A direct route to optimal parametric weighted least squares*

Saraswata Chaudhuri†

This version: August 28, 2019

Abstract

The parametric weighted least squares (WLS) estimator and Romano and Wolf (2017)'s adaptive LS (ALS) estimator are not optimal if the parametric model for heteroskedasticity of the regression error is misspecified. We propose a modified WLS (MWLS) estimator that is designed to be optimal. This modification involves obtaining the weights in WLS by directly minimizing the asymptotic variance of interest. By construction, the asymptotic variance of the MWLS estimator cannot exceed that of the WLS, ALS, and ordinary LS estimators. Simulations under Romano and Wolf (2017)'s design demonstrate the superior performance of the MWLS estimator even in relatively small samples.

JEL Classification: C12; C13; C21.

Keywords: asymptotic optimality; misspecification; nuisance parameters; weighted least squares

*Older versions of the paper were circulated until May 29, 2019 under a longer title: "Efficient estimation when the misspecification of nuisance parameters does not affect the consistency of the estimator for the parameters of interest: An application to FGLS". I thank E. Renault, J-M. Dufour, J. Galbraith, J. MacKinnon, R. Startz, S. Goncalves and V. Zinde-Walsh for their helpful comments.

†Department of Economics, McGill University & Cireq, Montreal. Email: saraswata.chaudhuri@mcgill.ca.

1 Introduction

It has been widely accepted since the seminal work of White (1980) that using the ordinary least squares (OLS) estimator with robust standard errors in a regression model is preferable to using the classical parametric weighted LS (henceforth, simply "WLS" for brevity) estimator. This is because a parametric misspecification of the unknown conditional variance of the regression error no longer guarantees that the WLS standard errors are smaller than that of the OLS. Consequently, the use of WLS is rare in modern economics and econometrics.[1]

In a recent paper, "Resurrecting weighted least squares", Romano and Wolf (2017) challenge this view (also see Leamer (2010)). Romano and Wolf (2017) propose an adaptive LS (ALS) estimator that is the OLS estimator if a pre-test cannot reject the conditional homoskedasticity of the regression error, but is the WLS estimator otherwise. They demonstrate that the ALS estimator can provide improvements over the OLS and WLS estimators. However, the asymptotic variance of the ALS estimator can exceed that of the OLS estimator when the regression error is actually conditionally heteroskedastic and this conditional variance is parametrically misspecified. This is because the ALS estimator is not designed for any kind of asymptotic optimality in such cases.

We wish to further advance this dissenting view on the neglect of WLS by proposing a modified parametric WLS (henceforth, "MWLS") estimator that directly addresses the above mentioned issue of non-optimality. Given the correct or incorrect parametric model for the conditional variance of the regression error, this MWLS estimator is designed such that its asymptotic variance is the smallest among a class of estimators that includes the OLS and WLS estimators as special cases.

The parameter of interest in our paper is $h(\beta)$ that is taken as a smooth function of the regression coefficients β. $h(\beta)$ includes β, elements of β, their sums, differences, products, etc. as special cases.

The idea behind MWLS: Let $\omega^2(X;\gamma)$ be the user-specified correct/incorrect parametric model for the variance of the regression error conditional on the regressors X. In a sample of size n, let $\widehat{\beta}_n(\gamma)$ be the weighted by $\omega^{-1}(X;\gamma)$ least squares estimator of β for each γ. Let $\Sigma(\gamma)$ be the asymptotic variance of $h(\widehat{\beta}_n(\gamma))$ for each γ. Let γ^* be an optimal γ in the sense that $\Sigma(\gamma)-\Sigma(\gamma^*)$ is positive semi-definite for all $\gamma \neq \gamma^*$. We define the MWLS estimator as $h(\widehat{\beta}_n(\widehat{\gamma}_n))$ where $\widehat{\gamma}_n$ is an estimator of an unknown optimal γ^*. If $\widehat{\gamma}_n$ is better than $n^{1/4}$-consistent for γ^* then, under standard conditions, the asymptotic variance of the MWLS estimator is $\Sigma(\gamma^*)$, which is optimal.

Comparison with WLS: If the parametric model $\omega^2(X;\gamma)$ is correctly specified then it is well known that WLS is the efficient estimator, and we show under standard conditions that $\Sigma(\gamma^*)$,

[1] To quote Angrist and Pischke (2010): "A legacy of White's (1980) paper on robust standard errors, one of the most highly cited from the period, is the near-death of generalized least squares in cross-sectional applied work."

which is the asymptotic variance of the MWLS estimator, is also the asymptotic variance of WLS. On the other hand, if $\omega^2(X;\gamma)$ is misspecified then WLS is no longer efficient, and now there is room for possible improvement given the user-specified parametric model $\omega^2(X;\gamma)$. The MWLS estimator captures this room for improvement completely since our proposal, by construction, leads to the smallest asymptotic variance $\Sigma(\gamma^*)$ given the user-specified parametric model $\omega^2(X;\gamma)$.

Comparison with OLS: In practice, parametric models $\omega^2(X;\gamma)$ allow for conditional homoskedasticity as a special case, i.e., a constant $\omega^2(X;\bar{\gamma})$ for some value $\bar{\gamma}$ of γ; see Wooldridge (2012), Romano and Wolf (2017), etc. Then, the asymptotic variance of the OLS estimator of $h(\beta)$ is $\Sigma(\bar{\gamma})$. By construction, $\Sigma(\bar{\gamma}) - \Sigma(\gamma^*)$ is positive semi-definite, i.e., the asymptotic variance of the MWLS estimator cannot exceed that of the OLS estimator. If the regression error actually happens to be conditionally homoskedastic, then OLS is efficient and hence, by construction, $\Sigma(\bar{\gamma}) = \Sigma(\gamma^*)$.

The fundamental difference between the idea behind the MWLS estimator and that for the other/existing parametric WLS estimators is as follows. The early papers like Carroll and Ruppert (1982) or the recent papers like Romano and Wolf (2017) and Spady and Stouli (2019), all seek a "best fit" of the parametric model $\omega^2(X;\gamma)$ to the true conditional variance. By contrast, as we will discuss in detail, we do not actively seek such a "best fit" to the true conditional variance because if $\omega^2(X;\gamma)$ is misspecified then this does not necessarily lead to the smallest asymptotic variance for the estimator of the parameters of interest $h(\beta)$. This is precisely why the asymptotic variance of estimators like WLS and ALS can be larger than that of OLS when $\omega^2(X;\gamma)$ is misspecified.

Instead, we first note that under standard conditions: (i) $h(\widehat{\beta}_n(\gamma))$ is consistent for $h(\beta)$ even if $\omega^2(X;\gamma)$ is misspecified, and (ii) plugging in an estimated γ in $h(\widehat{\beta}_n(\gamma))$ does not affect its asymptotic variance. This allows us to fit the model $\omega^2(X;\gamma)$ to directly minimize the asymptotic variance of $h(\widehat{\beta}_n(\gamma))$, and define the MWLS estimator based on this fit. Consequently, if the user happens to specify the model $\omega^2(X;\gamma)$ correctly, then we obtain the same asymptotic variance as the other methods. On the other hand, if the user misspecifies the model $\omega^2(X;\gamma)$, then we are still guaranteed the smallest asymptotic variance that is possible given this misspecified model.

We wish to emphasize regarding this idea behind the MWLS estimator that the general idea of targeting nuisance parameters to achieve some form of optimality of the estimator for the parameter of interest is very old.[2] This idea should apply to any estimation framework where a

[2]This idea has been practiced at least since the early days of the optimal design of experiments and sample surveys where the nuisance parameters are, respectively, the experiment-design and the sampling-design. Some of the directly relevant early works are cited in our paper (see, e.g., Elfving (1952) and Chernoff (1953) cited above). Numerous interesting papers in economics, statistics and biostatistics have recently studied similar optimality problems in contexts such as optimal design of experiments using large dimensional data, robust estimation using coarsened data, etc. The general idea is also related to that of finding the optimal estimating equations as described in Heyde (1997).

possible parametric misspecification of the nuisance parameters does not affect the consistency of the estimator for the parameters of interest; conditions for this property can be found in Theorem 6.2 of Newey and McFadden (1994). We introduce our proposed estimator in the specific context of weighted least squares because it is the leading example of such estimation frameworks and, thanks to Romano and Wolf (2017), there is a renewed interest in this universally familiar example. Extension of this idea to various other estimation frameworks is the topic of our ongoing research.

Finally, we note that three issues arise when implementing our idea for the MWLS estimator.

First, computation of an optimal γ^* can be difficult. We bypass this issue by instead minimizing the trace of $\Sigma(\gamma)$. This minimizer indeed leads to $\Sigma(\gamma^*)$ if an optimal γ^* exists. Otherwise, it gives a compromised notion of optimality: the A-optimality of Elfving (1952) and Chernoff (1953).

Second, while an optimal γ^* exists under standard conditions if the parametric model $\omega^2(X;\gamma)$ is correctly specified, the existence of γ^* is generally not guaranteed under misspecification unless $h(\beta)$ is a scalar. However, $h(\beta)$ is a scalar in most empirical studies since interest generally lies on the regression coefficients, their sums, differences, etc. *individually*. Empirical research generally reports estimators, standard errors, t-ratios, etc., i.e., quantities that correspond to distinct but inherently scalar $h(\beta)$'s (e.g., each coefficient is a distinct $h(\beta)$ in a regression output table). In light of this observation, one can always report distinct but scalar $h(\beta)$-specific MWLS estimators, standard errors, t-ratios, etc. since the optimal γ^*'s in these distinct scalar cases necessarily exist.

Third, the optimal γ^* may not be unique. However, this case of set-identified nuisance parameters (γ) is not a problem since we show that, by the definition of the set, $\Sigma(\gamma^*)$ is the same for all such optimal γ^*. Indeed, we do not even require the estimated γ to converge in probability to any of these optimal γ^*'s. Our results hold if, simply, the distance between a sequence of estimated γ and the possibly non-singleton set of optimal γ^*'s converges in probability to zero at the $n^{1/4}$ rate.

Our presentation of the MWLS estimator takes in to account all these issues carefully. We also provide a consistent estimator for the asymptotic variance of the MWLS estimator, using which one can directly obtain asymptotically normal t-ratios for the purpose of inference.

Our proposed MWLS estimator should be especially useful in applications where the OLS standard errors are large, the residual-plots indicate conditional heteroskedasticity, but, due to the usual limitations related to nonparametric estimation, the semiparametric WLS as in Carroll (1982), Robinson (1987) and Newey (1994) that involves estimating the conditional variance of the regression errors using kernel, nearest neighbor and series estimators respectively, is not practical.

Our paper proceeds as follows. Section 2 discusses the framework, introduces our proposed MWLS estimator, and establishes its first-order asymptotic properties. Section 3 provides further

discussion of the MWLS estimator drawing a sharp contrast with other estimators such as OLS, WLS, and the recently proposed estimators in Romano and Wolf (2017) and Spady and Stouli (2019). Section 4 is a Monte Carlo experiment that finds that our asymptotic propositions from Sections 2 and 3 generally hold up well in small samples. Under the design of Romano and Wolf (2017), the simulation results demonstrate the superior performance of our proposed MWLS estimator. Section 5 concludes. Appendix A contains the proofs of the results presented in Section 2. Appendix B collects supplementary materials related to Sections 2, 3 and 4.

We use the notation $:=$ for defined as, $\equiv$ for numerical equivalence, $\propto$ for proportional to, $\|.\|$ for the Euclidean norm, $\xrightarrow{d}$ and $\xrightarrow{p}$ for convergence in distribution and probability, and $O_p(a_n)$ and $o_p(a_n)$ for quantities that are, respectively, bounded in probability and $\xrightarrow{p} 0$ when divided by a_n.

2 The modified weighted least squares (MWLS) estimator

2.1 Framework and preliminaries

Consider a linear regression model with a dependent variable y and p regressors X:

$$\begin{aligned} y &= X'\beta + u && (1)\\ \text{where } E[u|X] &= 0, && (2)\\ E[u^2|X] &= \omega_0^2(X) && (3) \end{aligned}$$

almost surely in X. $\omega_0^2(X)$ is an unknown positive function of X. The parameter of interest is $h(\beta)$ that is a $d_h \times 1$ ($d_h \leq p$) function of β. The sample of observations is $(y_i, X_i')_{i=1}^n$.

Let $\omega^2(X;\gamma)$ be a positive function of X and γ where $\gamma \in \Gamma \subset \mathbb{R}^k$. $\omega^2(X;\gamma)$ is the user's attempt to parametrically model the unknown $\omega_0^2(X)$. This model may or may not be correct.

We define the weighted by $\omega^{-1}(X;\gamma)$ least squares estimator of β as:

$$\widehat{\beta}_n(\gamma) := \left(\frac{1}{n}\sum_{i=1}^n \frac{1}{\omega^2(X_i;\gamma)} X_i X_i'\right)^{-1} \left(\frac{1}{n}\sum_{i=1}^n \frac{1}{\omega^2(X_i;\gamma)} X_i y_i\right), \quad (4)$$

and index it by $\gamma \in \Gamma$. Accordingly, we define the estimator of $h(\beta)$, also indexed by $\gamma \in \Gamma$, as:

$$\widehat{h}_n(\gamma) := h(\widehat{\beta}_n(\gamma)). \quad (5)$$

We use the following notation throughout. The eigen values of an $a \times a$ matrix A are denoted by $\lambda_1(A) \leq \lambda_2(A) \leq \ldots \leq \lambda_a(A)$. For each $\gamma \in \Gamma$, we define:

$$B(\gamma) := E\left[\frac{1}{\omega^2(X;\gamma)} XX'\right], \quad \widehat{B}_n(\gamma) := \frac{1}{n}\sum_{i=1}^n \frac{1}{\omega^2(X_i;\gamma)} X_i X_i' \text{ and } C(\gamma) := E\left[\frac{\omega_0^2(X)}{(\omega^2(X;\gamma))^2} XX'\right].$$

Assumption (A1)-(A7):

(A1) β_0 is the true value of β. Hence, $h(\beta_0)$ is the true value of $h(\beta)$.

(A2) $(y_i, X_i')_{i=1}^n$ are i.i.d. copies of the random variables (y, X').

(A3) (a) $\omega_0^2(X) > 0$ with probability one, and (b) $\omega^2(X;\gamma) > 0$ with probability one for all $\gamma \in \Gamma$ where Γ is a compact subset of $\mathbb{R}^k$.

(A4) $\sup_{\gamma\in\Gamma} \left\| \widehat{B}_n(\gamma) - B(\gamma) \right\| = o_p(1)$.

(A5) $\frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{1}{\omega^2(X_i;\gamma)} X_i u_i \xrightarrow{d} N\left(0, C(\gamma)\right)$ for each $\gamma \in \Gamma$.

(A6) $\inf_{\gamma\in\Gamma} \lambda_1\left(B(\gamma)\right) > 0$, $\inf_{\gamma\in\Gamma} \lambda_1\left(C(\gamma)\right) > 0$, $\sup_{\gamma\in\Gamma} \lambda_p\left(B(\gamma)\right) < \infty$ and $\sup_{\gamma\in\Gamma} \lambda_p\left(C(\gamma)\right) < \infty$.

(A7) $h(\beta)$ is continuously differentiable in β in an open neighborhood of β_0, and $H := H(\beta_0)$ is a finite, full row-rank $d_h \times p$ matrix where $H(\beta) := \partial h(\beta)/\partial\beta'$.

Remarks: (A1)-(A7) are standard assumptions. (A1) states that the distribution of $(y, X')'$ is any member of the family of distributions for which (1) holds with $\beta = \beta_0$, and that satisfies (2), (3), conditions (A3)-(A6), and other standard conditions (stated later). (A2) simplifies the exposition of our key idea, but can be readily modified to accommodate for a non-stochastic design matrix $(X_1', \ldots, X_n')'$ or other types of commonly encountered samples $(y_i, X_i')_{i=1}^n$. Primitive conditions for the high-level assumptions (A4) and (A5) are well known. The implication on (A6) of the positivity instead of a bounded away from zero assumption in (A3b) is discussed later. (A7) allows for the use of the delta method. Some of (A2)-(A7) can be weakened for the individual results below.

Lemma 1

(i) $\widehat{\beta}(\gamma_1) \equiv \widehat{\beta}_n(\gamma_2)$ *for any* $\gamma_1 \in \Gamma$ *and* $\gamma_2 \in \Gamma$ *such that* $\omega^2(X;\gamma_1) \propto \omega^2(X;\gamma_2)$.

(ii) If assumptions (A1)-(A6) hold then, for each $\gamma \in \Gamma$,

$$\sqrt{n}\left(\widehat{\beta}_n(\gamma) - \beta_0\right) \xrightarrow{d} N\left(0, \Xi(\gamma) := B(\gamma)^{-1} C(\gamma) B(\gamma)^{-1}\right). \tag{6}$$

(iii) If assumptions (A1)-(A7) hold then, for each $\gamma \in \Gamma$,

$$\sqrt{n}\left(\widehat{h}_n(\gamma) - h(\beta_0)\right) \equiv \sqrt{n}\left(h\left(\widehat{\beta}_n(\gamma)\right) - h(\beta_0)\right) \xrightarrow{d} N\left(0, \Sigma(\gamma) := H\Xi(\gamma)H'\right). \tag{7}$$

There are two components in the design of our proposed MWLS estimator. First, the benchmark needs to be set by defining an "optimal" value of γ that minimizes the asymptotic variance $\Sigma(\gamma)$. Second, the "optimal" feasible estimator of $h(\beta)$, i.e., the MWLS estimator, needs to be obtained by plugging in an estimate (in a sense made precise later) of the unknown optimal value of γ in $\widehat{h}_n(\gamma)$. Now, we go over these two components systematically describing the underlying non-trivial issues such as existence and non-uniqueness, and our proposed solutions that address these issues.

2.2 Optimal but possibly non-unique γ

We define an optimal γ as any $\gamma^* \in \Gamma$ that satisfies the condition that:

$$\Sigma(\gamma) - \Sigma(\gamma^*) \text{ is positive semi-definite for all } \gamma \in \Gamma \text{ where } \gamma \neq \gamma^*. \tag{8}$$

If $h(\beta)$ is a scalar then γ^* exists by the extreme value theorem provided that $\Sigma(\gamma)$ is continuous in γ (as is typically the case) and Γ is compact in $\mathbb{R}^k$ (assumed in (A3b)). However, two problems arise when $h(\beta)$ is not a scalar, i.e., $d_h > 1$, specifically in the case where $\omega^2(X;\gamma)$ is misspecified. First, it is difficult to directly work with the general definition of optimality in (8). Second, γ^* may not exist. To avoid this second problem we assume the existence of γ^*. This assumption is not needed if $\omega^2(X;\gamma)$ is correctly specified or if, as in almost all cases of empirical interest, $d_h = 1$.

Assumption (A8):

(A8) There exists a γ^* satisfying (8).

Now, the first problem, i.e., computational difficulty, can be addressed by instead defining γ^* as a minimizer of Trace$(\Sigma(\gamma))$ with respect to $\gamma \in \Gamma$, a criterion function that is easier to minimize.

However, before formally introducing this convenient criterion, we note that there is another important issue with γ^*: it is generally non-unique.[3] The criterion in (8) accounts for this non-uniqueness by only requiring positive semi-definiteness instead of positive definiteness since the former allows for $\Sigma(\gamma) = \Sigma(\gamma^*)$ for $\gamma \neq \gamma^*$ and $\gamma \in \Gamma$. To incorporate this explicitly into our analysis, we characterize the non-uniqueness by defining a possibly non-singleton set $\Gamma^* \subseteq \Gamma$ as:

$$\Gamma^* := \left\{ \gamma^* \in \Gamma \,\middle|\, \Gamma_{\text{pd}}(\gamma^*) \bigcup \Gamma_{\text{eq}}(\gamma^*) = \Gamma \right\} \tag{9}$$

$$\text{where } \Gamma_{\text{pd}}(\gamma^*) := \{\gamma \in \Gamma \mid \Sigma(\gamma) - \Sigma(\gamma^*) \text{ is positive definite}\}, \tag{10}$$

$$\Gamma_{\text{eq}}(\gamma^*) := \{\gamma \in \Gamma \mid \Sigma(\gamma) = \Sigma(\gamma^*)\}. \tag{11}$$

Γ^* is the set of optimal γ^*'s: elements of this set have equal $\Sigma(\gamma^*)$'s, while $\Sigma(\gamma) - \Sigma(\gamma^*)$ is positive definite for elements $\gamma \in \Gamma \backslash \Gamma^*$ outside. Assumption (A8) ensures that Γ^* is nonempty. If $\Gamma_{\text{eq}}(\gamma) = \Gamma$ for any $\gamma \in \Gamma$ then $\Gamma^* = \Gamma$, which is unlikely except in pathological cases. If Γ^* is a singleton $\{\gamma^*\}$ then $\Gamma_{\text{eq}}(\gamma^*) = \{\gamma^*\}$ and $\Gamma_{\text{pd}}(\gamma^*) = \Gamma \setminus \{\gamma^*\}$, which is also unlikely. We allow for all possibilities.

Lemma 2 *If assumptions (A6)-(A8) hold then* Γ^* *defined in* (9) *satisfies:*

$$\Gamma^* = \left\{ \gamma^* \in \Gamma \,\middle|\, \mathit{Trace}(\Sigma(\gamma^*)) - \min_{\gamma \in \Gamma} \mathit{Trace}(\Sigma(\gamma)) = 0 \right\}. \tag{12}$$

[3] This can arise trivially, e.g., Lemma 1(i), that can often be ruled out by careful parameterization (up to scale factor) of $\omega^2(X;\gamma)$ by the user. However, especially when $\omega^2(X;\gamma)$ is misspecified, non-uniqueness can also arise in more involved ways that depend on the underlying distribution of (y, X); and this cannot be ruled out without very strong assumptions that severely restrict the premise of our paper. Naturally, we do not make any such assumption.

Remark: The equivalence relation in Lemma 2 greatly simplifies the computation of the optimal γ^*'s irrespective of whether $h(\beta)$ is a scalar. However, when $h(\beta)$ is not a scalar, our key assumption continues to be (A8) that enabled the existence of Γ^* in (9). Without it, the set characterized by the right-hand side (RHS) of (12), that is nonempty under continuity of Trace$(\Sigma(\gamma))$ and compactness of Γ, is simply the set of the so-called A-optimal γ's in the terminology of the design of experiments literature; see Elfving (1952) and Chernoff (1953). The notion of A-optimality is not going to be attractive in empirical work unless backed by assumption (A8) that elevates it to the more general notion of optimality in (8). Moreover, without assumption (A8), the set on the RHS of (12) would generally be non-singleton with elements γ's leading to possibly different $\Sigma(\gamma)$'s. Hence, the asymptotic properties of a feasible optimal estimator of $h(\beta)$, which requires estimating the unknown optimal γ, could be invalid. Of course, these concerns are all moot when $h(\beta)$ is a scalar, i.e., in almost all common cases of empirical interest, since assumption (A8) then holds trivially.[4]

2.3 A feasible optimal estimator of $h(\beta)$

2.3.1 Estimation of an optimal γ

Having characterized the set of optimal γ's as Γ^*, which is unknown, next we turn to the feasible optimal estimator of $h(\beta)$ that would involve estimating γ^* in a sense made precise now. First, let us describe this estimation of γ^* which, as we noted before, is not necessarily unique. Define:

$$\begin{aligned} m(\gamma) &:= \text{Trace}(\Sigma(\gamma)) \quad \text{if } \Sigma(\gamma) \text{ is positive definite} \\ &:= +\infty \qquad\qquad\quad \text{if } \Sigma(\gamma) \text{ is not positive definite.} \end{aligned} \tag{13}$$

The constraint of positive definiteness is redundant under assumptions (A6)-(A7), but can be useful for the sample analog $\widehat{m}_n(\gamma)$ of $m(\gamma)$, which, with an estimator $\widehat{\Sigma}_n(\gamma)$ of $\Sigma(\gamma)$, is defined as follows:

$$\begin{aligned} \widehat{m}_n(\gamma) &:= \text{Trace}(\widehat{\Sigma}_n(\gamma)) \quad \text{if } \widehat{\Sigma}_n(\gamma) \text{ is positive definite} \\ &:= +\infty \qquad\qquad\quad \text{if } \widehat{\Sigma}_n(\gamma) \text{ is not positive definite.} \end{aligned} \tag{14}$$

Since it is generally convenient when objective functions to be minimized have a zero minimum, define, respectively, the population and sample objective functions to be minimized with respect

[4] We use Trace$(\Sigma(\gamma))$ as the criterion to minimize *only* for computational ease. Our exposition does not specifically exploit any special properties of trace, and is generally applicable to the other "compromised" notions of optimality such as the D-optimality of Wald (1943), the E-optimality of Ehrenfeld (1956), the L-optimality (generalization of A-optimality) due to Karlin and Studden (1966) and Federov (1971), Kiefer (1974)'s general optimality of which the A/L, D and E optimalities are limiting cases, etc. The empirical attractiveness of the interpretability of these notions are also dependent on (A8). Also, they may not agree unless $h(\beta)$ is a scalar or, more generally, without the existence assumption (A8). Our results below are readily applicable if the user prefers to use these other notions of optimality.

to $\gamma \in \Gamma$ as follows:

$$Q(\gamma) := m(\gamma) - \min_{g \in \Gamma} m(g) \text{ and } \widehat{Q}_n(\gamma) := \widehat{m}_n(\gamma) - \min_{g \in \Gamma} \widehat{m}_n(g). \tag{15}$$

The population objective function $Q(\gamma)$ does not entail any additional loss of generality since $m(\gamma)$ is finite for $\gamma \in \Gamma$ by assumptions (A6)-(A7). A similar finiteness and positive definiteness assumption (in (A9) below) on $\widehat{\Sigma}_n(\theta)$ would ensure the same for the sample objective function $\widehat{Q}_n(\gamma)$.

Any $\gamma^* \in \Gamma^*$ is, by definition, a minimizer of $Q(\gamma)$ with respect to $\gamma \in \Gamma$. Now, consider any sequence $\widehat{\gamma}_n \in \Gamma$ of minimizer of $\widehat{Q}_n(\gamma)$ (allowing for non-unique minimizers at each $n \geq 1$):

$$\widehat{\gamma}_n \in \Gamma \text{ is such that } \widehat{Q}_n(\widehat{\gamma}_n) = \min_{\gamma \in \Gamma} \widehat{Q}_n(\gamma). \tag{16}$$

Subsequently, define the distance measure of any $\gamma \in \Gamma$ from Γ^*, which is non-empty by (A8), as:

$$d(\gamma, \Gamma^*) = \inf_{\bar{\gamma} \in \Gamma^*} \|\gamma - \bar{\gamma}\|. \tag{17}$$

Finally, based on (17), define $\Gamma^*_\delta := \{\gamma \in \Gamma \,|d(\gamma, \Gamma^*) \leq \delta\}$ as the δ-expansion of Γ^* for a $\delta \geq 0$.

Remarks: It is worth noting the following regarding the minimizer $\widehat{\gamma}_n$ and the distance $d(\widehat{\gamma}_n, \Gamma^*)$.

1. At any $n \geq 1$, there may exist multiple elements in closure(Γ^*) in Γ that have the smallest distance from $\widehat{\gamma}_n$. However, if, e.g., there exist $\gamma_{a,n}, \gamma_{b,n} \in$ closure(Γ^*) in Γ such that $\|\widehat{\gamma}_n - \gamma_{a,n}\| = \|\widehat{\gamma}_n - \gamma_{b,n}\| = d(\widehat{\gamma}_n, \Gamma^*)$ for all n large enough, then eventually $\gamma_{a,n}$ and $\gamma_{b,n}$ also have to be arbitrarily close if $d(\widehat{\gamma}_n, \Gamma^*) = o_p(1)$. On the other hand, much differently from the standard cases with a unique "(psuedo) true value", our asymptotic results below do not require that any such sequence $\gamma_{a,n}$ or $\gamma_{b,n}$ actually converges to a point ("(psuedo) true value") in Γ^* (or Γ) as $n \to \infty$.

2. A different issue arises because, additionally, the sequence $\widehat{\gamma}_n$ is also non-unique. For two sequences $\widehat{\gamma}_{n,c}$ and $\widehat{\gamma}_{n,d}$ satisfying (16), it is possible that the sequences $\gamma_{n,c}, \gamma_{n,d} \in$ closure(Γ^*) in Γ are such that $\|\widehat{\gamma}_{n,c} - \gamma_{n,c}\| = d(\widehat{\gamma}_{n,c}, \Gamma^*)$ and $\|\widehat{\gamma}_{n,c} - \gamma_{n,d}\| = d(\widehat{\gamma}_{n,d}, \Gamma^*)$. In this case, $\gamma_{n,c}$ and $\gamma_{n,d}$ may not be eventually close even if $\|\widehat{\gamma}_{n,c} - \gamma_{n,c}\| = o_p(1)$ and $\|\widehat{\gamma}_{n,d} - \gamma_{n,d}\| = o_p(1)$.

However, these two issues do not cause much problems with our asymptotic results. Lastly, we note that the discussion below follows trivially in the pathological case $\Gamma = \Gamma^*$. We highlight this in Lemma 3 (i), but ignore it otherwise and take $\Gamma^* \subset \Gamma$. As such, some of the assumptions below (e.g., (A11)) only make sense when $\Gamma \setminus \Gamma^*$ is non-empty, which, we hope, is clear from the context.

Assumption (A9)-(A13):

(A9) $\liminf_n \inf_{\gamma \in \Gamma} \lambda_1 \left(\widehat{\Sigma}_n(\gamma)\right) > 0$ and $\limsup_n \sup_{\gamma \in \Gamma} \lambda_{d_h} \left(\widehat{\Sigma}_n(\gamma)\right) < \infty$ with probability one.

(A10) $\sup_{\gamma \in \Gamma} |\widehat{Q}_n(\gamma) - Q(\gamma)| = o_p(1)$.

(A11) For all $\delta > 0$, there exists $\epsilon(\delta) > 0$ such that $\inf_{\gamma \in \Gamma^*_\delta \setminus \Gamma^*} Q(\gamma) \geq \epsilon(\delta)$.

(A12) $\sup_{\gamma \in \Gamma^*_\delta} \sqrt{n}|\widehat{Q}_n(\gamma) - Q(\gamma)| = O_p(1)$ for some $\delta > 0$.

(A13) $\inf_{\gamma \in \Gamma^*_\delta \setminus \Gamma^*} [\{Q(\gamma) - 0\} - D \times d(\gamma, \Gamma^*)] \geq 0$ for some $\delta > 0$ and $D > 0$.

Remarks: (A9) is the sample counterpart of (A6), and ensures that $\widehat{Q}_n(\gamma)$ is finite for all n large enough. (A10) is the standard uniform convergence of the sample objective function to the population objective function, for which, given assumption (A4) (and (A7)), it is sufficient that $C(\gamma)$ be estimated uniformly consistently. (A11) is the well-separability assumption, in our case, for Γ^*. Lemma 3 below uses ((A9),) (A10) and (A11) to obtain that $d(\widehat{\gamma}_n, \Gamma^*) = o_p(1)$. On the other hand, to obtain the rate of convergence of $d(\widehat{\gamma}_n, \Gamma^*)$ to zero, the lemma additionally uses (A12) that is a strengthened local version of (A10), and (A13) that ensures the local identification of Γ^*. Note that, (A10) and (A12) depend intrinsically on the estimator $\widehat{\Sigma}_n(\gamma)$ that forms the basis for our proposed MWLS estimator. We provide a self-contained discussion of $\widehat{\Sigma}_n(\gamma)$ in Section 2.3.3.

Lemma 3 *Consider any sequence $\widehat{\gamma}_n$ satisfying* (16)*.*

(i) Let $\Gamma = \Gamma^*$. Then $d(\widehat{\gamma}_n, \Gamma^*) \equiv 0$ trivially.

(ii) Let $\Gamma \setminus \Gamma^*$ be non-empty. If assumptions (A7)-(A13) hold then $\sqrt{n}d(\widehat{\gamma}_n, \Gamma^*) = O_p(1)$.

Remark: Note that, we are not after estimating the set Γ^* as in, e.g., Chernozhukov et al. (2007). Rather, we are interested in the rate at which the closeness between the point estimator $\widehat{\gamma}_n$ and the set Γ^*, in terms of the distance $d(\widehat{\gamma}_n, \Gamma^*)$ in (17), converges to zero in probability. A rate that is too slow will affect the asymptotic behavior of our feasible optimal estimator of $h(\beta)$. However, the rate $\sqrt{n}d(\widehat{\gamma}_n, \Gamma^*) = O_p(1)$, that follows from (A12), is stronger than needed; $n^{1/4}d(\widehat{\gamma}_n, \Gamma^*) = o_p(1)$ would have sufficed since our framework satisfies the "orthogonality condition" of Andrews (1994).

2.3.2 Feasible optimal estimator of $h(\beta)$: the MWLS estimator

Finally, by plugging in $\widehat{\gamma}_n$ defined in (16) in the expressions for $\widehat{\beta}_n(\gamma)$ and $\widehat{h}_n(\gamma)$ defined in (4) and (5) respectively, we define our proposed MWLS estimator of β and, thereby, of $h(\beta)$ as:

$$\widehat{\beta}_n := \widehat{\beta}_n(\widehat{\gamma}_n) \quad \text{and} \quad \widehat{h}_n := \widehat{h}_n(\widehat{\gamma}_n) \equiv h\left(\widehat{\beta}_n\right). \tag{18}$$

We present the asymptotic properties of $\widehat{h}_n$ under standard assumptions made on a sequence:

$$\gamma_n \in \text{ closure}(\Gamma^*) \text{ in } \Gamma \text{ satisfying } \|\widehat{\gamma}_n - \gamma_n\| = d(\widehat{\gamma}_n, \Gamma^*) \text{ for all } n \geq 1. \tag{19}$$

As noted before, these asymptotic properties are not affected by the issue that the sequence γ_n is not necessarily unique, and not much affected by the issue that the sequence γ_n may not converge.

A third issue is that γ_n can be on the boundary of Γ. Our heuristic argument for addressing this issue appeals to: (i) our assumption that Γ is compact in $\mathbb{R}^k$, and (ii) the observation that, in practice, the boundary of Γ is primarily dictated by the condition $\omega^2(X;\gamma) = 0$. First, recall that, our assumption (A3b) states that $\omega^2(X;\gamma) > 0$ with probability one for $\gamma \in \Gamma$ where Γ is compact in $\mathbb{R}^k$.[5] Hence, there exists a $\delta > 0$ such that still $\omega^2(X;\gamma) > 0$ with probability one for $\gamma \in \Gamma_\delta := \{\bar{\gamma} \in \mathbb{R}^k : d(\bar{\gamma}, \Gamma) \leq \delta\}$ provided that the functional form of $\omega^2(X;\gamma)$ is compatible with the extended domain Γ_δ. (Such compatibility holds for the commonly used functional forms of $\omega^2(X;\gamma)$; see, e.g., Wooldridge (2012), Romano and Wolf (2017), etc.) Since $\gamma_n \in \text{interior}(\Gamma_\delta)$ for any $\delta > 0$ and all $n \geq 1$ by the definition in (19), we can now assume differentiability of $\omega^2(X;\gamma)$ in $\gamma \in \text{interior}(\Gamma_\delta)$ (avoiding the boundary problem) that enables the use of the delta method.[6]

This assumption and others are listed in (A14)-(A19). We make these assumptions on the sequence γ_n (instead of $\gamma \in \Gamma_\delta^*$ or $\gamma \in \Gamma_\delta$ for some $\delta > 0$) simply to make their usage more explicit.

Assumption (A14)-(A19): [on the sequence γ_n defined in (19)]

(A14) There exist a sequence of $k \times 1$ vector $\nabla_\gamma(X;\gamma_n)$ and a sequence of positive scalar $\Delta(X;\gamma_n)$ such that the following inequality holds with probability one *for all* n large enough and some $\delta > 0$:

$$\sup_{\gamma \in \Gamma : \|\gamma - \gamma_n\| < \delta} \left\{ \left| \frac{1}{\omega^2(X;\gamma)} - \frac{1}{\omega^2(X;\gamma_n)} - \nabla_\gamma(X;\gamma_n)'(\gamma - \gamma_n) \right| - \frac{1}{2}\|\gamma - \gamma_n\|^2 \Delta(X;\gamma_n) \right\} \leq 0.$$

(A15) $W_{i,n} := X_i u_i \nabla_\gamma(X_i;\gamma_n)'$ satisfies $\frac{1}{n}\sum_{i=1}^n W_{i,n} = o_p(1)$.

(A16) $V_{i,n} := X_i u_i \Delta(X_i;\gamma_n)$ satisfies $\frac{1}{n}\sum_{i=1}^n \|V_{i,n}\| = O_p(1)$.

(A17) $\sup_{\gamma \in \Gamma : \|\gamma - \gamma_n\| < \delta} \|B(\gamma) - B(\gamma_n)\| = o(1)$ for some $\delta > 0$.

(A18) $\sup_{\gamma \in \Gamma : \|\gamma - \gamma_n\| < \delta} \|C(\gamma) - C(\gamma_n)\| = o(1)$ for some $\delta > 0$.

(A19) $Z_{i,n} := X_i u_i / \omega^2(X_i;\gamma_n)$ satisfies the following central limit theorems (CLTs):

(a) $\Xi(\gamma_n)^{-1/2} \frac{1}{\sqrt{n}} \sum_{i=1}^n B(\gamma_n)^{-1} Z_{i,n} \xrightarrow{d} N(0, I_p)$,

(b) $\Sigma(\gamma_n)^{-1/2} \frac{1}{\sqrt{n}} \sum_{i=1}^n HB(\gamma_n)^{-1} Z_{i,n} \xrightarrow{d} N(0, I_{d_h})$

where $\Xi(\gamma)^{1/2}$ and $\Sigma(\gamma)^{1/2}$ are such that $\Xi(\gamma) = \Xi(\gamma)^{1/2}\Xi(\gamma)^{1/2'}$ and $\Sigma(\gamma) = \Sigma(\gamma)^{1/2}\Sigma(\gamma)^{1/2'}$.

Remarks: Assumption (A14) is the smoothness assumption on $\omega^2(X;\gamma)$ as just discussed above. Assumptions (A15) and (A16) help to ensure that the difference $\frac{1}{\sqrt{n}}\sum_{i=1}^n \left[\frac{X_i u_i}{\omega^2(X_i;\widehat{\gamma}_n)} - \frac{X_i u_i}{\omega^2(X_i;\gamma_n)} \right]$ is $o_p(1)$. $W_{i,n}$ in assumption (A15) is a row-wise i.i.d. triangular array with $E[W_{i,n}] = 0$ by (2). Hence,

[5] The positivity assumption for $\omega^2(X;\gamma)$ is standard in the literature. For example, condition A6 in Romano and Wolf (2017) assumes that $\omega^2(X;\gamma) > 0$. The compactness assumption on Γ is difficult to avoid in our paper since the criterion for minimization, $Q(\gamma)$, cannot be shown to be convex although it is based on the trace (see Appendix B.1). Romano and Wolf (2017) do not explicitly assume a compact Γ, but Wooldridge (2010)(chapter 12) does.

[6] One could also appeal to Theorem 6 in Andrews (1997) for Taylor expansion at boundary points and then deal with this issue after imposing restrictions on the left/right partial derivatives of $1/\omega^2(X;\gamma)$.

for the weak law of large number (WLLN) in (A15) it is sufficient that $\sup_n E[\|W_{i,n}\|^2]/n \to 0$. Similar conditions would give what is essentially a WLLN for the row-wise i.i.d. triangular array $\|V_{i,n}\|$ in assumption (A16). (Since $\sqrt{n}\|\widehat{\gamma}_n - \gamma_n\| = O_p(1)$, a weaker version of (A16) with $\frac{1}{n}\sum_{i=1}^n \|V_{i,n}\| = o_p(\sqrt{n})$ suffices for our results.) (A17) and (A18) are continuity assumptions and, together with the CLT assumptions in (A19), they give the asymptotic distribution of the suitably scaled $\frac{1}{\sqrt{n}}\sum_{i=1}^n \frac{X_i u_i}{\omega^2(X_i;\gamma_n)}$ by exploiting that $\Sigma(\gamma^*)$ is equal for all $\gamma^* \in \Gamma^*$. Assumptions (A17) and (A18) follow from (A14) under standard conditions such as $\limsup_n E[\|X\|^2\|\nabla_\gamma(X;\gamma_n)\|] < \infty$ and $\limsup_n E[\|X\|^2\Delta(X;\gamma_n)] < \infty$ on the existence of moments. Assumption (A19) locally strengthens (A5). The row-wise i.i.d. triangular array CLTs in (A19) hold if, e.g., $\sup_n E\|Z_{i,n}\|^{2+\varepsilon} < \infty$ for some $\varepsilon > 0$ as the Lyapounov condition will then be satisfied under assumptions (A6) and (A7).

Theorem 4 *Consider from* (18) *the estimator* $\widehat{h}_n := \widehat{h}_n(\widehat{\gamma}_n) \equiv h(\widehat{\beta}_n)$ *where* $\widehat{\beta}_n := \widehat{\beta}_n(\widehat{\gamma}_n)$. *Then the following results hold under assumptions (A15)-(A19) and the conditions for Lemma 3.*

(i) $\sqrt{n}\left(\widehat{h}_n - h(\beta_0)\right) = \sqrt{n}\left(\widehat{h}_n(\gamma^*) - h(\beta_0)\right) + o_p(1) \xrightarrow{d} N(0, \Sigma(\gamma^*))$ *if there exists a* $\gamma^* \in \Gamma^*$ *such that* $n^{1/4}(\widehat{\gamma}_n - \gamma^*) = o_p(1)$. *(This existence condition is unnecessary for our subsequent results.)*

(ii) $\sqrt{n}\left(\widehat{h}_n - h(\beta_0)\right) \xrightarrow{d} N(0, \Sigma(\gamma^*))$ *where* $\Sigma(\gamma^*)$ *is the same for all* $\gamma^* \in \Gamma^*$.

Remarks: Theorem 4 is our main result. Theorem 4(i) gives an asymptotic equivalence between the feasible and infeasible (in terms of γ) estimators. This is similar to the well known asymptotic equivalence between feasible and infeasible WLS estimators; see Carroll and Ruppert (1982) for the parametric case, and Carroll (1982), Robinson (1987) and Newey (1994) for the semiparametric case. The asymptotic distribution of the MWLS estimator then follows directly from Lemma 1(iii).

Now, consider the result in Theorem 4(ii) that is obtained under the more general conditions and do not even require $\widehat{\gamma}_n$ to converge in probability to any element of Γ^*.[7] Note that, an asymptotic equivalence result similar to that in Theorem 4(i) does not hold in this general case. This is expected because, since $\widehat{\gamma}_n$ does not necessarily converge in probability to any given point in Γ^*, there is no point of reference (like $\widehat{h}_n(\gamma^*)$ in Theorem 4(i)) with which the asymptotic equivalence would hold.

Importantly, however, there is no other cost asymptotically to our proposed estimator since Theorem 4(ii) implies that the MWLS estimator $\widehat{h}_n$ is still asymptotically optimal. Specifically, $\widehat{h}_n$ is asymptotically unbiased and converges in distribution to a normal random variable with variance $\Sigma(\gamma^*)$ that is the optimal variance that was set as the target in (8). It is in this sense that the MWLS estimator is asymptotically optimal given the parametric model $\omega^2(X;\gamma)$, irrespective of whether this model is correctly or incorrectly specified for the true conditional variance $\omega_0^2(X)$.

[7] As noted before and also evident from the proof of this result, the convergence rate in Lemma 3 that is assumed for this result is stronger than needed; $n^{1/4}d(\widehat{\gamma}_n, \Gamma^*) = o_p(1)$ suffices. This is well known and hence not made explicit.

Algorithm: We summarize below a simple algorithm to obtain the MWLS estimator:

- Step 1: Using the OLS estimator $\widetilde{\beta}_n$ of β and the residuals $(y_i - X_i'\widetilde{\beta}_n)$ for $i = 1, \ldots, n$, obtain the estimator $\widehat{\Sigma}_n(\gamma|\widetilde{\beta}_n) := H(\widetilde{\beta}_n)\widehat{B}_n(\gamma)^{-1}\widehat{C}_n(\gamma|\widetilde{\beta}_n)\widehat{B}_n(\gamma)^{-1}H(\widetilde{\beta}_n)'$ as a function of $\gamma \in \Gamma$. This is done exactly in the same way one estimates the robust asymptotic variance for OLS; the only difference is that here we divide the regressors X_i and the residual $(y_i - X_i'\widetilde{\beta}_n)$ by $\omega(X_i;\gamma)$ since we need to estimate the variance of a weighted by $\omega^{-1}(X;\gamma)$ least squares estimator. (Section 2.3.3 discusses the estimator $\widehat{\Sigma}_n(\gamma|\widetilde{\beta}_n)$ based on the so-called HC0 form of $\widehat{C}_n(\gamma|\widetilde{\beta}_n)$.)
- Step 2: Using this $\widehat{\Sigma}_n(\gamma|\widetilde{\beta}_n)$, obtain the minimizer $\widetilde{\gamma}_n$ as in (16) by minimizing $\widehat{Q}_n(\gamma) = \widehat{m}_n(\gamma) - \min_{g\in\Gamma}\widehat{m}_n(g)$ or, equivalently and more conveniently, by minimizing $\widehat{m}_n(\gamma)$ where, at each γ, $\widehat{m}_n(\gamma) = \text{Trace}(\widehat{\Sigma}_n(\gamma|\widetilde{\beta}_n))$ if $\widehat{\Sigma}_n(\gamma|\widetilde{\beta}_n)$ is positive definite, but $\widehat{m}_n(\gamma) = +\infty$ otherwise.[8]
- Step 3: Using this $\widetilde{\gamma}_n$, obtain $\widehat{h}_n(\widetilde{\gamma}_n)$ as a simple version of the MWLS estimator following (18):

$$\widehat{h}_n(\widetilde{\gamma}_n) = h\left(\widehat{\beta}_n(\widetilde{\gamma}_n)\right) \text{ where } \widehat{\beta}_n(\widetilde{\gamma}_n) = \left(\frac{1}{n}\sum_{i=1}^{n}\frac{1}{\omega^2(X_i;\widetilde{\gamma}_n)}X_iX_i'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\frac{1}{\omega^2(X_i;\widetilde{\gamma}_n)}X_iy_i\right).$$

  One might choose to stop the algorithm here, and our experience so far is that this works well.
- Step 4 (optional): Repeat Steps 1-3 but now using $\widehat{\beta}_n(\widetilde{\gamma}_n)$ obtained in Step 3 instead of $\widetilde{\beta}_n$.

One might iterate on Step 4. One might also set up the algorithm to simultaneously obtain estimates of β and γ solving their respective profiled estimating equations. However, as is well known more generally: Step 4, or iterations of Step 4, or the simultaneous estimation of β and γ do not provide improvement in the first-order asymptotics over the simple implementation involving Steps 1-3 only. Higher order improvements leading to better finite-sample properties might be possible. We have not explored this analytically here since in (unreported) simulations we find rather mixed evidence of improvement over the simple implementation (Steps 1-3) in our Monte Carlo experiment.

2.3.3 Consistent estimation of $\Sigma(.)$

An estimator of $\Sigma(.)$ is needed in at least two occasions: (i) in Step 1, where it should be consistent for $\Sigma(\gamma)$ uniformly for $\gamma \in \Gamma$, and (ii) for inference based on $\widehat{h}_n$ obtained in Step 3, where it should be consistent for $\Sigma(\gamma^*)$. It is useful to recall that although no $\gamma^* \in \Gamma^*$ is known nor can they be consistently estimated in general, $\Sigma(\gamma^*)$ is by definition the same, denote it by Σ^*, for all $\gamma^* \in \Gamma^*$.

[8]The Monte Carlo study in Section 4 uses a stricter requirement that $\widehat{B}_n(\gamma)^{-1}\widehat{C}_n(\gamma|\widetilde{\beta}_n)\widehat{B}_n(\gamma)^{-1}$ be positive definite. In both contexts, positive definiteness is equivalent to non-zero determinant or minimum eigen value that could be checked in practice by seeing if either of them is bigger than some small and positive c_n where $\lim_n c_n \to 0$.

Now we discuss the estimation of $\Sigma(\gamma)$ and $\Sigma^*(:=\Sigma(\gamma^*))$. Our exposition is generic enough to be readily employed for occasions (i) and (ii) if one also performs Step 4 or iterations of Step 4. On the other hand, if one simultaneously estimates γ and $h(\beta)$, then our exposition can be readily employed for occasion (ii), and can be extended for occasion (i) by using the standard tools.

Accordingly, first consider any estimator $\check{\beta}_n$ of β. For occasions (i) and (ii) in the case of the simple implementation (Steps 1-3), $\check{\beta}_n$ is respectively the OLS estimator of β obtained in Step 1 and the estimator $\widehat{\beta}_n(\widetilde{\gamma}_n)$ obtained in Step 3. Second, consider any estimator $\check{\gamma}_n$, such as $\widetilde{\gamma}_n$ obtained in Step 2, to be used for occasion (ii). Finally, define the estimators of $\Sigma(\gamma)$ and Σ^* respectively for occasions (i) and (ii) as:

$$\widehat{\Sigma}_n(\gamma) := \widehat{\Sigma}_n(\gamma|\check{\beta}_n) \text{ and } \widehat{\Sigma}_n^* := \widehat{\Sigma}_n(\check{\gamma}_n|\check{\beta}_n) \tag{20}$$

where, for all $g \in \Gamma$:

$$\widehat{\Sigma}_n(g|\check{\beta}_n) := H(\check{\beta}_n)\widehat{\Xi}_n(g|\check{\beta}_n)H(\check{\beta}_n)', \tag{21}$$

$$\widehat{\Xi}_n(g|\check{\beta}_n) := \widehat{B}_n(g)^{-1}\widehat{C}_n(g|\check{\beta}_n)\widehat{B}_n(g)^{-1},$$

$$\widehat{C}_n(g|\check{\beta}_n) := \frac{1}{n}\sum_{i=1}^{n}\frac{1}{(\omega^2(X_i;g))^2}X_iX_i'\left(y_i - X_i'\check{\beta}_n\right)^2, \tag{22}$$

$\widehat{B}_n(g)$ is as defined earlier, and $H(\check{\beta}_n)$ is as defined in assumption (A7) but by plugging in $\check{\beta}_n$ instead of the unknown β_0. The estimator $\widehat{C}_n(g|\check{\beta}_n)$ in (22) is in the so-called HC0 form of White (1980), but can be readily modified to the various other forms as in MacKinnon (2012), if so desired.

Although similar conditions are well known even at a more primitive level at least since White (1980), for completeness we provide a set of standard conditions, adapted to our framework, for consistent estimation of $\Sigma(\gamma)$ and Σ^* by $\widehat{\Sigma}_n(\gamma) := \widehat{\Sigma}_n(\gamma|\check{\beta}_n)$ and $\widehat{\Sigma}_n^* := \widehat{\Sigma}_n(\check{\gamma}_n|\check{\beta}_n)$ respectively.

Assumption (A20)-(A23):

(A20) $\check{\beta}_n - \beta_0 = o_p(1)$.

(A21) (a) $\sup_{\gamma\in\Gamma} E\left[\left(\frac{\|X\|}{\omega(X;\gamma)}\right)^4\right] < \infty$, and (b) $\sup_{\gamma\in\Gamma} E\left[\left(\frac{u}{\omega(X;\gamma)}\right)^4\right] < \infty$.

(A22) $1/\omega^2(X;\gamma)$ is continuous with probability one at each $\gamma \in \Gamma$.

(A23) $\bar{\gamma}_n$ is a sequence in the closure(Γ^*) in Γ such that:

(a) $d(\check{\gamma}_n, \Gamma^*) = \|\check{\gamma}_n - \bar{\gamma}_n\| = o_p(1)$, and (b) (A17) and (A18) hold for $\bar{\gamma}_n$ (in place of γ_n).

Remarks: Assumption (A20) imposes consistency on the generic estimator $\check{\beta}_n$ to accommodate for the two occasions (i) and (ii) for estimating $\Sigma(.)$. Assumption (A21) is well known to lead to the

point-wise convergence for the concerned quantities (in the estimator of $\Sigma(\gamma)$) by routinely using the Cauchy-Schwartz and Holder inequalities as appropriate. (A21) also provides the integrable dominating functions that, along with the continuity assumption in (A22) and also the other assumptions in (A4), (A7) and (A20), enable a direct application of the uniform WLLN in Theorem 2 of Jenrich (1969) to extend the point-wise convergence of an estimator of $\Sigma(\gamma)$ to uniform convergence in $\gamma \in \Gamma$.[9] The focus of our discussion is primarily on the consistent estimation of $C(\gamma)$ since consistent estimation of the other parts — $B(\gamma)$ and H — already follows from the well-understood and widely used assumptions (A4), and (A7) and (A20) respectively. Assumption (23) is used to show that $\widehat{\Sigma}_n^* := \widehat{\Sigma}_n(\check{\gamma}_n|\check{\beta}_n) = \Sigma^* + o_p(1)$, i.e., for the consistent estimation of Σ^* for occasion (ii).

Lemma 5 *The following results hold under assumptions (A4), (A7) and (A20)-(A22).*

(i) $\sup_{\gamma\in\Gamma} \|\widehat{\Sigma}_n(\gamma) - \Sigma(\gamma)\| = o_p(1)$ *where* $\Sigma(\gamma)$ *and* $\widehat{\Sigma}_n(\gamma)$ *are defined in* (7) *and* (20) *respectively.*

(ii) Additionally if (A23) hold then $\widehat{\Sigma}_n^* = \Sigma^* + o_p(1)$ *where* $\widehat{\Sigma}_n^*$ *is defined in* (20) *and* $\Sigma^* := \Sigma(\gamma^*)$.

Remarks: Lemma 5(i) provides justification for assumption (A10) that is needed for our results in Lemma 3 and Theorem 4.[10,11] Lemma 5(ii), along with Theorem 4(ii), provides justification for standard Wald-type inference on $h(\beta)$ that we employ in the Monte Carlo experiment in Section 4.

3 Discussion of the main result: Theorem 4

To avoid notational clutter we take $h(\beta) \equiv \beta$ as the identity function, and hence $\Sigma(\gamma) \equiv \Xi(\gamma)$.

[9] The representation of (A21) may appear unusual because of the uniformity condition. However, since we are after uniform consistency of $\widehat{\Sigma}_n(\gamma)$, such a representation cannot be avoided without explicitly imposing restrictions on the tail behavior of $\omega^{-1}(X;\gamma)$ uniformly in γ. This is because our assumption (A3b), with only a positivity instead of the bounded away from zero condition on $\omega^2(X;\gamma)$, was weaker than usual (c.f. condition B.4 in Carroll and Ruppert (1982)). We did not impose a bounded away from zero restriction on $\omega^2(X;\gamma)$ because: (a) this does not hold for the commonly used functional forms of $\omega^2(X;\gamma)$ as in Wooldridge (2012), Romano and Wolf (2017), etc. without a bounded support for X, and (b) this would cause assumption (A3b) to be incompatible with our assumption (A3a).

[10] Assumption (A10) states that $\sup_{\gamma\in\Gamma} |\widehat{Q}_n(\gamma) - Q(\gamma)| = o_p(1)$. First, note from (13) that $m(\gamma) = \text{Trace}(\Sigma(\gamma))$ under assumption (A6). Then, note from (14) that for all n large enough, $\widehat{m}_n(\gamma) = \text{Trace}(\widehat{\Sigma}_n(\gamma))$ under assumption (A9). Combining these two observations with the result in Lemma 5(i) implies that $|\min_{g\in\Gamma} \widehat{m}_n(g) - \min_{g\in\Gamma} m(\gamma)| = o_p(1)$. Hence, it follows from the definition in (15) that $\sup_{\gamma\in\Gamma} |\widehat{Q}_n(\gamma) - Q(\gamma)| = o_p(1)$ as in assumption (A10).

[11] The other key condition on the estimator $\widehat{\Sigma}_n(\gamma)$ is assumption (A12) that leads to the desired rate of convergence in Lemma 3(ii). Given the conditions maintained in Lemma 5(i), it is clear that the rate-condition in (A12) follows if $\widehat{C}_n(\gamma|\check{\beta}_n)$ defined in (22) satisfies $\sqrt{n}(\widehat{C}_n(\gamma|\check{\beta}_n) - C(\gamma)) = O_p(1)$ uniformly in $\gamma \in \Gamma_\delta^*$ for some $\delta > 0$. Note that,

$$\widehat{C}_n(\gamma|\check{\beta}_n) - C(\gamma) \quad = \quad \frac{1}{n}\sum_{i=1}^{n} \left\{ \frac{X_i X_i' u_i^2}{\left(\omega^2(X_i;\gamma)\right)^2} - E\left[\frac{X_i X_i' u_i^2}{\left(\omega^2(X_i;\gamma)\right)^2}\right] \right\} + \frac{1}{n}\sum_{i=1}^{n} \frac{X_i X_i'}{\left(\omega^2(X_i;\gamma)\right)^2} \left\{ (y_i - X_i'\check{\beta}_n)^2 - u_i^2 \right\}.$$

While the proof of Lemma 5(i) involves showing that both terms on the RHS are $o_p(1)$ uniformly in $\gamma \in \Gamma$, it is immediately clear that a suitable functional CLT for the first term on the RHS can match the rate-condition required in (A12). On the hand, although it is less immediate from the coarse representation of the second term on the RHS above, the proof of Lemma 5(i) (see the steps in equation (35)) decomposes this second term in a way that makes it clear that similar functional CLTs can match this second term to the rate-condition required in (A12) provided that $\sqrt{n}(\check{\beta}_n - \beta_0) = O_p(1)$. This rate-condition on $\check{\beta}_n$ strengthens our assumption (A20) and, given our other assumptions, this is obviously satisfied by our choice of the OLS estimator of β as $\check{\beta}_n$. In summary, the message of footnotes 10 and 11 is that the estimator $\widehat{\Sigma}_n(\gamma)$ for $\Sigma(\gamma)$ can easily meet the requirements for our results in Lemma 3 and Theorem 4.

3.1 What happens when the parametric specification for $\omega_0^2(X)$ is correct?

If the parametric model $\omega^2(X;\gamma)$ is correctly specified for $\omega_0^2(X)$, then there exists a $\gamma_0 \in \Gamma$ such that almost surely in X:

$$\omega_0^2(X) = \omega^2(X;\gamma_0). \tag{23}$$

Under (23), Lemma 1 implies that $\sqrt{n}\left(\widehat{\beta}_n(\gamma_0) - \beta_0\right) \xrightarrow{d} N\left(0, \Sigma(\gamma_0)\right)$ where $\Sigma(\gamma_0) \equiv C(\gamma_0)^{-1} \equiv B(\gamma_0)^{-1} = \left(E\left[XX'/\omega_0^2(X)\right]\right)^{-1}$ which, if it exists, gives the well known efficiency bound for the regression coefficients in the model (1) under (2) and (3). Now, recall that, we have allowed for quite general user-specified functional forms $\omega^2(X;\gamma)$ for which γ_0 may not be necessarily unique.

First, consider the case where γ_0 is unique. Note that, for any $\gamma \in \Gamma$ we have under (23) that: $\Sigma(\gamma) - \Sigma(\gamma_0)$ is positive semi-definite since $\Sigma(\gamma_0)^{-1} - \Sigma(\gamma)^{-1} = E\left[e(\gamma)e'(\gamma)\right]$ is positive semi-definite where $e(\gamma)$ is the residual from the population least square regression of $\frac{X}{\omega(X;\gamma_0)}$ on $\frac{\omega^2(X;\gamma_0)}{\omega^2(X;\gamma)}\frac{X}{\omega(X;\gamma_0)}$. Under our assumptions, $e(\gamma)$ is zero for a $\gamma \in \Gamma$ only if $\omega^2(X;\gamma) \propto \omega^2(X;\gamma_0)$. Regardless of such proportionality, our optimal γ^* defined in (8) exists in this case since γ_0 is by definition such an optimal γ^*. Importantly, however, $\Sigma(\gamma^*) = \Sigma(\gamma_0)$ for any such optimal γ^*. Hence, Theorem 4 implies that our proposed MWLS estimator $\widehat{\beta}_n$ is asymptotically efficient.

Now, consider the case where γ_0 is not unique, and let Γ_0 denote the set of all such γ_0's. By the definition in (23), however, all $\gamma_0 \in \Gamma_0$ lead to the same asymptotic variance $\left(E\left[XX'/\omega_0^2(X)\right]\right)^{-1} = \Sigma(\gamma_0)$. Exactly by the same reason used above, it follows again that our optimal γ^* defined in (8) exists, and $\Gamma_0 \subseteq \Gamma^*$.[12] Hence, Theorem 4 implies that our proposed MWLS estimator $\widehat{\beta}_n$ is efficient.

3.2 Relation with the OLS estimator

If $\omega^2(X;\bar{\gamma})$ is identically a constant for some $\bar{\gamma} \in \Gamma$, then it is easy to see that the OLS estimator of β_0, i.e.,

$$\widetilde{\beta}_{n,\text{OLS}} = \left(\frac{1}{n}\sum_{i=1}^{n} X_i X_i'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} X_i y_i\right) \tag{24}$$

falls under the general class of estimators $\widehat{\beta}_n(\gamma)$ in (4) since $\widetilde{\beta}_{n,\text{OLS}} \equiv \widehat{\beta}_n(\bar{\gamma})$. As noted before, a provision for a constant $\omega^2(X;\gamma)$, i.e., conditional homoskedasticity, is made in all commonly used parametric models; e.g., $\omega^2(X;\gamma) = \tau(r(X)'\gamma)$ where $\tau(.) : \mathbb{R} \mapsto (0,\infty)$, and $r(X)$ is a $k \times 1$ function of X and the first element of $r(X)$ is taken as 1. Hence, in such cases, it follows that:

$$\sqrt{n}\left(\widetilde{\beta}_{n,\text{OLS}} - \beta_0\right) \equiv \sqrt{n}\left(\widehat{\beta}_n(\bar{\gamma}) - \beta_0\right) \xrightarrow{d} N\left(0, \Sigma(\bar{\gamma}) := \left(E\left[XX'\right]\right)^{-1} E\left[\omega_0^2(X)XX'\right]\left(E\left[XX'\right]\right)^{-1}\right)$$

under standard conditions. Therefore, Theorem 4 implies that the asymptotic variance $\Sigma(\gamma^*)$ of our proposed MWLS estimator $\widehat{\beta}_n$ cannot exceed that of $\widetilde{\beta}_{n,\text{OLS}}$ as long as the user-specified parametric

[12] $\Gamma_0 \subset \Gamma^*$ if $\omega^2(X;\gamma) \propto \omega^2(X;\gamma_0)$ for some $\gamma_0 \in \Gamma_0$ and $\gamma \in \Gamma \setminus \Gamma_0$. For some functional forms of $\omega^2(X;\gamma)$, an equality $\Gamma_0 = \Gamma^*$ follows if we broaden the definition in (23) by instead letting γ_0 such that $\omega_0^2(X) \propto \omega^2(X;\gamma_0)$.

model $\omega^2(X;\gamma)$ used by the MWLS estimator allows for a constant value of $\omega^2(X;\gamma)$ for some γ.

This observation is important in practice. Since the seminal work of White (1980), it has gradually been widely accepted in the profession and even recommended in standard undergraduate textbooks (see, e.g., page 695 of Stock and Watson (2011)) that using the OLS estimator with robust standard errors is preferable to using WLS since a parametric misspecification in the latter no longer guarantees that the WLS (robust) standard errors are smaller than that of the OLS. This is perhaps the main reason why the use of WLS is rare in modern econometrics/economics as highlighted by Angrist and Pischke (2010): "A legacy of White's (1980) paper ... is the near-death of generalized least squares in cross-sectional applied work." (fully quoted in our footnote 1).

Our results show that there is no reason to be pessimistic about WLS. While it is true that WLS or even the ALS estimator of Romano and Wolf (2017) (see Sections 3.3 and 3.4 below) may have larger standard error than OLS when $\omega^2(X;\gamma)$ is misspecified, one can use our proposed MWLS estimator and do at least as well as OLS asymptotically irrespective of correct or misspecification.

3.3 Relation with the conventional parametric WLS (simply "WLS") estimator

The WLS estimator of β_0 plugs in an estimator $\widetilde{\gamma}_{n,\mathrm{WLS}}$ for γ in $\widehat{\beta}_n(\gamma)$ defined in (4) where this estimator $\widetilde{\gamma}_{n,\mathrm{WLS}}$ is obtained by a least squares fit of the user-specified parametric model $\omega^2(X;\gamma)$ to the squared OLS residual $(y - X'\widetilde{\beta}_{n,\mathrm{OLS}})^2$ that proxies for $\omega_0^2(X)$. This fit is obtained in general by nonlinear least squares or, for specific functional forms $\omega^2(X;\gamma)$, by OLS in a transformed model; see, Wooldridge (2012) and Romano and Wolf (2017). Thus, the WLS estimator of β_0 is:

$$\widetilde{\beta}_{n,\mathrm{WLS}} \equiv \widehat{\beta}_n(\widetilde{\gamma}_{n,\mathrm{WLS}}) = \left(\frac{1}{n}\sum_{i=1}^{n}\frac{1}{\omega^2(X_i;\widetilde{\gamma}_{n,\mathrm{WLS}})}X_iX_i'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\frac{1}{\omega^2(X_i;\widetilde{\gamma}_{n,\mathrm{WLS}})}X_iy_i\right). \tag{25}$$

Minor extension of well known results such as Carroll and Ruppert (1982) gives that if $\sqrt{n}(\widetilde{\gamma}_{n,\mathrm{WLS}} - \gamma^\dagger) = O_p(1)$ (or, even $n^{1/4}(\widetilde{\gamma}_{n,\mathrm{WLS}} - \gamma^\dagger) = o_p(1)$) for some so-called psuedo-true value $\gamma^\dagger \in \Gamma$, then:

$$\sqrt{n}\left(\widetilde{\beta}_{n,\mathrm{WLS}} - \beta_0\right) \equiv \sqrt{n}\left(\widehat{\beta}_n(\widetilde{\gamma}_{n,\mathrm{WLS}}) - \beta_0\right) = \sqrt{n}\left(\widehat{\beta}_n(\gamma^\dagger) - \beta_0\right) + o_p(1) \xrightarrow{d} N\left(0, \Sigma(\gamma^\dagger)\right).$$

If the parametric model $\omega^2(X;\gamma)$ is correct in the sense that (23) holds for a *unique* $\gamma_0 \in \Gamma$ such that $\omega_0^2(X) = \omega^2(X;\gamma_0)$ almost surely in X, then it can be shown under standard conditions that $\gamma^\dagger = \gamma_0$ and hence WLS achieves asymptotic efficiency, i.e., $\Sigma(\gamma^\dagger) = \Sigma(\gamma_0)$. Therefore, in this case, Theorem 4 implies that our proposed MWLS estimator $\widehat{\beta}_n$ is asymptotically normal with the same asymptotic variance as that of the asymptotically efficient WLS estimator $\widetilde{\beta}_{n,\mathrm{WLS}}$.

Otherwise, Theorem 4 establishes that, by construction, the asymptotic variance $\Sigma(\gamma^*)$ of the MWLS estimator cannot exceed that of the WLS estimator $\widetilde{\beta}_{n,\mathrm{WLS}}$. If $\gamma^\dagger = \gamma^*$ defined in (8), then

of course $\Sigma(\gamma^\dagger) = \Sigma(\gamma^*)$. However, as will be seen in Section 3.4, unless the parametric specification is indeed correct, i.e., (23) holds, it is unlikely that $\gamma^\dagger = \gamma^*$ except by happenstance.

Romano and Wolf (2017)'s ALS estimator is the OLS estimator $\widetilde{\beta}_{n,\text{OLS}}$ if a pre-test cannot reject the conditional homoskedasticity of the regression error, but is the WLS estimator $\widetilde{\beta}_{n,\text{WLS}}$ otherwise. Provided that this pre-test for conditional homoskedasticity is consistent (against the user-specified, not necessarily correct, alternative), the asymptotic variance of ALS is the same as that of WLS, i.e., $\Sigma(\gamma^\dagger)$, if the parametric model $\omega^2(X;\gamma)$ is such that $\omega^2(X;\bar{\gamma})$ is identically a constant for some $\bar{\gamma} \in \Gamma$ (see Section 3.2). Indeed, this condition is maintained in Romano and Wolf (2017). Therefore, by the same reasons as above, (i) the asymptotic variance of our MWLS estimator is equal to that of ALS if the parametric model $\omega^2(X;\gamma)$ is correct or if, by chance, $\Sigma(\gamma^\dagger) = \Sigma(\gamma^*)$, (ii) but it can never exceed the asymptotic variance of ALS irrespective of if the parametric model $\omega^2(X;\gamma)$ is correct or not. Simulation results (under Romano and Wolf (2017)'s design) in Section 4 show that these asymptotic results are also well-reflected in small samples.

3.4 More primitive characterization of the variance-minimizer γ^*: An example

γ^* was characterized directly in (8) as the asymptotic variance minimizer of the weighted by $\omega^{-1}(X;\gamma)$ estimator $\widehat{\beta}_n(\gamma)$ in (4), where the asymptotic variance $\Sigma(\gamma)$ was expressed as a function of γ. It may be useful to dig deeper in an effort to characterize γ^* at a more primitive level. While this can be done under full generality, we focus on the simple case with $p = 1$ and no intercept in (1) to make this characterization more transparent by avoiding messy notation. Accordingly, taking logarithm and assuming differentiability, the first-order condition for minimizing $\Sigma(\gamma)$ implies that:

$$0 = \frac{\partial}{\partial\gamma}\log\left(\Sigma(\gamma^*)\right) \;\Rightarrow\; 0 = \frac{\partial}{\partial\gamma}\log\left(\frac{C(\gamma^*)}{B(\gamma^*)^2}\right) \;\Rightarrow\; \frac{\partial}{\partial\gamma}\log(B(\gamma^*)) = \frac{1}{2}\frac{\partial}{\partial\gamma}\log(C(\gamma^*))$$

where $B(\gamma) = E\left[\frac{1}{\omega^2(X;\gamma)}X^2\right]$ and $C(\gamma) = E\left[\frac{\omega_0^2(X)}{(\omega^2(X;\gamma))^2}X^2\right]$ in this simple case ($p = 1$). Then, it follows (see Appendix B.2) that γ^* satisfies the orthogonality conditions:[13]

$$\begin{aligned}
0 &= E\left[\frac{X^2}{\omega^2(X;\gamma^*)}\frac{\partial}{\partial\gamma}\log\left(\omega^2(X;\gamma^*)\right)\left\{\frac{u^2}{\omega^2(X;\gamma^*)} - \frac{C(\gamma^*)}{B(\gamma^*)}\right\}\right], \\
\Rightarrow 0 &= E\left[\frac{1}{\omega^2(X;\gamma^*)}\frac{X^2}{\omega^2(X;\gamma^*)}\frac{\partial}{\partial\gamma}\log\left(\omega^2(X;\gamma^*)\right)\left\{u^2 - \frac{C(\gamma^*)}{B(\gamma^*)}\omega^2(X;\gamma^*)\right\}\right], \qquad (26) \\
\Rightarrow 0 &= E\left[\frac{1}{\omega^2(X;\gamma^*)}\frac{X^2}{\omega^2(X;\gamma^*)}\frac{\partial}{\partial\gamma}\log\left(\omega^2(X;\gamma^*)\right)\left\{\omega_0^2(X) - \frac{C(\gamma^*)}{B(\gamma^*)}\omega^2(X;\gamma^*)\right\}\right].
\end{aligned}$$

[13] In this simple case, the Step 2 of the algorithm for the MWLS estimator involves obtaining $\widetilde{\gamma}_n$ by solving the sample analog of this (population) first order condition in (26):

$$0 = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{\omega^2(X_i;\widetilde{\gamma}_n)}\frac{X_i^2}{\omega^2(X;\widetilde{\gamma}_n)}\frac{\partial}{\partial\gamma}\log\left(\omega^2(X_i;\widetilde{\gamma}_n)\right)\left\{(y_i - X_i\widetilde{\beta}_n)^2 - \frac{\widehat{C}_n(\widetilde{\gamma}_n|\widetilde{\beta}_n)}{\widehat{B}_n(\widetilde{\gamma}_n)}\omega^2(X_i;\widetilde{\gamma}_n)\right\}.$$

The first line expresses the terms X^2 and u^2 weighted by $\omega^{-2}(X;\gamma^*)$. The second line, in particular the term inside the braces on the RHS of it, makes the "tilting" of $\omega^2(X;\gamma^*)$ by $C(\gamma^*)/B(\gamma^*)$ explicit. The third line is presented to make the connection with the other estimators explicit.

Under the correct specification (23) there exists a γ_0 such that $\omega^2(X;\gamma_0) = \omega_0^2(X)$ (see Section 3.1), and it is easy to see that the tilting term $C(\gamma_0)/B(\gamma_0) = 1$ at γ_0. Hence, at γ_0, we have that:

$$E\left[\frac{1}{\omega^2(X;\gamma_0)}\frac{X^2}{\omega^2(X;\gamma_0)}\frac{\partial}{\partial\gamma}\log\left(\omega^2(X;\gamma_0)\right)\left\{\omega_0^2(X)-\omega^2(X;\gamma_0)\right\}\right]=0,$$

which means that $\gamma^* = \gamma_0$ makes (26) hold. Since $\omega^2(X;\gamma^*) = \omega_0^2(X)$ when $\gamma^* = \gamma_0$, it follows that the condition (26) holds irrespective of the terms outside the braces in the expectation. Hence, under correct specification, the optimal γ, i.e., the minimizer of the asymptotic variance $\Sigma(\gamma) = C(\gamma)/B(\gamma)^2$ of $\widehat{\beta}_n(\gamma)$, leads to this same well known result of targeting $\omega_0^2(X)$ by $\omega^2(X;\gamma)$.

This discussion also applies to the case where the regression error u is actually conditionally homoskedastic (see Section 3.2).[14] Recall that the parametric model $\omega^2(X;\gamma)$ makes provision for conditional homoskedasticity by maintaining that $\omega^2(X;\gamma)$ is constant at some $\gamma = \bar{\gamma}$. Now, when represented in terms of the representation of correct specification in (23) (Section 3.1), this means that $\omega^2(X;\bar{\gamma}) = c \times \omega^2(X;\gamma_0)$ for some constant $c > 0$. Hence, at $\bar{\gamma}$, the tilting term $C(\bar{\gamma})/B(\bar{\gamma}) = (1/c)\times C(\gamma_0)/B(\gamma_0) = 1/c$, that implies that $\omega_0^2(X) - C(\bar{\gamma})/B(\bar{\gamma})\omega^2(X;\bar{\gamma}) = \omega_0^2(X) - \omega^2(X;\gamma_0) = 0$. This means that $\gamma^* = \bar{\gamma}$ makes (26) hold exactly in the same way described in the last paragraph.

Finally, consider the case of misspecification which is where the novel feature of (26) comes into play. Note that, the tilting term $C(\gamma^*)/B(\gamma^*)$ in (26) means that under misspecification we are not necessarily looking for a "best fit" (like least squares) of the user-specified model $\omega^2(X;\gamma)$ to the true conditional variance $\omega_0^2(X)$. This is where MWLS differs from the existing methods below:

(a) The conventional parametric WLS estimator for β characterizes its target $\gamma^\dagger$ (see Section 3.3) by the (un-weighted) least squares "best fit" of $\omega^2(X;\gamma)$ to $\omega_0^2(X)$ as:

$$\gamma^\dagger = \arg\min_{\gamma\in\Gamma} E\left[(u^2-\omega^2(X;\gamma))^2\right] \;\Rightarrow\; 0 = E\left[\frac{\partial}{\partial\gamma}\omega^2(X;\gamma^\dagger)\left(\omega_0^2(X)-\omega^2(X;\gamma^\dagger)\right)\right]. \tag{27}$$

This characterization also applies to the case of Romano and Wolf (2017)'s ALS (see Section 3.3).

(b) More recently, Spady and Stouli (2019) consider the joint estimation of the conditional mean and variance models in a linear regression, and show that (similar to our MWLS estimator) their estimator performs as well as the OLS estimator under misspecification of the conditional variance. Imposing $E[u|X] = 0$ in the first order condition for minimizing the criterion in their equation (2.2)

[14] This discussion would be redundant if one followed footnote 12 to broaden the definition of correct specification in (23). We prefer to treat conditional homoskedasticity separately because this case is important in its own right.

with respect to γ implies that their "best fit" of $\omega^2(X;\gamma)$ to $\omega_0^2(X)$ is characterized by a $\gamma^{\dagger\dagger} \in \Gamma$ that satisfies:

$$0 = E\left[\frac{\partial}{\partial\gamma}\omega(X;\gamma^{\dagger\dagger})\frac{\left(\omega_0^2(X) - \omega^2(X;\gamma^{\dagger\dagger})\right)}{\omega^2(X;\gamma^{\dagger\dagger})}\right]. \tag{28}$$

((28) uses our notation and ignores the fact that Spady and Stouli (2019) take $\omega^2(X;\gamma) = \omega^2(X'\gamma)$, i.e., as a linear index, but is otherwise the same as equation (3.9) of their Corollary 2.)

A key observation is that there is no tilting term like $C(\gamma)/B(\gamma)$ in (27) and (28). Specifically, irrespective of correct or misspecification of $\omega_0^2(X)$ by $\omega^2(X;\gamma)$, the $\gamma^\dagger$ in (27) (i.e., for WLS) and the $\gamma^{\dagger\dagger}$ in (28) (i.e., for Spady and Stouli (2019)) equate to zero the expectation of the respective weighted difference $\{\omega_0^2(X) - \omega^2(X;\gamma)\}$. By contrast, for MWLS, the γ^* in (26) does the same but for the weighted difference $\{\omega_0^2(X) - C(\gamma)/B(\gamma)\omega^2(X;\gamma)\}$ where $\omega^2(X;\gamma)$ is tilted by $C(\gamma)/B(\gamma)$.

As noted above, this distinction with (26) would be immaterial under correct specification, i.e., if $\omega_0^2(X) = \omega^2(X;\gamma_0)$ for some $\gamma = \gamma_0$. In this case, all the methods are equivalent in the sense that $\gamma^* = \gamma_0$ (as $C(\gamma^*)/B(\gamma^*) = 1$), $\gamma^\dagger = \gamma_0$ and $\gamma^{\dagger\dagger} = \gamma_0$ respectively solve (26), (27) and (28).

However, this distinction does matter when the model $\omega^2(X;\gamma)$ is misspecified for $\omega_0^2(X)$. The tilting term in (26) arises directly from the fact that our target γ^* minimizes the asymptotic variance $\Sigma(\gamma) = C(\gamma)/B(\gamma)^2$ of $\widehat{\beta}_n(\gamma)$. Its absence from (27) and (28) means that $\gamma^\dagger$ and $\gamma^{\dagger\dagger}$ that are used by these other methods cannot minimize this asymptotic variance except by happenstance.

Thus, the central feature in the characterization (26) of γ^* is the tilting term $C(\gamma^*)/B(\gamma^*)$. It plays a fundamental role for the optimality of γ^* and, hence, that of MWLS by switching between one and not-one automatically depending on the data, the true conditional variance $\omega_0^2(X)$, and the user-specified model $\omega^2(X;\gamma)$, and with the purpose of minimizing the asymptotic variance $\Sigma(\gamma)$.

4 Monte Carlo experiment

We conduct a small scale Monte Carlo experiment to investigate how well the asymptotic results discussed so far are reflected in finite samples. Based on our discussion above, we focus primarily on the following three asymptotic propositions of our paper for this finite-sample investigation.

P1. Section 3.2 suggests that if there exists a γ such that $\omega^2(X;\gamma)$ is constant for all X, then our proposed MWLS estimator cannot have larger asymptotic variance than the OLS estimator.

P2. Section 3.3 suggests that our proposed MWLS estimator cannot have larger asymptotic variance than the conventional parametric WLS (simply "WLS") estimator.

P3. Section 3.3 also suggests that our proposed MWLS estimator cannot have larger asymptotic

variance than the ALS estimator proposed by Romano and Wolf (2017).

The question is how well do these propositions hold up in finite samples? To investigate this (and related issues), and given P3, we consider the simulation design of Romano and Wolf (2017).

4.1 Simulation design

Romano and Wolf (2017)'s design takes $p = 2$ in (1): more precisely, $y = \beta_1 X_{(1)} + \beta_2 X_{(2)} + u$, with $X_{(1)} = 1$, $X = (X_{(1)}, X_{(2)})'$, $\beta = (\beta_1, \beta_2)'$, $\beta_0 = (0, 0)'$, and the error $u = \omega_0(X)Z$ where $Z \sim N(0, 1)$ is independent of $X_{(2)} \sim U(1, 4)$, and considers the following $4 + 2 + 2 + 2 = 10$ cases:

DGP 1: (a) $\omega_0^2(X) = 1$; (b) $\omega_0^2(X) = X_{(2)}$; (c) $\omega_0^2(X) = X_{(2)}^2$; (d) $\omega_0^2(X) = X_{(2)}^4$.

DGP 2: (a) $\omega_0^2(X) = \left(\log(X_{(2)})\right)^2$; (b) $\omega_0^2(X) = \left(\log(X_{(2)})\right)^4$.

DGP 3: (a) $\omega_0^2(X) = \exp\left(.1(X_{(2)} + X_{(2)}^2)\right)$; (b) $\omega_0^2(X) = \exp\left(.15(X_{(2)} + X_{(2)}^2)\right)$.

DGP 4: (a) $\omega_0^2(X) = \begin{cases} 1 & \text{if } X_{(2)} < 2 \\ 2 & \text{if } 2 \leq X_{(2)} < 3 \\ 3 & \text{if } X_{(2)} \geq 3 \end{cases}$; (b) $\omega_0^2(X) = \begin{cases} 1 & \text{if } X_{(2)} < 2 \\ 2^2 & \text{if } 2 \leq X_{(2)} < 3 \\ 3^2 & \text{if } X_{(2)} \geq 3 \end{cases}$.

Following Romano and Wolf (2017), we consider two parametric models $\omega^2(X; \gamma)$ for $\omega_0^2(X)$:

Model 1: $\omega^2(X; \gamma) := \exp(\gamma_1 + \gamma_2 \log(X_{(2)}))$

Model 2: $\omega^2(X; \gamma) := \exp(\gamma_1 + \gamma_2 X_{(2)})$.

We consider two choices for the parameter of interest $h(\beta)$ — β_1 and β_2 — separately. We consider four types of estimators: (i) the OLS estimator in (24), (ii) the WLS estimator in (25), (iii) the ALS estimator of Romano and Wolf (2017), and (iv) the MWLS estimator in (18).

The WLS, ALS and MWLS estimators are obtained based on the two parametric models Model 1 and Model 2 separately and, to signify this, their respective versions are henceforth referred to as WLS1 and WLS2, ALS1 and ALS2, and MWLS1 and MWLS2. ALS1 and ALS2 are obtained following Romano and Wolf (2017). While WLS1 and WLS2 can be obtained in different ways, we again follow Romano and Wolf (2017) because this makes our simulation results most favorable to WLS1 and WLS2. MWLS1 and MWLS2 are obtained following Steps 1-3 of the algorithm in Section 2.3.2. For completeness, we provide the details of these implementations in Appendix B.3.

The asymptotic propositions P1, P2 and P3 involve a comparison of the variability of MWLS1 relative that of OLS, WLS1 and ALS1 respectively when the user specifies Model 1, and, similarly, a comparison of the variability of MWLS2 relative to that of OLS, WLS2 and ALS2 respectively when the user specifies Model 2. To investigate if propositions P1-P3 also hold up in finite samples,

we consider two measures of variability (Var): (i) VarAS: average variance based on the asymptotic (AS) variance formula, and (ii) VarMC: variance based on Monte Carlo (MC). VarAS for all estimators is obtained using the HC0 form: as in White (1980) for OLS, and as in Section 2.3.3 for the other estimators. VarMC is infeasible in practice but is more reliable than VarAS as a measure of the true variability, and hence serves as a benchmark for the reliability of VarAS in our simulations.

4.2 Simulation results

We report results for samples sizes $n = 25, 50, 100, 200$ and 400, and based on 25,000 Monte Carlo trials. Table 1 reports the ratios of VarMC for the OLS, WLS and ALS estimators for β_1 and β_2 with respect to the VarMC of the MWLS estimator; and does this in both cases: Models 1 and 2. Table 2 reports the same but with VarAS instead of VarMC, and is subsequently used to discuss the implications of these ratios in practice in terms of inference and confidence intervals.[15]

Evidence for proposition P1 in terms of VarMC in finite samples: MWLS relative to OLS

First, note that the asymptotic variance of OLS cannot be less than that of MWLS since the functional forms of $\omega^2(X;\gamma)$ in Models 1 and 2 allow for conditional homoskedasticity with $\gamma_2 = 0$.

Now, consider DGP 1a that represents conditional homoskedasticity. OLS is efficient here. Under DGP 1a, Table 1 shows that in very small samples $n = 25$ or 50: the ratio of VarMC for OLS with respect to MWLS ranges from .89 to .94 in Model 1 and from .91 to .95 in Model 2, i.e., MWLS suffers a little loss in efficiency with respect to OLS. However, when the sample size is a little bigger, i.e., $n = 100, 200$, or 400, there is practically no loss in efficiency under DGP 1a.

All other DGPs represent conditional heteroskedasticy. Hence, OLS is no longer efficient. Under these DGPs, the smallest value of the ratio of VarMC for OLS with respect to MWLS in Models 1 and 2 is .99, and this occurs only two times — under DGPs 1b and 4a when estimating β_1 with $n = 25$ and using Model 1. In both cases the ratio quickly exceeds 1 with an increase in n, as it should, since OLS is not efficient in either case. Otherwise, the ratio is always bigger than 1, and often much bigger; see, e.g., the large values of this ratio under DGP 1c (ranging from 1.20 to 1.61), DGP 1d (ranging from 1.72 to 4.77) , DGP 2a (ranging from 1.37 to 5.43), DGP 2b (ranging from 1.76 to 13.19), DGP 3b (ranging from 1.27 to 1.65) and DGP 4b (ranging from 1.17 to 1.51).

Based on the above observations, we conclude that in the context of our simulation experiment we find strong evidence suggesting that proposition P1 can very well hold up in small samples.

[15]The actual VarMC and VarAS (that give these ratios) of all estimators are reported in Tables 5, 6 and 7 in Appendix B.4. We also note that the empirical bias is very small and similar for all estimators in our simulations, and we find the empirical mean squared error for the estimators to be essentially the same as their respective VarMC (divided by n). Hence, empirical bias and mean squared error are not reported to avoid clutter and redundancy.

Evidence for proposition P2 in terms of VarMC in finite samples: MWLS relative to WLS

Since the functional forms of $\omega^2(X;\gamma)$ in Models 1 and 2 allow for conditional homoskedasticity with $\gamma_2 = 0$, WLS is asymptotically efficient under DGP 1a. This is reflected in small samples by the ratio of VarMC in Table 1 for WLS with respect to MWLS (the ratio of the ratios WLS/MWLS and OLS/MWLS gives WLS/OLS) in both Models 1 and 2. The ratio of VarMC for WLS with respect to MWLS reveals that while MWLS can suffer a little loss in efficiency relative to WLS in very small samples under DGP 1a, the loss vanishes quickly as the sample size increases.

When the sample size is very small, MWLS can also suffer a little loss in efficiency under other DGPs. However, as the sample size increases, this efficiency loss always either vanishes or turns into efficiency gains. This switch to efficiency gains is very interesting and can be seen, e.g., under DGP 2a: for β_2 in Model 1, and under DGP 2b: for both β_1 and β_2 and in both Models 1 and 2.

Indeed, under DGPs 2a and 2b, the efficiency gain resulting from MWLS is quite substantial. For example, the WLS variability can be even about one and half times that of MWLS. For convenience of the reader, we highlight in Table 1 with blue color the cases of such efficiency gains.

Based on the above observations, we conclude that there is strong evidence in the simulation results suggesting that proposition P2 can hold up very well even in reasonably small samples.

Evidence for proposition P3 in terms of VarMC finite samples: MWLS relative to ALS

In Table 1, the ratio of VarMC for ALS with respect to MWLS is always either equal to or slightly bigger than the ratio of VarMC for WLS with respect to MWLS. This holds for the estimation of both β_1 and β_2 in both Models 1 and Models 2 (i.e., (WLS1/MWLS1) compared to (ALS1/MWLS1), and (WLS2/MWLS2) compared to (ALS2/MWLS2)). Hence, our discussion of efficiency gains of MWLS over WLS in the context of proposition P2 above also applies here. Therefore, we again conclude that there is strong evidence in the simulation results suggesting that proposition P3 can hold up very well even in reasonably small samples.

Evidence for propositions P1, P2 and P3 in terms of VarAS finite samples

The ratios of VarAS for OLS, WLS and ALS with respect to MWLS in Table 2 show that there is *never* any loss in efficiency of MWLS in terms of VarAS. While looking at Table 2, please note that the asymptotic variances of WLS and ALS are identical by definition, and so are their VarAS and, hence, the ratio of their VarAS with respect to MWLS. Therefore, to avoid redundancy, the corresponding ratios for WLS and ALS are presented together under the heading of Models 1 and 2, i.e., as WLS1 & ALS1 to MWLS1, and WLS2 & ALS2 to MWLS2.

We also note that the efficiency gains due to MWLS relative to OLS, WLS and ALS that were

observed in Table 1 in terms of VarMC are also all observed at similar levels of magnitude in terms of VarAS in Table 2. Therefore, we conclude that there is strong evidence in the simulation results suggesting that propositions P1-P3, in terms of VarAS, hold up very well even in small samples.

Inference in finite samples

Having seen the strong evidence of efficiency gains by MWLS in small samples in the context of propositions P1-P3, it is then natural to ask what their implications are for inference. For the sake of practicality, we explore this in terms of VarAS since it is feasible in practice while VarMC is not. We can still appeal to our discussion so far since, as documented in Tables 1 and 2, the ratios of VarAS for the different estimators with respect to MWLS follow the same pattern as those of VarMC. Indeed, these two sets of ratios are essentially the same in relatively less small samples.

The main implication of this efficiency gain is that it reduces the length of confidence intervals or, equivalently, increases the power of tests. For example, if the ratio of AvarAS for (OLS/MWLS) is a ($\in (0, \infty)$) then, roughly speaking, the confidence intervals obtained by inverting the t-test based on the OLS estimator are on average $\sqrt{a}$ times longer than the intervals of the same nominal level but based on the MWLS estimator. (This only gives a rough back-of-the envelope idea since it does not distinguish between the ratio of averages and the average of ratios.)

Accordingly, our discussion of proposition P1 suggests that the OLS and MWLS intervals are, on average, of almost equal length under DGP 1a (conditional homoskedasticity), while the MWLS intervals are shorter, often much so, under the other DGPs. Similarly, our discussion of propositions P2 and P3 respectively suggests that the MWLS intervals are, on average, roughly equal to or shorter (sometimes much so) than the WLS and ALS intervals.

However, shorter confidence intervals based on MWLS will be useless without a good coverage probability. Tables 3 and 4 present for β_1 and β_2 respectively the empirical non-coverage probability (in percentage) of such intervals of nominal level 95%. Specifically, these numbers are the empirical rejection rate of the true parameter value by a two-sided t-test using the 5%-$N(0,1)$ critical value.

In all cases, the empirical size for OLS is always the closest to the 5% level. This closeness is evident even in very small samples except under DGP 1a for which, however, it is well known that using another HC form (instead of the HC0 form used here) of the asymptotic variance estimator that is unbiased under conditional homoskedasticity would largely resolve this issue.

The empirical size for all estimators is generally closer to the 5% level for β_2 than for β_1. There is almost always a noticeable upward distortion in the empirical size for MWLS when the sample size is very small. This is also the case for WLS and ALS but to a slightly lesser degree.[16] However,

[16] As can be seen from Tables 8-9 (Appendix B.4) where the same empirical size is presented but using the respective

these distortions for WLS, ALS and MWLS vanish as the sample size increases and the empirical size is always reasonably close to the 5% level when sample size increases to $n = 100$.

In summary, the empirical size of MWLS (and OLS, WLS and ALS) and, hence, the coverage probability of the MWLS (and OLS, WLS and ALS) intervals are generally good under this simulation design of Romano and Wolf (2017). Coupled with the efficiency gains due to MWLS, this suggests that the proposed MWLS estimator can be an attractive choice to users in applied work.

5 Conclusion

Our paper was inspired by Romano and Wolf (2017) who challenged the conventional wisdom of preferring OLS estimation with robust standard errors over (parametric) WLS estimation. OLS estimation has always been attractive because, unlike WLS, it does not require the user to postulate parametric models for the conditional variance of the regression error. Hence, WLS estimation naturally fell out of favor among practitioners following the seminal work of White (1980).

Practitioners could only be encouraged to use WLS if, along with the well known efficiency gains that result from WLS in some cases, it can also be guaranteed that WLS will never suffer any efficiency loss relative to OLS in other cases. However, this is not possible. It is well known that the asymptotic variance of WLS can be more than that of OLS if the postulated parametric model for the conditional variance is misspecified. Unfortunately, in spite of its desirable performance otherwise, the ALS estimator proposed by Romano and Wolf (2017) does not solve this problem.

Our paper solved this problem head-on by taking a new and direct route to optimality. Our proposed MWLS estimator is guaranteed to have an asymptotic variance that is less than or equal to that of: (i) the OLS estimator and, more generally, (ii) any weighted by $\omega^{-1}(X;\gamma)$ least squares estimator including the WLS estimator and, therefore, the ALS estimator. The MWLS estimator is semiparametrically efficient if the postulated parametric model $\omega^2(X;\gamma)$ happens to be correct.

We provided a thorough discussion of the MWLS estimator and its superior asymptotic properties. A Monte Carlo experiment under the simulation design of Romano and Wolf (2017) demonstrated the excellent performance of the MWLS estimator even in reasonably small samples.

We conclude by emphasizing again that the key feature that we exploited in proposing the MWLS estimator is that: a parametric misspecification of the nuisance parameters in this estimation framework does not affect the consistency of the estimator for the parameters of interest.

test statistics that are instead based on the VarMC (infeasible), there is never any such distortion even when the sample size is very small. Hence, the size distortion in very small samples in Tables 3-4 stems from the poor approximation of the true variance by the asymptotic variance formula in such small samples (also see Tables 5-7 in Appendix B.4). This is a common issue and to fix it there exists a vast array of methods on robust variance estimation; see MacKinnon (2012). An analytical study of this and related finite-sample issues is left for future work.

There are various other frameworks that also enjoy this feature. The same idea should, in principle, be applicable in all those cases. An exploration of this idea under these other frameworks and the search for optimality even in the higher-order properties are left for our ongoing research.

References

Andrews, D. W. K. (1994). Asymptotics for Semiparametric Econometric Models Via Stochastic Equicontinuity. *Econometrica*, 62:43–72.

Andrews, D. W. K. (1997). Estimation when a parameter is on a boundary: Theory and applications. Yale University.

Angrist, J. D. and Pischke, J. S. (2010). The credibility revolution in empirical economics: How better research design is taking the con out of econometrics. *Journal of Economic Perspective*, 24: 3–30.

Carroll, R. J. (1982). Adapting for heteroscedasticity in linear models. *The Annals of Statistics*, 10: 1224–1233.

Carroll, R. J. and Ruppert, D. (1982). Robust estimation in heteroscedastic linear models. *The Annals of Statistics*, 10: 429–441.

Chernoff, H. (1953). Locally optimal designs for estimating parameters. *Annals of Mathematical Statistics*, 24: 586–602.

Chernozhukov, V., Hong, H., and Tamer, E. (2007). Estimation and confidence regions for parameter sets in econometric models. *Econometrica*, 75: 1243–1284.

Ehrenfeld, S. (1956). Complete class theorems in experimental design. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 57–67. University of California Press.

Elfving, G. (1952). Optimum allocation in linear regression theory. *Annals of Mathematical Statistics*, 23: 255–262.

Federov, V. V. (1971). Design of experiments for linear optimality criteria. *Theory of Probability and its Applications*, 16: 189–195.

Heyde, C. C. (1997). *Quasi-Likelihood And Its Application: A General Approach to Optimal Parameter Estimation*. Springer.

Jenrich, R. L. (1969). Asymptotic Properties of Non-Linear Least Squares Estimators. *The Annals of Mathematical Statistics*, 40:633–643.

Karlin, S. and Studden, W. J. (1966). Optimal experimental designs. *Annals of Mathematical Statistics*, 37: 783–815.

Kiefer, J. (1974). General equivalence theory for optimum designs (approximate theory). *Annals of Statistics*, 2: 849–879.

Leamer, E. E. (2010). Tantalus on the Road to Asymptotia. *Journal of Economic Perspective*, 24: 31–46.

MacKinnon, J. G. (2012). Thirty Years of Heteroskedasticity-Robust Inference. In Chen, X. and Swanson, N. R., editors, *Recent Advances and Future Directions in Causality, Prediction, and Specification Analysis*, pages 437–461. Springer.

Newey, W. K. (1994). Series Estimation of Regression Functionals. *Econometric Theory*, 10: 1–28.

Newey, W. K. and McFadden, D. L. (1994). Large Sample Estimation and Hypothesis Testing. In Engle, R. F. and McFadden, D., editors, *Handbook of Econometrics*, volume IV, chapter 36, pages 2212–2245. Elsevier Science Publisher.

Robinson, P. M. (1987). Asymptotically Efficient Estimation in the Presence of Heteroskedasticity of Unknown Form. *Econometrica*, 55: 875–891.

Romano, J. P. and Wolf, M. . (2017). Resurrecting Weighted Least Squares. *Journal of Econometrics*, 197: 1–19.

Spady, R. and Stouli, S. (2019). Simultaneous Mean-Variance Regression. Working paper.

Stock, J. H. and Watson, M. W. (2011). *Introduction to Econometrics*. Pearson, 3 edition.

Wald, A. (1943). On the efficient design of statistical investigations. *Annals of Mathematical Statistics*, 14: 134–140.

White, H. (1980). A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heterogeneity. *Econometrica*, 48:817–838.

Wooldridge, J. M. (2010). *Econometric Analysis of Cross Section and Panel Data*. The MIT Press.

Wooldridge, J. M. (2012). *Introductory Econometrics*. South-Western, Mason, Ohio.

DGP	n	Estimating β_1: Intercept						Estimating β_2: Slope					
		Model 1 for $\omega^2(X;\gamma)$			Model 2 for $\omega^2(X;\gamma)$			Model 1 for $\omega^2(X;\gamma)$			Model 2 for $\omega^2(X;\gamma)$		
		OLS	WLS1	ALS1	OLS	WLS2	ALS2	OLS	WLS1	ALS1	OLS	WLS2	ALS2
1a	25	.89	.95	.95	.91	.96	.96	.91	.97	.97	.93	.99	.99
	50	.93	.97	.97	.94	.97	.97	.94	.98	.98	.95	.99	.99
	100	.96	.98	.98	.97	.98	.98	.97	.99	.99	.97	.99	.99
	200	.98	1.00	1.00	.98	1.00	1.00	.98	1.00	1.00	.98	1.00	1.00
	400	.99	1.00	1.00	.99	1.00	1.00	.99	1.00	1.00	.99	1.00	1.00
1b	25	.99	.97	.98	1.01	.98	.99	1.01	.98	.99	1.01	.99	1.00
	50	1.07	.98	1.00	1.07	.99	1.00	1.06	.99	1.01	1.06	.99	1.00
	100	1.06	.98	1.00	1.07	.99	1.00	1.06	.99	1.00	1.06	1.00	1.01
	200	1.10	.99	.99	1.10	1.00	1.00	1.09	1.00	1.00	1.09	1.00	1.00
	400	1.13	1.00	1.00	1.13	1.00	1.00	1.11	1.00	1.00	1.11	1.00	1.00
1c	25	1.30	.96	.99	1.32	.99	1.01	1.21	.98	1.00	1.20	.99	1.01
	50	1.56	.98	1.00	1.55	.99	1.01	1.39	.99	1.00	1.36	.98	1.00
	100	1.47	.98	.98	1.47	.99	.99	1.37	.99	.99	1.36	.99	1.00
	200	1.60	.99	.99	1.59	1.00	1.00	1.46	1.00	1.00	1.44	1.00	1.00
	400	1.61	1.00	1.00	1.59	1.01	1.01	1.45	1.00	1.00	1.43	1.00	1.00
1d	25	2.48	.87	.87	2.57	.94	.94	1.72	.94	.94	1.73	.95	.95
	50	4.05	.88	.88	4.16	.94	.94	2.55	.96	.96	2.53	.95	.95
	100	3.86	.89	.89	3.93	.96	.96	2.78	.96	.96	2.77	.98	.98
	200	4.77	.94	.94	4.68	1.00	1.00	3.10	.97	.97	3.02	.99	.99
	400	4.66	.99	.99	4.47	1.01	1.01	3.02	1.00	1.00	2.89	1.00	1.00
2a	25	2.18	1.04	1.04	2.00	1.07	1.07	1.44	.97	.97	1.37	.98	.98
	50	2.99	1.15	1.15	2.62	1.18	1.18	1.92	1.01	1.01	1.69	.98	.98
	100	5.43	1.30	1.30	4.54	1.36	1.36	2.39	1.06	1.06	2.16	1.05	1.05
	200	4.86	1.38	1.38	3.45	1.23	1.23	2.16	1.06	1.06	1.81	1.00	1.00
	400	4.58	1.39	1.39	3.15	1.20	1.20	2.17	1.07	1.07	1.83	1.01	1.01
2b	25	3.36	.91	.91	3.31	.98	.99	1.87	.96	.96	1.76	.94	.94
	50	5.07	.80	.80	5.39	.98	.98	3.23	.93	.93	3.14	.97	.97
	100	9.14	1.12	1.12	8.91	1.23	1.23	4.01	1.14	1.14	3.67	1.09	1.09
	200	10.94	1.31	1.31	9.49	1.45	1.45	4.66	1.26	1.26	3.81	1.19	1.19
	400	13.19	1.45	1.45	10.82	1.57	1.57	5.11	1.28	1.28	3.97	1.17	1.17
3a	25	1.04	.95	.99	1.07	.96	1.00	1.08	.98	1.02	1.10	.99	1.03
	50	1.11	.98	1.02	1.14	.98	1.03	1.14	.99	1.03	1.15	.98	1.03
	100	1.14	.99	1.00	1.16	.99	1.01	1.16	1.00	1.01	1.18	1.00	1.02
	200	1.19	1.00	1.00	1.21	1.00	1.00	1.20	1.00	1.00	1.22	1.00	1.00
	400	1.20	1.00	1.00	1.22	1.00	1.00	1.21	1.00	1.00	1.23	1.00	1.00
3b	25	1.27	.94	.98	1.32	.96	1.00	1.27	.97	1.01	1.30	.98	1.02
	50	1.45	.97	1.04	1.52	.97	1.06	1.45	.99	1.06	1.51	.99	1.07
	100	1.44	.99	1.00	1.48	.99	1.00	1.43	1.00	1.01	1.47	1.00	1.01
	200	1.53	1.00	1.00	1.58	1.00	1.00	1.50	1.00	1.00	1.56	1.00	1.00
	400	1.60	1.00	1.00	1.65	1.00	1.00	1.56	1.00	1.00	1.62	1.00	1.00
4a	25	.99	.97	.98	1.01	.98	.99	1.00	.99	1.00	1.01	1.00	1.01
	50	1.02	.99	.99	1.03	.99	.99	1.02	.99	1.00	1.01	.98	.99
	100	1.07	1.00	1.00	1.08	.99	1.00	1.07	1.00	1.00	1.07	1.00	1.00
	200	1.09	1.00	1.00	1.10	1.00	1.00	1.08	1.00	1.00	1.08	1.00	1.00
	400	1.09	1.01	1.01	1.09	1.00	1.00	1.08	1.01	1.01	1.08	1.00	1.00
4b	25	1.26	.97	1.03	1.31	.99	1.06	1.17	.98	1.03	1.17	.99	1.03
	50	1.39	.98	1.03	1.44	.97	1.02	1.30	.98	1.02	1.31	.97	1.01
	100	1.46	1.00	1.00	1.50	1.00	1.00	1.38	1.00	1.00	1.40	1.00	1.00
	200	1.46	1.02	1.02	1.49	1.01	1.01	1.35	1.01	1.01	1.36	1.01	1.01
	400	1.49	1.03	1.03	1.51	1.01	1.01	1.36	1.02	1.02	1.36	1.01	1.01

Table 1: Ratios of VarMC of OLS, WLS1 and ALS1 with respect to that of MWLS1 are listed under the heading "Model 1 for $\omega^2(X;\gamma)$". Ratios of VarMC of OLS, WLS2 and ALS2 with respect to that of MWLS2 are listed under the heading "Model 2 for $\omega^2(X;\gamma)$". WLS1, ALS1 and MWLS1 use Model 1, i.e., $\exp(\gamma_1 + \log(X_{(2)}))$. WLS2, ALS2 and MWLS2 use Model 2, i.e., $\exp(\gamma_1 + X_{(2)})$.

DGP	n	Estimating β_1: Intercept				Estimating β_2: Slope			
		Model 1 for $\omega^2(X;\gamma)$		Model 2 for $\omega^2(X;\gamma)$		Model 1 for $\omega^2(X;\gamma)$		Model 2 for $\omega^2(X;\gamma)$	
		OLS	WLS1 & ALS1	OLS	WLS2 & ALS2	OLS	WLS1 & ALS1	OLS	WLS2 & ALS2
1a	25	1.23	1.20	1.13	1.12	1.14	1.12	1.09	1.08
	50	1.10	1.08	1.07	1.06	1.07	1.06	1.06	1.05
	100	1.04	1.04	1.04	1.03	1.04	1.03	1.03	1.03
	200	1.02	1.02	1.02	1.02	1.02	1.02	1.02	1.01
	400	1.01	1.01	1.01	1.01	1.01	1.01	1.01	1.01
1b	25	1.41	1.21	1.29	1.13	1.25	1.11	1.20	1.08
	50	1.28	1.09	1.24	1.08	1.21	1.06	1.18	1.05
	100	1.19	1.05	1.17	1.04	1.16	1.04	1.14	1.03
	200	1.17	1.03	1.16	1.02	1.15	1.02	1.13	1.02
	400	1.15	1.01	1.14	1.01	1.12	1.01	1.12	1.01
1c	25	1.87	1.16	1.70	1.11	1.49	1.08	1.44	1.07
	50	1.90	1.08	1.81	1.08	1.56	1.05	1.52	1.05
	100	1.68	1.06	1.62	1.05	1.50	1.04	1.45	1.03
	200	1.71	1.03	1.66	1.03	1.52	1.02	1.48	1.02
	400	1.64	1.01	1.61	1.01	1.46	1.01	1.43	1.01
1d	25	3.81	1.02	3.52	1.05	2.24	1.04	2.18	1.07
	50	5.35	1.00	5.04	1.04	3.00	1.03	2.87	1.03
	100	4.72	1.01	4.38	1.04	3.12	1.02	2.93	1.02
	200	5.44	1.01	5.03	1.04	3.37	1.01	3.15	1.01
	400	4.90	1.01	4.59	1.02	3.08	1.01	2.91	1.01
2a	25	3.01	1.22	2.57	1.17	1.73	1.08	1.63	1.07
	50	4.07	1.34	3.27	1.33	2.28	1.11	1.99	1.09
	100	6.21	1.31	4.87	1.32	2.50	1.06	2.23	1.04
	200	5.01	1.37	3.48	1.22	2.19	1.06	1.85	1.02
	400	4.67	1.37	3.24	1.20	2.21	1.07	1.86	1.02
2b	25	4.98	1.07	4.60	1.11	2.43	1.11	2.28	1.11
	50	7.96	1.03	7.47	1.22	4.20	1.09	3.87	1.14
	100	11.88	1.23	10.32	1.27	4.51	1.18	3.96	1.11
	200	12.57	1.43	9.96	1.48	4.95	1.29	3.95	1.21
	400	13.97	1.51	10.63	1.54	5.16	1.28	3.95	1.17
3a	25	1.51	1.18	1.39	1.11	1.37	1.11	1.32	1.08
	50	1.42	1.11	1.39	1.09	1.38	1.09	1.36	1.08
	100	1.27	1.04	1.27	1.04	1.26	1.04	1.26	1.03
	200	1.24	1.02	1.25	1.02	1.24	1.02	1.26	1.02
	400	1.24	1.01	1.25	1.01	1.25	1.01	1.26	1.01
3b	25	1.87	1.14	1.74	1.09	1.63	1.10	1.60	1.08
	50	1.85	1.10	1.84	1.09	1.72	1.07	1.76	1.08
	100	1.62	1.04	1.62	1.03	1.57	1.04	1.58	1.03
	200	1.62	1.02	1.65	1.02	1.58	1.02	1.62	1.02
	400	1.64	1.02	1.68	1.01	1.60	1.01	1.65	1.01
4a	25	1.38	1.18	1.27	1.11	1.23	1.09	1.19	1.07
	50	1.27	1.09	1.24	1.08	1.22	1.08	1.20	1.07
	100	1.18	1.04	1.17	1.03	1.15	1.03	1.15	1.03
	200	1.13	1.02	1.13	1.01	1.11	1.02	1.11	1.02
	400	1.12	1.02	1.12	1.01	1.10	1.01	1.10	1.01
4b	25	1.76	1.10	1.65	1.07	1.43	1.06	1.39	1.06
	50	1.76	1.08	1.72	1.06	1.56	1.07	1.53	1.06
	100	1.64	1.04	1.64	1.03	1.50	1.03	1.49	1.03
	200	1.54	1.04	1.55	1.02	1.40	1.03	1.40	1.02
	400	1.52	1.04	1.54	1.02	1.39	1.03	1.38	1.02

Table 2: Ratios of VarAS of OLS and WLS1 with that of MWLS1 (resp., of OLS and WLS2 with that of MWLS2) are listed under "Model 1 for $\omega^2(X;\gamma)$" (resp., "Model 2 for $\omega^2(X;\gamma)$"). Asymptotic variances of WLS1 and ALS1 (resp., WLS2 and ALS2) are identical, and so are their VarAS's. So, the ratios of their VarAS to that of MWLS1 (resp., MWLS2) are presented together.

DGP	n	OLS	WLS1	ALS1	MWLS1	WLS2	ALS2	MWLS2
	25	10.3	12.2	12.3	18.3	11.7	11.7	15.2
	50	6.9	8.0	8.0	10.0	7.7	7.7	9.2
1a	100	5.8	6.1	6.1	7.2	6.0	6.0	6.8
	200	5.3	5.7	5.7	5.9	5.5	5.5	5.7
	400	5.3	5.4	5.4	5.4	5.3	5.3	5.4
	25	8.2	12.0	12.3	17.7	11.3	11.5	14.4
	50	6.5	8.3	8.5	10.4	7.8	8.0	9.5
1b	100	5.5	6.4	6.5	7.5	6.2	6.3	6.9
	200	5.2	5.9	5.9	6.3	5.7	5.7	6.0
	400	5.0	5.1	5.1	5.3	5.1	5.1	5.2
	25	6.5	11.6	12.1	16.0	10.5	10.9	13.0
	50	6.2	8.4	8.6	10.0	7.7	7.9	9.0
1c	100	5.8	6.9	6.9	8.1	6.4	6.4	7.2
	200	5.3	5.8	5.8	6.4	5.5	5.5	5.9
	400	5.5	5.4	5.4	5.6	5.4	5.4	5.5
	25	6.1	11.0	11.0	13.6	9.2	9.2	11.2
	50	7.0	8.4	8.4	9.9	7.2	7.2	8.5
1d	100	6.1	7.3	7.3	9.0	6.2	6.2	7.5
	200	5.3	6.0	6.0	7.0	5.5	5.5	6.0
	400	5.1	5.6	5.6	5.8	5.3	5.3	5.4
	25	5.6	7.5	7.6	10.2	7.7	7.7	9.5
	50	6.7	6.6	6.6	7.9	6.3	6.3	7.2
2a	100	5.6	6.0	6.0	6.4	5.8	5.8	5.9
	200	5.4	5.5	5.5	5.3	5.5	5.5	5.1
	400	5.3	5.5	5.5	5.5	5.4	5.4	5.4
	25	6.0	7.9	7.9	10.5	8.1	8.1	10.4
	50	8.0	6.6	6.6	8.8	6.2	6.2	8.1
2b	100	5.8	6.6	6.6	7.5	6.2	6.2	6.8
	200	5.3	5.4	5.4	6.1	5.3	5.3	5.6
	400	5.3	5.3	5.3	5.4	5.3	5.3	5.1
	25	8.1	12.8	13.5	18.4	11.8	12.5	15.0
	50	7.0	9.6	10.2	12.2	8.9	9.4	10.8
3a	100	5.8	7.3	7.5	8.0	7.0	7.1	7.6
	200	5.6	5.8	5.8	6.2	5.7	5.7	6.1
	400	5.4	5.8	5.8	6.0	5.7	5.7	5.9
	25	7.3	12.7	13.3	17.4	11.4	12.1	14.1
	50	7.5	10.3	11.3	12.9	9.0	10.0	11.1
3b	100	5.9	7.3	7.4	8.2	6.9	7.0	7.6
	200	5.2	5.8	5.8	6.2	5.7	5.7	6.0
	400	5.2	5.3	5.3	5.6	5.3	5.3	5.4
	25	8.5	12.5	12.8	17.9	11.7	11.9	14.4
	50	6.7	9.5	9.5	11.5	8.8	8.8	10.2
4a	100	5.8	7.2	7.3	7.8	7.0	7.0	7.5
	200	5.6	5.7	5.7	6.1	5.7	5.7	5.9
	400	5.4	5.8	5.8	5.9	5.7	5.7	5.7
	25	6.8	12.2	13.1	15.5	10.8	11.6	12.5
	50	7.1	10.5	11.0	12.0	9.0	9.5	10.2
4b	100	5.9	7.2	7.3	7.9	6.9	7.0	7.3
	200	5.1	5.9	5.9	6.0	5.6	5.6	5.8
	400	5.3	5.5	5.5	5.5	5.3	5.3	5.4

Table 3: Empirical size of a nominal 5% two-sided test for β_1 using the $N(0,1)$ critical value, and the t-ratio based on the OLS, WLS1, ALS1, MWLS1, WLS2, ALS2 and MWLS2 estimators for β_1.

DGP	n	OLS	WLS1	ALS1	MWLS1	WLS2	ALS2	MWLS2
	25	9.2	10.6	10.6	13.3	10.3	10.3	11.9
	50	6.5	7.4	7.4	8.5	7.3	7.3	8.2
1a	100	5.9	6.2	6.2	6.9	6.2	6.2	6.7
	200	5.3	5.5	5.5	5.8	5.6	5.6	5.8
	400	5.3	5.4	5.4	5.4	5.3	5.3	5.4
	25	7.8	9.3	9.5	11.4	9.1	9.3	10.4
	50	6.7	7.5	7.6	8.5	7.3	7.4	8.1
1b	100	5.6	6.2	6.3	6.7	6.1	6.2	6.6
	200	5.6	5.9	5.9	6.3	5.8	5.8	6.1
	400	5.2	5.2	5.2	5.2	5.2	5.2	5.2
	25	7.6	9.1	9.3	10.6	8.7	8.9	10.1
	50	6.9	6.9	7.1	8.0	6.7	6.9	7.7
1c	100	6.0	6.3	6.3	7.0	6.1	6.1	6.7
	200	5.3	5.6	5.6	5.9	5.4	5.4	5.7
	400	5.6	5.3	5.3	5.4	5.1	5.1	5.3
	25	8.0	9.6	9.6	11.2	8.9	8.9	10.8
	50	7.5	7.8	7.8	8.6	7.1	7.1	8.1
1d	100	6.4	6.6	6.6	7.4	6.0	6.0	6.6
	200	5.6	5.7	5.7	6.2	5.5	5.5	5.9
	400	5.4	5.5	5.5	5.6	5.4	5.4	5.5
	25	7.4	7.9	7.9	9.6	8.0	8.0	9.5
	50	7.6	7.1	7.1	8.2	7.0	7.0	8.4
2a	100	6.1	6.2	6.2	6.3	6.1	6.1	6.2
	200	5.5	5.6	5.6	5.6	5.6	5.6	5.7
	400	5.4	5.5	5.5	5.5	5.5	5.5	5.5
	25	8.1	8.5	8.5	11.2	8.7	8.7	11.3
	50	8.3	7.2	7.2	8.9	7.0	7.0	8.6
2b	100	6.1	6.5	6.5	7.1	6.3	6.3	6.7
	200	5.5	5.5	5.5	5.7	5.4	5.4	5.7
	400	5.4	5.2	5.2	5.2	5.2	5.2	5.2
	25	7.9	9.9	10.4	12.6	9.6	10.1	11.2
	50	7.5	8.9	9.4	10.4	8.4	8.9	9.9
3a	100	6.0	6.8	7.0	7.2	6.6	6.8	6.9
	200	5.6	5.8	5.8	6.1	5.7	5.7	5.9
	400	5.5	5.8	5.8	6.0	5.6	5.6	5.8
	25	7.9	9.6	10.2	11.8	9.1	9.8	10.7
	50	8.0	9.1	10.0	10.7	8.4	9.4	9.7
3b	100	6.1	6.6	6.6	7.1	6.3	6.4	6.7
	200	5.3	5.5	5.5	5.8	5.5	5.5	5.8
	400	5.3	5.6	5.6	5.7	5.5	5.5	5.6
	25	7.9	9.8	9.9	11.9	9.5	9.6	10.8
	50	7.1	8.6	8.6	9.7	8.3	8.3	9.5
4a	100	5.9	6.7	6.7	7.0	6.5	6.5	6.9
	200	5.5	5.7	5.7	5.9	5.6	5.6	5.9
	400	5.4	5.8	5.8	5.8	5.7	5.7	5.8
	25	7.7	9.2	9.9	10.6	8.8	9.5	9.9
	50	7.7	8.9	9.3	10.0	8.0	8.4	9.3
4b	100	6.0	6.5	6.5	6.8	6.3	6.3	6.7
	200	5.3	5.6	5.6	5.8	5.5	5.5	5.6
	400	5.4	5.5	5.5	5.7	5.5	5.5	5.6

Table 4: Empirical size of a nominal 5% two-sided test for β_2 using the $N(0,1)$ critical value, and the t-ratio based on the OLS, WLS1, ALS1, MWLS1, WLS2, ALS2 and MWLS2 estimators for β_2.

A Appendix A: Proofs for the results in Section 2

Proof of Lemma 1:

(i) Follows by inspection.

(ii) Follows by assumptions (A1)-(A6), continuity of inverse (since it exists), and Slutsky's theorem once we note using (1) that for any given $\gamma \in \Gamma$:

$$\sqrt{n}\left(\widehat{\beta}_n(\gamma) - \beta_0\right) = \left(\frac{1}{n}\sum_{i=1}^{n}\frac{1}{\omega^2(X_i;\gamma)}X_iX_i'\right)^{-1}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{1}{\omega^2(X_i;\gamma)}X_iu_i\right).$$

(iii) Follows from (ii) under assumption (A7) by using the delta method. ■

Proof of Lemma 2: First, note that assumptions (A6) and (A7) imply that $\Sigma(\gamma)$ is a finite, positive definite matrix for $\gamma \in \Gamma$, and assumption (A8) implies that Γ^* defined in (9) is nonempty. Now, define the set $\bar{\Gamma}$ as the right hand side of (12):

$$\bar{\Gamma} := \left\{\gamma \in \Gamma \,\middle|\, \text{Trace}(\Sigma(\gamma)) - \min_{g\in\Gamma}\text{Trace}(\Sigma(g)) = 0\right\}.$$

Recall that $\bar{\Gamma}$ is nonempty because $\arg\min_{g\in\Gamma}\text{Trace}(\Sigma(g))$ exists in Γ as trace is a scalar criterion and as Γ is compact in $\mathbb{R}^k$. Now, we show that $\bar{\Gamma} = \Gamma^*$ where Γ^* is as defined in (9).

Step 1 [$\gamma^* \in \Gamma^* \Rightarrow \gamma^* \in \bar{\Gamma}$]: Take any $\gamma^* \in \Gamma^*$. This is possible by assumption (A8). We show that $\gamma^* \in \bar{\Gamma}$. Suppose that this is not true. Then there exists $\bar{\gamma} \in \bar{\Gamma}$ such that $\text{Trace}(\Sigma(\gamma^*) - \Sigma(\bar{\gamma})) > 0$. That implies that at least one diagonal element of $\Sigma(\gamma^*)$ is strictly greater than the corresponding diagonal element of $\Sigma(\bar{\gamma})$. This implies that neither is $\Sigma(\bar{\gamma}) - \Sigma(\gamma^*)$ positive definite (i.e., $\bar{\gamma} \notin \Gamma_{\text{pd}}(\gamma^*)$ defined in (10)) nor is $\Sigma(\bar{\gamma}) = \Sigma(\gamma^*)$ (i.e., $\bar{\gamma} \notin \Gamma_{\text{eq}}(\gamma^*)$ defined in (11)). Since $\bar{\gamma} \in \Gamma$, the above implies that $\Gamma_{\text{pd}}(\gamma^*) \cup \Gamma_{\text{eq}}(\gamma^*) \neq \Gamma$ and hence $\gamma^* \notin \Gamma^*$ by (9), which is contradiction. Hence, $\gamma^* \in \bar{\Gamma}$.

Step 2 [$\bar{\gamma} \in \bar{\Gamma} \Rightarrow \bar{\gamma} \in \Gamma^*$]: Take any $\bar{\gamma} \in \bar{\Gamma}$, which is equivalent to the condition that: $\text{Trace}(\Sigma(\gamma) - \Sigma(\bar{\gamma})) \geq 0$ for all $\gamma \in \Gamma$. We show that $\bar{\gamma} \in \Gamma^*$ where Γ^* is defined in (9). Suppose that this is not true. Then, if we consider any $\gamma^* \in \Gamma^*$ (nonempty by (A8)), any one of the following two mutually exclusive conditions has to hold: (i) $\bar{\gamma} \in \Gamma_{\text{pd}}(\gamma^*)$ or (ii) $\bar{\gamma} \in \Gamma_{\text{eq}}(\gamma^*)$. If (i) holds then $\text{Trace}(\Sigma(\bar{\gamma}) - \Sigma(\gamma^*)) > 0$ which is a contradiction to the condition on the trace stated at the top of Step 2, and hence is not possible. On the other hand, if (ii) holds then $\Sigma(\bar{\gamma}) = \Sigma(\gamma^*)$. Now, recalling that $\gamma^* \in \Gamma$ is equivalent to saying $\Gamma_{\text{pd}}(\gamma^*) \cup \Gamma_{\text{eq}}(\gamma^*) = \Gamma$, the implication of (ii) (i.e., $\Sigma(\bar{\gamma}) = \Sigma(\gamma^*)$) is that $\Gamma_{\text{pd}}(\bar{\gamma}) \cup \Gamma_{\text{eq}}(\bar{\gamma}) = \Gamma$, which would imply that $\bar{\gamma} \in \Gamma^*$ by (9). This is a contradiction to our supposition. Hence, $\bar{\gamma} \in \Gamma^*$. ■

Proof of Lemma 3:

(i) Follows trivially.

(ii) Take any $\delta > 0$. Assumption (A11) implies that $P(d(\widehat{\gamma}_n, \Gamma^*) > \delta) \leq P(Q(\widehat{\gamma}_n) \geq \epsilon(\delta))$. However, for any $\epsilon(\delta) > 0$, we have $P(Q(\widehat{\gamma}_n) \geq \epsilon(\delta)) \to 0$ because: (a) $|Q(\widehat{\gamma}_n)| \leq |Q(\widehat{\gamma}_n) - \widehat{Q}_n(\widehat{\gamma}_n)| + |\widehat{Q}_n(\widehat{\gamma}_n)|$ by the triangle inequality, and (b) since $\widehat{\gamma}_n \in \Gamma$, the first term on the RHS in (a) is $o_p(1)$ by assumption (A10) while the second term on the RHS in (a) is identically 0 by its definition in (15). Hence, $d(\widehat{\gamma}_n, \Gamma^*) = o_p(1)$. This implies that $\widehat{\gamma}_n \in \Gamma^*_\delta$ with probability approaching one. Hence, we can use assumption (A12), and the steps (a) and (b) in the consistency proof above to obtain that $|Q(\widehat{\gamma}_n)| = O_p(1/\sqrt{n})$. Then, using assumption (A13), it follows that $d(\widehat{\gamma}_n, \Gamma^*) \leq Q(\widehat{\gamma}_n)/D = O_p(1/\sqrt{n})$. ■

Proof of Theorem 4: The proof proceeds by obtaining the following four intermediate results.

First, note that, adding and subtracting terms to match assumption (A14) gives:

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{X_i u_i}{\omega^2(X_i; \widehat{\gamma}_n)} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{X_i u_i}{\omega^2(X_i; \gamma_n)} + R_{1,n} + R_{2,n}$$

where

$$R_{1,n} := \left(\frac{1}{n} \sum_{i=1}^{n} X_i u_i \nabla_\gamma(X_i; \gamma_n)' \right) \sqrt{n}(\widehat{\gamma}_n - \gamma_n) = \left(\frac{1}{n} \sum_{i=1}^{n} W_{i,n} \right) \sqrt{n}(\widehat{\gamma}_n - \gamma_n) = o_p(1)$$

by assumption (A15) and Lemma 3 respectively; and

$$\begin{aligned}
R_{2,n} &:= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i u_i \left[\frac{1}{\omega^2(X_i; \widehat{\gamma}_n)} - \frac{1}{\omega^2(X_i; \gamma_n)} - \nabla_\gamma(X_i; \gamma_n)'(\widehat{\gamma}_n - \gamma_n) \right] \\
\Rightarrow |R_{2,n}| &\leq \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \|X_i u_i\| \times \left| \frac{1}{\omega^2(X_i; \widehat{\gamma}_n)} - \frac{1}{\omega^2(X_i; \gamma_n)} - \nabla_\gamma(X_i; \gamma_n)'(\widehat{\gamma}_n - \gamma_n) \right| \\
&\leq \frac{1}{2\sqrt{n}} \sum_{i=1}^{n} \|X_i u_i\| \times |\Delta(X; \gamma_n)| \times \|\widehat{\gamma}_n - \gamma_n\|^2 \\
&\leq \left(\frac{1}{2n} \sum_{i=1}^{n} \|X_i u_i \Delta(X; \gamma_n)\| \right) \left(n^{1/4} \|\widehat{\gamma}_n - \gamma_n\| \right)^2 = o_p(1)
\end{aligned}$$

where the first inequality follows by the Cauchy-Schwartz inequality, the second inequality by assumption (A14), the third inequality by noting that $\Delta(X; \gamma_n)$ is a positive scalar random variable, and the last equality by assumption (A16) and Lemma 3 respectively. Therefore, we obtain that:

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{X_i u_i}{\omega^2(X_i; \widehat{\gamma}_n)} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{X_i u_i}{\omega^2(X_i; \gamma_n)} + o_p(1). \tag{29}$$

Second, note that, by using assumptions (A4) and (A17) respectively in the inequality on the second line below, it follows that:

$$\begin{aligned}\left\|\widehat{B}_n(\widehat{\gamma}_n) - B(\gamma_n)\right\| &\leq \left\|\widehat{B}_n(\widehat{\gamma}_n) - B(\widehat{\gamma}_n)\right\| + \|B(\widehat{\gamma}_n) - B(\gamma_n)\| \\ &\leq \sup_{\gamma\in\Gamma}\left\|\widehat{B}_n(\gamma) - B(\gamma)\right\| + \|B(\widehat{\gamma}_n) - B(\gamma_n)\| = o_p(1). \end{aligned} \tag{30}$$

Third, note that:

$$\begin{aligned}\sqrt{n}\left(\widehat{\beta}_n - \beta_0\right) &= \widehat{B}_n(\widehat{\gamma}_n)^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{X_i u_i}{\omega^2(X_i;\widehat{\gamma}_n)} = \left[B(\gamma_n)^{-1} + o_p(1)\right]\left[\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{X_i u_i}{\omega^2(X_i;\gamma_n)} + o_p(1)\right] \\ &= B(\gamma_n)^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{X_i u_i}{\omega^2(X_i;\gamma_n)} + o_p(1). \end{aligned} \tag{31}$$

In the second equality on the first line, the first bracket follows by (30) and assumption (A6) while the second bracket follows by (29). The second line follows by assumptions (A6) and (A19)(a).

Fourth, note that, by using (31) and assumption (A7), it follows that:

$$H(\widehat{\beta}_n) - H(\beta_0) \equiv H(\widehat{\beta}_n) - H = o_p(1). \tag{32}$$

Therefore, from (18), (29), (30), (31), (32) and assumptions (A6) and (A7), we obtain by a mean-value expansion around a (element-by-element) mean-value $\bar{\beta}_n$ that:

$$\begin{aligned}\sqrt{n}(\widehat{h}_n - h(\beta_0)) &\equiv \sqrt{n}(h(\widehat{\beta}_n) - h(\beta_0)) \\ &= H(\bar{\beta}_n)\sqrt{n}\left(\widehat{\beta}_n - \beta_0\right) \\ &= [H + o_p(1)]\left[B(\gamma_n)^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{X_i u_i}{\omega^2(X_i;\gamma_n)} + o_p(1)\right] \\ &= HB(\gamma_n)^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{X_i u_i}{\omega^2(X_i;\gamma_n)} + o_p(1) \\ &= \Sigma(\gamma^*)^{1/2}\Sigma(\gamma_n)^{-1/2}HB(\gamma_n)^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{X_i u_i}{\omega^2(X_i;\gamma_n)} + o_p(1) \\ &\xrightarrow{d} N(0,\Sigma(\gamma^*)) \end{aligned} \tag{33}$$

for any $\gamma^* \in \Gamma^*$. The first equality follows by a mean-value expansion around β_0 and with the element-by-element mean-value $\bar{\beta}_n$. The second equality follows by (31) and (32). The third equality follows by assumptions (A7) and (A19)(a). The fourth equality follows by assumptions (A17) and (A18) by virtue of assumptions (A6) and (A19b). The last line in the display follows by

assumption (A19)(b) and the definition of Γ^* in (9). This concludes the proof of Theorem 4(ii).

The special case in part (i) of Theorem 4 follows from (33) and Lemma 1(iii). ■

Proof of Lemma 5:

(i) Note that:

$$\|\widehat{C}_n(\gamma|\check{\beta}_n) - C(\gamma)\| = \|T_{1n}(\gamma) + T_{2n}(\gamma)\| \leq \|T_{1n}(\gamma)\| + \|T_{2n}(\gamma)\|$$

where:

$$\begin{aligned} T_{1n}(\gamma) &= \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{X_i X_i' u_i^2}{(\omega^2(X_i;\gamma))^2} - E\left[\frac{X_i X_i' u_i^2}{(\omega^2(X_i;\gamma))^2}\right]\right\}, \\ T_{2n}(\gamma) &= \frac{1}{n}\sum_{i=1}^{n}\frac{X_i X_i'}{(\omega^2(X_i;\gamma))^2}\left\{(y_i - X_i'\check{\beta}_n)^2 - u_i^2\right\}. \end{aligned}$$

Now, note that, uniformly in $\gamma \in \Gamma$,

$$\|T_{1n}(\gamma)\| = \left\|\frac{1}{n}\sum_{i=1}^{n}\left\{\frac{X_i X_i' u_i^2}{(\omega^2(X_i;\gamma))^2} - E\left[\frac{X_i X_i' u_i^2}{(\omega^2(X_i;\gamma))^2}\right]\right\}\right\| = o_p(1) \tag{34}$$

since the compactness of Γ, and assumptions (A21) and (A22) imply that Jenrich (1969)'s uniform weak law of large numbers (UWLLN) applies. On the other hand, letting $X_i \equiv (X_{1,i},\ldots,X_{p,i})'$, $\beta_0 \equiv (\beta_{0,1},\ldots,\beta_{0,p})'$ and $\check{\beta}_n \equiv (\check{\beta}_{n,1},\ldots,\check{\beta}_{n,p})'$, it follows that:

$$\begin{aligned} \|T_{2n}(\gamma)\| &= \left\|\frac{1}{n}\sum_{i=1}^{n}\frac{X_i X_i'}{(\omega^2(X_i;\gamma))^2}\left\{(y_i - X_i'\check{\beta}_n)^2 - u_i^2\right\}\right\| \\ &= \left\|\frac{1}{n}\sum_{i=1}^{n}\frac{X_i X_i'}{(\omega^2(X_i;\gamma))^2}\left\{(\check{\beta}_n - \beta_0)' X_i X_i' (\check{\beta}_n - \beta_0) - 2(\check{\beta}_n - \beta_0)' X_i u_i\right\}\right\| \\ &= \left\|\sum_{j=1}^{p}\sum_{l=1}^{p}(\check{\beta}_{n,j} - \beta_{0,j})(\check{\beta}_{n,l} - \beta_{0,l})\frac{1}{n}\sum_{i=1}^{n}\frac{X_i X_i' X_{j,i} X_{l,i}}{(\omega^2(X_i;\gamma))^2} - 2\sum_{j=1}^{p}(\check{\beta}_{n,j} - \beta_{0,j})\frac{1}{n}\sum_{i=1}^{n}\frac{X_i X_i' X_{j,i} u_i}{(\omega^2(X_i;\gamma))^2}\right\| \\ &\leq \sum_{j=1}^{p}\sum_{l=1}^{p}|\check{\beta}_{n,j} - \beta_{0,j}||\check{\beta}_{n,l} - \beta_{0,l}|\left\|\frac{1}{n}\sum_{i=1}^{n}\left\{\frac{X_i X_i' X_{j,i} X_{l,i}}{(\omega^2(X_i;\gamma))^2} - E\left[\frac{X_i X_i' X_{j,i} X_{l,i}}{(\omega^2(X_i;\gamma))^2}\right]\right\}\right\| \\ &\quad + \sum_{j=1}^{p}\sum_{l=1}^{p}|\check{\beta}_{n,j} - \beta_{0,j}||\check{\beta}_{n,l} - \beta_{0,l}|\left\|E\left[\frac{X_i X_i' X_{j,i} X_{l,i}}{(\omega^2(X_i;\gamma))^2}\right]\right\| \\ &\quad + 2\sum_{j=1}^{p}|\check{\beta}_{n,j} - \beta_{0,j}|\left\|\frac{1}{n}\sum_{i=1}^{n}\left\{\frac{X_i X_i' X_{j,i} u_i}{(\omega^2(X_i;\gamma))^2} - E\left[\frac{X_i X_i' X_{j,i} u_i}{(\omega^2(X_i;\gamma))^2}\right]\right\}\right\| \\ &\quad + 2\sum_{j=1}^{p}|\check{\beta}_{n,j} - \beta_{0,j}|\left\|E\left[\frac{X_i X_i' X_{j,i} u_i}{(\omega^2(X_i;\gamma))^2}\right]\right\| \\ &= o_p(1) + o_p(1) + o_p(1) + o_p(1) \quad \text{uniformly in } \gamma \in \Gamma, \end{aligned} \tag{35}$$

where, on the RHS of the last inequality, (i) the first term is $o_p(1)$ by assumption (A20) and then applying the UWLLN that follows from the compactness of Γ, assumptions (A21)(a) and (A22); (ii) the second term is $o_p(1)$ by assumptions (A20) and (A21)(a); (iii) the third term is $o_p(1)$ by assumption (A20) and then applying the UWLLN that follows from the compactness of Γ, and assumptions and (A21) and (A22); and (iv) the fourth term is $o_p(1)$ by assumptions (A20) and (A21). (The integrable dominating function in the cases of all these four terms was obtained by the Holder's inequality with the help of assumption (A21).) Therefore, (34) and (35) jointly imply that: (a) $\sup_{\gamma\in\Gamma} \|\widehat{C}_n(\gamma|\check{\beta}_n) - C(\gamma)\| = o_p(1)$. Then, applying assumptions (A7) and (A20), it follows that: (b) $H(\check{\beta}_n) - H = o_p(1)$ irrespective of γ, and hence uniformly in $\gamma \in \Gamma$. Therefore, combining (a), (b) and assumption (A4), it follows that $\sup_{\gamma\in\Gamma} \|\widehat{\Sigma}_n(\gamma|\check{\beta}_n) - \Sigma(\gamma)\| = o_p(1)$.

(ii) Note that:

$$\begin{aligned}
\|\widehat{\Sigma}_n(\check{\gamma}_n|\check{\beta}_n) - \Sigma(\gamma^*)\| &\leq \|\widehat{\Sigma}_n(\check{\gamma}_n|\check{\beta}_n) - \Sigma(\check{\gamma}_n)\| + \|\Sigma(\check{\gamma}_n) - \Sigma(\gamma_n)\| + \|\Sigma(\gamma_n) - \Sigma(\gamma^*)\| \\
&\leq \sup_{\gamma\in\Gamma} \|\widehat{\Sigma}_n(\gamma|\check{\beta}_n) - \Sigma(\gamma)\| + \|\Sigma(\check{\gamma}_n) - \Sigma(\gamma_n)\| + \|\Sigma(\gamma_n) - \Sigma(\gamma^*)\| \\
&= o_p(1) + o_p(1) + o_p(1)
\end{aligned}$$

where the first $o_p(1)$ follows by Lemma 5(i), the second $o_p(1)$ term by assumptions (A7), (A17), (A18) and (A23), while the third $o_p(1)$ term by (A7), (A17) and (A18). ■

B Appendix B: Supplementary materials

B.1 The problem with convexity of $m(\gamma)$ in γ mentioned in footnote 5

Without further assumption on the parameterization $\omega^2(X;\gamma)$, it is not possible to gainfully exploit the simplicity of the trace function to establish properties such as convexity for $m(\gamma)$. In fact, convexity may not hold in general. To simplify the exposition, instead of the minimization $m(\gamma) := \text{Trace}\left((B(\gamma)^{-1}C(\gamma)B(\gamma)^{-1}\right) \equiv \text{Trace}\left(C(\gamma)\right) / \left[\text{Trace}\left(B(\gamma)\right)\right]^2$ with respect to γ, use assumption (A6) and consider the minimization of $\log(m(\gamma)) \equiv \log\left(\text{Trace}\left(C(\gamma)\right)\right) - 2\log\left(\text{Trace}\left(B(\gamma)\right)\right)$. Then, note that, strict convexity of $\log(m(\gamma))$ in Γ follows if $\sum_{j=1}^{p}[a_j(\gamma) - b_j(\gamma)] > \sum_{j=1}^{p}[c_j(\gamma) - d_j(\gamma)]$ for $\gamma \in \Gamma$ where $a_j(\gamma) = (\nabla_{\gamma\gamma} f_j(\gamma))/\text{Trace}(C(\gamma))$, $b_j(\gamma) = (\nabla_\gamma f_j(\gamma))(\nabla_\gamma f_j(\gamma))'/\text{Trace}^2(C(\gamma))$, $c_j(\gamma) = (\nabla_{\gamma\gamma} g_j(\gamma))/\text{Trace}(B(\gamma))$, $d_j(\gamma) = (\nabla_\gamma g_j(\gamma))(\nabla_\gamma g_j(\gamma))'/\text{Trace}^2(B(\gamma))$ and, for $j = 1, \ldots, p$, $\nabla_\gamma f_j(\gamma)$ and $\nabla_{\gamma\gamma} f_j(\gamma)$ (respectively, $\nabla_\gamma g_j(\gamma)$ and $\nabla_{\gamma\gamma} g_j(\gamma)$) are the first and the second derivative of the j-th diagonal element of $C(\gamma)$ (respectively $B(\gamma)$). Such convexity may not hold even with a scalar β and γ, i.e., $p = k = 1$, without a very restrictive parameterization $\omega^2(X;\gamma)$.

B.2 Steps leading to equations (26) in Section 3.4

Denoting $\frac{\partial}{\partial\gamma}\omega^2(X;\gamma^*)$ by $\nabla_\gamma(X;\gamma^*)$, (26) follows from the first order condition because:

$$
\begin{aligned}
& \frac{\partial}{\partial\gamma}\log(B(\gamma^*)) = \frac{1}{2}\frac{\partial}{\partial\gamma}\log(C(\gamma^*)) \\
\Rightarrow \quad & E\left[\frac{X^2}{\omega^2(X;\gamma^*)}\frac{\nabla_\gamma(X;\gamma^*)}{\omega^2(X;\gamma^*)}\right] \Big/ B(\gamma) = E\left[\frac{u^2}{\omega^2(X;\gamma^*)}\frac{X^2}{\omega^2(X;\gamma^*)}\frac{\nabla_\gamma(X;\gamma^*)}{\omega^2(X;\gamma^*)}\right] \Big/ C(\gamma) \\
\Rightarrow \quad & 0 = \frac{1}{C(\gamma)}E\left[\frac{X^2}{\omega^2(X;\gamma^*)}\frac{\nabla_\gamma(X;\gamma^*)}{\omega^2(X;\gamma^*)}\left\{\frac{u^2}{\omega^2(X;\gamma^*)} - \frac{C(\gamma)}{B(\gamma)}\right\}\right] \\
\Rightarrow \quad & 0 = E\left[\frac{X^2}{\omega^2(X;\gamma^*)}\frac{\nabla_\gamma(X;\gamma^*)}{\omega^2(X;\gamma^*)}\left\{\frac{u^2}{\omega^2(X;\gamma^*)} - \frac{C(\gamma)}{B(\gamma)}\right\}\right] \\
\Rightarrow \quad & 0 = E\left[\frac{1}{\omega^2(X;\gamma^*)}\frac{X^2}{\omega^2(X;\gamma^*)}\frac{\nabla_\gamma(X;\gamma^*)}{\omega^2(X;\gamma^*)}\left\{u^2 - \frac{C(\gamma)}{B(\gamma)}\omega^2(X;\gamma^*)\right\}\right] \\
\Rightarrow \quad & 0 = E\left[\frac{1}{\omega^2(X;\gamma^*)}\frac{X^2}{\omega^2(X;\gamma^*)}\frac{\nabla_\gamma(X;\gamma^*)}{\omega^2(X;\gamma^*)}\left\{\omega_0^2(X) - \frac{C(\gamma)}{B(\gamma)}\omega^2(X;\gamma^*)\right\}\right]
\end{aligned}
$$

where the third line follows by rearranging terms, the fourth line by using that $C(\gamma) \neq 0$, the fifth line by factoring out $\frac{1}{\omega^2(X;\gamma^*)}$ from the terms inside the braces to make the tilting of the term $\omega^2(X;\gamma^*)$ inside the braces explicit, and the sixth line by the law of iterated expectations.

B.3 Implementation of WLS, ALS and MWLS estimators

The WLS and ALS estimators of β are obtained as in Romano and Wolf (2017) (RW). In particular, first the estimators $\widehat{\gamma}_{\text{RW1}}$ and $\widehat{\gamma}_{\text{RW2}}$ of $\gamma = (\gamma_1, \gamma_2)'$ in Models 1 and 2 respectively are obtained as:

- $\widehat{\gamma}_{\text{RW1}}$ is the coefficient from a regression (Regression 1) of $\log\left(\max(\delta^2, \widehat{u}^2)\right)$ on $(1, \log(X_{(2)}))$;
- $\widehat{\gamma}_{\text{RW2}}$ is the coefficients from a regression (Regression 2) of $\log\left(\max(\delta^2, \widehat{u}^2)\right)$ on $(1, X_{(2)})$;

where $\delta = .1$ and $\widehat{u}$'s are the OLS residuals from (1).

The WLS1 and WLS2 estimators are then obtained respectively as the weighted by $\omega^{-1}(X; \widehat{\gamma}_{\text{RW1}})$ and $\omega^{-1}(X; \widehat{\gamma}_{\text{RW2}})$ least squares estimators of β (see (4)).

The ALS1 and ALS2 estimators are obtained as follows. For ALS1, first a test for conditional homoskedasticity of u is performed by using the $n \times R^2$ from Regression 1 as the test statistic. If the null hypothesis of conditional homoskedasticity is rejected at the 10% level then ALS1 is the WLS1 estimator and otherwise ALS1 is the OLS estimator of β. Exactly in the same way for ALS2, first a test for conditional homoskedasticity of u is performed by using the $n \times R^2$ from Regression 2 as the test statistic. If the null hypothesis of conditional homoskedasticity is rejected at the 10% level then ALS2 is the WLS2 estimator and otherwise ALS2 is the OLS estimator of β.

Finally, consider obtaining the MWLS estimator of β_1 following the Algorithm in Section 2.3.2. (Estimation of and inference on β_2 follow similarly and are omitted.) In Step 1, we use the OLS estimator $\widetilde{\beta}_{n,\text{OLS}}$, the expression for $\widehat{B}_n(\gamma)$, and the formulae in (20) and (21) to obtain for each $\gamma \in \Gamma$ the scalar function of γ:

$$\widehat{\Sigma}_n(\gamma) \equiv \widehat{\Sigma}_n(\gamma|\widetilde{\beta}_{n,\text{OLS}}) = \text{the } (1,1)\text{-th element of the } 2 \times 2 \text{ matrix } \widehat{\Xi}_n(\gamma|\widetilde{\beta}_{n,\text{OLS}}) \text{ where}$$

$$\widehat{\Xi}_n(\gamma|\widetilde{\beta}_{n,\text{OLS}}) = \left(\frac{1}{n}\sum_{i=1}^n \frac{X_i X_i'}{\omega^2(X_i;\gamma)}\right)^{-1} \left(\frac{1}{n}\sum_{i=1}^n \frac{(y_i - X_i'\widetilde{\beta}_{n,\text{OLS}})_i^2 X_i X_i'}{(\omega^2(X_i;\gamma))^2}\right) \left(\frac{1}{n}\sum_{i=1}^n \frac{X_i X_i'}{\omega^2(X_i;\gamma)}\right)^{-1}$$

for the given model, Model 1 or Model 2, for $\omega^2(X;\gamma)$. If the determinant of $\widehat{\Xi}_n(\gamma|\widetilde{\beta}_{n,\text{OLS}})$ is zero for any γ then we redefine $\widehat{\Sigma}_n(\gamma) = +\infty$ for that γ to ensure non-degeneracy. In Step 2, we obtain $\widehat{\gamma}_n$ as a minimizer of $\widehat{\Sigma}_n(\gamma)$ with respect to $\gamma \in \Gamma$. In Step 3 we obtain the MWLS estimator for β_1 as the first element of the weighted by $\omega^{-1}(X; \widehat{\gamma}_n)$ least squares estimator, say, $\widehat{\beta}_{\text{n, [for } \beta_1]}$ of β.

For inference on β_1 based on this MWLS estimator, we obtain the estimator $\widehat{\Sigma}_n(.)$ of its asymptotic variance similarly as above but now by using $\widehat{\Xi}_n(\widehat{\gamma}_n|\widehat{\beta}_{n,[\text{for } \beta_1]})$ instead of $\widehat{\Xi}_n(\gamma|\widetilde{\beta}_{n,\text{OLS}})$.

B.4 Supplemental Monte Carlo results referred to in Section 4.2

Tables 5-6 report the VarMC of the estimators for β_1 and β_2 respectively. Table 7 reports the VarAS of the estimators for β_1 and β_2. Tables 8-9 report the empirical size of 5% two-sided tests for β_1 and β_2 respectively with t-ratios using the VarMC instead of the estimated asymptotic variance.

DGP	n	OLS	WLS1	ALS1	MWLS1	WLS2	ALS2	MWLS2
	25	10.27	11.00	10.99	11.56	10.82	10.82	11.32
	50	8.74	9.11	9.11	9.42	9.06	9.06	9.34
1a	100	10.28	10.52	10.52	10.74	10.47	10.47	10.63
	200	10.18	10.36	10.36	10.39	10.33	10.33	10.35
	400	8.83	8.89	8.89	8.90	8.88	8.88	8.88
	25	18.12	17.64	17.86	18.23	17.60	17.81	18.01
	50	15.91	14.56	14.87	14.80	14.60	14.88	14.82
1b	100	21.31	19.72	19.96	20.05	19.76	19.98	19.91
	200	20.40	18.38	18.40	18.55	18.41	18.43	18.47
	400	16.79	14.87	14.87	14.86	14.96	14.96	14.91
	25	39.71	29.36	30.30	30.54	29.74	30.59	30.14
	50	37.38	23.37	23.87	23.90	23.89	24.34	24.05
1c	100	55.29	36.80	36.84	37.69	37.44	37.48	37.71
	200	50.69	31.27	31.27	31.62	31.97	31.97	31.95
	400	41.52	25.63	25.63	25.71	26.19	26.19	26.03
	25	351.23	123.78	123.78	141.76	127.94	127.94	136.49
	50	371.55	80.96	80.96	91.70	83.70	83.70	89.30
1d	100	574.57	132.35	132.35	148.85	140.61	140.61	146.12
	200	485.74	95.83	95.83	101.84	103.53	103.53	103.73
	400	464.87	98.53	98.53	99.76	105.36	105.36	104.08
	25	4.04	1.93	1.93	1.85	2.16	2.16	2.02
	50	6.47	2.48	2.48	2.16	2.93	2.93	2.48
2a	100	5.17	1.24	1.24	.95	1.55	1.55	1.14
	200	4.63	1.31	1.31	.95	1.66	1.66	1.34
	400	5.35	1.63	1.63	1.17	2.04	2.04	1.70
	25	5.15	1.39	1.40	1.53	1.53	1.54	1.55
	50	8.88	1.40	1.40	1.75	1.61	1.61	1.65
2b	100	6.49	.80	.80	.71	.90	.90	.73
	200	6.32	.76	.76	.58	.97	.97	.67
	400	7.49	.82	.82	.57	1.09	1.09	.69
	25	21.88	20.01	20.81	20.98	19.78	20.62	20.54
	50	28.56	25.06	26.20	25.64	24.52	25.77	25.08
3a	100	24.58	21.26	21.59	21.50	20.97	21.31	21.13
	200	22.06	18.44	18.45	18.51	18.17	18.18	18.21
	400	24.92	20.83	20.83	20.82	20.49	20.49	20.49
	25	37.89	28.07	29.18	29.91	27.58	28.71	28.78
	50	54.54	36.37	39.28	37.64	34.89	38.03	35.80
3b	100	42.28	29.00	29.31	29.43	28.36	28.69	28.56
	200	40.08	26.13	26.13	26.20	25.32	25.32	25.34
	400	47.06	29.44	29.44	29.45	28.48	28.48	28.51
	25	13.87	13.61	13.75	14.06	13.47	13.62	13.71
	50	17.47	16.95	17.05	17.16	16.65	16.76	16.90
4a	100	16.56	15.45	15.46	15.52	15.24	15.26	15.33
	200	13.82	12.72	12.72	12.69	12.57	12.57	12.57
	400	15.19	14.10	14.10	13.98	13.91	13.91	13.88
	25	23.57	18.04	19.26	18.66	17.83	18.98	17.98
	50	36.12	25.59	26.85	26.05	24.46	25.71	25.12
4b	100	31.85	21.76	21.86	21.78	21.17	21.26	21.22
	200	26.76	18.79	18.79	18.35	18.09	18.09	17.94
	400	30.17	20.85	20.85	20.27	20.07	20.07	19.95

Table 5: VarMC of OLS, WLS1, ALS1, MWLS1, WLS2, ALS2 and MWLS2 estimators for β_1. WLS1, ALS1, MWLS1 (resp. WLS2, ALS2, MWLS2) use Model 1 (resp. Model 2) for $\omega^2(X;\gamma)$.

DGP	n	OLS	WLS1	ALS1	MWLS1	WLS2	ALS2	MWLS2
	25	1.09	1.16	1.16	1.20	1.15	1.15	1.17
	50	1.16	1.21	1.21	1.23	1.21	1.21	1.23
1a	100	1.41	1.44	1.44	1.46	1.44	1.44	1.45
	200	1.45	1.48	1.48	1.48	1.47	1.47	1.47
	400	1.26	1.27	1.27	1.27	1.26	1.26	1.26
	25	2.47	2.42	2.45	2.46	2.43	2.45	2.44
	50	2.78	2.58	2.63	2.62	2.60	2.64	2.63
1b	100	3.62	3.38	3.41	3.40	3.39	3.42	3.40
	200	3.65	3.34	3.35	3.35	3.35	3.35	3.35
	400	3.08	2.78	2.78	2.78	2.79	2.79	2.79
	25	7.08	5.77	5.89	5.87	5.84	5.95	5.92
	50	8.32	5.91	6.00	6.00	6.00	6.08	6.11
1c	100	11.29	8.16	8.17	8.22	8.26	8.27	8.31
	200	10.87	7.44	7.44	7.46	7.55	7.55	7.56
	400	9.53	6.57	6.57	6.58	6.66	6.66	6.67
	25	86.61	47.30	47.30	50.28	47.32	47.32	50.06
	50	102.82	38.91	38.91	40.38	38.59	38.59	40.66
1d	100	140.23	48.16	48.16	50.43	49.54	49.54	50.57
	200	125.68	39.54	39.54	40.55	41.42	41.42	41.67
	400	112.44	37.17	37.17	37.19	38.87	38.87	38.88
	25	.81	.55	.55	.56	.58	.58	.59
	50	1.50	.79	.79	.78	.87	.87	.89
2a	100	1.23	.55	.55	.52	.60	.60	.57
	200	1.17	.58	.58	.54	.65	.65	.65
	400	1.26	.62	.62	.58	.70	.70	.69
	25	1.30	.67	.67	.70	.70	.70	.74
	50	2.29	.66	.66	.71	.71	.71	.73
2b	100	1.76	.50	.50	.44	.52	.52	.48
	200	1.82	.49	.49	.39	.57	.57	.48
	400	2.01	.51	.51	.39	.59	.59	.51
	25	3.47	3.16	3.29	3.23	3.14	3.28	3.17
	50	5.71	4.98	5.20	5.03	4.88	5.13	4.97
3a	100	4.77	4.11	4.18	4.12	4.06	4.13	4.06
	200	4.55	3.78	3.79	3.79	3.73	3.73	3.73
	400	4.94	4.09	4.09	4.09	4.02	4.02	4.02
	25	7.38	5.67	5.86	5.83	5.57	5.77	5.66
	50	12.06	8.24	8.82	8.32	7.89	8.53	7.98
3b	100	9.18	6.39	6.45	6.41	6.24	6.31	6.23
	200	9.48	6.32	6.32	6.30	6.10	6.10	6.07
	400	10.64	6.82	6.82	6.82	6.57	6.57	6.56
	25	1.93	1.91	1.93	1.93	1.91	1.93	1.91
	50	3.35	3.26	3.28	3.30	3.24	3.26	3.31
4a	100	3.09	2.89	2.89	2.89	2.88	2.88	2.88
	200	2.67	2.48	2.48	2.47	2.46	2.46	2.46
	400	2.80	2.62	2.62	2.60	2.61	2.61	2.59
	25	4.40	3.69	3.86	3.75	3.71	3.87	3.75
	50	8.29	6.26	6.50	6.37	6.12	6.36	6.33
4b	100	7.26	5.24	5.26	5.26	5.20	5.22	5.21
	200	6.48	4.87	4.87	4.81	4.82	4.82	4.78
	400	6.92	5.17	5.17	5.09	5.12	5.12	5.08

Table 6: VarMC of OLS, WLS1, ALS1, MWLS1, WLS2, ALS2 and MWLS2 estimators for β_2. WLS1, ALS1, MWLS1 (resp. WLS2, ALS2, MWLS2) use Model 1 (resp. Model 2) for $\omega^2(X;\gamma)$.

		β_1: Intercept					β_2: Slope				
DGP	n	OLS	WLS1	MWLS1	WLS2	MWLS2	OLS	WLS1	MWLS1	WLS2	MWLS2
	25	8.90	8.74	7.26	8.81	7.84	.95	.94	.84	.94	.87
	50	8.32	8.21	7.59	8.24	7.76	1.11	1.10	1.04	1.10	1.05
1a	100	10.05	9.98	9.61	9.99	9.70	1.38	1.37	1.33	1.37	1.33
	200	10.08	10.05	9.87	10.06	9.91	1.44	1.44	1.41	1.44	1.42
	400	8.76	8.75	8.68	8.76	8.70	1.24	1.24	1.23	1.24	1.23
	25	16.53	14.23	11.72	14.57	12.84	2.27	2.01	1.82	2.04	1.89
	50	15.20	12.93	11.85	13.17	12.23	2.65	2.33	2.19	2.36	2.24
1b	100	21.09	18.58	17.67	18.79	18.05	3.57	3.20	3.09	3.23	3.13
	200	20.16	17.63	17.19	17.79	17.43	3.58	3.19	3.13	3.21	3.16
	400	16.75	14.77	14.62	14.87	14.72	3.08	2.76	2.74	2.78	2.76
	25	37.59	23.35	20.14	24.59	22.14	6.66	4.82	4.47	4.97	4.63
	50	35.89	20.44	18.92	21.39	19.83	7.92	5.33	5.07	5.48	5.23
1c	100	54.00	33.89	32.05	35.07	33.42	10.99	7.61	7.33	7.79	7.56
	200	49.92	30.08	29.15	30.95	30.12	10.71	7.20	7.05	7.34	7.23
	400	40.46	24.92	24.64	25.50	25.13	9.30	6.45	6.39	6.54	6.50
	25	336.43	90.13	88.41	100.12	95.71	80.38	37.44	35.84	39.31	36.82
	50	350.25	65.49	65.44	72.19	69.43	95.83	32.97	31.98	34.51	33.43
1d	100	546.24	117.09	115.77	130.04	124.66	133.29	43.48	42.76	46.45	45.43
	200	487.25	90.80	89.57	100.26	96.85	125.35	37.73	37.24	40.29	39.74
	400	464.85	95.79	94.81	103.48	101.38	111.55	36.56	36.26	38.57	38.35
	25	3.85	1.56	1.28	1.76	1.50	.76	.47	.44	.50	.46
	50	6.11	2.02	1.50	2.48	1.87	1.40	.68	.61	.77	.70
2a	100	5.02	1.06	.81	1.36	1.03	1.19	.50	.48	.56	.53
	200	4.55	1.25	.91	1.59	1.31	1.15	.56	.52	.63	.62
	400	5.28	1.55	1.13	1.95	1.63	1.24	.60	.56	.68	.66
	25	4.97	1.07	1.00	1.20	1.08	1.22	.56	.50	.59	.54
	50	8.20	1.06	1.03	1.34	1.10	2.11	.55	.50	.62	.55
2b	100	6.39	.66	.54	.79	.62	1.73	.45	.38	.48	.44
	200	6.33	.72	.50	.94	.64	1.81	.47	.37	.56	.46
	400	7.38	.80	.53	1.07	.69	1.98	.49	.38	.59	.50
	25	19.68	15.40	13.05	15.75	14.18	3.15	2.54	2.30	2.57	2.38
	50	26.64	20.85	18.79	20.95	19.24	5.28	4.19	3.83	4.18	3.87
3a	100	23.69	19.44	18.66	19.39	18.72	4.58	3.76	3.63	3.74	3.63
	200	21.68	17.81	17.44	17.64	17.36	4.45	3.65	3.58	3.60	3.55
	400	24.30	19.79	19.55	19.54	19.36	4.81	3.90	3.86	3.84	3.81
	25	35.20	21.48	18.85	22.07	20.25	6.82	4.58	4.18	4.59	4.25
	50	50.25	29.79	27.13	29.71	27.33	11.05	6.86	6.42	6.76	6.27
3b	100	40.99	26.41	25.37	26.13	25.28	8.95	5.93	5.71	5.83	5.66
	200	40.29	25.41	24.82	24.79	24.35	9.48	6.15	6.01	5.96	5.85
	400	46.62	28.85	28.43	27.98	27.71	10.49	6.66	6.57	6.43	6.36
	25	12.42	10.63	9.01	10.86	9.76	1.75	1.55	1.42	1.57	1.47
	50	16.48	14.21	12.99	14.33	13.32	3.14	2.78	2.58	2.81	2.62
4a	100	15.98	14.14	13.60	14.11	13.66	2.97	2.65	2.57	2.66	2.58
	200	13.61	12.32	12.05	12.22	12.04	2.61	2.39	2.35	2.39	2.35
	400	14.80	13.41	13.20	13.28	13.17	2.73	2.50	2.47	2.49	2.47
	25	22.21	13.88	12.61	14.42	13.44	4.11	3.04	2.87	3.14	2.95
	50	33.85	20.81	19.23	20.95	19.69	7.73	5.27	4.94	5.36	5.05
4b	100	31.09	19.78	18.98	19.50	18.95	7.12	4.88	4.74	4.88	4.76
	200	26.93	18.21	17.50	17.68	17.32	6.49	4.75	4.62	4.72	4.62
	400	29.82	20.38	19.58	19.67	19.38	6.81	5.04	4.91	5.01	4.92

Table 7: VarAS of OLS, WLS1, ALS1, MWLS1, WLS2, ALS2 and MWLS2 estimators for β_1 and β_2. WLS1, ALS1, MWLS1 (resp. WLS2, ALS2, MWLS2) are based on Model 1 (resp. Model 2) for $\omega^2(X;\gamma)$. Asymptotic variance of WLS1 and ALS1 (resp. WLS2 and ALS2) are identical, and so are their VarAS. So, the VarAS for ALS1 and ALS2 are omitted from the table to avoid repetition.

DGP	n	OLS	WLS1	ALS1	MWLS1	WLS2	ALS2	MWLS2
	25	5.1	5.2	5.2	5.1	5.2	5.2	5.1
	50	4.9	5.1	5.1	4.9	5.0	5.0	4.8
1a	100	5.0	4.9	4.9	5.0	4.9	4.9	5.0
	200	4.8	4.9	4.9	4.9	4.9	4.9	4.9
	400	5.1	5.1	5.1	5.1	5.1	5.1	5.0
	25	5.0	5.1	5.1	5.0	5.2	5.2	5.0
	50	5.0	5.1	5.1	5.2	5.1	5.1	5.1
1b	100	4.9	5.0	5.0	5.1	5.1	5.0	5.2
	200	4.9	5.0	5.0	5.0	5.1	5.1	5.0
	400	4.9	5.0	5.0	4.9	4.9	4.9	4.9
	25	5.1	5.2	5.1	5.0	5.1	5.0	5.0
	50	4.9	4.9	5.0	5.1	5.0	5.1	5.1
1c	100	5.0	5.2	5.1	5.1	5.1	5.1	5.1
	200	4.8	4.8	4.8	4.9	4.8	4.8	4.8
	400	5.0	5.0	5.0	5.0	4.9	4.9	5.0
	25	5.0	4.9	4.9	5.2	5.0	5.0	5.1
	50	5.0	5.0	5.0	5.1	5.0	5.0	5.0
1d	100	4.8	5.0	5.0	5.0	5.0	5.0	5.1
	200	5.2	5.0	5.0	4.9	4.9	4.9	4.9
	400	5.0	5.0	5.0	5.0	5.0	5.0	5.0
	25	4.9	5.1	5.0	5.4	5.1	5.1	5.2
	50	5.0	5.2	5.2	5.5	5.1	5.1	5.3
2a	100	4.8	5.1	5.1	5.1	5.1	5.1	5.0
	200	4.9	5.1	5.1	5.2	5.0	5.0	5.0
	400	5.1	5.0	5.0	5.3	4.9	4.9	5.1
	25	5.0	5.2	5.1	5.7	5.1	5.0	5.2
	50	5.1	5.4	5.4	6.2	5.2	5.2	5.9
2b	100	5.0	5.0	5.0	5.2	4.9	4.9	5.0
	200	4.9	5.0	5.0	5.1	5.0	5.0	5.1
	400	4.9	5.0	5.0	5.3	5.0	5.0	5.3
	25	5.0	5.1	5.0	4.9	5.0	5.0	4.9
	50	4.9	4.9	4.9	5.0	4.9	4.9	4.9
3a	100	4.9	4.8	4.9	4.9	4.9	4.9	4.9
	200	5.0	5.1	5.0	5.1	5.0	5.0	5.1
	400	5.0	5.0	5.0	5.1	5.1	5.1	5.0
	25	5.0	5.1	5.1	5.0	5.0	5.1	5.1
	50	5.0	4.9	4.9	5.1	5.0	4.9	5.0
3b	100	5.1	5.1	5.1	5.1	5.1	5.1	5.2
	200	4.9	5.0	5.0	5.1	5.0	5.0	5.1
	400	5.0	4.9	4.9	4.9	4.8	4.8	4.9
	25	5.0	5.0	4.9	5.0	5.0	5.0	5.0
	50	4.9	5.0	5.0	5.0	5.0	5.0	4.9
4a	100	4.8	4.8	4.8	4.8	4.8	4.8	4.8
	200	5.0	5.0	5.0	5.0	5.0	5.0	5.0
	400	5.0	5.1	5.1	5.1	5.1	5.1	5.1
	25	4.9	5.1	5.0	5.0	5.1	5.0	5.0
	50	5.1	5.0	5.0	5.0	5.0	4.9	5.1
4b	100	5.0	5.1	5.0	5.1	5.1	5.1	5.0
	200	4.9	5.0	5.0	5.0	5.0	5.0	5.1
	400	5.1	4.9	4.9	5.0	4.9	4.9	4.9

Table 8: Empirical size of a nominal 5% two-sided test for β_2 using the $N(0,1)$ critical value, and the t-ratio using the VarMC instead of the estimated asymptotic variance of the estimators for β_1.

DGP	n	OLS	WLS1	ALS1	MWLS1	WLS2	ALS2	MWLS2
	25	5.1	5.1	5.1	5.1	5.1	5.1	5.1
	50	4.9	5.0	5.0	4.8	4.9	4.9	4.8
1a	100	5.0	4.9	4.9	5.0	4.9	4.9	4.9
	200	5.0	5.0	5.0	5.0	5.0	5.0	5.0
	400	5.0	5.0	5.0	5.0	5.0	5.0	5.0
	25	5.2	5.1	5.1	5.2	5.1	5.1	5.2
	50	5.0	5.0	5.0	4.9	5.0	5.0	4.9
1b	100	4.9	5.0	5.0	5.0	5.0	5.0	5.0
	200	5.0	5.1	5.1	5.0	5.1	5.1	5.0
	400	5.0	5.0	5.0	5.0	5.0	5.0	5.0
	25	5.1	5.0	5.0	5.1	5.0	5.0	5.1
	50	4.9	4.9	5.0	4.9	4.9	4.9	4.8
1c	100	5.1	4.9	4.9	5.0	4.9	4.9	5.1
	200	4.8	4.9	4.9	4.9	4.9	4.9	5.0
	400	5.1	5.0	5.0	5.0	4.9	4.9	4.9
	25	5.0	5.0	5.0	5.0	5.0	5.0	5.0
	50	4.9	5.1	5.1	5.0	5.1	5.1	5.1
1d	100	4.9	5.0	5.0	5.0	5.0	5.0	5.1
	200	5.1	5.0	5.0	5.0	5.0	5.0	5.1
	400	5.1	5.1	5.1	5.2	5.2	5.2	5.3
	25	5.0	4.9	4.9	5.0	5.0	5.0	5.0
	50	4.9	5.1	5.1	5.1	5.1	5.1	5.1
2a	100	5.0	5.1	5.1	5.2	5.2	5.2	5.2
	200	4.9	5.1	5.1	5.0	5.1	5.1	5.0
	400	5.1	5.0	5.0	5.1	5.0	5.0	5.0
	25	5.1	5.1	5.1	5.2	5.0	5.0	5.1
	50	4.9	5.3	5.3	5.5	5.2	5.2	5.4
2b	100	5.1	5.0	5.0	5.1	5.0	5.0	5.1
	200	4.9	5.0	5.0	5.0	5.0	5.0	5.0
	400	4.9	4.9	4.9	5.1	4.9	4.9	5.0
	25	4.9	5.1	5.0	5.1	5.1	5.0	5.1
	50	4.9	5.0	4.9	4.8	5.0	5.0	4.9
3a	100	4.9	5.0	5.0	4.8	4.9	5.0	4.8
	200	4.9	5.1	5.1	5.0	5.1	5.1	5.0
	400	5.0	5.1	5.1	5.1	5.0	5.0	5.0
	25	5.0	5.1	5.2	5.0	5.2	5.3	5.2
	50	5.0	4.9	4.9	5.0	5.0	4.9	5.0
3b	100	5.2	5.1	5.1	5.0	5.1	5.0	5.1
	200	4.9	4.9	4.9	5.0	5.0	5.0	5.1
	400	5.0	5.1	5.1	5.0	5.1	5.1	5.1
	25	4.9	5.1	5.0	5.0	5.1	4.9	5.1
	50	4.9	5.0	5.0	4.9	4.9	4.9	4.9
4a	100	4.9	4.9	4.9	4.8	5.0	5.0	4.8
	200	5.0	5.1	5.1	5.0	5.1	5.1	5.0
	400	5.0	5.0	5.0	4.9	5.0	5.0	4.9
	25	5.1	5.1	5.0	5.1	5.1	5.1	5.1
	50	4.9	5.0	4.9	4.8	4.9	4.9	4.9
4b	100	5.2	4.9	5.0	5.0	5.1	5.1	5.0
	200	4.9	5.1	5.1	5.1	5.0	5.0	5.1
	400	4.9	5.0	5.0	5.0	5.1	5.1	5.1

Table 9: Empirical size of a nominal 5% two-sided test for β_2 using the $N(0,1)$ critical value, and the t-ratio using the VarMC instead of the estimated asymptotic variance of the estimators for β_2.