Estimation and Inference in a Panel Data Model of Social Interactions

Long Hong and Mikkel Sølvsten[*]

November 11, 2022

Abstract

This paper proposes a framework for estimation and inference in a panel data model of peer and spillover effects. We consider a linear-in-means model that may include social influences through three channels: outcomes, observed characteristics, and an unobserved individual-specific characteristic. To learn about the magnitude of effects in models that include some but not all of these social influences, the existing econometrics literature has adopted estimators based on least squares, maximum likelihood, and two-step instrumental variables ideas. Neither of these approaches will, in general, yield consistent estimators in a panel data context. Instead, we propose estimation and inference based on a novel objective function. We illustrate the ideas using the universal transcript data from the University of Wisconsin-Madison and explore the classroom peer effects during the semester when the university switched its learning mode to online during the recent pandemic. We show that the existing method estimates a positive and significant peer effect, while our method finds it to be close to zero and statistically insignificant.

Keywords: Social interactions, Spillover effects, Panel data, Non-linear regression, Cross-fitting, Leave-out estimation, Heteroskedasticity.
JEL codes: C18, C23, I21, J31.

[*] Hong: long.hong@wisc.edu, University of Wisconsin-Madison, USA. Sølvsten: miso@econ.au.dk, Aarhus University, Denmark. We acknowledge the Office of the Registrar at the University of Wisconsin-Madison for data access.

1 Introduction

Many empirical studies of social interactions document strong correlations between individuals' outcomes and those of their peers. In an influential paper, Manski (1993) delineates three plausible social mechanisms that may induce such dependence. These three channels are typically referred to as contextual spillover effects, which include influence generated by exogenous peer characteristics, endogenous peer effects, which encompass influence generated by peer outcomes, and correlated effects, which capture that individuals in the same reference group may behave similarly because they have similar individual characteristics or face a common environment. Credibly distinguishing between these potential explanations is a formidable challenge that has been the topic of immense literature in economics (see, e.g., De Paula, 2017, for a recent survey).

Even in the absence of correlated effects, Manski (1993) and Moffitt (2001) argue that it may be impossible to distinguish contextual spillovers from endogenous peer effects when social interactions take place in groups, as such patterns create a *reflection* problem. In adding nuance to this negative observation, Lee (2007) and Graham (2008) note that identification is possible under a homoskedasticity assumption and variation in group sizes, while Bramoullé, Djebbari and Fortin (2009) shows that the reflection problem disappears when social interactions are structured through a network. In such settings, consistent estimation can be facilitated by maximizing a Gaussian likelihood function (Lee, Liu and Lin, 2010) or by a two-step instrumental variables estimator (Kelejian and Prucha, 1998; Lee, 2003). The availability of panel data provides the potential to also allow for correlated effects that operate through an unobserved individual-specific characteristic. Assuming homoskedasticity and an absence of endogenous peer effects, Arcidiacono, Foster, Goodpaster and Kinsler (2012) shows that identification of contextual spillovers operating through the individual characteristic requires that there is mobility between reference groups and that consistent estimation can be facilitated by the use of non-linear least squares.[1]

In this paper, we build on the existing literature by relaxing two critical assumptions imposed by Arcidiacono et al. (2012). First, we allow for endogenous peer effects

[1] Mas and Moretti (2009) proposes a two-step estimator in a model that excludes endogenous peer effects and considers a so-called long panel data setting, where the number of observations per individual approaches infinity. Their two-step estimator is not consistent in short panels or in the presence of endogenous effects.

so that social influences may operate through all of the three channels outlined by Manski (1993). Second, we allow for heteroskedasticity in the unobserved errors so that identification (and consistent estimation) does not rest on a strong assumption regarding the error terms, which is often not motivated in applied work. The importance of relaxing these two assumptions was also stressed in the recent review paper Bramoullé, Djebbari and Fortin (2020).

The paper establishes three primary results. First, we show that mobility between peer groups and the corresponding network are the key conditions required for identification. In addition, we discuss how mobility-induced variation in peer group quality can serve as a sufficient condition for identification. These observations are natural extensions of the identification results presented in Bramoullé et al. (2009) and Arcidiacono et al. (2012). Second, we illustrate that estimation methods used in the existing literature can not guarantee consistent estimation of the model parameters. This observation leads us to propose a novel cross-fit objective function that can be used to construct consistent point estimators. Finally, we also provide accompanying standard errors that can be used to conduct valid inferences. The last contribution goes beyond most of the existing literature, which has focused on identification and point estimation but *not* on valid inference.

Apart from contributing to the economic literature on social influences and providing an econometric tool for applied economists to use in such settings, this paper also relates to a broader literature in econometrics that seeks to improve on estimation and inference methods in models that include a large number of parameters. So far, this literature has focused on models that are linear in the parameters, and a concise review can be found in Anatolyev (2019). The model considered in this paper not only includes a large number of parameters but also introduces non-linearity in these parameters. Therefore, the paper also seeks to expand this literature beyond the confines of linear models by drawing insights from it. In particular, our proposed framework borrows ideas from Hausman, Newey, Woutersen, Chao and Swanson (2012); Kline, Saggio and Sølvsten (2020); Anatolyev and Sølvsten (2020) and shows how these ideas can be adapted to a non-linear model.

We illustrate the ideas by estimating the classroom peer effect for all the freshmen at the University of Wisconsin - Madison, where a similar application is also used in Arcidiacono et al. (2012). In particular, we explore a special semester where the university switched the teaching mode to be entirely online due to the recent pandemic. We find that the peer effect is estimated to be positive and statistically significant using the existing method. However, our new method shows that the peer effect is close to zero and statistically insignificant.

The paper is organized as follows. Section 2 uses a simple motivating example to introduce our proposed estimator and contrast it with existing alternatives. Section 3 discusses the source of inconsistency in the non-linear least squares estimator and introduces our proposed alternative. Section 4 presents accompanying confidence sets, while Section 5 uses two canonical empirical examples to illustrate the use of our proposed inference method. Section 6 contains asymptotic theory, and Section 7 concludes. Some implementational details, proofs, and robustness checks are relegated to the Appendix.

2 Peer effects and non-linear regression

The primary theoretical contribution of the paper is to propose a novel cross-fit correction for the least squares estimator in a non-linear regression model where the number of regressors may be large. While the proposed approach applies broadly to such settings as described in Section 3, the current section introduces a motivating example which is the estimation of peer effects in a panel data model for wages. We return to this example in the empirical application of Section 5.

2.1 Contextual peer effects in unobservables

Consider the following framework. The outcome variable y_{it} denotes observed log wage for an individual i at time t. We are interested in the relationship between y_{it} and the average quality of individual i's contemporaneous group of peers. The peer group is observed by the researcher and is denoted by the index set $\mathcal{P}_{it} \subseteq \{1, \ldots, N\}$. The quality of each peer is unobserved but assumed to be captured by a measure of permanent ability α_i that also affects wages directly. Due to the possibility of endogenous sorting into peer groups, it is necessary to control for a vector of observed covariates w_{it}. As is common in applied practice, w_{it} may include a collection of group indicators. With additive separability, these considerations lead to a non-linear panel data regression

$$y_{it} = \alpha_i + \bar{\alpha}_{(i)t} \cdot \beta_0 + w_{it}'\gamma + \varepsilon_{it}, \qquad i = 1, \ldots, N,\ t = 1, \ldots, T_i, \tag{1}$$

where $\bar{\alpha}_{(i)t} = |\mathcal{P}_{it}|^{-1} \sum_{\iota \in \mathcal{P}_{it}} \alpha_\iota$ is the average of individual effects among i's peers. The object of interest is the coefficient on the average quality of the peers, $\beta_0 \in (-1, 1)$, while γ and $\alpha = (\alpha_1, \ldots, \alpha_n)'$ are non-random vectors of nuisance parameters.

Frameworks of the kind given above are widely used to determine the importance of peers in educational performance (e.g., Jackson and Bruegmann, 2009; Arcidiacono et al., 2012), wage settings (e.g., Lengermann, 2002; Cornelissen, Dustmann and Schönberg, 2017; Hong and Lattanzio, 2021), worker's productivity (e.g., Mas and Moretti, 2009; Guryan, Kroft and Notowidigdo, 2009; Brune, Chyn and Kerwin, 2020), firm revenues (e.g., Baum-Snow, Gendron-Carrier and Pavan, 2020).

The control variables w_{it} are included in the regression to ensure that no relevant confounders are excluded from the model so that the strict exogeneity of the peer group is satisfied. Letting $\mathcal{F}_i = \{w_{it}, \mathcal{P}_{it}\}_{t=1}^{T_i}$ collect individual i's observed history of peer groups and control variables, strict exogeneity can be formulated as

$$\mathbb{E}[\varepsilon_{it} \,|\, \mathcal{F}_i] = 0, \qquad i = 1, \ldots, N,\ t = 1, \ldots, T_i. \tag{2}$$

The set of control variables needed to ensure that (2) is satisfied depends on the specific context. We therefore have few further general comments about the choice of w_{it}.[2] Specifically, we consider the following specification for the controls.

$$w_{it}'\gamma = \psi_{j(i,t)} + \lambda_t + c_{it}'\gamma_c, \tag{3}$$

where $\psi_{j(i,t)}$ is the location effect (e.g., classroom or firm), λ_t is the time effect, and c_{it} is the observed time-varying individual characteristics. The inclusion of $\psi_{j(i,t)}$ controls for endogenous selection into firms or classrooms as in the seminal specification introduced by Abowd, Kramarz and Margolis (1999), which is originally used in wage

[2] When peers are completely randomly assigned, there is typically no need for any control variables. However, in many cases, random assignment is done conditional on a set of observed characteristics (e.g., Guryan et al., 2009), in which case it is most often necessary to include those characteristics in the model. With observational data, a judicious choice of control variables is typically required. Still, when interactions occur in groups, it may often be sensible to include a group fixed effect in $w_{it}'\gamma$ to further address the issue of correlated effects.

regression but has been widely adopted in many other settings.

In current empirical practice, estimation of (1) is commonly carried out by the use of (non-linear) least squares. Consistency of the resulting estimator for β_0 was established by Arcidiacono et al. (2012) in a setting with no control variables and serially uncorrelated, homoscedastic error terms. To shed light on the role played by these assumptions, Section 3 discuss why consistency of least squares fails when the error terms are not serially uncorrelated and homoscedastic. Given that the error covariance structure is rarely so well-behaved, we propose a cross-fit correction to the least squares estimator that allows for some dependence and unrestricted heteroscedasticity. An implication of our negative result regarding least squares and the structure of our proposed estimator is that researchers need to be explicit about their assumptions on the error variance structure when considering their choice of point estimator.

The regression structure of (1) makes the interpretation of β_0's magnitude and sign canonical: β_0 captures a return to peer quality in the sense that a one unit increase in average peer quality corresponds to a $\beta_0 \cdot 100\%$ increase in wages (on average). However, as peer quality is unobserved, the meaning of a one unit increase is ambiguous, so it is important to supplement any estimate of β_0 with a summary of possible changes in peer quality. Towards this end, we adapt the proposal in Kline et al. (2020) to provide an estimator of the overall variance in average peer group quality. We thereby facilitate that our proposed estimator of β_0 can be related to a standard deviation increase in average peer quality – a common way of grounding the interpretation of magnitudes in applied research. In practice, the mechanism through which peers affect outcomes is as interesting as the magnitude of the effect. Plausible mechanisms include knowledge spillover (Nix, 2020), peer pressure (Mas and Moretti, 2009), and promotion competition (Bianchi, Bovini, Li, Paradisi and Powell, 2021). The focus of this paper is the statistical problems of estimation and inference, so we will not delve further into the specific mechanisms that may drive the magnitude and sign of β_0.

2.2 Identifying variation and point estimation

Identification of β_0 requires variation in average peer group quality that cannot be predicted by the linear part of (1). In particular, it is necessary for identification that some individuals have a time-varying group of peers or equivalently that there is *mobility* between the observed peer groups. This requirement can, of course, be verified by inspecting the data at hand. Additionally, the mobility between peer groups must also induce variability in the average peer quality, $\bar{\alpha}_{(i)t} = |\mathcal{P}_{it}|^{-1} \sum_{\iota \in \mathcal{P}_{it}} \alpha_\iota$. Whether such variability is present cannot be verified ex-ante as the vector of individual heterogeneity α is unobserved. Instead, it must be maintained as an identifying assumption. In the extreme case of perfect homogeneity (identical entries in α), identification is bound to fail. When there is some individual heterogeneity and labor market frictions (e.g., Mortensen and Pissarides, 1994; Postel–Vinay and Robin, 2002) prevent perfect sorting, it is reasonable to expect that observed mobility must induce some identifying variation in the unobserved average peer quality.

To illustrate that the mobility-induced variation in average peer quality can lead to point identification, we consider a special case of (1). This special case also serves to highlight how the least squares estimator rely on homoskedasticity while our proposed cross-fit estimator does not. Suppose now that $w_{it}'\gamma$ is equal to $\psi_{j(i,t)}$, which was introduced in (3), that the time horizons T_i are all equal to two, that the error terms are independent across time and individuals, and that the data is generated in triplets of individuals where the first individual *stays* with the same employer for both periods, while the other two individuals *move* between two triplet-specific firms as depicted in Figure 1. The data for a generic triplet is then governed by the following

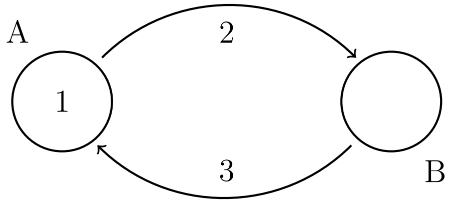

Figure 1: Depiction of three individuals (denoted 1, 2, and 3) and their mobility among two firms (denoted A and B). In the first period, individuals 1 and 2 are peers, while individuals 1 and 3 are peers in the second period. In both periods, firm A has two employees and firm B has one employee.

six equations

$$\begin{aligned} y_{11} &= \alpha_1 + \alpha_2\beta_0 + \psi_A + \varepsilon_{11}, & y_{12} &= \alpha_1 + \alpha_3\beta_0 + \psi_A + \varepsilon_{12}, \\ y_{21} &= \alpha_2 + \alpha_1\beta_0 + \psi_A + \varepsilon_{21}, & y_{22} &= \alpha_2 + \psi_B + \varepsilon_{22}, \\ y_{31} &= \alpha_3 + \psi_B + \varepsilon_{31}, & y_{32} &= \alpha_3 + \alpha_1\beta_0 + \psi_A + \varepsilon_{32}. \end{aligned}$$

The only part of this data that contains information about β_0 is a first difference for the stayer, $\mathcal{Y} = y_{12} - y_{11}$, and two differences involving both the movers, $\mathcal{X} = y_{32} - y_{21}$ and $\mathcal{Z} = y_{31} - y_{22}$. We can view $\mathcal{Y}$ as an outcome in a regression model that includes the unobserved regressor $\alpha_3 - \alpha_2$,

$$\mathcal{Y} = (\alpha_3 - \alpha_2)\beta_0 + (\varepsilon_{12} - \varepsilon_{11}),$$

while $\mathcal{X}$ and $\mathcal{Z}$ are independent noisy measurements of the unobserved regressor,

$$\mathcal{X} = \alpha_3 - \alpha_2 + (\varepsilon_{32} - \varepsilon_{21}) \quad \text{and} \quad \mathcal{Z} = \alpha_3 - \alpha_2 + (\varepsilon_{31} - \varepsilon_{22}).$$

Is it therefore immediate that a necessary and sufficient condition for point identification of β_0 is that $\alpha_3 - \alpha_2$ is not zero across all the triplets in the data. Unless the data is segregated into homogenous groups, such identifying variation will be present.

A particularly simple estimator of β_0 can be constructed from $(\mathcal{Y}, \mathcal{X}, \mathcal{Z})$ by averaging the two instrumental variables estimators that let $\mathcal{X}$ and $\mathcal{Z}$ take turns as noisy regressor and instrument. Equivalently, this estimator is the sample analog of the moment condition

$$\beta_0 = \frac{1}{2}\frac{\mathbb{E}[\mathcal{Z}\mathcal{Y}]}{\mathbb{E}[\mathcal{Z}\mathcal{X}]} + \frac{1}{2}\frac{\mathbb{E}[\mathcal{X}\mathcal{Y}]}{\mathbb{E}[\mathcal{X}\mathcal{Z}]}. \tag{4}$$

The resulting simple estimator is the cross-fit estimator introduced in the next Section.

To highlight the differences and similarities between the cross-fit and least squares estimators, we first re-express the denominator of (4) so that

$$\beta_0 = \frac{\mathbb{E}[(\mathcal{Z} + \mathcal{X})\mathcal{Y}]}{\mathbb{E}[\mathcal{Z}^2 + \mathcal{X}^2] - \mathbb{E}[(\mathcal{Z} - \mathcal{X})^2]}. \tag{5}$$

If the error terms are homoskedastic so that the unexplained variance in $\mathcal{Y}$ is the same as in $\mathcal{X}$ and $\mathcal{Z}$, we can alternatively express the variance $\mathbb{E}[(\mathcal{Z} - \mathcal{X})^2]$ appearing in the denominator of (5) as $4\sigma^2(\beta_0)$ where $\sigma^2(\beta)$ is a variance function that draws on both the movers and the stayer:

$$\sigma^2(\beta) = \frac{\mathbb{E}[(\mathcal{Z} - \mathcal{X})^2 + (\mathcal{Y} - \mathcal{X}\beta)^2 + (\mathcal{Y} - \mathcal{Z}\beta)^2]}{4(2 + \beta^2)}. \tag{6}$$

The non-linear least squares estimator minimizes the sample analog of (6). Additionally, we can describe the least squares estimator as a solution to the sample analog of a first order condition for minimization of (6), which looks exactly like (5), except that it uses $4\sigma^2(\beta)$ instead of $\mathbb{E}[(\mathcal{Z}-\mathcal{X})^2]$:

$$\beta = \frac{\mathbb{E}[(\mathcal{Z}+\mathcal{X})\mathcal{Y}]}{\mathbb{E}[\mathcal{Z}^2+\mathcal{X}^2]-4\sigma^2(\beta)}. \tag{7}$$

In the presence of heteroskedasticity, it is problematic to rely on an estimator that solves the sample analog of (7) as $4\sigma^2(\beta_0)$ will then be different from $\mathbb{E}[(\mathcal{Z}-\mathcal{X})^2]$ and this will in turn lead to an inconsistent estimator. In that case, the least squares estimator will be amplified (attenuated) relative to the truth if $\mathcal{Y}$ has higher (lower) unexplained variance than $\mathcal{X}$ and $\mathcal{Z}$. In this example, a natural assertion would be that the log-wage difference for the single stayer, $\mathcal{Y}$, has lower unexplained variance than the two log-wage differences involving the two movers, $\mathcal{X}$ and $\mathcal{Z}$. If that assertion holds true, the least squares estimator will understate the magnitude of the peer effects.

Remark 1. It may not be immediately obvious to every reader that the sample analogs of the simple expressions in (4)–(7) are special cases of the general formulas introduced in Section 3. Derivations that connect the two are provided in Appendix B where we also derive the bias direction for the least squares estimator as discussed above and show that $(\mathcal{Y},\mathcal{X},\mathcal{Z})$ contains all information about β_0.

3 Estimation

This section starts by characterizing the source of inconsistency in the least squares estimator when applied to a generic regression model with multiplicative non-linearity. Described at a high level, the source of inconsistency is that the least squares objective function is not minimized near the truth or equivalently that the gradient of the objective does not have a zero near β_0. Using this observation as a starting point, the section then proposes a new estimator which sets a recentered gradient of the least squares objective function equal to zero.[3]

[3] As discussed further below, consistency also requires that there is sufficient identifying variation in average peer quality.

3.1 Framework

We now suppress the multiple subscripts that were used to facilitate an economic discussion of the peer effects example introduced in Section 2. We therefore consider a regression model with multiplicative non-linearity of the form

$$y_\ell = x_\ell'\delta + a_\ell'\delta \cdot \beta_0 + \varepsilon_\ell, \qquad \ell = 1, \ldots, n. \tag{8}$$

Here x_ℓ and a_ℓ are observed K-dimensional vectors, δ is a vector of nuisance parameters, and β_0 remains the object of interest. In a peer effects setting, the vector a_ℓ is a function of the peer group that observation ℓ belongs to and thus dependent across ℓ. To encompass this example, we therefore do not impose restrictions on the dependence in x_ℓ and a_ℓ across ℓ. Instead, we conduct the analysis conditional on the regressors, $A = (a_1, \ldots, a_n)'$ and $X = (x_1, \ldots, x_n)'$, so that a_ℓ (and x_ℓ) may be arbitrarily dependent across observations.

The primary maintained assumptions are strict exogeneity, compactness of the parameter space for β_0, and a collection of full rank conditions.

Assumption 1. *(i)* $\mathbb{E}[\varepsilon_\ell \,|\, X, A] = 0$ *for all* ℓ *and* $\operatorname{range}(X)$ *contains the constant vectors, (ii)* $\beta_0 \in$ *interior*$(\mathcal{B})$ *where* $\mathcal{B} \subseteq \mathbb{R}$ *is compact, (iii)* $(X + A\beta, A\delta)$ *has full rank for any* $\beta \in \mathcal{B}$*, (iv)* $\max_\ell \mathbb{E}[\varepsilon_\ell^4 \,|\, X, A] < C$ *for some* $C < \infty$ *not depending on* n*.*

Part (i) is a strict exogeneity condition, part (ii) restricts the true β_0 to be in the interior of a compact set $\mathcal{B}$ as is standard for non-linear models, and part (iii) is a collection of full rank conditions on implied matrices of regressors. Part (iii) encapsulates two restrictions on the design. The first restriction excludes multicollinearity among the entries in $x_\ell + a_\ell\beta$ for any β in the parameter space and this condition ensures invertibility of the design matrix for estimation of δ when β_0 is equal to β: $S(\beta) = \sum_{\ell=1}^n (x_\ell + a_\ell\beta)(x_\ell + a_\ell\beta)'$. The second restriction is that the "unobserved regressor" $a_\ell'\delta$ contains identifying variation, i.e., that $a_\ell'\delta$ varies in ways that are not fully captured by a linear combination of $x_\ell + a_\ell\beta$ for any $\beta \in \mathcal{B}$. In the context of peer effects models, this part of Assumption 1 was discussed in Section 2.2 and requires that the sample contains variation in peer group quality that is not completely explained by the control variables. If a researcher is concerned about imposing this identifying restriction, it is possible to conduct inference by test inversion as opposed to by parameter estimation (see also Section 4). Part (iv) is a standard regularity condition.

Remark 2. In the peer effects setting considered in Section 2, the vector x_ℓ contains individual indicators and control variables while the vector a_ℓ contain indicators for peers divided by the peer group size.[4] In this setting, it is often natural to consider $\mathcal{B} \subset (-1, 1)$. This parameter space restricts the impact of the average peer quality to be smaller in magnitude than the individuals own effect α_i. With this restriction on $\mathcal{B}$, the design matrix $S(\beta)$ has full rank as long as $S(0)$ has full rank and it is therefore easy to verify or impose the part of Assumption 1, part (iii), that does not involve the unobserved regressor $a_\ell'\delta$.

3.2 Inconsistency of least squares

The least squares estimator applied to (8) yields the following estimator of β_0:

$$\hat{\beta}^{\text{LS}} = \arg\min_{\beta \in \mathcal{B}} \min_{\delta \in \mathbb{R}^k} \sum_{\ell=1}^{n} \left(y_\ell - x_\ell'\delta - a_\ell'\delta \cdot \beta \right)^2.$$

To give a representation of $\hat{\beta}^{\text{LS}}$ that is more amenable to analysis and intuition, we eliminate the nuisance vector δ using the blockwise matrix inversion formula that underpins the Frisch–Waugh–Lovell theorem. To do so, we define the entries of the matrix that residualizes against the regressor $x_\ell + a_\ell\beta$ as $M_{\ell k}(\beta) = \mathbf{1}\{\ell = k\} - (x_\ell + a_\ell\beta)'S(\beta)^{-1}(x_k + a_k\beta)'$. We can then represent $\hat{\beta}^{\text{LS}}$ as the solution to a minimization problem that does not involve δ:

$$\hat{\beta}^{\text{LS}} = \arg\min_{\beta \in \mathcal{B}} \hat{Q}_n(\beta) \qquad \text{where } \hat{Q}_n(\beta) = \sum_{\ell=1}^{n} \sum_{k=1}^{n} M_{\ell k}(\beta) y_\ell y_k.$$

The representation of the least squares estimator as a minimizer of the objective function $\hat{Q}_n$ implies that an almost necessary condition for consistency of $\hat{\beta}^{\text{LS}}$ is that the population analog $Q_n(\beta) = \mathbb{E}[\hat{Q}_n(\beta) \mid X, A]$ has a unique minimum at β_0. However, even under the strict exogeneity imposed in Assumption 1, we can determine this expectation to be composed of two terms where only the first has a unique minimum at β_0. To illustrate this point, let $\sigma_{\ell k} = \mathbb{E}[\varepsilon_\ell \varepsilon_k \mid X, A]$ be the covariance between the ℓ-th and the k-th error terms and define the part of $a_\ell'\delta$ that provides

[4] Additionally, a_ℓ is appended with a vector of zeroes in place of the control variables so that x_ℓ and a_ℓ are both K-dimensional vectors.

identifying variation when $\beta_0 = \beta$ as $\tilde{a}_\ell(\beta)'\delta$ where $\tilde{a}_\ell(\beta) = \sum_{k=1}^n M_{\ell k}(\beta)a_k'$. We then have

$$Q_n(\beta) = (\beta - \beta_0)^2 \sum_{\ell=1}^n (\tilde{a}_\ell(\beta)'\delta)^2 + \sum_{\ell=1}^n \sum_{k=1}^n M_{\ell k}(\beta)\sigma_{\ell k}.$$

The full rank restrictions of Assumption 1 imply that $\sum_{\ell=1}^n (\tilde{a}_\ell(\beta)'\delta)^2 > 0$ for all $\beta \in \boldsymbol{\mathcal{B}}$ so that the first part of Q_n is uniquely minimized at β_0. However, the part of Q_n that involves the error covariances is not, in general, minimized at the truth. The presence of the second part will therefore lead to inconsistency of the least squares estimator except in special cases. We therefore propose, in the next subsection, an alternative estimator based on a bias-corrected gradient of the objective function $\hat{Q}_n$.

Before proceeding to our proposed estimator, it is useful to highlight why least squares remain consistent with serially uncorrelated and homoscedastic error terms. In this case, we have that $\sigma_{\ell k} = \sigma^2 \mathbf{1}\{\ell = k\}$. This property implies that the second part of Q_n simplifies substantially. In fact, this second part becomes independent of β as the matrix function $M = (M_{\ell k})_{\ell,k}$ is a projection onto a linear space of dimension $n - K$, which in turn yields that the sum of the diagonal elements of $M(\beta)$ is $n - K$. We therefore have

$$\sum_{\ell=1}^n \sum_{k=1}^n M_{\ell k}(\beta)\sigma_{\ell k} = \sigma^2 \sum_{\ell=1}^n M_{\ell\ell}(\beta) = \sigma^2(n - K),$$

so that Q_n is uniquely minimized at β_0 when the error terms are serially uncorrelated and homoscedastic. In the special case of the peer effects model (1) without additional control variables w_{it}, this observation was also made by Arcidiacono et al. (2012).

Remark 3. Since the least squares estimator has seen widespread use in peer effects models, it would have been empirically relevant if it was possible to provide definite insights about the direction or sign of the bias in the least squares estimator under serially correlated or heteroskedastic errors. As the regression model in (8) involves an unobserved regressor $a_\ell'\delta$, which is itself estimated from the data, it might have been tempting to apply standard measurement error logic and conjecture that the least squares estimator is attenuated, i.e., biased towards zero. Unfortunately, as we show below the bias in the least squares estimator can have either sign and need not be towards zero. We therefore caution that least squares estimates that are available in the literature may differ from the underlying truth in systematic but unknown directions.

3.3 Cross-fit correction to least squares

Our proposed estimator relies on the standard cross-sectional assumption of conditionally independent errors in model (8). While such an assumption places restrictions on the patterns of dependence that can be allowed for in the data, it does not rule out dependence across observations at a lower level of aggregation as discussed next.

Assumption 2. *Conditional on X and A, $\{\varepsilon_\ell\}_{\ell=1}^n$ are jointly independent.*

Assumption 2 implies that the error variances are of the form $\sigma_{\ell k} = \sigma_\ell^2 \mathbf{1}\{\ell = k\}$. Specifically, we allow for heteroscedasticity in the non-linear regression model of (8).

Remark 4. In the peer effects model of equation (1), there are reasons to be wary of assuming independence among the error terms. For example, it seems reasonable to allow for wage errors to be serially dependent within a particular employment spell. Such dependence can be allowed for by writing down a model as in (1), collapsing the data to the level of employment spells, and then considering the resulting version of model (8) for the collapsed data. Section 5 further illustrates this approach and the biases that can arise from ignoring serial dependence in the micro data. We thereby highlight that in choosing the particular point estimator used (through the level at which the data is collapsed), the researcher needs to take into account the dependence structure of the error terms.

To introduce our proposed estimator, it is useful to describe the inconsistency of least squares in terms of derivatives of the objective functions introduced previously. Viewed through this lens, the least squares estimator is a zero of the sample moment function $\hat{m}_n = \nabla_\beta \hat{Q}_n$, and the source of inconsistency in least squares is that the population analog $m_n = \nabla_\beta Q_n$ is not equal to zero at β_0. Under Assumption 2, the gradient m_n at β_0 deviates from zero by

$$m_n(\beta_0) = \sum_{\ell=1}^{n} \nabla_\beta M_{\ell\ell}(\beta_0)\sigma_\ell^2. \tag{9}$$

Our proposed estimator is a zero of a sample moment function which is comprised of the difference between $\hat{m}_n$ and an estimator of the part that leads to the non-zero expectation in (9). We construct this sample moment using cross-fit or leave-one-out estimators of the individual error variances:

$$\hat{\sigma}_\ell^2(\beta) = \frac{y_\ell \hat{\varepsilon}_\ell(\beta)}{M_{\ell\ell}(\beta)} \tag{10}$$

where $\hat{\varepsilon}_\ell(\beta) = y_\ell - (x_\ell + a_\ell\beta)'\hat{\delta}^{\mathrm{LS}}(\beta)$ is the regression residual at β and $\hat{\delta}^{\mathrm{LS}}(\beta) = S(\beta)^{-1}\sum_{\ell=1}^n (x_\ell + a_\ell\beta)y_\ell$ is the corresponding least squares estimator of δ. That $\hat{\sigma}_\ell^2$ is a leave-one-out estimator follows from the equivalent representation

$$\hat{\sigma}_\ell^2(\beta) = y_\ell \left(y_\ell - (x_\ell + a_\ell\beta)'\hat{\delta}_{(\ell)}^{\mathrm{LS}}(\beta) \right),$$

in which $\hat{\delta}_{(\ell)}^{\mathrm{LS}}$ is the least squares estimator of δ applied to the sample that excludes the ℓ-th observation: $\hat{\delta}_{(\ell)}^{\mathrm{LS}}(\beta) = \big(\sum_{k\neq\ell}(x_k + a_k\beta)(x_k + a_k\beta)'\big)^{-1}\sum_{k\neq\ell}(x_k + a_k\beta)y_k$.

We use the leave-one-out individual error variance estimators to recenter the moment function $\hat{m}_n$. This leads to our proposed estimator

$$\hat{\beta}^{\mathrm{CF}} = \underset{\beta\in\mathcal{B}}{\arg\operatorname{zero}}\ \hat{m}_n^{\mathrm{CF}}(\beta) \qquad \text{where } \hat{m}_n^{\mathrm{CF}}(\beta) = \hat{m}_n(\beta) - \sum_{\ell=1}^n \nabla_\beta M_{\ell\ell}(\beta)\hat{\sigma}_\ell^2(\beta).$$

Because the cross-fit estimators $\{\hat{\sigma}_\ell^2(\beta_0)\}_{\ell=1}^n$ are *unbiased* for their respective error variances, the recentered moment function $\hat{m}_n^{\mathrm{CF}}$ has a mean of zero at β_0. Such a property is essential when establishing consistency of the resulting estimator $\hat{\beta}^{\mathrm{CF}}$.

Remark 5. The cross-fit variance estimators $\{\hat{\sigma}_\ell^2\}_{\ell=1}^n$ that underpins the recentered moment function $\hat{m}_n^{\mathrm{CF}}$ has previously been used in the context of linear regression to bias-correct non-linear functions of the least squares estimator or estimators of its variance: Kline et al. (2020); Anatolyev and Sølvsten (2020); Matsushita and Otsu (2019); Mikusheva and Sun (2020); Jochmans (2020). The use here is starkly different as we consider a non-linear regression model and cross-fitting is being used to bias-correct the least squares estimator itself.

Remark 6. There is a long tradition in econometrics of bias-correcting objective functions (instead of their gradients) to in an attempt at ensuring that their population counterparts are minimized at the true value for the parameter of interest (e.g., Han and Phillips, 2006; Hausman et al., 2012). Translating that approach to the current context would suggest that we considered a penalized objective function of the form

$$\hat{Q}_n^{\mathrm{CF}}(\beta) = \hat{Q}_n(\beta) - \textstyle\sum_{\ell=1}^n M_{\ell\ell}(\beta)\hat{\sigma}_\ell^2(\beta).$$

However, in contrast to the recentered sample moment $\hat{m}_n^{\mathrm{CF}}$, this objective function can not yield a consistent estimator of β_0 as $\hat{Q}_n^{\mathrm{CF}}(\beta)$ is zero for any value of β.

4 Inference

Hypothesis tests regarding the value of β_0 or a confidence interval for β_0 can be constructed using a Wald approach based on a normal approximation to the distribution of $\hat{\beta}^{\mathrm{CF}}$. In this case, inference requires an estimator for the variance $V_n(\beta) = \mathbb{V}\big[\hat{m}_n^{\mathrm{CF}}(\beta) \mid X, A\big]$. This section introduces our proposed variance estimator $\hat{V}_n(\beta)$ and briefly provides its use for inference using a Wald approach.

4.1 Variance estimator

In order to introduce and motivate our proposed variance estimator $\hat{V}_n$, it is useful first to give two separate U-statistic representations of $\hat{m}_n^{\mathrm{CF}}$. The first representation is a symmetric one in the sense that we write $\hat{m}_n^{\mathrm{CF}} = \sum_{\ell=1}^n \sum_{k\neq\ell} U_{\ell k}^{\mathrm{S}} y_\ell y_k$ and the order of the subscripts on the kernel function $U_{\ell k}^{\mathrm{S}}$ does not matter. This representation immediately follows from the definition of $\hat{Q}_n$ and the formulation of $\{\hat{\sigma}_\ell^2\}_{\ell=1}^n$ given in (10), which yields

$$U_{\ell k}^{\mathrm{S}}(\beta) = \nabla_\beta M_{\ell k}(\beta) - M_{\ell k}(\beta)\left(\nabla_\beta \log M_{\ell\ell}(\beta) + \nabla_\beta \log M_{kk}(\beta)\right)/2.$$

The second representation is asymmetric, i.e., $\hat{m}_n^{\mathrm{CF}} = \sum_{\ell=1}^n \sum_{k\neq\ell} U_{\ell k}^{\mathrm{A}} y_\ell y_k$ where $U_{\ell k}^{\mathrm{A}} \neq U_{k\ell}^{\mathrm{A}}$. To define $U_{\ell k}^{\mathrm{A}}$ and connect the two representations, we will rely on a small amount of matrix algebra. The projection M is idempotent, $M = M^2$, and differentiating each entry of this identity therefore yields that $\nabla_\beta M = M(\nabla_\beta M) + (\nabla_\beta M)M$. Because the derived matrix identity relates $\nabla_\beta M_{\ell k}$ to the two sums, $\sum_{m=1}^n M_{\ell m}\nabla_\beta M_{mk}$ and $\sum_{m=1}^n M_{km}\nabla_\beta M_{m\ell}$, we can therefore decompose $U_{\ell k}^{\mathrm{S}}$ into $\big(U_{\ell k}^{\mathrm{A}} + U_{k\ell}^{\mathrm{A}}\big)/2$ where

$$U_{\ell k}^{\mathrm{A}}(\beta) = 2\sum_{m=1}^n M_{\ell m}(\beta)\nabla_\beta M_{mk}(\beta) - M_{\ell k}(\beta)\nabla_\beta \log M_{kk}(\beta).$$

The usefulness of providing two U-statistic representations of $\hat{m}_n^{\text{CF}}$ is evident from the following expression, which describes the variance $V_n(\beta_0)$ in terms of both the symmetric and asymmetric kernel functions:

$$V_n(\beta_0) = 2\mathbb{E}\left[\sum_{\ell=1}^{n}\left(\sum_{k\neq\ell} U_{\ell k}^{\text{S}}(\beta_0)y_k\right)\left(\sum_{m\neq\ell} U_{\ell m}^{\text{A}}(\beta_0)y_m\right)\sigma_\ell^2 \,|\, X, A\right] - \mathbb{E}\left[\hat{m}_n^{\text{CF}}(\beta_0)\right]^2 \tag{11}$$

The squared expectation of the sample moment function, which is included at the end of this equation, is zero. It is nevertheless included here to motivate that our variance estimator will subtract $\big(\hat{m}_n^{\text{CF}}(\beta)\big)^2$ whenever it is evaluated at a β where the sample moment is non-zero.

Our proposed variance estimator drops the expectations present in (11) and replaces the unknown individual error variances $\{\sigma_\ell^2\}_{\ell=1}^n$ with cross-fit analogs. However, as there are already outcome variables entering the expression in (11), use of the leave-one-out cross-fit estimators may not suffice for consistent variance estimation. We therefore rely on leave-three-out estimators in construction of $\hat{V}_n$. Specifically, we use

$$\hat{\sigma}_{\ell,-km}^2(\beta) = y_\ell\big(y_\ell - (x_\ell + a_\ell\beta)'\hat{\delta}_{(\ell km)}(\beta)\big) \tag{12}$$

where the leave-three-out estimator of δ is $\hat{\delta}_{(\ell km)}(\beta) = (\sum_{s\neq\ell,k,m}(x_s + a_s\beta)(x_s + a_s\beta)')^{-1}\sum_{s\neq\ell,k,m}(x_s + a_s\beta)y_s$.[5] Our proposed variance estimator is then

$$\hat{V}_n(\beta) = 2\sum_{\ell=1}^{n}\sum_{k\neq\ell}\sum_{m\neq\ell} U_{\ell k}^{\text{S}}(\beta)U_{\ell m}^{\text{A}}(\beta)\left(y_k y_m \hat{\sigma}_{\ell,-km}^2(\beta)\right) - \left(\hat{m}_n^{\text{CF}}(\beta)\right)^2.$$

The leave-three-out cross-fit estimators in (12) are due to Anatolyev and Sølvsten (2020) who introduced them in the context of hypothesis testing in a linear regression.

Remark 7. For samples with about 10, 000 or fewer observations, exact computation of $\hat{V}_n$ is feasible (and fast) if one relies on the matrix representations derived in Anatolyev and Sølvsten (2020). With substantially larger samples, exact computation appears to be infeasible. For this reason, we introduce a recursive representation of the product $y_k y_m \hat{\sigma}_{\ell,-km}^2$, which we truncate to approximate $\hat{V}_n$. Defining $r_{\ell k} = M_{\ell k}/\sqrt{M_{\ell\ell}M_{kk}}$,

[5] When k is equal to m, $\hat{\delta}_{(\ell kk)}(\beta)$ is a leave-two-out estimator since it only drops observations ℓ and k. For this reason, we also use $\hat{\sigma}_{\ell,-k}^2$ when describing $\hat{\sigma}_{\ell,-kk}^2$.

which is bounded by one in absolute value, we can write

$$\begin{aligned} y_k y_m \hat{\sigma}^2_{\ell,-km} = {} & y_k y_m \hat{\sigma}^2_{\ell} - M_{\ell k}\frac{y_\ell y_m \hat{\sigma}^2_k}{M_{\ell\ell}} - \left(M_{\ell m} - \frac{M_{\ell k}M_{km}}{M_{kk}}\right)\frac{y_\ell y_k \hat{\sigma}^2_m}{M_{\ell\ell}} \\ & + (r^2_{\ell k} + r^2_{\ell m} - r_{\ell k} r_{\ell m} r_{km}) y_k y_m \hat{\sigma}^2_{\ell,-km} + \left(M_{\ell m} - \frac{M_{\ell k}M_{km}}{M_{jj}}\right) M_{km}\frac{y_\ell y_m \hat{\sigma}^2_{k,-\ell m}}{M_{\ell\ell}M_{mm}}. \end{aligned} \tag{13}$$

The truncated approximation $\hat{V}_n^{(tr)}$ uses the right hand side of (13) to approximate $y_k y_m \hat{\sigma}^2_{\ell,-km}$.

4.2 Confidence interval and hypothesis test

Here we briefly review the usage of $\hat{\beta}^{\text{CF}}$, $\hat{m}_n$, and $\hat{V}_n$ for inference. For a pre-specified level $\alpha \in (0,1)$, the Wald $(1-\alpha)\cdot 100\%$ confidence interval takes the form

$$C^{\text{W}}_\alpha = \left\{\beta \in \mathcal{B} : \left(\nabla_\beta \hat{m}_n^{\text{CF}}(\hat{\beta}^{\text{CF}})\right)^2 \hat{V}_n(\hat{\beta}^{\text{CF}})^{-1}(\hat{\beta}^{\text{CF}} - \beta)^2 \leq z^2_{\alpha/2}\right\}.$$

Here $z_{\alpha/2}$ is the $\alpha/2$-th quantile of the standard normal distribution. For this confidence interval to have correct asymptotic coverage, certain regularity conditions are required. Chief among them is that $\hat{\beta}^{\text{CF}}$ is consistent, and β_0 belongs to the interior of the parameter space.

The confidence intervals can equivalently be described through the inversion of a hypothesis test. For testing of a simple hypothesis $H_0 : \beta_0 = c$ against the two-sided alternative $H_A : \beta_0 \neq c$, the test whose inversion yields C^{W}_α is

$$\phi^{\text{W}}_\alpha(c) = \mathbf{1}\left\{\left(\nabla_\beta \hat{m}_n^{\text{CF}}(\hat{\beta}^{\text{CF}})\right)^2 \hat{V}_n(\hat{\beta}^{\text{CF}})^{-1}(\hat{\beta}^{\text{CF}} - c)^2 > z^2_{\alpha/2}\right\}$$

In the applied literature that relies on the non-linear least squares estimator $\hat{\beta}^{\text{LS}}$, the standard practice for inference is to rely on bootstrapping (e.g., Arcidiacono et al., 2012; Cornelissen et al., 2017). Unfortunately, there is no theoretical justification for the use of the bootstrap in models like (8) and, in fact, there is theoretical justification for why the bootstrap might fail to yield valid inference (Bickel and Freedman, 1983). To illustrate that the wild bootstrap is indeed not a viable option in the current context, we consider it in the simulations reported in Section 5.

5 Applications and simulations

In this section, we illustrate our ideas using an application similar to that one used in Arcidiacono et al. (2012). Specifically, we estimate the classroom peer effects using the universal transcript data from a large flagship university. We also provide simulation exercises that compare the non-linear least squared estimator and the wild bootstrap standard error with our cross-fit estimator and the Wald standard error.

5.1 Sample selection and econometric specification

We use the administrative student-level records from the Registrar's Office at the University of Wisconsin-Madison, which has a universal coverage of all the students. The database contains multiple records, including demographics, high school test scores, and transcript information.

Sample selection We focus on all the courses that are where the students are all undergraduate students. Under the UW system, it means all the courses with code under 300, e.g., Econ 101, because the university allows graduate students to take courses with code above 300. Given the university has a large body of graduate students, we exclude such courses to avoid noisy interactions coming from graduate students. We keep students who have valid A–F grade information for a given course. We assign numeric grade equivalents to the letter grades following the university GPA system: A = 4, AB = 3.5, B = 3, BC = 2.5, C = 2, D = 1, and F = 0.[6]

In particular, we define the *peer group* as all the students in the same discussion section in the same course. Given that undergraduate courses are typically large in class sizes, the university typically assigns multiple teaching assistants to hold discussion sections for each course. The discussion section typically has fewer than 20 students and allows students to interact and discuss problems with the guidance of the teaching assistant.

We are particularly interested in the semester when the Covid-19 pandemic hit

[6] Arcidiacono et al. (2012) uses a similar database from the University of Maryland from 1999 to 2001 – a much older period than what we focus on. They also divide their main sample into three categories: humanities, social science, and math and science, according to the official course types. The UW system does not have a similar corresponding classification system. As our main purpose is to demonstrate the ideas of our proposed estimator and inference, we keep the sample selection simple by pooling all the students together.

the university during the Spring semester of the academic year 2019/2020. Because the university decided to shift all the undergraduate courses entirely online. Given this situation is unprecedented and online learning tools are new to every student, we expect that the peer effect from student interactions may be largely reduced, if not disappeared. In fact, this is widely observed and discussed among teaching assistants and professors during the semester that nearly all students have turned off their cameras during the discussion section, and classroom interaction almost vanished. We also explore the Spring semester of the academic year 2018/2019 as our placebo semester, as the courses offered for these two semesters are almost identical.

Econometric specification We use the following regression specification following equation (1), which is similar to the specification in Arcidiacono et al. (2012).

$$y_{ij} = \alpha_i + \bar{\alpha}_{(i)j} \cdot \beta + \psi_j + \varepsilon_{ij}, \tag{14}$$

where α_i is the student fixed effect, which measures the ability of a student. We define ψ_j as a course-professor pair fixed effect. For example, if Econ 101 is taught by three professors, Alan, Bob, and Cathy, we define them as three different courses. The reason is straightforward: each professor typically makes their own syllabi and exams. The peer quality $\bar{\alpha}_{(i)j}$ is defined as the average students' ability within the same peer group, excluding the student i.

Note that we do not include the time dimension in this specification because we estimate each semester separately. One can still identify the student fixed effects using data from one semester because students must take at least one course to maintain their full-time student status. Also, each student chooses different lists of courses every semester, and the mobility across different courses is massive, which is important as it serves as the key identifying variation for the peer effect β_0.

Besides, we estimate an extensive form of the equation (14) by incorporating the endogenous effect as proposed by Bramoullé et al. (2020).

$$y_{ij} = \alpha_i + \bar{\alpha}_{(i)j} \cdot \beta + \bar{y}_{(i)j} \cdot \lambda + \psi_j + \varepsilon_{ij}, \tag{15}$$

where $\bar{y}_{(i)j}$ is the average grade for the course j in the peer group, excluding the student i. Note this $\bar{y}_{(i)j}$ is not observed by the student during the semester, although it is observed by the researchers (ex-post). One can consider it as a measure of the average effects of the peer, which is observed by the students, which can also affect their effect on the course.

5.2 Peer effects estimates

Table 1 reports the estimates from equations (14) and (15) in the Spring semesters of the year 2019 and year 2020. Specifically, we estimate equations (14) by using the method from Arcidiacono et al. (2012) and our new proposed method.

Let us first compare the results in the Spring semester of 2019 when the Covid-19 pandemic has not taken place. We find the non-linear least squared (NLLS) estimate for β is around 0.25, and the wild bootstrapping standard error is around 0.03. These figures are in a similar range to what was found in Arcidiacono et al. (2012) although they use a sample from a different university during a much older period. Our proposed cross-fit (CF) estimate is 0.17, which is around 30 percent smaller than the NLLS estimate. The Wald standard error is 0.033, which is 10 percent larger than the wild bootstrap standard error. The results suggest there is a large bias correction using our method, which is both economically and statistically meaningful. As discussed above, the variance of the plug-in fixed effects can be biased, $\hat{\sigma}_{\bar{\alpha}_{(i)t}}$ is estimated to be much smaller using the technique adopted from Kline, Saggio and Sølvsten (2019). As a result, the one-standard-deviation effect is 0.052 under NLLS and plug-in estimator of $\sigma_{\bar{\alpha}_{(i)t}}$ while the counterpart effect is 0.032 under our method, which is about 39 percent smaller.

The last column of the estimates from Spring 2019 is from equation (15), which includes the endogenous effect. With both peer effect β and endogenous effect γ, we use the multiplier effect $\frac{\hat{\beta}+\hat{\lambda}}{1-\hat{\lambda}}$ as the main effect from the peer group, which is estimated to be around 0.11. We use the delta method to infer its resulting standard error using the standard errors from both β and γ. As expected, the one-standard-deviation of the peer ability $\hat{\sigma}_{\bar{\alpha}_{(i)t}}$ is estimated to be similar as before as the inclusion of the endogenous should not dramatically alter the estimation of α_i. The corresponding one-standard-deviation effect is 0.021, which is about 35 percent smaller than the one without the endogenous effect. Given the large difference, including the endogenous effect can also be important to understand the peer effect.

Now, we turn our focus on the Spring semester of 2020 when the Covid-19 pandemic first hit. As we conjectured above, since the pandemic shifted the teaching

Table 1: UW-Madison register data for spring semesters in 2019 and 2020

		Spring 2019			Spring 2020		
		NLLS	CF	CF	NLLS	CF	CF
$\hat{\beta}$		0.249	0.169	0.026	0.049	-0.002	-0.076
		(0.030)	(0.033)	(0.042)	(0.026)	(0.033)	(0.036)
$\hat{\lambda}$				0.075			0.043
				(0.021)			(0.017)
$\frac{\hat{\beta}+\hat{\lambda}}{1-\hat{\lambda}}$				0.110			-0.034
				(0.034)			(0.033)
$\hat{\sigma}_{\bar{\alpha}_{(i)t}}$	plug-in	0.210			0.157		
$\hat{\sigma}_{\bar{\alpha}_{(i)t}}$	KSS		0.188	0.190		0.143	0.143
1-sd effect		0.052	0.032	0.021	0.008	-0.000	-0.005

Notes: Wild bootstrap standard errors for NLLS. Approximate leave-three-out standard errors for cross-fit.

mode to entirely online and social interactions during the discussion sections are very limited, the classroom peer effect should reduce substantially, if not completely disappeared. The NLLS estimate suggests that there is still a positive effect, although much smaller, which is also statistically significant. A one-standard-deviation increase in peer ability is associated with a 0.8 percent rise in the course grade. Using our method, the CF estimate is close to zero with a point estimate of -0.002. The corresponding standard error is 0.033, suggesting the point estimate is statistically insignificant. As a result, the one-standard-deviation effect is essentially zero. A similar result is also estimated under the model with endogenous effect where the effect is tiny and statistically insignificant.

Although not reported here, we also estimated equation (14) using the Spring semester of 2021, which was the last semester when the university imposed campus-wide online learning for undergraduates. We find that the estimated peer effect $\hat{\beta}$ is slightly larger with a point estimate of 0.05 and statistically significant. We believe that the students gradually got adapted to online tools over the past semesters. Still, interaction during the discussion sections is much less frequent than that in the in-person mode. It would be interesting to see how the peer effect changes when the university switches back to the in-person teaching mode. However, we are still applying to access the data from the most recent spring semester of 2022, when the university completely made all the courses in person.

5.3 Simulation exercises

The application above provides a detailed comparison between the NLLS estimator and our CF estimator in one sample. In this subsection, we further compare their performance using simulation exercises.

Specifically, we focus on the equation (14) using our CF parameter estimates during the Spring semester of 2019. We simulate the new outcome variable y by drawing heteroskedastic normal errors and the "true" underlying peer effect $\beta_0 = 0.169$. We conduct the same estimation and inference as in the previous subsection for NLLS and CF using 1,000 simulations, and Table 2 shows the results. We are particularly interested in three sets of performance from the simulations: (i) point estimator, (ii) standard errors, and (iii) coverage.

Table 2: Simulations using UW-Madison 2019 spring semester

	NLLS	CF
Point estimator:		
Bias	0.041	−0.005
Standard deviation	0.028	0.026
Bias/SD	1.440	−0.183
Standard error:		
Standard error/SD - 1	−19.3%	0.3%
Coverage:		
Nominal 95% CI	56.5%	92.9%

Notes: Wild bootstrap standard errors for NLLS. Approximate leave-three-out standard errors for cross-fit.

First, compared to the true β_0, the average bias from NLLS is about 0.041, and the CF estimator gives a much smaller bias of -0.005. Both point estimators have a similar standard deviation of 0.028 and 0.026, corresponding to a bias/SD ratio of 1.44 and -0.183, respectively. The finding suggests that the point estimator of CF performs much better than NLLS, mainly because NLLS is only consistently estimated under homoscedasticity and the simulated error term is heteroscedastic.

Second, we calculate the average standard errors from the simulations. Specifically, the standard errors from NLLS use the wild bootstrapping method, and CF uses the Wald standard errors. We find that the average standard error from CF is 20 percent smaller than the standard deviation of the point estimates. It suggests that wild bootstrapping may not be an ideal way to infer the asymptotic performance of the estimator. On the other hand, the average of our Wald standard errors is almost identical to the standard deviation of the point estimator.

Finally, we obtain a 56.5% coverage rate for the NLLS estimator under the 95% confidence interval. However, the coverage rate for the CF estimator is around 93%, which is very close to 95%. We conclude from the simulation exercise under heteroscedasticity that our proposed method outperforms the NLLS estimator, which is widely adopted in applied studies.

6 Conclusion

In this paper, we propose a framework for estimation and inference in a panel data model of peer and spillover effects. In particular, we consider a linear-in-means model that may include social influences through three channels: outcomes, observed characteristics, and an unobserved individual-specific characteristic. We find that the existing approaches, in general, do not yield consistent estimators in a panel data context. Instead, we propose a new estimator based on a novel objective function that corrects the bias coming from the inconsistency of the existing estimators. Moreover, we provide analytical inference based on a Wald approach, while the current studies mainly focus on identification and point estimator instead of valid inference.

We illustrate the ideas using the universal transcript data from the University of Wisconsin-Madison and explore the classroom peer effects during the semester when the university switched its learning mode to online. We show that the existing method estimates a positive and significant peer effect, while our method finds it to be close to zero and statistically insignificant. Also, simulation exercises suggest that our proposed method yields substantially better estimation and inference in terms of bias and coverage.

Our proposed method is, of course, not perfect. The current main drawback is that it may require more computational power than the existing methods, especially when the dataset is large (e.g., one million workers in a matched employer-employee database). We offer stochastic approximations in Appendix A that largely reduce the computational burden and make the method feasible to be implemented in a large dataset.

References

Abowd, John M., Francis Kramarz, and David N. Margolis, "High Wage Workers and High Wage Firms," *Econometrica*, 1999, *67* (2), 251–333.

Achlioptas, Dimitris, "Database-friendly random projections: Johnson-Lindenstrauss with binary coins," *Journal of computer and System Sciences*, 2003, *66* (4), 671–687.

Anatolyev, Stanislav, "Many instruments and/or regressors: a friendly guide," *Journal of Economic Surveys*, 2019, *33* (2), 689–726.

_ and Mikkel Sølvsten, "Testing Many Restrictions Under Heteroskedasticity," *arXiv preprint arXiv:2003.07320*, 2020.

Arcidiacono, Peter, Gigi Foster, Natalie Goodpaster, and Josh Kinsler, "Estimating spillovers using panel data, with an application to the classroom," *Quantitative Economics*, 2012, *3* (3), 421–470.

Baum-Snow, Nathaniel, Nicolas Gendron-Carrier, and Ronni Pavan, "Local Productivity Spillovers," 2020.

Bianchi, Nicola, Giulia Bovini, Jin Li, Matteo Paradisi, and Michael L Powell, "Career spillovers in internal labor markets," Technical Report, National Bureau of Economic Research 2021.

Bickel, Peter J and David A Freedman, "Bootstrapping regression models with many parameters," in "A festschrift for Erich L. Lehmann," CRC Press, 1983, pp. 28–48.

Bramoullé, Yann, Habiba Djebbari, and Bernard Fortin, "Identification of peer effects through social networks," *Journal of econometrics*, 2009, *150* (1), 41–55.

_ , _ , and _ , "Peer effects in networks: A survey," *Annual Review of Economics*, 2020, *12*, 603–629.

Brune, Lasse, Eric Chyn, and Jason Kerwin, "Peers and Motivation at Work," *Journal of Human Resources*, 2020, pp. 0919–10416R2.

Cornelissen, Thomas, Christian Dustmann, and Uta Schönberg, "Peer effects in the workplace," *American Economic Review*, 2017, *107* (2), 425–456.

Graham, Bryan S, "Identifying social interactions through conditional variance restrictions," *Econometrica*, 2008, *76* (3), 643–660.

Guryan, Jonathan, Kory Kroft, and Matthew J Notowidigdo, "Peer effects in the workplace: Evidence from random groupings in professional golf tournaments," *American Economic Journal: Applied Economics*, 2009, *1* (4), 34–68.

Han, Chirok and Peter CB Phillips, "GMM with many moment conditions," *Econometrica*, 2006, *74* (1), 147–192.

Hausman, Jerry A, Whitney K Newey, Tiemen Woutersen, John C Chao, and Norman R Swanson, "Instrumental variable estimation with heteroskedasticity and many instruments," *Quantitative Economics*, 2012, *3* (2), 211–255.

Hong, Long and Salvatore Lattanzio, "The Peer Effect on Future Wages in the Workplace," 2021.

Hutchinson, Michael F, "A stochastic estimator of the trace of the influence matrix for Laplacian smoothing splines," *Communications in Statistics-Simulation and Computation*, 1989, *18* (3), 1059–1076.

Jackson, C Kirabo and Elias Bruegmann, "Teaching students and teaching each other: The importance of peer learning for teachers," *American Economic Journal: Applied Economics*, 2009, *1* (4), 85–108.

Jochmans, Koen, "Heteroscedasticity-Robust Inference in Linear Regression Models With Many Covariates," *Journal of the American Statistical Association*, 2020, pp. 1–10.

Johnson, William B and Joram Lindenstrauss, "Extensions of Lipschitz mappings into a Hilbert space," *Contemporary mathematics*, 1984, *26* (189-206), 1.

Kelejian, Harry H and Ingmar R Prucha, "A generalized spatial two-stage least squares procedure for estimating a spatial autoregressive model with autoregressive disturbances," *The Journal of Real Estate Finance and Economics*, 1998, *17* (1), 99–121.

Kline, Patrick, Raffaele Saggio, and Mikkel Sølvsten, "Leave-out estimation of variance components," *Econometrica*, 2020, *88* (5), 1859–1898.

_ , _ , and _ , "Vignette for Leave-out estimation of variance components," https://github.com/rsaggio87/LeaveOutTwoWay/blob/master/doc/VIGNETTE.pdf 2021.

_ , _ , and Mikkel Sølvsten, "Leave-out estimation of variance components," 2019.

Lee, Lung-Fei, "Best spatial two-stage least squares estimators for a spatial autoregressive model with autoregressive disturbances," *Econometric Reviews*, 2003, *22* (4), 307–335.

_ , "Identification and estimation of econometric models with group interactions, contextual factors and fixed effects," *Journal of Econometrics*, 2007, *140* (2), 333–374.

_ , Xiaodong Liu, and Xu Lin, "Specification and estimation of social interaction models with network structures," *The Econometrics Journal*, 2010, *13* (2), 145–176.

Lengermann, Paul A., "Is it Who You Are, Where You Work, or With Whom You Work? Reassessing the Relationship Between Skill Segregation and Wage Inequality," Longitudinal Employer-Household Dynamics Technical Papers 2002-10, Center for Economic Studies, U.S. Census Bureau June 2002.

Manski, Charles F., "Identification of Endogenous Social Effects: The Reflection Problem," *The Review of Economic Studies*, 07 1993, *60* (3), 531–542.

Mas, Alexandre and Enrico Moretti, "Peers at work," *American Economic Review*, 2009, *99* (1), 112–145.

Matsushita, Yukitoshi and Taisuke Otsu, "Jackknife, small bandwidth and high-dimensional asymptotics," Technical Report, Suntory and Toyota International Centres for Economics and Related . . . 2019.

Mikusheva, Anna and Liyang Sun, "Inference with many weak instruments," *arXiv preprint arXiv:2004.12445*, 2020.

Moffitt, Robert A, "Policy interventions, low-level equilibria, and social interactions," *Social dynamics*, 2001, *4* (45-82), 6–17.

Mortensen, Dale T. and Christopher A. Pissarides, "Job Creation and Job Destruction in the Theory of Unemployment," *The Review of Economic Studies*, 07 1994, *61* (3), 397–415.

Nix, Emily, "Learning Spillovers in the Firm," *Job Market Paper*, 2020.

Paula, Aureo De, "Econometrics of network models," in "Advances in Economics and Econometrics: Theory and Applications: Eleventh World Congress," Vol. 1 Cambridge University Press Cambridge 2017, pp. 268–323.

Postel–Vinay, Fabien and Jean–Marc Robin, "Equilibrium Wage Dispersion with Worker and Employer Heterogeneity," *Econometrica*, 2002, *70* (6), 2295–2350.

Appendix A Computation on large datasets

This section spells out the key algebraical details that are needed for the implementation of the proposed point estimator and confidence sets. It first provides a matrix representation of the moment function $\hat{m}_n^{\mathrm{CF}}$, its derivative $\nabla_\beta \hat{m}_n^{\mathrm{CF}}$, and the truncated variance estimator $\hat{V}_n^{(tr)}$. Afterward, it introduces the stochastic approximations of these three functions. Finally, it spells out the implementational details regarding the computation of $\hat{\beta}^{\mathrm{CF}}$, C_α^{W}, and C_α^{LM}.

To facilitate a description of $\hat{m}_n^{\mathrm{CF}}$, $\nabla_\beta \hat{m}_n^{\mathrm{CF}}$, and $\hat{V}_n^{(tr)}$ that uses matrix algebra, define $y = (y_1, \dots, y_n)'$, $X = (x_1, \dots, x_n)'$, $A = (a_1, \dots, a_n)'$, $\sigma^2 = (\sigma_1^2, \dots, \sigma_n^2)'$, $\hat{\sigma}^2 = (\hat{\sigma}_1^2, \dots, \hat{\sigma}_n^2)'$ and $M^{(d)} = (M_{11}, \dots, M_{nn})'$. Note that $\hat{\sigma}^2$ and $M^{(d)}$ are functions of β and recall that $S(\beta) = (X + A\beta)'(X + A\beta)$ while $\hat{\varepsilon}(\beta) = M(\beta)y$ and $\hat{\delta}^{\mathrm{LS}}(\beta) = S(\beta)^{-1}(X + A\beta)y$ are the residuals and estimated coefficients from a linear regression of y on $X + A\beta$. For any vector v, diag$[v]$ is the diagonal matrix with v along its main diagonal. Elementwise products and ratios are denoted by $\odot$ and $\oslash$, respectively.

A.1 Matrix representations

Upon utilizing the matrix derivative relationship $\nabla(F^{-1}) = -F^{-1}(\nabla F)F^{-1}$, we find that $\nabla_\beta M = -(D + D')$ for the nilpotent matrix $D(\beta) = M(\beta)AS(\beta)^{-1}(X + A\beta)'$. Letting $\Lambda(\beta) = \mathrm{diag}\big[\nabla_\beta \log\big(M^{(d)}(\beta)\big)\big]$, we can then write $\hat{m}_n^{\mathrm{CF}} = -2\hat{\varepsilon}'A\hat{\delta}^{\mathrm{LS}} - \hat{\varepsilon}'\Lambda y$. Similarly, we have the representation

$$\begin{aligned}\nabla_\beta \hat{m}_n^{\mathrm{CF}} = {} & 2\left(\|MA\hat{\delta}^{\mathrm{LS}}\|^2 + \|A\hat{\delta}^{\mathrm{LS}}\|^2 - \|A\hat{\delta}^{\mathrm{LS}} - D'y\|^2\right) \\ & + (MA\hat{\delta}^{\mathrm{LS}} + D'y)'\Lambda y - \hat{\varepsilon}'(\nabla_\beta \Lambda)y.\end{aligned}$$

After collection of the asymmetric kernel weights $U_{\ell k}^{\mathrm{A}}$ in the matrix $U^{\mathrm{A}} = \{U_{\ell k}^{\mathrm{A}}\}_{\ell,k}$, we have the representation $U^{\mathrm{A}} = -(2D + M\Lambda)$ where $\hat{m}_n^{\mathrm{CF}} = y'U^{\mathrm{A}}y$. Similarly, the matrix of symmetric kernel weights is $U^{\mathrm{S}} = (U^{\mathrm{A}} + U^{\mathrm{A}\prime})/2$ where again $\hat{m}_n^{\mathrm{CF}} = y'U^{\mathrm{S}}y$.

We then have

$$
\begin{aligned}
\hat{V}_n^{(tr)}/2 &= y'U^{\mathrm{S}}\mathrm{diag}\big[\hat{\sigma}^2\big]U^{\mathrm{A}}y - \Big(\hat{m}_n^{\mathrm{CF}}\Big)^2/2 \\
&\quad - \mathrm{trace}\Big(\mathrm{diag}\big[\hat{\sigma}^2\big]M\mathrm{diag}\big[U^{\mathrm{A}}y \odot y \oslash M^{(d)}\big]U^{\mathrm{S}}\Big) \\
&\quad - \mathrm{trace}\Big(\mathrm{diag}\big[\hat{\sigma}^2\big]M\mathrm{diag}\big[U^{\mathrm{S}}y \odot y \oslash M^{(d)}\big]U^{\mathrm{A}}\Big) \\
&\quad + \mathrm{trace}\Big(\mathrm{diag}\big[\hat{\sigma}^2\big]M\mathrm{diag}\big[y \oslash M^{(d)}\big]\big(U^{\mathrm{S}} \odot M\big)\mathrm{diag}\big[y \oslash M^{(d)}\big]U^{\mathrm{A}}\Big).
\end{aligned}
$$

A.2 Stochastic approximations

Exact computation of the log-derivative matrix functions Λ and $\nabla_\beta \Lambda$ appearing in $\hat{m}_n^{\mathrm{CF}}$, U^{A}, U^{S}, and $\nabla_\beta \hat{m}_n^{\mathrm{CF}}$ is challenging in large samples as it typically requires evaluation and storage of the two $n \times n$ matrix functions D and M. We therefore rely on a stochastic approximation to the numerical derivative. Exact evaluation of the three matrix traces appearing in $\hat{V}_n^{(tr)}$ is similarly challenging, and we therefore rely on a related stochastic approximation to those traces. For these approximations, we let p be a large even integer and ϵ a small positive real. Our implementation of the approximation uses $p = 200$ and $\epsilon = 0.005$. Furthermore, we let $(r_1, \ldots, r_p) \in \mathbb{R}^{n \times p}$ be a random matrix with *i.i.d.* Rademacher entries (discrete uniform random variables with support $\{-1, 1\}$).

Our stochastic approximation to the vector $M^{(d)}$ is[7]

$$
\check{M}^{(d)} = \Big\{\sum\nolimits_{s=1}^{p}(Mr_s \odot Mr_s)\Big\} \oslash \Big\{\sum\nolimits_{s=1}^{p}(Mr_s \odot Mr_s) + (Pr_s \odot Pr_s)\Big\}
$$

where $P = I - M$. Each entry in the sums that enter $\check{M}^{(d)}(\beta)$ are squares of the residuals and fitted values in a regression of r_s on $X + A\beta$.

Remark 8. The approximation to $M^{(d)}$ is motivated by the following mean relationships for the numerator and denominator in $\check{M}^{(d)}$: $\mathbb{E}[\sum_{s=1}^{p}(Mr_s \odot Mr_s)] = pM^{(d)}$ and $\mathbb{E}[\sum_{s=1}^{p}(Mr_s \odot Mr_s) + (Pr_s \odot Pr_s)] = p\mathbf{1}_n$ where $\mathbf{1}_n = (1, \ldots, 1)' \in \mathbb{R}^n$. A related version of $\check{M}^{(d)}$ that instead uses the non-random denominator p was suggested in Achlioptas (2003) in the spirit of Johnson and Lindenstrauss (1984); see also Kline et al. (2020). $\check{M}^{(d)}$ improves on that version by enforcing the shape constraints that the entries of $M^{(d)}$ has support on $[0, 1]$. See also Kline, Saggio and Sølvsten (2021)

[7] For any observations where the entries of $\check{M}^{(d)}$ are below 0.01, we replace them by 0.01.

for a derivation of $\check{M}^{(d)}$ as a feasible version of a minimum variance combination of two separate stochastic approximations to $\check{M}^{(d)}$.

The approximation to Λ is based on a finite difference and $\check{M}^{(d)}$:

$$\check{\Lambda}(\beta) = \epsilon^{-1}\mathrm{diag}\Big[\log\Big(\check{M}^{(d)}(\beta+\epsilon) \oslash \check{M}^{(d)}(\beta)\Big)\Big].$$

The approximation to $\nabla_\beta \Lambda$ is similarly

$$\check{\nabla}_\beta \check{\Lambda} = \epsilon^{-1}(\check{\Lambda}(\beta+\epsilon) - \check{\Lambda}(\beta))$$

All simulations and empirical results use the stochastic approximations $\check{m}_n^{\mathrm{CF}} = -2\hat{\varepsilon}' A\hat{\delta} - \hat{\varepsilon}'\check{\Lambda}y$, $\check{U}^{\mathrm{A}} = -(2D + M\check{\Lambda})$, $\check{U}^{\mathrm{S}} = (\check{U}^{\mathrm{A}} + \check{U}^{\mathrm{A}\prime})/2$, $\check{\sigma}^2 = y \odot \hat{\varepsilon} \oslash \check{M}^{(d)}$,

$$\begin{aligned}\check{\nabla}_\beta \check{m}_n^{\mathrm{CF}} = {} & 2\left(\|MA\hat{\delta}^{\mathrm{LS}}\|^2 + \|A\hat{\delta}^{\mathrm{LS}}\|^2 - \|A\hat{\delta}^{\mathrm{LS}} - D'y\|^2\right) \\ & + (MA\hat{\delta}^{\mathrm{LS}} + D'y)'\check{\Lambda}y - \hat{\varepsilon}'(\check{\nabla}_\beta\check{\Lambda})y,\end{aligned}$$

and the stochastic approximation to $\hat{V}_n^{(tr)}$:

$$\begin{aligned}\check{V}_n^{(tr)}/2 = {} & y'\check{U}^{\mathrm{S}}\mathrm{diag}[\check{\sigma}^2]\check{U}^{\mathrm{A}}y - \left(\check{m}_n^{\mathrm{CF}}\right)^2 \\ & - \tfrac{1}{p}\sum\nolimits_{s=1}^{p}\Big(M\big(\check{\sigma}^2 \odot r_s\big) \odot \check{U}^{\mathrm{A}}y \odot y \oslash \check{M}^{(d)}\Big)'\check{U}^{\mathrm{S}}r_s \\ & - \tfrac{1}{p}\sum\nolimits_{s=1}^{p}\Big(M\big(\check{\sigma}^2 \odot r_s\big) \odot \check{U}^{\mathrm{S}}y \odot y \oslash \check{M}^{(d)}\Big)'\check{U}^{\mathrm{A}}r_s \\ & + \tfrac{1}{p}\sum\nolimits_{s=1}^{p/2}\Big(M\big(\check{\sigma}^2 \odot r_s\big) \odot r_{p/2+s} \odot y \oslash \check{M}^{(d)}\Big)'M\Big(\check{U}^{\mathrm{A}}r_s \odot \check{U}^{\mathrm{S}}r_{p/2+s} \odot y \oslash \check{M}^{(d)}\Big) \\ & + \tfrac{1}{p}\sum\nolimits_{s=1}^{p/2}\Big(M\big(\check{\sigma}^2 \odot r_{p/2+s}\big) \odot r_s \odot y \oslash \check{M}^{(d)}\Big)'M\Big(\check{U}^{\mathrm{A}}r_{p/2+s} \odot \check{U}^{\mathrm{S}}r_s \odot y \oslash \check{M}^{(d)}\Big).\end{aligned}$$

Remark 9. Our approximations to the first two traces in $\hat{V}_n^{(tr)}$ are called Hutchinson approximations (Hutchinson, 1989). They utilize that an unbiased estimator for the trace of a matrix F is the quadratic form $r_1' F r_1$ and that this quadratic form is easy to evaluate numerically. For the third trace entering $\hat{V}_n^{(tr)}$ the relevant quadratic form for use with the Hutchinson approximation is numerically challenging to evaluate due to the matrix Hadamard product $U^{\mathrm{S}} \odot M$. For this trace, we combine the Hutchinson approximation with "sample splitting" in the sense that we utilize $(r_1 \odot r_2)'(F_1 r_1 \odot F_2 r_2)$ as an unbiased estimator for the trace of $F_1 \odot F_2$.

Remark 10. The computationally most demanding part of evaluating $\check{m}_n^{\mathrm{CF}}(\beta)$ and

$\check{V}_n^{(tr)}(\beta)$ are to find the solutions to linear systems of equations of the kind $S(\beta)x = b$ for various values of b. To construct $\check{m}_n^{\mathrm{CF}}(\beta)$ there are $2p + 1$ such systems to solve while there is an additional $2 \cdot (2p + 1)$ systems involved in computing $\check{V}_n^{(tr)}(\beta)$. The time it takes to solve those systems of equations depends in large part on the number of regressors present in the model.[8]

Appendix B Derivations

As a service to the reader we quickly state the main definitions used throughout the paper and proofs. Here y, X, and A stacks the observations for y_ℓ, x_ℓ', and a_ℓ' and σ^2 stacks the individual error variances σ_ℓ^2. Furthermore, $R(\beta) = X + A\beta$, $S = R'R$, $P = RS^{-1}R'$, $M = I - P$, $M^{(d)}$ is the diagonal of M, $\Lambda = \mathrm{diag}\big[\nabla_\beta \log\big(M^{(d)}\big)\big]$, $D = MAS^{-1}R'$, $U^{\mathrm{A}} = -(2D + M\Lambda)$, and $U^{\mathrm{S}} = (U^{\mathrm{A}} + U^{\mathrm{A}\prime})/2$. The non-linear least squares estimator is $(\hat{\beta}^{\mathrm{LS}}, \hat{\delta}^{\mathrm{LS}}) = \arg\min_{\beta \in \mathcal{B}, \delta \in \mathbb{R}^k} \sum_{\ell=1}^n \big(y_\ell - x_\ell'\delta - a_\ell'\delta \cdot \beta\big)^2$ while the key objective and moments functions are $\hat{Q}_n = y'My$, $\hat{m}_n = \nabla_\beta Q_n$, $\hat{m}_n^{\mathrm{CF}} = \hat{m}_n - y'M\Lambda y$, $Q_n = \mathbb{E}[\hat{Q}_n \mid X, A]$, and $m_n^{\mathrm{CF}} = \mathbb{E}[\hat{m}_n^{\mathrm{CF}} \mid A, X]$. Finally, $\mathbf{1} = (1, \ldots, 1)' \in \mathbb{R}^n$.

Appendix material for Section 2.2 Suppose that data is generated according to the setup of Section 2.2. We first use the sufficiency principle together with an added assumption that the error terms are homoskedastic normal, to argue that $(\mathcal{Y}, \mathcal{X}, \mathcal{Z})$ contains all the information about β_0. Removing this assumption of homoskedastic normality (without adding other assumptions) can never lead the discarded data to become informative. The original data can be recovered from $(\mathcal{Y}, \mathcal{X}, \mathcal{Z})$ and $(\tilde{\mathcal{Y}}, \tilde{\mathcal{X}}, \tilde{\mathcal{Z}})$ where $\tilde{\mathcal{Y}} = y_{12} + y_{11}$, $\tilde{\mathcal{X}} = y_{32} + y_{21}$, and $\tilde{\mathcal{Z}} = y_{31} + y_{22}$. Furthermore, it follows from standard variance calculations that $(\mathcal{Y}, \mathcal{X}, \mathcal{Z})$ and $(\tilde{\mathcal{Y}}, \tilde{\mathcal{X}}, \tilde{\mathcal{Z}})$ are independent (conditionally on explanatory variables). Finally, the mean of $(\mathcal{Y}, \mathcal{X}, \mathcal{Z})$ depends only on β_0 and $\alpha_3 - \alpha_2$, and even when those two parameters are known, the mean of $(\tilde{\mathcal{Y}}, \tilde{\mathcal{X}}, \tilde{\mathcal{Z}})$ is unrestricted in $\mathbb{R}^3$. It therefore follows that $(\tilde{\mathcal{Y}}, \tilde{\mathcal{X}}, \tilde{\mathcal{Z}})$ only contains information about its mean, or in other words, that $(\mathcal{Y}, \mathcal{X}, \mathcal{Z})$ is sufficient for β_0, $\alpha_3 - \alpha_2$, and the unknown error variance.

We now show that the least squares estimator will be amplified (attenuated) relative to the truth if $\mathcal{Y}$ has higher (lower) unexplained variance than $\mathcal{X}$ and $\mathcal{Z}$. Towards this end, define $(\sigma_{\mathcal{Y}}^2, \sigma_{\mathcal{X}}^2, \sigma_{\mathcal{Z}}^2)$ as the unexplained variance of $(\tilde{\mathcal{Y}}, \tilde{\mathcal{X}}, \tilde{\mathcal{Z}})$. Furthermore,

[8] It also depends on the structure of the model through the degree of sparsity in S.

suppose that β^* is the unique global minimizer of $\sigma^2(\beta)$ and that the least squares estimator converge in probability to β^*. We can write

$$4\sigma^2(\beta) = \frac{2(\beta - \beta_0)^2\mathbb{E}[(\alpha_3 - \alpha_2)^2] + 2\sigma_{\mathcal{Y}}^2 + \mathbb{E}[(\mathcal{Z} - \mathcal{X})^2](1 + \beta^2)}{2 + \beta^2}.$$

If $\sigma_{\mathcal{Y}}^2 \leq \min\{\sigma_{\mathcal{X}}^2, \sigma_{\mathcal{Z}}^2\}$, then we have the ordering $4\sigma^2(\beta^*) \leq 4\sigma^2(\beta_0) \leq \mathbb{E}[(\mathcal{Z} - \mathcal{X})^2] < \mathbb{E}[\mathcal{Z}^2 + \mathcal{X}^2]$ which together with the first order condition in (7) yields

$$\begin{aligned}|\beta^*| &= \frac{|\mathbb{E}[(\mathcal{Z} + \mathcal{X})\mathcal{Y}]|}{\mathbb{E}[\mathcal{Z}^2 + \mathcal{X}^2] - 4\sigma^2(\beta^*)} \leq \frac{|\mathbb{E}[(\mathcal{Z} + \mathcal{X})\mathcal{Y}]|}{\mathbb{E}[\mathcal{Z}^2 + \mathcal{X}^2] - 4\sigma^2(\beta_0)} \\ &\leq \frac{|\mathbb{E}[(\mathcal{Z} + \mathcal{X})\mathcal{Y}]|}{\mathbb{E}[\mathcal{Z}^2 + \mathcal{X}^2] - \mathbb{E}[(\mathcal{Z} - \mathcal{X})^2]} = |\beta_0|.\end{aligned}$$

If instead we have $\sigma_{\mathcal{Y}}^2 \geq \max\{\sigma_1^2, \sigma_2^2\}$, then we have $\mathbb{E}[\mathcal{Z}^2 + \mathcal{X}^2] = \lim_{|\beta|\to\infty} 4\sigma^2(\beta) > 4\sigma^2(\beta^*) \geq \mathbb{E}[(\mathcal{Z} - \mathcal{X})^2]$ and in turn

$$|\beta^*| = \frac{|\mathbb{E}[(\mathcal{Z} + \mathcal{X})\mathcal{Y}]|}{\mathbb{E}[\mathcal{Z}^2 + \mathcal{X}^2] - 4\sigma^2(\beta^*)} \geq \frac{|\mathbb{E}[(\mathcal{Z} + \mathcal{X})\mathcal{Y}]|}{\mathbb{E}[\mathcal{Z}^2 + \mathcal{X}^2] - \mathbb{E}[(\mathcal{Z} - \mathcal{X})^2]} = |\beta_0|.$$

Next, we connect the expressions in (4)–(7) with the general formulas introduced in Section 3. As the model is over parameterized, we drop ψ_B from the model so that the design has full rank. For the six observations in a given triplet, the corresponding part of the matrix $M(\beta)$ is

$$I - \frac{1}{2(2+\beta^2)}\begin{bmatrix} 2(1+\beta^2) & 2 & \beta & \beta & -\beta & -\beta \\ 2 & 2(1+\beta^2) & -\beta & -\beta & \beta & \beta \\ \beta & -\beta & 3+\beta^2 & 1 & -1 & 1+\beta^2 \\ \beta & -\beta & 1 & 3+\beta^2 & 1+\beta^2 & -1 \\ -\beta & \beta & -1 & 1+\beta^2 & 3+\beta^2 & 1 \\ -\beta & \beta & 1+\beta^2 & -1 & 1 & 3+\beta^2 \end{bmatrix}$$

and in analogy with the sufficiency argument above, the formulation in terms of the full data leads to one half times the objective defined for $(\mathcal{Y}, \mathcal{X}, \mathcal{Z})$ where the corresponding matrix $M(\beta)$ is

$$I - \frac{1}{2+\beta^2}\begin{bmatrix} \beta^2 & \beta & \beta \\ \beta & 1 & 1 \\ \beta & 1 & 1 \end{bmatrix} = \frac{1}{2+\beta^2}\begin{bmatrix} 2 & -\beta & -\beta \\ -\beta & 1+\beta^2 & -1 \\ -\beta & -1 & 1+\beta^2 \end{bmatrix}.$$

Thus the contribution of any particular triplet to the least squares objective function is four times the sample analog of (6) since

$$\mathcal{Y}^2 + \mathcal{X}^2 + \mathcal{Z}^2 - \frac{(\mathcal{Y}\beta + \mathcal{X} + \mathcal{Z})^2}{2+\beta^2} = \frac{(\mathcal{Z}-\mathcal{X})^2 + (\mathcal{Y}-\mathcal{X}\beta)^2 + (\mathcal{Y}-\mathcal{Z}\beta)^2}{2+\beta^2}.$$

The cross-fit objective function contribution defined for a single triplet $(\mathcal{Y}, \mathcal{X}, \mathcal{Z})$ at (β_1, β) is similarly found as

$$\frac{-2\mathcal{Y}(\mathcal{Z}+\mathcal{X})\beta_1 - 2\mathcal{Z}\mathcal{X}}{2+\beta_1^2} + \frac{(1+\frac{1+\beta_1^2}{1+\beta^2})\mathcal{Y}(\mathcal{Z}+\mathcal{X})\beta + 2\frac{1+\beta_1^2}{1+\beta^2}\mathcal{Z}\mathcal{X}}{2+\beta_1^2}.$$

Therefore, each triplet contributes $(-\mathcal{Y}(\mathcal{Z}+\mathcal{X})+2\mathcal{Z}\mathcal{X}\beta)\frac{2}{(2+\beta^2)(1+\beta^2)}$ to the first order condition for a minimum of the cross-fit objective in β_1 at $\beta_1 = \beta$. Thus, we obtain the sample analog of the first order condition in (4).