Standard Errors When a Regressor is Randomly Assigned

Denis Chetverikov[*] UCLA

Jinyong Hahn[†] UCLA

Zhipeng Liao[‡] UCLA

Andres Santos[§] UCLA

March 17, 2023

Abstract

We examine asymptotic properties of the OLS estimator when the values of the regressor of interest are assigned randomly and independently of other regressors. We find that the OLS variance formula in this case is often simplified, sometimes substantially. In particular, when the regressor of interest is independent not only of other regressors but also of the error term, the textbook homoskedastic variance formula is valid even if the error term and auxiliary regressors exhibit a general dependence structure. In the context of randomized controlled trials, this conclusion holds in completely randomized experiments with constant treatment effects. When the error term is heteroscedastic with respect to the regressor of interest, the variance formula has to be adjusted not only for heteroscedasticity but also for correlation structure of the error term. However, even in the latter case, some simplifications are possible as only a part of the correlation structure of the error term should be taken into account. In the context of randomized control trials, this implies that the textbook homoscedastic variance formula is typically not valid if treatment effects are heterogenous but heteroscedasticity-robust variance formulas are valid if treatment effects are independent across units, even if the error term exhibits a general dependence structure. In addition, we extend the results to the case when the regressor of interest is assigned randomly at a group level, such as in randomized control trials with treatment assignment determined at a group (e.g., school/village) level.

JEL Classification: C14, C31, C32

Keywords: Cluster Robust Inference; Randomized Control Trial

1 Introduction

Textbook discussion of linear regression usually begins with a standard model of the form $\mathbf{Y} = \mathbf{X}\theta + \boldsymbol{\epsilon}$, where it is assumed that $\mathbf{X}$ is a nonstochastic matrix (with full column rank) of regressors

[*] Department of Economics, UCLA, Los Angeles, CA 90095-1477 USA. Email: chetverikov@econ.ucla.edu

[†] Department of Economics, UCLA, Los Angeles, CA 90095-1477 USA. Email: hahn@econ.ucla.edu

[‡] Department of Economics, UCLA, Los Angeles, CA 90095-1477 USA. Email: zhipeng.liao@econ.ucla.edu

[§] Department of Economics, UCLA, Los Angeles, CA 90095-1477 USA. Email: andres@econ.ucla.edu

and the error vector $\boldsymbol{\epsilon}$ has mean zero and variance matrix proportional to an identity matrix. As is well known, such an assumption justifies the formula $s^2\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}$ as an estimator of the variance of the OLS estimator, where s^2 is equal to the sum of squares of the estimated residuals divided either by the sample size or the degrees of freedom. This formula is easy to use but, as is typically taught, may not be valid if the variance matrix Ω of the error vector $\boldsymbol{\epsilon}$ is not proportional to an identity matrix. In such cases, the variance of the OLS estimator should be based on the formula $\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\Omega\mathbf{X}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}$ to reflect the heteroscedasticity and dependence structure of the error vector. An important practical challenge in implementing such an approach is that the matrix Ω may be hard to estimate if the dependence structure of the error vector $\boldsymbol{\epsilon}$ is unknown. In this paper, we study the implications for the variance of OLS estimators of having a regressor of interest whose values are i.i.d. across units/time periods and are independent of values of other regressors. The primary motivating applications for our analysis are randomized controlled trials in which units are independently assigned to some treatment level without a connection to observable characteristics. The main finding of the paper is that variance estimation in this case is often simplified, sometimes substantially.

Let $\mathbf{D}$ be the column of $\mathbf{X}$ corresponding to the regressor of interest and let $\mathbf{W}$ be the remaining columns of $\mathbf{X}$; i.e. columns corresponding to controls. Our first main result shows that when the vector $\mathbf{D}$ has i.i.d. components and is *strongly exogenous* in the sense of being independent not only of $\mathbf{W}$ but also of $\boldsymbol{\epsilon}$, the OLS estimator is asymptotically normally distributed and the formula $s^2\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}$ actually yields a valid variance estimator for the coefficient of $\mathbf{D}$ even if Ω is not proportional to the identity matrix. This result, which superficially contradicts the lessons of elementary linear regression analysis, is due to the randomness of the $\mathbf{X}$ matrix in our analysis. While the textbook analysis assumes away the randomness by conditioning on $\mathbf{X}$, we instead obtain our result by recognizing that the randomness of the $\mathbf{X}$ matrix delivers a suitable martingale structure.[1] We recognize that a version of this result in some simple contexts is well understood in the profession in the sense that many can anticipate such a result when the entire matrix of regressors is strongly exogenous; see references below. We go one step further, however, and establish our result for the case when (i) only *one* regressor is strongly exogenous (e.g., treatment in a randomized controlled trial); and/or (ii) the error vector is subject to some generalized dependence more complicated than what is commonly understood to be the cluster structure, e.g. temporal/spatial autocorrelation or a network structure. This result is important because it facilitates inference on the coefficient of the regressor of interest even if the researcher does not know the dependence structure of the error vector $\boldsymbol{\epsilon}$, which is useful from the pragmatic point of view. We emphasize, however, that while our conclusions hold for the coefficient corresponding to

[1] The validity of the formula $s^2\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}$ does not mean that the variance estimators based on the formula $\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\Omega\mathbf{X}\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}$ are invalid; see Lemma 1 in the Appendix for the asymptotic equivalence of these estimators in our setting.

a strongly exogenous regressor, they need not hold for other coefficients in the regression.

Our second main result shows that when the vector $\mathbf{D}$ has i.i.d. components and is independent of $\mathbf{W}$ but $\boldsymbol{\epsilon}$ is *conditionally heteroscedastic* with respect to $\mathbf{D}$, the formula $s^2(\mathbf{X}^\top\mathbf{X})^{-\mathbf{1}}$ is actually not valid and has to be adjusted not only for heteroscedasticity *but also*, rather surprisingly, for the dependence structure of the vector $\boldsymbol{\epsilon}$. Nevertheless, a simplified variance formula can often be used in this case as well. For example, conditional heteroscedasticity arises when the regression model $\mathbf{Y} = \mathbf{X}\theta + \boldsymbol{\epsilon}$ is taken from the potential outcomes framework with heterogeneous treatment effects. If the researcher is concerned about clustering in this case, our results imply that it is sufficient to adjust the variance formula for clustering of the treatment effects only.[2] In contrast, for example, there is no need to adjust the variance formula for clustering of the potential outcomes in any given treatment arm. We also note that neither of our results restrict or exclude conditional heteroscedasticity of $\boldsymbol{\epsilon}$ with respect to $\mathbf{W}$.

In addition, we extend both results to the case when the values of the regressor of interest are independent only *across groups* of units/time periods, such as is the case in randomized controlled trials in which treatment assignment is determined at a group (e.g., school/village) level. We show that in the strongly exogenous case, it suffices to take into account only the within-group correlation of the error vector $\boldsymbol{\epsilon}$. In other words, it suffices to use variance estimators that are clustered at the level at which treatment is assigned. In the conditional heteroscedasticity case, the variance formula still requires some adjustments for both heteroscedasticity and dependence but often allows for some simplifications relative to the standard textbook formula mentioned above.

Our first main result and its extension to the group-level assignment are related to but different from those in Barrios, Diamond, Imbens, and Kolesar (2012), who came to the same conclusions for regressions without controls and in which a fixed fraction of units/clusters is randomly assigned to be treated. To the best of our knowledge, however, there are no results in the literature related to our second main result. Our analysis is also related to Abadie, Athey, Imbens, and Wooldridge (2017), who presented a new clustering framework that emphasizes a finite population perspective as well as interactions between the sampling and assignment parts of the data-generating process. They established in particular that there is no need to cluster when estimating the variance if the randomization is conducted at the individual level and there is no heteregeneity in the treatment effects. Our first main result echoes and complements their findings in the following aspects. First, unlike them, we do not impose a particular structure on the sampling process, which allows us to cover general forms of time series or even network dependence in addition to the cluster-type dependence. Second, our analysis goes beyond the binary treatment framework and accommodates a general strongly exogenous regressor as well as the inclusion of additional controls in the regression.

[2] When the regression model $\mathbf{Y} = \mathbf{X}\theta + \boldsymbol{\epsilon}$ is taken from the potential outcomes framework with heterogeneous treatment effects, the case of strongly exogenous regressor corresponds to the assumption of constant treatment effects.

In particular, we allow for general dependence structures in the control variables, which makes it ex-ante unclear at what level one should cluster. Third, we rely on a traditional asymptotic framework, which may make our analysis more familiar to the reader. We recognize, however, that the third aspect may be a weakness in the sense that our framework is unable to address the finite population adjustment that plays an important role in Abadie, Athey, Imbens, and Wooldridge (2017). Finally, we note Bloom (2005) who considered a random effects type cluster structure and produced a variance estimator for the simple difference of means estimator that is quoted in Duflo, Glennerster, and Kremer (2007). His equation (4.3), which is presented without proof, is in fact a special case of the $s^2 \left(\mathbf{X}^\top \mathbf{X}\right)^{-1}$ formula. The cluster structure that he analyzed has a built-in correlation among observations, and as such, it would be tempting to think that variance estimation would require Moulton (1986)'s clustering formula – a conclusion that can be motivated if inference is to be conditioned on the $\mathbf{X}$ matrix. Hence, in our view, his equation (4.3) can only be motivated by explicitly recognizing the randomness of the $\mathbf{X}$ matrix.

Our results are not particularly difficult to derive. On the other hand, we are unaware of any systematic discussion of results along this line in the literature besides Barrios, Diamond, Imbens, and Kolesar (2012) and Abadie, Athey, Imbens, and Wooldridge (2017), especially in models where the control variables and treatment variables are both present. Our results have convenient pragmatic implications, which we hope are helpful to some empirical researchers.

Outline. We present the basic intuition underlying our results in Section 2. The intuitive discussion there suggests that asymptotic normality and the formula $s^2 \left(\mathbf{X}^\top \mathbf{X}\right)^{-1}$ are valid under fairly general dependence structure in $\boldsymbol{\epsilon}$ provided that the randomness of $\mathbf{X}$ generates a suitable martingale structure in $\mathbf{X}^\top \boldsymbol{\epsilon}$. We also explain how conditional heteroscedasticity breaks down this martingale structure. We formalize our discussion in Section 3, where our main restriction on the dependence of $\boldsymbol{\epsilon}$ is that it be weak enough for its sample average to converge in probability to zero – a condition that further emphasizes that our analysis is driven by the randomness in $\mathbf{X}$ and not $\boldsymbol{\epsilon}$. We provide an extension to the case of group-level random assignment in Section 4.

Notation. We use K to denote a generic strictly positive constant that may change from place to place but is independent of the sample size n. For any positive integer k, let $\mathbf{I}_k$, $\mathbf{1}_{k\times 1}$, and $\mathbf{0}_{k\times 1}$ denote the $k \times k$ identity matrix, $k \times 1$ vector of ones, and $k \times 1$ vector of zeros. For any real square matrix A, $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ denote its smallest and largest eigenvalues. We use $A \equiv B$ to denote that A is defined as B, and $(A_j)_{j\leq J}$ to denote the matrix composed by sequentially stacking matrices $A_1, \ldots, A_J$ with equal number of columns.

2 Intuition

In this section, we provide intuition for the validity of the formula $s^2 \left(\mathbf{X}^\top \mathbf{X}\right)^{-1}$ in the case of strongly exogenous regressors. We first consider the case of a simple univariate regression model and then extend the result to the case of a multivariate regression model. At the end of this section, we explain complications arising conditional heteroscedasticity and how they break the validity of the formula $s^2\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}$.

2.1 Case of Strong Exogeneity

We start by considering a simple univariate linear regression time series model in which we have

$$y_i = d_i\beta + \varepsilon_i, \quad i = 1, \ldots, n, \tag{1}$$

where $(\varepsilon_i)_{i\leq n}$ is a second-order, possibly autocorrelated, stationary time series – we employ the index i, rather than t, to emphasize our analysis is not confined to the time series context. We depart from the textbook time series model by assuming that the regressors $(d_i)_{i\leq n}$ are: (i) Independent and identically distributed (i.i.d.) with mean zero, and (ii) Strongly exogenous in the sense that $(d_i)_{i\leq n}$ is independent of the time series process $(\varepsilon_i)_{i\leq n}$.

As is well-known, the least squares estimator $\hat{\beta}$ of β in model (1) satisfies the equality

$$\sqrt{n}(\hat{\beta} - \beta) = \frac{n^{-1/2}\sum_{i=1}^n d_i\varepsilon_i}{n^{-1}\sum_{i=1}^n d_i^2}. \tag{2}$$

In many standard time series textbooks, the asymptotic distribution of $\hat{\beta}$ is thus derived by imposing sufficiently strong conditions to ensure that the score $n^{-1/2}\sum_{i=1}^n d_i\varepsilon_i$ is asymptotically normal and the Hessian $n^{-1}\sum_{i=1}^n d_i^2$ converges in probability to a non-stochastic matrix. In order to derive the standard error for $\hat{\beta}$, we therefore only need a consistent estimator of the long-run variance of the score; i.e., a heteroscedasticity and autocorrelation consistent (HAC) variance estimator, such as those proposed by Newey and West (1987) and Andrews (1991).

On the other hand, if the regressors $(d_i)_{i\leq n}$ are i.i.d. and strongly exogenous with mean zero, the independence of $(d_i)_{i\leq n}$ and $(\varepsilon_i)_{i\leq n}$ implies that for any $1 \leq i_1, i_2 \leq n$ with $i_1 \neq i_2$, we must have

$$\mathbb{E}\left[d_{i_1}\varepsilon_{i_1}d_{i_2}\varepsilon_{i_2}\right] = \mathbb{E}\left[d_{i_1}\right]\mathbb{E}\left[d_{i_2}\right]\mathbb{E}\left[\varepsilon_{i_1}\varepsilon_{i_2}\right] = 0,$$

and also $\mathbb{E}[(d_i\varepsilon_i)^2] = \mathbb{E}[d_i^2]\mathbb{E}[\varepsilon_i^2]$ for all $1 \leq i \leq n$. Hence, as long as some version of the central limit theorem is applicable to the score $n^{-1/2}\sum_{i=1}^n d_i\varepsilon_i$ and a law of large numbers is applicable to the Hessian $n^{-1}\sum_{i=1}^n d_i^2$, we can conclude that the asymptotic distribution of $\hat{\beta}$ is given by

$$\sqrt{n}(\hat{\beta} - \beta) \to_d N\left(0, \ \frac{\mathbb{E}\left[\varepsilon_i^2\right]}{\mathbb{E}\left[d_i^2\right]}\right).$$

In particular, statistical inference on β can be conducted as if it were not a time series model, i.e. using the formula $s^2\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}$, where $\mathbf{X} \equiv (d_i)_{i\leq n}$ in this case.

The main takeaway of the preceding example is that the strong exogeneity and i.i.d. nature of the regressors $(d_i)_{i\leq n}$ imply that the sequence $(d_i\varepsilon_i)_{i\leq n}$ is homoscedastic *even if* the errors

$(\varepsilon_i)_{i\leq n}$ are arbitrarily autocorrelated. Is this simplification confined to the time series model? Our preceding discussion suggests that this is not the case. Indeed, provided the regressors $(d_i)_{i\leq n}$ are i.i.d., mean zero, and strongly exogenous, the score $n^{-1/2}\sum_{i=1}^n d_i\varepsilon_i$ has a built-in martingale structure vis-à-vis the filtration $\mathcal{F}_i$ generated by $(d_j,\varepsilon_j)_{j\leq i}^\top$ because:

$$\mathbb{E}\left[d_i\varepsilon_i\middle|\mathcal{F}_{i-1}\right]=\mathbb{E}\left[\mathbb{E}\left[d_i\middle|\mathcal{F}_{i-1},\varepsilon_i\right]\varepsilon_i\middle|\mathcal{F}_{i-1}\right]=\mathbb{E}[d_i]\mathbb{E}[\varepsilon_i|\mathcal{F}_{i-1}]=0. \tag{3}$$

Therefore, assuming that the random pairs (d_i,ε_i) satisfy certain moment conditions, the martingale central limit theorem will be applicable regardless of the dependence structure of $(\varepsilon_i)_{i\leq n}$ and the long run variance of the score will reduce to $\mathbb{E}[d_i^2]\mathbb{E}[\varepsilon_i^2]$. In particular, the variance formula $s^2\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}$ will remain valid despite the dependence present in the variables $(\varepsilon_i)_{i\leq n}$. Thus, spatial correlation, network dependence, and/or a cluster structure in the variables $(\varepsilon_i)_{i\leq n}$ are all accommodated by the standard homoscedastic standard errors. Moreover, we note that a quick inspection at the preceding argument reveals that the assumptions we have imposed so far are stronger than necessary for the desired conclusion to hold.

We next build on our preceding discussion by considering the multivariate linear regression model

$$y_i=d_i\beta+w_i^\top\gamma+\varepsilon_i,\quad i=1,\dots,n, \tag{4}$$

where d_i is a scalar regressor of interest, w_i is a d_w-vector of controls, and $(w_i,\varepsilon_i)_{i\leq n}$ is a second-order, possibly autocorrelated, stationary time series satisfying $\mathbb{E}[w_i\varepsilon_i]=\mathbf{0}_{d_w\times 1}$ for all $1\leq i\leq n$. We continue to assume that the regressors $(d_i)_{i\leq n}$ are i.i.d. with mean zero, and strongly exogenous in the sense that $(d_i)_{i\leq n}$ is independent of the time series process $(\varepsilon_i,w_i^\top)_{i\leq n}$. The parameter of interest continues to be β.

For this model, the Frisch-Waugh-Lovell theorem implies the least squares estimator $\hat\beta$ satisfies

$$\sqrt{n}\left(\hat\beta-\beta\right)=\frac{n^{-1/2}\sum_{i=1}^n\left(d_i-w_i^\top\hat\alpha\right)\varepsilon_i}{n^{-1}\sum_{i=1}^n\left(d_i-w_i^\top\hat\alpha\right)^2}\qquad \hat\alpha\equiv\left(\sum_{i=1}^n w_iw_i^\top\right)^{-1}\sum_{i=1}^n w_id_i. \tag{5}$$

Hence, under appropriate regularity conditions the estimator $\hat\beta$ admits the asymptotic expansion

$$\sqrt{n}\left(\hat\beta-\beta\right)=\frac{n^{-1/2}\sum_{i=1}^n\left(d_i-w_i^\top\alpha\right)\varepsilon_i}{n^{-1}\sum_{i=1}^n\left(d_i-w_i^\top\alpha\right)^2}+o_p(1)\qquad \alpha\equiv(\mathbb{E}[w_iw_i^\top])^{-1}\mathbb{E}[w_id_i]. \tag{6}$$

In particular, if $(d_i)_{i\leq n}$ is mean zero and independent of $(w_i^\top)_{i\leq n}$, then $\alpha=0$ and the asymptotic expansion reduces to the univariate setting – i.e. the right-hand side of (6) is (asymptotically) equivalent to the right-hand side of (2). It therefore follows that the end result is the same as in the univariate case: The variance formula $s^2\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}$, where $\mathbf{X}\equiv(d_i,w_i^\top)_{i\leq n}$ in this case, remains valid for $\hat\beta$. We emphasize, however, that this formula is not necessarily justified for conducting inference on the coefficient γ.

As a preview of results in the next section, we again note that the conditions we have imposed so far are stronger than required. For instance, suppose that instead of demanding that the regressors $(d_i)_{i\leq n}$ and $(w_i^\top)_{i\leq n}$ be fully independent, we assume that they are related according to the model

$$d_i = \alpha w_i + \eta_i,$$

for $(\eta_i)_{i\leq n}$ i.i.d. and independent of $(\varepsilon_i, w_i^\top)_{i\leq n}$. The asymptotic expansion in (6), combined with the same arguments employed in the univariate case, then continue to imply that

$$\sqrt{n}(\hat{\beta} - \beta) \to_d N\left(0,\ \frac{\mathbb{E}\left[\varepsilon_i^2\right]}{\mathbb{E}\left[\eta_i^2\right]}\right)$$

under mild moment conditions – i.e. when computing standard errors for $\hat{\beta}$ we can continue to pretend that the time series process fits the textbook homoscedastic model. Setting w_i to be a constant, for instance, reveals that the mean zero assumption on $(d_i)_{i\leq n}$ is superfluous.

2.2 Case of Conditional Heteroscedasticity

The preceding martingale argument relies crucially on two key assumptions: (i) The regressors $(d_i)_{i\leq n}$ are independent of each other, and (ii) The regressors $(d_i)_{i\leq n}$ are independent of the errors $(\varepsilon_i)_{i\leq n}$. A challenge to our martingale argument arises when $(d_i)_{i\leq n}$ and $(\varepsilon_i)_{i\leq n}$ are not independent. Within the potential outcome framework, for instance, this full independence requirement is violated in the presence of heterogenous treatment effects. More precisely, heterogenous treatment effects render ε_i conditionally heteroscedastic with respect to the treatment status d_i. Motivated by this observation, we also study a model in which $\varepsilon_i = \sum_{l=1}^{L} \sigma_l(d_i)\varepsilon_{l,i}^*$ with $e_i \equiv (\varepsilon_{l,i}^*)_{l\leq L}$ possibly correlated across i and l, but $(e_i)_{i\leq n}$ fully independent of $(d_i)_{i\leq n}$. In the potential outcome framework, with $d_i \in \{0,1\}$ indicating treatment status, we would have

$$\varepsilon_i = (y_i(0) - \mathbb{E}[y_i(0)]) + d_i((y_i(1) - y_i(0)) - \mathbb{E}[y_i(1) - y_i(0)]), \tag{7}$$

where $y_i(d)$ denotes the potential outcome for unit i under treatment status $d \in \{0,1\}$.

To see the problem for the martingale structure in this model, observe that for any $1 \leq i_1, i_2 \leq n$, we now have

$$\begin{aligned}\mathbb{E}[d_{i_1}\varepsilon_{i_1}d_{i_2}\varepsilon_{i_2}] &= \mathbb{E}\left[d_{i_1}\sum_{l_1=1}^{L}\sigma_{l_1}(d_{i_1})\varepsilon_{l_1,i_1}^* d_{i_2}\sum_{l_2=1}^{L}\sigma_{l_2}(d_{i_2})\varepsilon_{l_2,i_2}^*\right] \\ &= \sum_{l_1=1}^{L}\sum_{l_2=1}^{L}\mathbb{E}\left[d_{i_1}\sigma_{l_1}(d_{i_1})\right]\mathbb{E}\left[d_{i_2}\sigma_{l_2}(d_{i_2})\right]\mathbb{E}[\varepsilon_{l_1,i_1}^*\varepsilon_{l_2,i_2}^*],\end{aligned}$$

which is not necessarily zero even if d_i's are mean zero. The variance formula for the sum $\sum_{i\leq n} d_i\varepsilon_i$ therefore must include interactions terms as long as the random vectors $(\varepsilon_{l,i_1}^*)_{l\leq L}$ and $(\varepsilon_{l,i_2}^*)_{l\leq L}$ are correlated.

We will present a detailed analysis of the conditional heteroscedasticity case in Section 3 but the main takeaways from our results are: (i) The independence of $(d_i)_{i\leq n}$ from controls $(w_i^\top)_{i\leq n}$ still simplifies the asymptotic variance for $\hat{\beta}$; and (ii) Conditional heteroscedasticity yields a break in the martingale structure that requires us to adjust standard errors *not only* for heteroscedasticity, but *also* for correlation of the errors across units. For example, in the context of (7), our analysis implies the asymptotic variance of $\sqrt{n}(\hat{\beta}-\beta)$ equals (the probability limit of)

$$\frac{1}{n\sigma_d^4}\sum_{i=1}^{n}\mathbb{E}[(d_i-\mathbb{E}[d_i])^2\varepsilon_i^2]+\mathrm{Var}\left(n^{-1/2}\sum_{i=1}^{n}(y_i(1)-y_i(0))\right)-n^{-1}\sum_{i=1}^{n}\mathrm{Var}\left(y_i(1)-y_i(0)\right). \tag{8}$$

In particular, we note that standard errors may need to be adjusted for correlation if we are concerned the treatment effects are correlated. On the other hand, the correlation between components of the vector $(y_i(0))_{i\leq n}$ plays no role for the standard errors, and neither does correlation between components of the vector $(y_i(1))_{i\leq n}$.

3 Main Theory

We now present a rigorous theory showing that the variance formula $s^2\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}$ for the OLS estimator is valid in the strongly exogenous case even if the errors in the regression model are correlated. We also derive the variance formula in the case of conditional heteroscedasticity. In what follows, we suppose an outcome y_i satisfies

$$y_i = x_i^\top\theta+\varepsilon_i, \quad i=1,\ldots,n, \tag{9}$$

where $x_i=(d_i,w_i^\top)^\top$ is a vector of regressors, with d_i being a key regressor and w_i being a vector of controls, $\theta=(\beta,\gamma^\top)^\top$ is a vector of parameters, with β being a parameter of interest and γ being a vector of nuisance parameters, and ε_i is an error term with mean zero. More explicitly, the regression model (9) can be rewritten as

$$y_i = d_i\beta+w_i^\top\gamma+\varepsilon_i, \quad i=1,\ldots,n. \tag{10}$$

We assume that the first component of the vector w_i is a (non-zero) constant, meaning that the regression model (10) contains the intercept term. For notational simplicity, we set $\mathbf{Y}\equiv(y_i)_{i\leq n}$, $\mathbf{D}\equiv(d_i)_{i\leq n}$, $\mathbf{W}\equiv(w_i^\top)_{i\leq n}$, $\mathbf{X}\equiv(x_i^\top)_{i\leq n}$, and $\boldsymbol{\epsilon}\equiv(\varepsilon_i)_{i\leq n}$.

Under a suitable exogeneity assumption on $(d_i,w_i^\top)^\top$, the unknown parameter $\theta\equiv(\beta,\gamma^\top)^\top$ can be estimated by OLS:

$$\hat{\theta}\equiv(\hat{\beta},\hat{\gamma}^\top)^\top\equiv(\mathbf{X}^\top\mathbf{X})^{-1}(\mathbf{X}^\top\mathbf{Y}).$$

Standard estimators of the asymptotic variance of $\hat{\theta}$ rely on the asymptotic variance of the "score" $n^{-1/2}\sum_{i=1}^{n}(d_i,w_i^\top)^\top\varepsilon_i$, which may take a complicated form due to possible dependence between observations. This relationship makes standard error estimation and statistical inference challenging in practice. For instance, cluster-robust standard errors, such as those proposed by Moulton (1986), Liang and Zeger (1986), and Arellano (1987), are predicated on knowledge of the relevant group structure at which to cluster; see Hansen (2007), Ibragimov and Müller (2010) and Ibragimov and Müller (2016) for related discussion. Similarly, spatial standard errors, such as those proposed by Conley (1999), often require knowledge of a measure of "economic distance" that is relates to the degree of dependence across observations. However, we next show that when $\mathbf{D}$ is independent of $\mathbf{W}$, such as in randomized control trials, the asymptotic variance for $\hat{\beta}$ simplifies significantly. As a result, estimation of standard errors and asymptotically valid inference simplify as well.

3.1 Case of Strong Exogeneity

We first study the case that $\mathbf{D}$ is independent of $\boldsymbol{\epsilon}$, and hence $\boldsymbol{\epsilon}$ is conditionally homoscedastic with respect to $\mathbf{D}$. Together with independence of $\mathbf{D}$ from $\mathbf{W}$, this means that $\mathbf{D}$ is strongly exogenous. The conditional heteroscedasticity case is discussed later in this section. In the assumptions that follow, recall that K should be interpreted to denote a sufficiently large constant that can change from place to place but is independent of the sample size n.

Assumption 1. *(i) The random variables d_i, $1 \leq i \leq n$, are i.i.d. with mean μ_d and variance σ_d^2; (ii) $(d_i)_{i \leq n}$ is independent of $(w_i^\top)_{i \leq n}$; (iii) there exists a constant $\delta_1 > 0$ such that* $\max_{i \leq n} \mathbb{E}[|d_i|^{2+\delta_1}] \leq K$*; (iv) $\sigma_d^2 \geq K^{-1}$; (v) $(d_i)_{i \leq n}$ is independent of $(\varepsilon_i)_{i \leq n}$.*

Assumption 2. *(i) $\lambda_{\min}(n^{-1} \sum_{i=1}^n w_i w_i^\top) \geq K^{-1} + o_p(1)$; (ii) $\max_{i \leq n} \mathbb{E}[||w_i||^2] \leq K$; (iii) the first component of the vectors w_i, $1 \leq i \leq n$, is a non-zero constant.*

Assumption 3. *(i) $n^{-1} \sum_{i=1}^n w_i \varepsilon_i = o_p(1)$; (ii) $n^{-1} \sum_{i=1}^n \varepsilon_i^2 = n^{-1} \sum_{i=1}^n \mathbb{E}[\varepsilon_i^2] + o_p(1)$; (iii) there exists a constant $\delta_2 > 0$ such that* $\max_{i \leq n} \mathbb{E}[|\varepsilon_i|^{2+\delta_2}] \leq K$*; (iv) $n^{-1} \sum_{i=1}^n \mathbb{E}[\varepsilon_i^2] \geq K^{-1}$.*

Assumption 1 contains our main requirements for the regressor of interest. In particular, Assumptions 1(i) and 1(ii) are our key conditions that are satisfied in many randomized control trials. Assumptions 1(iii) and 1(iv) are mild moment conditions. Assumption 1(v) means that $(d_i)_{i \leq n}$ is strongly exogenous. Assumption 2 contains our main requirements for the controls. In particular, Assumption 2(i) means that there is no multicollinearity among controls. Assumption 2(ii) is a mild moment condition. Assumption 2(iii) means that we study regressions with an intercept. Assumption 3 contains our main requirements for the regression error. Assumption 3(i) holds if ε_i's are uncorrelated with w_i's and a law of large numbers applies to the product $w_i \varepsilon_i$. Assumption 3(ii) is essentially a law of large numbers for ε_i^2's. Assumptions 3(iii) and 3(iv) are mild moment conditions. We highlight that our assumptions allow for a wide array of dependence structures in the matrix $(\varepsilon_i, w_i^\top)_{i \leq n}$, with the main condition in this regard intuitively being that dependence be “weak” enough for the law of large numbers imposed in Assumptions 3(i) and 3(ii) to apply.

For all $1 \leq i \leq n$, denote $d_i^* \equiv d_i - \mu_d$. The following theorem derives the asymptotic distribution of the OLS estimator $\hat{\beta}$ in the strongly exogenous case.

Theorem 1. *Let Assumptions 1, 2 and 3 hold. Then*

$$\frac{\sqrt{n}(\hat{\beta}-\beta)}{\sigma_\varepsilon/\sigma_d} = \frac{n^{-1/2}\sum_{i=1}^n d_i^*\varepsilon_i}{\sigma_d\sigma_\varepsilon} + o_P(1) \to_d N(0,1),$$

where $\sigma_\varepsilon^2 \equiv n^{-1}\sum_{i=1}^n \mathrm{Var}(\varepsilon_i)$.

This theorem establishes two key facts. First, it shows that, given the strong exogeneity of $(d_i)_{i\leq n}$, our mild requirements on the dependence structure of $(\varepsilon_i, w_i^\top)_{i\leq n}$ suffice for establishing asymptotic normality of $\hat{\beta}$. To establish such a conclusion, we rely on a martingale construction that generalizes our discussion in Section 2. Second, Theorem 1 establishes that the asymptotic variance of $\hat{\beta}$ is not affected by the possible correlation across vectors $w_i\varepsilon_i$ since it only depends on the variance of d_i and the averaged variance of the error terms $(\varepsilon_i)_{i\leq n}$. We emphasize that neither of these conclusions need hold for the estimator $\hat{\gamma}$ of the coefficient γ corresponding to the vectors of controls w_i.

In addition, Theorem 1 suggests that we can estimate the variance of $\hat{\beta}$ by $s^2(\breve{\mathbf{D}}^\top\breve{\mathbf{D}})^{-1}$, where

$$s^2 \equiv n^{-1}(\mathbf{Y}-\mathbf{D}\hat{\beta}-\mathbf{W}\hat{\gamma})^\top(\mathbf{Y}-\mathbf{D}\hat{\beta}-\mathbf{W}\hat{\gamma}), \tag{11}$$

$\breve{\mathbf{D}} \equiv \mathbf{M}_W\mathbf{D}$ and $\mathbf{M}_W \equiv \mathbf{I}_n - \mathbf{W}(\mathbf{W}^\top\mathbf{W})^{-1}\mathbf{W}^\top$. The following corollary confirms this conjecture.

Corollary 1. *Let Assumptions 1, 2 and 3 hold. Then*

$$s^2(n^{-1}\breve{\mathbf{D}}^\top\breve{\mathbf{D}})^{-1} = \sigma_\varepsilon^2/\sigma_d^2 + o_p(1) \quad \textit{and} \quad \frac{\hat{\beta}-\beta}{\sqrt{s^2(\breve{\mathbf{D}}^\top\breve{\mathbf{D}})^{-1}}} \to_d N(0,1). \tag{12}$$

Together with Theorem 1, this corollary is our first main result. Indeed, it is well-known that the top left element of the matrix $(\mathbf{X}^\top\mathbf{X})^{-1}$ coincides with $(\breve{\mathbf{D}}^\top\breve{\mathbf{D}})^{-1}$. Therefore, this corollary justifies using the classic standard error formula $s^2\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}$ for inference on β, even though we allow for general dependence structures in the errors $(\varepsilon_i)_{i\leq n}$ and controls $(w_i^\top)_{i\leq n}$.

Remark. Theorem 1 and Corollary 1 can be extended to allow for dependence between d_i and w_i. Indeed, suppose that d_i depends linearly on w_i:

$$d_i = w_i^\top\alpha + \eta_i,$$

where η_i's satisfy the conditions of Assumption 1 imposed on d_i. By arguments that are similar to those in the proof of Theorem 1 and Corollary 1, it is then straightforward to show that

$$\frac{\sqrt{n}(\hat{\beta}-\beta)}{\sigma_\varepsilon/\sigma_\eta} \to_d N(0,1),$$

where σ_η^2 denotes the variance of η_i's, and that the convergence results (12) still hold. □

Remark. Theorem 1 and Corollary 1 can also be extended to instrumental variable (IV) estimators. Indeed, suppose that

$$d_i = v_i \rho + w_i^\top \alpha + \eta_i, \tag{13}$$

where v_i is an instrumental variable satisfying the conditions of Assumption 1 imposed on d_i and η_i is a (first-stage) estimation error satisfying the conditions of Assumption 3 imposed on ε_i. In addition denote $\mathbf{V} \equiv (v_i)_{i \le n}$ and define $\mathbf{M}_{V,W}$ and $\mathbf{M}_V$ the same way as $\mathbf{M}_W$ with $\mathbf{W}$ replaced by $(\mathbf{V}, \mathbf{W})$ and $\mathbf{V}$, respectively. The two-stage least squared (2SLS) estimator of β then satisfies

$$\hat{\beta}_{2sls} = \frac{\mathbf{D}^\top (\mathbf{M}_W - \mathbf{M}_{V,W}) \mathbf{Y}}{\mathbf{D}^\top (\mathbf{M}_W - \mathbf{M}_{V,W}) \mathbf{D}}.$$

Using Assumptions 2 and 3, it is then possible to show that

$$\sqrt{n}(\hat{\beta}_{2sls} - \beta) = \frac{n^{-1/2} \sum_{i=1}^n (v_i - \mu_v) \varepsilon_i}{\rho \sigma_v^2} + o_p(1),$$

where μ_v is the mean of v_i's and σ_v^2 is the variance of v_i's. Therefore,

$$\frac{\sqrt{n}(\hat{\beta}_{2sls} - \beta)}{\sigma_\varepsilon / (\rho \sigma_v)} \to_d N(0,1).$$

It is clear that σ_ε^2 can be estimated by the same s^2 as that in (11) with $\hat{\beta}$ replaced by $\hat{\beta}_{2sls}$, ρ can be estimated by OLS on the (first-stage) regression (13), and σ_v^2 can be estimated by $n^{-1} \tilde{\mathbf{V}}^\top \tilde{\mathbf{V}}$, where $\tilde{\mathbf{V}} = \mathbf{V} - \mathbf{1}_{n,1}(n^{-1} \mathbf{V}^\top \mathbf{1}_{n \times 1})$. □

3.2 Case of Conditional Heteroscedasticity

Next, we derive the asymptotic variance for $\hat{\beta}$ in the case of conditional heteroscedasticity, i.e. when $(\varepsilon_i)_{i \le n}$ is conditionally heteroscedastic with respect to $(d_i)_{i \le n}$. Following the notation introduced in Section 2, we focus on the case in which $\varepsilon_i \equiv \sum_{l \le L} \sigma_l(d_i) \varepsilon_{l,i}^*$, where $e_i \equiv (\varepsilon_{l,i}^*)_{l \le L}$ is a vector with mean zero, for all $1 \le i \le n$.

Let $A_i \equiv (d_i - \mu_d)(\sigma_l(d_i))_{l \le L}$ for all $1 \le i \le n$ and observe that under Assumption 1(i), the random vectors A_i are i.i.d. Denote their common mean vector by μ_A. In addition, denote $\sigma_{e,1}^2 \equiv n^{-1} \sum_{i=1}^n \mathbb{E}[((A_i - \mu_A)^\top e_i)^2]$ and $\sigma_{e,2}^2 \equiv \mathbb{E}[(n^{-1/2} \sum_{i=1}^n \mu_A^\top e_i)^2]$. Within this context, we impose the following assumptions.

Assumption 4. *(i) $(d_i)_{i \le n}$ is independent of $(e_i)_{i \le n}$; (ii) the functions σ_l, $1 \le l \le L$, are bounded; (iii) $\sigma_{e,1}^2 \ge K^{-1}$.*

Assumption 5. *(i) $\| n^{-1} \sum_{i=1}^n e_i e_i^\top - n^{-1} \sum_{i=1}^n \mathbb{E}[e_i e_i^\top] \| = o_p(1)$; (ii) there exists a constant $\delta_3 > 0$ such that $n^{-1} \sum_{i=1}^n \|e_i\|^{2+\delta_3} = O_p(1)$; (iii) $\sigma_{e,2}^{-1} n^{-1/2} \sum_{i=1}^n \mu_A^\top e_i \to_d N(0,1)$.*

Assumption 4 is mainly used to replace Assumption 1(v) and accounts for the conditional heteroscedasticity of $(\varepsilon_i)_{i\leq n}$ with respect to $(d_i)_{i\leq n}$. Assumption 4(i) requires that $(d_i)_{i\leq n}$ are strongly exogenous with respect to the"scaled" error vector $(e_i)_{i\leq n}$ – a requirement that, as discussed in Section 2 maps well into a potential outcome framework with heterogeneous treatment effects. Assumption 4(ii) imposes upper bounds on the functions $\sigma_l^2(\cdot)$, which we view as a mild regularity condition. Assumption 4(iii) is a mild moment condition. Assumption 5 contains further restrictions on the vectors e_i. Assumption 5(i) is essentially a law of large numbers for $e_i e_i^\top$'s. Assumption 5(ii) is essentially a moment condition for the random variables $\|e_i\|$. Assumption 5(iii) limits the amount of dependence among the vectors e_i to ensure convergence in distribution.

The following theorem, which is our second main result, derives the asymptotic distribution of the OLS estimator $\hat{\beta}$ in the case of conditional heteroscedasticity.

Theorem 2. *Let Assumptions 1(i)-(iv), 2, 3, 4, and 5 hold. Then*

$$\frac{\sqrt{n}(\hat{\beta}-\beta)}{\sigma_{d\varepsilon}/\sigma_d^2} = \frac{n^{-1/2}\sum_{i=1}^n d_i^*\varepsilon_i}{\sigma_{d\varepsilon}} + o_p(1) \to_d N(0,1), \tag{14}$$

where $\sigma_{d\varepsilon}^2 \equiv \sigma_{e,1}^2 + \sigma_{e,2}^2$.

To establish this theorem, we decompose the score $n^{-1/2}\sum_{i=1}^n d_i^*\varepsilon_i$ into the sum of two uncorrelated terms,

$$n^{-1/2}\sum_{i=1}^n d_i^*\varepsilon_i = n^{-1/2}\sum_{i=1}^n (A_i-\mu_A)^\top e_i + n^{-1/2}\sum_{i=1}^n \mu_A^\top e_i,$$

and observe that the first term on the right-hand side here forms a martingale difference sequence while the second term is asymptotically normal by assumption. Combining these facts, we are able to obtain asymptotic normality of the sum; see the detailed proof in the Appendix.

Like Theorem 1, this theorem establishes two key facts as well. To see both of them, observe that the term $\sigma_{d\varepsilon}^2$ appearing in the convergence result (14) can be more explicitly rewritten as

$$\sigma_{d\varepsilon}^2 = n^{-1}\sum_{i=1}^n \mathbb{E}\left[(d_i^*)^2\varepsilon_i^2\right] + \mu_A^\top\left[\operatorname{Var}\left(n^{-1/2}\sum_{i=1}^n e_i\right) - n^{-1}\sum_{i=1}^n \operatorname{Var}(e_i)\right]\mu_A. \tag{15}$$

This expression in turn implies that the asymptotic variance of the OLS estimator now depends on the correlation across the vectors e_i, which means that *heteroscedasticity* of the regression errors forces OLS variance estimators to be adjusted for *correlation* across the errors. On the other hand, the expression (15) also demonstrates that it suffices to adjust the variance estimators only for correlation across the random variables $\mu_A^\top e_i$, instead of correlation across the full vectors e_i. The latter point might seem like a minor technicality but it in fact plays an interesting role in models with heterogeneous treatment effects. Indeed, when $d_i \in \{0,1\}$ represents the treatment assignment status, $(y_i(1), y_i(0))$ represents the pair of potential outcomes with and without treatment, so that $y_i = y_i(0) + d_i(y_i(1) - y_i(0))$, and w_i consists of a non-zero constant only, it follows that ε_i takes the form (7), which matches the setting here with $e_i = (y_i(0) - \mathbb{E}[y_i(0)], y_i(1) - y_i(0) - \mathbb{E}[y_i(1) - y_i(0)])^\top$ and $\sigma(d_i) = (1, d_i)^\top$. Hence, $\mu_A = (0, \sigma_d^2)^\top$, and so the expression for the asymptotic variance of $\sqrt{n}(\hat{\beta} - \beta)$ reduces to the probability limit of

$$\frac{1}{n\sigma_d^4}\sum_{i=1}^n \mathbb{E}[(d_i - \mu_d)^2\varepsilon_i^2] + \text{Var}\left(n^{-1/2}\sum_{i=1}^n (y_i(1) - y_i(0))\right) - n^{-1}\sum_{i=1}^n \text{Var}\left(y_i(1) - y_i(0)\right),$$

as previewed in the previous section; see expression (8) there. In turn, this expression means that it suffices to adjust OLS variance estimation for correlation across treatment effect, and there is no need to worry about correlation of potential outcomes within any given treatment arm. For example, whenever treatment effects $y_i(1) - y_i(0)$ are independent across i, it suffices to use the usual heteroscedasticity-robust variance formulas, even if regression errors ε_i are correlated. Note, however, that it is *necessary* to use heteroscedasticity-robust variance formulas even if the treatment effects are i.i.d. (and not just independent), as the formula $s^2\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}$ is not valid in this case. Moreover, the same results apply even if w_i's include non-constant controls as well, as the second component of e_i's remains the same in this case.

Finally, we note that as long as the form of correlation across random variables $\mu_A^\top e_i$ is known, estimation of the asymptotic variance based on Theorem 2 is conceptually straightforward. For example, if treatment effects are clustered, it suffices to use Moulton (1986)'s or Liang and Zeger (1986)'s formulas assuming that regression errors are clustered at the same level (even though they could be clustered at a different level because of the clustering of potential outcomes without treatment, for example). For brevity of the paper, we do not provide formal statement of such results, as they are case-specific and depend on the form of the correlation structure, e.g. time series versus cluster versus spatial dependence.

4 Group-Level Randomization

A key assumption behind the results of Section 3 is that the regressor of interest is independent and identically distributed across units. In randomized controlled trials, such an assumption is satisfied for completely randomized assignments but can fail under other randomization protocols. For instance, in randomized controlled trials the i.i.d. assumption on the treatment fails when treatment is assigned at a group level. Motivated by this challenge, we suppose in this section that

$$y_{i,j} = d_j\beta + w_{i,j}^\top\gamma + \varepsilon_{i,j}, \tag{16}$$

where the index $j = 1, \ldots n_g$ denotes the group membership, the index $i = 1, \ldots, n_j$ denotes units within group j, n_g is the number of groups, n_j is the number of units within group j, d_j denotes the regressor of interest which is invariant within group j, $w_{i,j}$ denotes the vector of controls, and $\varepsilon_{i,j}$ is a mean-zero regression error. We continue to assume that the first component of each $w_{i,j}$ is a non-zero constant and also continue to employ n as the total number of observations, $n = \sum_{j \leq n_g} n_j$.

We emphasize at this point that the index j distinguishes the level at which the regressor d_j varies, but has no special significance for other variables. In particular, we do not insist that any (potential) clustering structure of the errors $((\varepsilon_{i,j})_{i \leq n_j})_{j \leq n_g}$ be the same as the group structure specified by the regressor d_j. In other words, $((\varepsilon_{i,j})_{i \leq n_j})_{j \leq n_g}$ may have a dependence structure completely different from group structure determined by j – e.g., ε_{i_1,j_1} may be correlated with ε_{i_2,j_2} when $j_1 \neq j_2$. This discussion will be formalized in the nature of the regularity conditions below, which do not include, for example, the random effects type specification as in Moulton (1986). Instead, as in the previous section, we will rely on a martingale structure based on $(d_j)_{j \leq n_g}$ that will be crucial for understanding the asymptotic distribution of $\hat{\beta}$.

Following our discussion in the previous section, we consider cases of strong exogeneity and conditional heteroscedasticity separately. We first consider the case of strong exogeneity. In order to study the asymptotic properties of $\hat{\beta}$, we first need to revise Assumptions 1 and 3 to account for the group structure and a possible within-group correlation. To this end, we let $\kappa_n \equiv (\sum_{j \leq n_g} n_j^2)^{1/2}$ and $S_\varepsilon^2 \equiv n^{-1} \sum_{j \leq n_g} \mathbb{E}[(\sum_{i \leq n_j} \varepsilon_{i,j})^2]$ and impose the following assumptions.

Assumption 6. *(i) The random variables d_j, $1 \leq j \leq n_g$, are i.i.d. with mean μ_d and variance σ_d^2; (ii) $(d_j)_{j \leq n_g}$ is independent of $((w_{i,j}^\top)_{i \leq n_j})_{j \leq n_g}$; (iii) $\max_{j \leq n_g} \mathbb{E}[|d_j|^4] \leq K$; (iv) $\sigma_d^2 \geq K^{-1}$; (v) $(d_j)_{i \leq n_g}$ is independent of $((\varepsilon_{i,j})_{i \leq n_j})_{j \leq n_g}$.*

Assumption 7. *(i) $n^{-1} \sum_{j=1}^{n_g} \sum_{i=1}^{n_j} w_{i,j} \varepsilon_{i,j} = o_p(S_\varepsilon n^{1/2}/\kappa_n)$; (ii) $S_\varepsilon^{-2} n^{-1} \sum_{j=1}^{n_g} (\sum_{i=1}^{n_j} \varepsilon_{i,j})^2 \to_p 1$; (iii) there exists a constant $\delta_4 \in (0, 2]$ such that $n^{-1-\delta_4/2} \sum_{j=1}^{n_g} (|\sum_{i=1}^{n_j} \varepsilon_{i,j}|)^{2+\delta_4} = o_p(S_\varepsilon^{2+\delta_4})$; (iv) $\kappa_n/n = o(1)$.*

Assumption 6 requires mild moment restrictions and that the regressor of interest d_j be strongly exogenous in the sense that it be independent of the errors and other regressors. To make sense of Assumption 7, assume that each group j has the same size n_j that is independent of n. In this case, κ_n is of order $\sqrt{n}$ and S_ε^2 is typically of order one. In turn, the latter implies that Assumption 7(i) reduces to $n^{-1} \sum_{j \leq n_g} \sum_{i \leq n_j} w_{i,j} \varepsilon_{i,j} = o_p(1)$, which is similar to Assumption 3(i), and Assumption 7(iii) reduces to $n^{-1-\delta_4/2} \sum_{j=1}^{n_g} \sum_{i=1}^{n_j} |\varepsilon_{i,j}|^{2+\delta_4} = o_p(1)$, which is satisfied as long as $\max_{1 \leq j \leq n_g} \max_{1 \leq i \leq n_j} |\varepsilon_{i,j}|^{2+\delta_4} \leq K$. In addition, Assumption 7(iv) reduces $n^{-1/2} = o(1)$ and is satisfied automatically and Assumption 7(ii) can be regarded as a law of large numbers. Thus, Assumption 7 in general requires that the size of groups does not increase too fast. Note also that, as previously claimed, this assumption does not require $((w_{i,j}^\top, \varepsilon_{i,j})_{i \leq n_j})_{j \leq n_g}$ to be independent across j.

We are now ready to derive the asymptotic distribution of the OLS estimator $\hat{\beta}$ in the strongly exogenous case with a group-level assignment. In the statement of the result, we impose Assumption 2, which should be interpreted to hold with $\sum_{j=1}^{n_g}\sum_{i=1}^{n_j}$ in place of $\sum_{i=1}^{n}$ in Assumption 2(i) and $\max_{j\leq n_g}\max_{i\leq n_j}$ in place of $\max_{i\leq n}$ in Assumption 2(ii). Also, we denote $d_j^* \equiv d_j - \mu_d$ for all $1 \leq j \leq n_g$.

Theorem 3. *Let Assumptions 2, 6 and 7 hold. Then*

$$\frac{\sqrt{n}(\hat{\beta}-\beta)}{S_\varepsilon/\sigma_d} = \frac{n^{-1/2}\sum_{j=1}^{n_g} d_j^* \sum_{i=1}^{n_j}\varepsilon_{i,j}}{\sigma_d S_\varepsilon} + o_p(1) \to_d N(0,1). \tag{17}$$

The implications of this theorem are following. First, the OLS estimator $\hat{\beta}$ is asymptotically normally distributed under mild moment conditions and under fairly general assumptions on the dependence structure of $((w_{i,j}^\top, \varepsilon_{i,j})_{i\leq n_j})_{j\leq n_g}$. Second, the asymptotic variance of $\hat{\beta}$ is given by $S_\varepsilon^2/\sigma_d^2$. In particular, since $S_\varepsilon^2 = n^{-1}\sum_{j\leq n_g}\mathrm{Var}(\sum_{i\leq n_j}\varepsilon_{i,j})$, when computing standard errors for $\hat{\beta}$ we need only account for possibly within-group j correlation *even if* $((\varepsilon_{i,j})_{i\leq n_j})_{j\leq n_g}$ is dependent across j. Importantly, the group structure is determined solely by the variables $(d_j)_{j\leq n_g}$ and hence is known, considerably simplifying estimation. For instance, in a randomized controlled trial with constant effects, Theorem 3 implies we may, for example, employ Moulton (1986)'s or Liang and Zeger (1986) formulas clustered at the level at which treatment was assigned. We again emphasize, however, that similar conclusions do not apply for $\hat{\gamma}$ whose standard errors and asymptotic normality may depend on the dependence structure of $((w_{i,j}^\top, \varepsilon_{i,j})_{i\leq n_j})_{j\leq n_g}$ across j.

Next, we consider the case of conditional heteroscedasticity. Denote $S_{e,1}^2 \equiv n^{-1}\sum_{j=1}^{n_g}\mathbb{E}[((A_j - \mu_A)^\top\sum_{i=1}^{n_j} e_{i,j})^2]$ and $S_{e,2}^2 \equiv \mathbb{E}[(n^{-1/2}\sum_{i=1}^{n}\mu_A^\top e_i)^2]$. Note that $S_{e,2}$ here actually coincides with $\sigma_{e,2}$ in the previous section. Within this context, we impose the following assumptions.

Assumption 8. *(i) $(d_j)_{j\leq n_g}$ is independent of $((e_{i,j}^\top)_{i\leq n_j})_{j\leq n_g}$; (ii) the functions σ_l, $1\leq l\leq L$, are bounded; (iii) $S_{e,1}/S_\varepsilon \geq K^{-1}$.*

Assumption 9. *(i) $\|n^{-1}\sum_{j=1}^{n_g}(\sum_{i=1}^{n_j} e_{i,j})(\sum_{i=1}^{n_j} e_{i,j})^\top - n^{-1}\sum_{j=1}^{n_g}\mathbb{E}[(\sum_{i=1}^{n_j} e_{i,j})(\sum_{i=1}^{n_j} e_{i,j})^\top]\| = o_p(S_{e,1}^2)$; (ii) $n^{-2}\sum_{j=1}^{n_g}\|\sum_{i=1}^{n_j} e_{i,j}\|^4 = o_p(S_{e,1}^4)$; (iii) $S_{e,2}^{-1}n^{-1/2}\sum_{i=1}^{n}\mu_A^\top e_i \to_d N(0,1)$.*

These assumptions naturally extend Assumptions 4 and 5 in the previous section to allow for group-level assignments.

The next theorem derives the asymptotic distribution of the OLS estimator $\hat{\beta}$ in the case of conditional heteroscedasticity with a group-level assignment.

Theorem 4. *Let Assumptions 2, 6(i)-(iv), 7, 8, and 9 hold. Then*

$$\frac{\sqrt{n}(\hat{\beta}-\beta)}{S_{d\varepsilon}/\sigma_d^2} = \frac{n^{-1/2}\sum_{j=1}^{n_g} d_j^* \sum_{i=1}^{n_j}\varepsilon_{i,j}}{S_{d\varepsilon}} + o_p(1) \to_d N(0,1), \tag{18}$$

where $S_{d\varepsilon} \equiv S_{e,1}^2 + S_{e,2}^2$.

This theorem relates to Theorem 3 in the same way as Theorem 2 relates to Theorem 1. In particular, noting that the term $S^2_{d\varepsilon}$ appearing in this theorem can be more explicitly rewritten as

$$S^2_{d\varepsilon} \equiv n^{-1} \sum_{j=1}^{n_g} \mathbb{E}\left[\left(d^*_j \sum_{i=1}^{n_j} \varepsilon_{i,j}\right)^2\right] + \mu_A^\top \left[\operatorname{Var}\left(n^{-1/2} \sum_{j=1}^{n_g} \sum_{i=1}^{n_j} e_{i,j}\right) - n^{-1} \sum_{j=1}^{n_g} \operatorname{Var}\left(\sum_{i=1}^{n_j} e_{i,j}\right)\right] \mu_A,$$

we conclude that because of conditional heteroscedasticity, variance estimators that are clustered at the group level at which the regressor d_j is assigned may not be valid if the errors $e_{i,j}$ are correlated across groups j. On the other hand, in the context of estimation with heterogeneous treatment effects, such estimators are valid if the treatment effects are uncorrelated across these groups.

References

Abadie, A., S. Athey, G. W. Imbens, and J. Wooldridge (2017): "When should you adjust standard errors for clustering?," Discussion paper, National Bureau of Economic Research.

Andrews, D. W. K. (1991): "Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimation," *Econometrica*, 59(3), 817–858.

Arellano, M. (1987): "Computing robust standard errors for within-groups estimators," *Oxford bulletin of Economics and Statistics*, 49(4), 431–434.

Barrios, T., R. Diamond, G. Imbens, and M. Kolesar (2012): "Clustering, spacial correlations, and randomization inference," *Journal of American Statistical Association*, 107, 578–591.

Bloom, H. S. (2005): *Learning more from social experiments: Evolving analytic approaches.* Russell Sage Foundation.

Conley, T. G. (1999): "GMM estimation with cross sectional dependence," *Journal of econometrics*, 92(1), 1–45.

DasGupta, A. (2008): *Asymptotic Theory of Statistics and Probability.* Springer.

Duflo, E., R. Glennerster, and M. Kremer (2007): "Using randomization in development economics research: A toolkit," *Handbook of development economics*, 4, 3895–3962.

Hall, P., and C. C. Heyde (1980): *Martingale Limit Theory and Its Application.* Academic Press, Inc. New York.

Hansen, C. B. (2007): "Asymptotic properties of a robust variance matrix estimator for panel data when T is large," *Journal of Econometrics*, 141(2), 597–620.

Ibragimov, R., and U. K. Müller (2010): "t-Statistic based correlation and heterogeneity robust inference," *Journal of Business & Economic Statistics*, 28(4), 453–468.

——— (2016): "Inference with few heterogeneous clusters," *Review of Economics and Statistics*, 98(1), 83–96.

Liang, K.-Y., and S. Zeger (1986): "Longitudinal Data Analysis Using Generalized Linear Models," *Biometrica*, 73, 13–22.

Moulton, B. R. (1986): "Random group effects and the precision of regression estimates," *Journal of econometrics*, 32(3), 385–397.

Newey, W. K., and K. D. West (1987): "A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix," *Econometrica*, 55(3), 703–708.

White, H. (2014): *Asymptotic theory for econometricians*. Academic press.

Appendix

A Proof of the main results

Proof of Theorem 1. Define $\tilde{\mathbf{D}} \equiv \mathbf{D} - \mathbf{1}_{n\times 1}(n^{-1}\mathbf{D}^\top \mathbf{1}_{n\times 1})$ and $\mathbf{M}_W \equiv \mathbf{I}_n - \mathbf{W}(\mathbf{W}^\top\mathbf{W})^{-1}\mathbf{W}^\top$. Then, given that the matrix $\mathbf{W}$ includes a non-zero constant column by Assumption 2(iii), it follows that $\mathbf{1}_{n\times 1}^\top \mathbf{M}_W = 0$. Therefore, by the Frisch-Waugh-Lovell theorem,

$$\hat{\beta} - \beta = (\mathbf{D}^\top \mathbf{M}_W \mathbf{D})^{-1}(\mathbf{D}^\top \mathbf{M}_W \boldsymbol{\epsilon}) = (\tilde{\mathbf{D}}^\top \mathbf{M}_W \tilde{\mathbf{D}})^{-1}(\tilde{\mathbf{D}}^\top \mathbf{M}_W \boldsymbol{\epsilon}). \tag{19}$$

Also, denoting $\overline{d}_n \equiv n^{-1}\sum_{i=1}^n d_i$, we have

$$\mathbb{E}[|\overline{d}_n - \mu_d|^2] = \sigma_d^2/n \leq K/n = o(1)$$

by Assumptions 1(i,iii), and so $\overline{d}_n - \mu_d = o_p(1)$ by Markov's inequality. Hence,

$$\overline{d}_n^2 - \mu_d^2 = (\overline{d}_n - \mu_d)(\overline{d}_n + \mu_d) = o_p(1)$$

by Assumption 1(iii) again. In addition,

$$\mathbb{E}\left[\left|n^{-1}\sum_{i=1}^n (d_i^2 - \mathbb{E}[d_i^2])\right|^{1+\delta_1/2}\right] \leq 2\sum_{i=1}^n \mathbb{E}\left[|d_i^2 - \mathbb{E}[d_i^2]|^{1+\delta_1/2}\right]/n^{1+\delta_1/2} = o(1)$$

by the von Bahr-Esseen Inequality (see Section 35.1.5 in DasGupta (2008)) and Assumptions 1(i,iii). Hence, by Markov's inequality,

$$n^{-1}\sum_{i=1}^n d_i^2 = n^{-1}\sum_{i=1}^n \mathbb{E}[d_i^2] + o_p(1),$$

and so

$$n^{-1}\tilde{\mathbf{D}}^\top\tilde{\mathbf{D}} = n^{-1}\sum_{i=1}^n \mathbb{E}[d_i^2] - \mu_d^2 + o_p(1) = \sigma_d^2 + o_p(1). \tag{20}$$

Further,

$$n^{-1}\tilde{\mathbf{D}}^\top \mathbf{W} = n^{-1}\sum_{i=1}^n d_i^* w_i - \left(n^{-1}\sum_{i=1}^n d_i^*\right)\left(n^{-1}\sum_{i=1}^n w_i\right). \tag{21}$$

By Assumptions 1(i, ii) and 2(iii), we further obtain

$$\mathbb{E}\left[\left\|n^{-1/2}\sum_{i=1}^n d_i^* w_i\right\|^2\right] = n^{-1}\sum_{i=1}^n \mathbb{E}\left[||d_i^* w_i||^2\right] \leq n^{-1}\sum_{i=1}^n \mathbb{E}[d_i^2]\mathbb{E}\left[||w_i||^2\right] \leq K.$$

Hence, by Markov's inequality,

$$n^{-1}\sum_{i=1}^n d_i^* w_i = O_p(n^{-1/2}). \tag{22}$$

Similarly, we can use Assumptions 1(i, iii) and 2(ii) to show

$$\left(n^{-1}\sum_{i=1}^{n} d_i^*\right)\left(n^{-1}\sum_{i=1}^{n} w_i\right) = O_p(n^{-1/2}), \tag{23}$$

which together with (21) and (22) implies that

$$n^{-1}\tilde{\mathbf{D}}^\top \mathbf{W} = O_p(n^{-1/2}). \tag{24}$$

Combining this result with (20) and using Assumptions 2(i), we then have

$$n^{-1}\tilde{\mathbf{D}}^\top \mathbf{M}_W \tilde{\mathbf{D}} = n^{-1}\tilde{\mathbf{D}}^\top\tilde{\mathbf{D}} - n^{-1}\tilde{\mathbf{D}}^\top \mathbf{W}(n^{-1}\mathbf{W}^\top\mathbf{W})^{-1}n^{-1}\mathbf{W}^\top\tilde{\mathbf{D}} = \sigma_d^2 + o_p(1). \tag{25}$$

For the term $\tilde{\mathbf{D}}^\top \mathbf{M}_W \boldsymbol{\epsilon}$ in the numerator of $\hat{\beta} - \beta$ in (19), we have

$$\begin{aligned} n^{-1/2}\tilde{\mathbf{D}}^\top \mathbf{M}_W \boldsymbol{\epsilon} &= n^{-1/2}\tilde{\mathbf{D}}^\top \boldsymbol{\epsilon} - n^{-1/2}\tilde{\mathbf{D}}^\top \mathbf{W}(\mathbf{W}^\top\mathbf{W})^{-1}\mathbf{W}^\top\boldsymbol{\epsilon} \\ &= n^{-1/2}\sum_{i=1}^{n} d_i^*\varepsilon_i - \left(n^{-1/2}\sum_{i=1}^{n} d_i^*\right)\left(n^{-1}\sum_{i=1}^{n}\varepsilon_i\right) + o_p(1) \\ &= n^{-1/2}\sum_{i=1}^{n} d_i^*\varepsilon_i + o_p(1), \end{aligned} \tag{26}$$

where the second equality follows by the definition of $\tilde{\mathbf{D}}$, (24), and Assumptions 2(i) and 3(i), and the third equality follows by Assumptions 1(i, iii) and 3(iii) and Markov's inequality. Therefore,

$$\sqrt{n}(\hat{\beta} - \beta) = \frac{n^{-1/2}\sum_{i=1}^{n} d_i^*\varepsilon_i}{\sigma_d^2} + o_p(1) \tag{27}$$

by Assumption 1(iv), and so

$$\frac{\sqrt{n}(\hat{\beta} - \beta)}{\sigma_\varepsilon/\sigma_d} = \frac{n^{-1/2}\sum_{i=1}^{n} d_i^*\varepsilon_i}{\sigma_d\sigma_\varepsilon} + o_p(1) \tag{28}$$

by Assumptions 1(iii) and 3(iv).

We next derive the asymptotic distribution of $n^{-1/2}\sum_{i=1}^{n} d_i^*\varepsilon_i/(\sigma_d\sigma_\varepsilon)$. Let $\mathcal{F}_{i,n}$ denote the filtration generated by $(\boldsymbol{\epsilon}^\top, ((d_j)_{j\leq i})^\top)$. Then by Assumptions 1(iii,iv) and 3(iii, iv), $n^{-1/2}d_i^*\varepsilon_i/(\sigma_d\sigma_\varepsilon)$ has finite second moment and

$$\mathbb{E}\left[\frac{n^{-1/2}d_i^*\varepsilon_i}{\sigma_d\sigma_\varepsilon}|\mathcal{F}_{i-1,n}\right] = \frac{n^{-1/2}\varepsilon_i}{\sigma_d\sigma_\varepsilon}\mathbb{E}\left[d_i^*|\,\mathcal{F}_{i-1,n}\right] = \frac{n^{-1/2}\varepsilon_i}{\sigma_d\sigma_\varepsilon}\mathbb{E}\left[d_i^*\right] = 0 \tag{29}$$

almost surely, which implies that $n^{-1/2}d_i^*\varepsilon_i/(\sigma_d\sigma_\varepsilon)$ is a martingale difference array with respect to $\mathcal{F}_{i,n}$. Next, observe that Assumptions 1(i,v) and 3(ii,iv) yield

$$\sum_{i=1}^{n}\mathbb{E}\left[\left(\frac{n^{-1/2}d_i^*\varepsilon_i}{\sigma_d\sigma_\varepsilon}\right)^2|\mathcal{F}_{i-1,n}\right] = \sigma_\varepsilon^{-2}n^{-1}\sum_{i=1}^{n}\varepsilon_i^2 \to_p 1. \tag{30}$$

Moreover for any $\eta > 0$, Assumptions 1(iii,iv,v) and 3(iii,iv) allow us to conclude that for $\delta \equiv \min(\delta_1, \delta_2)$,

$$\begin{aligned}
&\sum_{i=1}^{n} \mathbb{E}\left[\left(\frac{n^{-1/2} d_i^* \varepsilon_i}{\sigma_d \sigma_\varepsilon}\right)^2 1\left\{\left|\frac{n^{-1/2} d_i^* \varepsilon_i}{\sigma_d \sigma_\varepsilon}\right| > \eta\right\}\middle| \mathcal{F}_{i-1,n}\right] \\
&\le \frac{1}{\eta^\delta} \sum_{i=1}^{n} \mathbb{E}\left[\left|\frac{n^{-1/2} d_i^* \varepsilon_i}{\sigma_d \sigma_\varepsilon}\right|^{2+\delta}\middle| \mathcal{F}_{i-1,n}\right] \\
&= \frac{1}{\eta^\delta (\sigma_d \sigma_\varepsilon)^{2+\delta} n^{1+\delta/2}} \sum_{i=1}^{n} \mathbb{E}\left[|d_i^*|^{2+\delta}\right] |\varepsilon_i|^{2+\delta} \le \frac{K}{\eta^\delta n^{\delta/2}} n^{-1} \sum_{i=1}^{n} |\varepsilon_i|^{2+\delta} = o_p(1). \qquad (31)
\end{aligned}$$

In view of (30) and (31), we can invoke the martingale central limit theorem (see, e.g., Corollary 3.1 in Hall and Heyde (1980)) to conclude that

$$\frac{n^{-1/2} \sum_{i=1}^{n} d_i^* \varepsilon_i}{\sigma_d \sigma_\varepsilon} \to_d N(0,1). \qquad (32)$$

The claim of the theorem follows from combining this result with (28). *Q.E.D.*

Proof of Corollary 1. We first proof that $\lambda_{\min}(n^{-1}\mathbf{X}^\top \mathbf{X}) \ge K^{-1} + o_p(1)$. To do so, observe that

$$n^{-1}\mathbf{D}^\top \mathbf{D} = \mu_d^2 + \sigma_d^2 + o_p(1)$$

by Assumptions 1(i,iii) and the law of large numbers. Also, denoting $d_i^* = d_i - \mu_d$ for all $1 \le i \le n$, we have

$$n^{-1}\mathbf{D}^\top \mathbf{W} = n^{-1}\sum_{i=1}^{n} d_i w_i = \mu_d n^{-1}\sum_{i=1}^{n} w_i + n^{-1}\sum_{i=1}^{n} d_i^* w_i = \mu_d n^{-1}\sum_{i=1}^{n} w_i + o_p(1),$$

where the last equality follows from (22) in the proof of Theorem 1. Hence,

$$n^{-1}\mathbf{X}^\top \mathbf{X} = n^{-1}(\mathbf{D}, \mathbf{W})^\top (\mathbf{D}, \mathbf{W}) = n^{-1}\sum_{i=1}^{n} (\mu_d, w_i^\top)^\top (\mu_d, w_i^\top) + (\sigma_d, \mathbf{0}_{d_w \times 1}^\top)^\top (\sigma_d, \mathbf{0}_{d_w \times 1}^\top) + o_p(1),$$

where d_w is the dimension of the vectors w_i. Now, denote

$$R_n \equiv \min\left(\frac{\sqrt{3\lambda_{\min}(n^{-1}\sum_{i=1}^{n} w_i w_i^\top)}}{8|\mu_d|}, \frac{1}{2}\right)$$

and fix any $a_1 \in \mathbb{R}$ and $a_2 \in \mathbb{R}^{d_w}$ such that $a_1^2 + \|a_2\|^2 = 1$. If $|a_1| > R_n$, then

$$(a_1, a_2^\top)\left(n^{-1}\sum_{i=1}^{n} (\mu_d, w_i^\top)^\top (\mu_d, w_i^\top) + (\sigma_d, \mathbf{0}_{d_w \times 1}^\top)^\top (\sigma_d, \mathbf{0}_{d_w \times 1}^\top)\right)(a_1, a_2^\top)^\top \ge a_1^2 \sigma_d^2 \ge R_n^2 \sigma_d^2 \ge K^{-1}$$

by Assumptions 1(iii,iv) and 2(i). If, on the other hand, $|a_1| \leq R_n$, then $\|a_2\|^2 = 1 - a_1^2 \geq 1 - 1/4 = 3/4$, and so

$$|a_1| \leq \frac{\|a_2\|\sqrt{\lambda_{\min}(n^{-1}\sum_{i=1}^n w_i w_i^\top)}}{4|\mu_d|}.$$

The latter in turn implies via Jensen's inequality that

$$\begin{aligned}
&(a_1, a_2^\top)\left(n^{-1}\sum_{i=1}^n (\mu_d, w_i^\top)^\top(\mu_d, w_i^\top) + (\sigma_d, \mathbf{0}_{d_w\times 1}^\top)^\top(\sigma_d, \mathbf{0}_{d_w\times 1}^\top)\right)(a_1, a_2^\top)^\top \\
&\geq n^{-1}\sum_{i=1}^n (a_1\mu_d + a_2^\top w_i)^2 \geq n^{-1}\sum_{i=1}^n (a_2^\top w_i)^2 - 2|a_1\mu_d| n^{-1}\sum_{i=1}^n |a_2^\top w_i| \\
&\geq n^{-1}\sum_{i=1}^n (a_2^\top w_i)^2 - 2|a_1\mu_d|\sqrt{n^{-1}\sum_{i=1}^n (a_2^\top w_i)^2} \\
&= \sqrt{n^{-1}\sum_{i=1}^n (a_2^\top w_i)^2}\left(\sqrt{n^{-1}\sum_{i=1}^n (a_2^\top w_i)^2} - 2|a_1\mu_d|\right) \\
&\geq \|a_2\|\sqrt{\lambda_{\min}\left(n^{-1}\sum_{i=1}^n w_i w_i^\top\right)}\left(\|a_2\|\sqrt{\lambda_{\min}\left(n^{-1}\sum_{i=1}^n w_i w_i^\top\right)} - 2|a_1\mu_d|\right) \\
&\geq 2^{-1}\|a_2\|^2\lambda_{\min}\left(n^{-1}\sum_{i=1}^n w_i w_i^\top\right) \geq (3/8)\lambda_{\min}\left(n^{-1}\sum_{i=1}^n w_i w_i^\top\right) \geq K^{-1}
\end{aligned}$$

by Assumption 2(i). Hence, it follows that

$$\lambda_{\min}\left(n^{-1}\sum_{i=1}^n (\mu_d, w_i^\top)^\top(\mu_d, w_i^\top) + (\sigma_d, \mathbf{0}_{d_w\times 1}^\top)^\top(\sigma_d, \mathbf{0}_{d_w\times 1}^\top)\right) \geq K^{-1},$$

and so

$$\lambda_{\min}(n^{-1}\mathbf{X}^\top\mathbf{X}) \geq K^{-1} + o_p(1), \tag{33}$$

as required.

Next, we prove that s^2 is consistent for σ_ε^2. To this end, note that by Assumptions 1(i, iii, v) and 3(iii),

$$\mathbb{E}\left[\left(n^{-1}\sum_{i=1}^n d_i^*\varepsilon_i\right)^2\right] = n^{-2}\sum_{i=1}^n \mathbb{E}\left[d_i^{*2}\varepsilon_i^2\right] \leq Kn^{-1}. \tag{34}$$

Hence, by Assumptions 2(iii) and 3(i) and Markov's inequality

$$n^{-1}\mathbf{D}^\top\boldsymbol{\epsilon} = n^{-1}\sum_{i=1}^n d_i^*\varepsilon_i + \mathbb{E}[d_i]n^{-1}\sum_i \varepsilon_i = o_p(1),$$

which combined with Assumption 3(i) further implies that

$$n^{-1}\mathbf{X}^\top\boldsymbol{\epsilon} = n^{-1}(\mathbf{D}, \mathbf{W})^\top\boldsymbol{\epsilon} = o_p(1). \tag{35}$$

Combining this bound with (33) gives

$$\hat{\theta} - \theta = (n^{-1}\mathbf{X}^\top \mathbf{X})^{-1}(n^{-1}\mathbf{X}^\top \boldsymbol{\epsilon}) = o_p(1). \tag{36}$$

In addition, $\mathbf{Y} - \mathbf{X}\hat{\theta} = \boldsymbol{\epsilon} - \mathbf{X}(\hat{\theta} - \theta)$. Therefore,

$$\begin{aligned} s^2 &= n^{-1}(\mathbf{Y} - \mathbf{X}\hat{\theta})^\top(\mathbf{Y} - \mathbf{X}\hat{\theta}) = n^{-1}(\boldsymbol{\epsilon} - \mathbf{X}(\hat{\theta} - \theta))^\top(\boldsymbol{\epsilon} - \mathbf{X}(\hat{\theta} - \theta)) \\ &= n^{-1}\boldsymbol{\epsilon}^\top\boldsymbol{\epsilon} - 2n^{-1}(\hat{\theta} - \theta)^\top \mathbf{X}^\top \boldsymbol{\epsilon} + n^{-1}(\hat{\theta} - \theta)^\top \mathbf{X}^\top \mathbf{X}(\hat{\theta} - \theta) \\ &= \sigma_\varepsilon^2 + o_p(1), \end{aligned} \tag{37}$$

where the last equality follows from (35) and (36) and Assumptions 1(iii), 2(ii) and 3(ii). This finishes the proof of consistency of s^2.

Now, defining $\tilde{\mathbf{D}} \equiv \mathbf{D} - \mathbf{1}_{n\times 1}(n^{-1}\mathbf{D}^\top \mathbf{1}_{n\times 1})$ to match the proof of Theorem 1 and recalling that $\breve{\mathbf{D}} = \mathbf{M}_W \mathbf{D}$, we have

$$\begin{aligned} n^{-1}\breve{\mathbf{D}}^\top \breve{\mathbf{D}} &= n^{-1}\tilde{\mathbf{D}}^\top \mathbf{M}_W \tilde{\mathbf{D}} = n^{-1}\tilde{\mathbf{D}}^\top \tilde{\mathbf{D}} - n^{-1}(\tilde{\mathbf{D}}^\top \mathbf{W})(\mathbf{W}^\top \mathbf{W})^{-1}(\mathbf{W}^\top \tilde{\mathbf{D}}) \\ &= n^{-1}\tilde{\mathbf{D}}^\top \tilde{\mathbf{D}} + o_p(1) = \sigma_d^2 + o_p(1), \end{aligned}$$

where the third equality follows from (24) and Assumption 2(i), and the fourth from (20). Together with (37), this gives the first convergence result in (12) since $\sigma_d^2 \geq K^{-1}$ by Assumption 1(iii). The second convergence result follows from the first one, Theorem 1, and Slutsky's theorem. *Q.E.D.*

Proof of Theorem 2. As in the proof of Theorem 1, we have

$$\sqrt{n}(\hat{\beta} - \beta) = \frac{n^{-1/2}\sum_{i=1}^n d_i^* \varepsilon_i}{\sigma_d^2} + o_p(1); \tag{38}$$

see Equation (27) there and note that the derivation of (27) did not rely on Assumption 1(v), which we are not imposing here. Since $\sigma_{d\varepsilon}^2 \geq \sigma_{e,1}^2 \geq K^{-1}$ by Assumption 4(iii) and $\sigma_d^2 \leq K$ by Assumption 1(iii), (38) implies that

$$\frac{\sqrt{n}(\hat{\beta} - \beta)}{\sigma_{d\varepsilon}/\sigma_d^2} = \frac{n^{-1/2}\sum_{i=1}^n d_i^* \varepsilon_i}{\sigma_{d\varepsilon}} + o_p(1),$$

which yields the equality in (14).

We next derive the convergence result in (14), i.e. we show that $\sigma_{d\varepsilon}^{-1} n^{-1/2}\sum_{i=1}^n d_i^* \varepsilon_i \to_d N(0,1)$. To do so, we write

$$n^{-1/2}\sum_{i=1}^n d_i^* \varepsilon_i = n^{-1/2}\sum_{i=1}^n A_i^\top e_i = n^{-1/2}\sum_{i=1}^n (A_i - \mu_A)^\top e_i + n^{-1/2}\sum_{i=1}^n \mu_A^\top e_i$$

and denote the first and the second terms on the right-hand side by M_1 and M_2, respectively. Also, we denote $\tilde{A}_i \equiv A_i - \mu_A$ for all $1 \leq i \leq n$. In addition, denote $\hat{\sigma}_{e,1}^2 \equiv n^{-1}\sum_{i=1}^n \mathbb{E}[(\tilde{A}_i^\top e_i)^2 \mid \mathcal{F}_e]$, where $\mathcal{F}_e$ is the filtration generated by $(e_i^\top)_{i\leq n}$. Then

$$\hat{\sigma}_{e,1}^2 = n^{-1}\sum_{i=1}^n e_i^\top \mathbb{E}[\tilde{A}_i \tilde{A}_i^\top] e_i = n^{-1}\sum_{i=1}^n \operatorname{tr}(\mathbb{E}[\tilde{A}_i \tilde{A}_i^\top] e_i e_i^\top) = \operatorname{tr}\left(n^{-1}\sum_{i=1}^n \mathbb{E}[\tilde{A}_i \tilde{A}_i^\top] e_i e_i^\top\right)$$

$$= \operatorname{tr}\left(n^{-1}\sum_{i=1}^{n}\mathbb{E}[\tilde{A}_i\tilde{A}_i^\top]\mathbb{E}[e_ie_i^\top]\right) + o_p(1) = n^{-1}\sum_{i=1}^{n}\mathbb{E}[e_i^\top\tilde{A}_i\tilde{A}_i^\top e_i] + o_p(1) = \sigma_{e,1}^2 + o_p(1), \quad (39)$$

where the first equality follows from Assumption 4(i), the second and the third from properties of the trace operator $\operatorname{tr}(\cdot)$, the fourth from Assumptions 1(i,iii), 4(ii), and 5(i), the fifth from Assumption 4(i) and properties of the trace operator $\operatorname{tr}(\cdot)$, and the sixth from the definition of $\sigma_{e,1}^2$.

Next, let (Z_1, Z_2) be a pair of independent standard normal random variables that is independent of everything else. Then for $\delta = \min(\delta_1, \delta_3)$,

$$\sup_{t\in\mathbb{R}}\left|\mathbb{P}\left(\hat{\sigma}_{e,1}^{-1}n^{-1/2}\sum_{i=1}^{n}\tilde{A}_i^\top e_i \leq t \mid \mathcal{F}_e\right) - P(Z_1 \leq t)\right|$$
$$\leq \frac{n^{-1-\delta/2}\sum_{i=1}^{n}\mathbb{E}[|\tilde{A}_i^\top e_i|^{2+\delta} \mid \mathcal{F}_e]}{\hat{\sigma}_{e,1}^{2+\delta}} \leq \frac{Kn^{-1-\delta/2}\sum_{i=1}^{n}\|e_i\|^{2+\delta}}{\hat{\sigma}_{e,1}^{2+\delta}} = o_p(1), \quad (40)$$

where the first inequality follows from a version of the Berry-Esseen theorem (see Section 35.1.9 in DasGupta (2008)), the second inequality from the Cauchy-Schwarz inequality and Assumptions 1(iii) and 4(ii), and the last bound from (39) and Assumptions 4(iii) and 5(ii). Therefore, for any $t \in \mathbb{R}$,

$$\begin{aligned}
&\mathbb{P}(\sigma_{d\varepsilon}^{-1}(M_1 + M_2) \leq t)\\
&\quad= \mathbb{E}[\mathbb{P}(\hat{\sigma}_{e,1}^{-1}M_1 \leq \hat{\sigma}_{e,1}^{-1}(\sigma_{d\varepsilon}t - M_2) \mid \mathcal{F}_e)] = \mathbb{E}[\mathbb{P}(Z_1 \leq \hat{\sigma}_{e,1}^{-1}(\sigma_{d\varepsilon}t - M_2) \mid \mathcal{F}_e)] + o(1)\\
&\quad= \mathbb{P}(\hat{\sigma}_{e,1}Z_1 + M_2 \leq \sigma_{d\varepsilon}t) + o(1) = \mathbb{P}(\sigma_{e,1}Z_1 + M_2 \leq \sigma_{d\varepsilon}t) + o(1)\\
&\quad= \mathbb{E}[\mathbb{P}(\sigma_{e,2}^{-1}M_2 \leq \sigma_{e,2}^{-1}(\sigma_{d\varepsilon}t - \sigma_{e,1}Z_1) \mid Z_1)] + o(1) = \mathbb{E}[\mathbb{P}(Z_2 \leq \sigma_{e,2}^{-1}(\sigma_{d\varepsilon}t - \sigma_{e,1}Z_1) \mid Z_1)] + o(1)\\
&\quad= \mathbb{P}(\sigma_{d\varepsilon}^{-1}(\sigma_{e,1}Z_1 + \sigma_{e,2}Z_2) \leq t) + o(1) = \mathbb{P}(Z_1 \leq t) + o(1), \qquad (41)
\end{aligned}$$

where the first equality follows from the law of iterated expectations (LIE), the second from (40) by noting that the difference of two probabilities is always a number between zero and one to conclude that $o_p(1)$ in (40) satisfies $\mathbb{E}[o_p(1)] = o(1)$, the third from the LIE, the fourth from (39) and Assumption 4(iii), the fifth from the LIE, the sixth from Assumption 5(iii), the seventh from the LIE, and the eighth from noting that $\sigma_{e,1}Z_1 + \sigma_{e,2}Z_2$ is a normal random variable with mean zero and variance $\sigma_{d\varepsilon}^2 = \sigma_{e,1}^2 + \sigma_{e,2}^2$. This gives $\sigma_{d\varepsilon}^{-1}n^{-1/2}\sum_{i=1}^{n}d_i^*\varepsilon_i \to_d N(0,1)$ and completes the proof of the theorem. *Q.E.D.*

Proof of Theorem 3. For this proof, it will be convenient to denote $\mathbf{W} \equiv ((w_{i,j}^\top)_{i\leq n_j})_{j\leq n_g}$, $\mathbf{Y} \equiv ((y_{i,j})_{i\leq n_j})_{j\leq n_g}$, and $\boldsymbol{\epsilon} \equiv ((\varepsilon_{i,j})_{i\leq n_j})_{j\leq n_g}$. In addition, denote $\mathbf{D} \equiv (d_j\mathbf{1}_{n_j\times 1})_{j\leq n_g}$ and $\tilde{\mathbf{D}} \equiv \mathbf{D} - (n^{-1}\mathbf{D}^\top\mathbf{1}_{n\times 1})$. Moreover, denote $\bar{d}_n \equiv n^{-1}\sum_{j=1}^{n_g}n_jd_j$.

As in the proof of Theorem 1, given that the matrix $\mathbf{W}$ includes a non-zero constant column by Assumption 2(iii), it follows that $\mathbf{1}_{n\times 1}^\top\mathbf{M}_W = 0$. Therefore, by the Frisch-Waugh-Lovell theorem,

$$\hat{\beta} - \beta = (\mathbf{D}^\top\mathbf{M}_W\mathbf{D})^{-1}(\mathbf{D}^\top\mathbf{M}_W\boldsymbol{\epsilon}) = (\tilde{\mathbf{D}}^\top\mathbf{M}_W\tilde{\mathbf{D}})^{-1}(\tilde{\mathbf{D}}^\top\mathbf{M}_W\boldsymbol{\epsilon}). \quad (42)$$

We first consider the denominator $\tilde{\mathbf{D}}^\top \mathbf{M}_W \tilde{\mathbf{D}}$. Since $n = \sum_{j \le n_g} n_j$, under Assumptions 6(i,iii) we have

$$\mathbb{E}\left[|\bar{d}_n - \mu_d|^2\right] = n^{-2} \sum_{j=1}^{n_g} n_j^2 \mathbb{E}\left[(d_j^*)^2\right] \le K n^{-2} \sum_{j=1}^{n_g} n_j^2 = K\kappa_n^2/n^2. \tag{43}$$

We therefore obtain from Markov's inequality that

$$\bar{d}_n - \mu_d = O_p(\kappa_n/n), \tag{44}$$

and so

$$\bar{d}_n^2 - \mu_d^2 = (\bar{d}_n - \mu_d)(\bar{d}_n + \mu_d) = O_p(\kappa_n/n). \tag{45}$$

by Assumption 6(iii). In addition, again under Assumptions 6(i,iii) we have

$$\mathbb{E}\left[\left|n^{-1} \sum_{j=1}^{n_g} n_j (d_j^2 - \mathbb{E}[d_j^2])\right|^2\right] = n^{-2} \sum_{j=1}^{n_g} n_j^2 \mathbb{E}\left[|d_j^2 - \mathbb{E}[d_j^2]|^2\right] \le K n^{-2} \sum_{j=1}^{n_g} n_j^2 = K\kappa_n^2/n^2,$$

and so

$$n^{-1} \sum_{j=1}^{n_g} n_j (d_j^2 - \mathbb{E}[d_j^2]) = O_p(\kappa_n/n) \tag{46}$$

by Markov's inequality. Combining results (45) and (46) then yields

$$n^{-1} \tilde{\mathbf{D}}^\top \tilde{\mathbf{D}} = n^{-1} \sum_{j=1}^{n_g} n_j (d_j - \bar{d}_n)^2 = n^{-1} \sum_{j=1}^{n_g} n_j d_j^2 - (\bar{d}_n)^2 = \sigma_d^2 + O_p(\kappa_n/n). \tag{47}$$

Further, since d_j does not depend on i, we have

$$\begin{aligned} n^{-1} \tilde{\mathbf{D}}^\top \mathbf{W} &= n^{-1} \sum_{j=1}^{n_g} d_j^* \sum_{i=1}^{n_j} w_{i,j} - \left(n^{-1} \sum_{j=1}^{n_g} n_j d_j^*\right)\left(n^{-1} \sum_{j=1}^{n_g} \sum_{i=1}^{n_j} w_{i,j}\right) \\ &= n^{-1} \sum_{j=1}^{n_g} d_j^* \sum_{i=1}^{n_j} w_{i,j} + O_p(\kappa_n/n), \end{aligned} \tag{48}$$

where the second equality is by Assumption 2(ii) and (44). Moreover, by Assumptions 2(ii) and 6(i, ii, iii),

$$\mathbb{E}\left[\left\|n^{-1} \sum_{j=1}^{n_g} d_j^* \sum_{i=1}^{n_j} w_{i,j}\right\|^2\right] = \sigma_d^2 n^{-2} \sum_{j=1}^{n_g} \mathbb{E}\left[\left\|\sum_{i=1}^{n_j} w_{i,j}\right\|^2\right] \le K n^{-2} \sum_{j=1}^{n_g} n_j^2 = K\kappa_n^2/n^2.$$

Therefore, by Markov's inequality we obtain

$$n^{-1} \sum_{j=1}^{n_g} d_j^* \sum_{i=1}^{n_j} w_{i,j} = O_p(\kappa_n/n),$$

which together with (48) further shows that

$$n^{-1}\tilde{\mathbf{D}}^\top \mathbf{W} = O_p(\kappa_n/n). \tag{49}$$

Collecting the results (47), and (49) and using Assumptions 2(i) and 7(iv), we get

$$n^{-1}\tilde{\mathbf{D}}^\top \mathbf{M}_W \tilde{\mathbf{D}} = n^{-1}\tilde{\mathbf{D}}^\top \tilde{\mathbf{D}} - n^{-1}\tilde{\mathbf{D}}^\top \mathbf{W}(n^{-1}\mathbf{W}^\top \mathbf{W})^{-1} n^{-1}\mathbf{W}^\top \tilde{\mathbf{D}} = \sigma_d^2 + o_p(1). \tag{50}$$

Next, we consider the numerator $\tilde{\mathbf{D}}^\top \mathbf{M}_W \boldsymbol{\epsilon}$ in (42). By (44) and Assumptions 7(i) and 2(iii),

$$\frac{\overline{d}_n - \mu_d}{S_\varepsilon} n^{-1/2} \sum_{j=1}^{n_g} \sum_{i=1}^{n_j} \varepsilon_{i,j} = \frac{O_p(\kappa_n/n)}{S_\varepsilon} n^{-1/2} \sum_{j=1}^{n_g} \sum_{i=1}^{n_j} \varepsilon_{i,j} = o_p(1).$$

Therefore,

$$\begin{aligned}
\frac{n^{-1/2}\tilde{\mathbf{D}}^\top \boldsymbol{\epsilon}}{S_\varepsilon} &= \frac{n^{-1/2} \sum_{j=1}^{n_g} d_j^* \sum_{i=1}^{n_j} \varepsilon_{i,j}}{S_\varepsilon} - \frac{(\overline{d}_n - \mu_d)}{S_\varepsilon} n^{-1/2} \sum_{j=1}^{n_g} \sum_{i=1}^{n_j} \varepsilon_{i,j} \\
&= \frac{n^{-1/2} \sum_{j=1}^{n_g} d_j^* \sum_{i=1}^{n_j} \varepsilon_{i,j}}{S_\varepsilon} + o_p(1). \qquad (51)
\end{aligned}$$

In addition,

$$\frac{n^{-3/2}\kappa_n \mathbf{W}^\top \boldsymbol{\epsilon}}{S_\varepsilon} = o_p(1)$$

by Assumption 7(i), which together with (49) and (51) and Assumption 2(i) shows

$$\begin{aligned}
\frac{n^{-1/2}\tilde{\mathbf{D}}^\top \mathbf{M}_W \boldsymbol{\epsilon}}{S_\varepsilon} &= \frac{n^{-1/2}\tilde{\mathbf{D}}^\top \boldsymbol{\epsilon}}{S_\varepsilon} - \frac{n^{-1/2}\tilde{\mathbf{D}}^\top \mathbf{W}(\mathbf{W}^\top \mathbf{W})^{-1}\mathbf{W}^\top \boldsymbol{\epsilon}}{S_\varepsilon} \\
&= \frac{n^{-1/2} \sum_{j=1}^{n_g} d_j^* \sum_{i=1}^{n_j} \varepsilon_{i,j}}{S_\varepsilon} + o_p(1). \qquad (52)
\end{aligned}$$

Given that $S_\varepsilon^{-1} n^{-1/2} \sum_{j=1}^{n_g} (d_j - \mu_d) \sum_{i=1}^{n_j} \varepsilon_{i,j} = O_p(1)$ by Assumptions 6(i,iii) and 7(ii) and Markov's inequality and that $\sigma_d^2 \geq K^{-1}$ by Assumption 6(iv), we obtain from (42), (50), and (52) that

$$\frac{\sqrt{n}(\hat{\beta} - \beta)}{S_\varepsilon} = \frac{n^{-1/2} \sum_{j=1}^{n_g} d_j^* \sum_{i=1}^{n_j} \varepsilon_{i,j}}{\sigma_d^2 S_\varepsilon} + o_p(1)$$

and so

$$\frac{\sqrt{n}(\hat{\beta} - \beta)}{S_\varepsilon/\sigma_d} = \frac{n^{-1/2} \sum_{j=1}^{n_g} d_j^* \sum_{i=1}^{n_j} \varepsilon_{i,j}}{\sigma_d S_\varepsilon} + o_p(1) \tag{53}$$

by Assumption 6(iii).

We next derive the asymptotic distribution of $n^{-1/2} \sum_{j=1}^{n_g} d_j^* \sum_{i=1}^{n_j} \varepsilon_{i,j}/(\sigma_d S_\varepsilon)$. Let $\mathcal{F}_{j,n}$ denote the filtration generated by $(\boldsymbol{\epsilon}^\top, ((d_m)_{m\leq j})^\top)$. Then by Assumptions 6(i,v),

$$\mathbb{E}\left[\frac{n^{-1/2} d_j^* \sum_{i=1}^{n_j} \varepsilon_{i,j}}{\sigma_d S_\varepsilon} | \mathcal{F}_{j-1,n}\right] = \frac{n^{-1/2} \sum_{i=1}^{n_j} \varepsilon_{i,j}}{\sigma_d S_\varepsilon} \mathbb{E}\left[d_j^* \middle| \mathcal{F}_{j-1,n}\right] = \frac{n^{-1/2} \sum_{i=1}^{n_j} \varepsilon_{i,j}}{\sigma_d S_\varepsilon} \mathbb{E}\left[d_j^*\right] = 0$$

almost surely, which implies that $n^{-1/2} d_j^* \sum_{i=1}^{n_j} \varepsilon_{i,j} / (\sigma_d S_\varepsilon)$ is a martingale difference array with respect to $\mathcal{F}_{j,n}$. Next, observe that Assumptions 6(i,v) and 7(ii) yield

$$\sum_{j=1}^{n_g} \mathbb{E}\left[\left(\frac{n^{-1/2} d_j^* \sum_{i=1}^{n_j} \varepsilon_{i,j}}{\sigma_d S_\varepsilon}\right)^2 |\mathcal{F}_{j-1,n}\right] = S_\varepsilon^{-2} n^{-1} \sum_{j=1}^{n_g} \left(\sum_{i=1}^{n_j} \varepsilon_{i,j}\right)^2 \to_p 1. \tag{54}$$

Moreover for any $\eta > 0$, Assumptions 6(iii,iv,v) and 7(iii) allow us to conclude that

$$\begin{aligned}
&\sum_{j=1}^{n_g} \mathbb{E}\left[\left(\frac{n^{-1/2} d_j^* \sum_{i=1}^{n_j} \varepsilon_{i,j}}{\sigma_d S_\varepsilon}\right)^2 1\left\{\left|\frac{n^{-1/2} d_j^* \sum_{i=1}^{n_j} \varepsilon_{i,j}}{\sigma_d S_\varepsilon}\right| > \eta\right\}\middle| \mathcal{F}_{j-1,n}\right] \\
&\leq \frac{1}{\eta^{\delta_4}} \sum_{j=1}^{n_g} \mathbb{E}\left[\left|\frac{n^{-1/2} d_j^* \sum_{i=1}^{n_j} \varepsilon_{i,j}}{\sigma_d S_\varepsilon}\right|^{2+\delta_4} \middle| \mathcal{F}_{j-1,n}\right] \\
&= \frac{1}{\eta^{\delta_4} (\sigma_d S_\varepsilon)^{2+\delta_4} n^{1+\delta_4/2}} \sum_{j=1}^{n_g} \mathbb{E}\left[|d_i^*|^{2+\delta_4}\right] \left|\sum_{i=1}^{n_j} \varepsilon_{i,j}\right|^{2+\delta_4} \leq \frac{K}{\eta^{\delta_4} S_\varepsilon^{2+\delta_4} n^{1+\delta_4/2}} \sum_{j=1}^{n_g} \left|\sum_{i=1}^{n_j} \varepsilon_i\right|^{2+\delta_4} = o_p(1).
\end{aligned}$$

Combining this bound with (54), we can invoke the martingale central limit theorem (see, e.g., Corollary 3.1 in Hall and Heyde (1980)) to conclude that

$$\frac{n^{-1/2} \sum_{j=1}^{n_g} d_j^* \sum_{i=1}^{n_j} \varepsilon_{i,j}}{\sigma_d S_\varepsilon} \to_d N(0,1). \tag{55}$$

The claim of the theorem follows from combining this result with (53). *Q.E.D.*

Proof of Theorem 4. As in the proof of Theorem 3, we have

$$\frac{\sqrt{n}(\hat{\beta} - \beta)}{S_\varepsilon / \sigma_d} = \frac{n^{-1/2} \sum_{j=1}^{n_g} d_j^* \sum_{i=1}^{n_j} \varepsilon_{i,j}}{\sigma_d S_\varepsilon} + o_p(1); \tag{56}$$

see Equation (53) there and note that the derivation of (53) did not rely on Assumption 6(v), which we are not imposing here. Since $S_{d\varepsilon} \geq S_{e,1}$ and $S_{e,1} \geq S_\varepsilon K^{-1}$ by Assumption 8(iii) and $\sigma_d \leq K$ by Assumption 6(iii), (56) implies that

$$\frac{\sqrt{n}(\hat{\beta} - \beta)}{S_{d\varepsilon} / \sigma_d^2} = \frac{n^{-1/2} \sum_{j=1}^{n_g} d_j^* \sum_{i=1}^{n_j} \varepsilon_{i,j}}{S_{d\varepsilon}} + o_p(1),$$

which yields the equality in (18).

We next derive the convergence result in (18), i.e. we show that $S_{d\varepsilon}^{-1} n^{-1/2} \sum_{j=1}^{n_g} d_j^* \sum_{i=1}^{n_j} \varepsilon_{i,j} \to_d N(0,1)$. To do so, we write

$$n^{-1/2} \sum_{j=1}^{n_g} d_j^* \sum_{i=1}^{n_j} \varepsilon_{i,j} = n^{-1/2} \sum_{j=1}^{n_g} A_j^\top \sum_{i=1}^{n_j} e_{i,j} = \sum_{j=1}^{n_g} (A_j - \mu_A)^\top \sum_{i=1}^{n_j} e_{i,j} + n^{-1/2} \sum_{j=1}^{n_g} \mu_A^\top \sum_{i=1}^{n_j} e_{i,j}$$

and denote the first and the second terms on the right-hand side by M_1 and M_2, respectively. Also, we denote $\tilde{A}_j \equiv A_j - \mu_A$ for all $1 \leq j \leq n_g$. In addition, denote $\hat{\S}^2_{e,1} \equiv n^{-1} \sum_{i=1}^n \mathbb{E}[(\tilde{A}_j^\top e_{i,j})^2 \mid \mathcal{F}_e]$, where $\mathcal{F}_e$ is the filtration generated by $((e_{i,j}^\top)_{i \leq n_j})_{j \leq n_g}$.

Then $\hat{S}^2_{e,1} - S^2_{e,1} = o_p(S^2_{e,1})$ by the same argument as that to used to derive (39) in the proof of Theorem 2 with Assumption 9(i) replacing Assumption 5. In addition, letting Z be a standard normal random variable that is independent of everything else, we have

$$\sup_{t \in \mathbb{R}} \left| \mathbb{P}\left(\hat{\S}^{-1}_{e,1} n^{-1/2} \sum_{j=1}^{n_g} \tilde{A}_j^\top \sum_{i=1}^{n_j} e_{i,j} \leq t \mid \mathcal{F}_e \right) - P(Z \leq t) \right| = o_p(1)$$

by the same argument as that used to derive (40) in the proof of Theorem 2 with $\delta = 2$ and Assumption 9(ii) replacing Assumption 5(ii). Finally,

$$\mathbb{P}(S^{-1}_{d\varepsilon}(M_1 + M_2) \leq t) = \mathbb{P}(Z \leq t) + o(1)$$

for all $t \in \mathbb{R}$ by the same argument as that used to derive (41) in the proof of Theorem 2 with Assumption 9(iii) replacing Assumption 5(iii). This gives $S^{-1}_{d\varepsilon} n^{-1/2} \sum_{j=1}^{n_g} d_j^* \sum_{i=1}^{n_j} \varepsilon_{i,j} \to_d N(0,1)$ and completes the proof of the theorem. *Q.E.D.*

B Asymptotic equivalence of two variance formula

Lemma 1. *Consider the linear regression model in (1). Suppose: (i) $(d_i)_{i \leq n}$ are i.i.d. with zero mean, finite and nonzero variance σ_d^2; (ii) $(\varepsilon_i)_{i \leq n}$ is covariance stationary with auto-covariance function $\Gamma_\varepsilon(\cdot)$ satisfying $\Gamma_\varepsilon(0) > 0$ and $\sum_{j=1}^\infty \Gamma_\varepsilon(j)^2 < \infty$. For $\mathbf{D} \equiv (d_i)_{i \leq n}$, we then have*

$$\frac{\mathbf{D}^\top \Omega \mathbf{D}}{\Gamma_\varepsilon(0) \mathbf{D}^\top \mathbf{D}} \to 1 \textit{ almost surely as } n \to \infty, \tag{57}$$

where Ω is the covariance matrix of $(\varepsilon_i)_{i \leq n}$.

Proof of Lemma 1. Note that

$$n^{-1} \mathbf{D}^\top \Omega \mathbf{D} = n^{-1} \sum_{i_1=1}^n \sum_{i_2=1}^n d_{i_1} d_{i_2} \Gamma_\varepsilon(i_1 - i_2) = \Gamma_\varepsilon(0) n^{-1} \mathbf{D}^\top \mathbf{D} + 2n^{-1} \sum_{i=2}^n U_i \tag{58}$$

where $U_i \equiv \sum_{i'=1}^{i-1} x_i x_{i'} \Gamma_\varepsilon(i - i')$. Let $\tilde{\mathcal{F}}_i$ denote the natural filtration generated by $(d_j)_{j \leq i}$. Then, under the assumption that $(d_i)_{i \leq n}$ is i.i.d., it follows that $\{U_i, \tilde{\mathcal{F}}_i\}$ is a martingale difference sequence with variance $\mathbb{E}[U_i^2] = \sigma_d^4 \sum_{j=1}^{i-1} \Gamma_\varepsilon(j)^2$. Therefore we obtain that

$$\sum_{i=2}^n i^{-2} \mathbb{E}\left[U_i^2\right] = \sigma_d^4 \sum_{i=2}^n i^{-2} \sum_{j=1}^{i-1} \Gamma_\varepsilon(j)^2 = \sigma_d^4 \sum_{j=1}^{n-1} \Gamma_\varepsilon(j)^2 \sum_{m=j+1}^n m^{-2} \leq K \sigma_d^4 \sum_{j=1}^\infty \Gamma_\varepsilon(j)^2 < \infty \tag{59}$$

where the first inequality follows from $\sum_{m=1}^{\infty} m^{-2} < K$ and the last inequality is due to $\sigma_d^2 < \infty$ and $\sum_{j=1}^{\infty} \Gamma_\varepsilon(j)^2 < \infty$ by assumption. Hence, (59) establishes that $\sum_{i=2}^{\infty} i^{-2}\mathbb{E}\left[U_i^2\right] < \infty$. By the martingale strong law of large numbers (see, e.g., Theorem 3.76 in White (2014)) we can therefore deduce that $n^{-1}\sum_{i=2}^{n} U_i \to 0$ almost surely as $n \to \infty$, which together with (58) yields

$$n^{-1}\mathbf{D}^\top \Omega \mathbf{D} - \Gamma_\varepsilon(0) n^{-1}\mathbf{D}^\top \mathbf{D} \to 0 \text{ almost surely as } n \to \infty. \tag{60}$$

Moreover, by condition (i) of the lemma and Kolmogorov's strong law of large numbers (see, e.g., Theorem 3.1 in White (2014)), $n^{-1}\mathbf{D}^\top \mathbf{D} \to \sigma_d^2$ almost surely as $n \to \infty$. Since $\sigma_d^2 > 0$ and $\Gamma_\varepsilon(0) > 0$, the claim of the lemma then follows from (60). *Q.E.D.*