discarded and is the default setting for many multivariate procedures and regressions analysis in popular statistical packages such as SAS, SPSS, SYSTAT and R. AC method computes statistics using the rows for which every constituent variables are observed and is the default setting for descriptive, correlation, and regression analysis when using either correlation or covariance matrices in SAS, SPSS and SYSTAT. The `cov` function and `regtools` package in R also provide AC analysis for correlation estimation and linear regression. The goal of this article is to compare the performance of these two methods. Particularly, we mainly focus on the classical low-dimensional settings with the assumption that the proportion of complete observation is positive (to ensure CC method is feasible) and some typical block-wise missing patterns.

Other mainstream treatments for missing data in regression analysis include: 1) imputation, 2) weighting, and 3) maximum-likelihood based methods (Lang and Little, 2018; Little and Rubin, 2019). Imputation methods try to impute the missing part of the dataset. Single imputation often imputes the missing values with some fixed values (e.g., mean values), randomly drawn values from the same variable (simple hot-deck) or predictive values from other variables. Multiple imputation (MI) imputes the missing data while acknowledging the uncertainty associated with the imputed values (Rubin, 1977, 1996). Weighting approaches discard incomplete samples and assign weights to each subject according to some missing features to reduce the bias and variance of final inference (Seaman and White, 2013).

arXiv: 2010.13332
*We thank Dr. Rod Little for helpful comments.
†Corresponding author.

Maximum-likelihood methods such as information maximum likelihood (FIML, also known as direct maximum likelihood) consider only the observed samples when calculating the sample log-likelihood function and maximize it using EM algorithm to estimate parameters (Enders and Bandalos, 2001; Olinsky, Chen and Harlow, 2003). Generally speaking, no technique is universally the best, but they are better than the complete case analysis (CC) in most cases (Tsikriktsis, 2005; Xu et al., 2015; Gao et al., 2018).

Under the missing completely at random (MCAR) assumption, deletion-based methods are the fully automatic methods, which makes them popular even there is no consensus about whether AC is better than CC or not. There have been intense debates in the literature about the merits and flaws of different deletion-based methods. In particular, AC and CC are often in the center of the controversy. Glasser (1964) is the first researcher (as far as we know) who systematically introduced the AC estimator in the context of linear regression. He argued that the AC estimator is consistent and derived its asymptotic variance. In the simulation study with two predictors ($p = 2$), Glasser (1964) concluded that the AC estimator is in general better than the CC estimator if the correlation between two predictors is less than 0.58. However, Haitovsky (1968) pointed out that Glasser (1964)'s asymptotic result which does not involve the true regression coefficients ($\boldsymbol{\beta}$) was not accurate and provided the right asymptotic covariance. He reached an opposite conclusion that "listwise deletion (the CC estimator) is judged superior in almost all the cases" by considering nine simulation scenarios. In contrast to Haitovsky (1968)'s findings, Kim and Curry (1977) did another simulation study and claimed their setting is more typical in sociological studies. The result indicated that the AC method performs better than the CC estimator by using the correlation structure among predictors in Blau and Duncan (1967)'s book. In the following decades, these contradictory papers were frequently cited by researchers to show the comparison between two methods had not been fully settled (Little, 1992; Allison, 2001; Pigott, 2001). Finally, we must point out that all these comparisons mentioned above are under MCAR, since deletion-based method are asymptotically unbiased under this assumption. MCAR is generally a strong and unrealistic assumption which is unlikely to hold in many studies (Little et al., 2012). However, for the simplicity of theoretical results, we still focus on the performance comparison under this assumption in this article.

The rest of the paper is organized as follows. In Section 2, we review the existing results of both methods. In Section 3, we compare the performance of any scalar regression coefficient estimator in realistic situations. We show that the estimators from both methods are asymptotically unbiased and using more data (i.e., AC) does not necessarily lead to better asymptotic performance. It is necessary to look into the missing patterns, covariance structure and true regression coefficients together to determine which method is better. In Section 4, we conduct simulation studies based on Kim and Curry (1977)'s settings to verify our theoretical propositions and validate our findings in Section 3. With the guidance of the theoretical results, we are able to find out what was missed or misinterpreted in the previous work and provide our suggestions. In the last section, we discuss further research directions.

2. BACKGROUND

2.1 Asymptotic results for complete case

Let $\boldsymbol{X} = (X_1, X_2, \cdots X_p)^T \in \mathbb{R}^p$ be a random vector. Let $Y \in \mathbb{R}$ be a random variable such that

$$Y = \boldsymbol{X}^T\boldsymbol{\beta} + \varepsilon,$$

where $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_p)^T \in \mathbb{R}^p$ is a coefficient vector, $\varepsilon \in \mathbb{R}$ is a random variable with mean 0 and variance σ^2. Furthermore, we assume X_j is independent of ε.

Let $\boldsymbol{Z} = (Z_1, \cdots, Z_{p+1})^T \triangleq \left(\boldsymbol{X}^T, Y\right)^T$ to be a $(p+1)$-dimensional random vector with mean $\boldsymbol{\mu}$ and non-singular covariance matrix $\boldsymbol{\Sigma}$. Assume all fourth-order moments of $\boldsymbol{Z}$ are finite. Partition $\boldsymbol{\Sigma}$ conformably as follows:

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_x & \boldsymbol{\Sigma}_{xy} \\ \boldsymbol{\Sigma}_{xy}^T & \boldsymbol{\Sigma}_y \end{pmatrix},$$

where $\boldsymbol{\Sigma}_x = \mathrm{Cov}(\boldsymbol{X})$, $\boldsymbol{\Sigma}_{xy} = \mathrm{Cov}(\boldsymbol{X}, Y) = \boldsymbol{\Sigma}_x\boldsymbol{\beta}$, $\boldsymbol{\Sigma}_y = \mathrm{Var}(Y) = \boldsymbol{\beta}^T\boldsymbol{\Sigma}_x\boldsymbol{\beta} + \sigma^2$. Let μ_j denote the jth element in $\boldsymbol{\mu}$. Let σ_{jk} denote the (j,k)th element in $\boldsymbol{\Sigma}$, and conventionally we use σ_j^2 to denote the elements on the diagonal of $\boldsymbol{\Sigma}$ (i.e., $\sigma_j^2 = \sigma_{jj}$).

We collect a set of observation data $\{Z_{1i}, \cdots Z_{p+1,i}\}_{i=1,\cdots,n}$ from n independent samples and assume there are not any missing data in this section. Define the sample covariance matrix $\boldsymbol{S} = [s_{jk}]$ with entries:

$$s_{jk} = \frac{1}{n}\sum_{i=1}^{n}(Z_{ji} - \bar{Z}_j)(Z_{ki} - \bar{Z}_k),$$

where $\bar{Z}_j$ is the sample mean of Z_j $(j = 1, \cdots, p+1)$. Similar to $\boldsymbol{\Sigma}$, we also partition $\boldsymbol{S}$ into four parts correspondingly:

$$\boldsymbol{S} = \begin{pmatrix} \boldsymbol{S}_x & \boldsymbol{S}_{xy} \\ \boldsymbol{S}_{xy}^T & S_y \end{pmatrix},$$

where $\boldsymbol{S}_x$, $\boldsymbol{S}_{xy}$, S_y are the sample covariance/variance of $\boldsymbol{X}$, $(\boldsymbol{X}, Y)$ and Y. Then the least-squares estimator of $\boldsymbol{\beta}$ is well known:

$$\hat{\boldsymbol{\beta}} = \boldsymbol{S}_x^{-1}\boldsymbol{S}_{xy}.$$

The sample quantities $\boldsymbol{S}$ and $\hat{\boldsymbol{\beta}}$ are consistent estimators of their theoretical counterparts $\boldsymbol{\Sigma}$, $\boldsymbol{\beta}$ respectively and are asymptotically normally distributed (Rao, 1973). The

former is guaranteed by the Lindeberg-Levy Central Limit Theorem and the latter can be derived from the Delta method.

Proposition 2.1 (Rao, 1973)**.** Let $\boldsymbol{S}$ be the sample covariance matrix of r.v $\boldsymbol{Z} = \left(\boldsymbol{X}^T, Y\right)^T$, then

$$\sqrt{n}(\text{vec}(\boldsymbol{S}) - \text{vec}(\boldsymbol{\Sigma})) \xrightarrow{d} N\left(\boldsymbol{0}, \boldsymbol{\Phi}\right),$$

where the asymptotic covariance $\boldsymbol{\Phi}$ consists of elements $\phi_{(ij)(mn)} \triangleq \text{Cov}(s_{jk}, s_{mn})$:

$$\phi_{(jk)(mn)} = \text{E}(Z_j - \mu_j)(Z_k - \mu_k)(Z_m - \mu_m)(Z_n - \mu_n) - \sigma_{jk}\sigma_{mn}.$$

Proposition 2.2 (Rao, 1973)**.** Let $\hat{\boldsymbol{\beta}}$ be the least-squares estimator of $\boldsymbol{\beta}$ in the aforementioned regression, then

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{d} N_p(\boldsymbol{0}, \boldsymbol{\Delta}\boldsymbol{\Phi}\boldsymbol{\Delta}^T),$$

where $\boldsymbol{\Delta}$ denotes the matrix of partial derivatives of function $\hat{\boldsymbol{\beta}}(\boldsymbol{S})$ evaluated in $\boldsymbol{\Sigma}$.

The form of $\boldsymbol{\Delta}$ and $\boldsymbol{\Phi}$ depends on the way of vectorizing $\boldsymbol{S}$. In Appendix A (online supplemental material), we provide an example of vectorizing $\boldsymbol{S}$ in columns (i.e., vec($\boldsymbol{\Sigma}$)). Similar results have been obtained in the literature; see White (1980); Van Praag (1981); Bentler (1985) for example.

2.2 Asymptotic results for incomplete case

Suppose there are missing values in predictor matrix $\boldsymbol{X}$. Following Little (1982), the missing pattern is independent of the values of predictors (i.e., missing completely at random, MCAR). Let $\boldsymbol{R} = [R_{ji}]$ $(j = 1, \cdots, p+1; i = 1, \cdots, n)$ be an indicator matrix that

$$R_{ji} = \begin{cases} 0 & \text{if } Z_{ji} \text{ is not observed,} \\ 1 & \text{if } Z_{ji} \text{ is observed.} \end{cases}$$

2.2.1 Available-case analysis

"Available-case analysis (AC) tries to use the largest possible sets of available cases to estimate individual parameters" (Little, 1992; Pigott, 2001). Define the sample covariance matrix in AC method $\boldsymbol{S}_{AC} = [s^{AC}_{jk}]$ with entries:

$$s^{AC}_{jk} = \frac{1}{n_{jk}} \sum_{i \in \tau_{jk}} \left(Z_{ji} - \frac{1}{n_{jk}} \sum_{l \in \tau_{jk}} Z_{jl}\right)\left(Z_{ki} - \frac{1}{n_{jk}} \sum_{l \in \tau_{jk}} Z_{kl}\right),$$

where $\tau_{jk} = \{i : R_{ji}R_{ki} = 1\}$ is the index set of samples that both Z_j and Z_k are observed; n_{jk} is the size of τ_{jk} (i.e., $n_{jk} = \sum_{i=1}^{n} R_{ji}R_{ki}$). A defect of AC method is that the estimated covariance matrix $\boldsymbol{S}_{AC}$ might not be positive definite. However, Van Praag, Dijkstra and Van Velzen (1985) pointed out that the probability of $\boldsymbol{S}_{AC}$ being positive definite tends to 1 as the sample size increases. Similar to $\boldsymbol{S}$, we partition $\boldsymbol{S}_{AC}$ into $\boldsymbol{S}^{AC}_x$, $\boldsymbol{S}^{AC}_{xy}$, S^{AC}_y and define the AC estimator $\hat{\boldsymbol{\beta}}_{AC}$ as follows:

$$\hat{\boldsymbol{\beta}}_{AC} = \left\{\boldsymbol{S}^{AC}_x\right\}^{-1} \boldsymbol{S}^{AC}_{xy}.$$

Let q_j be the proportion of the cases with Z_j observed (i.e., $q_j = \frac{1}{n}\sum_{i=1}^{n} R_{ji}$), and q_{jk} be the proportion of the cases with both Z_j and Z_k observed (i.e., $q_{jk} = \frac{1}{n}\sum_{i=1}^{n} R_{ji}R_{ki}$). Similarly, we also define q_{jkm} and q_{jkmn}. For the AC estimator, the following proposition holds:

Proposition 2.3 (Van Praag, Dijkstra and Van Velzen 1985)**.** Under the MCAR assumption, assuming that the observing proportions (i.e., $q_j, q_{jk}, q_{jkm}, q_{jkmn}$) are not zero and remain the same as sample size n goes to infinity, the asymptotic distribution of $\hat{\boldsymbol{\beta}}_{AC}$ is given by:

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{AC} - \boldsymbol{\beta}) \xrightarrow{d} N_p(\boldsymbol{0}, \boldsymbol{\Delta}(\boldsymbol{\Phi} \circ \boldsymbol{Q})\boldsymbol{\Delta}^T),$$

where $\boldsymbol{Q}$ consists of elements $q_{(jk)(mn)} = \frac{q_{jkmn}}{q_{jk}q_{mn}}$ corresponding to $\phi_{(jk)(mn)}$; $\circ$ represents the Hadamard product.

From the proposition, we conclude that $\boldsymbol{\beta}_{AC}$ is asymptotically unbiased and its asymptotic variance is $\boldsymbol{\Delta}(\boldsymbol{\Phi} \circ \boldsymbol{Q})\boldsymbol{\Delta}^T/n$, obtained by multiplying a specific factor $q_{(jk)(mn)}$ to $\phi_{(jk)(mn)}$ in $\boldsymbol{\Phi}$ that is from the variance of $\hat{\boldsymbol{\beta}}$ in the complete case.

2.2.2 Complete-case analysis

Complete-case analysis (CC) only utilizes the complete samples without any missing data. The CC estimator $\hat{\boldsymbol{\beta}}_{CC}$ is exactly the same as $\hat{\boldsymbol{\beta}}$ in Section 2.1 except that the dataset is constrained to complete samples. Therefore, CC method is feasible only when there exist a sufficient number of complete cases.

Proposition 2.4. Let $\tilde{q}$ denote the proportion of samples that have complete observations, and assume $\tilde{q} > 0$ is a constant. Under the MCAR assumption, the CC estimator $\hat{\boldsymbol{\beta}}_{CC}$ follows:

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{CC} - \boldsymbol{\beta}) \xrightarrow{d} N_p(\boldsymbol{0}, \boldsymbol{\Delta}\boldsymbol{\Phi}\boldsymbol{\Delta}^T/\tilde{q}),$$

Similar to the AC estimator, $\hat{\boldsymbol{\beta}}_{CC}$ is also asymptotically unbiased and its asymptotic variance is $\boldsymbol{\Delta}\boldsymbol{\Phi}\boldsymbol{\Delta}^T/(n\tilde{q})$.

3. COMPARISON BETWEEN AC AND CC

Somewhat surprisingly, although AC makes better use of data by accounting for all available data points, many simulation studies show that AC is markedly inferior to CC on highly correlated data and can be superior to CC on weakly correlated data (Haitovsky, 1968; Kim and Curry, 1977; Little and Rubin, 1989). Since both $\hat{\boldsymbol{\beta}}_{AC}$ and $\hat{\boldsymbol{\beta}}_{CC}$ are consistent estimators of $\boldsymbol{\beta}$, we compare their asymptotic

variances in the article. Let V_{CC}, V_{AC} denote the asymptotic variance of $\hat{\boldsymbol{\beta}}_{CC}$, $\hat{\boldsymbol{\beta}}_{AC}$ respectively. Then the difference V_D is:

$$V_D = V_{CC} - V_{AC} = \frac{1}{n\tilde{q}} \boldsymbol{\Delta} \left\{ \boldsymbol{\Phi} \circ (\mathbf{1} - \boldsymbol{Q}\tilde{q}) \right\} \boldsymbol{\Delta}^T,$$

where $\mathbf{1}$ is a $\frac{p^2+3p}{2}$ by $\frac{p^2+3p}{2}$ matrix with 1 being all of its entries.

Neither method is uniformly better than the other with any fixed missing pattern (see more detailed explanation in Appendix B (online supplemental material)). It turns out that we have to look into the covariance structure $\boldsymbol{\Sigma}$, true coefficient $\boldsymbol{\beta}$ together with missing pattern $\boldsymbol{Q}$ to determine which method is better.

3.1 Asymptotic variance of estimating an individual coefficient

Comparing asymptotic covariance matrices of all coefficients is rather complicated. We can gain insights by focusing on comparing the variance of estimating an individual coefficient using either AC or CC method. This is a relevant task in many real applications. For example, in genetics, we often want to test the association of a disease and a genetic locus while adjusting for additional clinical covariates. Here we assume $\boldsymbol{Z}$ follows an elliptical distribution (Owen and Rabinovitch, 1983) and obtain the asymptotic variance of $\hat{\beta}_1$ in both methods without loss of generality. Under the MCAR and elliptical distribution assumption, the asymptotic variance of $\hat{\beta}_1$ is as follows:

$$n \cdot V_{AC,\hat{\beta}_1} = (1+\kappa) \left\{ \sum_{g=1}^{p} c_g \beta_g^2 + \sum_{g=1}^{p} \sum_{h=g+1}^{p} d_{gh} \beta_g \beta_h + \left(\sum_{j=1}^{p} \frac{r_{1j}^2 \sigma_j^2}{q_j} + \sum_{j=1}^{p} \sum_{k=j+1}^{p} 2 r_{1j} r_{1k} \sigma_{jk} \frac{q_{jk}}{q_j q_k} \right) \sigma^2 \right\}, \tag{1}$$

$$n \cdot V_{CC,\hat{\beta}_1} = (1+\kappa) \left(\sum_{j=1}^{p} r_{1j}^2 \sigma_j^2 + \sum_{j=1}^{p} \sum_{k=j+1}^{p} 2 r_{1j} r_{1k} \sigma_{jk} \right) \frac{\sigma^2}{\tilde{q}} = (1+\kappa) \frac{r_{11} \sigma^2}{\tilde{q}}, \tag{2}$$

where c_g, d_{gh} are defined in Appendix C (online supplemental material); κ is a kurtosis parameter that will be introduced in the following remark; r_{jk} is the (j,k)th element in $\boldsymbol{\Sigma}^{-1}$ (e.g., r_{1j} is the jth element in the first row of $\boldsymbol{\Sigma}^{-1}$). We also notice when all proportions (i.e., q_j, q_{jk}, etc) are equal, namely there are no mismatched observations, then $c_g = d_{gh} = 0$ and the variance of the AC estimator coincides with that of the CC estimator as expected.

Remark: The reason for assuming an elliptical distribution of $\boldsymbol{Z}$ is to simplify the fourth central moments involved in V_D. A special case is to assume $\boldsymbol{Z}$ follow a multivariate normal distribution for which the fourth central moments can be expressed in terms of its covariance matrix by Isserlis' theorem (Isserlis, 1918). In this article, we adopt a more general assumption that $\boldsymbol{Z}$ follows an elliptically contoured distribution (Owen and Rabinovitch, 1983) that includes not only multivariate normal distribution, but also fatter-tailed distributions such as multivariate t-distribution, multivariate logistic distribution, and thinner-tailed distributions such as sub-Gaussian α-stable distribution. Bentler (1983) introduced a kurtosis parameter κ to link the fourth moments with the covariance matrix:

$$\begin{aligned} &\mathrm{E}(Z_j - \mu_j)(Z_k - \mu_k)(Z_m - \mu_m)(Z_n - \mu_n) \\ &= (1+\kappa)(\sigma_{jk}\sigma_{mn} + \sigma_{jm}\sigma_{kn} + \sigma_{jn}\sigma_{km}) \end{aligned}$$

where $\kappa = \frac{\mathrm{E}(Z_j-\mu_j)^2}{3\{\mathrm{E}(Z_j-\mu_j)^2\}^2} - 1$ is one-third of the excess kurtosis for each marginal r.v Z_j. In our regression setting, κ is always larger than $-1/2$ (Bentler and Berkane, 1986). For normal distribution, $\kappa = 0$. There are several ways to estimate the common kurtosis parameter from the data (See Appendix D (online supplemental material)).

3.2 Comparison of $\mathrm{Var}(\hat{\beta}_1)$ under special missing patterns

As we can see from expressions (1), (2), a very general missing pattern results in a complex formula. In this section, we assume X_2 to X_p follow the same missing pattern and explore the asymptotic variance of $\hat{\beta}_1$ in both methods. As shown in Figure 1, we focus on two missing patterns. The pattern (a) is a *unit monotone missing pattern* and the pattern (b) is a *univariate missing pattern* when predictors X_2 to X_p are complete (Little, 1992).

3.2.1 Missing pattern (a)

Consider the unit monotone missing pattern (a) shown in Figure 1(a). Let q_1 denote the observed proportion of X_1; q_{-1} be the observed proportion of X_j $(j \geq 2)$. In addition,

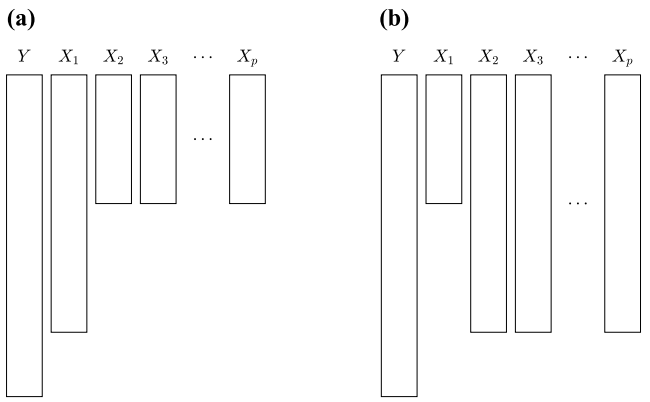

Figure 1. Illustration of two missing patterns.

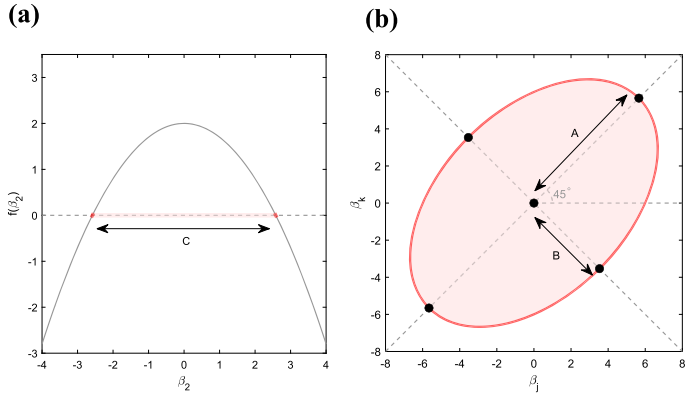

Figure 2. (a) $f(\beta_2)$ in scenario 1 ($p = 2$); (b) the projection of $f(\boldsymbol{\beta}) = 0$ in scenario 2 ($p \geq 3$).

we assume available samples in X_2 to X_p are a subset of X_1 ($q_1 > q_{-1}$) to have the monotone missing.

According to expressions (1) and (2), we obtain the asymptotic variance of $\hat{\beta}_1$ in both methods and calculate the difference $V_{D,\hat{\beta}_1}$. Let $f(\boldsymbol{\beta}) \triangleq n \cdot V_{D,\hat{\beta}_1}$ denote the difference of asymptotic variance as a function of $\boldsymbol{\beta}$:

$$f(\boldsymbol{\beta}) = -r_{11}^2\left(\frac{1}{q_{-1}} - \frac{1}{q_1}\right)\sum_{g=2}^{p}(\sigma_{1g}^2 + 2\kappa\sigma_{1g}^2 + \sigma_1^2\sigma_g^2 + \kappa\sigma_1^2\sigma_g^2)\beta_g^2$$
$$- 2r_{11}^2\left(\frac{1}{q_{-1}} - \frac{1}{q_1}\right)\sum_{g=2}^{p}\sum_{h=g+1}^{p}(\sigma_{1g}\sigma_{1h} + 2\kappa\sigma_{1g}\sigma_{1h} + \sigma_1^2\sigma_{gh}$$
$$+ \kappa\sigma_1^2\sigma_{gh})\beta_g\beta_h + (1+\kappa)\left(\frac{1}{q_{-1}} - \frac{1}{q_1}\right)\left(2r_{11} - r_{11}^2\sigma_1^2\right)\sigma^2.$$

The true coefficient β_1 is not involved in this expression. If $f(\boldsymbol{\beta}) > 0$, then AC estimator is better.

We find that $\left(\frac{1}{q_{-1}} - \frac{1}{q_1}\right)$ is a key quantity in $f(\boldsymbol{\beta})$. No matter which method is better, when we fix all other parameters, the larger the difference between $1/q_{-1}$ and $1/q_1$, the larger the difference between the two methods.

A special case is that all predictors are independent:

$$f(\boldsymbol{\beta}) = \frac{1}{\sigma_1^2}(1+\kappa)\left(\frac{1}{q_{-1}} - \frac{1}{q_1}\right)\left(\sigma^2 - \sum_{g=2}^{p}\sigma_g^2\beta_g^2\right).$$

AC estimator is better when $f(\boldsymbol{\beta}) > 0$, so we have the following proposition:

Proposition 3.1. In missing pattern (a), assuming all predictors are independent, the AC estimator is asymptotically better if and only if:

$$\sum_{g=2}^{p}\sigma_g^2\beta_g^2 < \sigma^2.$$

We can rewrite the inequality as $\sum_{g=2}^{p}\left(\frac{\sigma_g\beta_g}{\sigma}\right)^2 < 1$, which means when the sum of squares of the standardized coefficients (except for X_1) are less than 1, the AC estimator of β_1 is better.

For the general case that predictors are not independent, we further discuss the behavior of $f(\boldsymbol{\beta})$ under two scenarios where $p = 2$ or $p \geq 3$.

Scenario 1, p = 2:
In this scenario, we only have predictors X_1, X_2 in our model. Then $f(\boldsymbol{\beta})$ is simplified as:

$$f(\beta_2) = -r_{11}^2(\sigma_{12}^2 + 2\kappa\sigma_{12}^2 + \sigma_1^2\sigma_2^2 + \kappa\sigma_1^2\sigma_2^2)\left(\frac{1}{q_{-1}} - \frac{1}{q_1}\right)\beta_2^2$$
$$+ (1+\kappa)\left(\frac{1}{q_{-1}} - \frac{1}{q_1}\right)\left(2r_{11} - r_{11}^2\sigma_1^2\right)\sigma^2.$$

It is obvious that when the constant term (that does not involve β_2) is negative, $f(\beta_2)$ is always less than 0 (i.e., CC is better). Therefore, we have the following proposition:

Proposition 3.2 (See Appendix E (online supplemental material) for proof)**.** In missing pattern (a) with two predictors, a sufficient condition that the CC estimator of β_1 is asymptotically better than AC is:

$$\frac{\sigma_1^2\sigma_2^2}{\sigma_{12}^2} < 2.$$

This proposition shows that if the correlation between two predictors is strong (i.e., $|\rho_{12}| > \frac{\sqrt{2}}{2}$), the AC estimator is always worse.

When $\frac{\sigma_1^2\sigma_2^2}{\sigma_{12}^2} \geq 2$ (i.e., $|\rho_{12}| \leq \frac{\sqrt{2}}{2}$), the AC estimator has the possibility to be better than CC only if β_2 is not too far from 0. In Figure 2(a), we plot function $f(\beta_2)$ and find that $f > 0$ iff β_2 lies in the interval between two intersections (the pink interval). This interval is symmetric around 0 and

Table 1. How C changes with parameters

Parameter	Segment 1	C	Segment 2	C
κ	$(-1/2, +\infty)$	$\searrow$		
σ_{12}	$(-\sqrt{\sigma_1^2\sigma_2^2/2}, 0)$	$\nearrow$	$(0, \sqrt{\sigma_1^2\sigma_2^2/2})$	$\searrow$
(ρ_{12})	$(-\sqrt{2}/2, 0)$	$\nearrow$	$(0, \sqrt{2}/2)$	$\searrow$
σ_2^2	$(2\sigma_{12}^2/\sigma_1^2, M_0$[a]$)$	$\nearrow$	$(M_0, +\infty)$	$\searrow$
σ_1^2	$(2\sigma_{12}^2/\sigma_2^2, +\infty)$	$\nearrow$		
σ^2	$(0, +\infty)$	$\nearrow$		

[a] $M_0 = \left(2 + \sqrt{8 - 2/(1+\kappa)}\right)\sigma_{12}^2/\sigma_1^2$.

we denote its length as C:

$$C = \sqrt{\frac{4\left(\sigma_1^2\sigma_2^2 - 2\sigma_{12}^2\right)\sigma^2}{\left(\frac{1+2\kappa}{1+\kappa}\sigma_{12}^2 + \sigma_1^2\sigma_2^2\right)\sigma_2^2}}.$$

We list how C changes with different parameters in Table 1. When the kurtosis parameter κ increases, the interval length C decreases, which means a heavy-tailed dataset favors CC method. For the covariance structure, we find that a larger σ_1^2, σ^2 and a smaller $|\sigma_{12}|$ favors the AC estimator, but the effect of σ_2^2 is not monotone when fixing other parameters. In other words, decreasing the variance of X_1 or the residual, and increasing the correlation between X_1, X_2 make the AC estimator of β_1 a worse estimator.

Scenario 2, p ≥ 3:

In this scenario, we assume that $X_j\,(j \geq 2)$ are homoscedastic and have an exchangeable covariance structure. Their correlation with X_1 is exchangeable as well. Specifically, we assume that the variance of X_1 is σ_1^2; the variance of $X_j\,(j \geq 2)$ is $\sigma_{2'}^2$; the covariance between X_1 and $X_j\,(j \geq 2)$ is $\sigma_{12'}$; and the covariance between $X_j\,(j \geq 2)$ and $X_k\,(k \geq 2, k \neq j)$ is $\sigma_{2'3'}$. Then $f(\boldsymbol{\beta})$ is simplified as:

$$\begin{aligned} f(\boldsymbol{\beta}) = &-r_{11}^2(\sigma_{12'}^2 + 2\kappa\sigma_{12'}^2 + \sigma_1^2\sigma_{2'}^2 + \kappa\sigma_1^2\sigma_{2'}^2)\\ &\left(\frac{1}{q_{-1}} - \frac{1}{q_1}\right)\sum_{g=2}^{p}\beta_g^2 - 2r_{11}^2(\sigma_{12'}^2 + 2\kappa\sigma_{12'}^2 + \sigma_1^2\sigma_{2'3'}\\ &+ \kappa\sigma_1^2\sigma_{2'3'})\left(\frac{1}{q_{-1}} - \frac{1}{q_1}\right)\sum_{g=2}^{p}\sum_{h=g+1}^{p}\beta_g\beta_h\\ &+ (1+\kappa)\left(\frac{1}{q_{-1}} - \frac{1}{q_1}\right)\left(2r_{11} - r_{11}^2\sigma_1^2\right)\sigma^2. \end{aligned}$$

We find that $f(\boldsymbol{\beta})$ is an elliptic paraboloid $\mathbb{R}^p$. When the constant term (that does not involve $\boldsymbol{\beta}$) in $f(\boldsymbol{\beta})$ is negative, $f(\boldsymbol{\beta})$ is always negative (See Appendix F (online supplemental material) for proof). So we have the following proposition:

Proposition 3.3 (See Appendix E (online supplemental material) for proof)**.** In missing pattern (a) with all assumptions above, a sufficient condition that CC estimator of β_1 is asymptotically better than AC is:

$$\{(p-2)\sigma_{2'3'} + \sigma_{2'}^2\}\sigma_1^2 < 2(p-1)\sigma_{12'}^2.$$

As $p \to \infty$, this condition becomes:

$$\frac{\sigma_{2'3'}\sigma_1^2}{\sigma_{12'}^2} < 2.$$

The condition $\sigma_{2'3'}\sigma_1^2 < 2\sigma_{12'}^2$ is equivalent to $\rho_{2'3'} < 2\rho_{12'}^2$, where $\rho_{12'}$, $\rho_{2'3'}$ is the correlation between X_1, $X_j\,(j \geq 2)$, and $X_j\,(j \geq 2)$, $X_k\,(k > 1, k \neq j)$ respectively. This proposition shows that in a high dimensional dataset (p is large) with missing pattern (a), if the correlation between X_1 and $X_j\,(j \geq 2)$ is too strong ($|\rho_{12'}| > \sqrt{\frac{|\rho_{2'3'}|}{2}}$), the AC estimator is always worse.

In Figure 2(b), we plot this ellipse whose center is at the origin and the major axis is rotated 45° around the origin. When point (β_j, β_k) lies in the ellipse (the pink region), then the AC estimator is better than CC. Let A and B denote the length of the semi-major and semi-minor axes:

$$A = \sqrt{\frac{(-2(p-1)\sigma_{12'}^2 + (p-2)\sigma_1^2\sigma_{2'3'} + \sigma_1^2\sigma_{2'}^2)\,\sigma^2}{((p-2)\sigma_{2'3'} + \sigma_{2'}^2)\left(\frac{2+4\kappa}{1+\kappa}\sigma_{12'}^2 + \sigma_1^2\sigma_{2'}^2 + \sigma_1^2\sigma_{2'3'}\right)}},$$

$$B = \sqrt{\frac{(-2(p-1)\sigma_{12'}^2 + (p-2)\sigma_1^2\sigma_{2'3'} + \sigma_1^2\sigma_{2'}^2)\,\sigma^2}{((p-2)\sigma_{2'3'} + \sigma_{2'}^2)\,(\sigma_1^2\sigma_{2'}^2 - \sigma_1^2\sigma_{2'3'})}}.$$

Similar to scenario 1, when $((p-2)\sigma_{2'3'} + \sigma_{2'}^2)\sigma_1^2 \geq 2(p-1)\sigma_{12'}^2$, AC method has a potential to be better than CC. To be more specific, if setting $f(\boldsymbol{\beta}) = 0$, we get an ellipsoid in $\mathbb{R}^{p-1}$ space. This ellipsoid is symmetric around the origin and its projection onto any (β_j, β_k)-plane has the same shape and size. The projection curve on the (β_j, β_k)-plane is an ellipse and described by the following expression:

$$\begin{aligned} &\left(\frac{1+2\kappa}{1+\kappa}\sigma_{12'}^2 + \sigma_1^2\sigma_{2'}^2\right)(\beta_j^2 + \beta_k^2)\\ &\quad + 2\left(\frac{1+2\kappa}{1+\kappa}\sigma_{12'}^2 + \sigma_1^2\sigma_{2'3'}\right)\beta_j\beta_k = \left(\frac{2}{r_{11}} - \sigma_1^2\right)\sigma^2. \end{aligned}$$

We list how A, B change with different parameters in Table 2 and 3. In particular, when the number of predictors p increases, both axes get shorter, resulting in a smaller ellipse that favors CC method. Larger kurtosis parameter κ also shrinks the ellipse, which means a heavy-tailed dataset impairs the performance of AC. In addition, we find that a larger σ_1^2, σ^2 and a smaller $|\sigma_{12'}|$ favor AC estimator. The effect of $\sigma_{2'}^2$, $\sigma_{2'3'}$ is not monotone. Overall, we conclude that a higher correlation between X_1 and other predictors, a lower variance of X_1 or the residual make the AC estimator of β_1 a worse estimator.

Table 2. How A changes with parameters

Parameter	Condition	Segment 1	A	Segment 2	A
p		(I_L, I_R)[a]	$\searrow$		
κ		$(I_L, +\infty)$	$\searrow$		
$\sigma_{12'}$		$(I_L, 0)$	$\nearrow$	$(0, I_R)$	$\searrow$
$\sigma_{2'3'}$	$I_L < M_1$[b] $< I_R$	$(I_L, 0)$	$\nearrow$	$(0, I_R)$	$\searrow$
	$I_R < M_1$	(I_L, I_R)	$\nearrow$		
	$I_L > M_1$	(I_L, I_R)	$\searrow$		
$\sigma^2_{2'}$	con_1[c] > 0 and $I_L < M_2$[b]	(I_L, M_2)	$\nearrow$	$(M_2, +\infty)$	$\searrow$
	$con\_1 < 0$ or $I_L > M_2$	$(I_L, +\infty)$	$\searrow$		
σ^2_1		$(I_L, +\infty)$	$\nearrow$		
σ^2		$(0, +\infty)$	$\nearrow$		

[a] I_L, I_R are the minimum/maximum value for this parameter to take (See Appendix G (online supplemental material)).
[b] The expressions of M_1, M_2 are in Appendix G (online supplemental material).
[c] $con\_1 = (2 + 2p - 2/(1+\kappa))\sigma^2_{12'} + (3-p)\sigma^2_1\sigma_{2'3'}$.

Table 3. How B changes with parameters

Parameter	Condition	Segment 1	B	Segment 2	B
p		(I_L, I_R)[a]	$\searrow$		
κ		$(I_L, +\infty)$	$\rightarrow$		
$\sigma_{12'}$		$(I_L, 0)$	$\nearrow$	$(0, I_R)$	$\searrow$
$\sigma_{2'3'}$	$2\sigma^2_{12'} > \sigma^2_1\sigma^2_{2'}$ and $I_R > M_3$[b]	(I_L, M_3)	$\nearrow$	(M_3, I_R)	$\searrow$
	$2\sigma^2_{12'} < \sigma^2_1\sigma^2_{2'}$ or $I_R < M_3$	(I_L, I_R)	$\nearrow$		
$\sigma^2_{2'}$	$2\sigma^2_{12'} > \sigma^2_1\sigma_{2'3'}$ and $I_L < M_4$[b]	(I_L, M_4)	$\nearrow$	$(M_4, +\infty)$	$\searrow$
	$2\sigma^2_{12'} < \sigma^2_1\sigma_{2'3'}$ or $I_L > M_4$	$(I_L, +\infty)$	$\searrow$		
σ^2_1		$(I_L, +\infty)$	$\nearrow$		
σ^2		$(0, +\infty)$	$\nearrow$		

[a] I_L, I_R are the minimum/maximum value for this parameter to take (See Appendix G (online supplemental material)).
[b] The expressions of M_3, M_4 are in Appendix G (online supplemental material).

3.2.2 Missing pattern (b)

This missing pattern is shown in Figure 1(b). Let q_1 denote the observed proportion of X_1; q_{-1} be the observed proportion of $X_j (j \geq 2)$. In addition, we assume available samples in X_1 are a subset of X_2 to X_p ($q_1 < q_{-1}$). A special case is that only variable X_1 has missing values ($q_{-1} = 1$) which is called *univariate missing*. With expressions (1), (2), we obtain the asymptotic variance of $\hat{\beta}_1$ of two methods and the difference $V_{D,\hat{\beta}_1}$ is as follows:

$$n \cdot V_{D,\hat{\beta}_1} = \left(\frac{1}{q_1} - \frac{1}{q_{-1}}\right)\left\{n(r_{11} - r^2_{11}\sigma^2_1)\sigma^2 - c_1\beta^2_1\right\},$$

where

$$\begin{aligned} c_1 = &\sum_{j=2}^{p} r^2_{1j}\left(\sigma^2_{1j} + \sigma^2_1\sigma^2_j\right) + \kappa r^2_{1j}\sigma^2_1\sigma^2_j + 2\kappa r^2_{1j}\sigma^2_{1j} \\ &+ \sum_{j=2}^{p}\sum_{k=j+1}^{p}\Big\{2r_{1j}r_{1k}\left(\sigma_{1j}\sigma_{1k} + \sigma^2_1\sigma_{jk}\right) \\ &+ 2\kappa r_{1j}r_{1k}\sigma^2_1\sigma_{jk} + 4\kappa r_{1j}r_{1k}\sigma_{1j}\sigma_{1k}\Big\}. \end{aligned}$$

The asymptotic variance of $\hat{\beta}_1$ in CC method is always equal to or smaller than AC method (See Appendix H (online supplemental material) for proof). The quantity $\left(\frac{1}{q_1} - \frac{1}{q_{-1}}\right)$ determines the difference of performance between two methods.

Proposition 3.4. In missing pattern (b), CC estimator of β_1 is asymptotically equal to or better than the AC estimator.

This proposition implies that using extra data from X_2 to X_p does not improve the estimation of β_1 asymptotically. The special case is that when X_1 is independent of other predictors, then both $(r_{11} - r^2_{11}\sigma^2_1) = 0$ and $c_1 = 0$ and thus we have the following proportion:

Proposition 3.5 (See Appendix H (online supplemental material) for proof)**.** In missing pattern (b), AC and CC have the same asymptotic performance if and only if X_1 is independent of other predictors.

3.3 Summary

The main results of Subsection 3.2 are listed in Table 4. In missing pattern (a) that available samples in other predictors are a subset of X_1, the CC estimator of β_1 outperforms

Table 4. Summary table of the comparison in different scenarios.

Missing Pattern	p	Condition	AC	CC
pattern (a)		all predictors are independent, $\sum_{g=2}^{p}\sigma_g^2\beta_g^2 < \sigma^2$	✓	
	2	$\lvert\rho_{12}\rvert > \sqrt{2}/2$		✓
	2	large κ, $\lvert\rho_{12}\rvert$		★
	2	large σ_1^2, σ^2; small $\lvert\beta_2\rvert$	★	
	≥ 3	special covariance[a]; $\rho_{12'}^2 > \rho_{2'3'}/2$[b](large p)		✓
	≥ 3	special covariance; large p, κ, $\lvert\sigma_{12'}\rvert$		★
	≥ 3	special covariance; large σ_1^2, σ^2; small $\lvert\beta_g\rvert$ $(g \geq 2)$	★	
pattern (b)		X_1 is independent of other predictors	same	
		X_1 is not independent of other predictors		✓

[a] $X_j (j \geq 2)$ are homoscedastic and have an exchangeable covariance structure. Their correlation with X_1 is exchangeable as well.
[b] This condition becomes $\{(p-2)\sigma_{2'3'} + \sigma_{2'}^2\}\sigma_1^2 < 2(p-1)\sigma_{12'}^2$ when p is not large enough.
✓ represents the better estimator in this condition; ★ represents this condition flavors the method, but it is not guaranteed to be better.

the AC estimator when the correlation between X_1 and others are large (i.e., $|\rho_{12}| > \sqrt{2}/2$ for two-dimensional predictors; $\rho_{12'}^2 > \rho_{2'3'}/2$ for very high dimensional predictors with special covariance structure). A larger variance of X_1 and residual, and true coefficients β_g $(g \geq 2)$ that is closer to 0 increase the relative performance of AC; while higher dimension of predictors, heavier distribution tails, and larger correlation between X_1 and other predictors flavor the performance CC. In missing pattern (b) that available samples in X_1 are a subset of other predictors, the CC estimator of β_1 is better, except for the scenario that when X_1 is independent of other predictors, both methods have the same asymptotic performance.

4. SIMULATION STUDY

In the introduction part, we mentioned several simulation studies that tried to evaluate AC method. In this section, we take Kim and Curry (1977)'s paper as an example to illustrate the performance of AC method comparing with CC in more details with the help of our theoretical results from the last section.

The simulation studies are based on the correlation matrix on page 196 in Blau and Duncan (1967)'s book. All results from Kim and Curry (1977) showed that AC method is superior to CC. For example, there is a regression analysis of education status. Response variable U is education status and predictors are V (father's education) and X (father's occupational status):

$$U = 0.310V + 0.279X + \varepsilon.$$

The variance of two predictors and the residual is 1 ($\sigma_V^2 = \sigma_X^2 = \sigma_\varepsilon^2 = 1$). The covariance between V, X is 0.516 ($\sigma_{VX} = 0.516$). So we obtain the covariance matrix of random vector $(V, X)^T$ and $(V, X, U)^T$ as follows:

$$\mathbf{\Sigma}_{VX} = \begin{pmatrix} 1 & 0.516 \\ 0.516 & 1 \end{pmatrix}, \mathbf{\Sigma}_{VXU} = \begin{pmatrix} 1 & 0.516 & 0.454 \\ 0.516 & 1 & 0.439 \\ 0.454 & 0.439 & 1.263 \end{pmatrix}.$$

4.1 Finite-sample performance

In the first part of the simulation study, we set up five settings to examine the asymptotic property of our theoretical results.

- **Setting (1)**. $(V, X, U)^T \sim N(\mathbf{0}, \mathbf{\Sigma}_{VXU})$. This setting assumes predictors and errors are normally distributed.
- **Setting (2)**. $(V, X, U)^T \sim t_5(\mathbf{0}, \frac{3}{5}\mathbf{\Sigma}_{VXU})$. In this setting, the covariance matrix of (V, X, U) is $\mathbf{\Sigma}_{VXU}$. The response and predictors follow a multivariate t distribution with a degree of freedom of 5, which meets the elliptic distribution assumption.
- **Setting (3)**. $(V, X)^T \sim$ Multivariate Bernoulli with covariance matrix $\mathbf{\Sigma}_{VX}/6$. The response $U = 0.310V + 0.279X + \varepsilon$ where $\varepsilon \sim N(0, 1)$. This is a typical case where predictors are categorical variables but the error term is normally distribution. This setting violates the elliptic distribution assumption.
- **Setting (4)**. $(V, X)^T \sim$ Multivariate Poisson with covariance matrix $\mathbf{\Sigma}_{VX}$. The response $U = 0.310V + 0.279X + \varepsilon$ where $\varepsilon \sim N(0, 1)$. Another setting that violates the elliptical distribution assumption. The setting is similar to Setting (3), but the predictors follow a multivariate Poisson distribution that has a larger kurtosis than Bernoulli.
- **Setting (5)**. $(V, X, U)^T \sim$ Multivariate Poisson with covariance matrix $\mathbf{\Sigma}_{VXU}$. Comparing with Setting (4), the error does not follow a normal distribution. This also violates the elliptical distribution assumption.

In each setting, each predictor has 10% of random missing cases. Without loss of generality, we focus on the variance of the coefficient estimator for predictor V (i.e., $\hat{\beta}_V$). We calculate the variance of $\hat{\beta}_V$ with 10,000 estimated $\hat{\beta}_V$ in AC method and repeat the simulations 100 times to obtain the standard deviation. The sample size varies from 50 to 250. The theoretical kurtosis parameter κ is estimated from the samples using the second approach in Appendix D (online supplemental material).

In Figure 3, we plot the theoretical results in solid lines and simulated results in dashed lines. In the first two settings

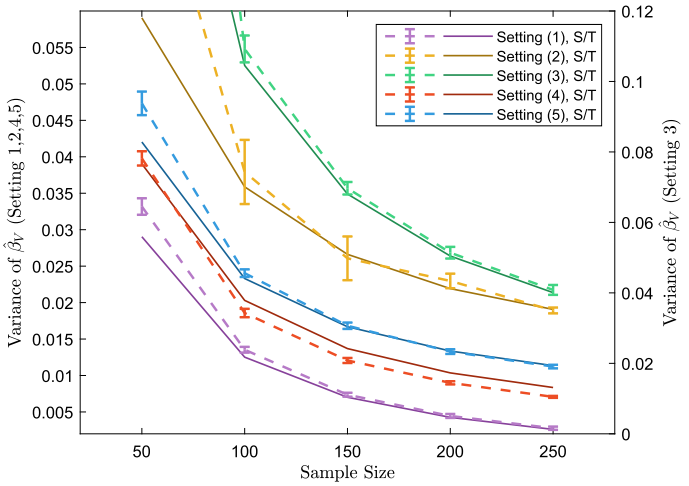

Figure 3. Asymptotic performance of the theoretical results. "S/T" stand for simulated/theoretical result respectively.

where the elliptical distribution assumption holds, the theoretical variance converges to the simulated variance quickly and can be used to represent the true variance accurately when sample size is larger than 150. For the rest of settings, the theoretical result slightly overestimates the variance in Setting (4) and performs well in Setting (3) and (5). The main reason is that the kurtosis of response U is quite different from that of V, X in setting (4). The kurtosis parameter κ is estimated as one third of the mean excess kurtosis of each variable and thus its value is not accurate in setting (4), which results in a worse convergence property. Overall, we conclude that it is safe to utilize the theoretical asymptotic variance to analyze the true estimator variance when the elliptical distribution assumption holds and sample size is not too small. When the assumption severely violates, the derived asymptotic variance may be inaccurate in some cases, especially when the kurtosis of each variable varies much.

4.2 Comparison of AC and CC

In the second part of the simulation study, we still focus on the variance of $\hat{\beta}_V$. Here, we try to explore how the performance of AC changes with different model parameters when comparing with the CC estimator. In addition, we also include FIML and MI into the comparison. The MI uses the default 'predictive mean matching' (pmm) imputation method in the `MICE` package in R.

The response and predictors are simulated from a multivariate-normal distribution as in Setting (1) with 1000 ($n = 1000$) samples.

In Table 5 row 1, we reproduce the result of the Kim and Curry (1977)'s setting where they deleted 10% of the cases from both predictors randomly and obtain their finding that AC method has the smaller variance. Then we explore the simpler settings that only one predictor has missing values. In row 2–3, only V has missings and as we discussed in Subsection 3.2.2, AC method will not improve the efficiency of $\hat{\beta}_V$ by using extra data from X. Therefore, CC method is always better in this missing pattern. The last two rows are the settings where X has missings. AC method even has larger advantages than Kim and Curry (1977)'s setting. In addition, we observe that no matter which method is better, the performance difference gets larger when the missing proportion increases, which is consistent with the finding that the inverse of the observation proportion serves as a scalar in V_D in Subsection 3.2.1, Subsection 3.2.2.

Table 5. The MSE of $\hat{\beta}_V$ in different missing patterns

q_V	q_X	q_{VX}	$CC^{(s)a}$	$CC^{(t)a}$	$AC^{(s)a}$	$AC^{(t)a}$	FIML[a]	MI[a]
0.9	0.9	0.81	1.682	1.683	**1.631**	**1.652**	1.544	1.585
0.9	1	0.9	**1.495**	**1.514**	1.554	1.576	1.474	1.500
0.8	1	0.8	**1.704**	**1.704**	1.841	1.842	1.671	1.716
1	0.9	0.9	1.495	1.514	**1.431**	**1.438**	1.415	1.423
1	0.8	0.8	1.704	1.704	**1.529**	**1.532**	1.481	1.499

[a](s) means simulation results; (t) means theoretical results using expressions (1), (2) in Section 3.1. All the variances are in the order of magnitude of −3.
Note: the smaller MSE in AC or CC is in boldface; the smallest MSE in all methods is in red.

We further investigate the influence of different parameters on the performance of different methods. We focus on the setting with complete V and incomplete X with 10% missing (i.e., row 4 in Table 5). The results are presented in Appendix Table 1 to 4 (online supplemental material) and the relative efficiency of $\hat{\beta}_V$ between CC, AC is shown in Figure 4. In Appendix Table 1 (online supplemental material), we change the correlation between V, X in different ways. We fix the variance of V and X in the first part of the table, and it turns out that CC method will outperform AC when the covariance increases (yellow line in Figure 4(a)). Some articles misinterpret this result and claim that it is always better to use AC method when the correlation between predictors is small (Glasser, 1964; Kim and Curry, 1977). The counterexample is in the second part of the table, where we increase the correlation by decreasing the variance of X. As we point out in Subsection 3.2.1, the effect of σ_X^2 is not monotone, so that we see CC method beats AC in both low and high correlation situations (red line in Figure 4(a)). Lastly, we increase the correlation by decreasing the variance of V in the third part. As expected, the effect of σ_V^2 is monotone. A smaller variance of V flavors CC method (blue line in Figure 4(a)).

The effect of σ_ε^2 is simple. As shown in Appendix Table 2 (online supplemental material) and Figure 4(b), a smaller variance of the residual makes CC method more advantageous.

The most severe problem of AC method is showed in Appendix Table 3 (online supplemental material). We use different true coefficients to compare two methods. From the theoretical results (expressions (1), (2)), we know that the

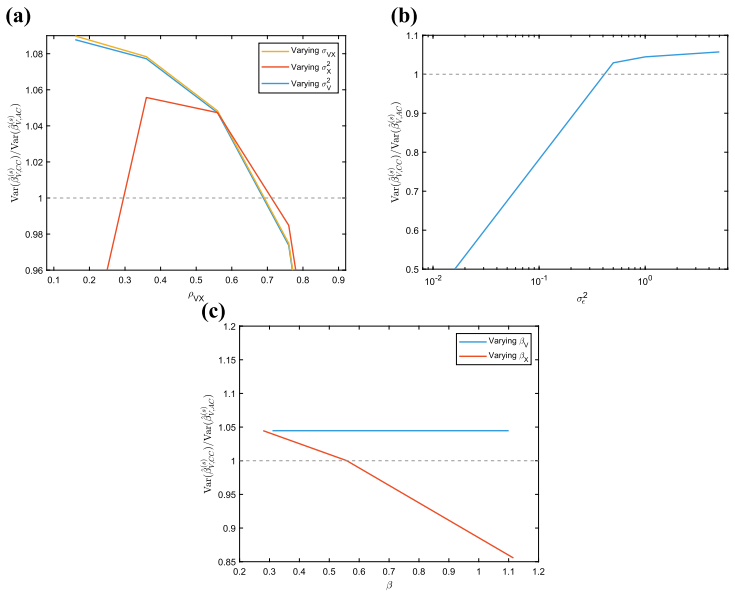

Figure 4. The relative efficiency of $\hat{\beta}_V$ between CC, AC. Single parameter changes each time. (a) Different correlations of V, X. (b) Different residual variances. (c) Different coefficients.

asymptotic variance of the coefficients is not related to the true coefficients in CC method, but that will change in AC method. In this setting, the variance of estimated $\hat{\beta}_V$ in AC method increases with true β_X increases and thus relative efficiency decreases (red line in Figure 4(c)). Therefore, AC method will be inferior when the effect size of other predictors is large. Especially when there are several predictors, AC method is worse as long as any of the other coefficients is large.

In Appendix Table 4, we increase the dimension of the simulated data by adding more covariates Z_i that has the same variance, missing pattern and association with response U as covariate X. The covariance between Z_i and V, X is 0.516 in covariance structure 1 and covariance between Z_i and V, X is 0 in covariance structure 2. We can see in both settings, the AC estimator loses its power. Especially in covariance structure 1, the AC estimator becomes very unstable when the dataset dimension reaches to 50.

Finally, comparing with CC or AC, modern techniques such as FIML and MI have a similar or better performance in almost all cases. This finding is in our expectation since these modern techniques already showed their superiority to CC in literatures (Rubin, 1987; Schenker et al., 1988; Rubin, 1996; Anderson, 1957; Rigdon, 1998) and AC can only be worse or slightly better than CC.

5. DISCUSSION

Since both the AC estimator and the CC estimator are asymptotically unbiased under the MCAR assumption, the comparison of their asymptotic variance helps us determine which one is better. The missing pattern, data covariance structure, true coefficients, etc together influence the performance comparison and their effects on the estimator efficiency under two special missing patterns are summarized in Subsection 3.3. Generally speaking, the AC estimator may have advantages when the predictors are not strongly correlated while variances of other predictors are small. A heavier tailed distribution, higher predictor dimension and larger true predictor coefficients flavor the CC estimator.

We find that the comparison outcome is quite complicated and hope to have some more intuitive explanations. Tarpey et al. (2014) pointed out that the least-squares estimator $\hat{\boldsymbol{\beta}} = \boldsymbol{S}_x^{-1}\boldsymbol{S}_{xy}$ enjoys special superiority because it is a ratio estimator that exploits the dependence between $\boldsymbol{S}_x$ and $\boldsymbol{S}_{xy}$, so that even the estimator $\tilde{\boldsymbol{\beta}} = \boldsymbol{\Sigma}_x^{-1}\boldsymbol{S}_{xy}$ that uses the true $\boldsymbol{\Sigma}_x$ to replace $\boldsymbol{S}_x$ is inferior to the least-squares estimate $\hat{\boldsymbol{\beta}}$. Therefore, though the AC estimator utilizes more data than CC method, it is not always better because it may impair the benefits from utilizing dependence between $\boldsymbol{S}_x$ and $\boldsymbol{S}_{xy}$. This argument provides a potential perspective on the intuitive explanation and needs to be formalized in the further work.

Based on these results, we believe that AC is not a promising method and the effort in missing field should be devoted to making modern methods such as FIML, MI more accessible and user-friendly. For example, researches on GWAS (genome-wide association studies) show that a proper imputation method could greatly increase the study power even for extremely low-coverage sequencing, which obviously far outperforms AC or CC method (Pasaniuc et al., 2012; Höglund et al., 2019).

One future possible research direction is to extend the work to more scenarios. We focus on the comparison un-

der MCAR assumption, but this assumption is often regarded to be very strong and unrealistic (Little et al., 2012). The performance of deletion-based methods will be different under other more general missing patterns. We can also further relax the elliptical assumption or compare several coefficients at the same time. These works will help us to gain insight into the whole picture and lead to more elegant solutions. Besides, the AC estimator has many improved versions (Yu et al., 2020; Lorenzo-Seva and Ferrando, 2020) that possibly have better performance than the original one. Future researches can evaluate these variants and further optimize the method.

ACKNOWLEDGEMENTS

We thank Dr. Rod Little for helpful comments. Gen Li's work was partially supported by National Institutes of Health [grant number R01HG010731]; Kun Chen's work is partially supported by National Science Foundation, Alexandria, VA [grant number IIS-1718798].

SUPPLEMENTARY MATERIALS

Proof and Simulation: It mainly contains the proof of some propositions in Section 3, detailed simulation results in Section 4. (http://www.intlpress.com/site/pub/files/_supp/sii/2022/0015/0004/sii-2022-0015-0004-s002.pdf)

Received 26 June 2021

REFERENCES

Allison, P. D. (2001). *Missing Data* **136**. Sage publications.

Anderson, T. W. (1957). Maximum likelihood estimates for a multivariate normal distribution when some observations are missing. *Journal of the American Statistical Association* **52** 200–203. MR0087286

Bentler, P. M. (1983). Some Contributions to Efficient Statistics in Structural Models: Specification and Estimation of Moment Structures. *Psychometrika* **48** 493–517. MR0731206

Bentler, P. M. (1985). Efficient Estimation via Linearization in Structural Models. *Multivariate Analysis* 9–42. MR0822286

Bentler, P. M. and Berkane, M. (1986). Greatest Lower Bound to the Elliptical Theory Kurtosis Parameter. *Biometrika* **73** 240–241. MR0836456

Blau, P. M. and Duncan, O. D. (1967). *The American Occupational Structure*. Education Resources Information Center.

Dong, Y. and Peng, C.-Y. J. (2013). Principled Missing Data Methods for Researchers. *Springer Plus* **2** 222.

Enders, C. K. and Bandalos, D. L. (2001). The Relative Performance of Full Information Maximum Likelihood Estimation for Missing Data in Structural Equation Models. *Structural Equation Modeling* **8** 430–457. MR1836195

Gao, Y., Merz, C., Lischeid, G. and Schneider, M. (2018). A review on missing hydrological data processing. *Environmental earth sciences* **77** 1–12.

Glasser, M. (1964). Linear Regression Analysis with Missing Observations Among the Independent Variables. *Journal of the American Statistical Association* **59** 834–844. MR0165628

Haitovsky, Y. (1968). Missing Data in Regression Analysis. *Journal of the Royal Statistical Society: Series B (Methodological)* **30** 67–82.

Höglund, J., Rafati, N., Rask-Andersen, M., Enroth, S., Karlsson, T., Ek, W. E. and Johansson, Å. (2019). Improved power and precision with whole genome sequencing data in genome-wide association studies of inflammatory biomarkers. *Scientific reports* **9** 1–14.

Isserlis, L. (1918). On a Formula for the Product-moment Coefficient of any Order of a Normal Frequency Distribution in any Number of Variables. *Biometrika* **12** 134–139.

Kim, J.-O. and Curry, J. (1977). The Treatment of Missing Data in Multivariate Analysis. *Sociological Methods & Research* **6** 215–240.

Lang, K. M. and Little, T. D. (2018). Principled Missing Data Treatments. *Prevention Science* **19** 284–294.

Little, R. J. (1982). Models for Nonresponse in Sample Surveys. *Journal of the American Statistical Association* **77** 237–250. MR0664675

Little, R. J. (1992). Regression with Missing X's: a Review. *Journal of the American Statistical Association* **87** 1227–1237.

Little, R. J. and Rubin, D. B. (1989). The Analysis of Social Science Data with Missing Values. *Sociological Methods & Research* **18** 292–326.

Little, R. J. and Rubin, D. B. (2019). *Statistical Analysis with Missing Data* **793**. John Wiley & Sons. MR1925014

Little, R. J., D'Agostino, R., Cohen, M. L., Dickersin, K., Emerson, S. S., Farrar, J. T., Frangakis, C., Hogan, J. W., Molenberghs, G., Murphy, S. A. et al. (2012). The prevention and treatment of missing data in clinical trials. *New England Journal of Medicine* **367** 1355–1360.

Lorenzo-Seva, U. and Ferrando, P. J. (2020). Not Positive Definite Correlation Matrices in Exploratory Item Factor Analysis: Causes, Consequences and a Proposed Solution. *Structural Equation Modeling: A Multidisciplinary Journal* 1–10. MR4229710

Olinsky, A., Chen, S. and Harlow, L. (2003). The Comparative Efficacy of Imputation Methods for Missing Data in Structural Equation Modeling. *European Journal of Operational Research* **151** 53–79. MR1997008

Owen, J. and Rabinovitch, R. (1983). On the Class of Elliptical Distributions and Their Applications to the Theory of Portfolio Choice. *The Journal of Finance* **38** 745–752.

Pasaniuc, B., Rohland, N., McLaren, P. J., Garimella, K., Zaitlen, N., Li, H., Gupta, N., Neale, B. M., Daly, M. J., Sklar, P. et al. (2012). Extremely low-coverage sequencing and imputation increases power for genome-wide association studies. *Nature Genetics* **44** 631–635.

Peng, C.-Y. J., Harwell, M., Liou, S.-M. and Ehman, L. H. (2006). Advances in Missing Data Methods and Implications for Educational Research. *Real Data Analysis* 31–78.

Pigott, T. D. (2001). A Review of Methods For Missing Data. *Educational Research and Evaluation* **7** 353–383.

Rao, C. R. (1973). *Linear Statistical Inference and its Applications* **2**. Wiley New York. MR0346957

Rigdon, E. E. (1998). Advanced structural equation modeling: issues and techniques. *Applied Psychological Measurement* **22** 85–87.

Rubin, D. B. (1977). Formalizing Subjective Notions About the Effect of Nonrespondents in Sample Surveys. *Journal of the American Statistical Association* **72** 538–543. MR0445668

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. John Wiley & Sons. MR0899519

Rubin, D. B. (1996). Multiple Imputation After 18+ Years. *Journal of the American Statistical Association* **91** 473–489. MR1294072

Schenker, N., Welsh, A. H. et al. (1988). Asymptotic results for multiple imputation. *Annals of Statistics* **16** 1550–1566. MR0964938

Seaman, S. R. and White, I. R. (2013). Review of Inverse Probability Weighting for Dealing With Missing Data. *Statistical Methods in Medical Research* **22** 278–295. MR3190658

Tarpey, T., Ogden, R. T., Petkova, E. and Christensen, R. (2014). A Paradoxical Result in Estimating Regression Coefficients. *The American Statistician* **68** 271–276. MR3280614

Tsikriktsis, N. (2005). A review of techniques for treating missing data in OM survey research. *Journal of Operations Management* **24** 53–62.

Van Praag, B. M. S. (1981). Model Free Regression. *Economics Letters* **7** 139–144.
Van Praag, B. M. S., Dijkstra, T. and Van Velzen, J. (1985). Least-Squares Theory Based on General Distributional Assumptions with an Application to the Incomplete Observations Problem. *Psychometrika* **50** 25–36. MR0789216
White, H. (1980). Using Least Squares to Approximate Unknown Regression Functions. *International Economic Review* 149–170. MR0572464
Xu, S., Lu, B., Baldea, M., Edgar, T. F., Wojsznis, W., Blevins, T. and Nixon, M. (2015). Data cleaning in the process industries. *Reviews in Chemical Engineering* **31** 453–490.
Xue, F. and Qu, A. (2020). Integrating Multisource Block-Wise Missing Data in Model Selection. *Journal of the American Statistical Association* 1–14. MR4353722
Yu, G., Li, Q., Shen, D. and Liu, Y. (2020). Optimal Sparse Linear Prediction for Block-missing Multi-modality Data Without Imputation. *Journal of the American Statistical Association* **115** 1406–1419. MR4143474

Tianchen Xu
Mailman School of Public Health, Columbia University
New York, NY, 10032, USA
E-mail address: tx2155@columbia.edu

Kun Chen
Department of Statistics, University of Connecticut
Storrs, CT 06269, USA
E-mail address: kun.chen@uconn.edu

Gen Li
School of Public Health, University of Michigan
Ann Arbor, MI 48109, USA
E-mail address: ligen@umich.edu