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

We consider a situation where there are p predictor variables $X_1, \dots, X_p$ and a response variable Y, linked through a linear regression model, and we are interested in the regression coefficients. There are K observational studies, where for each $k = 1, \dots, K$, study k does a linear regression of Y on only a subset of the predictors, and reports point estimates and their estimated standard errors of the regression coefficients in the reduced model. The subset of predictors used varies from study to study. From this information, the regression coefficients in the full model are not identifiable. It is often the case that from external sources we have estimates of the expectations $E(X_i X_j)$, $i, j = 1, \dots, p$. For this case, we construct estimates $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p$ of the regression coefficients in the full model which are consistent and asymptotically normal, and we propose a method for estimating the covariance matrix of $(\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p)$, thus enabling us to form confidence regions or intervals for the β_j's. Our development does not require any parametric assumptions, Gaussian or otherwise, on the error terms, nor that these be homoscedastic. We show good performance of our methodology through extensive simulations and an illustration on real data.

Key words and phrases: causal effect; data integration; meta-analysis; observational studies

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

1 Introduction

The following situation arises frequently in observational studies in epidemiological, medical and social sciences. There is an outcome variable Y and p predictors $X_1, \dots, X_p$, and we wish to assess the relationship between Y and the X's through a linear regression model of the form

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon, \tag{1.1}$$

where ϵ has mean 0 and finite variance. We are interested in the regression coefficients $\beta_0, \beta_1, \dots,$ β_p. There are K independent studies, each of which investigates the relationship between Y and the predictors, but for various reasons, for example the interest focus of the investigators or cost, some studies (possibly all) do not use the complete set of predictors. For example, for study 1 the data might be n_1 iid replicates of the pair $(Y, (X_1, X_2))$, while for study 2 the data could be n_2 iid replicates of the pair $(Y, (X_1, X_3))$. Study 1 then gives an estimate of the regression coefficients for the model $Y = \beta_0^{(1)} + \beta_1^{(1)} X_1 + \beta_2^{(1)} X_2 + \epsilon$, and likewise study 2 gives an estimate of the regression coefficients for the model $Y = \beta_0^{(2)} + \beta_2^{(2)} X_2 + \beta_3^{(2)} X_3 + \epsilon$.

To set notation to handle the general case, let X denote the full (row) vector of predictors, and in our framework of observational studies, X is random. To avoid excessive notation, we take $X = (1, X_1, \dots, X_p)$, i.e. X contains the intercept term. For each $k = 1, \dots, K$, let $X^{(k)}$ denote the subvector of X used by study k, and let $\beta^{(k)}$ denote the vector of regression coefficients corresponding to $X^{(k)}$. Although we allow the possibility that none of the studies use the complete set of predictors, we assume that for each component of $(1, X_1, \dots, X_p)$, there is at least one study that uses that component. For each k, study k gives an estimate $\hat{\beta}^{(k)}$ of $\beta^{(k)}$, together with an estimate S_k of the variance matrix of $\hat{\beta}^{(k)}$.

As is well known, for any study for which $X^{(k)}$ is not the full set of covariates, the true β's for that study are not necessarily equal to the true β's for the full model (1.1), so even if the sample size for study k was huge, $\hat{\beta}^{(k)}$ would not necessarily be close to the corresponding subvector of the full vector $\beta = (\beta_0, \beta_1, \dots, \beta_p)^\top$. As a consequence, to establish a causal relationship between predictor

X_j and the outcome Y, it is necessary to estimate β_j in a full model that includes all potential confounding variables. Unfortunately, as is well known (and is easy to see through simple examples), the parameter β_j in the full model is unidentifiable from the K pairs $(\hat{\beta}^{(k)}, S_k)$, $k = 1, \ldots, K$, and so is the full vector β. However, it turns out that with certain additional information, namely "knowledge" of the matrix $E(X^\top X)$, it is possible to arrive at a consistent and asymptotically normal estimate of the full vector β. Here, "knowledge" means that we have access to an estimate of the matrix that is consistent and asymptotically normal; a more precise statement is given in Section 2.

The problem of estimating the vector of regression coefficients in the full model from summary data of several studies each of which considers only a subset of the predictors was also considered by Kundu et al. (2019), which inspired the present work. Their framework is more general than ours in that they are not limited to a linear model, but rather allow quite general parametric models; these include the logistic and probit models, and also the linear model (1.1) with Gaussian errors. Their approach is based on the generalized method of moments and uses an iterative scheme for solving a system of equations. It requires an estimate of the joint distribution of the full predictor vector X in the form of an iid sample $X^{(1)}, \ldots, X^{(N)}$, where the $X^{(j)}$'s have the same distribution as X. Notwithstanding the generality of their setup, our development has several advantages: (i) We do not make any parametric assumptions at all, and in particular we do not assume that the ϵ's in (1.1) are either Gaussian or homoscedastic (in fact, we do not even assume that $E(Y \mid X = x)$ is a linear function of x, as we explain in Section 2). (ii) At a practical level, we do not need to estimate the joint distribution of X, and all we need is an estimate of the matrix $E(X^\top X)$. The individual entries $E(X_i X_j)$ can be estimated separately from different data sets, so we do not need a single data set containing a large number of replicates of the entire vector X (this is an important point because typically studies measure only a subset of the covariate X). (iii) Additionally, we do not need an iterative scheme to compute our estimate of β; the estimate emerges in closed form through generalized least squares.

Assumptions regarding knowledge of the joint distribution of the vector of covariates X have been made in the causal inference literature before. In Candès et al. (2018), which deals with false discovery rates in high-dimensional regression, the authors assume that the joint distribution of X is known, or at least can be accurately estimated from unlabelled data. Berrett et al. (2020) consider a problem of conditional independence testing in a high-dimensional setting, which is stated as follows. There are univariate variables Y_1 and Y_2 of principal interest, and another variable X, which is possibly high dimensional and may contain confounders. We wish to test conditional independence of Y_1 and Y_2 given X. Berrett et al. (2020) assume that the conditional distribution of Y_1 given X is known, or at least can be accurately estimated. The point here is that in the problems considered in these papers and in the present paper there is an unidentifiability issue, and to address it, it seems inevitable that one will need further information on the vector of covariates.

The rest of this paper is structured as follows. In Section 2 we present our approach for obtaining an estimate of the full vector β and an estimate of its variance matrix, and we provide a result giving theoretical support for the approach. In Section 3 we present simulation studies that confirm our theoretical results regarding the estimates of and confidence sets for β, and we a give an illustration on real data, and in Section 4 we make a few concluding remarks.

2 Valid Estimation of the Vector of Regression Coefficients in the Full Model

Throughout this paper we assume that for each study the model contains an intercept term. In this section we show how from the summary statistics $(\hat{\beta}^{(k)}, S_k),\ k = 1, \dots, K$ and knowledge of $E(X^\top X)$, we can obtain a valid estimate of the full vector β together with a variance estimate. In order to focus on the substance, we first consider the idealized situation in which $E(X^\top X)$ is known exactly, and for $k = 1, \dots, K$ the estimate S_k of $\mathrm{Var}(\hat{\beta}^{(k)})$ is exact.

2.1 An Idealized Situation

We assume that the fourth moments of the components of (Y, X) are all finite and that the matrix $E(X^\top X)$ is positive definite, so is invertible. Define $f\colon \mathbb{R}^{p+1} \to \mathbb{R}$ by $f(b) = E(\|Y - Xb\|^2)$, where $\|\cdot\|$ is Euclidean norm. Via simple calculus, we see that the minimizing value of b is $\beta = [E(X^\top X)]^{-1} E(X^\top Y)$. We may write $Y = X\beta + (Y - X\beta)$, so defining $\epsilon = Y - X\beta$, we have the well-known decomposition

$$Y = X\beta + \epsilon, \tag{2.1}$$

and in the usual L_2 space of random variables having finite second moments and inner product $\langle V, W\rangle = E(VW)$, ϵ is orthogonal to the linear space spanned by $X_0, X_1, \ldots, X_p$ (where $X_0 = 1$), so in particular,

$$E(\epsilon X_j) = 0 \qquad \text{for every } j = 0, 1, \ldots, p. \tag{2.2}$$

Also, ϵ may be correlated with X, so model (2.1) does not correspond to the usual homoscedastic regression model. In fact, the decomposition (2.1) is valid without any assumption that the distribution of (Y, X) is Gaussian, or that $E(Y \mid X = x)$ is a linear function of x. Whether or not $E(Y \mid X = x)$ is a linear function of x, β is well defined as $\arg\min_b E(\|Y - Xb\|^2)$ and provides the best linear relationship between Y and X in the L_2 sense.

We will need the condition that $E(\hat{\beta}^{(k)}) = \beta^{(k)}$ for every k. For the full model, this condition would be $E(\hat{\beta}) = \beta$. From (2.2) with $j = 0$ we get $E(\epsilon) = 0$ (which does not necessarily hold if the model does not include an intercept term). We will now show that when $\hat{\beta}$ is taken to be the least squares estimate, if we strengthen the condition $E(\epsilon) = 0$ to

$$E(\epsilon \mid X) = 0 \qquad \text{almost surely}, \tag{2.3}$$

then indeed $E(\hat{\beta}) = \beta$. Statement (2.3) is certainly true when the joint distribution of (Y, X) is Gaussian, but is true for many other distributions as well. Let (Y_i, X_i), $i = 1, \ldots, n$ be n independent hypothetical copies of (Y, X), and let $\epsilon_i = Y_i - X_i\beta$. Denote $\mathsf{Y} = (Y_1, \ldots, Y_n)^\top$, $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^\top$, and let X denote the $n \times p$ design matrix formed from the X_i's. The least squares estimate of β is

$\hat{\beta} = (\mathsf{X}^\top \mathsf{X})^{-1}\mathsf{X}^\top \mathsf{Y}$, and we have $\hat{\beta} = (\mathsf{X}^\top \mathsf{X})^{-1}\mathsf{X}^\top[\mathsf{X}\beta + \boldsymbol{\epsilon}]$, so

$$E(\hat{\beta}) = \beta + E[(\mathsf{X}^\top \mathsf{X})^{-1}\mathsf{X}^\top \boldsymbol{\epsilon}] = \beta + E\big[(\mathsf{X}^\top \mathsf{X})^{-1}\mathsf{X}^\top E(\boldsymbol{\epsilon} \,|\, \mathsf{X})\big] = \beta$$

(by slight abuse of notation $E(\boldsymbol{\epsilon} \,|\, \mathsf{X})$ is taken to mean $(E(\epsilon_1 \,|\, X_1), \dots, E(\epsilon_n \,|\, X_n))^\top$), where the last equality follows from (2.3). To recapitulate: under (2.3), the least squares estimate satisfies $E(\hat{\beta}) = \beta$.

For $k = 1, \dots, K$, let d_k denote the dimension of $X^{(k)}$, the (row) vector of predictors for study k, and define $f_k \colon \mathbb{R}^{d_k} \to \mathbb{R}$ by $f_k(b) = E(\|Y - X^{(k)}b\|^2)$. We define the true parameter for study k by $\beta^{(k)} = \arg\min_b f_k(b)$ and, as before, we have a closed-form expression for $\beta^{(k)}$, namely $\beta^{(k)} = [E(X^{(k)\top}X^{(k)})]^{-1}E(X^{(k)\top}Y)$. Suppose the data for study k are $(Y_{ki}, X_i^{(k)})$, $i = 1, \dots, n_k$. Let $\mathsf{X}^{(k)}$ denote the observed $n_k \times d_k$ design matrix, for which the rows are $X_i^{(k)}$, $i = 1, \dots, n_k$, and let $\mathsf{Y}_k = (Y_{k1}, \dots, Y_{kn_k})^\top$. We will assume that the estimate of $\beta^{(k)}$ given by study k satisfies $E(\hat{\beta}^{(k)}) = \beta^{(k)}$. As before, this assumption holds for the conventional least squares estimate $\hat{\beta}^{(k)} = (\mathsf{X}^{(k)\top}\mathsf{X}^{(k)})^{-1}\mathsf{X}^{(k)\top}\mathsf{Y}_k$ under the condition that

$$E(\epsilon \,|\, X^{(k)}) = 0 \text{ almost surely, where now } \epsilon \text{ corresponds to (2.1) but for study } k. \tag{2.4}$$

Now

$$\begin{aligned} E(\hat{\beta}^{(k)}) &= [E(X^{(k)\top}X^{(k)})]^{-1}E(X^{(k)\top}Y) && \text{(2.5a)} \\ &= [E(X^{(k)\top}X^{(k)})]^{-1}E(X^{(k)\top}X\beta) + [E(X^{(k)\top}X^{(k)})]^{-1}E(X^{(k)\top}\epsilon) && \text{(2.5b)} \\ &= [E(X^{(k)\top}X^{(k)})]^{-1}E(X^{(k)\top}X\beta) + 0 && \text{(2.5c)} \\ &=: U_k\beta, \end{aligned}$$

where (2.5a) is simply the statement that $E(\hat{\beta}^{(k)}) = \beta^{(k)}$, (2.5b) follows from (2.1), and (2.5c) is a consequence of the orthogonality relations (2.2).

Note that the left and right sides of (2.5) are (column) vectors of length d_k. Thus, $\hat{\beta}^{(k)}$ is a d_k-dimensional vector each of whose entries is an unbiased estimator of a linear combination of the components of β, and because we know $E(X^\top X)$ exactly, we know both $[E(X^{(k)\top}X^{(k)})]^{-1}$ and

$E(X^{(k)\top}X)$ exactly. Therefore, $(\hat{\beta}^{(1)\top}, \dots, \hat{\beta}^{(K)\top})^\top$ is a vector of length $d = d_1 + \cdots + d_K$ for which each entry is an unbiased estimator of a linear combination of the components of β where the coefficients are known. In other words, we have the familiar linear model $(\hat{\beta}^{(1)\top}, \dots, \hat{\beta}^{(K)\top})^\top = U\beta + \varepsilon$, where ε has mean zero and U is a *known* matrix. Furthermore, because the K studies are independent, $\text{Var}(\varepsilon)$ has a simple structure: it is block diagonal, with K blocks, where for $k = 1, \dots, K$, block k is the $d_k \times d_k$ matrix $\text{Var}(\hat{\beta}^{(k)})$. Because $\text{Var}(\hat{\beta}^{(k)})$ is assumed known (in this section), $\text{Var}(\varepsilon)$ is *known*. Therefore, the full vector β may be estimated by the usual generalized least squares estimator $\tilde{\beta}_{\text{GLS}}$, and additionally $\text{Var}(\tilde{\beta}_{\text{GLS}})$ may be estimated by the usual generalized least squares estimator of variance.

We now express the development above more explicitly. Let $Z = (\hat{\beta}^{(1)\top}, \dots, \hat{\beta}^{(K)\top})^\top$, let V be the $d \times d$ block-diagonal matrix with K blocks, where the k^{th} block is $\text{Var}(\hat{\beta}^{(k)})$, and let

$$U = (U_1^\top, \dots, U_K^\top)^\top, \text{ where } U_k = [E(X^{(k)\top}X^{(k)})]^{-1}E(X^{(k)\top}X) \tag{2.6}$$

(U is simply a vertical stacking of $U_1, \dots, U_K$). Our development gives the linear model

$$Z = U\beta + \varepsilon \qquad (E(\varepsilon) = 0 \text{ and } \text{Var}(\varepsilon) = V), \tag{2.7}$$

but where V is not a constant times the $d \times d$ identity matrix. The full vector $\beta = (\beta_0, \beta_1, \dots, \beta_p)^\top$ may therefore be estimated by the generalized least squares estimator

$$\tilde{\beta}_{\text{GLS}} = (U^\top V^{-1} U)^{-1} U^\top V^{-1} Z, \tag{2.8}$$

for which the variance is $\text{Var}(\tilde{\beta}_{\text{GLS}}) = (U^\top V^{-1} U)^{-1}$. Furthermore, without any Gaussian assumption on the error term ϵ in (2.1), when the study sample sizes n_k are all large, we expect that $[\text{Var}(\tilde{\beta}_{\text{GLS}})]^{-1/2}(\tilde{\beta}_{\text{GLS}} - \beta)$ will be approximately distributed according to the $(p+1)$-dimensional normal distribution with mean 0 and identity as variance matrix (details are provided in the formal development given in Section 2.2 and the Appendix). We now summarize. *We can form a valid estimate of the full vector $\beta = (\beta_0, \beta_1, \dots, \beta_p)^\top$ from the K studies, possibly none of which use the full set of predictors $X_1, \dots, X_p$, if we know the matrix $E(X^\top X)$ and our variance estimates for the*

K studies are exact. When the n_k's are all large, the estimate is approximately normal, and consequently we can form a confidence ellipse for β and, significantly, we can form individual confidence intervals for any components of β.

Implicit in the expressions for $\tilde{\beta}_{\text{GLS}}$ and $\text{Var}(\tilde{\beta}_{\text{GLS}})$ is that U is of full rank, so that $U^\top V^{-1} U$ is invertible. Proposition 1 states that under our assumptions, U is indeed of full rank.

Proposition 1 *Under the assumptions that $E(X^\top X)$ is positive definite, and that for each component of $(1, X_1, \dots, X_p)$ there is at least one study which uses that component, the matrix U is of full rank.*

The proof is in the supplement Doss and Joo (2025).

2.2 The General Case

The development above applies to a situation in which there are two idealizations: (1) the matrix $E(X^\top X)$ is known exactly, and (2) the study variances $\text{Var}(\hat{\beta}^{(k)})$ are known exactly. In reality, these quantities must be estimated and this will inflate the variance of $\tilde{\beta}_{\text{GLS}}$. We now discuss the details.

Let $V_k = \text{Var}(\hat{\beta}^{(k)})$, so V is the $d \times d$ block-diagonal matrix with the V_k's as the blocks. Recall that S_k is an estimate of V_k, and is now no longer assumed exact. Let S denote the $d \times d$ block-diagonal matrix with the S_k's as the blocks. Let $E_{ij} = E(X_i X_j)$, $i, j = 0, 1, \dots, p$, and let $\widehat{E}_{ij}$'s be estimates of the E_{ij}'s obtained from some source possibly external to the K studies. The matrix U defined in (2.6) is based on the E_{ij}'s, and we will write $U = g(\{E_{ij},\, i, j = 0, 1, \dots, p\})$, where g is some (complicated) function. Let $\widehat{U} = g(\{\widehat{E}_{ij},\, i, j = 0, 1, \dots, p\})$. The "ideal" generalized least squares estimate is $\tilde{\beta}_{\text{GLS}}$ given by (2.8), whose variance is $(U^\top V^{-1} U)^{-1}$. The estimate that will be used is

$$\hat{\beta}_{\text{GLS}} = (\widehat{U}^\top S^{-1} \widehat{U})^{-1} \widehat{U}^\top S^{-1} Z, \tag{2.9}$$

and we will denote this estimate simply by $\hat{\beta}$.

Unfortunately, $\text{Var}(\hat{\beta})$ is not the simple expression $(\widehat{U}^\top \widehat{V}^{-1} \widehat{U})^{-1}$) that one might obtain from a naive and incorrect use of Slutsky's theorem, but is much more complicated. In order to state a theoretical result about it we will impose some conditions on the rate at which the S_k's, the $\widehat{E}_{ij}$'s, and the $\hat{\beta}^{(k)}$'s converge to the V_k's, the E_{ij}'s, and the $\beta^{(k)}$'s, respectively. The regularity conditions that we will ultimately impose (C1–C3 below) involve the joint behavior of the S_k's, the $\widehat{E}_{ij}$'s, and the $\hat{\beta}^{(k)}$'s, but it will be helpful to first discuss these three sets of quantities individually. We discuss C1–C3 as a whole after the statement of Theorem 1. Let $n = n_1 + \cdots + n_K$, where we recall that the n_k's are the study sample sizes.

Regarding the S_k's, we will assume that the study sample sizes are comparable: we will assume that there exist constants $\lambda_1, \ldots, \lambda_K \in (0,1)$ such that for each k, $n_k/n \to \lambda_k$, and that "$n_k^{1/2}(S_k - V_k)$ is asymptotically normally distributed." We need to express the statement in quotes more carefully, because S_k and V_k are symmetric matrices. To this end, let vech denote the operator on symmetric matrices that creates a column vector whose elements are the stacked columns of the lower triangular elements of the matrix. (This operator is invertible, i.e. if A is symmetric and we know $\mathrm{vech}(A)$, then we know A; so on occasion we will use A and $\mathrm{vech}(A)$ interchangeably.) The statement in quotes is replaced by the statement that for some positive-definite $[(d_k+1)d_k/2] \times [(d_k+1)d_k/2]$ matrices Ω_k the estimates S_k satisfy

$$n_k^{1/2}(\mathrm{vech}(S_k) - \mathrm{vech}(V_k)) \xrightarrow{d} \mathcal{N}_{(d_k+1)d_k/2}(0, \Omega_k) \qquad \text{as } n \to \infty. \tag{2.10}$$

The convergence statement (2.10) is typical for variance estimates based on a sample of size n_k.

Regarding the $\widehat{E}_{ij}$'s, we will assume that their rates of convergence to the E_{ij}'s are comparable across all $i, j \in \{0, 1, \ldots, p\}$ (here and throughout, we implicitly exclude the trivial case $(i,j) = (0,0)$, for which $X_i X_j = 1$). Specifically, we will assume that there exist integers m_{ij} (to be viewed as sample sizes) such that if $m := \sum_{i=1}^p \sum_{j=1}^p m_{ij}$, then in our asymptotic regime in which $m \to \infty$, there exist numbers $\gamma_{ij} \in (0,1)$ such that $m_{ij}/m \to \gamma_{ij}$. Furthermore, we will assume that there exist $\tau_{ij} > 0$ such that

$$m_{ij}^{1/2}(\widehat{E}_{ij} - E_{ij}) \xrightarrow{d} \mathcal{N}(0, \tau_{ij}^2) \qquad \text{as } m \to \infty. \tag{2.11}$$

Because typically the $\widehat{E}_{ij}$'s are empirical moments, and such moments are subject to the central limit theorem, (2.11) is a reasonable assumption. If we have access to an external dataset $\tilde{X}^{(i)}$, $i =$

$1, \ldots, N$, where the $\tilde{X}^{(i)}$'s are iid, having the same distribution as X, then the $\widehat{E}_{ij}$'s can be taken to be the sample cross-moments from the dataset, and the m_{ij}'s are then all equal (and so are the γ_{ij}'s). However, having access to a single external dataset is not a necessary condition, and it may be possible to obtain estimates $\widehat{E}_{ij}$'s separately, using multiple datasets, none of which have all the components of X.

Regarding the $\hat{\beta}^{(k)}$'s, we will assume that there exist positive-definite $d_k \times d_k$ matrices Σ_k such that

$$n_k^{1/2}(\hat{\beta}^{(k)} - \beta^{(k)}) \xrightarrow{d} \mathcal{N}_{d_k}(0, \Sigma_k) \qquad \text{as } n \to \infty. \tag{2.12}$$

Asymptotic normality of least squares estimators when the predictors are random is a well-known fact; see page 1219 of Freedman (1981) for a brief proof (the case of non-random predictors, which is well understood and for which asymptotic normality has been established under very general conditions, is not relevant here).

Theorem 1 refers to the conditions below. These are discussed right after the statement of the theorem. The main one is C2, which is a strengthening of (2.10), (2.11) and (2.12). C1 is a condition about the relative sizes of the samples sizes n_k and m_{ij} appearing in (2.10), (2.11) and (2.12).

C1 There exist positive constants λ_k, $k = 1, \ldots, K$ and m_{ij}, $i, j \in \{0, 1, \ldots, p\}$, such that $n_k/n \to \lambda_k \in (0, 1)$ and $m_{ij}/m \to \gamma_{ij} \in (0, 1)$. Here, $n = n_1 + \cdots + n_K$ and $m = \sum_{i=1}^p \sum_{j=1}^p m_{ij}$. Also, $n/m \to c \in [0, \infty)$ (note that this interval is closed on the left).

C2 The joint convergence statement below is written in an asymptotic regime where $n \to \infty$, as opposed to $n_k \to \infty$ and $m_{ij} \to \infty$. For some positive-definite matrix Ψ, the joint distribution of the S_k's, $\widehat{E}_{ij}$'s and $\hat{\beta}^{(k)}$'s satisfies

$$n^{1/2}\big(\{\mathrm{vech}(S_k) - \mathrm{vech}(V_k)\}_{k=1}^K, \{\widehat{E}_{ij} - E_{ij}\}_{i,j=0}^p, \{\hat{\beta}^{(k)} - \beta^{(k)}\}_{k=1}^K\big) \xrightarrow{d} \mathcal{N}(0, \Psi) \quad \text{as } n \to \infty. \tag{2.13}$$

In (2.13), the limiting variance of $n^{1/2}(\mathrm{vech}(S_k) - \mathrm{vech}(V_k))$ is Ω_k/λ_k (see (2.10)), the limiting variance of $n^{1/2}(\widehat{E}_{ij} - E_{ij})$ is $\tau_{ij}^2 c/\gamma_{ij}$ (see (2.11)), and the limiting variance of $n^{1/2}(\hat{\beta}^{(k)} - \beta^{(k)})$

is Σ_k/λ_k (see (2.12)). If the constant c is 0, then the interpretation of the statement $n^{1/2}(\widehat{E}_{ij} - E_{ij}) \xrightarrow{d} \mathcal{N}(0, 0)$ is that $\widehat{E}_{ij} - E_{ij} = o_p(n^{-1/2})$.

C3 $E(\hat{\beta}^{(k)}) = \beta^{(k)}$ for every k. (Recall that this condition holds under no assumptions on the joint distribution of (Y, X), parametric or otherwise (save for finiteness of fourth moments), if the $\hat{\beta}^{(k)}$'s are the least squares estimates and (2.4) holds.)

Theorem 1 *Under conditions C1–C3, $n^{1/2}(\hat{\beta} - \beta) \xrightarrow{d} \mathcal{N}_{p+1}(0, \Phi)$ as $n \to \infty$ for some variance matrix Φ.*

Before presenting the proof of the theorem, we discuss our regularity conditions.

Remarks on Conditions C1–C3 Condition C2 is a strengthening of (2.10), (2.11) and (2.12) to a statement about joint convergence. Here, joint means that in particular, (A) not only do the standardized S_k's converge to a Gaussian limit, but they converge jointly, likewise for the standardized $\widehat{E}_{ij}$'s, and likewise for the standardized $\hat{\beta}^{(k)}$'s; and (B) not only do the three sequences $\{S_k\}_{k=1}^K$, $\{\widehat{E}_{ij}\}_{i,j=0}^p$, and $\{\hat{\beta}^{(k)}\}_{k=1}^K$ converge, but they converge jointly.

Regarding point (A) there is nothing complicated. When (2.12) holds for every k, joint convergence of the sequence $\{\hat{\beta}^{(k)}\}_{k=1}^K$ holds simply because the K studies are assumed independent; and likewise, when (2.10) holds for every k, joint convergence of the sequence $\{S_k\}_{k=1}^K$ holds for the same reason. As for the $\widehat{E}_{ij}$'s, if there is an unlabelled sequence of vectors $(X_{\ell i_1}, \ldots, X_{\ell i_t})$, $\ell = 1, \ldots, L$, and the estimates $\{\widehat{E}_{i_r j_s},\, r, s = 1, \ldots, t\}$ are obtained as empirical moments from the unlabelled sequence, then these estimates, when standardized, converge jointly; and when all the estimates $\widehat{E}_{ij}$ are obtained from multiple independent unlabelled sequences, we have joint convergence also.

Regarding point (B), the estimates $\widehat{E}_{ij}$ will typically be obtained from sources that are independent of the K studies, so the issue of substance regarding point (B) is joint convergence of $\hat{\beta}^{(k)}$ and S_k, which we now discuss, for the case where $\hat{\beta}^{(k)}$ is the least squares estimate and S_k is the conventional corresponding estimate of variance. For the standard linear model with Gaussian errors and non-random predictors, it is well known that the least squares estimate and the usual estimate of variance are independent. When the errors are not Gaussian, independence need not hold, even asymptotically. However, condition C2 does not require independence of $\hat{\beta}^{(k)}$ and S_k; it requires only that the standardized versions of these two estimators are asymptotically jointly normal, and we now show that this is the case. To ease the notation, in the paragraph below we drop the subscript/superscript k.

It is well known that for the case of random predictors, the least squares estimator is asymptotically normal with mean 0 and variance matrix $Q^{-1}MQ^{-1}$, where $Q = E(X^\top X)$ and M is the matrix for which $M_{rs} = E(X_r X_s \epsilon^2)$, where ϵ corresponds to the decomposition (2.1) but for study k; see page 1219 of Freedman (1981). From the vectors $X_1, \dots, X_n$, with $X_i = (X_{i1}, \dots, X_{id})$, consider the sums

$$\begin{gathered}\sum_{i=1}^n X_{ir}X_{is},\ r,s = 1,\dots,d, \quad \sum_{i=1}^n X_{ir}Y_i,\ r = 1,\dots,d, \quad \sum_{i=1}^n X_{ir}X_{is}Y_i^2,\ r,s = 1,\dots,d, \\ \sum_{i=1}^n X_{ir}X_{is}X_{it}Y_i,\ r,s,t = 1,\dots,d, \quad \sum_{i=1}^n X_{ir}X_{is}X_{it}X_{iu},\ r,s,t,u = 1,\dots,d.\end{gathered} \tag{2.14}$$

The multivariate central limit theorem asserts that these sums, when standardized, are jointly asymptotically normal. Consider the natural estimates of $E(X_r X_s)$ and $E(X_r X_s \epsilon^2)$ formed from the empirical averages $(1/n)\sum_{i=1}^n X_{ir}X_{is}$ and $(1/n)\sum_{i=1}^n X_{ir}X_{is}\epsilon_i^2$, where $\epsilon_i = Y_i - X_i\hat{\beta}$. It is clear that $\hat{\beta}$ and $(1/n)\sum_{i=1}^n X_{ir}X_{is}$ are functions of the sums in (2.14), and it is not difficult to see that $(1/n)\sum_{i=1}^n X_{ir}X_{is}\epsilon_i^2$ is also a function of the sums in (2.14) ($(1/n)\sum_{i=1}^n X_{ir}X_{is}\epsilon_i^2$ also depends on $\hat{\beta}$, but $\hat{\beta}$ is a function of the sums in (2.14)). Thus, if we take $\widehat{Q}$ and $\widehat{M}$ to be the natural estimates of Q and M obtained by replacing $E(X_r X_s)$ and $E(X_r X_s \epsilon^2)$ by their empirical averages, we see that $\widehat{Q}$ and $\widehat{M}$ are (differentiable) functions of the sums in (2.14), and hence so is $\widehat{Q}^{-1}\widehat{M}\widehat{Q}^{-1}$. So, by the multivariate delta method, $\hat{\beta}$ and $\widehat{Q}^{-1}\widehat{M}\widehat{Q}^{-1}$ are jointly asymptotically normal.

A curious fact is that if for every pair $i, j \in \{0, 1, \dots, p\}$ there exists a $k = k_{ij}$ such that study k includes covariates X_i and X_j, and if the study records the pairs $(X_{i1}, X_{j1}), \dots, (X_{in_k}, X_{jn_k})$, then there is no need to have any external data sets in order to estimate $E(X^\top X)$: E_{ij} may be estimated by the empirical joint moment $(1/n_k)\sum_{s=1}^{n_k} X_{is}X_{js}$. The estimates $\widehat{E}_{ij}$ would then not necessarily be independent of the $\hat{\beta}^{(k)}$'s and S_k's, but there is nothing in assumptions C1–C3 that requires such independence. (We write "curious fact" because it seems counter-intuitive that the full parameter β is identifiable from a set of studies none of which include the full covariate vector X.)

Proof of Theorem 1 From the formula for $\hat{\beta}$ given by (2.9) we see that $\hat{\beta}$ is a function of S, $\widehat{U}$ and Z. Now, S is given by the sequence $S_1, \ldots, S_K$, $\widehat{U}$ itself is a function of $\bigl\{\widehat{E}_{ij},\, i, j \in \{0, 1, \ldots, p\}\bigr\}$, and $Z = (\hat{\beta}^{(1)\top}, \ldots, \hat{\beta}^{(K)\top})^\top$. Therefore, we may write

$$\hat{\beta} = f\bigl(\{S_k\}_{k=1}^K,\, \{\widehat{E}_{ij}\}_{i,j=0}^p,\, \{\hat{\beta}^{(k)}\}_{k=1}^K\bigr), \tag{2.15}$$

where f is some function. We now argue that

$$\beta = f\bigl(\{V_k\}_{k=1}^K,\, \{E_{ij}\}_{i,j=0}^p,\, \{\beta^{(k)}\}_{k=1}^K\bigr), \tag{2.16}$$

where in (2.16) f is the same function that appears in (2.15). From (2.5) and the equation $E(\hat{\beta}^{(k)}) = \beta^{(k)}$ right above (2.4), we obtain $(\beta^{(1)\top}, \ldots, \beta^{(K)\top})^\top = U\beta$. Left-multiplying this equation by $(U^\top V^{-1} U)^{-1} U^\top V^{-1}$, we get

$$(U^\top V^{-1} U)^{-1} U^\top V^{-1} (\beta^{(1)\top}, \ldots, \beta^{(K)\top})^\top = (U^\top V^{-1} U)^{-1} U^\top V^{-1} U\beta = \beta.$$

This gives (2.16). Now, because U is of full rank (see Proposition 1), f is differentiable at β. Therefore, in view of (2.13) we may apply the delta method, giving

$$n^{1/2}(\hat{\beta} - \beta) \xrightarrow{d} \mathcal{N}_{p+1}\bigl(0, \nabla f(\theta)^\top\, \Psi\, \nabla f(\theta)\bigr),$$

thus proving the theorem with $\Phi = \nabla f(\theta)^\top\, \Psi\, \nabla f(\theta)$.

Estimation of Φ The conventional approach to estimating Φ, namely separately estimating $\nabla f(\theta)$ and Ψ, is difficult to implement for several reasons, including the fact that it requires us to take the gradient of a function of several matrices, which is very complicated. An alternative is to use the bootstrap, and we now discuss this.

Consider the idealized situation in which $E(X^\top X)$ and the variance matrices $V_1, \ldots, V_k$ are known exactly. Display (2.7) gives the equation $Z = U\beta + \varepsilon$, where ε has mean 0 and variance matrix V. We may re-express the equation as $V^{-1/2}Z = V^{-1/2}U\beta + V^{-1/2}\varepsilon$, or $\bar{Z} = \bar{U}\beta + \bar{\varepsilon}$, in self-explanatory notation, and we have $\mathrm{Var}(\bar{\varepsilon}) = I_{d\times d}$. Thus, if we let $\bar{Z}_i$ denote the i^{th} component of the d-dimensional vector $\bar{Z}$ and let $\bar{U}_i$ denote the i^{th} row of the $d \times (p+1)$ matrix $\bar{U}$, we have $\bar{Z}_i = \bar{U}_i^\top \beta + \bar{\varepsilon}_i$, and $\hat{\beta}$ emerges as the ordinary least squares estimate based on the pairs $(\bar{U}_i, \bar{Z}_i)$, $i = 1, \ldots, d$. The predictors $\bar{U}_i$ are random, not fixed, and the appropriate bootstrap scheme for this situation is to resample the pairs $(\bar{U}_i, \bar{Z}_i)$, $i = 1, \ldots, d$ (this is as opposed to resampling the residuals from the fitted model, which would be appropriate if the $\bar{U}_i$'s were fixed). Freedman (1981) discusses this scheme of resampling the pairs $(\bar{U}_i, \bar{Z}_i)$, and shows that it gives consistent estimates of the variance of the standardized $\hat{\beta}$ (a precise statement is given in Theorem 3.1 of Freedman (1981)).

As mentioned before, $E \coloneqq E(X^\top X)$ and V are not known exactly, and we must use estimates $\widehat{E}$ for E and S_k for V_k, $k = 1, \ldots, K$. Let $\widehat{\Phi}_B$ denote the bootstrap estimate of variance that is obtained from the resampling scheme described above, using the observed values $\widehat{E}$ and $S_1, \ldots, S_K$. There are two points to note:

1. Since $\widehat{E}$ is included, intact, in the entire execution of the bootstrap scheme, $\widehat{\Phi}_B$ is really an estimate of the conditional variance $\mathrm{Var}(\hat{\beta} \,|\, \widehat{E})$. Now because of the decomposition $\mathrm{Var}(\hat{\beta}) = E(\mathrm{Var}(\hat{\beta} \,|\, \widehat{E})) + \mathrm{Var}(E(\hat{\beta} \,|\, \widehat{E}))$, we see that $\widehat{\Phi}_B$ underestimates $\mathrm{Var}(\hat{\beta})$ on average. However, this effect is small if the estimate $\widehat{E}$ is accurate, as would be the case when $\widehat{E}$ is formed from a very large external data set.

2. Consider a parametric setting in which θ is a parameter of interest and ψ is a vector of nuisance parameters, and assume that standard regularity conditions hold. Suppose that $\mathcal{I}_o$ is a confidence interval for θ with asymptotic coverage probability 0.95 in the oracle model in which ψ is known, and let $\mathcal{I}$ be the same interval, except that a $\sqrt{n}$-consistent estimate of ψ is used. A heuristic, supported by numerous examples, is that $\mathcal{I}$ also has asymptotic coverage probability equal to 0.95. Based on this heuristic, we expect that confidence sets for β using $\widehat{\Phi}_B$ will have asymptotic coverage probability equal to 0.95 even though $\widehat{\Phi}_B$ uses the S_k's instead of the unknown V_k's. This heuristic is strongly confirmed in our setting; see Section 3.1.

It is worth emphasizing that typically the external data set used to estimate E is very large, because the data can be taken from electronic records unrelated to any of the studies and can be unlabelled.

Bootstrap confidence interval methodology provides intervals that are more elaborate than the simple confidence intervals of the sort "point estimate $\pm$ 1.96 $\times$ bootstrap standard error estimate", and for these, convergence of the coverage probability to 0.95 is faster; see Hall (1988) for a review. In our simulation studies we have not used these because the meta-analyses we have in mind involve large observational studies, so the gains offered by better small-sample asymptotics are quite minor. However, in a setting where the study sample sizes are small, a user may prefer these more elaborate intervals.

3 Numerical Studies

This section consists of two parts. Section 3.1 presents the results of a simulation study whose aim is to evaluate the performance of our methodology on several criteria and to give an empirical check on the asymptotic normality asserted in Theorem 1. Section 3.2 illustrates the methodology on a real dataset.

3.1 Simulation Study

Our simulations are conducted using the framework of the standard linear model (1.1) with homoscedastic errors. Their purpose is to assess the bias and variance of $\hat{\beta}$, evaluate the coverage probability of the bootstrap confidence intervals, and give empirical support to the statement regarding asymptotic normality of $\hat{\beta}$ made in Theorem 1, all of these under a range of settings.

The simulations were run under the following specifications. In a single data set, the number of studies was $K = 20$, and the dimension of the covariate vector was $p = 8$. For study 1, the covariate vector was $X^{(1)} = (1, X_1, X_2)$, for study 2 it was $X^{(2)} = (1, X_3, X_4)$, for study 3 it was $X^{(3)} =$

$(1, X_5, X_6)$, for study 4 it was $X^{(4)} = (1, X_7, X_8)$, for study 5 it was $X^{(5)} = (1, X_1, X_2, X_7, X_8)$, and for study 6 it was $X^{(6)} = (1, X_3, X_4, X_5, X_6)$. The covariates for studies 7–12 and studies 13–18 followed exactly the same pattern, and for studies 19 and 20 they were $(1, X_1, X_2)$, and $(1, X_3, X_4)$, respectively. The sample sizes for the 20 studies were as follows: $n_1 = \cdots = n_5 = 300$, $n_6 = \cdots = n_{10} = 400$, $n_{11} = \cdots = n_{15} = 500$, and $n_{16} = \cdots = n_{20} = 600$. The entire vectors $(X_1, \ldots, X_8)$ for the 20 studies were generated independently from the multivariate normal distribution with mean vector $\mu_X = (10, -20, 30, -10, 20, -30, 10, -20)$ and variance matrix Σ_X, for which the diagonal entries are all equal to 10, and all the off-diagonal entries are equal to the same value ρ_X. There are four scenarios, corresponding to $\rho_X = 0.3$, 0.6, 0.8, and 0.9. The external data set was generated from the same multivariate normal distribution, with a sample size of 50,000 (we discuss the effect of reducing the size of the external data set later in this section). The error term ϵ was generated from a scaled t-distribution with 4.1 degrees of freedom; specifically, $\epsilon/3 \sim t_{4.1}$. (The t-distribution with a degrees of freedom has a finite moment of order b for any $b < a$, but does not have a moment of order a. Our theory requires that the predictors have a finite fourth moment, and we took the degrees of freedom parameter to be close to 4 in order to evaluate "robustness" of our theoretical results.) Once the predictors and the error term are generated, the response Y is determined by the linear model in (1.1) with the true β equal to $(-4, 2, 0, -2, 4, -2, 0, 2)$. For each of the four values of ρ_X, 5,000 data sets were generated independently, where a single data set consists of the n_k (covariate, response) pairs for $k = 1, \ldots, 20$, with the n_k's specified above. For each data set, the estimates $\hat{\beta}^{(k)}$ and S_k were obtained through standard least squares, for $k = 1, \ldots, 20$, and from these the estimate $\hat{\beta}$ of the full vector β was calculated through our methodology. Thus, for each of the 5,000 data sets, the calculation of $\hat{\beta}$ is based on only the summary statistics $(\hat{\beta}^{(k)}, S_k)$, $k = 1, \ldots, 20$.

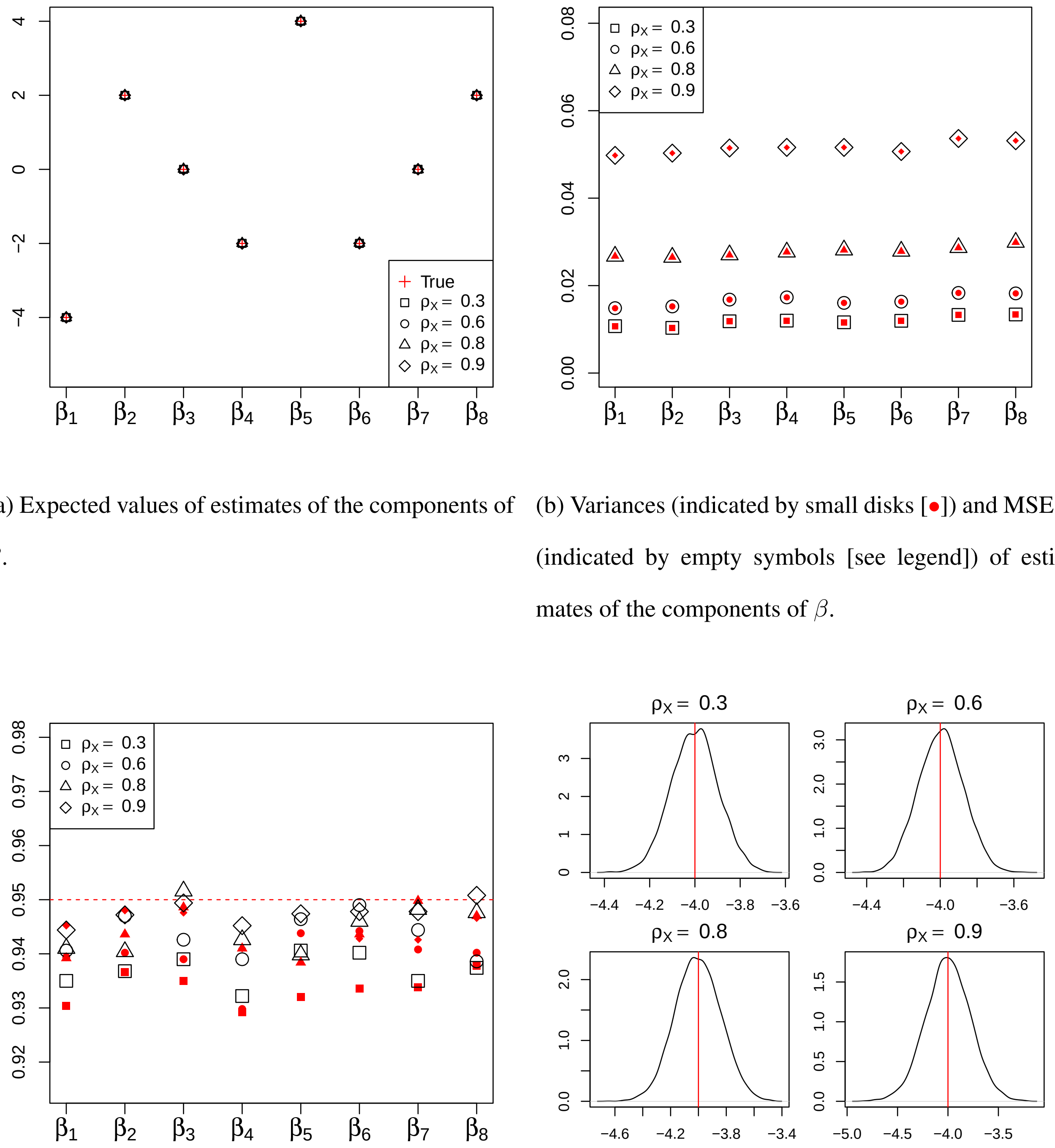
Figure 1 summarizes the results. Panel (a) shows, for each component of the full vector β, the mean of the 5,000 estimates of that component, and this for each of the four values of ρ_X; the panel also displays the true value of the component. In the panel we see that for each component of β the five values are visually virtually indistinguishable, which strongly suggests that our estimates are essentially unbiased (at least in the setting described above). The Monte Carlo standard error associated with the mean is less than 0.237 across all $4 \times 8 = 32$ estimates. Panel (b) pertains to the variance and mean squared error (MSE). From the panel, we see that these two quantities are visually indistinguishable across all components of β and all values of ρ_X (Monte Carlo standard errors for the 64 estimates are all less than 0.241). Our intuition is that β becomes harder to estimate when the predictors become more highly correlated, and panel (b) is consistent with this intuition.

Next, we discuss bootstrap confidence sets for β. Let $\rho_X \in \{0.3, 0.6, 0.8, 0.9\}$ be fixed. The bootstrap confidence intervals for the components of β were constructed as follows. For each of the $5{,}000$ simulated data sets (each data set consists of the data generated for 20 studies), using the resampling scheme described in the paragraph "Estimation of Φ" in Section 2.2, we generated $B = 1{,}000$ bootstrap samples. From the $1{,}000$ bootstrap samples we formed the usual sample covariance matrix, call it R. A 95% bootstrap confidence interval for the i^{th} component of β is $\hat{\beta}_i \pm 1.96 R_{ii}$. We now review and elaborate on two points raised in Section 2.2. First, R is really an estimate of $\text{Var}(\hat{\beta} \mid \widehat{E})$, which is smaller than $\text{Var}(\hat{\beta})$ in expectation. This could result in a coverage probability that is smaller than the nominal value of 0.95, although we do not expect the discrepancy to be significant if the external data set is large. Second, in the theoretical justification for the bootstrap estimate of variance (Theorem 3.1 of Freedman (1981)), we assume that the matrix $V^{-1/2}$ is known exactly, whereas we only have an estimate of it. Again, we do not expect this to be a significant problem if the data sets are large (see point (2) in the paragraph "Estimation of Φ"). To investigate the potential undercoverage of our bootstrap confidence intervals, for each of the $5{,}000$ simulated data sets we constructed two confidence intervals. One of them, which we call the "oracle confidence interval," is obtained by using the true values of E and V; and the other, which we call the "real confidence interval," is obtained by using the estimates of E and V.

Panel (c) of Figure 1 shows the coverage probabilities of both the oracle and the real confidence intervals. From the panel we see that across the 4×8 cases, the coverage probabilities of the oracle intervals are close to the nominal value (always between 0.932 and 0.952). The coverage probabilities of the real confidence intervals are almost always lower, ranging from 0.93 to 0.95. However, the difference in the coverage probabilities between the real and oracle intervals is essentially negligible across all cases. Figures S-1 and S-2 in Doss and Joo (2025) localize the effects of using the estimated values of E and using the estimated values of V separately. Figure S-1 in Doss and Joo (2025) 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. Figure S-2 in Doss and Joo (2025) 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.

Panel (d) shows density estimates for β_1. These were constructed by applying the R function `density` (with all options set to default) to the first component of the $5{,}000$ simulated values of $\hat{\beta}$. The four density estimates, and similar ones for the other seven components of β (plots not shown), are all close to normal, giving empirical validation for Theorem 1.

Section S-2 of Doss and Joo (2025) reports results of additional simulation studies, including simulations to evaluate the effect K, $d_1, \dots, d_K$, and p (recall that d_k is the number of predictors used in study k). These simulations suggest that our methodology performs well when we vary these configuration parameters. The most interesting conclusion from these additional simulations is that the bootstrap estimate of variance becomes unstable (and hence bootstrap confidence intervals become unreliable) when the d_k's are small, p is large, and K is small. Intuitively, it is to be expected that problems will arise when the d_k's are small, p is large, and K is small, but Doss and Joo (2025) provide a theoretical explanation for this, and through an expression that involves K, the d_k's and p, they provide guidance on identifying the extreme cases where the methodology will not work well.



(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 1: Aspects of the distribution of $\hat{\beta}$, and coverage probabilities of bootstrap confidence intervals.

Code for calculating $\hat{\beta}$, forming the bootstrap estimate of variance and bootstrap confidence intervals, and also for reproducing all the results in this section is provided in the supplementary material.

3.2 Illustration on an Analysis of the Association Between High Blood Pressure and a Liver Function Enzyme

The evaluation of the performance of our methodology on simulated data is certainly useful, but has the obvious limitation that the distributions we use in the simulations may be different from the underlying distributions in real data. On the other hand, evaluation of the performance using real data (for example using coverage probability of confidence intervals as the criterion) can be difficult because, in contrast to the situation that arises when carrying out simulations, the true value of β is not known. A reasonable approach for dealing with this problem is to consider a very large data set for which there is an outcome variable Y and several predictor variables $X_1, \ldots, X_p$, of which one (say X_1) is of principal interest: we wish to determine the causal effect of X_1 on Y, with $X_2, \ldots, X_p$ being potential confounders. The data set should be so large that in the regression model $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon$, the coefficient β_1 is essentially known, meaning that a confidence interval for it is very narrow. We then split the data set into two parts. One serves as the external data set used to estimate the matrix $E(X^\top X)$. The other part is split into K subparts, with each subpart providing data for a study. The K studies will use only a subset of the predictors, not the full set. This situation then mimics the setup considered in the present paper. We can therefore form an estimate and a confidence interval for β_1 based on the K studies and the external data set using the methods of this paper, and determine whether the confidence interval contains the true value of β_1 (more precisely, determine whether our confidence interval overlaps with the confidence interval for β_1 formed from the entire data set using the full model).

To carry out this approach we considered the 2013 Korean health checkup dataset (https://www.data.go.kr/en/data/15007122/fileData.do) which gives detailed health measurements for approximately one million national health insurance subscribers in South Korea who underwent health checkups in 2013. Our general objective is to quantify the causal effect of hypertension (high blood pressure) on the health of the liver. Gamma-glutamyl transferase (GGT) is an enzyme which plays a key role in liver function and is commonly used as a biomarker to assess liver health. Previous studies, including Lee et al. (2002), Sakboonyarat et al. (2023), Liu et al. (2012), and Yi et al. (2017), have found a significant association between blood pressure and GGT levels, suggesting that hypertension may lead to changes in GGT activity. Our specific objective is to evaluate the causal effect of mean arterial pressure (MAP, see Section S-3 of Doss and Joo (2025) for a precise definition) on GGT. Other variables included in the dataset which are potential confounders are body mass index (BMI); smoking status (recorded as a categorical variable with values 1 (not a smoker), 2 (previously smoked), and 3 (presently smoking)); drinking status, recorded simply as a binary variable; pre-meal blood glucose level; and total cholesterol. We will denote these by X_{bmi}, X_{smoke}, X_{drink}, X_{glucose} and X_{tc}, respectively, and we will denote BMI by X_{bmi}. The full model has GGT level as the response, and the vector $X = (X_{\text{map}}, X_{\text{bmi}}, X_{\text{smoke}}, X_{\text{drink}}, X_{\text{glucose}}, X_{\text{tc}})$ as the predictor. Our goal is to make inference on β_{map}, the regression coefficient for X_{map}.

Following the procedure described in the beginning of this subsection, the entire data set was randomly split into two parts. One, consisting of approximately half a million points, serves as the external data set used to estimate $E(X^\top X)$. The remainder of the dataset was randomly divided into 20 subparts, each consisting of approximately $25{,}000$ points, and viewed as data for a study. The studies were designed as follows: studies 1 and 2 use only X_{map}, excluding all other predictors. In constrast, studies 3–6 include X_{map} along with additional variables, and studies 7–20 exclude X_{map} but include other predictor variables. In more detail, study 3 includes X_{map} and X_{bmi}; study 4 includes X_{map} and X_{smoke}; study 5 includes X_{map} and X_{drink}; study 6 includes X_{map}, X_{glucose}, and X_{tc}; studies 7–10 include X_{bmi} and X_{smoke}; studies 11–14 include X_{drink}, X_{glucose}, and X_{tc}; studies 15–18 include X_{bmi}, X_{smoke}, and X_{drink}; and studies 19 and 20 include X_{glucose} and X_{tc}. We constructed three confidence intervals for β_{map}:

- An interval based on the entire data set formed from the model that uses all the predictors; we denote this interval by I_{full}.
- An interval based on the 20 studies and the estimate of $E(X^\top X)$, using the meta-analysis method developed in this paper; we denote it by I_{meta}.
- An interval based on the entire data set, but formed from the model that uses only X_{map}; we denote it by I_{naive}.

Evidently, we expect that I_{full} and I_{naive} will be very short, because they are based on a million data points; and we expect that I_{meta} will be considerably wider, if only because only 6 of the 20 studies include the variable X_{map}. We will make two comparisons: I_{full} and I_{naive}, and I_{full} and I_{meta}. The three intervals are clearly dependent, with a dependence structure having a complex form, so the most convenient way to account for multiplicity is to apply the Bonferroni correction. Thus, to make 95% statements, we take the confidence level of the intervals to be 98.33%.

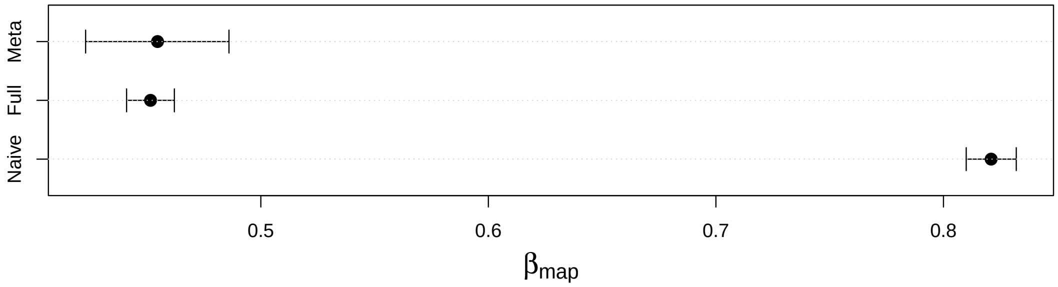


Figure 2: 98.33% confidence intervals for β_{map} for the Korean health checkup data set formed in three ways. The "Meta" interval is formed using our meta-analytic methodology; the "Full" interval is the standard interval based on all the subjects in the dataset in a regression model that uses all the predictors; and the "Naive" interval is the standard interval based on all the subjects in the dataset, but in a regression model that uses only X_{map} as the predictor.

Figure 2 displays the intervals in graphical form. From the figure we see that I_{naive} does not overlap with I_{full}, and our only purpose in calculating I_{naive} was to show this, in order to demonstrate the need to account for confounders in this problem. (In this regard, we note that the first two papers cited in the second paragraph of the present subsection consider only the predictor X_{map}, and thus inappropriately ignore confounding factors.) The main point of the figure is to show that I_{meta} and I_{full} overlap (in fact, I_{meta} contains I_{full} in its entirety), which is the principal goal of the illustration. We mention in passing that both I_{full} and I_{meta} do not include 0, indicating a causal relationship between β_{map} and GGT levels, a point on which we do not elaborate, as our focus is not on medical conclusions, but rather on statistical methodology. Some details regarding the definition of the variables in the dataset and regarding our analysis are given in the supplement Doss and Joo (2025).

4 Discussion

The problem considered in this paper is described briefly as follows. There are p predictor variables $X_1, \dots, X_p$ and an outcome variable Y. Let P be the joint distribution of $(Y, X_1, \dots, X_p)$. There is a regression model linking Y and $X_1, \dots, X_p$, and we would like to estimate the regression coefficients $\beta_0, \beta_1, \dots, \beta_p$ in this model, but we do not observe replicates $(Y_i, X_{i1}, \dots, X_{ip}) \overset{\text{iid}}{\sim} P$, $i = 1, \dots, n$. Rather, we observe summary statistics (point estimates and their standard error estimates) from K studies, where for $k = 1, \dots, K$, study k uses data of the form $(Y, X^{(k)})$, in which $X^{(k)}$ is a subvector of $(X_1, \dots, X_p)$. We also observe data giving information on the joint distribution of $(X_1, \dots, X_p)$. This problem has been considered before, as mentioned in the Introduction. Kundu et al. (2019) have considered it in a parametric framework, and their methodology requires that we can estimate the entire joint distribution of $(X_1, \dots, X_p)$. Other papers have considered the problem without imposing parametric assumptions, but in a framework where each study does a univariate regression of Y on a single predictor. See, for example, Li et al. (2020) (Li et al. (2020) state that they make a Gaussian assumption on $(Y, X_1, \dots, X_p)$, but in a close look at their proofs, we see that the assumption is not actually used.) The methodology developed in these papers does not seem to extend to the case where the K studies use multiple predictors, and more importantly, these papers do not establish asymptotic normality of $(\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p)$, so they do not develop a rigorous method for obtaining confidence regions or confidence intervals.

Our methodology does not require that we can estimate the joint distribution of the entire vector $(X_1, \dots, X_p)$; we need only estimate the covariances $\mathrm{Cov}(X_i, X_j)$, and estimates of $\mathrm{Cov}(X_i, X_j)$ can be obtained separately for different pairs i and j. We also do not make any parametric assumptions, and we do not require that each of the K studies does a univariate regression (in our development, any study can use any subset of the predictors). On the other hand, our development is for the linear model, and our methodology does not handle generalized linear models, for example logistic regression.

An interesting research direction is to consider the problem described in the first paragraph of this section for the case where the K studies evaluate causal effects through the use of propensity scores, and we now elaborate. Suppose that the variable of principal interest is X_1, and one is interested in assessing the causal effect of X_1. Recall that for study k, only the subvector $X^{(k)}$ is measured. Write $X^{(k)} = (X_1, X^{(k)}_{(-1)})$, where $X^{(k)}_{(-1)}$ denotes the vector $X^{(k)}$ except for the entry X_1. It is often the case that when estimating the regression coefficient of X_1, study k takes into account $X^{(k)}_{(-1)}$ through propensity scores, as opposed to through a regression of Y on $X^{(k)}$. In the situation where one or more of the studies use propensity scores instead of using a standard regression model, how do we combine the data (i.e. the summary statistics) from all the studies plus information on the joint distribution of the full covariate vector to arrive at a valid estimate of the regression coefficient of X_1?

References

Berrett, T. B., Wang, Y., Barber, R. F. and Samworth, R. J. (2020). The conditional permutation test for independence while controlling for confounders. *Journal of the Royal Statistical Society,* Series B **82** 175–197.

Candès, E., Fan, Y., Janson, L. and Lv, J. (2018). Panning for gold: ‘model-X’ knockoffs for high dimensional controlled variable selection. *Journal of the Royal Statistical Society,* Series B **80** 551–577.

Doss, H. and Joo, J. (2025). Supplement to “Valid causal inference in linear regression from several observational studies none of which account for all confounders”.

Freedman, D. A. (1981). Bootstrapping regression models. *The Annals of Statistics* **9** 1218–1228.

Hall, P. (1988). Theoretical comparison of bootstrap confidence intervals. *The Annals of Statistics* **16** 927–953.

Kundu, P., Tang, R. and Chatterjee, N. (2019). Generalized meta-analysis for multiple regression models across studies with disparate covariate information. *Biometrika* **106** 567–585.

Lee, D.-H., Ha, M.-H., Kim, J.-R., Gross, M. and Jacobs Jr, D. R. (2002). Gamma-glutamyltransferase, alcohol, and blood pressure: A four year follow-up study. *Annals of Epidemiology* **12** 90–96.

Li, H., Miao, W., Cai, Z., Liu, X., Zhang, T., Xue, F. and Geng, Z. (2020). Causal data fusion methods using summary-level statistics for a continuous outcome. *Statistics in Medicine* **39** 1054–1067.

Liu, C.-F., Gu, Y.-T., Wang, H.-Y. and Fang, N.-Y. (2012). Gamma-glutamyltransferase level and risk of hypertension: A systematic review and meta-analysis. *PloS One* **7** e48878.

Sakboonyarat, B., Poovieng, J., Lertsakulbunlue, S., Jongcherdchootrakul, K., Srisawat, P., Mungthin, M. and Rangsin, R. (2023). Association between raised blood pressure and elevated serum liver enzymes among active-duty Royal Thai Army personnel in Thailand. *BMC Cardiovascular Disorders* **23** 143.

Yi, S.-W., Lee, S.-H., Hwang, H.-J. and Yi, J.-J. (2017). Gamma-glutamyltransferase and cardiovascular mortality in Korean adults: A cohort study. *Atherosclerosis* **265** 102–109.