Discussion Paper Series A No.747

LINEAR REGRESSIONS, SHORTS TO LONG

Toru Kitagawa
(Department of Economics, Brown University,
and Institute of Economic Research, Hitotsubashi University)
and
Masayuki Sawada
(Institute of Economic Research, Hitotsubashi University)

August 2023

Institute of Economic Research
Hitotsubashi University
Kunitachi, Tokyo, 186-8603 Japan

LINEAR REGRESSIONS, SHORTS TO LONG[1]

Toru Kitagawa[ab] and Masayuki Sawada[b]

We study the identification problem of the *linear* long regression coefficients by data combination. Unlike the usual data combination problem, we consider combining multiple *short* regressions of the same outcome with different regressors. For this conceptually novel problem, we provide partial identification results for the long regression coefficients under a restriction on the unknown correlation structure. Specifically, we employ an elliptic constraint from the relations among the explained variations of the regressions to induce the bounds.

Keywords: Data combination, Linear regression, Elliptic constraint.

1. INTRODUCTION

Suppose that there are two studies on the same outcome y sampled from the same population and each study sample contains different key variables x_1 and x_2 that are unavailable jointly. Namely, we obtain the summary statistics from two studies of (y, x_1) and (y, x_2) separately, but not of (y, x_1, x_2). In this study, we consider the identification problem of the long regression $E[y|x_1, x_2]$ when the short regressions from two separate studies $E[y|x_1]$ and $E[y|x_2]$ are available.

This problem is similar to but different from the existing identification problem under data combination. Cross and Manski (2002) study nonparametric identification of the long regression $E[y|x_1, x_2]$ from the short conditional distributions $P[y|x_1]$ and $P[x_1|x_2]$. Ichimura and Martinez-Sanchis (2009) consider a semiparametric estimation problem of a related data combination problem. Molinari and Peski (2006) generalized the result of Cross and Manski (2002) for the infinitely supported y. Pacini (2019) further studies the relation in long linear projections and short linear projections. We focus on a different problem of different short regressions of $E[y|x_1]$ and $E[y|x_2]$. Such a problem appears in a frequent context but little has been done in the literature.

This identification problem appears frequently as an empirical challenge. For example, a study of y on x_2 suggests that x_2 is a key confounder for the causal relation between

[1] First Version: August, 2023.

[a] Department of Economics, Brown University.

[b] Institute of Economic Research, Hitotsubashi University.

y and x_1. However, the confounder (x_2) is unobserved in the study sample of y on x_1. This unobserved variable (x_2) may also be a candidate for the mediation analysis in the relation between y and x_1. For the example below, we illustrate the problem of combining two short regressions that suffer from the omitted variable bias.

Example 1.1 Suppose that one is interested in the impact of the years of schooling (x_0) on the earnings (y). The leading concern is the ability bias: the unobserved ability is correlated with both education level and earnings. Certain survey data contain a usually unavailable confounding variable such as the IQ test score (x_1); however, another key confounder of the birth location (x_2) may be limited because of confidentiality reasons. Nevertheless, the geographical information (x_2) with the earnings data (y) may be publicly available in a different data source. One may run two separate regressions of earnings (y) on schooling (x_0) and IQ score (x_1), and earnings (y) on schooling (x_0) and a location variable (x_2). Both regressions suffer from omitted variable biases; hence, we aim to recover the coefficient on x_0 in the long-regression of earnings (y) on schooling (x_0), and both IQ score (x_1) as well as the location variable (x_2).

Objectives in these examples require the coefficients of the long regression $E[y|x_1, x_2]$; however, only the short regressions $E[y|x_1]$ and $E[y|x_2]$ may be available.

In this study, we propose partial identification strategies for the *linear* long regression coefficients. First, we consider the case when all the variables in the long regression appear in some of the study samples. If the correlation structure of x_1 and x_2 is known, then the short regression coefficients point identify the long regression coefficients. If the correlation structure is unknown, then the target parameter is only partially identified. We consider the bounds as a tool for the sensitivity analysis of the target parameter in the worst-case correlation coefficients.

Second, we consider an additional omitted variable that is never observed in any studies. In Example 1.1 above, we consider the IQ test score as a proxy for the unobserved ability. However, the residual ability net of the IQ score may still plague the identification. With such an omitted variable that is never observed, the target parameter is not point identified from the known correlation structure. Nevertheless, we show that the elliptic constraint on the explained variation of the long regression induces the bounds that are the intersection of bounds from each study. The use of the elliptic constraint is related to the seminal work of Leamer (1978) and its followers. This use of the elliptic constraint is conceptually new in the literature of data combination.

We also contribute to the literature of sensitivity analysis against omitted variables. The sensitivity of regression estimates are studied in a wide range of literature including Mauro (1990), Murphy and Topel (1990), Frank (2000), Imbens (2003), Altonji et al. (2005), Clarke (2009), Bellows and Miguel (2006), Hosman et al. (2010), González and Miguel (2015), Krauth (2016), Cinelli and Hazlett (2020). Recently, Masten and Poirier (2022) study the sensitivity measure proposed by Oster (2019) that follows the idea of Altonji et al. (2005) and show that the sign of the coefficient of interest is sensitive to the measure when the omitted variable is relatively unimportant than the observed control variables. Diegert et al. (2022) also consider the sensitivity analysis without assuming that control variables are exogenous. Those recent studies approach the omitted variable problem for a particular variable of interest in the long regression in a single study. We approach this problem from a different perspective of combining multiple short regressions. Specifically, we demonstrate that the combining multiple short regressions tighten the bounds for the long regression with omitted variables.

In the remainder of this paper, we present the identification result in Section 2. In Section 3, we introduce an omitted variable that is never observed in either of the two studies. In Section 4, we conclude the paper with discussion of future challenges.

2. IDENTIFYING LINEAR LONG REGRESSION FROM SHORT REGRESSIONS

Consider there are two studies $s \in \{1, 2\} \equiv \mathcal{S}$ on each study sample drawn from a single population. For example, the two studies may contain key omitted variables separately, the determinants of a treatment assignment separately, or a treatment variable and a mediating variable separately. Two studies share the same outcome of interest $y \in \mathbb{R}$, the *ideal* vector of L explanatory variables $z \in \mathcal{R}^L$ and a linear conditional expectation function (CEF) of y given z, $E[y|z] = z'\beta$. However, the whole vector z is never observed in each study. Alternatively, for each study s, we observe a sample of (y, z_s) where $z_s = [x_0', x_s']'$ is a subvector of z such that x_0 is common across studies but x_s is study specific. By construction, we have $x_1 \cap x_2 = \emptyset$ and $z = [x_0', x_1', x_2']'$. Our target parameter is $c'\beta$ for some non-zero vector $c \in \mathbb{R}^L$; namely, we aim to recover some of the long-regression coefficients when only the short-regression coefficients are available.

Such an evidence aggregation problem arises when a researcher *ex-post* realizes the use of two different study samples jointly. Consequently, the researcher may have assess only through the summary statistics from published studies, confidential administrative, or proprietary data. Hence, we limit the information available to the analyst to be the following: the short projection coefficients $\mathcal{L}(y|z_s) = z_s'\gamma_s$, $s \in \mathcal{S}$; the variance-covariance matrices of z_s, $\Sigma_{s,s'} \equiv E(z_s z_{s'})$ for $\{s, s'\} \in \mathcal{S}$; the residual variance of linear projection $\eta^2(z_s) \equiv Var(y - \mathcal{L}(y|z_s))$ for $s \in \mathcal{S}$; and the variance of y, $Var(y)$. Let $L_s < L$ be the length of x_s for each $s \in \{0, 1, 2\}$.

In addition to the Example 1.1 in Introduction, we illustrate the motivation further in the following two examples:

Example 2.1 Consider a hypothetical policy ($x_{1,1}$) offering tuition waivers for eligible students. Suppose the eligibility ($x_{1,1}$) is purely determined by the field of study ($x_{1,2}$) and parental income (x_2). One study collects a representative sample of schools and their students about their eligibility ($x_{1,1}$), their field of study ($x_{1,2}$) and earnings after graduation (y), but missing their parental income (x_2). Eligible students tend to be from households with lower parental income and parental income is a critical confounder for the regression of policy and field of study $x_1 = \{x_{1,1}, x_{1,2}\}$ on y. Suppose that another study collects another representative sample of households about their household income (x_2) and college graduates' earnings (y), but the eligibility of the program is unknown because their field of study is unknown. We aim to recover the coefficient on $x_{1,1}$ in the long-regression of earnings (y) on the policy eligibility ($x_{1,1}$), controlling for both students' field of study ($x_{1,2}$) and households income (x_2).

Example 2.2 Consider a randomized evaluation of a policy x_1 on an outcome y. A referee suggests that another background information x_2 may be the mechanism that drives the policy impact of x_1 on y. Unfortunately, such information x_2 is not available in the study at hand. Nevertheless, administrative data on the same study area contains both y and x_2, but not x_1 which is the intervention introduced by the researcher. We aim to recover the coefficient on x_1 after controlling for x_2 to compare the coefficient on x_1 without controlling for x_2.

We first consider identification of $c'\beta$ pretending that $E(x_1 x_2')$ is known. In other words, the whole variance matrix $W \equiv E[zz']$ is known.

For every $s \in \mathcal{S}$, the vector of linear projection coefficients of y onto z_s,

$$\gamma_s = E[z_s z_s']^{-1} E[z_s y] = E[z_s z_s']^{-1} E[z_s z']\beta,$$

is identified. Consequently, we obtain L_s number of linear constraints for $\beta \in \mathbb{R}^L$ for each $s \in \mathcal{S}$,

$$E(z_s z')\beta = E(z_s z_s')\gamma_s \tag{2.1}$$

because γ_s is the vector of linear projection coefficients of z_s, and the CEF is linear. If the whole vector of covariates z is observed *somewhere* in the studies, then the linear constraints (2.1) are sufficient for identification.

PROPOSITION 2.1 If $z_1 \cup z_2 = z$ and $W = E[zz']$ is a known positive-definite matrix, then β is point-identified.

PROOF: By stacking the matrices of (2.1) over studies, we have

$$\begin{bmatrix} E[x_0 z'] \\ E[x_1 z'] \\ E[x_0 z'] \\ E[x_2 z'] \end{bmatrix} \beta = \begin{bmatrix} E[x_0 z_1']\gamma_1 \\ E[x_1 z_1']\gamma_1 \\ E[x_0 z_2']\gamma_2 \\ E[x_2 z_2']\gamma_2 \end{bmatrix}.$$

Let

$$J = \begin{bmatrix} I_{L_0} & O_{L_1} & O_{L_0} & O_{L_2} \\ O_{L_0} & I_{L_1} & O_{L_0} & O_{L_2} \\ O_{L_0} & O_{L_1} & O_{L_0} & I_{L_2} \end{bmatrix}$$

where I_k is $k \times k$ identity matrix and O_k is $k \times k$ zero matrix. Multiplying J from left, we obtain

$$J \begin{bmatrix} E[x_0 z'] \\ E[x_1 z'] \\ E[x_0 z'] \\ E[x_2 z'] \end{bmatrix} \beta = E[zz']\beta = J \begin{bmatrix} E[x_0 z_1']\gamma_1 \\ E[x_1 z_1']\gamma_1 \\ E[x_0 z_2']\gamma_2 \\ E[x_2 z_2']\gamma_2 \end{bmatrix}.$$

Hence

$$\beta = W^{-1} J \begin{bmatrix} E[x_0 z_1']\gamma_1 \\ E[x_1 z_1']\gamma_1 \\ E[x_0 z_2']\gamma_2 \\ E[x_2 z_2']\gamma_2 \end{bmatrix}.$$

Q.E.D.

Remark 2.1 The choice of J matrix is not unique and β is over-identified. For example, taking

$$\tilde{J} = \begin{bmatrix} O_{L_0} & O_{L_1} & I_{L_0} & O_{L_2} \\ O_{L_0} & I_{L_1} & O_{L_0} & O_{L_2} \\ O_{L_0} & O_{L_1} & O_{L_0} & I_{L_2} \end{bmatrix}$$

obtains another identification formula of

$$\beta = W^{-1} \tilde{J} \begin{bmatrix} E[x_0 z_1']\gamma_1 \\ E[x_1 z_1']\gamma_1 \\ E[x_0 z_2']\gamma_2 \\ E[x_2 z_2']\gamma_2 \end{bmatrix}.$$

Hence, the underlying model has a testable restriction.

The above proposition assumes the knowledge of W but the cross moment between x_1 and x_2 is usually unknown because x_1 and x_2 are not observed for the same unit.

When the covariance of x_1 and x_2 are unknown, $c'\beta$ becomes a set-identified object. Let

$$b \equiv J \begin{bmatrix} E[x_0 z_1']\gamma_1 \\ E[x_1 z_1']\gamma_1 \\ E[x_0 z_2']\gamma_2 \\ E[x_2 z_2']\gamma_2 \end{bmatrix}.$$

Specifically, the upper and lower bounds for $c'\beta$ is attained from the maximum and minimum of the following objective function:

$$c'W^{-1}b$$

subject to

$$W = \begin{bmatrix} \tilde{A} & B & C \\ B' & R_{11} & R_{12} \\ C' & R_{12}' & R_{22} \end{bmatrix}$$

where $\tilde{A} = E[x_1 x_1'], R_{11} = E[x_0 x_0'], R_{12} = E[x_0 x_2'], R_{22} = E[x_2 x_2'], B = E[x_0 x_1']$ and $C = E[x_1 x_2']$ and C is the only unknown matrix, $c, b \neq 0$ are known vector of length L. Note that

$$(c-b)'W^{-1}(c-b) = c'W^{-1}c + b'W^{-1}b - b'W^{-1}c - c'W^{-1}b$$

and hence

$$c'W^{-1}b = \left[(c-b)'W^{-1}(c-b) - (c'W^{-1}c + b'W^{-1}b)\right]/2.$$

As in the following lemma, $c'W^{-1}c$ is a convex function with respect to C.

Lemma 2.1 For any non-zero vector a and a positive definite matrix

$$W = \begin{bmatrix} \tilde{A} & B & C \\ B' & R_{11} & R_{12} \\ C' & R_{12}' & R_{22} \end{bmatrix},$$

$a'W^{-1}a$ is a convex function with respect to C.

Proof: Proof is in the Appendix. *Q.E.D.*

Consequently, $(c-b)'W^{-1}(c-b), c'W^{-1}c$ and $b'W^{-1}b$ are convex function of C, the objective function is difference in convex functions. Consequently, the objective function is not necessarily convex.

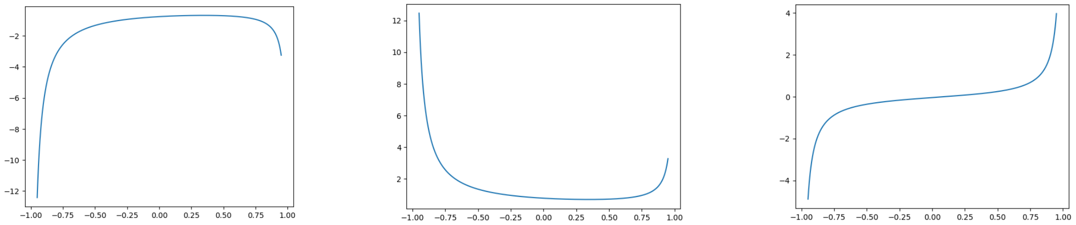
For an illustrative example, consider a scalar x_1 and x_2 case with empty x_0. Let

$$W = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}$$

where σ_1^2 and σ_2^2 are the variances of x_1 and x_2, and ρ is the correlation coefficient, which is the only unknown parameter. Figure 2.1 illustrate the objective function as a function of ρ in three different signs of $\beta = (\beta_0, \beta_1, \beta_2)$. As shown in the figure, the objective function can be either concave (Panel A), convex (Panel B), or non-convex (Panel C) in the value of ρ. An important remark of the illustration is that the objective function can have an informative upper or lower bound without restricting the unknown correlation coefficient ρ. Nevertheless, such an informative bound is not available in general.

As demonstrated in Figure 2.1, the objective function is not necessarily bounded. If ρ is not bounded, then the objective function $c'W^{-1}b$ is not necessarily bounded as W becomes singular when $\rho = 1$. As a result, the bounds are not informative without restricting W. Hence, we may use ρ as the sensitivity parameter to assess the sensitivity of the bounds against unknown correlation structure of two separate dataset.

In the earlier examples, we often have a particular few covariates that are relevant and not in common across two studies. If the number of non-common covariates is small, the non-convex optimization is straightforward with a brute-force search. For a larger



(A) $\beta_1 = -0.6, \beta_2 = -0.2$ (B) $\beta_1 = 0.6, \beta_2 = 0.2$ (C) $\beta_1 = 0.6, \beta_2 = -1.2$

FIGURE 2.1.— Plot of the objective function value as the function of ρ in illustrative numerical examples of different specifications over $-0.95 \leq \rho \leq 0.95$. All specifications are common in the distribution of covariates and the intercept $\beta_0 = 0.1$.

dimensional problem, we may resort to an optimization algorithm for a non-convex optimization. In particular, the objective function is a difference in convex functions and a DC (Difference of Convex functions) algorithm (Le Thi and Pham Dinh, 2018, for example) may be applicable.

3. ROBUSTNESS AGAINST COMMON OMITTED VARIABLES

In the aforementioned examples, all the relevant variables are observed somewhere in the study samples. Nevertheless, we may concern about omitted variables that are never observed in any of the study samples. For example, in Example 1.1, IQ score is not a perfect proxy of the unobserved ability and the residual ability net of the IQ score may still cause the omitted variable bias. The eligibility in Example 2.1 may be determined by an additional unobserved variable. In Example 2.2, x_1 is randomized, but x_2 may be correlated with an unobserved confounder.

An important example is the interaction term $x_1 x_2$ which is the unobserved omitted variable. For the purpose of mediation analysis, the interaction term plays a critical role in decomposing the direct and indirect effect of the randomized policy x_1, and the interaction term is not observed in any of the study samples.

Hence, we consider a vector of omitted variables from both studies. Now, the vector of L explanatory variables z is composed of the vector that is observed in some study and the other vector of the *omitted variables* w such that $z = [x_0', x_1', x_2', w']'$. Each study has its own vector of omitted variables $w_1 = [x_2', w']'$ and $w_2 = [x_1', w']'$.

With omitted variables w, the above point identification does not hold. Nevertheless, in addition to the linear constraints (2.1), there is a set of quadratic constraints for the variance of y in each study. For $\sigma_y^2 = Var(y)$, we have

$$\gamma_s' E(z_s z_s') \gamma_s \leq \beta' E(zz')\beta \leq \sigma_y^2. \tag{3.1}$$

Below, we exploit both equations (2.1) and (3.1) given the knowledge about γ_s and $E(x_s x_s')$. The unknown objects are β, $E(x_s w_s')$ and $E(w_s w_s')$. We consider constructing the bounds for the parameter of interest $\theta = c'\beta$ given the knowledge of $E(z_s w_s')$ and $E(w_s w_s')$ that are partially identified. Later, we aggregate the bounds for $c'\beta$ by intersecting them across $s = 1$ and $s = 2$. In the last step, we construct the bounds of $c'\beta$ by incorporating partial identification of $E(x_s w_s')$ and $E(w_s w_s')$.

The upper and lower bounds for θ are obtained by optimizing $c'\beta$ subject to (2.1) and (3.1). This optimization can be viewed as a series of linear optimizations with linear and quadratic constraints of the following form: for each $s \in \{1, 2\}$,

$$\begin{aligned} &\max / \min_{\beta \in \mathbb{R}^L} c'\beta, \\ \text{s.t. } \; &E(x_s z')\beta = E(z_s z_s')\gamma_s, \\ &\gamma_s' E(z_s z_s')\gamma_s \leq \beta' W \beta \leq \sigma_y^2. \end{aligned} \tag{3.2}$$

We illustrate the constraints in an illustration of scalar x_1, x_2 and w. Figure 3.2 illustrate the constraints as a cross section view along with the coefficient of x_2 that satisfies the linear constraint (2.1). As illustrated in the figure, we can ignore the lower bound of the quadratic inequality constraint because the optimum is attained when the upper bound of the elliptic inequality is binding.

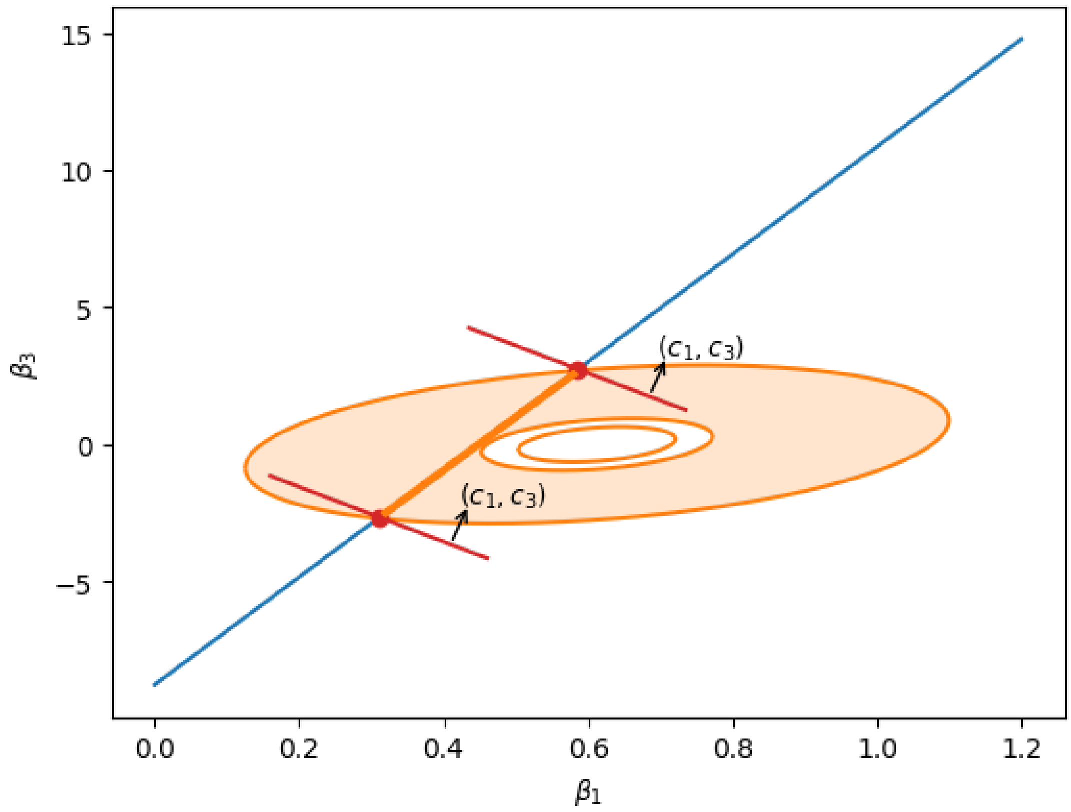


FIGURE 3.2.— A numerical illustration of two observed variables with coefficients (β_1, β_2) and an omitted variable with a coefficient β_3. The figure represents the two-dimensional (β_1, β_2) plot of the cross section view of the three dimensional space $(\beta_1, \beta_2, \beta_3)$ along the linear constraint that is solved for β_2. Given c (black arrows), the maximum and minimum of $c'\beta$ (red dots) are the intersections of the linear constraint (blue straight line) and the upper bound of the quadratic constraint (the outside orange ellipse). The identified set (bold orange line) is the interval between the red dots.

This problem (3.2) is abstracted into the following canonical form

$$\begin{aligned} &\max / \min_{\beta \in \mathbb{R}^L} c'\beta, \\ \text{s.t.}\ \ & A\beta = b, \\ & \ell \leq \beta' W \beta \leq u. \end{aligned} \tag{3.3}$$

where A is a $(\bar{L}, L)$ matrix, b is a vector of length $\bar{L}$ and W is a symmetric and positive semidefinite (L, L) matrix. The problem (3.3) has a closed-form solution as in the next lemma.

Lemma 3.1 Given W, the solutions to the optimization problem with respect to β in the canonical form (3.3) are given by

$$\beta = W^{-1}A'(AW^{-1}A')^{-1}b - \mu(W^{-1} - H)c, \tag{3.4}$$

where

$$\mu = \pm\sqrt{\frac{u - b'(AW^{-1}A')^{-1}b}{c'(W^{-1} - H)c}}, \tag{3.5}$$

$$H = W^{-1}A'(AW^{-1}A')^{-1}AW^{-1} \tag{3.6}$$

Proof: Noting that at the optimum the upper bound of the quadratic inequality constraint is binding, while the lower bound is not, we form the Lagrangean as

$$\mathcal{L} = c'\beta - \lambda'(A\beta - b) - \frac{1}{2\mu}(u - \beta' W\beta), \tag{3.7}$$

where $(\lambda, 1/2\mu) \in \mathbb{R}^{\bar{L}+1}$ are Lagrange multipliers. The first-order conditions in β give

$$\beta = \mu W^{-1}(A'\lambda - c). \tag{3.8}$$

The constraint $A\beta = b$ implies

$$\lambda = (AW^{-1}A')\left(AW^{-1}c + \frac{1}{\mu}b\right). \tag{3.9}$$

Combining (3.8) and (3.9) yields β in the lemma. To pin down μ, we plug in (3.8) and (3.9) into $\beta' W\beta = u$. It leads to (3.5). *Q.E.D.*

To apply the lemma, for each A, observe that AW^{-1} can be seen as the coefficient matrix of the linear projections of x_s onto z, which is, for each $s \in \{0, \mathcal{S}\}$,

$$AW^{-1} = \begin{pmatrix} I_{L_s\times L_s} & O_{L_s\times(L-L_s)} \end{pmatrix}. \tag{3.10}$$

Hence, we have

$$AW^{-1}A' = E(x_s x_s') \tag{3.11}$$

$$H = \begin{pmatrix} E(x_s x_s')^{-1} & O_{L_s \times (L-L_s)} \\ O_{(L-L_s)\times L_s} & O_{(L-L_s)\times(L-L_s)} \end{pmatrix}. \tag{3.12}$$

Consider the linear projection of w_s onto x_s,

$$w_s = \Pi^s x_s + v_s, \tag{3.13}$$

where $\Pi^s = E(w^s x_s')E(x_s x_s')^{-1}$. Then, by the inverse formula for a block matrix,

$$W^{-1} = \begin{pmatrix} E(x_s x_s')^{-1} + \Pi^{s\prime} E(v_s v_s{\prime})^{-1}\Pi^s & -\Pi^{s\prime}E(v_s v_s')^{-1} \\ -E(v_s v_s')^{-1}\Pi^s & E(v_s v_s')^{-1} \end{pmatrix} \tag{3.14}$$

Hence, we obtain

$$\begin{aligned} W^{-1} - H &= \begin{pmatrix} \Pi^{s\prime}E(v_s v_s')^{-1}\Pi^s & -\Pi^{s\prime}E(v_s v_s')^{-1} \\ -E(v_s v_s')^{-1}\Pi^s & E(v_s v_s')^{-1} \end{pmatrix} \\ &= \begin{pmatrix} \Pi^{s\prime} \\ I_{(L-L_s)\times(L-L_s)} \end{pmatrix} E(v_s v_s')^{-1} \begin{pmatrix} \Pi^s & I_{(L-L_s)\times(L-L_s)} \end{pmatrix} \end{aligned} \tag{3.15}$$

Putting altogether, Lemma 3.1 gives the following proposition:

Proposition 3.1 Suppose that $W = E(zz')$ is constrained in an identified set $IS_{zz'}$, then the identified set for $\theta = c'\beta$ is the convex interval $[\max_s \theta_\ell^s, \min_s \theta_u^s]$ where

$$\theta_\ell^s = c_s'\gamma_s - \eta(z_s)\sqrt{\max_{W\in IS_{zz'}} c'(W^{-1}-H)c}, \tag{3.16}$$

$$\theta_u^s = c_s'\gamma_s + \eta(z_s)\sqrt{\max_{W\in IS_{zz'}} c'(W^{-1}-H)c}, \tag{3.17}$$

where $W^{-1} - H$ is as shown in (3.15) and c_s is the subvector of c corresponding to the elements of x_s in z, Proof: The first terms in the right-hand sides of (3.16) and (3.17) follow by combining $b = E(x_s z')\gamma_s$, (3.10), and (3.11). Note also that the numerator in (3.5) equals to $\eta(z_s) = \sqrt{Var(y - L(y|x_s))}$, which is an identified quantity in study s. From the bounds for each s, we attain the stated bounds because

$$\begin{aligned}
&\max_{W \in IS_{zz'}} \min_s \theta_u^s \\
&\quad = \max_{W \in IS_{zz'}} \min_s \left(c_s' \gamma_s + \eta(z_s) \sqrt{c' W^{-1} c - c' H^s c} \right) \\
&\quad = \min_s \max_{W \in IS_{zz'}} \left(c_s' \gamma_s + \eta(z_s) \sqrt{c' W^{-1} c - c' H^s c} \right) \\
&\quad = \min_s \left(c_s' \gamma_s + \eta(z_s) \sqrt{\max_{c' W^{-1} c : W \in IS_{zz'}} c' W^{-1} c - c' H^s c} \right)
\end{aligned}$$

from $\eta(z_s) \geq 0$ for every s, and because

$$\begin{aligned}
&\min_{W \in IS_{zz'}} \max_s \theta_l^s \\
&\quad = \min_{W \in IS_{zz'}} \max_s \left(c_s' \gamma_s - \eta(z_s) \sqrt{c' W^{-1} c - c' H^s c} \right) \\
&\quad = \max_s \min_{W \in IS_{zz'}} \left(c_s' \gamma_s - \eta(z_s) \sqrt{c' W^{-1} c - c' H^s c} \right) \\
&\quad = \max_s \left(c_s' \gamma_s - \eta(z_s) \sqrt{\max_{W \in IS_{zz'}} c' W^{-1} c - c' H^s c} \right).
\end{aligned}$$

Q.E.D.

3.1. *Implementation of the bounds*

The above bounds $[\max_s \theta_l^s, \min_s \theta_u^s]$ for $c'\beta$ are subject to the identified set of the unknown matrix W. As shown in Proposition 3.1, we need to obtain the maximum value of $c' W^{-1} c$.

Below, we consider a scalar common omitted variable w. Note that the scalar unobservable can be a general structure given the additive model of the conditional expectation function. Its variance matrix is

$$W = \begin{bmatrix} \tilde{A} & B & C & D_1 \\ B' & R_{11} & R_{12} & D_2 \\ C' & R_{12}' & R_{22} & D_3 \\ D_1' & D_2' & D_3' & \tilde{r} \end{bmatrix},$$

and the matrix C, the vector $D = [D_1, D_2, D_3]$, and a scalar $\tilde{r}$ are the unknown parameters. We consider a sequential optimization of the objective function $c'Wc$. In particular, we fix C in the the upper-block submatrix

$$\bar{A} \equiv \begin{bmatrix} \tilde{A} & B & C \\ B' & R_{11} & R_{12} \\ C' & R_{12}' & R_{22} \end{bmatrix}$$

and maximizing

$$c' \begin{bmatrix} \bar{A} & D \\ D' & \tilde{r} \end{bmatrix}^{-1} c$$

over D and $\tilde{r}$, and search for the maximizing C by a brute-force search as in the case of no omitted variable.

Given the submatrix $\bar{A}$, $c'W^{-1}c$ is a convex function with respect to the elements that correspond D and $\tilde{r}$ as we show in the upcoming lemma below. Finding the maximum of a convex function over a convex constraint is known to have its solutions on the boundary of the constraint (Kubo et al., 2012, Chapter 12, for example). Hence, the optimization is nonetheless infeasible for a large number of the omitted variables because we need to list all the vertex of the convex constraints. Conversely, the sequential optimization is feasible when the first stage optimization is feasible.

Note that if $\bar{A}$ and the whole matrix W are non-singular and $\tilde{r} > 0$, we have

$$W^{-1} = \begin{bmatrix} \bar{A}^{-1} + \bar{A}^{-1}DD'\bar{A}^{-1}/k & -\bar{A}^{-1}D/k \\ -D'\bar{A}^{-1}/k & 1/k \end{bmatrix}.$$

where $k \equiv \tilde{r} - D'\bar{A}^{-1}D$. Since $\bar{A}$ is known, our objective is to find D and k that maximizes $c'W^{-1}c$. We parameterize the objective function by $\tilde{y} \equiv [d', k]'$ where $d = [d_{11}, d_{12}, \ldots, d_{1L_2}, d_{21}, d_{22}, \ldots, d_{L_1L_2}]'$ so that

$$D(d) = [d_{ij}]_{1 \leq i \leq L_1, 1 \leq j \leq L_2}$$

as

$$h(\tilde{y}) = c' \begin{bmatrix} A^{-1} + A^{-1}DD'A^{-1}/k & -A^{-1}D/k \\ -D'A^{-1}/k & 1/k \end{bmatrix} c.$$

This function h is a weakly convex function.

Lemma 3.2 The Hessian matrix of $h(\tilde{y})$ is positive semi-definite.

Note again that the convexity of the objective function implies that the optimization problem is the minimization of a concave function. Since the constraint set is convex, a solution exists on the boundary of the constraint set.

Hence, we propose a nested-loop optimization: We fix the correlation structure among the observed non-common covariates and find $\tilde{y}$ that maximizes the objective function value, and iterate over the restrictions on the correlation structure.

The inner loop optimization involves optimization over the boundary of the constraints. If all the constraints are linear, then an optimal solution exists at one of vertices of a polyhedron generated from the linear constraints. However, the set of feasible D and k is constrained by the elliptic constraint of $k = \tilde{R} - D'\tilde{A}^{-1}D$. Hence, we may approximate the elliptic constraint with a polyhedron (Ben-Tal and Nemirovski, 2001).

Below, we illustrate the optimization procedure when there is a scalar omitted variable w. To obtain an informative bound, one needs to specify three quantities: $Var(w)$, $Corr(x_1, x_2)$ and $Corr(x', w)$ where $x = [x_0, x_1, x_2]'$. Consider that one has the upper bound of $Var(w) \leq \bar{\sigma}_w^2$ and worst-case correlation $\bar{\rho}$ such that the absolute values of $Corr(x_1, x_2)$ and $Corr(x, w)$ fall below $\bar{\rho}$.

If one consider a specific omitted variable w, then one may have its summary statistics from other data sources that do not contain neither y nor (x_1, x_2). The variance bound for $V(w)$ may be constructed from the support of w when w is discrete. For example, if the omitted variable w is the cross term x_1x_2 and both x_1 and x_2 are binary, then the correlation of x_1 and x_2 through the cross moment $E[x_1x_2]$ characterizes the unknown part of the variance matrix.

Note that normalization $Var(w) = 1$ as in Masten and Poirier (2022) is possible and the objective function is convex in the remaining unobserved vector D from a similar argument as in Lemma 2.1. Given $\bar{\sigma}_w^2$, one may consider the sensitivity analysis with respect to the value of $\bar{\rho}$.

Note that given C of the covariance of x_1 and x_2, the unknown parameters are D of the covariance of x and w and a scalar $k = \sqrt{Var(w) - D'\tilde{A}D}$. Hence, the restrictions on $\{D, k\}$ is characterized by the following set of constraints: for each $x^i \in x$ and the corresponding correlation d_i of $Corr(x^i, w)$

$$-\bar{\rho}\sqrt{Var(x^i)\bar{\sigma}_w^2} \leq d_i \leq \bar{\rho}\sqrt{Var(x^i)\bar{\sigma}_w^2}$$

and

$$\sqrt{D'\tilde{A}D} \leq \sqrt{\bar{\sigma}_w^2 - k}$$

where the latter elliptic constraint can be approximated by a polyhedron following Ben-Tal and Nemirovski (2001). Once the constraint is represented by a polyhedron, we may elicit the vertices of the polyhedron by the double description method (Motzkin et al., 1953).

4. CONCLUSION

In this study, we propose the partial identification for the linear long regression coefficients from short regression coefficients by aggregating multiple sources. Specifically, we consider two studies that share the same outcome y but contain different key variables x_1 and x_2 separately. Our target parameter is the long regression coefficients $E[y|x_1, x_2]$ but we observe the short regressions $E[y|x_1]$ and $E[y|x_2]$ only.

We show that the target parameter is point identified when the covariance of x_1

and x_2 are known and the long regression regresses only on the variables that are observed in either of two study samples. Oftentimes, the covariance of x_1 and x_2 are unknown. Hence, we propose bounds for the target parameter by specifying the worst-case correlation coefficients.

Furthermore, we show that the target parameter is only partially identified when the covariance of x_1 and x_2 are known but the long regression contains an omitted variable that is never observed in two study samples. Specifically, we show that an elliptic constraint is critical for the partial identification. The elliptic constraint arises from the constraint on the explained variation of the long regression. We demonstrate that the data combination tightens the bounds as the bounds are constructed as the intersection bounds of each study.

There are a few problems left unresolved. First, the optimization of the objective function for relatively many omitted variables can be challenging. There are a few non-convex optimization problems proposed including the DC algorithm (Le Thi and Pham Dinh, 2018), but we avoid these optimization procedures by focusing on a low-dimensional problem that is typically concerned in Economics studies. Second, we focus on the restrictions up to second moments for the partial identification and exclude other higher-order moments and distributional information. It is possible that this information tightens the bounds. Nevertheless, we limit our focus on the linear projection and the first and second moments because they are relatively accessible than other higher-order moments. Third, we assume that the short regressions represent the same population as the long regression of interest. Considering data combinations from different populations complicates the analysis, and we defer the analysis for future research.

REFERENCES

Altonji, J. G., T. E. Elder, and C. R. Taber (2005): "Selection on Observed and Unobserved Variables: Assessing the Effectiveness of Catholic Schools," *Journal of Political Economy*, 113, 151–184.

Bellows, J. and E. Miguel (2006): "War and Institutions: New Evidence from Sierra Leone," *American Economic Review*, 96, 394–399.

Ben-Tal, A. and A. Nemirovski (2001): "On Polyhedral Approximations of the Second-Order Cone," *Mathematics of Operations Research*, 26, 193–205.

Cinelli, C. and C. Hazlett (2020): "Making Sense of Sensitivity: Extending Omitted Variable Bias," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 82, 39–67.

Clarke, K. A. (2009): "Return of the Phantom Menace: Omitted Variable Bias in Political Research," *Conflict Management and Peace Science*, 26, 46–66.

Cross, P. J. and C. F. Manski (2002): "Regressions, Short and Long," *Econometrica*, 70, 357–368.

Diegert, P., M. A. Masten, and A. Poirier (2022): "Assessing Omitted Variable Bias When the Controls Are Endogenous," https://arxiv.org/abs/2206.02303v4.

Frank, K. A. (2000): "Impact of a Confounding Variable on a Regression Coefficient," *Sociological Methods & Research*, 29, 147–194.

González, F. and E. Miguel (2015): "War and Local Collective Action in Sierra Leone: A Comment on the Use of Coefficient Stability Approaches," *Journal of Public Economics*, 128, 30–33.

Hosman, C. A., B. B. Hansen, and P. W. Holland (2010): "The Sensitivity of Linear Regression Coefficients' Confidence Limits to the Omission of a Confounder," *The Annals of Applied Statistics*, 4, 849–870.

Ichimura, H. and E. Martinez-Sanchis (2009): "Estimation and Inference of Models with Incomplete Data by Combining Two Data Sets," *Working Paper*.

Imbens, G. W. (2003): "Sensitivity to Exogeneity Assumptions in Program Evaluation," *American Economic Review*, 93, 126–132.

Krauth, B. (2016): "Bounding a Linear Causal Effect Using Relative Correlation Restrictions," *Journal of Econometric Methods*, 5, 117–141.

Kubo, M., T. Akihisa, and T. Matsui, eds. (2012): *Ouyou Suuri Keikaku Handbook (In Japanese)*, Tōkyō-to Shinjuku-ku: Asakura Shoten, fukyūban ed.

Le Thi, H. A. and T. Pham Dinh (2018): "DC Programming and DCA: Thirty Years of Developments," *Mathematical Programming*, 169, 5–68.

Leamer, E. E. (1978): *Specification Searches : Ad Hoc Inference with Nonexperimental Data*, New York: Wiley.

Masten, M. A. and A. Poirier (2022): "The Effect of Omitted Variables on the Sign of Regression Coefficients," https://arxiv.org/abs/2208.00552v2.

Mauro, R. (1990): "Understanding L.O.V.E. (Left out Variables Error): A Method for Estimating the Effects of Omitted Variables," *Psychological Bulletin*, 108, 314–329.

Molinari, F. and M. Peski (2006): "GENERALIZATION OF A RESULT ON "REGRESSIONS, SHORT AND LONG"," *Econometric Theory*, 22, 159–163.

Motzkin, T., H. Raiffa, GL. Thompson, and R. Thrall (1953): "The Double Description Method," in *Contributions to Theory of Games*, ed. by H. Kuhn and A. Tucker, Princeton, RI: Princeton University Press, vol. 2.

Murphy, K. M. and R. H. Topel (1990): "Efficiency Wages Reconsidered: Theory and Evidence," in *Advances in the Theory and Measurement of Unemployment*, ed. by Y. Weiss and G. Fishelson, London: Palgrave Macmillan UK, 204–240.

Oster, E. (2019): "Unobservable Selection and Coefficient Stability: Theory and Evidence," *Journal of Business & Economic Statistics*, 37, 187–204.

Pacini, D. (2019): "Two-Sample Least Squares Projection," *Econometric Reviews*, 38, 95–123.

APPENDIX A: PROOFS

Proof of Lemma 2.1:

$$W = \begin{bmatrix} A & B & C \\ B' & R_{11} & R_{12} \\ C' & R_{12}' & R_{22} \end{bmatrix},$$

If W is positive definite, then there is an upper triangular matrix

$$M = \begin{bmatrix} M_1 & M_2 & M_3 \\ 0 & M_4 & M_5 \\ 0 & 0 & M_6 \end{bmatrix}$$

such that

$$W = M'M.$$

Note that

$$\begin{bmatrix} M_1' & 0 & 0 \\ M_2' & M_4' & 0 \\ M_3' & M_5' & M_6' \end{bmatrix} \begin{bmatrix} M_1 & M_2 & M_3 \\ 0 & M_4 & M_5 \\ 0 & 0 & M_6 \end{bmatrix} = \begin{bmatrix} M_1'M_1 & M_1'M_2 & M_1'M_3 \\ M_2'M_1 & M_2'M_2 + M_4'M_4 & M_2'M_3 + M_4'M_5 \\ M_3'M_1 & M_3'M_2 + M_5'M_4 & M_3'M_3 + M_5'M_5 + M_6'M_6. \end{bmatrix}$$

or

$$W = \begin{bmatrix} A & B & M_1'M_3 \\ B' & R_{11} & R_{12} \\ M_3'M_1 & R_{12}' & R_{22} \end{bmatrix}$$

where M_1 is known as Cholesky decomposition of A matrix: $A = M_1'M_1$. Consequently,

M_3 is the only free parameter to optimize and the elements of C maps the elements of M_3 uniquely given the other parts of the matrix W. Let $m_3 = vec(M_3)$ with its length K. For each element m_{3j} of m_3, the derivative of $c'W^{-1}c$ is

$$\frac{\partial c'W^{-1}c}{\partial m_{3j}} = -c'W^{-1}\begin{bmatrix} 0 & 0 & M_1'M_{3j} \\ 0 & 0 & 0 \\ M_{3j}'M_1 & 0 & 0 \end{bmatrix} W^{-1}c$$

where M_{3j} is the matrix M_3 replacing m_{3j} element with 1 and all the other with 0. The cross-derivative with m_{3j} and m_{3i} is

$$\begin{aligned}\frac{\partial^2 c'W^{-1}c}{\partial m_{3j}\partial m_{3i}} =& c'W^{-1}\begin{bmatrix} 0 & 0 & M_1'M_{3i} \\ 0 & 0 & 0 \\ M_{3i}'M_1 & 0 & 0 \end{bmatrix} W^{-1}\begin{bmatrix} 0 & 0 & M_1'M_{3j} \\ 0 & 0 & 0 \\ M_{3j}'M_1 & 0 & 0 \end{bmatrix} W^{-1}c \\ &+ c'W^{-1}\begin{bmatrix} 0 & 0 & M_1'M_{3j} \\ 0 & 0 & 0 \\ M_{3j}'M_1 & 0 & 0 \end{bmatrix} W^{-1}\begin{bmatrix} 0 & 0 & M_1'M_{3i} \\ 0 & 0 & 0 \\ M_{3i}'M_1 & 0 & 0 \end{bmatrix} W^{-1}c.\end{aligned}$$

Let H be the Hessian matrix of $c'W^{-1}c$ with respect to m_3. For any non-zero vector $\tilde{c}$ of length K,

$$\tilde{c}'H\tilde{c} = \sum_{i=1}^{K}\sum_{j=1}^{K} \tilde{c}_i\tilde{c}_j \frac{\partial^2 c'W^{-1}c}{\partial m_{3j}\partial m_{3i}} = 2c'W^{-1}\begin{bmatrix} 0 & 0 & M_1'\tilde{M}_3 \\ 0 & 0 & 0 \\ \tilde{M}_3'M_1 & 0 & 0 \end{bmatrix} W^{-1}\begin{bmatrix} 0 & 0 & M_1'\tilde{M}_3 \\ 0 & 0 & 0 \\ \tilde{M}_3'M_1 & 0 & 0 \end{bmatrix} W^{-1}c$$

where $\tilde{M}_3 = \sum_{i=1}^{K} \tilde{c}_i M_{3i}$. Consequently, the matrix

$$W^{-1}\begin{bmatrix} 0 & 0 & M_1'\tilde{M}_3 \\ 0 & 0 & 0 \\ \tilde{M}_3'M_1 & 0 & 0 \end{bmatrix} W^{-1}\begin{bmatrix} 0 & 0 & M_1'\tilde{M}_3 \\ 0 & 0 & 0 \\ \tilde{M}_3'M_1 & 0 & 0 \end{bmatrix} W^{-1}$$

is symmetric and has the Cholesky decomposition of

$$\left[W^{-1} \begin{bmatrix} 0 & 0 & M_1'\tilde{M}_3 \\ 0 & 0 & 0 \\ \tilde{M}_3'M_1 & 0 & 0 \end{bmatrix} W^{-1/2} \right] \left[W^{-1/2} \begin{bmatrix} 0 & 0 & M_1'\tilde{M}_3 \\ 0 & 0 & 0 \\ \tilde{M}_3'M_1 & 0 & 0 \end{bmatrix} W^{-1} \right]$$

from the positive-definiteness of W. This proves that $\tilde{c}'H\tilde{c} \geq 0$; hence, $c'W^{-1}c$ is a (weakly) convex function with respect to the elements of C. *Q.E.D.*

Proof of Lemma 3.2: Consider taking derivatives with respect to $D = \{d_1, \ldots, d_{\bar{L}}\}$ and k. For each $i \in \{1, \ldots, \bar{D}\}$,

$$\frac{\partial h}{\partial d_i} = c' \begin{bmatrix} A^{-1}(D_iD' + DD_i')A^{-1}/k & -A^{-1}D_i/k \\ -D_i'A^{-1}/k & 0 \end{bmatrix} c.$$

where D_i is a $(\bar{L} \times 1)$ vector with ith element being 1 and 0 otherwise, and

$$\frac{\partial h}{\partial k} = c' \begin{bmatrix} -A^{-1}DD'/k^2 & A^{-1}D/k^2 \\ D'A^{-1}/k^2 & -1/k^2 \end{bmatrix} c.$$

Taking second derivatives, we obtain

$$\frac{\partial^2 h}{\partial d_i \partial d_j} = c' \begin{bmatrix} A^{-1}(D_iD_j' + D_jD_i')A^{-1}/k & 0 \\ 0 & 0 \end{bmatrix} c,$$

$$\frac{\partial^2 h}{\partial d_i \partial k} = c' \begin{bmatrix} -A^{-1}(D_iD' + DD_i')A^{-1}/k^2 & A^{-1}D_i/k^2 \\ D_i'A^{-1}/k^2 & 0 \end{bmatrix} c,$$

and

$$\frac{\partial^2 h}{\partial k^2} = c' \begin{bmatrix} 2A^{-1}DD'A^{-1}/k^3 & -2A^{-1}D/k^3 \\ -2D'A^{-1}/k^3 & 2/k^3 \end{bmatrix} c.$$

To check the definiteness of the Hessian, $\nabla^2 h$, take a quadratic form with an arbitrary non-zero vector $\tilde{y}$ of the same shape as y,

$$\begin{aligned}
\tilde{y}'\nabla^2 h\tilde{y} =& \sum_i\sum_j \tilde{d}_i\tilde{d}_j\frac{\partial^2 h}{\partial d_i\partial d_j} + \tilde{k}^2\frac{\partial^2 h}{\partial k^2} + 2\sum_i \tilde{d}_i\tilde{k}\frac{\partial^2 h}{\partial d_i\partial k} \\
=& 2c'\begin{bmatrix} A^{-1}(k^{-1}(\tilde{D}\tilde{D}') - \tilde{k}k^{-2}(\tilde{D}D' + D\tilde{D}') + \tilde{k}^2k^{-3}(DD'))A^{-1} & A^{-1}(\tilde{D}\tilde{k}/k^2 - D\tilde{k}^2/k^3) \\ (\tilde{D}'\tilde{k}/k^2 - D'\tilde{k}^2/k^3)A^{-1} & \tilde{k}^2/k^3 \end{bmatrix} c \\
=& 2c'\begin{bmatrix} A^{-1}k^{-1}((\tilde{D}\tilde{D}') - \tilde{k}k^{-1}(\tilde{D}D' + D\tilde{D}') + \tilde{k}^2k^{-2}(DD'))A^{-1} & A^{-1}k^{-1}(\tilde{k}/k)(\tilde{D} - D\tilde{k}/k) \\ k^{-1}(\tilde{k}/k)(\tilde{D}' - D'\tilde{k}/k)A^{-1} & k^{-1}(\tilde{k}/k)^2 \end{bmatrix} c \\
=& 2c'\begin{bmatrix} A^{-1}k^{-1}(\tilde{D} - \tilde{k}k^{-1}D)(\tilde{D}' - \tilde{k}k^{-1}D')A^{-1} & A^{-1}k^{-1}(\tilde{k}/k)(\tilde{D} - D\tilde{k}/k) \\ k^{-1}(\tilde{k}/k)(\tilde{D}' - D'\tilde{k}/k)A^{-1} & k^{-1}(\tilde{k}/k)^2 \end{bmatrix} c \\
=& 2c'\begin{bmatrix} k^{-1/2}A^{-1}(\tilde{D} - \tilde{k}k^{-1}D) & 0 \\ k^{-1/2}(\tilde{k}/k) & 0 \end{bmatrix}\begin{bmatrix} k^{-1/2}(\tilde{D}' - \tilde{k}k^{-1}D')A^{-1} & k^{-1/2}(\tilde{k}/k) \\ 0 & 0 \end{bmatrix} c \geq 0.
\end{aligned}$$

because the matrix inside is Cholesky decomposed and therefore is positive semi-definite. Consequently, the Hessian matrix is positive semi-definite. *Q.E.D.*