Supplement to "Valid Causal Inference in Linear Regression From Several Observational Studies None of Which Account for All Confounders"

Hani Doss*
Department of Statistics
University of Florida

Jaewoong Joo
Department of Statistics
University of Florida

Abstract

This document provides supporting material for "Valid Causal Inference in Linear Regression From Several Observational Studies None of Which Account for All Confounders" by Hani Doss and Jaewoong Joo, specifically a proof of Proposition 1, additional simulations to augment those in Section 3.1, and details pertaining to the real-data illustration in Section 3.2.

Throughout this document, sections and figures are labelled with the prefix "S". We do this in order to avoid confusion with the sections and figures of the main paper.

S-1 Proof of Proposition 1

From its definition (see (2.6)), $U = (U_1^\top, \ldots, U_K^\top)^\top$, where $U_k = [E(X^{(k)\top} X^{(k)})]^{-1} E(X^{(k)\top} X)$ (invertibility of $E(X^{(k)\top} X^{(k)})$ follows from invertibility of $E(X^\top X)$). Let A be the $d \times d$ block-diagonal matrix consisting of K blocks, where the k^{th} block is $E(X^{(k)\top} X^{(k)})$. Now A is invertible, so if $\tilde{U} = AU$, then $\text{rank}(\tilde{U}) = \text{rank}(U)$. We have $\tilde{U} = (E(X^{(1)\top} X)^\top, \ldots, E(X^{(K)\top} X)^\top)$. The rows of $\tilde{U}$ span $\mathbb{R}^{p+1}$, because the rows of $E(X^\top X)$ span $\mathbb{R}^{p+1}$ (since $E(X^\top X)$ is invertible), and for each covariate X_j, there is at least one study which includes X_j, so that each row of $E(X^\top X)$ appears as a row of $E(X^{(k)\top} X)$ for some k. So $\text{rank}(\tilde{U}) = p + 1$, which implies that $\text{rank}(U) = p + 1$.

S-2 Additional Simulations to Augment Those in Section 3.1

In Section 2 we established that the coverage probabilities of confidence sets for β are asymptotically correct if we use the true values of E and V, but in practice we must use estimated values $\widehat{E}$ for E and S for V (the two kinds of confidence sets are called "oracle" and "real", respectively).

*Research supported by NSF grant DMS-1854476 and NIH grant 1R01NS121099-01A1

In Section 3.1, we investigated the coverage probabilities of the two kinds of confidence intervals for the components of β, and we showed empirically that the coverage probabilities of the real confidence intervals are slightly lower than those of the oracle confidence intervals. Here, we investigate, separately, the effect of using $\widehat{E}$ instead of E and using S instead of V.

Figure S-1 shows the effect of the size of the external data set on the coverage probability of the bootstrap confidence intervals. The figure shows a drop in coverage probability when we reduce the size from $50{,}000$ to $10{,}000$; the drop is consistent across all 4×8 cases, but is not very significant. Additional experiments not shown here show that the drop is more significant when we go from $50{,}000$ to a few thousand; however, we do not think that this is a problem because the external data set that we need consists of unlabelled data, and such data are typically plentiful.

Figure S-2 shows the effect of using the estimated values of V. The figure shows the coverage probabilities for intervals constructed using estimates of V, and coverage probabilities for intervals using the true values of V. For these two types of intervals, all other configuration parameters, including the size of the external data set (which is $50{,}000$) are identical. The figure strongly suggests that the effect of using estimates of V is minimal: the maximum of the deviations between the two types of intervals over all 4×8 cases is 0.005.

We now consider the effect of the number of studies K, the study sample sizes $n_1, \ldots, n_K$, and the total number of variables p. Section 3.1 gave information on the performance of our methodology for the case $K = 20$; see in particular Figure 1. Figures S-3, S-4 and S-5 are similar, but with K changed to 40, 80 and 10, respectively. For all three cases p is kept at 8. For the case $K = 40$, the pattern regarding which covariates are used is the same as for the case $K = 20$ for the first 20 studies and is repeated for the subsequent 20 studies. The sample sizes are $n_1 = \cdots = n_{10} = 300$, $n_{11} = \cdots = n_{20} = 400$, $n_{21} = \cdots = n_{30} = 500$, and $n_{31} = \cdots = n_{40} = 600$. For the case $K = 80$, the pattern regarding which covariates are used is again the same for the first 20 studies and is repeated three times. The sample sizes are $n_1 = \cdots = n_{20} = 300$, $n_{21} = \cdots = n_{40} = 400$, $n_{41} = \cdots = n_{60} = 500$, and $n_{61} = \cdots = n_{80} = 600$. For the case $K = 10$, the pattern regarding which covariates are used is the pattern for the first 10 studies for the case $K = 20$, and the sample sizes are $n_1 = n_2 = n_3 = 300$, $n_4 = n_5 = n_6 = 400$, $n_7 = n_8 = 500$, and $n_9 = n_{10} = 600$.

Figures S-3 and S-4 do not show any problems emerging as we increase the number of studies. But Figure S-5, panel (c), shows that for $K = 10$, the bootstrap confidence intervals have coverage probabilities that are significantly greater than the nominal value of 0.95, and we now explain this. The basic problem is that the bootstrap estimate of variance does not work well when $(p + 1)/d$ is large. Recall that we obtain the estimate $\hat{\beta}$ via the linear regression model $\bar{Z} = \bar{U}\beta + \epsilon$, where $\bar{U}$ is a $d \times (p + 1)$ matrix. Here, $d = d_1 + \cdots + d_K$, where d_k is the number of covariates (including the intercept) included in study k. To obtain the bootstrap estimate of the variance of $\hat{\beta}$ we proceed as follows. For $b = 1, \ldots, B$, we generate $(\bar{U}^{*(b)}, \bar{Z}^{*(b)})$ by sampling with replacement the d pairs $(\bar{U}_i, \bar{Z}_i)$, $i = 1, \ldots, d$. We compute $\hat{\beta}^{*(b)} = (\bar{U}^{*(b)\top}\bar{U}^{*(b)})^{-1}\bar{U}^{*(b)\top}\bar{Z}^{*(b)}$. And we use $\hat{\beta}^{*(1)}, \ldots, \hat{\beta}^{*(B)}$ to compute the estimate of variance. See the paragraph "Estimation of $\mathbf{\Phi}$" in Section 2.2. This procedure requires that for each b, $\bar{U}^{*(b)}$ has full rank, i.e., $\mathrm{rank}(\bar{U}^{*(b)}) = p + 1$, because we need the matrix $\bar{U}^{*(b)\top}\bar{U}^{*(b)}$ to be invertible.

When $(p+1)/d$ is large, the probability that $\bar{U}^{*(b)}$ is not of full rank is large. To see this, take the case $K = 10$, and consider an extreme case where each study includes only one predictor variable,

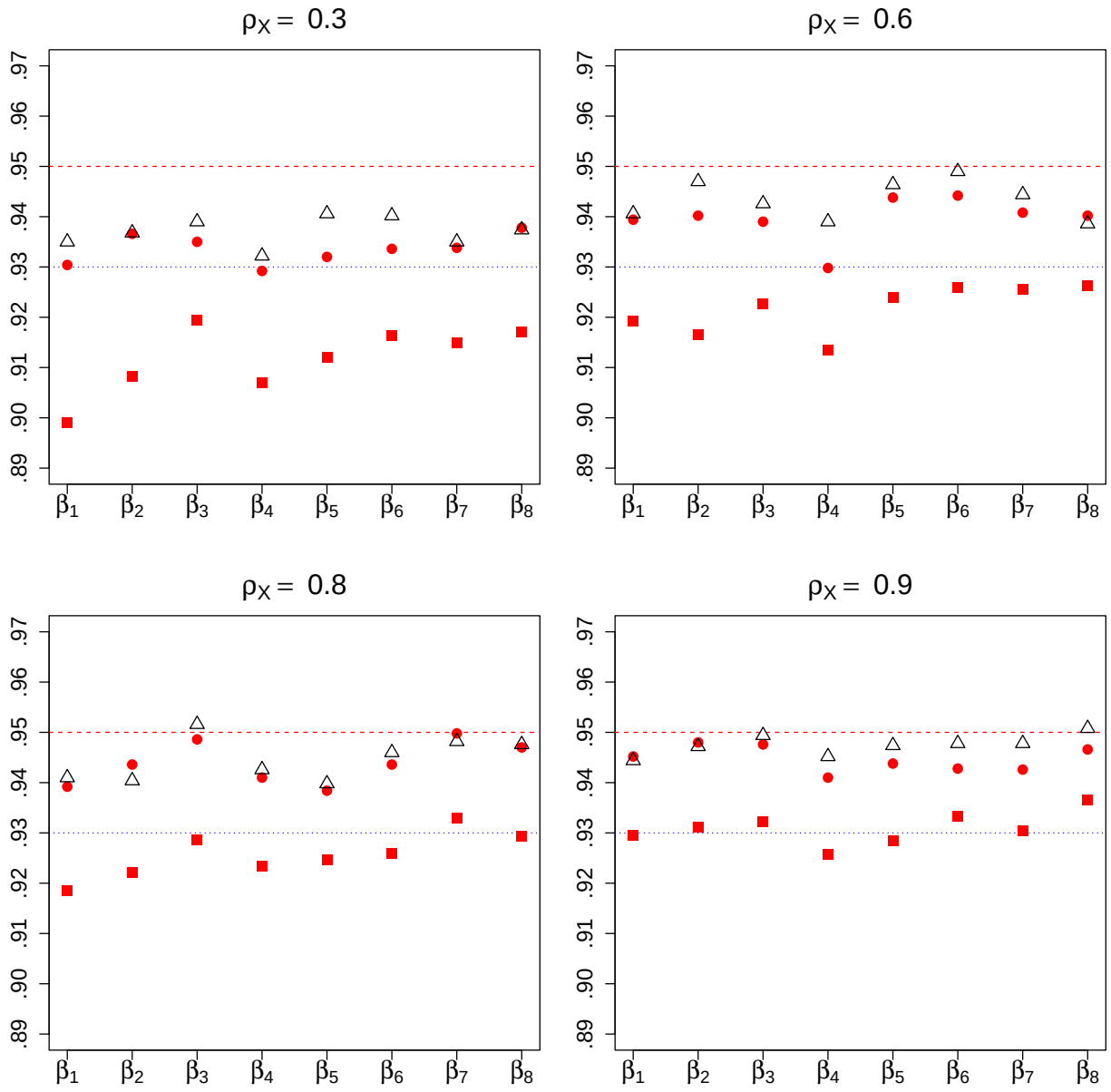


Figure S-1: Effect of size of external data set on coverage probabilities of bootstrap confidence intervals: △ is for oracle CI, ● is for real CI with external data set size equal to 50,000, and ■ is for real CI with external data set size equal to 10,000.

so $d_k = 2$ for all k, and hence $d = 20$; and suppose that $p = 19$. In bootstrap sampling, if any row is selected more than once, then the $d \times (p+1)$ matrix $\bar{U}^{*(b)}$ does not have full rank, so $\bar{U}^{*(b)\top}\bar{U}^{*(b)}$ is singular. In this situation, our algorithm does what bootstrap algorithms commonly do, which is to exclude the sample, and this causes instability in the overall estimate of variance. In the situation above, $(p+1)/d = 1$. However, even when $(p+1)/d < 1$, if $(p+1)/d$ is close to 1, there is a non-negligible probability that the smallest eigenvalue of $\bar{U}^{*(b)\top}\bar{U}^{*(b)}$ is very small, and this causes the bootstrap estimate of variance to be unstable. In our experiments with $K = 80, 40, 20, 10$, the

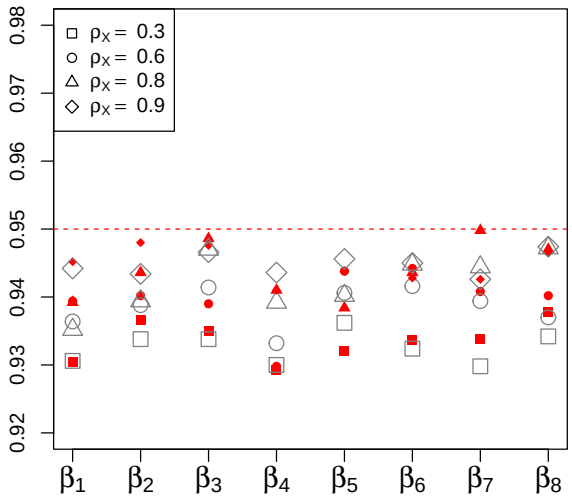


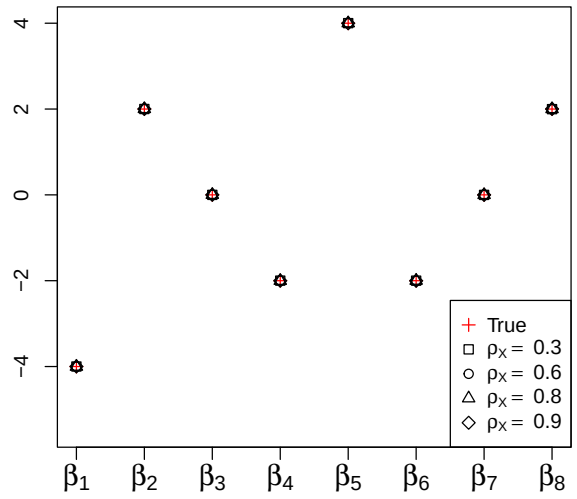
Figure S-2: Effect of using estimated variance matrix, as opposed to using the true variance matrix, on coverage probabilities of the bootstrap confidence intervals. Large empty symbols indicate that true variance matrix is used, and small filled-in symbols (e.g., ▪) indicate that the estimated variance matrix is used.

quantity $(p+1)/d$ is 0.031, 0.062, 0.125, and 0.265, respectively. We carried out an experiment with $K = 10$, and $p = 6$, 5, and 4. The values of $(p+1)/d$ for the three cases are 0.206, 0.200, and 0.167, respectively. Figure S-6 shows the coverage probabilities for the three cases. The figure confirms that the problem diminishes as p decreases (because then $(p+1)/d$ becomes smaller).

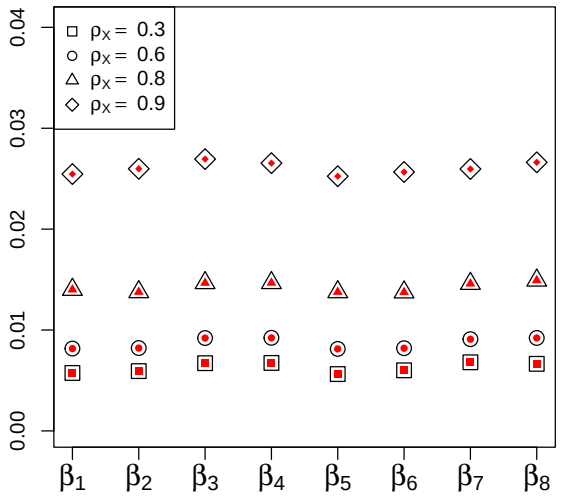
Details for the experiments with $p = 6$, 5, and 4 are as follows. For $p = 6$, the covariates used by studies 1–5 were $(1, X_1, X_2)$, $(1, X_3, X_4)$, $(1, X_5, X_6)$, $(1, X_1, X_6)$, and $(1, X_2, X_3, X_4, X_5)$, respectively. This pattern was repeated for studies 6–10. For $p = 5$, the covariates used by studies 1–10 were $(1, X_1, X_2)$, $(1, X_3)$, $(1, X_4, X_5)$, $(1, X_1)$, $(1, X_2, X_3)$, $(1, X_4, X_5)$, $(1, X_1, X_2, X_3)$, $(1, X_4, X_5)$, $(1, X_1, X_2)$, and $(1, X_3, X_4, X_5)$. For $p = 4$, the covariates used by studies 1–4 were $(1, X_1, X_2)$, $(1, X_3, X_4)$, $(1, X_1, X_4)$, and $(1, X_2, X_3)$, respectively; the covariates used by studies 5–8 followed exactly the same pattern; and the covariates used by studies 9 and 10 were $(1, X_1, X_2)$ and $(1, X_3, X_4)$, respectively.

From extensive experimentation with many values of $(K, d_1, \ldots, d_K, p)$, we have found good performance of the bootstrap estimate of variance (and hence of bootstrap confidence intervals) when $(p+1)/d \leq 0.15$. It is worth pointing out that, even when $(p+1)/d$ is large, our estimate $\hat{\beta}$ continues to be consistent and asymptotically normal, as asserted in Theorem 1, and a problem arises only with the bootstrap estimate of variance.

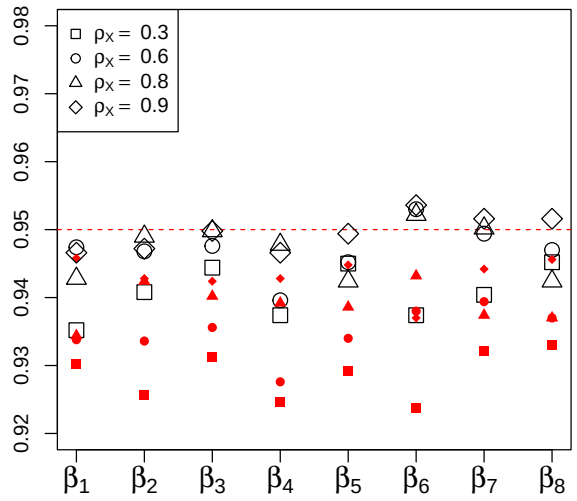
Instability of bootstrap confidence intervals in regression has been noted before. In a context that is different from ours, El Karoui and Purdom (2018) found that when p/n is large, where p is the number of predictors and n is the sample size, bootstrap resampling often fails to preserve the rank of the design matrix, leading to bootstrap confidence intervals that are overly conservative. Our findings are consistent with theirs.



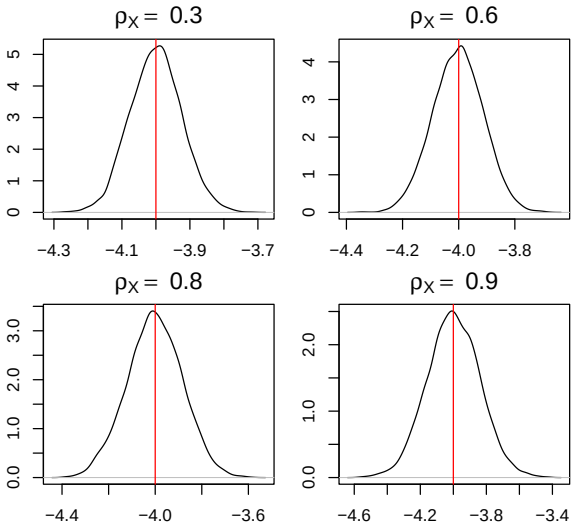
(a) Expected values of estimates of the components of β.



(b) Variances (indicated by small disks [•]) and MSEs (indicated by empty symbols [see legend]) of estimates of the components of β.



(c) Coverage probabilities of oracle (indicated by large empty symbols [see legend]) and real (indicated by small filled-in symbols [e.g., ▪]) 95% CIs.



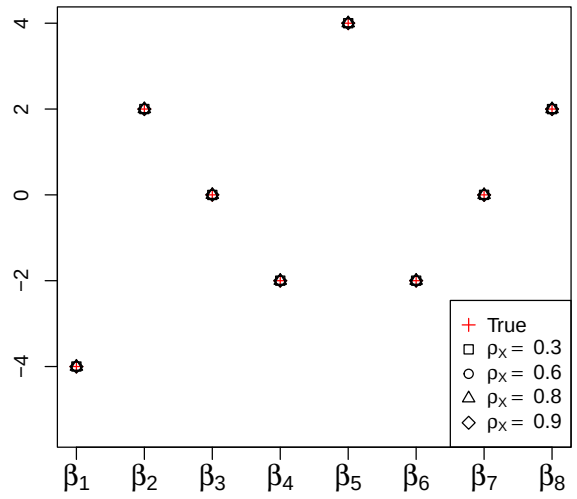
(d) Kernel density estimates of β_1; vertical lines indicate the true value of β_1.

Figure S-3: Aspects of the distribution of $\hat{\beta}$ and coverage probabilities of bootstrap confidence intervals for the case $K = 40$.

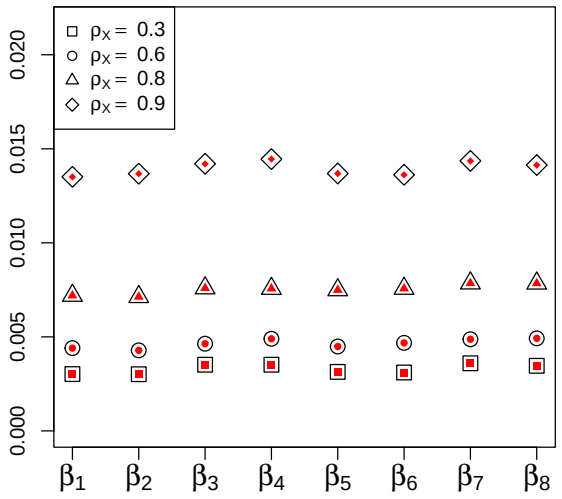
S-3 Some Details Pertaining to Section 3.2

The Korean health checkup dataset includes 38 cases with missing values. Since these cases constitute only a tiny fraction of the entire dataset, we simply excluded them in our analysis.

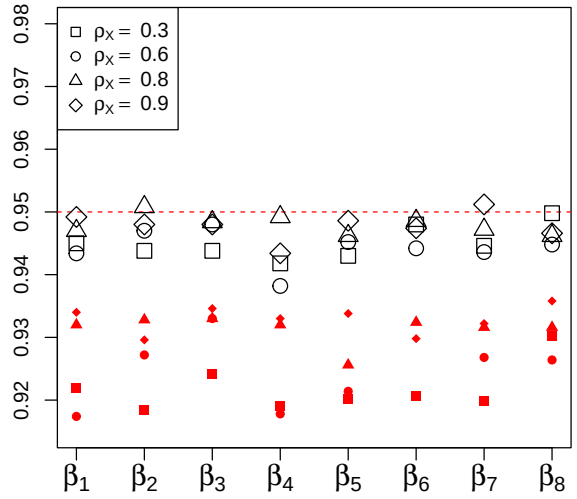
We now discuss two variables used in our analysis, "mean arterial pressure" (MAP) and "total



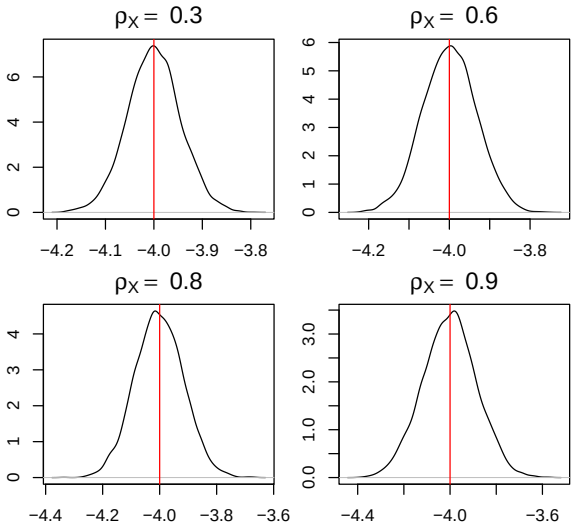
(a) Expected values of estimates of the components of β.



(b) Variances (indicated by small disks [•]) and MSEs (indicated by empty symbols [see legend]) of estimates of the components of β.



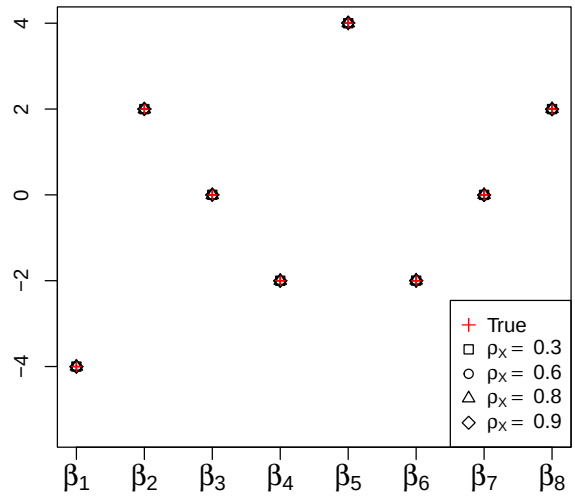
(c) Coverage probabilities of oracle (indicated by large empty symbols [see legend]) and real (indicated by small filled-in symbols [e.g., ▪]) 95% CIs.



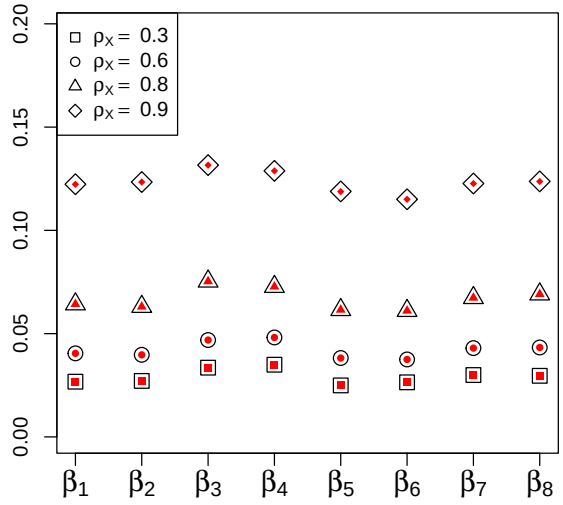
(d) Kernel density estimates of β_1; vertical lines indicate the true value of β_1.

Figure S-4: Aspects of the distribution of $\hat{\beta}$ and coverage probabilities of bootstrap confidence intervals for the case $K = 80$.

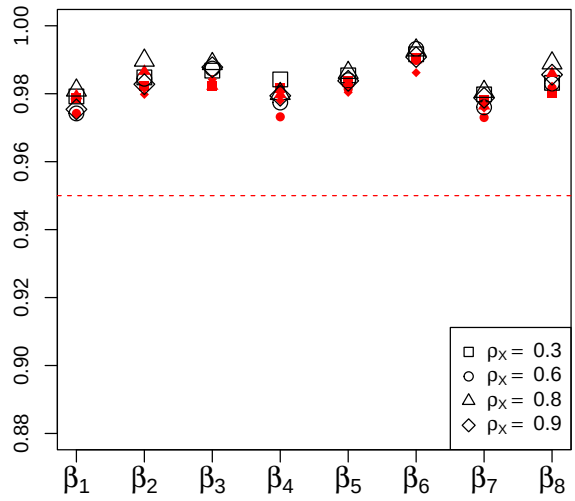
cholesterol" (TC), neither of which is given directly in the dataset. The blood pressure variables in the dataset are systolic pressure (SP) and diastolic pressure (DP). MAP is defined by the weighted average MAP $= (2/3)$DP$+(1/3)$SP. It is frequently used, and is considered a good single number to report blood pressure; see, e.g., Kandil et al. (2023). The lipids variables in the dataset are "high-



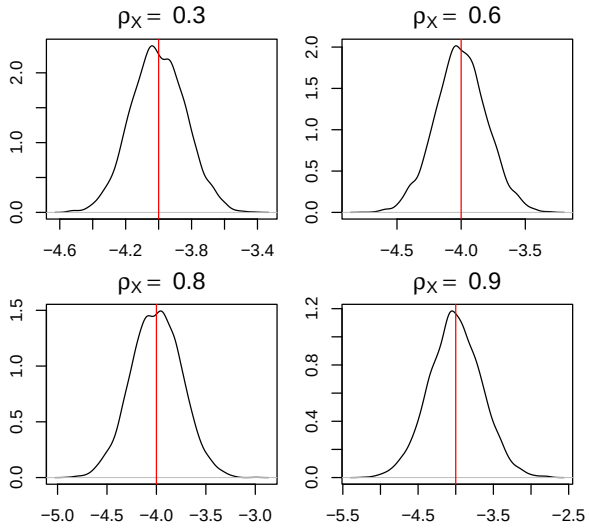
(a) Expected values of estimates of the components of β.



(b) Variances (indicated by small disks [•]) and MSEs (indicated by empty symbols [see legend]) of estimates of the components of β.



(c) Coverage probabilities of oracle (indicated by large empty symbols [see legend]) and real (indicated by small filled-in symbols [e.g., ▪]) 95% CIs.



(d) Kernel density estimates of β_1; vertical lines indicate the true value of β_1.

Figure S-5: Aspects of the distribution of $\hat{\beta}$ and coverage probabilities of bootstrap confidence intervals for the case $K = 10$.

density lipoprotein" (HDL), "low-density lipoprotein" (LDL), and "triglyceride". Total cholesterol is defined by TC $=$ HDL $+$ LDL $+$ $(.2 \times$ triglyceride$)$.

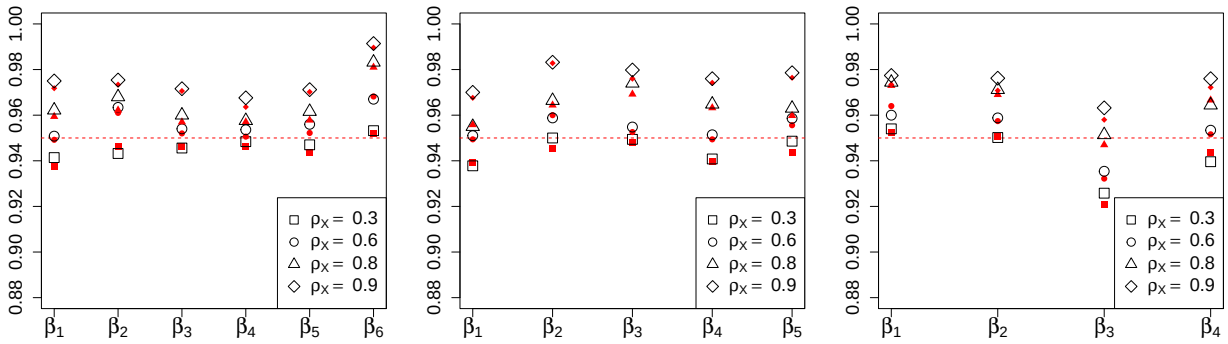


Figure S-6: Coverage probabilities of oracle (indicated by large empty symbols [see legend]) and real (indicated by small filled-in symbols [e.g., ▪]) 95% CIs, for $K = 10$ and $p = 6$ (left), $p = 5$ (center), and $p = 4$ (right).

References

El Karoui, N. and Purdom, E. (2018). Can we trust the bootstrap in high-dimensions? The case of linear models. *Journal of Machine Learning Research* **19** 1–66.

Kandil, H., Soliman, A., Alghamdi, N. S., Jennings, J. R. and El-Baz, A. (2023). Using mean arterial pressure in hypertension diagnosis versus using either systolic or diastolic blood pressure measurements. *Biomedicines* **11** 849.