

Estimation and Inference in a Peer Effects Model under Heteroskedasticity

Long Hong, Kensuke Sakamoto, and Mikkel Sølvsten*

August 5, 2026

Abstract

This paper develops estimation and inference for a panel-data peer-effects model in which social interactions operate through peers' unobserved characteristics and the errors are heteroskedastic. The non-linear least squares (NLLS) estimator widely used in this literature is inconsistent under heteroskedasticity, with bias of indeterminate sign. We propose a cross-fit correction that restores consistency under heteroskedasticity, and we develop the first analytic standard error for non-linear peer-effects estimators, where the literature has relied on the wild bootstrap. We apply the method to two empirical settings. In university transcript data from the COVID-19 online semester, NLLS finds a positive and significant classroom peer effect, whereas our estimator finds it close to zero and statistically insignificant. In the Italian matched employer–employee panel, by contrast, our estimate is about 15% larger than the NLLS estimate. With a consistent estimate of the peer effect, a bias-corrected wage-variance decomposition reveals that average coworker quality explains a share of wage variance comparable to that of firm effects — a channel the standard AKM model cannot see — with worker–coworker sorting as the dominant margin.

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@asu.edu, Arizona State University, USA. Sakamoto: McGill University, Canada, kenlfo2080@gmail.com. Sølvsten: miso@econ.au.dk, Aarhus University, Denmark. We acknowledge the Office of the Registrar at the University of Wisconsin-Madison and Fondazione Rodolfo Debenedetti and Bocconi University, for data access.

1 Introduction

Many empirical studies of social interactions document strong correlations between individuals’ outcomes and those of their peers. As [Manski \(1993\)](#) emphasizes, three distinct mechanisms can generate such patterns: contextual spillovers from exogenous peer characteristics, endogenous effects from peer outcomes, and correlated effects from shared characteristics or environments ([De Paula, 2017](#)). Contextual and endogenous spillovers cannot in general be separated when interactions take place within groups, as [Manski \(1993\)](#) and [Moffitt \(2001\)](#) argue — the reflection problem — while identification is restored by variation in group sizes ([Lee, 2007](#); [Graham, 2008](#)) or by structuring interactions through a network ([Bramoullé, Djebbari and Fortin, 2009](#)). Panel data extend this logic by allowing correlated effects to operate through an unobserved individual-specific characteristic, and [Mas and Moretti \(2009\)](#) and [Arcidiacono, Foster, Goodpaster and Kinsler \(2012\)](#) independently develop estimation procedures for this setting. The non-linear least squares (NLLS) estimator of [Arcidiacono et al. \(2012\)](#) has since become the workhorse of a large applied literature.¹

Peer and spillover effects of this kind have been documented across many domains — from classroom achievement and workplace productivity to urban agglomeration, regional labor markets, and firm-to-firm spillovers (e.g., [Mas and Moretti, 2009](#); [Burke and Sass, 2013](#); [Battisti, 2017](#); [Cornelissen, Dustmann and Schönberg, 2017](#); [De la Roca and Puga, 2017](#); [Dix-Carneiro and Kovak, 2017](#); [Holden, Keane and Lilley, 2021](#); [Neururer and Sun, 2021](#); [Thiemann, 2022](#); [Portugal, Reis, Guimarães and Cardoso, 2024](#); [Hong and Lattanzio, 2025](#); [Messina, Sanz-de Galdeano and Terskaya, 2026](#)). The consistency of the NLLS estimator, however, rests on an assumption of homoskedasticity that is untenable in the administrative-data settings where it is typically applied, in which error variances differ across individuals through mobility, selection, and measurement heterogeneity. Inference, likewise, has relied on wild-bootstrap procedures, with no analytical alternative available for this class of estimators.

This paper develops estimation and inference for the panel-data peer-effects model when the errors are heteroskedastic. We establish three main theoretical results. First, the NLLS estimator widely used in this literature is inconsistent under het-

¹In linear-in-means models, consistent estimation can also be facilitated by maximizing a Gaussian likelihood ([Lee, Liu and Lin, 2010](#)) or by a two-step instrumental-variables estimator ([Kelejian and Prucha, 1998](#); [Lee, 2003](#)).

eroskedasticity, and the direction of the bias is determined entirely by the empirical design — meaning practitioners cannot use NLLS as an upper or lower bound on the peer coefficient without correction. As we show in a simple three-worker example, the reason is concrete: even with each individual’s permanent level absorbed by a fixed effect, outcomes fluctuate around that level more for some individuals, and in some situations, than for others, and the bias tilts toward amplification or attenuation depending on whether this transitory volatility sits with the movers or with the stayers around them. Second, we propose a cross-fit (CF) correction to the NLLS moment that yields a consistent estimator under heteroskedasticity. Third, we propose an analytic standard error for the peer coefficient, while the literature has relied on wild-bootstrap procedures for this class of estimators. Additionally, although we develop them for peer effects, these results apply more broadly: the cross-fit correction works for any non-linear least squares problem in which a low-dimensional parameter multiplies a latent individual effect.

We first revisit the classroom peer-effect setting of [Arcidiacono et al. \(2012\)](#) using the universal transcript data from the University of Wisconsin–Madison. In a pre-pandemic benchmark sample, the CF estimate is about 30% smaller than the NLLS estimate (0.17 versus 0.25), and Monte Carlo simulations calibrated to this design show why the correction matters for inference as well: the nominal 95% confidence interval covers the truth in only 46% of draws under NLLS with the wild bootstrap, against 94% under our cross-fit estimator with the proposed standard error. As a further test, we turn to the Spring 2020 semester, when the pandemic moved all undergraduate teaching online. NLLS continues to return a positive and statistically significant classroom peer effect, whereas the CF estimate is close to zero and statistically insignificant.

We then apply our estimator to study workplace peer effects by incorporating a coworker component into a canonical AKM model ([Abowd, Kramarz and Margolis, 1999](#)) using the Italian matched employer–employee panel. The cross-fit estimate is about 15% *larger* than the iterated NLLS estimate on the same sample — the opposite direction from the classroom application. The workplace application is not merely a second illustration. A consistent peer coefficient lets us ask a question the two-way AKM model cannot answer — how much of wage inequality operates through a worker’s coworkers rather than through the firm — by extending the canonical worker–firm wage decomposition to incorporate average coworker quality. Two find-

ings emerge. First, coworker quality explains about 11% of wage variance — a share comparable to that explained by firm effects, but invisible to the two-way model. Second, the worker–firm sorting correlation falls from 0.21 to 0.08 once coworkers are included, while a worker–coworker correlation of 0.60 emerges in its place.

This paper contributes to three strands of the literature. The first is estimation *and* inference for peer effects operating through peers’ unobserved heterogeneity. On estimation, we propose a cross-fit correction that yields a consistent estimator under general heteroskedasticity, whereas the NLLS estimator of [Arcidiacono et al. \(2012\)](#) is consistent only under homoskedasticity. Two related estimators take different routes, each under conditions ours does not require. The two-step estimator of [Mas and Moretti \(2009\)](#) plugs estimated worker fixed effects into the peer average and is attenuated unless those fixed effects are estimated with a precision special to their ten-minute productivity data. [Braun and Verdier \(2023\)](#) proposes an instrumental-variables estimator, whereas ours requires no search for instruments. On inference, we provide the first analytic standard error for a non-linear peer-effects estimator, while the literature has relied on the wild bootstrap.

The second is a methodological literature on estimation and inference in models with many parameters (see [Anatolyev, 2019](#), for a survey). This work has focused on linear specifications, building on leave-one-out constructions ([Hausman, Newey, Woutersen, Chao and Swanson, 2012](#); [Kline, Saggio and Sølvesten, 2020](#); [Anatolyev and Sølvesten, 2023](#)) that purge incidental-parameter bias from second moments of estimated fixed effects. We contribute to this literature by carrying the leave-out approach beyond the linear models on which it has so far focused. In our setting the parameter of interest enters non-linearly, multiplying a latent peer average that is itself a function of the estimated individual effects — a structure the existing linear corrections do not cover. More broadly, the correction applies to any non-linear least squares problem in which a low-dimensional parameter multiplies a latent individual effect; models of heterogeneous teacher value-added ([Hahn, Singleton and Yildiz, 2023](#)) share this structure, so the method is not specific to peer averaging.

Finally, our workplace application contributes to a large literature on the sources of wage inequality, which has emphasized firm heterogeneity and worker–firm sorting as central drivers ([Card, Heining and Kline, 2013](#); [Card, Cardoso, Heining and Kline, 2018](#); [Song, Price, Guvenen, Bloom and Von Wachter, 2019](#)). Incorporating coworker quality into this decomposition has not previously been feasible: it requires

a consistent peer coefficient together with second moments of the estimated coworker quality that are purged of incidental-parameter bias — the coworker-channel analog of the limited-mobility bias correction in two-way models (Andrews, Gill, Schank and Upward, 2008; Kline et al., 2020). Existing methods provide neither. Ours provides both, which lets us split what the two-way model bundles into a single firm effect into two distinct channels — the influence of a worker’s coworkers and that of the firm itself. We find that coworker quality rivals firm effects in its contribution to wage variance and that the dominant sorting margin shifts from worker–firm to worker–coworker — a reallocation that revises the standard account of how labor markets generate wage inequality.

The paper is organized as follows. Section 2 sets out the model and the formal identification result. Section 3 discusses the source of inconsistency in NLLS under heteroskedasticity and introduces our cross-fit estimator with its consistency theorem. Section 4 presents the consistent variance estimator for valid inference. Section 5 revisits the classroom peer-effect setting of Arcidiacono et al. (2012) and reports Monte Carlo simulations comparing our cross-fit estimator with NLLS. Section 6 applies the method to workplace peer effects in Italian matched employer–employee data and presents the extended wage-variance decomposition. Section 7 concludes.

2 Model and identification

This section introduces the panel-data peer-effects model that is the focus of the paper and the variation that identifies the peer coefficient β_0 . The model is the leading case of a broader class of non-linear regression problems — those in which a low-dimensional parameter multiplies a latent individual effect — for which we develop estimation and inference in Section 3. We return to the peer-effects model in the empirical application from 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 $\mathcal{P}_{it} \subseteq \{1, \dots, N\}$.

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

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

where $\bar{\alpha}_{(i)t} = |\mathcal{P}_{it}|^{-1} \sum_{i \in \mathcal{P}_{it}} \alpha_i$ 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 $\alpha = (\alpha_1, \dots, \alpha_n)'$ 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 et al., 2017](#); [Hong and Lattanzio, 2025](#)), worker 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, 2024](#)).

Because $\bar{\alpha}_{(i)t}$ excludes α_i , one may ask whether (1) inherits the exclusion, or “omit-self,” bias studied by [Bayer, Ross and Topa \(2008\)](#), [Guryan et al. \(2009\)](#), and [Caeyers and Fafchamps \(2016\)](#), in which a leave-out peer average is mechanically negatively correlated with own type within a finite selection pool. It does not: α_i enters as a free parameter, so that correlation is one between included regressors rather than between a regressor and the error term. Model (1) therefore satisfies by construction the principle stated by [Fletcher, Ross and Zhang \(2020\)](#), that a component removed from a constructed index should be included as its own regressor rather than left in the error. The separate difficulty created by the fact that α must be estimated is the subject of Section 3.

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

$$\mathbb{E}[\varepsilon_{it} | \mathcal{F}_i] = 0, \quad i = 1, \dots, N, \quad t = 1, \dots, T_i. \quad (2)$$

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

$$w'_{it}\gamma = \psi_{j(i,t)} + \lambda_t + c'_{it}\gamma_c, \quad (3)$$

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 et al. \(1999\)](#), which was originally developed for wage regressions but has since 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, homoskedastic error terms. To shed light on the role played by these assumptions, Section 3 discusses why consistency of least squares fails when the error terms are not serially uncorrelated and homoskedastic. 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 heteroskedasticity. 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) gives β_0 a transparent interpretation: it 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 bias correction of [Kline et al. \(2020\)](#) to estimate the variance of average peer-group quality. Because the plug-in variance of estimated

²When 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.

peer quality is inflated by limited mobility, the applied peer-effects literature has typically reported the effect of a fixed-percentage increase (e.g., 10 percent in Mas and Moretti, 2009; Cornelissen et al., 2017) rather than a one-standard-deviation effect; our bias-corrected variance lets us report the latter directly. In practice, the mechanism through which peers affect outcomes is as interesting as the magnitude of the effect. Plausible mechanisms include knowledge spillover (Nix, 2016), 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 for the overall peer effect, so we will not delve further into the specific mechanisms that may drive the magnitude and sign of β_0 .

2.2 Identifying variation

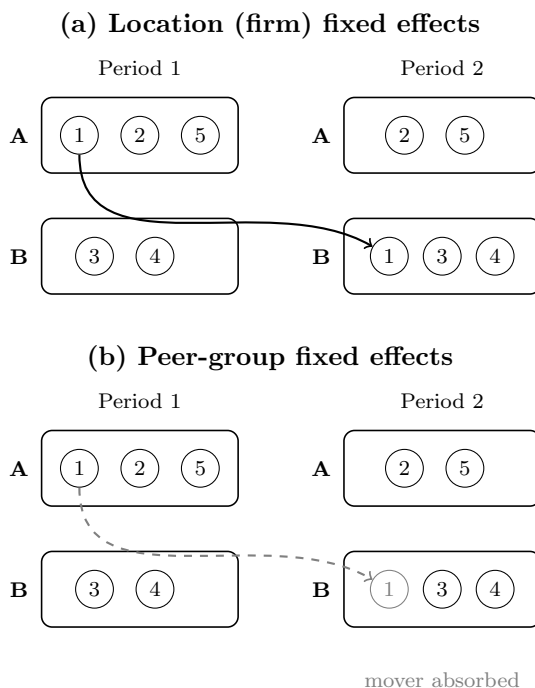
Identification of β_0 requires variation in average peer quality that the linear part of (1) cannot predict. Two conditions deliver it. First, workers must move between peer groups — mobility that is directly observable in the data. Second, that mobility must actually shift the unobserved average peer quality $\bar{\alpha}_{(i)t}$. This cannot be verified directly, since α is unobserved, but it holds in any setting with heterogeneous workers and imperfect sorting, except in the knife-edge case where everyone shares the same α .

Identification draws on two sources of this variation. *Movers* contribute when a worker changes peer groups, so that the average quality of their peers changes. *Stayers* contribute when a worker remains in a group whose composition changes as others join or leave. Figure 1 illustrates both variations in panel (a), with five workers and two firms over two periods: worker 1 moves from firm A to firm B, while the others stay. The single move changes three average peer qualities. The mover, worker 1, experiences a change from $\frac{1}{2}(\alpha_2 + \alpha_5)$ to $\frac{1}{2}(\alpha_3 + \alpha_4)$. Stayer 2, who loses worker 1 from firm A, experiences a change from $\frac{1}{2}(\alpha_1 + \alpha_5)$ to α_5 ; and stayer 3, who gains worker 1 at firm B, from α_4 to $\frac{1}{2}(\alpha_1 + \alpha_4)$.

One can also impose the most demanding control — a fixed effect for each peer group in each period — which absorbs any shock common to the group. The peer effect β_0 remains identified even then: because $\bar{\alpha}_{(i)t}$ excludes worker i , it varies within a peer group as the group’s composition and size change. Figure 1(b) shows the variation that survives: worker 1’s move between firms no longer contributes — the

peer-group effect absorbs it — but the variation from the stayers is unchanged.

Figure 1: Sources of identifying variation under two control specifications



Notes: Five workers and two firms over two periods; worker 1 moves from firm A to firm B. Under location (firm) fixed effects, panel (a), the move changes the average peer quality $\bar{\alpha}$ of the mover (worker 1) and of the stayers it leaves (worker 2) and joins (worker 3), and all three changes identify β_0 . Under a saturated peer-group fixed effect, panel (b), the mover's between-firm change is absorbed, and only the stayers' within-group changes survive.

3 Estimation

This section first characterizes the source of inconsistency in the least squares estimator applied to a generic regression model with multiplicative non-linearity. At a high level, the inconsistency is that the least squares objective is not minimized near the truth — equivalently, that its gradient does not vanish near β_0 . Building on this observation, we then propose an estimator that sets a recentered gradient of the least squares objective equal to zero.³

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

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

$$y_\ell = x'_\ell \delta + a'_\ell \delta \cdot \beta_0 + \varepsilon_\ell, \quad \ell = 1, \dots, n. \quad (4)$$

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, δ contains fixed effects and coefficients on control variables, x_ℓ contains dummy variables for fixed effects and control variables, and a_ℓ is a function of the peer group that observation ℓ belongs to and is 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, $A = (a_1, \dots, a_n)'$ and $X = (x_1, \dots, x_n)'$, so that a_ℓ (and x_ℓ) may be arbitrarily dependent across observations.

Our analysis maintains three primary conditions — strict exogeneity, compactness of the parameter space for β_0 , and a collection of full-rank conditions — which we now state formally.

Assumption 1 (Primary conditions). *(i) Strict exogeneity: $\mathbb{E}[\varepsilon_\ell | X, A] = 0$ for all ℓ , and $\text{range}(X)$ contains the constant vectors.*

(ii) Compact parameter space: $\beta_0 \in \text{interior}(\mathcal{B})$, where $\mathcal{B} \subseteq \mathbb{R}$ is compact.

(iii) Full rank: there exists $N < \infty$ such that $(X + A\beta, A\delta)$ has full rank for any $\beta \in \mathcal{B}$ and all $n \geq N$.

(iv) Bounded fourth moments: there exist constants $C_4 < \infty$ and $N < \infty$ such that, for all $n \geq N$,

$$\max_{\ell \leq n} \mathbb{E}[\varepsilon_\ell^4 | X, A] \leq C_4.$$

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, part (iii) is a collection of full-rank conditions on implied matrices of regressors, and part (iv) is a standard uniform boundedness condition on the fourth moments of the error terms. In the peer-effects model of Section 2, the parameter space $\mathcal{B}$ can be taken to be a compact subset of the interval $(-1, 1)$ introduced there. Part (i) is often

called “exogenous mobility” in the employer-employee matched data literature; it excludes sorting between individuals and peer groups based on latent effects that are not captured by fixed effects controlled for in the regression. 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 estimating 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.

Remark 1. The high-level model (4) is general; peer effects is its leading case, but not the only one. The estimation of heterogeneous teacher value-added is another instance. In the specification of Hahn et al. (2023), for example,

$$y_{js} = w_{js}\gamma + \alpha_j + \alpha_j z_s \beta_0 + \varepsilon_{js},$$

β_0 measures the heterogeneity in teacher value-added, α_j is teacher j ’s latent value-added, w_{js} collects observed covariates, and z_s denotes student s ’s attributes. This is a special case of (4), with δ containing the teacher effects $\{\alpha_j\}$ and γ , x_ℓ a teacher dummy and w_{js} , and a_ℓ a teacher dummy interacted with z_s . The cross-fit correction therefore applies here as well; we develop it for peer effects, the setting of our applications.

3.1 Inconsistency of least squares

The least squares estimator applied to (4) yields the following estimator of β_0 :

$$\hat{\beta}^{\text{NLLS}} = \arg \min_{\beta \in \mathcal{B}} \min_{\delta \in \mathbb{R}^k} \sum_{\ell=1}^n (y_\ell - x'_\ell \delta - a'_\ell \delta \cdot \beta)^2.$$

To give a representation of $\hat{\beta}^{\text{NLLS}}$ 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 $\hat{\beta}^{\text{NLLS}}$ as the solution to a minimization problem that does not involve δ :

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

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 $\hat{\beta}^{\text{NLLS}}$ is that the population analog $Q_n(\beta) = \mathbb{E}[\hat{Q}_n(\beta) | X, A]$ has a unique minimum at β_0 . However, even under the strict exogeneity imposed in Assumption 1, we can write this expectation as the sum of two terms, only the first of which has a unique minimum at β_0 . To illustrate this point, let $\sigma_{\ell k} = \mathbb{E}[\varepsilon_\ell \varepsilon_k | X, A]$ be the covariance between the ℓ -th and the k -th error terms and define the part of $a'_\ell \delta$ that provides 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

$$Q_n(\beta) = (\beta - \beta_0)^2 \sum_{\ell=1}^n (\tilde{a}_\ell(\beta)' \delta)^2 + \sum_{\ell=1}^n \sum_{k=1}^n M_{\ell k}(\beta) \sigma_{\ell k}. \quad (5)$$

The full-rank restrictions of Assumption 1 imply that $\sum_{\ell=1}^n (\tilde{a}_\ell(\beta)' \delta)^2 > 0$ for all $\beta \in \mathcal{B}$ 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 therefore leads to inconsistency of the least squares estimator except in special cases.

Before proceeding to our proposed estimator, it is useful to highlight why the least squares estimator remains consistent with serially uncorrelated and homoskedastic error terms. In this case, we have $\sigma_{\ell k} = \sigma^2 \mathbf{1}\{\ell = k\}$. This property implies that the second part of Q_n simplifies substantially. In fact, this second part becomes independent of β because the matrix function $M = (M_{\ell k})_{\ell, k}$ is a projection onto a linear space of dimension $n - K$, which in turn implies 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 in this case. 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); Hong and Lattanzio (2025) extends the consistency result under homoskedasticity to the model with controls.

As the regression model in (4) involves an unobserved regressor $a'_\ell \delta$ that is itself estimated from the data, one might be tempted to apply standard measurement-error logic and conjecture that the least squares estimator is attenuated toward zero. The bias term in (5), however, has no determinate sign: it depends on how the error variances covary with the leverage weights $M_{\ell k}(\beta)$, so least squares may be amplified or attenuated. Appendix B.2 confirms this in the three-worker example, where β admits a closed form: least squares is amplified (attenuated) when the stayer's wage difference has higher (lower) residual variance than the movers', so estimates in the literature may depart from the underlying truth in systematic but unknown directions.

3.2 Cross-fit correction to least squares

Our proposed estimator relies on the standard cross-sectional assumption that errors are conditionally independent in model (4). 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 independence). *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 heteroskedasticity in the non-linear regression model in (4).

Remark 2. In the peer effects model in 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 accommodated 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 (4) for the collapsed data. Section 5 further illustrates this approach and the biases that can arise from ignoring serial dependence in the microdata. We therefore highlight that, when 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. We note, however, that cross-sectional dependence across ε_ℓ when observations share a common peer group or a common employer/classroom is not allowed here, while the inclusion of common shocks as in Section 6.2.1 potentially relaxes this restriction. [Braun and Verdier \(2023\)](#) allows for such cross-sectional dependence in error terms following a different estimation strategy.

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

$$m_n(\beta_0) = \sum_{\ell=1}^n \nabla_\beta M_{\ell\ell}(\beta_0) \sigma_\ell^2. \quad (6)$$

Our proposed estimator is a zero of a sample moment function defined as $\hat{m}_n$ minus an estimate of the bias term in (6). We construct this sample moment using cross-fit, or leave-one-out, estimators of the individual error variances:

$$\hat{\sigma}_\ell^2(\beta) = \frac{y_\ell \hat{\varepsilon}_\ell(\beta)}{M_{\ell\ell}(\beta)} \quad (7)$$

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

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

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

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

$$\hat{\beta}^{\text{CF}} = \arg \text{zero}_{\beta \in \mathcal{B}} \hat{m}_n^{\text{CF}}(\beta) \quad \text{where } \hat{m}_n^{\text{CF}}(\beta) = \hat{m}_n(\beta) - \sum_{\ell=1}^n \nabla_\beta M_{\ell\ell}(\beta) \hat{\sigma}_\ell^2(\beta). \quad (8)$$

The recentered moment function is finite-sample unbiased at the truth:

Proposition 1 (Finite-sample unbiasedness of the CF moment). *Under Assumptions 1 and 2, if $\min_{l \leq n} M_l(\beta_0) > 0$,*

$$\mathbb{E}[\hat{m}_n^{CF}(\beta_0) \mid X, A] = 0.$$

The result follows because the cross-fit estimators $\{\hat{\sigma}_\ell^2(\beta_0)\}_{\ell=1}^n$ are unbiased for their respective error variances, so the recentering⁴ exactly removes the bias term in (6). The unbiasedness of $\hat{m}_n^{CF}$ at β_0 is the cornerstone property that motivates $\hat{\beta}^{CF}$ as a credible alternative to $\hat{\beta}^{NLLS}$ without requiring homoskedasticity.

The cross-fit variance estimators $\{\hat{\sigma}_\ell^2\}_{\ell=1}^n$ have previously been used in the linear-regression literature to bias-correct non-linear functions of the least squares estimator and to estimate its variance (Matsushita and Otsu, 2019; Anatolyev and S¸olvsten, 2023; Jochmans, 2020; Kline et al., 2020; Mikusheva and Sun, 2020); the use here differs in that we consider a non-linear regression and use cross-fitting to bias-correct the least squares estimator itself.

There is a long tradition in econometrics of bias-correcting objective functions, rather than their gradients, in an attempt to ensure that their population counterparts are minimized at the truth (e.g., Han and Phillips, 2006; Hausman et al., 2012). Translating that approach here would suggest a penalized objective function $\hat{Q}_n^{CF}(\beta) = \hat{Q}_n(\beta) - \sum_{\ell=1}^n M_{\ell\ell}(\beta)\hat{\sigma}_\ell^2(\beta)$; however, this objective is identically zero in β and so cannot yield a consistent estimator of β_0 . Recentering the gradient instead is what makes the construction work.

3.3 Consistency

Consistency of $\hat{\beta}^{CF}$ requires conditions beyond those used for identification and finite-sample unbiasedness. We collect them in the following regularity assumption, which is the non-linear analog of the design conditions imposed in Anatolyev and S¸olvsten (2023) for linear regression.

⁴The recentering has a transparent measurement-error interpretation: the leave-one-out projection supplies, for each observation, a second and independent noisy measurement of the latent regressor $a_\ell'\delta$, and combining the two views cancels the error-variance contamination that makes least squares inconsistent. Appendix B.1 develops this in the three-worker example.

Assumption 3 (Regularity for consistency). (i) Bounded leverage:

$$\liminf_{n \rightarrow \infty} \inf_{\beta \in \mathcal{B}} \min_{1 \leq \ell \leq n} M_{\ell\ell}(\beta) > 0.$$

(ii) Bounded design: *there exist constants $C_x, C_a, C_\delta < \infty$ and $N < \infty$ such that, for all $n \geq N$,*

$$\begin{aligned} \max_{\ell \leq n} \|x_\ell\|_1 &\leq C_x, & \max_{\ell \leq n} \|a_\ell\|_1 &\leq C_a, \\ \|\delta\|_\infty &\leq C_\delta, \end{aligned}$$

and

$$\liminf_{n \rightarrow \infty} \inf_{\beta \in \mathcal{B}} \lambda_{\min}(n^{-1}S(\beta)) > 0.$$

(iii) Restricted dimension: $K/n \rightarrow \rho \in [0, 1)$ as $n \rightarrow \infty$.

(iv) Strong identification: *for every $\varepsilon > 0$,*

$$\liminf_{n \rightarrow \infty} \inf_{\beta \in \mathcal{B}: |\beta - \beta_0| \geq \varepsilon} n^{-1} |m_n^{\text{CF}}(\beta)| > 0,$$

where $m_n^{\text{CF}}(\beta) = \mathbb{E}[\hat{m}_n^{\text{CF}}(\beta) \mid X, A]$ is the population counterpart of the cross-fit moment function.

Part (i) ensures that the leave-one-out variance estimator $\hat{\sigma}_\ell^2 = y_\ell \hat{\varepsilon}_\ell / M_{\ell\ell}$ is well-defined with a bounded denominator; this is the standard limited-mobility-type restriction used in the leave-out literature (Kline et al., 2020; Anatolyev and Sølvesten, 2023). Part (ii) ensures that the least squares estimator of δ for any $\beta \in \mathcal{B}$ is well-defined and imposes row-sparsity and bounded-coefficient conditions suited to high-dimensional fixed-effect designs. The row-sparsity conditions are satisfied in the peer effects setting when the sizes of peer groups are bounded or grow slowly with n , which is a common feature of sparsely matched teacher-student/worker-firm data. Part (iii) excludes the over-parameterized regime in which K grows as fast as n . Part (iv) is an identification condition that requires $m_n^{\text{CF}}(\beta_0) = 0$ to be a well-separated unique solution to the population moment condition. Roughly speaking, this condition is satisfied when the variation in $\tilde{a}_\ell(\beta)' \delta$ is sufficiently large relative to the terms capturing the sensitivity of $\tilde{a}_\ell(\beta)' \delta$ and $M_{\ell\ell}(\beta)$ to changes in β as β moves away from

β_0 . Otherwise, there may be another $\tilde{\beta} \neq \beta_0$ where the variation in $\tilde{a}_\ell(\tilde{\beta})'\delta$ is cancelled by the sensitivity terms to the extent that $m_n^{\text{CF}}(\tilde{\beta}) = 0$ as well, so that $\hat{\beta}^{\text{CF}}$ may not be consistent for β_0 .

In the peer effects setting, the identifying variation in $\tilde{a}_\ell(\beta)'\delta$ comes from changes in peer group quality that are not fully captured by a linear combination of $x_\ell + a_\ell\beta$. This variation is generated by job mobility and worker turnover, as described in Section 2.2. Assumption 3(iv), in particular, requires such variation to be large enough relative to the sensitivity of the nuisance-parameter estimation and bias correction to the choice of β .

Under these conditions, the cross-fit estimator is consistent:

Theorem 1 (Consistency of $\hat{\beta}^{\text{CF}}$). *Under Assumptions 1, 2, and 3,*

$$\hat{\beta}^{\text{CF}} \xrightarrow{p} \beta_0 \quad \text{as } n \rightarrow \infty.$$

Theorem 1 shows that the peer effect β_0 can be consistently estimated by the cross-fit correction. A practical advantage is that the correction does not depend on the empirical specification. By contrast, the estimator of Braun and Verdier (2023) requires instruments for peer-average outcomes $\bar{y}_{(i)t}$ that may not be available in every application.

4 Inference

Hypothesis tests regarding the value of β_0 and confidence intervals for β_0 can be constructed based on a normal approximation to the distribution of $\hat{\beta}^{\text{CF}}$. Such inference requires an estimator of the variance $V_n(\beta) = \mathbb{V}[\hat{m}_n^{\text{CF}}(\beta) | X, A]$. This section proposes the variance estimator $\hat{V}_n(\beta)$ and establishes the validity of the resulting inference procedure.

4.1 Variance estimator

To introduce and motivate our proposed variance estimator $\hat{V}_n$, it is useful first to give two separate U-statistic representations of $\hat{m}_n^{\text{CF}}$. The first representation is symmetric in the sense that we write $\hat{m}_n^{\text{CF}} = \sum_{\ell=1}^n \sum_{k \neq \ell} U_{\ell k}^S y_\ell y_k$ and the order of the subscripts on the kernel function $U_{\ell k}^S$ 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 (7), which yields

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

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

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

Both representations are needed for the following expression, which writes the variance $V_n(\beta_0)$ in terms of both the symmetric and asymmetric kernel functions:

$$V_n(\beta_0) = 2\mathbb{E} \left[\sum_{\ell=1}^n \left(\sum_{k \neq \ell} U_{\ell k}^S(\beta_0) y_k \right) \left(\sum_{m \neq \ell} U_{\ell m}^A(\beta_0) y_m \right) \sigma_\ell^2 | X, A \right] - \mathbb{E} \left[\hat{m}_n^{\text{CF}}(\beta_0) \right]^2. \quad (9)$$

The final term, the squared expectation of the sample moment, is zero at β_0 . We include it to clarify why our variance estimator subtracts $(\hat{m}_n^{\text{CF}}(\beta))^2$ when evaluated at a β where the sample moment is non-zero.

Our proposed variance estimator drops the expectations present in (9) and replaces the unknown individual error variances $\{\sigma_\ell^2\}_{\ell=1}^n$ with cross-fit analogs. However, because outcome variables already enter the expression in (9), the use of leave-one-out cross-fit estimators may not suffice for consistent variance estimation. We therefore rely on leave-three-out estimators in the construction of $\hat{V}_n$. Specifically, we use

$$\hat{\sigma}_{\ell, -km}^2(\beta) = y_\ell (y_\ell - (x_\ell + a_\ell \beta)' \hat{\delta}_{(\ell km)}(\beta)) \quad (10)$$

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

$$\hat{V}_n(\beta) = 2 \sum_{\ell=1}^n \sum_{k \neq \ell} \sum_{m \neq \ell} U_{\ell k}^S(\beta) U_{\ell m}^A(\beta) y_k y_m \hat{\sigma}_{\ell, -km}^2(\beta) - \left(\hat{m}_n^{\text{CF}}(\beta) \right)^2. \quad (11)$$

Because the leave-three-out variance estimator $\hat{\sigma}_{\ell, -km}^2(\beta_0)$ is unbiased for σ_ℓ^2 and independent of the outcomes y_k and y_m , the variance estimator is itself unbiased at the truth:

Proposition 2 (Unbiasedness of leave-three-out cross-fit estimators). *Under Assumptions 1 and 2, if $\hat{\sigma}_{\ell, -km}^2(\beta_0)$ is well-defined for all ℓ , $k \neq \ell$, and $m \neq \ell$,*

$$\mathbb{E}[\hat{V}_n(\beta_0) \mid X, A] = V_n(\beta_0).$$

4.2 Consistency and Asymptotic Normality

The consistency of $\hat{V}_n$ and the asymptotic normality of $\hat{\beta}^{\text{CF}}$ require additional moment and variance-estimator regularity conditions. Let $\mathcal{N}_0$ be a fixed neighborhood of β_0 contained in $\mathcal{B}$. Define the leave-two-out and leave-three-out determinants for distinct observations:

$$D_{\ell k}(\beta) = \begin{vmatrix} M_{\ell\ell}(\beta) & M_{\ell k}(\beta) \\ M_{\ell k}(\beta) & M_{kk}(\beta) \end{vmatrix},$$

$$D_{\ell km}(\beta) = \begin{vmatrix} M_{\ell\ell}(\beta) & M_{\ell k}(\beta) & M_{\ell m}(\beta) \\ M_{\ell k}(\beta) & M_{kk}(\beta) & M_{km}(\beta) \\ M_{\ell m}(\beta) & M_{km}(\beta) & M_{mm}(\beta) \end{vmatrix},$$

and $D_{kk}(\beta) = 0$ and $D_{\ell k k}(\beta) = D_{\ell k}(\beta)$ for $\ell \neq k$ by convention, where $|\cdot|$ denotes the determinant of a matrix.

Assumption 4 (Regularity for inference). (i) Bounded eighth moments: *there exist constants $C_8 < \infty$ and $N < \infty$ such that, for all $n \geq N$,*

$$\max_{\ell \leq n} \mathbb{E}[|\varepsilon_\ell|^8 \mid X, A] \leq C_8.$$

⁵When k is equal to m , $\hat{\sigma}_{(\ell kk)}(\beta)$ 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 $\hat{\sigma}_{\ell, -kk}^2$.

(ii) Leave-two-out and leave-three-out conditioning:

$$\liminf_{n \rightarrow \infty} \inf_{\beta \in \mathcal{N}_0} \min_{\ell, k, m: k, m \neq \ell} D_{\ell km}(\beta) > 0.$$

Part (i) strengthens the corresponding condition in Assumption 1 to account for higher-order interactions of the error terms reflected in the formula for $\hat{V}_n(\beta_0)$. Part (ii) requires that the leave-two-out estimators $\sigma_{(lk)}^2$ and the leave-three-out estimators $\sigma_{(lkm)}^2$ are well-defined uniformly over n and locally around β_0 . See Section 4.3 for a discussion of when this condition does not hold.

The variance estimator $\hat{V}_n$ is consistent, and when combined with the consistency of $\hat{\beta}^{\text{CF}}$ (Theorem 1), we obtain asymptotic normality for the studentized estimator. The following theorem summarizes this result, justifying a standard inference procedure for β_0 :

Theorem 2 (Asymptotic inference). *Suppose Assumptions 1, 2, 3, and 4 hold, and that*

$$\liminf_{n \rightarrow \infty} n^{-1} V_n(\beta_0) > 0 \quad \text{and} \quad \liminf_{n \rightarrow \infty} \inf_{\beta \in \mathcal{N}_0} n^{-1} |\nabla_{\beta} m_n^{\text{CF}}(\beta)| > 0.$$

Then:

(i) [Variance estimator consistency] $\hat{V}_n(\beta_0)/V_n(\beta_0) \xrightarrow{p} 1$ and $\hat{V}_n(\hat{\beta}^{\text{CF}})/V_n(\beta_0) \xrightarrow{p} 1$ as $n \rightarrow \infty$.

(ii) [Studentized asymptotic normality]

$$\frac{\nabla_{\beta} \hat{m}_n^{\text{CF}}(\hat{\beta}^{\text{CF}})}{\sqrt{\hat{V}_n(\hat{\beta}^{\text{CF}})}} (\hat{\beta}^{\text{CF}} - \beta_0) \xrightarrow{d} N(0, 1)$$

as $n \rightarrow \infty$.

For a pre-specified level $\alpha \in (0, 1)$, the Wald $(1 - \alpha) \cdot 100\%$ confidence interval for β_0 is

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

where $z_{\alpha/2}$ is the $\alpha/2$ quantile of the standard normal distribution; by Theorem 2, C_{α}^{W} has asymptotic coverage $1 - \alpha$. Equivalently, C_{α}^{W} inverts the two-sided test of

$H_0 : \beta_0 = c$ that rejects when $(\nabla_{\beta} \hat{m}_n^{\text{CF}}(\hat{\beta}^{\text{CF}}))^2 \hat{V}_n(\hat{\beta}^{\text{CF}})^{-1}(\hat{\beta}^{\text{CF}} - c)^2 > z_{\alpha/2}^2$.

Remark 3. In the applied literature that relies on the non-linear least squares estimator $\hat{\beta}^{\text{NLLS}}$, the standard practice for inference is to rely on bootstrapping (e.g., [Arcidiacono et al., 2012](#); [Cornelissen et al., 2017](#)). There is no theoretical justification for the bootstrap in settings like ours, and the simulations in Section 5 indicate that the wild bootstrap understates sampling variability and undercovers. Our proposed variance estimator $\hat{V}_n$, on the other hand, provides a viable alternative to the bootstrap, and Theorem 2 provides the theoretical justification for valid inference when $\sqrt{\hat{V}_n(\hat{\beta}^{\text{CF}})}/\nabla_{\beta} \hat{m}_n^{\text{CF}}(\hat{\beta}^{\text{CF}})$ is used as the standard error for $\hat{\beta}^{\text{CF}}$.

4.3 Practical considerations

As with the leave-one-out estimator for σ_l^2 (7), the leave-three-out estimator (10) can be computed without explicitly constructing leave-three-out samples; the following representations can be leveraged for computation. As shown in [Anatolyev and Sølvyten \(2023\)](#), we have

$$\begin{aligned} & y_l - (x_l + a_l \beta)' \hat{\delta}_{(lk)}(\beta) \\ &= \frac{M_{kk}(\beta)(y_l - (x_l + a_l \beta)' \hat{\delta}^{\text{LS}}(\beta)) - M_{lk}(\beta)(y_k - (x_k + a_k \beta)' \hat{\delta}^{\text{LS}}(\beta))}{D_{lk}(\beta)}, \end{aligned}$$

and

$$\begin{aligned} & y_l - (x_l + a_l \beta)' \hat{\delta}_{(lkm)}(\beta) \\ &= \frac{(y_l - (x_l + a_l \beta)' \hat{\delta}^{\text{LS}}(\beta))}{D_{lkm}(\beta)/D_{km}(\beta)} \\ &= \frac{M_{lk}(\beta)(y_k - (x_k + a_k \beta)' \hat{\delta}_{(lk)}(\beta)) + M_{lm}(\beta)(y_m - (x_m + a_m \beta)' \hat{\delta}_{(lm)}(\beta))}{D_{lkm}(\beta)/D_{km}(\beta)}. \end{aligned}$$

When the number of observations is moderate — n on the order of a few tens of thousands — these representations compute $\hat{\sigma}_{\ell, -km}^2(\beta)$, and hence $\hat{V}_n(\beta)$, more efficiently than directly constructing leave-three-out samples.

However, if the number of observations is considerably larger, exact computation of $\hat{V}_n$ becomes 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 $r_{\ell k} =$

$M_{\ell k}/\sqrt{M_{\ell\ell}M_{kk}}$, which is bounded by one in absolute value, we can write

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

The truncated approximation $\hat{V}_n^{(tr)}$ is then given by replacing $y_k y_m \hat{\sigma}_{\ell, -km}^2$ with $\Upsilon_{\ell km}$ in the expression for $\hat{V}_n$:

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

As shown in Section A, $\hat{V}_n^{(tr)}$ can be expressed in matrix form and computed more efficiently than the exact $\hat{V}_n$. We note, however, that the consistency of $\hat{V}_n^{(tr)}$ is not guaranteed in general, and we therefore recommend using $\hat{V}_n$ whenever computationally feasible. We leave the theoretical analysis of the consistency of $\hat{V}_n^{(tr)}$ to future work.

Finally, we note that some of the leave-three-out estimators $\hat{\sigma}_{\ell, -km}^2$ may be ill-defined or unstable when $D_{lkm}(\beta)$ is close to zero, i.e., when Assumption 4 (ii) is violated. It is still possible to conduct valid inference when such (l, k, m) combinations are not too common, either by (i) dropping them from the sample or (ii) replacing $\hat{\sigma}_{\ell, -km}^2$ with a conservative estimate. See Section D for a formal treatment of the second approach.

5 Revisiting Arcidiacono et al. (2012): classroom peer effects

We illustrate our method by replicating and extending the seminal application of Arcidiacono et al. (2012), who estimated classroom peer effects using student transcript data. We use the universal transcript data from the University of Wisconsin–Madison, with a sharp natural experiment: in Spring 2020, the COVID-19 pandemic forced all undergraduate teaching online, plausibly eliminating the in-classroom peer interactions that the canonical model captures. This section presents the classroom

benchmark; Section 6 applies the estimator to workplace peer effects using the Veneto matched employer–employee panel. We also provide Monte Carlo simulations calibrated to the UW Spring 2019 design that compare the non-linear least squares estimator with wild-bootstrap standard errors against our cross-fit estimator with the proposed standard errors.

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 universal coverage of students. The database contains multiple records, including demographics, high school test scores, and transcript information.

Sample selection We focus on courses taken only by undergraduates. In the UW system these are courses numbered below 300 (e.g., Econ 101), since graduate students may enroll in courses numbered 300 and above. Given that 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

Following Arcidiacono et al. (2012), we define the *peer group* as all the students in the same discussion section in the same course. Given that undergraduate courses typically have large class sizes, the university typically assigns multiple teaching assistants to hold discussion sections for each course. The discussion section usually 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 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 that

⁶Arcidiacono 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.

this situation is unprecedented and online learning tools are new to every student, we expect the peer effect to be sharply reduced, if not eliminated. Anecdotally, during that semester nearly all students kept their cameras off in discussion sections, and in-class interaction nearly disappeared. We also explore the Spring semester of the academic year 2018/2019 as our benchmark semester, as the courses offered for these two semesters are almost identical.

Summary statistics Table 1 reports summary statistics for the connected estimation sample. Peer-group sizes vary substantially (mean around 16, standard deviation near 28), and about 70% of students take two or more courses; the online term (Spring 2020) has higher and less variable grades than the pre-pandemic term (mean 3.68 versus 3.39), suggesting more lenient grading during the pandemic.

Table 1: Summary statistics, connected estimation sample

	Spring 2019		Spring 2020	
	Mean	Std. Dev.	Mean	Std. Dev.
<i>Student level</i>				
Grade points	3.39	0.80	3.68	0.55
Courses taken	2.46	1.29	2.35	1.25
Takes 2+ courses	0.70	0.46	0.67	0.47
<i>Peer-group level</i>				
Peer-group size	17.2	28.7	15.7	27.2
Student-course observations	54,944		52,009	
Number of students	22,311		22,098	
Number of courses	1,312		1,337	
Number of peer groups	3,186		3,309	

Notes: Connected estimation sample for each term (the input to the cross-fit estimator), after restricting to undergraduate letter grades and to peer groups of at least two students in a single connected component. Peer groups are discussion sections; “Takes 2+ courses” is the share of students enrolled in two or more courses; grade points are on the 4.0 scale.

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

$$y_{ij} = \alpha_i + \bar{\alpha}_{(i)j} \cdot \beta + \psi_j + \varepsilon_{ij}, \quad (12)$$

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, we treat each professor’s offering as a distinct course. 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 substantial, which is important as it serves as the key identifying variation for the peer effect β_0 .

An extension that adds an endogenous peer effect through peer outcomes (in the spirit of [Bramoullé et al., 2009](#); [Bramoullé, Djebbari and Fortin, 2020](#)) is provided in [Appendix E.2](#) as an empirical robustness exercise. Because the formal asymptotic theory for the joint estimator extends beyond the scope of this paper, we keep the main text focused on [\(12\)](#).

5.2 Peer effects estimates

[Table 2](#) reports estimates from equation [\(12\)](#) in the Spring semesters of 2019 and 2020, comparing the non-linear least squares (NLLS) estimator of [Arcidiacono et al. \(2012\)](#) with our cross-fit (CF) estimator.

Consider first the Spring 2019 semester, before the pandemic. The NLLS estimate of β is about 0.25, with a wild-bootstrap standard error of 0.024. 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.⁷ Our proposed cross-fit (CF) estimate is 0.17, which is around 30 percent smaller than the NLLS estimate. Our proposed standard error is 0.033, which is roughly a third 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. Consistent with [Section 2](#), the plug-in variance of peer quality is biased upward, so the bias-corrected estimate $\hat{\sigma}_{\bar{\alpha}}$ (computed following [Kline et al., 2020](#)) is considerably smaller.

⁷For a computational comparison of the cross-fit estimator, our NLLS estimator, and the iterative NLLS estimator of [Arcidiacono et al. \(2012\)](#), see [Appendix E.1](#).

The implied one-standard-deviation effect is 0.031 grade points under our method, against 0.052 under NLLS with the plug-in variance — about 40% smaller.

Table 2: UW-Madison register data for spring semesters in 2019 and 2020

	Spring 2019		Spring 2020	
	NLLS	CF	NLLS	CF
$\hat{\beta}$	0.249 (0.024)	0.167 (0.033)	0.049 (0.023)	−0.002 (0.032)
$\hat{\sigma}_{\hat{\alpha}_{(i)j}}$ plug-in	0.210		0.157	
$\hat{\sigma}_{\hat{\alpha}_{(i)j}}$ KSS		0.189		0.142
1-sd effect	0.052	0.031	0.008	−0.000

Notes: Wild bootstrap standard errors for NLLS (100 replications). The proposed standard errors based on $\hat{V}_n$ for cross-fit. The full table including the endogenous-peer-effect extension appears in Appendix E.2.

We now turn to the Spring 2020 semester, when the COVID-19 pandemic shifted all undergraduate teaching online. As we conjectured above, social interactions during discussion sections were sharply reduced, so the classroom peer effect should fall substantially, if not vanish. The NLLS estimate suggests that a small but statistically significant positive effect remains: a one-standard-deviation increase in peer ability raises the course grade by 0.008 grade points. Using our method, the CF estimate is close to zero with a point estimate of −0.002 and a standard error of 0.032, statistically insignificant. The implied one-standard-deviation effect is essentially zero.

5.3 Simulation exercises

The classroom estimates demonstrate the empirical bite of the bias correction in a single sample. The next step is to verify that the gap between $\hat{\beta}^{\text{NLLS}}$ and $\hat{\beta}^{\text{CF}}$ is a systematic consequence of heteroskedasticity. In this subsection, we conduct a Monte Carlo exercise calibrated to the Spring 2019 specification (12) to compare the two estimators across 1,000 simulated draws.

The calibration takes the CF point estimate $\beta_0 = 0.167$ as the data-generating parameter, with the fitted values $(\hat{\alpha}, \hat{\psi}_j)$ from the CF estimation supplying the linear component of the regression. Errors are drawn independently as $\varepsilon_\ell \sim N(0, \tilde{\sigma}_\ell^2)$, with the variance function calibrated to the data in two steps. First, we form the leave-out

variance estimates $\hat{\sigma}_\ell^2 = y_\ell \hat{\varepsilon}_\ell / M_{\ell\ell}$ from the CF fit, which are unbiased for σ_ℓ^2 — unlike squared residuals, which are shrunk by the high-dimensional fixed effects — but noisy at the observation level. Second, since the individual $\hat{\sigma}_\ell^2$ can be negative, we smooth them through the positive model $\tilde{\sigma}_\ell^2 = \exp(\eta_1 \nabla_\beta M_{\ell\ell} + \eta_2 M_{\ell\ell} + \eta_3 \ln|\mathcal{P}_\ell| + \eta_0)$, where $|\mathcal{P}_\ell|$ is the size of observation ℓ 's peer group, fit by least squares and evaluated at β_0 , with the resulting standard deviations capped at their 99th percentile. Modeling the variance as a function of the leverage $M_{\ell\ell}$, its sensitivity $\nabla_\beta M_{\ell\ell}$, and the peer-group size preserves the empirical dependence of error variances on leverage — precisely the feature that renders NLLS inconsistent through the second term of the decomposition (5) — so the design confronts both estimators with the form of heteroskedasticity that the data exhibit. For each draw, we re-estimate both $\hat{\beta}^{\text{NLLS}}$ (with wild-bootstrap standard errors, following [Arcidiacono et al., 2012](#)) and $\hat{\beta}^{\text{CF}}$ (with the proposed standard errors based on $\hat{V}_n$), and record three metrics in Table 3: (i) bias of the point estimator relative to β_0 , (ii) ratio of the average standard error to the realized standard deviation of estimates across draws, and (iii) coverage of the 95% nominal confidence interval.

Table 3: Simulations using UW-Madison 2019 spring semester

	NLLS	CF
Point estimator:		
Bias	0.042	−0.006
Standard deviation	0.028	0.025
Bias/SD	1.523	−0.226
Standard error:		
Standard error/SD - 1	−27.0%	6.4%
Coverage:		
Nominal 95% CI	45.7%	94.1%

Notes: Wild bootstrap standard errors for NLLS. Proposed standard errors based on $\hat{V}_n$ for cross-fit.

Point estimator The NLLS bias of 0.042 corresponds to a bias-to-SD ratio of 1.52: under heteroskedasticity, the NLLS estimator has an order-one bias relative to its sampling variability. The CF estimator delivers a bias of −0.006 (bias-to-SD ratio −0.23), consistent with the finite-sample-unbiasedness of $\hat{m}_n^{\text{CF}}(\beta_0)$ in Proposition 1 and the consistency in Theorem 1. The two point estimators have similar realized

standard deviations (0.028 vs. 0.025), so the bias-to-SD difference is driven by bias rather than by efficiency.

Standard error The average wild-bootstrap standard error for NLLS is 27% smaller than the realized standard deviation of $\hat{\beta}^{\text{NLLS}}$ across draws. The wild bootstrap is standard practice in the applied literature but lacks theoretical justification in non-linear models like (4) (see also [Bickel and Freedman, 1983](#)), and it systematically understates sampling variability in the heteroskedastic design simulated here. The proposed standard error based on $\hat{V}_n$ tracks the realized standard deviation of $\hat{\beta}^{\text{CF}}$ closely, deviating by 6.4% on the conservative side, consistent with the unbiasedness of $\hat{V}_n$ established in Proposition 2.

Coverage The 95% confidence interval based on NLLS + wild bootstrap covers β_0 in only 45.7% of draws — a substantial undercoverage combining the NLLS bias with the understated standard error, and almost exactly what those two distortions imply: a normal approximation with the bias-to-SD ratio and standard-error shortfall reported in Table 3 predicts coverage of 46%. The confidence interval based on $\hat{\beta}^{\text{CF}}$ and $\hat{V}_n$ covers 94.1% of draws, close to the 95% nominal target and supporting the use of our procedure for inference in heteroskedastic peer-effects designs.

6 Workplace peer effects

We apply the cross-fit estimator to study workplace peer effects, using the Veneto matched employer–employee panel. The application has two purposes. First, it delivers a consistent, heteroskedasticity-robust estimate of the workplace peer effect in a large labor market, the setting of a growing applied literature (e.g., [Cornelissen et al., 2017](#); [Hong and Lattanzio, 2025](#)). Second, with that estimate in hand, it extends the canonical two-way worker–firm wage decomposition ([Abowd et al., 1999](#); [Card et al., 2018](#); [Kline et al., 2020](#)) to admit average coworker quality as a third component. Section 6.1 describes the data and sample; Section 6.2 reports the specification and peer-effect estimates; Section 6.3 reports the variance decomposition.

6.1 Data and sample

We use the Veneto Worker History (VWH) database, an administrative matched employer–employee panel covering the entire private-sector workforce in the Veneto region of northern Italy. The database is constructed from social-security records reported to the Italian National Institute of Social Security (INPS) and contains accurate annual earnings without top-coding, weeks worked, occupation (manager, white-collar, blue-collar, apprentice), and contract type. We restrict attention to the years 1995–2001, a window over which the Veneto labor market was stable, with nearly full employment (Tattara and Valentini, 2010; Serafinelli, 2019).⁸

We apply the standard sample restrictions in the AKM literature: keep each worker’s primary job in each year, restrict to workers aged 16–65, drop part-time and apprentice contracts, drop firms with more than 5,000 employees, and require peer groups with at least two workers. To support both the cross-fit estimator and the bias-corrected variance estimator of Kline et al. (2020) used in Section 6.3, we further restrict to the leave-one-out connected set of firms.⁹ The resulting sample contains approximately five million person-year observations covering more than one million workers and around seventy thousand firms. The typical worker earns about 830 euros per week and is 35 years old with roughly five years of tenure; firms are small (median nine employees), and about a third of workers change firm at least once over the sample window. Appendix F reports the full descriptive statistics.

Peer groups are defined at the firm $\times$ broad occupation $\times$ year level, where broad occupation refers to one of three professional categories (blue-collar, white-collar, manager).¹⁰ Additional sample-selection details, descriptive statistics, mobility rates by firm size, and the institutional setting of the Italian labor market over the sample

⁸We follow the standard practice in the AKM literature of restricting to a moderately long but not exceedingly long panel; this lets us observe enough worker mobility for identification while limiting the time over which worker fixed effects must be treated as fixed.

⁹The largest connected set is the largest set of firms linked by worker mobility, which is required to identify worker and firm fixed effects (Abowd et al., 1999). The leave-one-out connected set strengthens this condition by requiring connectedness to survive the removal of any single mover, and is what is needed to apply the bias correction of Kline et al. (2020). The leave-one-out sample is similar to the full sample in observable characteristics; see Appendix F for a comparison.

¹⁰Portugal et al. (2024) find that workplace peer-effect estimates are quantitatively similar across alternative occupation definitions in comparable Portuguese data, and Appendix F reports that our estimates are essentially unchanged when managers are reassigned to their pre-promotion occupation. Because the peer-group definition imperfectly captures the true pattern of interaction, the resulting measurement error attenuates our estimates (Nix, 2016; Cornelissen et al., 2017), so our coefficients can be read as a lower bound on the true peer effect.

window are deferred to Appendix F.

6.2 Peer-effect estimates

We estimate the wage equation

$$y_{it} = \alpha_i + \beta \cdot \bar{\alpha}_{(i)t} + \psi_{j(i,t)} + w'_{it}\gamma + \varepsilon_{it}, \quad (13)$$

where y_{it} is the log weekly wage, α_i is the worker fixed effect, $\psi_{j(i,t)}$ is the firm fixed effect, $\bar{\alpha}_{(i)t}$ is the average peer quality in worker i 's peer group, and w_{it} contains age-squared (normalized as age minus 40), tenure, tenure squared, log firm size, occupation indicators, and year indicators. The specification is exactly (3) in Section 2, applied to the VWH sample.

Identification of β in (13) draws on two complementary sources of mobility-induced variation in $\bar{\alpha}_{(i)t}$. Job switchers contribute variation through changes in their peer group; job stayers contribute variation through the entry and exit of coworkers around them. Both are quantitatively important in our sample: approximately 8% of workers change firms in a given year, and the average firm replaces around 20% of its workforce annually, with this turnover rate remaining around 15% even among firms with more than 1,000 employees (see Appendix F for the firm-size profile). These mobility statistics support the strong-identification condition of Assumption 3(iv), which requires that mobility-induced variation in peer-group composition translates into non-trivial variation in the unobserved $\bar{\alpha}_{(i)t}$.

We estimate (13) by the cross-fit estimator $\hat{\beta}^{\text{CF}}$ developed in Section 3. For comparison, we also report the iterated NLLS estimator $\hat{\beta}^{\text{NLLS}}$ of Arcidiacono et al. (2012). Following Kline et al. (2020), we bias-correct the variance of the peer-quality regressor $\hat{\sigma}_{\bar{\alpha}_{(i)t}}$ so that the implied one-standard-deviation effect is comparable across specifications. Table 4 reports the baseline estimates in Column (1).

The cross-fit estimator returns $\hat{\beta}^{\text{CF}} = 0.413$ with a proposed standard error of 0.035, statistically significant at the 1% level. The bias-corrected standard deviation of average peer quality is $\hat{\sigma}_{\bar{\alpha}_{(i)t}} = 0.191$, implying that a one-standard-deviation increase in coworker quality raises a worker's wage by 7.8%, comparable in magnitude to the return to one year of schooling in Italy over the same period (Lucifora, Comi and Brunello, 2000). The iterated NLLS estimator returns $\hat{\beta}^{\text{NLLS}} = 0.351$, about 15% smaller than the cross-fit estimate. The classroom application in Section 5 gives the

opposite case — in Spring 2019 the NLLS estimate is 0.249 against a cross-fit estimate of 0.167, so NLLS overstates the peer effect by about 50% — while here NLLS understates it by about 15%. Our workplace point estimate lies in the upper part of the range reported in comparable matched employer–employee studies (Cornelissen et al., 2017; Hong and Lattanzio, 2025), with differences across settings likely reflecting genuine variation in workplace interaction.

Table 4: Workplace peer effects in Veneto

	Baseline (1)	Common shocks (2)	Placebo peers (3)
$\hat{\beta}^{\text{CF}}$	0.413 (0.035)	0.367 (0.028)	0.016 (0.005)
$\hat{\sigma}_{\bar{\alpha}_{(i)t}}$ (KSS)	0.191	0.198	0.210
1-SD effect	7.8%	7.3%	0.3%
Worker FE	x	x	x
Firm FE	x		x
Occupation FE, year FE, controls	x		x
Firm $\times$ Occupation $\times$ Year FE		x	

Notes: Column (1) is the baseline specification. Column (2) replaces firm, occupation, and year fixed effects with a saturated firm $\times$ occupation $\times$ year fixed effect (peer-group fixed effect) to absorb peer-group-specific common shocks. Column (3) defines the placebo peer group as workers in *other* occupations within the same firm-year.

6.2.1 Common shocks

A potential concern with (13) is that peer-group-specific time-varying shocks may correlate with shocks to peer quality. For example, a firm may adopt new technology specific to one occupation, simultaneously raising wages and (estimated) worker quality within that occupation. To address this, we re-estimate (13) replacing $\psi_{j(i,t)}$ together with the occupation and year fixed effects by a saturated firm $\times$ occupation $\times$ year fixed effect $\xi_{j(i,t),o(i),t}$. This saturated control is the time-varying peer-group fixed effect discussed in Section 2.2; as explained there, the peer effect remains identified through the within-cell variation of stayers, because $\bar{\alpha}_{(i)t}$ excludes worker i .

Column (2) of Table 4 reports $\hat{\beta}^{\text{CF}} = 0.367$, broadly similar to the baseline. We retain (13) as our preferred specification because it uses both job-mover and job-stayer variation and preserves the link to the canonical AKM decomposition that we

will exploit in Section 6.3.

6.2.2 Placebo peers

If the estimated peer effect reflects coworker influence rather than firm- or occupation-level common shocks not absorbed by the controls, then re-defining the peer group as workers in *other* occupations within the same firm-year should return an effect close to zero. Column (3) of Table 4 reports $\hat{\beta}^{\text{CF}} = 0.016$ under this placebo definition, implying a one-standard-deviation effect of 0.3%. The placebo coefficient is two orders of magnitude smaller than the baseline, supporting the interpretation that the baseline estimate captures within-occupation peer influence rather than common variation across occupations within the firm.

6.3 Variance decomposition

The consistent peer-effect estimate from (13) allows us to extend the canonical AKM wage-variance decomposition to admit average coworker quality as a third component. Following Card et al. (2018) and Sorkin (2018), we use the “ensemble” decomposition,

$$\begin{aligned} \text{Var}(y_{it}) = & \text{Cov}(y_{it}, \alpha_i) + \text{Cov}(y_{it}, \psi_{j(i,t)}) + \text{Cov}(y_{it}, \beta \cdot \bar{\alpha}_{(i)t}) \\ & + \text{Cov}(y_{it}, w'_{it}\gamma) + \text{Cov}(y_{it}, \varepsilon_{it}), \end{aligned} \quad (14)$$

which assigns to each component the share of wage variance it explains in the regression. The covariance terms between α_i and $\psi_{j(i,t)}$ and between α_i and $\bar{\alpha}_{(i)t}$, after appropriate expansion of (14), identify worker–firm sorting and worker–coworker sorting respectively. As is well known, plugging fixed-effect estimates directly into a variance decomposition produces sizable bias under limited mobility (Andrews et al., 2008; Kline et al., 2020). We apply the bias-correction technique of Kline et al. (2020) to all variance and covariance components in (14); in our setting this is the same correction that delivers the bias-corrected $\hat{\sigma}_{\bar{\alpha}_{(i)t}}$ used in Section 6.2.

Table 5 reports the decomposition. Column (1) shows the canonical AKM decomposition (omitting average coworker quality) as a benchmark; Column (2) shows the decomposition from the peer model in (13). In the AKM model, worker fixed effects explain 53.3% of wage variance and firm fixed effects explain 18%, with worker–firm sorting accounting for an additional 10.8% (correlation 0.21). These magnitudes are

in line with what the AKM literature has documented across European and U.S. labor markets (e.g., [Card et al., 2018](#); [Sorkin, 2018](#)).

Table 5: Wage variance decomposition: AKM versus peer model

<i>Share of $\text{Var}(y_{it})$ explained by</i>	AKM model (1)	Peer model (2)
Worker effects, α_i	53.3%	47.1%
Firm effects, $\psi_{j(i,t)}$	18.0%	13.0%
Coworker effects, $\beta \cdot \bar{\alpha}_{(i)t}$		11.0%
Sorting: $2 \text{Cov}(\alpha_i, \psi_{j(i,t)})$	10.8%	3.4%
Sorting: $2 \text{Cov}(\alpha_i, \beta \cdot \bar{\alpha}_{(i)t})$		27.6%
Correlation: (worker, firm)	0.211	0.080
Correlation: (worker, coworker)		0.596

Notes: Ensemble decomposition of $\text{Var}(y_{it})$ following [Card et al. \(2018\)](#). Variance and covariance components are bias-corrected following [Kline et al. \(2020\)](#). Column (1) is the canonical AKM specification without $\bar{\alpha}_{(i)t}$; Column (2) is the peer model in (13) with $\hat{\beta}^{\text{CF}}$ as the peer coefficient. The control-variate and residual rows are omitted.

Two findings emerge from the peer-model column. First, average coworker quality explains 11% of wage variance — a share comparable to the 13% explained by firm effects in the same specification. The coworker channel is therefore as economically large as the firm channel, but it is invisible to the two-way model.¹¹ [Portugal et al. \(2024\)](#) report a smaller coworker share in Portuguese matched data; the difference is consistent with the bias-corrected estimator used here yielding a larger consistent estimate of β than the homoskedastic NLLS estimator used in their work.

Second, the structure of labor-market sorting shifts substantially once coworkers are admitted. The worker–firm sorting correlation that has anchored the AKM literature falls from 0.21 to 0.08, while a worker–coworker sorting correlation of 0.60 emerges as the dominant sorting margin in the data. The worker–coworker correlation is consistent in magnitude with what [Lopes de Melo \(2018\)](#) reports in Brazilian data using an indirect plug-in approach. The reading we favor is not that the AKM literature has been measuring sorting incorrectly, but that what the two-way model

¹¹The interpretation that the firm channel “shrinks” from 18% in the AKM model to 13% in the peer model is suggestive but not literal: the AKM and peer models are different statistical models, and the comparison reflects what each can attribute to firms when coworker quality is or is not separately identified. The clean reading is that the peer model identifies an 11% coworker channel of comparable economic size to firms, which the AKM specification cannot separate.

attributes to worker–firm sorting is, in substantial part, worker–coworker sorting that the model has no language for. With a consistent peer-effect estimate in hand, the decomposition can separate the two for the first time.

7 Conclusion

This paper develops estimation and inference for panel-data peer-effects models in which social interactions operate through an unobserved individual characteristic and the errors are heteroskedastic. The non-linear least squares estimator that anchors this literature is inconsistent under heteroskedasticity, and the sign of its bias is governed by the empirical design rather than by any general principle. Our cross-fit correction removes this bias: it yields a moment that is unbiased in finite samples at the truth, an estimator that is consistent under standard regularity conditions, and an analytic standard error built from a leave-three-out variance estimator. The stochastic approximations in Appendix A make the procedure feasible on administrative-scale data.

We first revisit the canonical classroom setting of [Arcidiacono et al. \(2012\)](#), comparing our estimator and inference against non-linear least squares. In the University of Wisconsin transcript data, the cross-fit estimate is about 30% below NLLS in a pre-pandemic benchmark, and during the Spring 2020 online semester — when in-person classroom interaction was largely absent — NLLS still reports a positive and significant peer effect while our estimate is close to zero and statistically insignificant; Monte Carlo experiments calibrated to this design confirm that the gap is a systematic consequence of heteroskedasticity, and that our standard error restores the coverage the wild bootstrap loses. We then apply the method to workplace peer effects in the Italian matched employer–employee panel, where the cross-fit estimate is about 15% *larger* than the iterated NLLS benchmark. The correction runs in opposite directions across the two settings, confirming that the sign of the bias is empirically driven.

A consistent peer coefficient also makes possible a wage-variance decomposition that the canonical two-way model cannot deliver. Extending the AKM decomposition to incorporate average coworker quality, we find that coworkers account for a share of wage variance comparable to that of firms, and that the dominant axis of sorting in the labor market runs between workers and their coworkers rather than between workers and firms. To our knowledge this is the first decomposition to separate these

two sorting channels using a consistent estimate of the peer coefficient, and it suggests that a meaningful part of what two-way models attribute to firms in fact reflects who a worker’s colleagues are. Finally, the method is not specific to peer averaging: it applies to non-linear least squares problems in which a low-dimensional parameter multiplies a latent individual effect, a structure shared by models of heterogeneous teacher value-added ([Hahn et al., 2023](#)).

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,” *Journal of Econometrics*, 2023, *236*, 105473.
- Andrews, Martyn J, Len Gill, Thorsten Schank, and Richard Upward, “High wage workers and low wage firms: negative assortative matching or limited mobility bias?,” *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 2008, *171* (3), 673–697.
- 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.
- Battisti, Michele, “High wage workers and high wage peers,” *Labour Economics*, 2017, *46* (February 2015), 47–63.
- Baum-Snow, Nathaniel, Nicolas Gendron-Carrier, and Ronni Pavan, “Local Productivity Spillovers,” *American Economic Review*, 2024, *114* (4), 1030–1069.

- Bayer, Patrick, Stephen L Ross, and Giorgio Topa, “Place of work and place of residence: Informal hiring networks and labor market outcomes,” *Journal of Political Economy*, 2008, *116* (6), 1150–1196.
- 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.
- Braun, Martin and Valentin Verdier, “Estimation of spillover effects with matched data or longitudinal network data,” *Journal of Econometrics*, 2023, *233* (2), 689–714.
- Brune, Lasse, Eric Chyn, and Jason Kerwin, “Peers and Motivation at Work,” *Journal of Human Resources*, 2020, pp. 0919–10416R2.
- Burke, Mary A. and Tim R. Sass, “Classroom Peer Effects and Student Achievement,” *Journal of Labor Economics*, 2013, *31* (1), 51–82.
- Caeyers, Bet and Marcel Fafchamps, “Exclusion bias in the estimation of peer effects,” Technical Report, National Bureau of Economic Research 2016.
- Card, David, Ana Rute Cardoso, Joerg Heining, and Patrick Kline, “Firms and labor market inequality: Evidence and some theory,” *Journal of Labor Economics*, 2018, *36* (S1), S13–S70.
- , Jörg Heining, and Patrick Kline, “Workplace Heterogeneity and the Rise of West German Wage Inequality*,” *The Quarterly Journal of Economics*, 05 2013, *128* (3), 967–1015.
- Cornelissen, Thomas, Christian Dustmann, and Uta Schönberg, “Peer effects in the workplace,” *American Economic Review*, 2017, *107* (2), 425–456.

- de Jong, Peter, “A central limit theorem for generalized quadratic forms,” *Probability Theory and Related Fields*, 1987, *75* (2), 261–277.
- de Melo, Rafael Lopes, “Firm Wage Differentials and Labor Market Sorting: Reconciling Theory and Evidence,” *Journal of Political Economy*, 2018, *126* (1), 313 – 346.
- Dix-Carneiro, Rafael and Brian K. Kovak, “Trade Liberalization and Regional Dynamics,” *American Economic Review*, 2017, *107* (10), 2908–2946.
- Fletcher, Jason M, Stephen L Ross, and Yuxiu Zhang, “The consequences of friendships: Evidence on the effect of social relationships in school on academic achievement,” *Journal of Urban Economics*, 2020, *116*, 103241.
- Graham, Bryan S, “Identifying social interactions through conditional variance restrictions,” *Econometrica*, 2008, *76* (3), 643–660.
- Guiso, Luigi, Luigi Pistaferri, and Fabiano Schivardi, “Insurance within the firm,” *Journal of Political Economy*, 2005, *113* (5), 1054–1087.
- 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.
- Hahn, Jinyong, John D. Singleton, and Nese Yildiz, “Identification of Non-Additive Fixed Effects Models,” Working Paper 31384, National Bureau of Economic Research 2023.
- 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.
- Holden, Richard, Michael Keane, and Matthew Lilley, “Peer Effects on the United States Supreme Court,” *Quantitative Economics*, 2021, *12* (3), 981–1019.
- Hong, Long and Salvatore Lattanzio, “The Peer Effect on Future Wages in the Workplace,” *Journal of Applied Econometrics*, 2025, *40* (5), 521–539.

- 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.
- Iversen, Torben, “Wage bargaining, central bank independence, and the real effects of money,” *International Organization*, 1998, 52 (3), 469–504.
- 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.
- la Roca, Jorge De and Diego Puga, “Learning by Working in Big Cities,” *Review of Economic Studies*, 2017, 84 (1), 106–142.
- 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.
- Lucifora, Claudio, Simona Comi, and Giorgio Brunello, “The returns to education in Italy: A new look at the evidence,” Technical Report, IZA Discussion Papers 2000.
- 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.
- Messina, Julián, Anna Sanz de Galdeano, and Anastasia Terskaya, “Birds of a Feather Earn Together: Gender and Peer Effects at the Workplace,” *Journal of Human Resources*, 2026. Forthcoming.
- 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.
- Neururer, Thaddeus and Estelle Y. Sun, “Quantifying the Effect of Information and Ability Spillovers on Analyst Earnings Forecast Accuracy,” Working Paper 2021. SSRN Working Paper.
- Nix, Emily, “Learning Spillovers in the Firm,” *Job Market Paper*, 2016.

- 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.
- Portugal, Pedro, Hugo Reis, Paulo Guimarães, and Ana Rute Cardoso, “What Lies behind the Returns to Schooling: The Role of Labor Market Sorting and Worker Heterogeneity,” *The Review of Economics and Statistics*, 2024. Forthcoming.
- Serafinelli, Michel, ““Good” firms, worker flows, and local productivity,” *Journal of Labor Economics*, 2019, *37* (3), 747–792.
- Song, Jae, David J Price, Fatih Guvenen, Nicholas Bloom, and Till Von Wachter, “Firming up inequality,” *The Quarterly journal of economics*, 2019, *134* (1), 1–50.
- Sorkin, Isaac, “Ranking firms using revealed preference,” *The quarterly journal of economics*, 2018, *133* (3), 1331–1393.
- Tattara, Giuseppe and Marco Valentini, “Turnover and excess worker reallocation. The Veneto labour market between 1982 and 1996,” *Labour*, 2010, *24* (4), 474–500.
- Thiemann, Petra, “The Persistent Effects of Short-Term Peer Groups on Performance: Evidence from a Natural Experiment in Higher Education,” *Management Science*, 2022, *68* (2), 1131–1148.

Appendix A Computation on large datasets

This appendix spells out the algebraic details needed for implementation of the proposed point estimator and confidence set. We first provide matrix representations of the moment function $\hat{m}_n^{\text{CF}}$, its derivative $\nabla_\beta \hat{m}_n^{\text{CF}}$, and the truncated variance estimator $\hat{V}_n^{(tr)}$ introduced in Section 4; it is the variant used when exact computation of $\hat{V}_n$ is infeasible. We then introduce stochastic approximations to those quantities that make computation feasible at scale.

To facilitate a description of $\hat{m}_n^{\text{CF}}$, $\nabla_\beta \hat{m}_n^{\text{CF}}$, 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 $\hat{\delta}^{\text{LS}}(\beta) = S(\beta)^{-1}(X + A\beta)'y$ are the residuals and estimated coefficients from a linear regression of y on $X + A\beta$. For any vector v , $\text{diag}[v]$ 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 $\Lambda(\beta) = \text{diag}[\nabla_\beta \log(M^{(d)}(\beta))]$, we can then write $\hat{m}_n^{\text{CF}} = -2\hat{\varepsilon}'A\hat{\delta}^{\text{LS}} - \hat{\varepsilon}'\Lambda y$. Similarly, we have the representation

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

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

We then have

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

A.2 Stochastic approximations

Exact computation of the log-derivative matrix functions Λ and $\nabla_\beta \Lambda$ appearing in $\hat{m}_n^{\text{CF}}$, U^A , U^S , and $\nabla_\beta \hat{m}_n^{\text{CF}}$ 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 $(r_1, \dots, r_p) \in \mathbb{R}^{n \times p}$ be a random matrix with *i.i.d.* Rademacher entries (discrete uniform random variables with support $\{-1, 1\}$).

Our stochastic approximation to the vector $M^{(d)}$ is¹²

$$\check{M}^{(d)} = \left\{ \sum_{s=1}^p (Mr_s \odot Mr_s) \right\} \odot \left\{ \sum_{s=1}^p (Mr_s \odot Mr_s) + (Pr_s \odot Pr_s) \right\}$$

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 4. 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 $\mathbf{1}_n = (1, \dots, 1)' \in \mathbb{R}^n$. 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](#)

¹²For any observations where the entries of $\check{M}^{(d)}$ are below 0.01, we replace them by 0.01; at those observations we also set the corresponding entry of $\check{\Lambda}$ to zero, so that they contribute no bias correction, because a near-zero leverage renders the cross-fit variance estimate and its log-derivative numerically unstable.

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 Slvsten (2021) for a derivation of $\check{M}^{(d)}$ as a feasible version of a minimum variance combination of two separate stochastic approximations to $\check{M}^{(d)}$.

Writing the numerator and denominator of $\check{M}^{(d)}$ as $\check{N} = \sum_{s=1}^p (Mr_s \odot Mr_s)$ and $\check{d} = \sum_{s=1}^p [(Mr_s \odot Mr_s) + (Pr_s \odot Pr_s)]$, so that $\check{M}^{(d)} = \check{N} \oslash \check{d}$, we approximate Λ by the exact derivative of $\check{M}^{(d)}$ rather than by a finite difference:

$$\begin{aligned}\check{\Lambda}(\beta) &= \text{diag}[\check{N}' \oslash \check{N} - \check{d}' \oslash \check{d}], \\ \check{N}' &= -2 \sum_{s=1}^p (Mr_s) \odot (D + D')r_s, \\ \check{d}' &= \check{N}' + 2 \sum_{s=1}^p (Pr_s) \odot (D + D')r_s,\end{aligned}$$

where $\nabla_{\beta}(Mr_s) = -(D + D')r_s$. Each $(D + D')r_s$ reuses the factorization of $S(\beta)$ already formed for $\check{M}^{(d)}$, so $\check{\Lambda}$ is exact (no ϵ truncation) and needs no second factorization at $\beta + \epsilon$. 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 $\check{m}_n^{\text{CF}} = -2\hat{\epsilon}'A\hat{\delta}^{\text{LS}} - \hat{\epsilon}'\check{\Lambda}y$, $\check{U}^{\text{A}} = -(2D + M\check{\Lambda})$, $\check{U}^{\text{S}} = (\check{U}^{\text{A}} + \check{U}^{\text{A}'})/2$, $\check{\sigma}^2 = y \odot \hat{\epsilon} \oslash \check{M}^{(d)}$,

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

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

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

Remark 5. 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 $U^S \odot M$. 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 6. The computationally most demanding part of evaluating $\check{m}_n^{\text{CF}}(\beta)$ and $\check{V}_n^{(tr)}(\beta)$ is to find the solutions to linear systems of equations of the kind $S(\beta)x = b$ for various values of b . To construct $\check{m}_n^{\text{CF}}(\beta)$ 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.¹³

Appendix B Derivations from a simple example

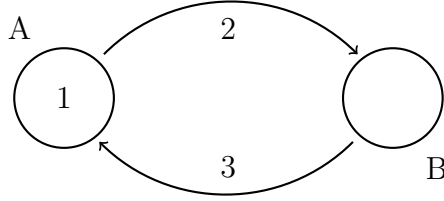
This appendix works through the simplest version of the model — three workers and two firms — to make concrete what the general treatment establishes abstractly: how worker mobility identifies β_0 ; how the cross-fit estimator specializes to a familiar averaged two-measurement (instrumental-variables) estimator; why least squares requires homoskedasticity and in which direction it is biased; and why the reduced data $(\mathcal{Y}, \mathcal{X}, \mathcal{Z})$ sacrifices no information about β_0 .

As a service to the reader we quickly state the main definitions used throughout the paper and proofs. Here y , X , and A stack the observations for y_ℓ , x'_ℓ , and a'_ℓ and σ^2 stack 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 , $\Lambda = \text{diag}[\nabla_\beta \log(M^{(d)})]$, $D = MAS^{-1}R'$, $U^A = -(2D + M\Lambda)$, and $U^S = (U^A + U^A')/2$. The non-linear least squares estimator is $(\hat{\beta}^{\text{NLLS}}, \hat{\delta}^{\text{LS}}) = \arg \min_{\beta \in \mathcal{B}, \delta \in \mathbb{R}^k} \sum_{\ell=1}^n (y_\ell - x'_\ell \delta - a'_\ell \delta \cdot \beta)^2$ while the key objective and moment functions are $\hat{Q}_n = y'My$, $\hat{m}_n = \nabla_\beta Q_n$, $\hat{m}_n^{\text{CF}} = \hat{m}_n - y'M\Lambda y$, $Q_n = \mathbb{E}[\hat{Q}_n | X, A]$, and $m_n^{\text{CF}} = \mathbb{E}[\hat{m}_n^{\text{CF}} | A, X]$. Finally, $\mathbf{1} = (1, \dots, 1)' \in \mathbb{R}^n$.

¹³It also depends on the structure of the model through the degree of sparsity in S .

B.1 Motivating example: three individuals, two firms

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 relies 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



Appendix Figure 1: Depiction of three individuals (denoted 1, 2, and 3) and their mobility between 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 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,

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

It is 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 homogeneous groups, such identifying variation will be present.

The reduction to $(\mathcal{Y}, \mathcal{X}, \mathcal{Z})$ discards no information about β_0 . We 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.

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

$$\beta_0 = \frac{1}{2} \frac{\mathbb{E}[\mathcal{Z}\mathcal{Y}]}{\mathbb{E}[\mathcal{Z}\mathcal{X}]} + \frac{1}{2} \frac{\mathbb{E}[\mathcal{X}\mathcal{Y}]}{\mathbb{E}[\mathcal{X}\mathcal{Z}]} \quad (15)$$

This simple estimator is a special case of the cross-fit estimator introduced in Section 3.

To highlight the differences and similarities between the cross-fit and least squares estimators, we first re-express the denominator of (15) so that

$$\beta_0 = \frac{\mathbb{E}[(\mathcal{Z} + \mathcal{X})\mathcal{Y}]}{\mathbb{E}[\mathcal{Z}^2 + \mathcal{X}^2] - \mathbb{E}[(\mathcal{Z} - \mathcal{X})^2]} \quad (16)$$

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 (16) as $4\sigma^2(\beta_0)$ where $\sigma^2(\beta)$ is a variance function that draws on both the movers and the stayer:

$$\sigma^2(\beta) = \frac{\mathbb{E}[(\mathcal{Z} - \mathcal{X})^2 + (\mathcal{Y} - \mathcal{X}\beta)^2 + (\mathcal{Y} - \mathcal{Z}\beta)^2]}{4(2 + \beta^2)}. \quad (17)$$

The non-linear least squares estimator minimizes the sample analog of (17). Additionally, we can describe the least squares estimator as a solution to the sample analog of a first order condition for minimization of (17), which looks exactly like (16), except that it uses $4\sigma^2(\beta)$ instead of $\mathbb{E}[(\mathcal{Z} - \mathcal{X})^2]$:

$$\beta = \frac{\mathbb{E}[(\mathcal{Z} + \mathcal{X})\mathcal{Y}]}{\mathbb{E}[\mathcal{Z}^2 + \mathcal{X}^2] - 4\sigma^2(\beta)}. \quad (18)$$

In the presence of heteroskedasticity, it is problematic to rely on an estimator that solves the sample analog of (18) 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.

The sample analogs of (15)–(18) are special cases of the general formulas in Section 3; the explicit connection is developed in Appendix B.3 below.

The structure exposed by (15) is the textbook measurement-error correction: $\mathcal{X}$ and $\mathcal{Z}$ are two independent noisy measurements of the same latent regressor $\alpha_3 - \alpha_2$, and each serves as instrument for the other. Section 3 explains how the general cross-fit moment $\hat{m}_n^{\text{CF}}$ in (8) extends this IV construction to a high-dimensional non-linear regression: leave-one-out projection plays the role of the “second noisy measurement,” and the recentering algebraically combines the two views as the IV moment in (15) does.

B.2 Bias direction of the least squares estimator in the simple example

We now show that the least squares estimator is 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, suppose that β^* is the unique global minimizer of $\sigma^2(\beta)$ and that the least squares estimator converges 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 (18) 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 $\sigma_{\mathcal{Y}}^2 \geq \max\{\sigma_{\mathcal{X}}^2, \sigma_{\mathcal{Z}}^2\}$, then we have $\mathbb{E}[\mathcal{Z}^2 + \mathcal{X}^2] = \lim_{|\beta| \rightarrow \infty} 4\sigma^2(\beta) > 4\sigma^2(\beta^*) \geq \mathbb{E}[(\mathcal{Z} - \mathcal{X})^2]$ and in turn

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

B.3 Connecting the simple example to the general formulas

Next, we connect the expressions in (15)–(18) 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 (17) 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 to be

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

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 (15).

Appendix C Proofs

This appendix collects formal proofs of the theorems and propositions stated in the main text.

C.1 Proof of Theorem 1

Throughout this proof all expectations and stochastic orders are conditional on (X, A) . Let

$$m_n^{\text{CF}}(\beta) = \mathbb{E}[\hat{m}_n^{\text{CF}}(\beta) \mid X, A].$$

By Proposition 1, $m_n^{\text{CF}}(\beta_0) = 0$. The argument is a Z-estimator consistency argument for the triangular array of moments.

Lemma 1 (Zero of a uniformly convergent moment). *Suppose that, for every $\varepsilon > 0$,*

$$\liminf_n \inf_{\beta \in \mathcal{B}: |\beta - \beta_0| \geq \varepsilon} n^{-1} |m_n^{\text{CF}}(\beta)| > 0,$$

and

$$\sup_{\beta \in \mathcal{B}} n^{-1} |\hat{m}_n^{\text{CF}}(\beta) - m_n^{\text{CF}}(\beta)| \xrightarrow{p} 0.$$

Then any exact zero $\hat{\beta}^{\text{CF}}$ of $\hat{m}_n^{\text{CF}}$ in $\mathcal{B}$ satisfies $\hat{\beta}^{\text{CF}} \xrightarrow{p} \beta_0$.

Proof. Fix $\varepsilon > 0$. By the separation condition, there are $\eta_\varepsilon > 0$ and N_ε such that, for $n \geq N_\varepsilon$,

$$\inf_{\beta \in \mathcal{B}: |\beta - \beta_0| \geq \varepsilon} n^{-1} |m_n^{\text{CF}}(\beta)| \geq \eta_\varepsilon.$$

On the event $\{|\hat{\beta}^{\text{CF}} - \beta_0| \geq \varepsilon\}$,

$$\begin{aligned} \eta_\varepsilon &\leq n^{-1} |m_n^{\text{CF}}(\hat{\beta}^{\text{CF}})| \\ &\leq \sup_{\beta \in \mathcal{B}} n^{-1} |\hat{m}_n^{\text{CF}}(\beta) - m_n^{\text{CF}}(\beta)| + n^{-1} |\hat{m}_n^{\text{CF}}(\hat{\beta}^{\text{CF}})|. \end{aligned}$$

The last term is zero by definition of $\hat{\beta}^{\text{CF}}$, and the first term is $o_p(1)$. Hence the probability of $\{|\hat{\beta}^{\text{CF}} - \beta_0| \geq \varepsilon\}$ converges to zero. $\square$

Assumption 3(iv) is the separation condition in Lemma 1. It remains to prove the uniform convergence condition.

For interpretation of this separation condition, define

$$q_\ell(\beta) = \tilde{a}_\ell(\beta)' \delta, \quad R_n(\beta) = 2 \sum_{\ell=1}^n q_\ell(\beta) \nabla_\beta q_\ell(\beta),$$

and

$$B_n(\beta) = \sum_{\ell=1}^n \frac{\nabla_\beta M_{\ell\ell}(\beta)}{M_{\ell\ell}(\beta)} \mu_\ell q_\ell(\beta), \quad \mu_\ell = \mathbb{E}[y_\ell \mid X, A].$$

Using

$$\mathbb{E}[\hat{\sigma}_\ell^2(\beta) \mid X, A] = \sigma_\ell^2 + (\beta_0 - \beta) \frac{\mu_\ell q_\ell(\beta)}{M_{\ell\ell}(\beta)},$$

the population CF moment can be written as

$$m_n^{\text{CF}}(\beta) = (\beta - \beta_0) \left[2 \sum_{\ell=1}^n q_\ell(\beta)^2 + (\beta - \beta_0) R_n(\beta) + B_n(\beta) \right].$$

Thus Assumption 3(iv) requires the leading identifying variation in $q_\ell(\beta)$ not to be cancelled by the sensitivity terms $R_n(\beta)$ and $B_n(\beta)$ outside any neighborhood of β_0 .

Lemma 2 (Uniform norm bounds). *Under Assumption 3(i)–(iii),*

$$\begin{aligned} \|A\|_2 &= O(\sqrt{n}), & \sup_{\beta \in \mathcal{B}} \|R(\beta)\|_2 &= O(\sqrt{n}), & \sup_{\beta \in \mathcal{B}} \|S(\beta)^{-1}\|_2 &= O(n^{-1}), \\ \sup_{\beta \in \mathcal{B}} \|D(\beta)\|_2 &= O(1), & \sup_{\beta \in \mathcal{B}} \|D(\beta)\|_F &= O(\sqrt{n}), \\ \sup_{\beta \in \mathcal{B}} \|\nabla_\beta M(\beta)\|_2 &= O(1), & \sup_{\beta \in \mathcal{B}} \|\nabla_\beta D(\beta)\|_2 &= O(1), \\ \sup_{\beta \in \mathcal{B}} \|\Lambda(\beta)\|_\infty &= O(1), & \sup_{\beta \in \mathcal{B}} \|\nabla_\beta \Lambda(\beta)\|_\infty &= O(1), \\ \sup_{\beta \in \mathcal{B}} \|U^A(\beta)\|_2 &= O(1), & \sup_{\beta \in \mathcal{B}} \|\nabla_\beta U^A(\beta)\|_2 &= O(1), & \sup_{\beta \in \mathcal{B}} \|U^A(\beta)\|_F &= O(\sqrt{n}), \\ \sup_{\beta \in \mathcal{B}} \|\nabla_\beta U^A(\beta)\|_F &= O(\sqrt{n}), & \sup_{\beta \in \mathcal{B}} \|\nabla_{\beta\beta} U^A(\beta)\|_2 &= O(1). \end{aligned}$$

Proof. Bounded row ℓ_1 norms imply bounded row ℓ_2 norms. Hence

$$\|A\|_2 \leq \|A\|_F = \left(\sum_{\ell=1}^n \|a_\ell\|_2^2 \right)^{1/2} \leq C_a \sqrt{n} = O(\sqrt{n}).$$

The same argument gives $\|X\|_2 = O(\sqrt{n})$, and compactness of $\mathcal{B}$ gives

$$\sup_{\beta \in \mathcal{B}} \|R(\beta)\|_2 \leq \|X\|_2 + \sup_{\beta \in \mathcal{B}} |\beta| \|A\|_2 = O(\sqrt{n}).$$

Assumption 3(ii) implies

$$\sup_{\beta \in \mathcal{B}} \|S(\beta)^{-1}\|_2 = O(n^{-1}).$$

Since $M(\beta)$ is an orthogonal projection, $\|M(\beta)\|_2 = 1$. Therefore

$$\|D(\beta)\|_2 \leq \|M(\beta)\|_2 \|A\|_2 \|S(\beta)^{-1}\|_2 \|R(\beta)\|_2 = O(1)$$

uniformly in β . Since $\text{rank}(D(\beta)) \leq K$,

$$\|D(\beta)\|_F \leq \sqrt{K} \|D(\beta)\|_2 = O(\sqrt{n}),$$

where the last equality uses Assumption 3(iii).

The identity $\nabla_\beta M(\beta) = -(D(\beta) + D(\beta)')$ gives

$$\sup_{\beta \in \mathcal{B}} \|\nabla_\beta M(\beta)\|_2 = O(1).$$

Next, differentiating $D = MAS^{-1}R'$ and using $\nabla_\beta R = A$ and

$$\nabla_\beta S^{-1} = -S^{-1}(A'R + R'A)S^{-1},$$

yields

$$\nabla_\beta D = (\nabla_\beta M)AS^{-1}R' - MAS^{-1}(A'R + R'A)S^{-1}R' + MAS^{-1}A'.$$

The three terms have operator norms bounded by

$$\begin{aligned} O(1) O(\sqrt{n}) O(n^{-1}) O(\sqrt{n}) &= O(1), \\ O(\sqrt{n}) O(n^{-1}) O(n) O(n^{-1}) O(\sqrt{n}) &= O(1), \\ O(\sqrt{n}) O(n^{-1}) O(\sqrt{n}) &= O(1), \end{aligned}$$

respectively, where $\|A'R + R'A\|_2 \leq 2\|A\|_2\|R\|_2 = O(n)$. Thus $\sup_{\beta \in \mathcal{B}} \|\nabla_\beta D(\beta)\|_2 = O(1)$.

For Λ , recall

$$\Lambda_{\ell\ell}(\beta) = \frac{\nabla_\beta M_{\ell\ell}(\beta)}{M_{\ell\ell}(\beta)}.$$

Assumption 3(i) bounds the denominator away from zero. Also

$$\nabla_\beta M_{\ell\ell}(\beta) = -2D_{\ell\ell}(\beta), \quad |D_{\ell\ell}(\beta)| \leq \|D(\beta)\|_2,$$

so $\sup_{\beta \in \mathcal{B}} \|\Lambda(\beta)\|_\infty = O(1)$. Moreover,

$$\nabla_\beta \Lambda_{\ell\ell}(\beta) = \frac{\nabla_{\beta\beta} M_{\ell\ell}(\beta)}{M_{\ell\ell}(\beta)} - \frac{\{\nabla_\beta M_{\ell\ell}(\beta)\}^2}{M_{\ell\ell}(\beta)^2}.$$

Since $\nabla_{\beta\beta} M = -(\nabla_\beta D + (\nabla_\beta D)')$, the previous bounds and bounded leverage imply $\sup_{\beta \in \mathcal{B}} \|\nabla_\beta \Lambda(\beta)\|_\infty = O(1)$.

Finally, $U^A = -(2D + M\Lambda)$. Hence

$$\|U^A\|_2 \leq 2\|D\|_2 + \|M\|_2\|\Lambda\|_\infty = O(1),$$

and

$$\|\nabla_\beta U^A\|_2 \leq 2\|\nabla_\beta D\|_2 + \|\nabla_\beta M\|_2\|\Lambda\|_\infty + \|M\|_2\|\nabla_\beta \Lambda\|_\infty = O(1).$$

For the Frobenius norm,

$$\|M\Lambda\|_F^2 = \text{tr}(\Lambda M' M \Lambda) = \text{tr}(\Lambda^2 M) \leq \|\Lambda\|_\infty^2 \text{tr}(M) = O(n),$$

and therefore $\|U^A\|_F \leq 2\|D\|_F + \|M\Lambda\|_F = O(\sqrt{n})$. The same product-rule calcula-

tions give

$$\sup_{\beta \in \mathcal{B}} \|\nabla_{\beta} D(\beta)\|_F = O(\sqrt{n}), \quad \sup_{\beta \in \mathcal{B}} \|\nabla_{\beta} M(\beta)\|_F = O(\sqrt{n}).$$

Combining these bounds with $\|\Lambda\|_{\infty} = O(1)$, $\|\nabla_{\beta} \Lambda\|_{\infty} = O(1)$, and $\text{tr}(M) = O(n)$ yields

$$\sup_{\beta \in \mathcal{B}} \|\nabla_{\beta} U^A(\beta)\|_F = O(\sqrt{n}).$$

Differentiating the expression for $\nabla_{\beta} U^A$ once more produces only finite sums of products of A , R , powers of S^{-1} , M , Λ , and their first derivatives. The preceding operator-norm bounds, together with bounded leverage, imply

$$\sup_{\beta \in \mathcal{B}} \|\nabla_{\beta\beta} U^A(\beta)\|_2 = O(1).$$

□

Lemma 3 (Uniform convergence of the CF moment). *Under Assumptions 1, 2, and 3,*

$$\sup_{\beta \in \mathcal{B}} n^{-1} |\hat{m}_n^{\text{CF}}(\beta) - m_n^{\text{CF}}(\beta)| \xrightarrow{p} 0.$$

Proof. Let $\mu = \mathbb{E}[y \mid X, A] = R(\beta_0)\delta$ and $\varepsilon = y - \mu$. Assumption 3(ii) implies

$$\max_{\ell \leq n} |\mu_{\ell}| \leq (C_x + |\beta_0|C_a) C_{\delta}, \quad \|\mu\|_2^2 = O(n).$$

The matrix representation in Section 4 gives

$$\hat{m}_n^{\text{CF}}(\beta) = y' U^A(\beta) y, \quad U^A(\beta) = -(2D(\beta) + M(\beta)\Lambda(\beta)).$$

The diagonal of U^A is zero because the diagonal of $2D$ is $-\nabla_{\beta} M_{\ell\ell}$ and the diagonal of $M\Lambda$ is $\nabla_{\beta} M_{\ell\ell}$. Hence, using conditional independence,

$$m_n^{\text{CF}}(\beta) = \mathbb{E}[y' U^A(\beta) y \mid X, A] = \mu' U^A(\beta) \mu.$$

Writing $U^S = (U^A + U^{A'})/2$, we have

$$\hat{m}_n^{\text{CF}}(\beta) - m_n^{\text{CF}}(\beta) = 2\mu'U^S(\beta)\varepsilon + \varepsilon'U^A(\beta)\varepsilon.$$

Both terms have conditional mean zero. For the linear term,

$$\text{Var}\left(2\mu'U^S(\beta)\varepsilon \mid X, A\right) \leq 4\|\mu\|_2^2\|U^S(\beta)\|_2^2 \max_{\ell \leq n} \sigma_\ell^2 = O(n),$$

uniformly in fixed β , since $\|U^S\|_2 \leq \|U^A\|_2 = O(1)$ and Assumption 1(iv) bounds $\max_\ell \sigma_\ell^2$. For the quadratic term, the diagonal-zero property, conditional independence, and bounded fourth moments imply the standard quadratic-form bound

$$\text{Var}\left(\varepsilon'U^A(\beta)\varepsilon \mid X, A\right) \leq C\|U^A(\beta)\|_F^2 = O(n).$$

Thus, for each fixed $\beta \in \mathcal{B}$,

$$n^{-1}|\hat{m}_n^{\text{CF}}(\beta) - m_n^{\text{CF}}(\beta)| \xrightarrow{p} 0.$$

We now upgrade pointwise convergence to uniform convergence. First,

$$\|y\|_2^2 \leq 2\|\mu\|_2^2 + 2\|\varepsilon\|_2^2 = O_p(n),$$

because $\mathbb{E}[\|\varepsilon\|_2^2 \mid X, A] = \sum_\ell \sigma_\ell^2 = O(n)$. By the mean-value theorem and Lemma 2,

$$\begin{aligned} n^{-1}|\hat{m}_n^{\text{CF}}(\beta) - \hat{m}_n^{\text{CF}}(\beta')| &\leq n^{-1}\|y\|_2^2 \sup_{\tilde{\beta} \in \mathcal{B}} \|\nabla_{\tilde{\beta}} U^A(\tilde{\beta})\|_2 |\beta - \beta'| \\ &\leq B_n |\beta - \beta'|, \end{aligned}$$

where $B_n = O_p(1)$. Similarly,

$$n^{-1}|m_n^{\text{CF}}(\beta) - m_n^{\text{CF}}(\beta')| \leq n^{-1}\|\mu\|_2^2 \sup_{\tilde{\beta} \in \mathcal{B}} \|\nabla_{\tilde{\beta}} U^A(\tilde{\beta})\|_2 |\beta - \beta'| \leq C|\beta - \beta'|.$$

Let $\delta_\eta > 0$ be a mesh size to be chosen and let $\{\beta_j\}_{j=1}^{J_\eta}$ be a finite δ_η -net of the

compact set $\mathcal{B}$. For any $\beta \in \mathcal{B}$, choose β_j with $|\beta - \beta_j| \leq \delta_\eta$. Then

$$\begin{aligned} n^{-1} |\hat{m}_n^{\text{CF}}(\beta) - m_n^{\text{CF}}(\beta)| \\ \leq \max_{1 \leq j \leq J_\eta} n^{-1} |\hat{m}_n^{\text{CF}}(\beta_j) - m_n^{\text{CF}}(\beta_j)| + (B_n + C)\delta_\eta. \end{aligned}$$

The maximum over the finite net is $o_p(1)$ by pointwise convergence. Since $B_n = O_p(1)$, choosing δ_η arbitrarily small gives the desired uniform convergence. $\square$

The theorem now follows immediately from Assumption 3(iv), Lemma 3, and Lemma 1. $\square$

C.2 Proof of Theorem 2

We prove the two claims in Theorem 2 in three steps. First, we establish the central limit theorem for the CF moment. Let $\varepsilon = y - \mu$. At β_0 , Proposition 1 gives $m_n^{\text{CF}}(\beta_0) = 0$, and the diagonal-zero property of U^{A} gives

$$\begin{aligned} \hat{m}_n^{\text{CF}}(\beta_0) &= 2\mu' U^{\text{S}}(\beta_0) \varepsilon + \varepsilon' U^{\text{S}}(\beta_0) \varepsilon \\ &= \sum_{\ell=1}^n c_{\ell n} \varepsilon_\ell + \sum_{\ell < k} 2U_{\ell k}^{\text{S}}(\beta_0) \varepsilon_\ell \varepsilon_k. \end{aligned}$$

This is a generalized quadratic form in the independent errors $\{\varepsilon_\ell\}$, consisting of a linear projection term and a degenerate quadratic term. Assumption 4(i) gives the required uniform moment bound. It remains to verify that no single observation dominates. We first claim

$$\max_{\ell \leq n} \sum_{k=1}^n |U_{\ell k}^{\text{S}}(\beta_0)| = O(1).$$

To see this, let $r_\ell(\beta) = x_\ell + a_\ell \beta$. Bounded row ℓ_1 norms imply $\sup_{\ell, \beta \in \mathcal{B}} \|r_\ell(\beta)\|_2 = O(1)$. Since $\|S(\beta)^{-1}\|_2 = O(n^{-1})$,

$$\sum_{k=1}^n |P_{\ell k}(\beta)| = \sum_{k=1}^n |r_\ell(\beta)' S(\beta)^{-1} r_k(\beta)| \leq \|S(\beta)^{-1} r_\ell(\beta)\|_2 \sum_{k=1}^n \|r_k(\beta)\|_2 = O(1),$$

uniformly in ℓ and β . Hence the row and column ℓ_1 norms of $M(\beta) = I - P(\beta)$ are uniformly bounded. It follows that

$$\max_{\ell \leq n} \|e'_\ell M(\beta)A\|_1 \leq \max_{\ell \leq n} \sum_{j=1}^n |M_{\ell j}(\beta)| \|a_j\|_1 = O(1),$$

where we use $\max_j \|a_j\|_1 = O(1)$. Using this bound and $\|S(\beta)^{-1}\|_2 = O(n^{-1})$ gives uniformly bounded row and column ℓ_1 norms for

$$D(\beta) = M(\beta)AS(\beta)^{-1}R(\beta)'.$$

Moreover, $\|\Lambda(\beta)\|_\infty = O(1)$ by Lemma 2, so $M(\beta)\Lambda(\beta)$ also has uniformly bounded row and column ℓ_1 norms. Since $U^A = -(2D + M\Lambda)$ and $U^S = (U^A + U^{A'})/2$, the displayed claim follows. The same calculations, using the derivative bounds in Lemma 2, give uniformly bounded row and column ℓ_1 norms for $U^A(\beta)$, $U^S(\beta)$, $\nabla_\beta U^A(\beta)$, and $\nabla_\beta U^S(\beta)$ on $\mathcal{N}_0$. Let

$$W_{\ell km}(\beta) = 2U_{\ell k}^S(\beta)U_{\ell m}^A(\beta),$$

$$W_{\ell km}^{(0)}(\beta) = W_{\ell km}(\beta), \quad W_{\ell km}^{(1)}(\beta) = \nabla_\beta W_{\ell km}(\beta),$$

$$U_{\ell k}^{(0)}(\beta) = U_{\ell k}^A(\beta), \quad U_{\ell k}^{(1)}(\beta) = \nabla_\beta U_{\ell k}^A(\beta),$$

and

$$h_{\ell s}^{(-\ell km)}(\beta) = \mathbb{I}\{s \notin I_{\ell km}\} r_\ell(\beta)' S_{-I_{\ell km}}(\beta)^{-1} r_s(\beta),$$

$$I_{\ell km} = \{\ell, k, m\}, \quad S_{-I}(\beta) = \sum_{j \notin I} r_j(\beta) r_j(\beta)'.$$

$$h_{\ell s}^{(-\ell km, 0)}(\beta) = h_{\ell s}^{(-\ell km)}(\beta), \quad h_{\ell s}^{(-\ell km, 1)}(\beta) = \nabla_\beta h_{\ell s}^{(-\ell km)}(\beta).$$

The row and column bounds for U^A and U^S imply, for $q = 0, 1$,

$$\sup_{\beta \in \mathcal{N}_0} \max_{\ell} \sum_{k, m} |W_{\ell km}^{(q)}(\beta)| + \sup_{\beta \in \mathcal{N}_0} \max_k \sum_{\ell, m} |W_{\ell km}^{(q)}(\beta)|$$

$$+ \sup_{\beta \in \mathcal{N}_0} \max_m \sum_{\ell, k} |W_{\ell km}^{(q)}(\beta)| = O(1).$$

We next verify the leave-out weight bounds. Let $R_I(\beta)$ be the matrix whose rows are $\{r_i(\beta)' : i \in I\}$. For a leave-out set $I = I_{\ell km}$, the Sherman–Morrison–Woodbury formula gives

$$S_{-I}(\beta)^{-1} = S(\beta)^{-1} + S(\beta)^{-1} R_I(\beta)' M_{II}(\beta)^{-1} R_I(\beta) S(\beta)^{-1},$$

where $M_{II}(\beta) = I_{|I|} - R_I(\beta) S(\beta)^{-1} R_I(\beta)'$ is the principal submatrix of $M(\beta)$ indexed by I . By the convention $D_{\ell k k}(\beta) = D_{\ell k}(\beta)$, $\det M_{I_{\ell km} I_{\ell km}}(\beta) = D_{\ell km}(\beta)$. Since $0 \preceq M_{II}(\beta) \preceq I_{|I|}$, Assumption 4(ii) implies $\sup_{\beta \in \mathcal{N}_0, \ell, k, m} \|M_{I_{\ell km} I_{\ell km}}(\beta)^{-1}\|_2 = O(1)$. Combining this with Assumption 3(ii) and bounded row ℓ_1 norms of x_ℓ and a_ℓ yields

$$\sup_{\beta \in \mathcal{N}_0, \ell, k, m} \|S_{-I_{\ell km}}(\beta)^{-1}\|_2 = O(n^{-1}).$$

Moreover, $\sup_{\beta \in \mathcal{N}_0, \ell, k, m} \|\nabla_\beta S_{-I_{\ell km}}(\beta)\|_2 = O(n)$, so

$$\nabla_\beta S_{-I}(\beta)^{-1} = -S_{-I}(\beta)^{-1} \{\nabla_\beta S_{-I}(\beta)\} S_{-I}(\beta)^{-1}$$

has operator norm $O(n^{-1})$ uniformly in β and I . Therefore

$$\sup_{\beta \in \mathcal{N}_0} \max_{\ell, k, m, s} |h_{\ell s}^{(-\ell km)}(\beta)| = O(n^{-1}), \quad \sup_{\beta \in \mathcal{N}_0} \max_{\ell, k, m} \sum_s |h_{\ell s}^{(-\ell km)}(\beta)| = O(1),$$

and the same bounds with $h_{\ell s}^{(-\ell km, 1)}(\beta)$ in place of $h_{\ell s}^{(-\ell km, 0)}(\beta)$.

These primitive bounds imply the overlap control needed for the leave-three-out variance estimator. For each observation i , define

$$\begin{aligned} \mathcal{K}_{q,i}(\beta) &= \sum_{\ell, k, m: i \in \{\ell, k, m\}} |W_{\ell km}^{(q)}(\beta)| + \sum_{\ell, k, m, s: i \in \{\ell, k, m, s\}} |W_{\ell km}^{(q)}(\beta) h_{\ell s}^{(-\ell km, 0)}(\beta)| \\ &+ \sum_{\ell, k, m, s: i \in \{\ell, k, m, s\}} |W_{\ell km}^{(0)}(\beta) h_{\ell s}^{(-\ell km, q)}(\beta)| + \sum_{j=1}^n \left\{ |U_{ij}^{(q)}(\beta)| + |U_{ji}^{(q)}(\beta)| \right\}. \end{aligned}$$

The first and last terms are $O(1)$ uniformly in i , q , and $\beta \in \mathcal{N}_0$ by the slice bounds

for W and the row and column bounds for U^A . For $q, r \in \{0, 1\}$,

$$\begin{aligned} & \sup_{\beta \in \mathcal{N}_0} \max_i \sum_{\ell, k, m, s: i \in \{\ell, k, m, s\}} |W_{\ell km}^{(q)}(\beta) h_{\ell s}^{(-\ell km, r)}(\beta)| \\ & \leq C + \sup_{\beta \in \mathcal{N}_0} \left(\sum_{\ell, k, m} |W_{\ell km}^{(q)}(\beta)| \right) \max_{\ell, k, m, s} |h_{\ell s}^{(-\ell km, r)}(\beta)| = O(1), \end{aligned}$$

because the total $W^{(q)}$ mass is $O(n)$ and $\max_{\ell, k, m, s} |h_{\ell s}^{(-\ell km, r)}(\beta)| = O(n^{-1})$. Therefore

$$\max_i \mathcal{K}_{0,i}(\beta_0) = O(1), \quad \sup_{\beta \in \mathcal{N}_0} \max_i \mathcal{K}_{1,i}(\beta) = O(1).$$

The claim and $\max_{\ell} |\mu_{\ell}| = O(1)$ imply

$$\max_{\ell \leq n} c_{\ell n}^2 \sigma_{\ell}^2 = O(1),$$

while Assumption 4(i) implies $\max_{\ell} \sigma_{\ell}^2 = O(1)$ and therefore

$$\max_{\ell \leq n} \sum_{k \neq \ell} \omega_{\min\{\ell, k\}, \max\{\ell, k\}, n}^2 = O(1).$$

Since $\liminf_n n^{-1} V_n(\beta_0) > 0$, both maxima are $o(V_n(\beta_0))$. Thus the Lindeberg/no-dominant-component condition for the linear-plus-degenerate-quadratic form holds. The central-limit theorem for generalized quadratic forms in independent variables (de Jong, 1987) therefore implies

$$\frac{\hat{m}_n^{\text{CF}}(\beta_0)}{\sqrt{\text{Var}(\hat{m}_n^{\text{CF}}(\beta_0) \mid X, A)}} \xrightarrow{d} N(0, 1).$$

The denominator is $V_n(\beta_0)$ by definition. Hence

$$\frac{\hat{m}_n^{\text{CF}}(\beta_0)}{\sqrt{V_n(\beta_0)}} \xrightarrow{d} N(0, 1). \tag{19}$$

Second, we prove consistency of the leave-three-out variance estimator. Let $r_{\ell} =$

$x_\ell + a_\ell \beta_0$ and $I_{\ell km} = \{\ell, k, m\}$. At the true value,

$$\hat{\delta}_{(\ell km)}(\beta_0) = \delta + S_{-I_{\ell km}}(\beta_0)^{-1} \sum_{s \notin I_{\ell km}} r_s \varepsilon_s,$$

and therefore

$$\hat{\sigma}_{\ell, -km}^2(\beta_0) = y_\ell \left(\varepsilon_\ell - \sum_{s=1}^n h_{\ell s}^{(-\ell km)}(\beta_0) \varepsilon_s \right).$$

Substituting this expression into (11), and using $\hat{m}_n^{\text{CF}}(\beta_0) = y' U^A(\beta_0) y$, shows that $\hat{V}_n(\beta_0) - V_n(\beta_0)$ is a centered fourth-degree polynomial in the independent errors. Its deterministic weights are finite sums of products formed from

$$W_{\ell km}(\beta_0), \quad W_{\ell km}(\beta_0) h_{\ell s}^{(-\ell km)}(\beta_0), \quad U_{ij}^A(\beta_0).$$

If two monomials in this polynomial involve disjoint sets of error indices, their covariance is zero. If they overlap, Assumption 4(i) bounds the covariance by a constant times the product of the corresponding absolute weights. Summing these possible overlaps gives

$$\text{Var}(\hat{V}_n(\beta_0) \mid X, A) \leq C \sum_{i=1}^n \mathcal{K}_{0,i}(\beta_0)^2.$$

The displayed overlap bound therefore gives

$$\text{Var}(\hat{V}_n(\beta_0) \mid X, A) = o(n^2).$$

Since $\liminf_n n^{-1} V_n(\beta_0) > 0$, this implies

$$\text{Var}(\hat{V}_n(\beta_0) \mid X, A) = o(V_n(\beta_0)^2).$$

By Proposition 2,

$$\mathbb{E}[\hat{V}_n(\beta_0) \mid X, A] = V_n(\beta_0).$$

Therefore, for any $\varepsilon > 0$, Chebyshev's inequality gives

$$\mathbb{P} \left(\left| \frac{\hat{V}_n(\beta_0)}{V_n(\beta_0)} - 1 \right| > \varepsilon \mid X, A \right) \leq \frac{\text{Var}(\hat{V}_n(\beta_0) \mid X, A)}{\varepsilon^2 V_n(\beta_0)^2} = o(1),$$

Hence

$$\hat{V}_n(\beta_0)/V_n(\beta_0) \xrightarrow{p} 1,$$

By Theorem 1, $\hat{\beta}^{\text{CF}} \xrightarrow{p} \beta_0$. Hence, with probability approaching one, $\hat{\beta}^{\text{CF}} \in \mathcal{N}_0$. The derivative of the centered variance-estimator process is again a finite polynomial in the errors. Its non-random terms are $O(n)$ by the uniform slice bounds for W , h , U^A , and their first derivatives established above. For the centered terms, the same overlap argument, now using $\sup_{\beta \in \mathcal{N}_0} \mathcal{K}_{1,i}(\beta) = O(1)$, gives

$$\mathbb{E} \left[\sup_{\beta \in \mathcal{N}_0} \left\{ n^{-1} |\nabla_{\beta} (\hat{T}_n(\beta) - \{\hat{m}_n^{\text{CF}}(\beta)\}^2) | \right\}^2 \mid X, A \right] = O(1).$$

Therefore

$$n^{-1} |\hat{V}_n(\hat{\beta}^{\text{CF}}) - \hat{V}_n(\beta_0)| \leq \sup_{\beta \in \mathcal{N}_0} n^{-1} |\nabla_{\beta} (\hat{T}_n(\beta) - \{\hat{m}_n^{\text{CF}}(\beta)\}^2)| \|\hat{\beta}^{\text{CF}} - \beta_0\| = o_p(1).$$

Since $\liminf_n n^{-1} V_n(\beta_0) > 0$, this implies

$$\hat{V}_n(\hat{\beta}^{\text{CF}})/V_n(\beta_0) \xrightarrow{p} 1.$$

This proves part (i).

Third, we prove the studentized normality. By Theorem 1, $\hat{\beta}^{\text{CF}} \xrightarrow{p} \beta_0$. Since β_0 is in the interior of $\mathcal{B}$, with probability approaching one the line segment between $\hat{\beta}^{\text{CF}}$ and β_0 lies in $\mathcal{N}_0$. A mean-value expansion of the sample moment around β_0 gives

$$0 = \hat{m}_n^{\text{CF}}(\beta_0) + \nabla_{\beta} \hat{m}_n^{\text{CF}}(\tilde{\beta})(\hat{\beta}^{\text{CF}} - \beta_0)$$

for some $\tilde{\beta}$ between $\hat{\beta}^{\text{CF}}$ and β_0 . Rearranging,

$$\hat{\beta}^{\text{CF}} - \beta_0 = -\frac{\hat{m}_n^{\text{CF}}(\beta_0)}{\nabla_{\beta} \hat{m}_n^{\text{CF}}(\tilde{\beta})}. \quad (20)$$

We first analyze the derivative terms. Differentiating the quadratic-form representation gives

$$\nabla_{\beta} \hat{m}_n^{\text{CF}}(\beta) = y' \nabla_{\beta} U^A(\beta) y.$$

Because the diagonal of $U^A(\beta)$ is identically zero in β , the diagonal of $\nabla_{\beta} U^A(\beta)$ is

also zero. Hence

$$\nabla_{\beta} m_n^{\text{CF}}(\beta) = \mathbb{E}[y' \nabla_{\beta} U^{\text{A}}(\beta) y \mid X, A] = \mu' \nabla_{\beta} U^{\text{A}}(\beta) \mu.$$

The proof of Lemma 3, applied with $\nabla_{\beta} U^{\text{A}}$ in place of U^{A} and using the bounds

$$\sup_{\beta \in \mathcal{B}} \|\nabla_{\beta} U^{\text{A}}(\beta)\|_2 = O(1), \quad \sup_{\beta \in \mathcal{B}} \|\nabla_{\beta} U^{\text{A}}(\beta)\|_F = O(\sqrt{n}), \quad \sup_{\beta \in \mathcal{B}} \|\nabla_{\beta\beta} U^{\text{A}}(\beta)\|_2 = O(1),$$

from Lemma 2, gives

$$\sup_{\beta \in \mathcal{B}} n^{-1} |\nabla_{\beta} \hat{m}_n^{\text{CF}}(\beta) - \nabla_{\beta} m_n^{\text{CF}}(\beta)| \xrightarrow{p} 0. \quad (21)$$

Moreover, for any two sequences $\beta_n, \beta'_n \in \mathcal{N}_0$ satisfying $|\beta_n - \beta'_n| = o_p(1)$,

$$n^{-1} |\nabla_{\beta} m_n^{\text{CF}}(\beta_n) - \nabla_{\beta} m_n^{\text{CF}}(\beta'_n)| \leq n^{-1} \|\mu\|_2^2 \sup_{\beta \in \mathcal{B}} \|\nabla_{\beta\beta} U^{\text{A}}(\beta)\|_2 |\beta_n - \beta'_n| = o_p(1),$$

because $\|\mu\|_2^2 = O(n)$. Combining this display with (21) yields

$$\frac{\nabla_{\beta} \hat{m}_n^{\text{CF}}(\hat{\beta}^{\text{CF}})}{\nabla_{\beta} \hat{m}_n^{\text{CF}}(\tilde{\beta})} \xrightarrow{p} 1.$$

To see this, the numerator of the difference between this ratio and one is $o_p(n)$ by the preceding two displays, while the local slope condition in the theorem and (21) imply that

$$\inf_{\beta \in \mathcal{N}_0} n^{-1} |\nabla_{\beta} \hat{m}_n^{\text{CF}}(\beta)|$$

is bounded away from zero with probability approaching one.

Using (20), we can write the studentized statistic as

$$\begin{aligned} \frac{\nabla_{\beta} \hat{m}_n^{\text{CF}}(\hat{\beta}^{\text{CF}})}{\sqrt{\hat{V}_n(\hat{\beta}^{\text{CF}})}} (\hat{\beta}^{\text{CF}} - \beta_0) &= -\frac{\hat{m}_n^{\text{CF}}(\beta_0)}{\sqrt{\hat{V}_n(\beta_0)}} \sqrt{\frac{V_n(\beta_0)}{\hat{V}_n(\hat{\beta}^{\text{CF}})}} \frac{\nabla_{\beta} \hat{m}_n^{\text{CF}}(\hat{\beta}^{\text{CF}})}{\nabla_{\beta} \hat{m}_n^{\text{CF}}(\tilde{\beta})} \\ &\xrightarrow{d} N(0, 1), \end{aligned}$$

where we use the central limit theorem in (19), the plug-in variance consistency in part (i), and the derivative ratio above. The minus sign does not affect the limit

because the standard normal distribution is symmetric. $\square$

C.3 Proof of Proposition 1

Observe that

$$\mathbb{E}[\hat{m}_n(\beta_0)|X, A] = m_n(\beta_0) = \sum_{l=1}^n \nabla_{\beta} M_l(\beta_0) \sigma_l^2.$$

Also, if $\min_{l \leq n} M_l(\beta_0) > 0$, each $\hat{\sigma}_l^2(\beta_0)$ is well-defined and unbiased for σ_l^2 :

$$\begin{aligned} \mathbb{E}[\hat{\sigma}_l^2(\beta_0)|X, A] &= \mathbb{E}[(\mu_l + \varepsilon_l)(\mu_l + \varepsilon_l - (x_l + a_l \beta_0)' \hat{\delta}_{(l)}^{\text{LS}}(\beta_0))|X, A] \\ &= \mathbb{E}[(\mu_l + \varepsilon_l)(\varepsilon_l + (x_l + a_l \beta_0)'(\delta - \hat{\delta}_{(l)}^{\text{LS}}(\beta_0)))|X, A] \\ &= \sigma_l^2 \end{aligned}$$

where the last equality uses the independence of ε_l from $\hat{\delta}_{(l)}^{\text{LS}}(\beta_0)$ and the fact that $\hat{\delta}_{(l)}^{\text{LS}}(\beta_0)$ is an unbiased estimator of δ under the leave-one-out procedure:

$$\mathbb{E}[\hat{\delta}_{(l)}^{\text{LS}}(\beta_0)|X, A] = \left(\sum_{j \neq l} (x_j + a_j \beta_0)(x_j + a_j \beta_0)' \right)^{-1} \sum_{j \neq l} (x_j + a_j \beta_0)(x_j + a_j \beta_0)' \delta = \delta.$$

Therefore,

$$\mathbb{E}[\hat{m}_n^{\text{CF}}(\beta_0)] = \mathbb{E}[\hat{m}_n(\beta_0)|X, A] - \sum_{l=1}^n \nabla_{\beta} M_l(\beta_0) \mathbb{E}[\hat{\sigma}_l^2(\beta_0)|X, A] = 0.$$

This completes the proof of Proposition 1. $\square$

C.4 Proof of Proposition 2

Observe that

$$\begin{aligned} \mathbb{E}[\hat{\sigma}_{l,-km}^2(\beta_0)|X, A] &= \mathbb{E}[(\mu_l + \varepsilon_l)(\mu_l + \varepsilon_l - (x_l + a_l \beta_0)' \hat{\delta}_{(lkm)}(\beta_0))|X, A] \\ &= \mathbb{E}[(\mu_l + \varepsilon_l)(\varepsilon_l + (x_l + a_l \beta_0)'(\delta - \hat{\delta}_{(lkm)}(\beta_0)))|X, A] \\ &= \sigma_l^2, \end{aligned}$$

where the last equality uses the independence of ε_l from $\hat{\delta}_{(lkm)}(\beta_0)$ and the fact that $\hat{\delta}_{(lkm)}(\beta_0)$ is an unbiased estimator of δ under the leave-three-out procedure:

$$\begin{aligned} & \mathbb{E}[\hat{\delta}_{(lkm)}(\beta_0)|X, A] \\ &= \left(\sum_{j \notin \{l, k, m\}} (x_j + a_j \beta_0)(x_j + a_j \beta_0)' \right)^{-1} \sum_{j \notin \{l, k, m\}} (x_j + a_j \beta_0)(x_j + a_j \beta_0)' \delta \\ &= \delta. \end{aligned}$$

Therefore,

$$\begin{aligned} & \mathbb{E}[\hat{V}_n(\beta_0)|X, A] \\ &= 2\mathbb{E} \left[\sum_{l=1}^n \sum_{k \neq l} \sum_{m \neq l} U_{lk}^S(\beta_0) U_{lm}^A(\beta_0) y_k y_m \hat{\sigma}_{l, -km}^2(\beta_0) | X, A \right] - \mathbb{E}[(\hat{m}_n^{\text{CF}}(\beta_0))^2 | X, A] \\ &= 2\mathbb{E} \left[\sum_{l=1}^n \sum_{k \neq l} \sum_{m \neq l} U_{lk}^S(\beta_0) U_{lm}^A(\beta_0) y_k y_m \sigma_l^2 | X, A \right] - \mathbb{E}[(\hat{m}_n^{\text{CF}}(\beta_0))^2 | X, A] \\ &= V_n(\beta_0) \end{aligned}$$

where the second equality uses the independence of $\hat{\sigma}_{l, -km}^2(\beta_0)$ from y_k and y_m for $k, m \neq l$, and the last equality uses the expressions of $V_n(\beta_0)$ in (9). This completes the proof of Proposition 2. $\square$

C.5 Proof of Proposition 3

Lemma 4 (Concentration of the conservative variance estimator). *Suppose Assumptions 1, 2, and 3 hold, Assumption 4(i) holds, Assumption 5 replaces Assumption 4(ii), and $\liminf_{n \rightarrow \infty} n^{-1} V_n(\beta_0) > 0$. Then for every $\eta > 0$,*

$$\mathbb{P} \left(\sup_{\beta \in \mathcal{N}_0} \frac{|\hat{V}_n^{\text{mod}}(\beta) - \bar{V}_n^{\text{mod}}(\beta)|}{V_n(\beta_0)} > \eta \mid X, A \right) \rightarrow 0.$$

Proof. The proof follows the overlap argument used for $\hat{V}_n$ in the proof of Theorem 2. Assumption 5, together with Assumption 3(i), gives the same row and column bounds for the leave-two-out and leave-three-out projection weights that are actually used in (22). The replacement y_ℓ^2 introduces no leave-out projection weights, and the

trimming indicators $G_{\ell,-km}$ are bounded by one. Hence the deterministic absolute-weight sums entering the overlap calculation are bounded by the same $\mathcal{K}_{q,i}$ envelopes constructed above, up to constants.

Using bounded eighth moments, covariance terms whose error indices overlap are bounded by a constant times the product of their absolute weights. Terms with disjoint relevant indices have zero covariance after conditioning on the variables entering the corresponding trimming indicators; the remaining covariance terms are covered by the same overlap envelope because each trimming indicator is attached to a single ℓ -slice. Therefore

$$\text{Var}\left(\hat{V}_n^{\text{mod}}(\beta) - \bar{V}_n^{\text{mod}}(\beta) \mid X, A\right) \leq C \sum_{i=1}^n \mathcal{K}_{0,i}(\beta)^2 = o(n^2)$$

uniformly on $\mathcal{N}_0$. Since $\liminf_n n^{-1}V_n(\beta_0) > 0$, Chebyshev's inequality gives pointwise convergence after normalization by $V_n(\beta_0)$. The derivative version of the same overlap bound gives stochastic equicontinuity on $\mathcal{N}_0$, so the pointwise convergence upgrades to the displayed uniform convergence. $\square$

Lemma 5 (Conservative centering). *For fixed β , write $\bar{V}_n^{\text{mod}}(\beta) = \mathbb{E}[\hat{V}_n^{\text{mod}}(\beta) \mid X, A]$ and $\Delta_n^{\text{mod}}(\beta) = \bar{V}_n^{\text{mod}}(\beta) - V_n(\beta)$. Under the conditions of Lemma 4,*

$$\Delta_n^{\text{mod}}(\beta_0) \geq 0.$$

Moreover, for every $\eta > 0$,

$$\mathbb{P}\left(\frac{\bar{V}_n^{\text{mod}}(\hat{\beta}^{\text{CF}})}{V_n(\beta_0)} < 1 - \eta \mid X, A\right) \rightarrow 0.$$

Proof. At β_0 , the first two lines of (22) give unbiased replacements for the corresponding terms in (9): the original leave-three-out estimator is unbiased when $D_{\ell km}(\beta_0) > 0$, and the leave-two-out replacement is unbiased when the leave-three-out failure is not caused by ℓ .

It remains to consider the terms for which $\bar{\sigma}_{\ell,-km}^2(\beta_0) = y_\ell^2$. For such terms, y_ℓ^2 is independent of the aggregate weight

$$B_\ell(\beta_0) = \sum_{k \neq \ell} \sum_{m \neq \ell} W_{\ell km}(\beta_0) \mathbb{I}\{\bar{\sigma}_{\ell,-km}^2(\beta_0) = y_\ell^2\},$$

conditional on X, A and on $\{y_j : j \neq \ell\}$. If $B_\ell(\beta_0) \geq 0$, then $G_{\ell, -km}(\beta_0) = 1$ for the biased terms attached to ℓ , and their conditional excess relative to using σ_ℓ^2 is

$$B_\ell(\beta_0)\mathbb{E}[y_\ell^2 - \sigma_\ell^2 \mid X, A, \{y_j : j \neq \ell\}] = B_\ell(\beta_0)\mu_\ell^2 \geq 0.$$

If $B_\ell(\beta_0) < 0$, then $G_{\ell, -km}(\beta_0) = 0$ for these biased terms. Omitting them removes the target contribution $B_\ell(\beta_0)\sigma_\ell^2$, which is nonpositive, and therefore again weakly increases the expectation of the variance estimator. Summing over ℓ gives $\Delta_n^{\text{mod}}(\beta_0) \geq 0$.

For the plug-in statement, Theorem 1 gives $\hat{\beta}^{\text{CF}} \xrightarrow{p} \beta_0$. The determinant condition in Assumption 5 makes the leave-out replacement rule locally stable except at bounded trimming boundaries, and the same derivative bounds used in the proof of Theorem 2 give

$$\frac{\bar{V}_n^{\text{mod}}(\hat{\beta}^{\text{CF}}) - \bar{V}_n^{\text{mod}}(\beta_0)}{V_n(\beta_0)} \xrightarrow{p} 0.$$

Since $\bar{V}_n^{\text{mod}}(\beta_0) \geq V_n(\beta_0)$, the displayed one-sided claim follows. $\square$

Because $\bar{V}_n^{\text{mod}}(\beta_0) \geq V_n(\beta_0)$ by Lemma 5,

$$\left(1 - \frac{\hat{V}_n^{\text{mod}}(\beta_0)}{V_n(\beta_0)}\right)_+ \leq \frac{|\hat{V}_n^{\text{mod}}(\beta_0) - \bar{V}_n^{\text{mod}}(\beta_0)|}{V_n(\beta_0)}.$$

The first claim therefore follows from Lemma 4.

For the plug-in statement, use the deterministic inequality

$$\begin{aligned} \left(1 - \frac{\hat{V}_n^{\text{mod}}(\hat{\beta}^{\text{CF}})}{V_n(\beta_0)}\right)_+ &\leq \left(1 - \frac{\bar{V}_n^{\text{mod}}(\hat{\beta}^{\text{CF}})}{V_n(\beta_0)}\right)_+ \\ &\quad + \frac{|\hat{V}_n^{\text{mod}}(\hat{\beta}^{\text{CF}}) - \bar{V}_n^{\text{mod}}(\hat{\beta}^{\text{CF}})|}{V_n(\beta_0)}. \end{aligned}$$

Lemma 5 makes the first term on the right negligible in probability. Theorem 1 implies $\hat{\beta}^{\text{CF}} \in \mathcal{N}_0$ with probability approaching one, so Lemma 4 makes the second term negligible in probability. $\square$

Appendix D Conservative variance estimation under leave-out rank deficiency

The full-rank requirement in Assumption 4(ii) can fail in fixed-effect designs with small cells, because deleting three observations may remove the last identifying observation for one or more nuisance parameters. Following the logic of Section 4 of [Anatolyev and Solvsten \(2023\)](#), the variance estimator can be modified so that the unbiased leave-three-out estimator is used whenever it exists, while failures are handled by replacements that are either still unbiased for the relevant product or upward biased.

To state the modification, define the leave-two and leave-three determinants for distinct observations:

$$D_{\ell k}(\beta) = \begin{vmatrix} M_{\ell\ell}(\beta) & M_{\ell k}(\beta) \\ M_{\ell k}(\beta) & M_{kk}(\beta) \end{vmatrix},$$

$$D_{\ell km}(\beta) = \begin{vmatrix} M_{\ell\ell}(\beta) & M_{\ell k}(\beta) & M_{\ell m}(\beta) \\ M_{\ell k}(\beta) & M_{kk}(\beta) & M_{km}(\beta) \\ M_{\ell m}(\beta) & M_{km}(\beta) & M_{mm}(\beta) \end{vmatrix}.$$

By the Sherman-Morrison-Woodbury formula, $D_{\ell k}(\beta) > 0$ and $D_{\ell km}(\beta) > 0$ characterize the existence of the corresponding leave-two-out and leave-three-out least squares estimators, given full rank of the full-sample design. We use the conventions $D_{\ell kk}(\beta) = D_{\ell k}(\beta)$ and $D_{kk}(\beta) = 0$.

Define the replacement for the leave-three-out variance estimator by

$$\bar{\sigma}_{\ell, -km}^2(\beta) = \begin{cases} \hat{\sigma}_{\ell, -km}^2(\beta), & \text{if } D_{\ell km}(\beta) > 0, \\ \hat{\sigma}_{\ell, -k}^2(\beta), & \text{if } D_{km}(\beta) = 0 \text{ and } D_{\ell k}(\beta)D_{\ell m}(\beta) > 0, \\ y_{\ell}^2, & \text{otherwise.} \end{cases} \quad (22)$$

The first line is the original leave-three-out estimator. The second line applies when the leave-three-out failure is not caused by observation ℓ : in that case, the leave-two-out estimator for ℓ is independent of y_k and y_m , so the term $y_k y_m \hat{\sigma}_{\ell, -k}^2(\beta_0)$ remains unbiased for the component of (9). The final line applies when the failure is caused by ℓ ; it replaces σ_{ℓ}^2 by y_{ℓ}^2 , whose conditional expectation is $\sigma_{\ell}^2 + \mu_{\ell}^2$ at the truth, where

$$\mu_\ell = (x_\ell + a_\ell \beta_0)' \delta.$$

Because the last line of (22) is upward biased only before it is multiplied by the possibly signed weight in (11), biased replacements with negative aggregate weight should be removed to ensure the conservativeness of the resulting variance estimator. To this end, define the term-specific weight

$$W_{\ell km}(\beta) = 2U_{\ell k}^S(\beta)U_{\ell m}^A(\beta)y_k y_m$$

and the aggregate weight on biased replacements for observation ℓ as

$$B_\ell(\beta) = \sum_{k \neq \ell} \sum_{m \neq \ell} W_{\ell km}(\beta) \mathbb{I}\{\bar{\sigma}_{\ell, -km}^2(\beta) = y_\ell^2\},$$

and set

$$G_{\ell, -km}(\beta) = \begin{cases} 0, & \text{if } \bar{\sigma}_{\ell, -km}^2(\beta) = y_\ell^2 \text{ and } B_\ell(\beta) < 0, \\ 1, & \text{otherwise.} \end{cases}$$

The conservative variance estimator is then

$$\hat{V}_n^{\text{mod}}(\beta) = \sum_{\ell=1}^n \sum_{k \neq \ell} \sum_{m \neq \ell} W_{\ell km}(\beta) G_{\ell, -km}(\beta) \bar{\sigma}_{\ell, -km}^2(\beta) - \left(\hat{m}_n^{\text{CF}}(\beta) \right)^2. \quad (23)$$

When all leave-three-out estimators exist, $\bar{\sigma}_{\ell, -km}^2 = \hat{\sigma}_{\ell, -km}^2$ and $G_{\ell, -km} = 1$, so (23) reduces to the original estimator (11). When some leave-three-out estimators fail, the construction keeps all unbiased replacements and retains upward-biased replacements only when their aggregate contribution is nonnegative.

To state a formal conservative consistency result, define $\bar{V}_n^{\text{mod}}(\beta) = \mathbb{E}[\hat{V}_n^{\text{mod}}(\beta) \mid X, A]$ and $\Delta_n^{\text{mod}}(\beta) = \bar{V}_n^{\text{mod}}(\beta) - V_n(\beta)$ for fixed β . The following condition replaces Assumption 4(ii).

Assumption 5 (Regularity with leave-out failures). *Conditional on (X, A) , the nonzero leave-out determinants are bounded away from zero: there exist constants $C_D < \infty$ and $N_D < \infty$ such that, for all $n \geq N_D$,*

$$\sup_{\beta \in \mathcal{N}_0} \left\{ \max_{\ell, k: D_{\ell k}(\beta) > 0} D_{\ell k}(\beta)^{-1} + \max_{\ell, k, m: D_{\ell km}(\beta) > 0} D_{\ell km}(\beta)^{-1} \right\} \leq C_D,$$

where maxima over empty sets are set equal to zero.

Assumption 5 is the analog of Assumption 4(ii): together with the bounded-leverage condition in Assumption 3(i), it permits $D_{\ell k}(\beta)$ and $D_{\ell km}(\beta)$ to be zero, but prevents the nonzero leave-out denominators that are actually used from becoming arbitrarily small.

The following proposition shows that the modified variance estimator is asymptotically conservative:

Proposition 3 (Conservative variance consistency when leave-three-out may fail). *Suppose Assumptions 1, 2, and 3 hold, Assumption 4(i) holds, and Assumption 5 replaces Assumption 4(ii). If $\liminf_{n \rightarrow \infty} n^{-1}V_n(\beta_0) > 0$, then, for every $\eta > 0$, the positive part of the shortfall from the target variance is asymptotically negligible:*

$$\begin{aligned} \left(1 - \frac{\hat{V}_n^{\text{mod}}(\beta_0)}{V_n(\beta_0)}\right)_+ &\rightarrow_p 0, \\ \left(1 - \frac{\hat{V}_n^{\text{mod}}(\hat{\beta}^{\text{CF}})}{V_n(\beta_0)}\right)_+ &\rightarrow_p 0 \end{aligned}$$

where $(t)_+ = \max\{t, 0\}$.

Proposition 3 shows that tests and confidence intervals based on

$$\frac{\nabla_{\beta} \hat{m}_n^{\text{CF}}(\hat{\beta}^{\text{CF}})}{\sqrt{\hat{V}_n^{\text{mod}}(\hat{\beta}^{\text{CF}})}}(\hat{\beta}^{\text{CF}} - \beta_0)$$

are asymptotically valid, albeit potentially conservative. This follows because by Theorem 2, this statistic is asymptotically normal when the denominator is replaced by $\sqrt{V_n(\beta_0)}$, and Proposition 3 implies that the denominator $\sqrt{\hat{V}_n^{\text{mod}}(\hat{\beta}^{\text{CF}})}$ is asymptotically no smaller than $\sqrt{V_n(\beta_0)}$ with probability approaching one.¹⁴

Appendix E Additional results for the classroom application

¹⁴The asymptotic normality of $\nabla_{\beta} \hat{m}_n^{\text{CF}}(\hat{\beta}^{\text{CF}})/\sqrt{V_n(\beta_0)}(\hat{\beta}^{\text{CF}} - \beta_0)$ does not depend on the existence of leave-three-out estimators, so it continues to hold under the conditions of Proposition 3.

This appendix reports two additional results for the classroom application of Section 5. Appendix E.1 compares the computational cost of our cross-fit estimator with the non-linear least squares estimator and the iterative procedure of Arcidiacono et al. (2012). Appendix E.2 then re-estimates the model allowing for an *endogenous* peer effect operating through peer outcomes, as a check on whether the contextual peer effect reported in the main text is confounded by this second channel.

E.1 Computational comparison

On the Spring 2019 classroom sample, we time three estimators of the contextual peer-effect model (12): our cross-fit estimator (root-finding on the cross-fit moment, with the analytic $\hat{V}_n$ for inference); our NLLS estimator (root-finding on the first-order condition, with an individual wild bootstrap for inference); and the iterative NLLS estimator of Arcidiacono et al. (2012), which alternates a closed-form fixed-effect update with an update of β until the relative change in β falls below a tolerance. All times are measured from the raw data file. Table 1 reports the results.

Appendix Table 1: Computational comparison on the Spring 2019 classroom sample

Method	$\hat{\beta}$	Point (s)	Inference (s)
Cross-fit (analytic V_n)	0.167	57.8	18.1
NLLS (wild bootstrap, $B = 100$)	0.249	37.0	1008.0
Arcidiacono iterative (tol = 10^{-3})	0.201	0.4	44.7
Arcidiacono iterative (tol = 10^{-5})	0.229	37.9	3786.5
Arcidiacono iterative (tol = 10^{-7})	0.248	229.1	22908.8

Notes: Wall-clock seconds from the raw data file; “Point” includes data preparation. Inference uses the analytic $\hat{V}_n$ for cross-fit and a wild bootstrap ($B = 100$) for the NLLS estimators; the iterative-NLLS bootstrap time is extrapolated as B times the point-estimate time.

Two points stand out. First, the iterative estimator’s point estimate is sensitive to the convergence tolerance: it rises from 0.201 at a loose tolerance, to 0.229 at an intermediate one, to 0.248 at a tight one, approaching the exact root (0.249) that our root-finding solver locates directly, so a practitioner who stops iterating too early obtains a visibly attenuated estimate. Second, inference is where the cross-fit estimator is decisively cheaper: the analytic $\hat{V}_n$ costs seconds, whereas the wild bootstrap re-estimates the model for every replication — about seventeen minutes for our NLLS estimator and, for the iterative estimator at a tolerance tight enough

to converge, several hours.

It is worth noting that these comparisons are specific to the classroom scale. The cross-fit point estimate requires forming and storing the leave-out objects (the matrices A and M), whose cost grows with the sample size; on much larger data — such as the matched employer–employee panel of Section 6 — this memory and time cost can dominate, and the small memory footprint of the iterative method becomes attractive. The natural combination there is an iterative or random-projection point estimate paired with the analytic $\hat{V}_n$, which preserves the advantage in inference cost documented above.

E.2 Empirical extension: endogenous peer effects

The empirical specification in the main text considers only contextual peer effects — influence operating through peer characteristics (here, peer ability $\bar{\alpha}_{(i)j}$). Since Manski (1993), a central concern in the peer-effects literature is that such contextual effects may be confounded with *endogenous* peer effects — influence operating through peer *outcomes* — because the two channels cannot in general be separated within a single interacting group (the reflection problem). It is therefore natural to ask whether the contextual peer effect estimated in the main text survives once the endogenous channel is allowed for. Bramoullé et al. (2009) establishes that the richer model with both channels is point-identified when peers interact through a network that is not perfectly transitive. The corresponding regression specification, building on (1), is

$$y_{ij} = \alpha_i + \bar{\alpha}_{(i)j} \cdot \beta + \bar{y}_{(i)j} \cdot \lambda + \psi_j + \varepsilon_{ij}, \quad (24)$$

where $\bar{y}_{(i)j}$ is the average grade in the peer group, excluding student i . Students do not observe $\bar{y}_{(i)j}$ when choosing effort (we observe it only ex post), so it enters as a realized peer outcome the group is collectively exposed to, rather than one each student best-responds to contemporaneously.

We estimate (24) using a natural extension of the cross-fit moment to the joint parameter (β, λ) , and report the social-multiplier summary $(\hat{\beta} + \hat{\lambda})/(1 - \hat{\lambda})$ — the total peer effect once the endogenous feedback through peer outcomes is propagated to equilibrium — with delta-method standard errors. Results appear in Appendix Table 2 alongside the main-text NLLS and CF baseline estimates for ease of comparison.

We caution that the formal asymptotic theory for the joint estimator extends the arguments of Section 3 and Section 4 but requires additional structure beyond Assumptions 1–3 — in particular, a quantitative identification condition on the network suitable to Bramoullé et al. (2009)-type arguments. Such an extension lies outside the scope of the present paper, and we report the estimates here as suggestive empirical evidence.

For Spring 2019 (pre-COVID), the multiplier effect $(\hat{\beta} + \hat{\lambda})/(1 - \hat{\lambda})$ is estimated at 0.106 with a delta-method standard error of 0.045, statistically distinguishable from zero. The endogenous coefficient $\hat{\lambda} = 0.075$ is also statistically significant. The corresponding one-standard-deviation effect is 0.020, about 36% smaller than under the contextual-only specification, suggesting that the endogenous channel absorbs part of what would otherwise be attributed to the contextual channel.

For Spring 2020 (COVID-online), the multiplier effect is estimated at -0.038 with a delta-method standard error of 0.037, statistically indistinguishable from zero. The conclusion of the main text — that the COVID-online semester eliminates the classroom peer effect — is robust to allowing endogenous peer effects.

Appendix Table 2: UW-Madison register data for spring semesters in 2019 and 2020 — full version including endogenous peer effects

		Spring 2019			Spring 2020		
		NLLS	CF	CF + endog.	NLLS	CF	CF + endog.
$\hat{\beta}$		0.249 (0.024)	0.167 (0.033)	0.024 (0.070)	0.049 (0.023)	-0.002 (0.032)	-0.085 (0.054)
$\hat{\lambda}$				0.075 (0.032)			0.049 (0.027)
$\frac{\hat{\beta} + \hat{\lambda}}{1 - \hat{\lambda}}$				0.106 (0.045)			-0.038 (0.037)
$\hat{\sigma}_{\bar{\alpha}_{(i)j}}$	plug-in	0.210			0.157		
$\hat{\sigma}_{\bar{\alpha}_{(i)j}}$	KSS		0.189	0.191		0.142	0.143
1-sd effect		0.052	0.031	0.020	0.008	-0.000	-0.005

Notes: Wild bootstrap standard errors for NLLS (100 replications); the proposed standard errors based on $\hat{V}_n$ for cross-fit (CF). The *CF + endog.* column reports the joint estimator of (β, λ) in (24) with delta-method standard errors for the multiplier $(\hat{\beta} + \hat{\lambda})/(1 - \hat{\lambda})$; these standard errors are preliminary, pending the joint asymptotic theory discussed above.

Appendix F Data and institutional setting for the workplace application

This appendix collects additional information on the Veneto Worker History (VWH) sample used in Section 6: the full set of sample-selection restrictions (Appendix F.1), descriptive statistics on the resulting sample (Appendix F.2), firm-level mobility rates by firm size (Appendix F.3), robustness of the peer-group definition (Appendix F.4), and the institutional setting of the Italian labor market over the sample window (Appendix F.5).

F.1 Sample selection

We construct the analysis sample from the VWH database with the following restrictions, applied in order:

1. Restrict the panel to calendar years 1995–2001.
2. For workers with multiple employment relationships in the same year, retain only the primary job (highest annual earnings and weeks worked); ties are broken at random.
3. Restrict to workers aged 16–65.
4. Drop part-time and apprentice contracts, because wages in those contracts are set under different rules than for full-time regular employment.
5. Drop firms with more than 5,000 employees, following [Cornelissen et al. \(2017\)](#); these firms account for less than 2% of the observations.
6. Require each peer group (firm $\times$ broad occupation $\times$ year) to contain at least two workers. Workers in one-worker firms account for under 2% of observations and earn about 30% below the labor-market average.
7. Restrict to firms in the largest connected set of the worker-firm mobility graph ([Abowd et al., 1999](#)); the largest connected set covers about 98% of the remaining observations.

8. Restrict further to the *leave-one-out* connected set: the set of firms that remains connected after the removal of any single mover. This is the condition required by the bias-corrected variance estimator of Kline et al. (2020) and is the sample used in both Sections 6.2 and 6.3.

In observable characteristics — mean wages, age, tenure, gender composition, and occupational distribution — the leave-one-out sample is similar to the full sample, with a modest reduction in the number of firms and a slight increase in average firm size that reflects the stronger connectedness requirement.

F.2 Descriptive statistics

Table 3 reports descriptive statistics for the leave-one-out sample. The final sample contains approximately 4.8 million person-year observations covering 1.2 million workers and 69 thousand firms. The mean weekly wage is 833 euros (2003 prices). The average worker is 35 years old with about 5 years of tenure; 36% of workers change firms at least once over the seven-year window, and women make up 35% of the sample. Workers are predominantly in blue-collar occupations (68%), with white-collar workers at 30% and managers at 2%. Firms are small on average, consistent with the structure of the Italian labor market: the mean firm size is 21 employees with a median of 9; mean peer-group size is 10 with a median of 3.

F.3 Firm mobility by firm size

The mobility statistics cited in Section 6.2 as evidence for the strong-identification condition (Assumption 3(iv)) are documented at the firm level in Figure 2. On average, about 8% of workers move from another firm in a given year, and the average firm replaces around 20% of its workforce annually. As shown in the figure, this turnover rate is broadly similar across firm-size bins: even firms with more than 1,000 employees see annual turnover of around 15%. The pattern is informative for identification because the cross-fit estimator draws on movers and on coworker entry/exit around stayers, and the latter scales with firm-level turnover.

Appendix Table 3: Descriptive statistics for the leave-one-out sample, 1995–2001

	Mean	Standard Dev.	Median
<i>Worker level</i>			
Weekly wage (euros)	832.91	2731.75	701
Age	34.79	9.85	33
Tenure	4.60	5.02	3
Mover	0.36	0.48	
Woman	0.35	0.48	
Blue-collar	0.68	0.47	
White-collar	0.30	0.46	
Manager	0.02	0.14	
<i>Firm level</i>			
Firm size	21	92	9
Peer-group size	10	51	3
Annual mobility (workers)	5	36	2
Person-year observations		4,828,066	
Number of workers		1,203,965	
Number of firms		68,883	

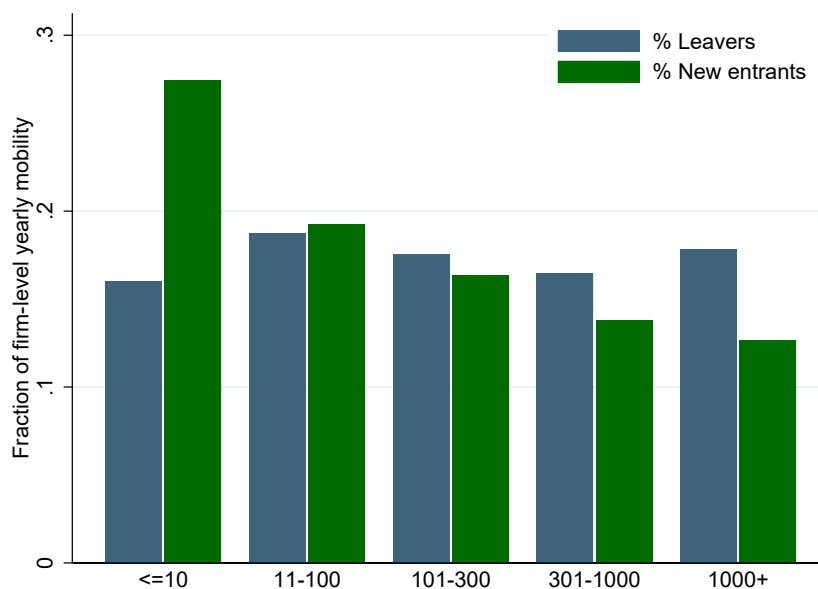
Notes: Sample is the leave-one-out connected set of VWH 1995–2001, after the sample restrictions enumerated in Appendix F.1. Wages are inflation-adjusted to 2003 prices.

F.4 Peer-group definition: robustness

The baseline peer-group definition is firm $\times$ broad occupation $\times$ year, with three broad occupational categories (blue-collar, white-collar, manager). Two concerns motivate robustness checks.

Managers as a separate group. Treating managers as a peer group distinct from blue- and white-collar workers may be conservative if managers in fact interact with the workers they supervise. As a robustness exercise, we reassign each manager to either blue- or white-collar based on the occupation she was promoted from. Where a worker has always been observed as a manager, we assign her to the modal non-manager occupation in the same firm-year (defaulting to blue-collar if all coworkers are blue-collar). Re-estimating equation (13) with this redefined peer group yields a peer coefficient nearly identical to the baseline, which is expected given that managers comprise only 2% of the sample.

Appendix Figure 2: Annual firm turnover by firm size



Notes: At the firm level, average fractions of newly entered and newly separated workers in a given year, plotted against firm size. VWH leave-one-out sample, 1995–2001.

Within-occupation heterogeneity. The broad-occupation definition can be too coarse if distinct occupations within a professional category (e.g., consultants and accountants within white-collar) do not in fact interact. Two observations bear on this concern. First, peer-group sizes are typically small (median 3), so cross-occupation interaction within a firm is plausible: in many firms, small white-collar teams span occupations and work jointly. Second, [Portugal et al. \(2024\)](#) document quantitatively similar peer-effect estimates under alternative occupation definitions using comparable Portuguese matched data. To the extent that residual measurement error in the peer-group definition remains, it generates the classical attenuation discussed in [Cornelissen et al. \(2017\)](#) and [Nix \(2016\)](#), and our reported coefficients can be read as a lower bound on the true peer effect.

F.5 Italian wage-setting institutions over the sample window

The sample window 1995–2001 sits within the relatively decentralized phase of Italian wage-setting institutions described in [Guiso, Pistaferri and Schivardi \(2005\)](#). A major restructuring of the Italian industrial-relations system in July 1993 reorganized

the relationship between sector-level and firm-level bargaining; we begin our sample in 1995 to allow the post-reform regime to settle and because firm-level information in the VWH is more accurate from 1995 onward. Within international comparisons, [Guiso et al. \(2005\)](#) report that the [Iversen \(1998\)](#) index of wage-bargaining centralization places Italy at the lower-centralization end of OECD countries over this period, with a value near 0.18 — closer to the United Kingdom (0.18), France (0.12), and Switzerland (0.25) than to the highly centralized Nordic economies (above 0.50). This decentralized institutional setting is the labor market in which we identify workplace peer effects, and our estimates should be read against that backdrop rather than against settings where wages are predominantly set at the national or sectoral level.