Efficient estimation of regression models with user-specified parametric model for heteroskedasticity*

Saraswata Chaudhuri† and Eric Renault‡

Abstract

We show that it is possible to estimate regression coefficients at least as precisely as Ordinary Least Squares (OLS), Weighted Least Squares (WLS) and its recently proposed refinements, regardless of the user's parametric model for heteroskedasticity. The key is to modify WLS depending on scalar target parameters, e.g., coefficients, marginal effects, predictions, etc. to obtain a targeted WLS that maximizes precision of estimation of the target given the user's model for heteroskedasticity. By construction, targeting improves precision over other estimators when the user's model is incorrect, and maintains semiparametric efficiency when it is correct. When the user's model is incorrect, targeted WLS can be improved by targeted combination of WLS and OLS building on the literature on "resurrecting WLS". Our targeted estimators are parametric and based on robust variance estimation. They are asymptotically optimal in important classes of estimators and show big improvements over others in simulations.

JEL Classification: C12; C13; C21.

Keywords: heteroskedasticity; misspecification; nuisance parameters; optimality; precision; weighted least squares

*We thank I. Andrews, F. Bugni, B. Gafarov, J. Galbraith, S. Goncalves, P. Guggenberger, J-M. Dufour, M. Kolesar, J. MacKinnon, P.C.B. Phillips, P. Rilstone, C. Rothe, P. Sant'Anna, A. Santos, R. Startz, Y. Shin, K. Xu, V. Zinde-Walsh, and various seminar and conference participants for helpful comments and discussions on various versions of the paper. The first author also thanks SSHRC for financial support.

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

‡Department of Economics, University of Warwick. Email: Eric.Renault@warwick.ac.uk.

1 Introduction

As announced by Stock and Watson (2011)'s popular textbook of econometrics, it has been widely accepted since the seminal work of White (1980) that "despite the theoretical appeal of WLS, heteroskedasticity robust standard errors provide a better way to handle potential heteroskedasticity in most applications". Angrist and Pischke (2010) go as far as reporting the "near-death of generalized least squares in cross sectional applied work". None of this is surprising because if the user's parametric model for heteroskedasticity is incorrect, which is perhaps the norm rather than the exception, then WLS can be even less precise than OLS.

In this paper we show that, irrespective of the correctness of the user's parametric model for heteroskedasticity, we always have a well-defined optimal weighting strategy that dominates OLS, WLS and its recently proposed refinements. Up to regularity conditions, the only constraint on the specification of the user's parametric model for heteroskedasticity is that it must contain conditional homoskedasticity as a particular case. Thus, while we agree with Stock and Watson (2011) that the functional form of conditional heteroskedasticity is rarely known, we demonstrate that it is not a sufficient reason to overlook possible improvements in accuracy brought by target-driven reweighting of the observations. Even though we could put forward a broader scope for validity (see Appendix B), our analysis is valid within a general regression framework — including instrumental variables or nonlinear regressions — as long as the regression is defined by a conditional expectation condition.

By: (i) referring to regressions defined by conditional expectations, (ii) using parametric models that may be incorrect for conditional heteroskedasticity but nest conditional homoskedasticity, and (iii) always using heteroskedasticity robust standard errors, we remain true to the research agenda of "resurrecting WLS"put forward by Romano and Wolf (2017).

Romano and Wolf (2017) propose an adaptive least squares (ALS) estimator that is WLS if a test of homoskedasticity based on the user's parametric model rejects homoskedasticity,

and is OLS otherwise. ALS can improve upon WLS in small samples under conditional homoskedasticity. However, ALS is asymptotically equivalent to WLS, and hence ALS can also be less precise than OLS under conditional heteroskedasticity if the user's parametric model for heteroskedasticity is incorrect. ALS is not designed for optimality in such cases.

DiCiccio et al. (2019) address this issue by focusing on the asymptotic variance of the estimator of a scalar component of the regression coefficients and choosing accordingly between OLS and WLS or some optimal combination thereof. The estimator choosing WLS or OLS depending on which one has MINimum standard error is their MIN estimator. The estimator constructing an optimal Convex Combination of OLS and WLS is their CC estimator.

In a different vein, Lu and Wooldridge (2020) build on Cragg (1983) to construct an optimal GMM estimator by combining the moment restrictions of OLS and WLS, to ensure at least as much precision as OLS or WLS estimators for the full vector of regression coefficients.

In our paper, we wish to further advance this new research agenda of "resurrecting WLS". We propose a targeted estimation method that takes a direct route to optimality and improves upon all the aforementioned methods (and others). Our key idea is twofold:

On the one hand, we follow DiCiccio et al. (2019) to acknowledge the need of targeting a scalar function of the regression coefficients to define our optimality criterion.

On the other hand, by contrast with Romano and Wolf (2017), DiCiccio et al. (2019) or Lu and Wooldridge (2020), we do not limit the search for optimality to only OLS and WLS. There is generally no reason to assume that the user's parametric model for heteroskedasticity is correct. Fortunately, an incorrect model does not cause biased estimates of regression coefficients. Therefore, we contend that the user's model should be utilized in a way such that the weights for reweighting observations are chosen to minimize the asymptotic variance of the estimator of the target. The goal is accurate estimation of the target and not necessarily least squares approximation of the true skedastic function, as implicitly performed by WLS.

We are not the first to follow this route. Cragg (1992) minimized the trace or determinant

of the asymptotic variance matrix of a weighted least squares estimator of all the regression coefficients with respect to the parameters in the user's model for heteroskedasticity. While Cragg (1992) does not discuss it, his minimization of trace or determinant corresponds to the notion of A or D optimality from the design of experiments literature. This could be appealing since a minimized asymptotic variance matrix for the full vector of regression coefficients may not even exist if the user's parametric model for heteroskedasticity is incorrect.

Unfortunately, without that existence, Cragg (1992)'s *standard errors may be larger than that of WLS or even OLS estimators of specific regression coefficients*. This is not desirable.

Therefore, while we also optimize with respect to the parameters in the user's model for heteroskedasticity, we argue that, unlike Cragg (1992)'s aggregate measures, the quantity to be optimized should be the estimation accuracy of scalar targets that directly align with the objects of empirical interest — individual coefficients, marginal effects, predictions, etc. We will show that this subtle difference due to our proposal delivers big improvements over all methods even when Cragg (1992)'s estimators themselves are less precise than WLS or OLS.

For our scalar target, we optimize the asymptotic variance of estimators following from three strategies that are in increasing order of precision and computational complexity. A common feature of these strategies is to refer to a given user-specified parametric model for heteroskedasticity, where the skedastic function is a known function of observations and of a finite dimensional vector $\gamma \in \Gamma$ of unknown parameters. To fix ideas without introducing many notation, for any scalar target parameter, let us generically denote by $WLS(\gamma)$ the weighted least squares estimator computed with weights defined by the value γ of the user's model parameters. Thus, the classical version of WLS, referred to as WLS throughout, is asymptotically equivalent to $WLS(\gamma_{WLS})$ where γ_{WLS} is the value of γ that makes the user's model the closest to the true unknown skedastic function in terms of mean squared error.

Our first strategy, that we refer to as "Targeted WLS" (TWLS), is asymptotically equivalent to $WLS(\gamma^*_{TWLS})$ where γ^*_{TWLS} is the asymptotic variance minimizer of $WLS(\gamma)$. The

term "Targeted"emphasizes that the choice of γ and, hence, the weights is target-driven. Thus, while γ_{WLS} does not change with targets, γ^*_{TWLS} will change with the target in general if the user's model for heteroskedasticity is incorrect. By construction, TWLS is at least as precise as OLS, $WLS(\gamma_{WLS})$, Cragg (1992)'s estimators, Romano and Wolf (2017)'s ALS estimator and DiCiccio et al. (2019)'s MIN estimator for the given scalar target.

Our second strategy improves upon the first by considering convex combinations of OLS and $WLS(\gamma)$. The variance minimizing convex combination of OLS and $WLS(\gamma_{WLS})$ was introduced by DiCiccio et al. (2019) as their CC estimator. By contrast, we do not limit ourselves to $WLS(\gamma_{WLS})$. We obtain our "Targeted Convex Combination" (TCC) estimator as follows. First we obtain an optimal convex combination of OLS and $WLS(\gamma)$ by finding the $\lambda^*(\gamma) \in [0,1]$ that minimizes the asymptotic variance of $(1-\lambda)OLS + \lambda WLS(\gamma)$. Then, we choose the optimal γ^*_{TCC} that minimizes the asymptotic variance of $(1-\lambda^*(\gamma))OLS + \lambda^*(\gamma)WLS(\gamma)$. DiCiccio et al. (2019)'s CC estimation does not do this key step of variance minimization with respect to γ but uses γ_{WLS} that, as we will show, can lead to substantial suboptimality when the user's model for heteroskedasticity is incorrect. By construction, TCC is at least as precise as CC and also all the estimators noted under strategy one.

Our third strategy, in the spirit of Chen et al. (2016), involves matrix extensions of convex combinations of estimators with matrix weights summing to the identity matrix. By contrast with TCC, now the first step minimizes the asymptotic variance of the matrix combination of the OLS and WLS(γ) estimators of the *entire vector of regression coefficients*, while the second step minimizes with respect to γ the asymptotic variance of the targeted scalar function of the optimal matrix combination estimator from step one. Since the first step turns out to be the optimal GMM-combination of the concerned moment functions, we call the resulting estimator the "Targeted GMM" (TGMM) estimator (and the minimizing γ, γ^*_{TGMM}). TGMM is an improvement over Lu and Wooldridge (2020)'s GMM estimator in the same way as TWLS is an improvement over $WLS(\gamma_{WLS})$, or TCC over CC.

Our three strategies are asymptotically equivalent to WLS if the user's model for heteroskedasticity is correct, in which case WLS is indeed optimal and $\gamma_{WLS} = \gamma^*_{TWLS} = \gamma^*_{TCC} = \gamma^*_{TGMM}$ for any target. On the other hand, if the user's model is incorrect, then this equality of the γ's may not hold and our strategies provide (weakly) more precision than WLS, with TGMM being the optimal one asymptotically. Our strategies, however, do not encompass Gourieroux et al. (1996), Spady and Stouli (2019), Papadopoulos and Tsionas (2022) and others who impose additional structure to propose estimators under heteroskedasticity.

Our takeaway message is that our proposed targeting is harmless, at least asymptotically, and when the user's model for heteroskedasticity is incorrect it can, in fact, increase precision. Under each of our three strategies, targeting delivers asymptotic variance and, incidentally, empirical (Monte Carlo) mean squared errors that are as good as and often much better than that of the contenders. Thus, targeting is ranked the best among the estimators under each strategy. On the other hand, there is a tension among our three strategies in the sense that while TWLS is the easiest and TGMM is the hardest computationally, we can formally establish that TWLS cannot be more precise than TCC which, in turn, cannot be more precise than TGMM. Also, our simulations find TGMM may sometimes be flawed by small-sample issues that warrants a maintained interest in TCC and TWLS.

Lastly, it is worth noting that targeting does not search for parametric models to obtain a significant result. Rather, it takes the user's/expert's model as given. And, then, it searches for the target-specific optimal γ-value (leading to optimal weights) given this model. Our estimate of the target parameter obtained under each strategy may or may not be significant, but is nevertheless the most precise one asymptotically that one could obtain under that strategy and given the user's/expert's choice of the parametric model for heteroskedasticity.

Our paper is organized as follows.

Section 2 defines targeted estimation in regression models, makes explicit the need to focus on scalar functions of the regression coefficients as target parameters of interest, in-

troduces the TWLS estimator, and characterizes its properties. To emphasize the benefit of our way of proper targeting without going into the more complicated estimation methods, Section 2 also performs Romano and Wolf (2017)'s Monte Carlo experiment using the well-known Boston housing data, and shows that improper targeting by Cragg (1992)'s estimators leads them to be much less precise than not only TWLS but also even WLS and OLS.

Section 3 builds on the targeting idea and applies it to the combination of estimators and moment restrictions proposed by DiCiccio et al. (2019) and Lu and Wooldridge (2020). While the same idea could be applied more generally (see Appendix B that builds on Chen et al. (2016)), here we focus on OLS and WLS following the recent literature on "resurrecting WLS". TCC and TGMM estimators are the products of our application of targeting to the combination of estimators and moment restrictions respectively.

Section 4 presents the superior small-sample empirical mean squared error of our proposed estimators over OLS, WLS, and the recently proposed ALS, MIN, CC, GMM estimators, under the Monte Carlo designs of Romano and Wolf (2017) and Lu and Wooldridge (2020). It also refers to favorable comparison with semiparametric WLS in Appendix C. There is not much cost to inference in terms of over-rejection of the truth. We also see precision gains by our proposed estimators when applied to Lu and Wooldridge (2020)'s empirical application.

Section 5 concludes. There is a Supplemental Appendix. Appendix A presents the well-known expressions for the quantities used in Section 3, and proves the optimality result from Section 3. Appendix B presents the theory connecting the affine combination estimators and the GMM estimators by revisiting a main result of Chen et al. (2016). It is worth noting that we also interpret these combinations through the lens of standard econometrics by casting them as regression of OLS on the difference between OLS and WLS. Appendix C considers the simulation designs in Miller and Startz (2019) and Gonzales-Coya and Perron (2024), and provides comparisons favoring our proposed estimators over the classical semiparametric WLS and the machine-learning-based semiparametric WLS estimators from those papers.

2 Targeted estimation

2.1 The basic regression model

Our linear model with the unknown regression coefficients $\beta = (\beta_1, \ldots, \beta_p)'$ takes the form:

$$y_i \;=\; \beta_1 + \beta_2 x_{i,2} + \ldots + \beta_p x_{i,p} + u_i(\beta) \;=\; x_i'\beta + u_i(\beta) \tag{1}$$

where $(y_i, x_{i,2}, \ldots, x_{i,p})$ is independent and identically distributed (i.i.d.) for $i = 1, \ldots, n$, and the regressors $x_i = (x_{i,1}, \ldots, x_{i,p})'$ with $x_{i,1} = 1$ are not redundant, i.e., $E[x_i x_i']$ is nonsingular. The true unknown value β^0 of β is well-defined by the conditional expectation condition:

$$E[u_i | x_i] = 0 \quad \text{where} \quad u_i := u_i(\beta^0). \tag{2}$$

Remark: Our results apply not only to the linear regression $E[y_i - x_i'\beta^0 | x_i] = 0$, but also to the nonlinear regression $E[y_i - m(x_i, \beta^0) | x_i] = 0$ where $m(.,.)$ is a known scalar function, and to the IV regression $E[y_i - x_i'\beta^0 | z_i] = 0$ with z_i as instruments. In this last case, conditional heteroskedasticity of the error term $u_i = y_i - x_i'\beta^0$ must be understood given z_i.

While setting the focus on the regression model (1), we define the unknown skedastic function $\omega_0^2(x_i)$ as the conditional heteroskedasticity, i.e., the conditional variance of the error u_i:

$$\omega_0^2(x_i) := E\left[u_i^2 | x_i\right] > 0.$$

On the other hand, the user's model for $\omega_0^2(x_i)$, which is very likely incorrect, is given by a parametric family:

$$\omega^2(x_i, \gamma) > 0, \gamma \in \Gamma \subset \mathbb{R}^{d_\gamma}. \tag{3}$$

As is standard, we will always assume that this family nests the case of conditional homoskedasticity, i.e., there exists $\gamma^{\text{hom}} \in \Gamma$ such that $\omega^2\left(x_i, \gamma^{\text{hom}}\right) \equiv \omega_{\text{hom}}^2$, a constant.

Examples: Romano and Wolf (2017) point out three commonly used parametric families (and prefer Ex1) for the user's model $\omega^2(x_i,\gamma)$ in (3) for conditional heteroskedasticity:

$$\text{Ex1: } \exp\left(\gamma_1+\sum_{j=2}^{p}\gamma_j\log\left(|x_{i,j}|\right)\right),\ \text{Ex2: } \gamma_1+\sum_{j=2}^{p}\gamma_j|x_{i,j}|,\ \text{Ex3: } \exp\left(\gamma_1+\sum_{j=2}^{p}\gamma_j x_{i,j}\right).$$

$\gamma^{\text{hom}}=\left(\gamma_1^{\text{hom}},0,...,0\right)'$ for Ex1-Ex3. The parametric models may not contain $\omega_0^2(x_i)$.

2.2 Weighted Least Squares

For any $\gamma\in\Gamma$ we define a weighted-by-$\omega^2(x_i,\gamma)$ estimator of β as:

$$\widehat{\beta}(\gamma):=\left(\sum_{i=1}^{n}\frac{x_ix_i'}{\omega^2(x_i,\gamma)}\right)^{-1}\sum_{i=1}^{n}\frac{x_iy_i}{\omega^2(x_i,\gamma)}. \tag{4}$$

The mean independence condition (2) implies that under standard regularity conditions, $\sqrt{n}\left(\widehat{\beta}(\gamma)-\beta^0\right)\xrightarrow{d}N(0,\Sigma(\beta^0,\gamma))$ with the asymptotic variance matrix $\Sigma(\beta^0,\gamma)$ given by:

$$\Sigma\left(\beta^0,\gamma\right):=\left(E\left[\frac{x_ix_i'}{\omega^2(x_i,\gamma)}\right]\right)^{-1}E\left[\frac{x_ix_i'\omega_0^2(x_i)}{\omega^4(x_i,\gamma)}\right]\left(E\left[\frac{x_ix_i'}{\omega^2(x_i,\gamma)}\right]\right)^{-1}. \tag{5}$$

Definition: The parametric model (3) for the conditional heteroskedasticity is well-specified, i.e., correct, if and only if for some parameter value $\gamma^0\in\Gamma$:

$$\omega^2\left(x_i,\gamma^0\right)=\omega_0^2(x_i). \tag{6}$$

In this case: $\Sigma(\beta^0,\gamma^0)=\left(E\left[\frac{x_ix_i'}{\omega^2(x_i,\gamma^0)}\right]\right)^{-1}\ll\Sigma(\beta^0,\gamma)$ for all $\gamma\in\Gamma$ with the notation $A\ll B$ meaning "$B-A$ positive semi-definite". It is the case in particular if u_i is conditionally homoskedastic, in which case $\gamma^0=\gamma^{\text{hom}}$ and $\omega^2(x_i,\gamma^0)=\omega_{\text{hom}}^2=\omega_0^2(x_i)$.

Our focus of interest is precisely in the common scenario where there is some conditional

heteroskedasticity and the user's parametric model for heteroskedasticity is misspecified (incorrect). Under this scenario, we cannot define a true unknown value γ^0 as the solution of (6). We can only define a pseudo-true value, denoted by γ_{WLS}, as the solution of:

$$\gamma_{WLS} := \arg\min_{\gamma\in\Gamma} E\left[\left(\omega_0^2(x_i) - \omega^2(x_i,\gamma)\right)^2\right] = \arg\min_{\gamma\in\Gamma} E\left[\left(u_i^2 - \omega^2(x_i,\gamma)\right)^2\right]. \tag{7}$$

The notation is motivated by the fact that the infeasible WLS corresponds to $\widehat{\beta}(\gamma_{WLS})$ while a feasible version is given by $\widehat{\beta}(\widehat{\gamma}_{WLS})$ for any consistent estimator $\widehat{\gamma}_{WLS}$ of γ_{WLS}. $\widehat{\gamma}_{WLS}$ is typically obtained by the sample analog of the minimizer of the minimization program (7) with $u_i := u_i(\beta^0)$ replaced by (some possibly trimmed version of) the OLS residual:

$$\widehat{u}_{i,OLS} := u_i(\widehat{\beta}_{OLS}) = y_i - x_i'\widehat{\beta}_{OLS} \quad \text{where} \quad \widehat{\beta}_{OLS} := \left(\sum_{i=1}^n x_i x_i'\right)^{-1} \sum_{i=1}^n x_i y_i = \widehat{\beta}(\gamma^{\text{hom}}).$$

It is however worth keeping in mind that when the first minimization in (7) does not provide a value function equal to zero, i.e., when the user's model for heteroskedasticity is misspecified, there is no strong argument in favor of the choice of the value γ_{WLS} of γ. Ideally, one would like to define an optimal value γ^* such that:

$$\Sigma\left(\beta^0,\gamma^*\right) \ll \Sigma\left(\beta^0,\gamma\right) \quad \text{for all} \quad \gamma\in\Gamma. \tag{8}$$

However, except in the well-specified case (then $\gamma^* = \gamma^0$), such an optimal value γ^* does not exist in general, and there is no compelling argument to consider that γ_{WLS} is "closer to optimality" than its contenders. The only way to escape this deadlock will be to replace the matrix-minimization problem (8) by a program of minimization of a scalar function.

This scalar minimization is the purpose of the next subsection. An estimator $\widehat{\beta}(\gamma)$ for some given user-specified γ, leading to weights $\omega^2(x_i,\gamma)$ in (4), will be dubbed "User-specified WLS" (UWLS). When computed with our preferred value of γ (see below), it will be dubbed "Targeted WLS" (TWLS). We will call both $\widehat{\beta}(\gamma_{WLS})$ and $\widehat{\beta}(\widehat{\gamma}_{WLS})$ WLS or classical WLS.

2.3 Targeting in regression

As explained above, we need to set the focus on estimation of a scalar target parameter of interest. We take this scalar target as a known and smooth function of β:

$$h(\beta) : \mathbb{R}^p \mapsto \mathbb{R} \quad \text{with } \frac{\partial h(\beta)}{\partial \beta} \text{ continuous at } \beta = \beta^0.$$

Common examples of $h(\beta)$ include individual regression coefficients, prediction of conditional mean $\tilde{x}'\beta$ at some given value $x = \tilde{x}$, linear combinations of β such as $\beta_2 + \beta_2$, $\beta_2 - \beta_3$, etc.

We consider substitution estimators of $h(\beta)$ based on UWLS estimators $\widehat{\beta}(\gamma)$ of β:

$$\widehat{h}_{UWLS}(\gamma) \ := \ h(\widehat{\beta}(\gamma)).$$

The mean independence in (2) and regularity conditions give $\sqrt{n}\left(\widehat{h}_{UWLS}(\gamma) - h(\beta^0)\right) \xrightarrow{d} N\left(0, \sigma^2_{h,UWLS}(\beta^0, \gamma)\right)$ with the asymptotic variance, which is a function of $\gamma \in \Gamma$, given by:

$$\sigma^2_{h,UWLS}(\beta^0, \gamma) \ := \ \frac{\partial h(\beta^0)}{\partial \beta'} \Sigma(\beta^0, \gamma) \frac{\partial h(\beta^0)}{\partial \beta}.$$

As noted by Cragg (1992), the classical WLS estimator "can lead to larger diagonal elements of $\Sigma(\beta^0, \gamma)$ than those of OLS" (Cragg (1992) used $V(\hat{\beta}(\theta))$ to denote $\Sigma(\beta^0, \gamma)$). Surprisingly, while Cragg (1992) was concerned by estimation of *specific* components of β, he never chose to minimize the asymptotic variance of a *specific* component but rather minimized aggregate measures viz. the determinant or trace of $\Sigma(\beta^0, \gamma)$. Consequently, as we will see, Cragg (1992)'s estimators themselves can be less precise than WLS and OLS.

Therefore, more generally and appropriately as far as precise estimation of $h(\beta)$ is concerned, we define our TWLS as a feasible version of $\widehat{h}_{UWLS}(\gamma^*_{h,TWLS}) = h(\widehat{\beta}(\gamma^*_{h,TWLS}))$ where:

$$\gamma^*_{h,TWLS} \ := \ \arg\min_{\gamma \in \Gamma} \ \sigma^2_{h,UWLS}(\beta^0, \gamma).$$

2.4 Feasible TWLS

The first task is to estimate for any $\gamma \in \Gamma$ the asymptotic variance matrix $\Sigma(\beta^0, \gamma)$ given in (5). Under regularity conditions, we have a consistent estimator (uniformly in $\gamma \in \Gamma$):

$$\widehat{\Sigma}(\widehat{\beta}_{OLS}, \gamma) \ := \ \left(\frac{1}{n} \sum_{i=1}^{n} \frac{x_i x_i'}{\omega^2(x_i;\gamma)} \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^{n} \frac{\widehat{u}_{i,OLS}^2 x_i x_i'}{\omega^4(x_i;\gamma)} \right) \left(\frac{1}{n} \sum_{i=1}^{n} \frac{x_i x_i'}{\omega^2(x_i;\gamma)} \right)^{-1}.$$

Consistency of the estimator for $\left(E\left[\frac{x_i x_i'}{\omega^2(x_i,\gamma)} \right] \right)^{-1}$ is implied by the uniform law of large numbers and the continuous mapping theorem. Consistency of the estimator for $E\left[\frac{x_i x_i' \omega_0^2(x_i)}{\omega^4(x_i,\gamma)} \right]$ is a consequence of the uniform consistency of an Eicker-Huber-White estimator when the vector x_i of explanatory variables is replaced by the pseudo-sphericized one $\widetilde{x}_i(\gamma) = x_i/\omega^2(x_i, \gamma)$.

We can now define:

$$\widehat{\sigma}_{h,UWLS}^2(\widehat{\beta}_{OLS}, \gamma) \ := \ \frac{\partial h(\widehat{\beta}_{OLS})}{\partial \beta'} \widehat{\Sigma}(\widehat{\beta}_{OLS}, \gamma) \frac{\partial h(\widehat{\beta}_{OLS})}{\partial \beta}$$

and a consistent estimator of the targeted γ:

$$\widehat{\gamma}_{h,TWLS} \ := \ \arg\min_{\gamma \in \Gamma} \ \widehat{\sigma}_{h,UWLS}^2(\widehat{\beta}_{OLS}, \gamma).$$

Thus, we obtain our feasible TWLS estimator and its standard error, respectively, as:

$$\widehat{h}_{TWLS} \ := \ \widehat{h}_{UWLS}(\widehat{\gamma}_{h,TWLS}) \ = \ h(\widehat{\beta}(\widehat{\gamma}_{h,TWLS})) \ \text{ and } \ se_{h,TWLS} \ := \ \left[\frac{1}{n} \widehat{\sigma}_{h,UWLS}^2(\widehat{\beta}_{OLS}, \widehat{\gamma}_{h,TWLS}) \right]^{1/2}.$$

Remark: One can plug in for β^0 any consistent estimator for β^0 when estimating quantities such as $\Sigma(\beta^0, \gamma)$, $\sigma_{h,UWLS}^2(\beta^0, \gamma)$, $\gamma_{h,TWLS}^*$, etc. without altering the first-order asymptotic properties of estimation and inference of $h(\beta)$ considered in our paper. Using $\widehat{\beta}_{OLS}$ as the plugin is neither a restriction nor our prescription. We write the plugin as $\widehat{\beta}_{OLS}$ only for the sake of uniformity in the presentation of the estimators of $h(\beta)$ and their standard errors.

2.5 The targets

By definition, the optimal weights for TWLS depend upon the target of interest $h(\beta)$.

A standard case where $h(\beta) = \beta_2$, a single coefficient, is the target is when we consider a regression model of y_i on a binary treatment variable $x_{i,2}$ and the covariates $w_{i,k}$'s:

$$y_i = \beta_1^0 + \beta_2^0 x_{i,2} + \sum_{k=1}^{K} \nu_{1,k}^0 w_{i,k} + \sum_{k=1}^{K} \nu_{2,k}^0 x_{i,2}(w_{i,k} - E[w_{i,k}]) + u_i.$$

(The ν's represent the $\beta_3, \ldots, \beta_p$ in regression model (1).) β_2 is meant to capture the average effect of $x_{i,2}$. Interest can also be on other $h(\beta)$'s, e.g., $h(\beta) = (\tilde{w}_2 - E[w_{i,2}])\nu_{2,2}$, i.e., the extra effect of $x_{i,2}$ over the average effect when the covariate $w_{i,2}$ is fixed at $\tilde{w}_2$.

A more restrictive case with $h(\beta) = \beta_2$ meant to capture the homogeneous effect of $x_{i,2}$ being the target would be the non-interactive ($x_{i,2}$ additively separable from $w_{i,k}$'s) model:

$$y_i = \beta_1^0 + \beta_2^0 x_{i,2} + \sum_{k=1}^{K} \nu_{1,k}^0 w_{i,k} + u_i.$$

Since the $\nu_{1,k}$'s are not of interest, one can work under the assumption $E[u_i|x_{i,2}, w_{i,1}, \ldots, w_{i,K}] = E[u_i|w_{i,1}, \ldots, w_{i,K}]$ (not necessarily 0) that is weaker than our assumption (2). This paper continues with (2) following the literature on "resurrecting WLS". Use of WLS under various weaker assumptions is pursued in a followup focusing on violations of assumptions like (2).[1]

Another case of standard regression follows if interest lies on a linear combination of

[1]WLS-type methods work under $E[u_i|x_{i,2}, W_i] = E[u_i|W_i]$ that is a weaker condition than (2). Let $x_i = (1, x_{i,2}, w_{i,1}, \ldots, w_{i,K})'$ and $W_i = (1, w_{i,1}, \ldots, w_{i,K})'$. To fix ideas, impose linearity, i.e., $E[u_i|x_{i,2}, W_i] = E[u_i|W_i] = W_i'\delta$ following Appendix 6.5 of Stock and Watson (2011). Using the Frisch-Waugh-Lovell Theorem, note that this linearity implies the effective WLS-type moment restriction for estimation of β_2 as:

$$E\left[\left\{\frac{x_{i,2}}{\sqrt{\omega^2(x_i;\gamma)}} - E\left[\frac{x_{i,2}W_i}{\omega^2(x_i;\gamma)}\right]\left(E\left[\frac{W_iW_i'}{\omega^2(x_i;\gamma)}\right]\right)^{-1}\frac{W_i}{\sqrt{\omega^2(x_i;\gamma)}}\right\}\frac{u_i}{\sqrt{\omega^2(x_i;\gamma)}}\right] = 0 \quad \text{for all } \gamma.$$

This robustness of the moment restriction to perturbations in γ implies that estimation of γ for feasible WLS, TWLS, etc. does not affect the first-order asymptotic variance of the estimators of β_2. Hence the standard theory still works when the only purpose of W_i is to make the causal variable $x_{i,2}$ "as if random".

$\beta_1, \ldots, \beta_p$'s, i.e., $h(\beta) = c_1\beta_1 + c_2\beta_2 + c_3\beta_3 + \ldots + c_p\beta_p$ for known scalar constants $c_1, c_2, c_3, \ldots, c_p$. Taking $c_2 \neq 0$ without loss of generality, one can then rewrite the original regression model in (1) such that the target $h(\beta)$ is now the coefficient of one transformed regressor $x_{i,2}/c_2$, and revert to focusing on a single scalar regression coefficient:

$$y_i = \beta_1^0 \left(1 - \frac{c_1}{c_2} x_{i,2}\right) + h(\beta^0) \frac{x_{i,2}}{c_2} + \beta_3^0 \left(x_{i,3} - \frac{c_3}{c_2} x_{i,2}\right) + \ldots + \beta_p^0 \left(x_{i,p} - \frac{c_p}{c_2} x_{i,2}\right) + u_i.$$

One might also be interested in prediction, i.e., $E[y|x = \tilde{x}]$, at a given $\tilde{x} = (1, \tilde{x}_2, \ldots, \tilde{x}_p)'$ with $\tilde{x}_2 \neq 0$ without loss of generality. Then the target is $h(\beta) = \beta_1 + \beta_2\tilde{x}_2 + \ldots + \beta_p\tilde{x}_p$, which is a special case of the last example with $c_1 = 1$ and $c_j = \tilde{x}_j$ for $j = 2, \ldots, p$.

On the other hand, if one is interested in the complete set of regression coefficients $\beta_1, \ldots, \beta_p$, then one must indeed estimate each of them with different optimal weights. The result of the regression would then come as follows:

$$\widehat{y}_i = \underset{(se_{\beta_1,TWLS})}{\widehat{\beta}_1} + \underset{(se_{\beta_2,TWLS})}{\widehat{\beta}_2} x_{i,2} + \ldots + \underset{(se_{\beta_p,TWLS})}{\widehat{\beta}_p} x_{i,p}.$$

This way of presenting regression results overlooks the cross-correlation between estimators of different components of β. But there is nothing restrictive in this respect by comparison with the standard practice of separately checking the Student t-values for each coefficient.

2.6 The importance of targeting

We present a numerical example based on well-known real-life data to display the quantitative benefit of "proper" targeting by TWLS in comparison with Cragg (1992)'s proposal.

We conduct Romano and Wolf (2017)'s simulation experiment based on the extremely well-known data from 1970 for $n = 506$ communities in the Boston area; also see DiCiccio et al. (2019) and Miller and Startz (2019). Consider the linear regression function $E[y_i|x_i] = \beta_1 + \beta_2 x_{i,2} + \beta_3 x_{i,3} + \beta_4 x_{i,4} + \beta_5 x_{i,5}$ for community $i = 1, \ldots, n$, where y_i is the log of the

median housing price, $x_{i,2}$ is the log of nitrogen oxide in the air, $x_{i,3}$ is the log of weighted distance from five employment centers, $x_{i,4}$ is the average number of rooms per house, and $x_{i,5}$ is the average student–teacher ratio in the community's schools; see Wooldridge (2012).

To resemble the true conditional heteroskedasticity in this data: (i) obtain $\widehat{e}_i = (y_i - x_i'\widetilde{\beta})/\sqrt{1 - x_i'(\sum_j x_j x_j')^{-1} x_i}$ for $i = 1, \ldots, n$ where $\widetilde{\beta}$ is the OLS estimator based on (y_i, x_i) for $i = 1, \ldots, n$; (ii) generate 10000 times artificial data sets (y_i^*, x_i^*) for $i = 1, \ldots, n$ where $x_i^* = x_i$ and $y_i^* = x_i'\widetilde{\beta} + \widehat{e}_i v_i$ where $v_i \sim N(0,1)$ i.i.d. $i = 1, \ldots, n$ independently of the system. Thus, the true β in each of this 10000 artificial data sets is $\beta^0 = \widetilde{\beta}$.

For each $h(\beta) = \beta_1, \ldots, \beta_5$, we report in Table 1 the ratio of the average standard error (ASE) for each of (i) WLS, (ii) Cragg's LTV (trace minimizing) estimator, (iii) Cragg's LGV (determinant minimizing) estimator, and (iv) our TWLS estimator, with respect to that for OLS. By construction, the standard errors of TWLS cannot exceed that of OLS, WLS or Cragg (1992)'s estimators for any target $h(\beta) = \beta_1, \ldots, \beta_5$. Hence, we also report the ratio of the empirical mean squared error (EMSE) of the estimators (i)-(iv) with respect to that of OLS. The ASEs and EMSEs are obtained as averages based on 10000 Monte Carlo trials.

Three observations are in order. First, WLS is more precise than OLS in this example. Second, while Cragg (1992)'s LTV and LGV estimators are more precise than OLS and WLS estimators for $\beta_1, \ldots, \beta_4$, LTV and LGV are in fact both less precise than OLS and WLS for β_5. This is a bad problem that makes Cragg (1992)'s LTV and LGV "inadmissible" in the sense they can perform worse than the estimators that they are meant to improve. This "inadmissibility" is not surprising theoretically but is worth documenting numerically. Third, and most importantly, we see that proper targeting of the object of interest by our TWLS estimator makes it by far the preferred estimator in this experiment.

While some other simulation designs of Romano and Wolf (2017) or Lu and Wooldridge (2020) lead to bigger benefits of proper targeting by TWLS, we presented this example since this simulation experiment is based on a widely familiar real-life data; see Wooldridge (2012).

	ASE(Estimator)/ASE(OLS)				EMSE(Estimator)/EMSE(OLS)			
$h(\beta)$	WLS	Cragg-LTV	Cragg-LGV	TWLS	WLS	Cragg-LTV	Cragg-LGV	TWLS
β_1	.779	.672	.704	.671	.613	.494	.513	.501
β_2	.812	.768	.789	.736	.676	.607	.637	.562
β_3	.713	.612	.650	.577	.506	.381	.427	.337
β_4	.710	.613	.591	.555	.500	.389	.340	.348
β_5	.953	1.154	1.037	.941	.927	1.427	1.139	.883

Table 1: The left and right panels report the ratio of the average standard error (ASE) and empirical MSE (EMSE) respectively of each estimator with respect to that of OLS. ASE amd EMSE are both obtained based on 10000 Monte Carlo trials. Romano and Wolf (2017)'s Model 1 $\omega^2(x_i;\gamma) = \exp\left(\gamma_1 + \sum_{j=2}^{p} \gamma_j \log\left(|x_{i,j}|\right)\right)$ is used as the user's model for heteroskedasticity.

Perhaps because of their "inadmissibility", Cragg (1992)'s estimators have been overlooked in the empirical literature and, surprisingly, even in the new literature on "resurrecting WLS" (except for a citation as earlier work in a sentence in footnote 2 of Romano and Wolf (2017)). We will also not consider estimators that can be less precise than both OLS and WLS. However, we must note that our TWLS estimator builds upon Cragg (1992) by modifying it with proper targeting to not only overcome "inadmissibility" but also improve over all the estimators like OLS, WLS, ALS, MIN that are encompassed by our first strategy.

3 Combine Estimators, Combine Moment Restrictions

We build on DiCiccio et al. (2019) and Lu and Wooldridge (2020) that combine OLS and the classical WLS (i.e., WLS(γ_{WLS})). We improve over them by the application of targeting.

The targeted combination of the OLS and UWLS(γ) estimators for $h(\beta)$ will lead to our TCC estimator for $h(\beta)$ in Section 3.1. The targeted matricial combination of the OLS and UWLS(γ) estimators/moment conditions for the entire vector β will lead to our TGMM estimator for $h(\beta)$ in Section 3.2. The motivation behind targeting is the same as before, i.e., we wish to fully exploit the user's model for heteroskedasticity and obtain the most precise estimator of the target not only when the user's model is correct but also when it is not.

3.1 Targeted Convex Combination (TCC) estimator

For any $\gamma \in \Gamma$, define the optimal convex combination of the two estimators $h(\widehat{\beta}_{OLS})$ and $h(\widehat{\beta}(\gamma))$, i.e., the OLS and UWLS(γ) estimators of $h(\beta)$, as:

$$\widehat{h}_{UCC}(\gamma) \;:=\; (1-\widehat{\lambda}_h(\gamma))h(\widehat{\beta}_{OLS}) + \widehat{\lambda}_h(\gamma)h(\widehat{\beta}(\gamma))$$

where, using $Avar$ to denote the variance of the asymptotic distribution and $\widehat{Avar}$ its estimator,

$$\widehat{\lambda}_h(\gamma) \;:=\; \arg\min_{\lambda\in[0,1]} \widehat{Avar}\left((1-\lambda)h(\widehat{\beta}_{OLS}) + \lambda h(\widehat{\beta}(\gamma))\right). \tag{9}$$

We will call $\widehat{h}_{UCC}(\gamma)$ the "User-specified CC" (UCC(γ)) and our proposed estimator "Targeted CC" (TCC). DiCiccio et al. (2019)'s CC estimator is $\widehat{h}_{CC} := \widehat{h}_{UCC}(\widehat{\gamma}_{WLS})$, which is one choice of UCC(γ). However, this choice involving $\gamma = \widehat{\gamma}_{WLS}$ does not exploit the source of optimality with respect to $\gamma \in \Gamma$ when the user's model $\omega^2(x_i;\gamma)$ is misspecified/incorrect for the true skedastic function $\omega_0^2(x_i)$. Thus, the need for our TCC estimator.

Following the same logic as in Section 2, we define our TCC estimator by means of a further minimization as:

$$\widehat{h}_{TCC} \;:=\; \widehat{h}_{UCC}(\widehat{\gamma}_{h,TCC})$$

where

$$\widehat{\gamma}_{h,TCC} \;:=\; \arg\min_{\gamma\in\Gamma} \widehat{Avar}\left(\widehat{h}_{UCC}(\gamma)\right). \tag{10}$$

It is well known that thanks to the conditional mean independence condition $E[y_i - x_i'\beta^0|x_i] = 0$ in (2), estimation of γ does not affect the first-order asymptotic properties of WLS under regularity conditions; see, e.g., Romano and Wolf (2017) for a recent reference. The same applied for our TWLS estimator. In this section, we additionally observe that:

$$E\left[(1-\lambda)\widehat{\beta}_{OLS} + \lambda\widehat{\beta}(\gamma)\right] \;=\; \beta^0 \;\text{ for any } \gamma\in\Gamma \text{ and any } \lambda.$$

Therefore, as shown in DiCiccio et al. (2019), estimation of γ and λ does not affect the first-order asymptotic properties of their CC estimator under very weak conditions; also see, e.g., Section 6 of Newey and McFadden (1994). The same applies for our TCC estimator.

Consequently, our TCC estimator $\widehat{h}_{TCC}$ will be asymptotically equivalent to an infeasible estimator $\widehat{h}^{\text{inf}}_{TCC}$ that is optimal in the class of convex combination estimators, i.e.,

$$\sqrt{n}(\widehat{h}_{TCC} - h(\beta^0)) \;=\; \sqrt{n}(\widehat{h}^{\text{inf}}_{TCC} - h(\beta^0)) + o_p(1) \;\xrightarrow{d}\; N(0, \sigma^2_{h,UCC}(\beta^0, \gamma^*_{h,TCC}, \lambda^*_h(\gamma^*_{h,TCC})))$$

where the key quantity, i.e., the asymptotic variance, is the minimum with respect to γ, λ of:

$$\sigma^2_{h,UCC}(\beta^0, \gamma, \lambda) \;:=\; AVar\left((1-\lambda)h(\widehat{\beta}_{OLS}) + \lambda h(\widehat{\beta}(\gamma))\right),$$

which gives the said optimality in the class. And, resembling the optimizations in (9) and (10) but now in the population (hence the infeasibility of the estimator $\widehat{h}^{\text{inf}}_{TCC}$),

$$\begin{aligned}
\lambda^*_h(\gamma) \;&:=\; \arg\min_{\lambda\in[0,1]} \; \sigma^2_{h,UCC}(\beta^0, \gamma, \lambda), \\
\gamma^*_{h,TCC} \;&:=\; \begin{cases} \arg\min_{\gamma\in\Gamma} \; \sigma^2_{h,UCC}(\beta^0, \gamma, \lambda^*_h(\gamma)) & \text{if} \;\; \lambda^*_h(\gamma) \neq 0, \\ \gamma^{\text{hom}} \;\; \text{if} \;\; \lambda^*_h(\gamma) = 0, \end{cases} \\
\widehat{h}^{\text{inf}}_{TCC} \;&:=\; (1-\lambda^*_h(\gamma^*_{h,TCC}))h(\widehat{\beta}_{OLS}) + \lambda^*_h(\gamma^*_{h,TCC})h(\widehat{\beta}(\gamma^*_{h,TCC})).
\end{aligned}$$

Thus, using (9) and (10), we obtain the standard error of our TCC estimator $\widehat{h}_{TCC}$ as:

$$se_{h,TCC} \;:=\; \left[\frac{1}{n}\widehat{\sigma}^2_{h,UCC}(\widehat{\beta}_{OLS}, \widehat{\gamma}_{h,TCC}, \widehat{\lambda}(\widehat{\gamma}_{h,TCC}))\right]^{1/2}$$

where $\widehat{\sigma}^2_{h,UCC}(.)$ is the sample analog of $\sigma^2_{h,UCC}(.)$ with each argument replaced by its consistent estimator. While the expressions for the quantities used here are well known (see, e.g., DiCiccio et al. (2019)), we collect them in Appendix A.1 for ready reference.

3.2 Targeted GMM (TGMM) estimator

TCC was concerned with combining scalar estimators $h(\widehat{\beta}_{OLS})$ and $h(\widehat{\beta}(\gamma))$. Now we focus on constructing a targeted matrix combination, with matrix weights adding up to identity, of the vector estimators $\widehat{\beta}_{OLS}$ and $\widehat{\beta}(\gamma)$ and then consider the $h(.)$ function of that estimator. One could call this a matricial convex combination (MCC) estimator and its targeted version the Targeted MCC (TMCC) estimator. However, since such matrix combinations can be efficiently obtained by GMM using the stacked OLS and UWLS(γ) moment vectors for β, and Lu and Wooldridge (2020) already consider the un-targeted GMM version combining OLS and UWLS(γ_{WLS}) and call it GMM, we will call TMCC the Targeted GMM (TGMM) estimator. We will follow their GMM representation. In Appendix B we discuss general combinations of exactly identifying moment restrictions revisiting Chen et al. (2016).

The OLS and UWLS(γ) moment vectors, and the stacked moment vector are:

$$g_1(y_i, x_i, \beta) := x_i(y_i - x_i'\beta), \quad g_2(y_i, x_i, \beta, \gamma) := \frac{x_i(y_i - x_i'\beta)}{\omega^2(x_i;\gamma)} \quad \text{and} \quad g(y_i, x_i, \beta, \gamma) := \begin{bmatrix} g_1(y_i, x_i, \beta) \\ g_2(y_i, x_i, \beta, \gamma) \end{bmatrix}.$$

As before, we will ignore in the sequel matters related to the estimation of γ since such estimation does not affect the first-order asymptotic variance of any estimators of β based on these moment vectors; see, e.g., Section 6 of Newey and McFadden (1994). This is because the conditional mean independence condition in (2) implies that the stacked moment vector satisfies:

$$E[g(y_i, x_i, \beta^0, \gamma)] \;=\; 0 \quad \text{any} \quad \gamma \in \Gamma.$$

Standard regularity conditions give the asymptotic distribution of this moment vector as:

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} g(y_i, x_i, \beta^0, \gamma) \xrightarrow{d} N\left(0, V\left(\beta^0, \gamma\right) := \begin{bmatrix} V_{11}\left(\beta^0\right) := E\left[x_i x_i' \omega_0^2\left(x_i\right)\right] & V_{12}\left(\beta^0, \gamma\right) := E\left[\frac{x_i x_i' \omega_0^2(x_i)}{\omega^2(x_i,\gamma)}\right] \\ V_{21}\left(\beta^0, \gamma\right) := V_{12}\left(\beta^0, \gamma\right) & V_{22}\left(\beta^0, \gamma\right) := E\left[\frac{x_i x_i' \omega_0^2(x_i)}{\omega^4(x_i,\gamma)}\right] \end{bmatrix}\right).$$

An example of a natural estimator of $V\left(\beta^0,\gamma\right)$ is:

$$\widehat{V}\left(\widehat{\beta}_{OLS},\gamma\right) := \left[\begin{array}{cc} \widehat{V}_{11}\left(\widehat{\beta}_{OLS}\right) := \frac{1}{n}\sum_{i=1}^{n} x_i x_i' \widehat{u}_{i,OLS}^2 & \widehat{V}_{12}\left(\widehat{\beta}_{OLS},\gamma\right) := \frac{1}{n}\sum_{i=1}^{n} \frac{x_i x_i' \widehat{u}_{i,OLS}^2}{\omega^2(x_i;\gamma)} \\ \widehat{V}_{21}\left(\widehat{\beta}_{OLS},\gamma\right) := \widehat{V}_{12}\left(\widehat{\beta}_{OLS},\gamma\right) & \widehat{V}_{22}\left(\widehat{\beta}_{OLS},\gamma\right) := \frac{1}{n}\sum_{i=1}^{n} \frac{x_i x_i' \widehat{u}_{i,OLS}^2}{\omega^4(x_i;\gamma)} \end{array}\right].$$

Since $V\left(\beta^0,\gamma\right)$ is singular under conditional homoskedasticity, we will use the Moore-Penrose (MP) inverse of $\widehat{V}\left(\widehat{\beta}_{OLS},\gamma\right)$, denoted by $\left[\widehat{V}\left(\widehat{\beta}_{OLS},\gamma\right)\right]^{+}$, as the weighting matrix in our targeted GMM strategy and appeal to Xiao (2020) for convergence of this MP inverse.[2]

Accordingly, we define the "User-specified GMM" (UGMM(γ)) estimator of $h(\beta)$ as:

$$\widehat{h}_{UGMM}(\gamma) \;:=\; h(\widehat{\beta}_{UGMM}(\gamma))$$

that is based on plugging in the efficient GMM estimator of β (for given γ):

$$\widehat{\beta}_{UGMM}(\gamma) \;:=\; \arg\min_{\beta}\left(\frac{1}{n}\sum_{i=1}^{n} g(y_i,x_i,\beta,\gamma)\right)' \left[\widehat{V}\left(\widehat{\beta}_{OLS},\gamma\right)\right]^{+} \left(\frac{1}{n}\sum_{i=1}^{n} g(y_i,x_i,\beta,\gamma)\right).$$

[2] Lu and Wooldridge (2020) propose to conduct efficient GMM estimation using $\left[\widehat{V}\left(\widehat{\beta}_{OLS},\gamma\right)\right]^{-1}$ as the weighting matrix, and plugging in for γ the WLS or a quasi-maximum likelihood estimator of γ. However, then the weighting matrix will not exist in the limit under the very important case of conditional homoskedasticity because $V\left(\beta^0,\gamma_{WLS}\right)$ becomes singular as, now, $\gamma_{WLS} = \gamma^{\text{hom}}$ and that implies $\omega^2(x_i;\gamma_{WLS}) = \omega_0^2(x_i) = \omega_{\text{hom}}^2$ (a constant), and hence:

$$V\left(\beta^0,\gamma_{WLS}\right) \;=\; V\left(\beta^0,\gamma^{\text{hom}}\right) \;=\; \left[\begin{array}{cc} \omega_{\text{hom}}^2 & 1 \\ 1 & 1/\omega_{\text{hom}}^2 \end{array}\right] \otimes E[x_i x_i'].$$

The possibility that $V\left(\beta^0,\gamma\right)$ may become (near) singular may lead, following White (1986), to use a MP inverse of $V\left(\beta^0,\gamma\right)$. However, it has been known since Stewart (1969) that the generalized inverse is not always continuous, and that a sequence of pseudo-inverses A_n^{+} converges toward the pseudo-inverse A^{+} of the limit if and only if $rank(A_n) = rank(A)$ for n large enough. Andrews (1987) points out the restrictiveness of this condition for the practice of Wald testing. Fortunately, Xiao (2020) has shown that using generalized inverses in the context of GMM optimal weighting matrix is sound because Stewart's continuity condition is fulfilled in this case. A caveat noted by Xiao (2020) is that using generalized inverse for efficient GMM, although theoretically sound, may be unstable since a "small perturbation of a singular matrix may result in large deviations for its generalized inverses". The bottom line is that it is theoretically justified to solve the problem with Lu and Wooldridge (2020)'s weighting matrix by using $\left[\widehat{V}\left(\widehat{\beta}_{OLS},\widehat{\gamma}_{WLS}\right)\right]^{+}$ as the weighting matrix. Using $\left[\widehat{V}\left(\widehat{\beta}_{OLS},\gamma\right)\right]^{+}$ as the weighting matrix in our targeted GMM strategy also works well in our simulations. In the sequel we will use the superscript "+" to denote the MP inverse of matrices.

Thanks to Theorem 3.1 and the discussion below Theorem 4.1 in Xiao (2020), we have:

$$\sqrt{n}(\widehat{h}_{UGMM}(\gamma) - h(\beta^0)) \xrightarrow{d} N\left(0, \sigma^2_{h,UGMM}\left(\beta^0, \gamma, \left[V(\beta^0,\gamma)\right]^+\right)\right)$$

under GMM-regularity conditions and mean independence (2). The asymptotic variance is:

$$\sigma^2_{h,UGMM}\left(\beta^0, \gamma, \left[V(\beta^0,\gamma)\right]^+\right) := \frac{\partial h(\beta^0)}{\partial \beta'}\left(E\left[\frac{\partial g'(y_i,x_i;\beta^0,\gamma)}{\partial \beta}\right]\left[V(\beta^0,\gamma)\right]^+ E\left[\frac{\partial g(y_i,x_i;\beta^0,\gamma)}{\partial \beta'}\right]\right)^{-1}\frac{\partial h(\beta^0)}{\partial \beta}$$

with $E\left[\frac{\partial g'(y_i,x_i;\beta^0,\gamma)}{\partial \beta}\right] = \left[E\left[x_i x_i'\right], E\left[\frac{x_i x_i'}{\omega^2(x_i,\gamma)}\right]\right]$. Let $\widehat{\sigma}^2_{h,UGMM}\left(\widehat{\beta}_{OLS}, \gamma, \left[\widehat{V}(\widehat{\beta}_{OLS},\gamma)\right]^+\right)$ be its estimator with the population quantities replaced by sample analogs and β^0 by $\widehat{\beta}_{OLS}$.

We can now define our TGMM estimator as:

$$\widehat{h}_{TGMM} := \widehat{h}_{UGMM}(\widehat{\gamma}_{h,TGMM})$$

where:

$$\widehat{\gamma}_{h,TGMM} := \arg\min_{\gamma\in\Gamma} \widehat{\sigma}^2_{h,UGMM}\left(\widehat{\beta}_{OLS}, \gamma, \left[\widehat{V}(\widehat{\beta}_{OLS},\gamma)\right]^+\right).$$

Like the other feasible targeted estimators, TGMM is also asymptotically equivalent to its infeasible version, in this case the estimator $\widehat{h}^{\text{inf}}_{TGMM}$, that is optimal in its class. That is:

$$\sqrt{n}\left(\widehat{h}_{TGMM} - h(\beta^0)\right) = \sqrt{n}\left(\widehat{h}^{\text{inf}}_{TGMM} - h(\beta^0)\right) + o_p(1) \xrightarrow{d} N\left(0, \sigma^2_{h,UGMM}\left(\beta^0, \gamma^*_{h,TGMM}, \left[V(\beta^0,\gamma^*_{h,TGMM})\right]^+\right)\right)$$

where:

$$\widehat{h}^{\text{inf}}_{TGMM} := \widehat{h}_{UGMM}(\gamma^*_{h,TGMM})$$

with:

$$\gamma^*_{h,TGMM} := \arg\min_{\gamma\in\Gamma} \sigma^2_{h,UGMM}\left(\beta^0, \gamma, \left[V(\beta^0,\gamma)\right]^+\right).$$

The standard error of our TGMM estimator $\widehat{h}_{TGMM}$ is:

$$se_{h,TGMM} := \left[\frac{1}{n}\widehat{\sigma}^2_{h,UGMM}\left(\widehat{\beta}_{OLS}, \widehat{\gamma}_{h,TGMM}, \left[\widehat{V}(\widehat{\beta}_{OLS},\widehat{\gamma}_{h,TGMM})\right]^+\right)\right]^{1/2}.$$

3.3 Optimality among the three strategies for targeting

Optimality of asymptotic variance with respect to γ is the main message of our paper. TWLS, TCC and TGMM are respectively optimal in this sense in the three classes of estimators $\widehat{h}_{UWLS}(\gamma)$, $\widehat{h}_{UCC}(\gamma)$ and $\widehat{h}_{UGMM}(\gamma)$. But which estimator among TWLS, TCC and TGMM is preferred? Proposition 1 answers this question as far as the asymptotic variance is concerned.

Proposition 1 *Under standard conditions maintained throughout our paper:*

$$AVar\left(\widehat{h}_{TGMM}\right) \leq AVar\left(\widehat{h}_{TCC}\right) \leq AVar\left(\widehat{h}_{TWLS}\right).$$

Appendix A.2 proves this result. An extensive discussion on related matters can be found in Appendix B.3. Despite the clean result in Proposition 1, two reasons — (i) computational convenience and (ii) the fact that our simulation results find TGMM may be flawed by some small-sample issues — warrant a maintained interest in TCC and TWLS.

4 Numerical evidence of small-sample properties

We explore numerically the small-sample performance of our proposed targeted estimators TWLS, TCC and TGMM, and compare them with that of their main competitors — OLS, WLS, ALS, MIN, CC and GMM — under various designs recently used in this literature.

The message of our numerical results is that if the user's model $\omega^2(x_i;\gamma)$ for $\omega_0^2(x_i)$ allows for improvement in precision over others then our proposed targeting estimators achieve it. Improvements in empirical mean squared error (EMSE) by our proposed estimators can be huge. There does not seem to be any major cost in terms of empirical bias (unreported), empirical size (reported), etc. to using our proposed estimators. Comparison among our proposed estimators TWLS, TCC and TGMM does not however give a clear winner, although simplicity of computation might lead some user to prefer TCC or TWLS in practice.

4.1 Under the design in Romano and Wolf (2017)

Let $y_i = \beta_1^0 + \beta_2^0 x_{i,2} + u_i$ with $x_{i,2} \sim U(1,4)$ and $u_i|x_i \sim N(0, \omega_0^2(x_i))$ i.i.d. for $i = 1, \ldots, n$. Let $\beta^0 = (0,0)'$. Romano and Wolf (2017) consider 10 cases for $\omega_0^2(x_i)$:

Case 1: (a) $\omega_0^2(x_i) = 1$; (b) $\omega_0^2(x_i) = x_{i,2}$; (c) $\omega_0^2(x_i) = x_{i,2}^2$; (d) $\omega_0^2(x_i) = x_{i,2}^4$.

Case 2: (a) $\omega_0^2(x_i) = (\log(x_{i,2}))^2$; (b) $\omega_0^2(x_i) = (\log(x_{i,2}))^4$.

Case 3: (a) $\omega_0^2(x_i) = \exp\left(.1(x_{i,2} + x_{i,2}^2)\right)$; (b) $\omega_0^2(x_i) = \exp\left(.15(x_{i,2} + x_{i,2}^2)\right)$.

Case 4: (a) $\omega_0^2(x_i) = \begin{cases} 1 & \text{if } x_{i,2} < 2 \\ 2 & \text{if } 2 \leq x_{i,2} < 3 \\ 3 & \text{if } x_{i,2} \geq 3 \end{cases}$; (b) $\omega_0^2(x_i) = \begin{cases} 1 & \text{if } x_{i,2} < 2 \\ 2^2 & \text{if } 2 \leq x_{i,2} < 3 \\ 3^2 & \text{if } x_{i,2} \geq 3 \end{cases}$.

Romano and Wolf (2017) consider two parametric models $\omega^2(x_i;\gamma)$ — Model 1: $\omega^2(x_i;\gamma) = \exp(\gamma_1 + \gamma_2 \log(x_{i,2}))$ and Model 2: $\omega^2(x_i;\gamma) = \exp(\gamma_1 + \gamma_2 x_{i,2})$ — for $\omega_0^2(x_i)$. Like theirs, our results here are also very similar for both models. For brevity we will report any results related to Romano and Wolf (2017) based on Model 1 only since it is their preferred model.

Taking sample size $n = 50, 100, 200, 400$, we will, like Romano and Wolf (2017), report results on the target $h(\beta) = \beta_2$, i.e., the slope. The results are similar for $h(\beta) = \beta_1$.

As a measure of precision, Table 2 presents the ratio of the EMSE of each estimator with respect to that of OLS. All estimators behave similar to OLS in Case 1(a) (conditional homoskedasticity, i.e., when OLS is efficient), especially for $n \geq 100$. In other cases the other estimators lead to smaller, sometimes much smaller, EMSE than OLS. Model 1 is correct for $\omega_0^2(x_i)$ under Cases 1(a)-1(d) with $\gamma_2^0 = 0, 1, 2, 4$ respectively, and hence WLS is efficient in these cases (jointly with OLS in Case 1(a)). All recently proposed estimators perform similar to WLS in these cases, especially if n is not too small. We observe that our proposed estimators either perform very similar to the other estimators that they are supposed to improve upon, or lead to really big gains in precision as in Cases 2 (a) and (b). These gains

in precision can be more dramatic if heteroskedasticity is more severe, and to demonstrate that we added a case, Case 2(c): $\omega_0(x_i) = [\log(x_i)]^6$, in this tabular display in Table 2.

It is interesting that simply by targeting, TWLS can have smaller EMSE in some cases than the un-targeted CC and GMM estimators of DiCiccio et al. (2019) and Lu and Wooldridge (2020) respectively; e.g., Cases 2 (a)-(c). To put it in context, recall that TWLS utilizes the WLS framework only whereas CC and GMM combine WLS with OLS. The impressive improvement due to TWLS over CC or GMM, and similarly due to TCC over GMM is case-specific and not part of our general theory. Hence, it is noteworthy that we observe such improvements in the simulations due to targeted management of the nuisance parameters γ.

Table 3 presents the empirical size (empirical rejection probability of the truth) of the 5% Wald tests based on each estimator. Any size distortions vanish as sample size increases.

4.2 Under the design in Lu and Wooldridge (2020)

Let $y_i = \beta_1^0 + \beta_2^0 x_{i,2} + \beta_3^0 x_{i,3} + \beta_4^0 x_{i,4} + u_i$ with $x_{i,2} \sim N(1,1)$, $x_{i,3} = .8 + .2x_{i,2} + e_{i,1}$, $x_{i,4} = 1(x_{i,5} > x_{i,3})$, $u_i = s(x_i)e_{i,3}$ where $e_{i,1}, e_{i,2}, e_{i,3}$ are independent $N(0,1)$, and $x_{i,5} = .3 + .1x_{i,2} + .1x_{i,3} + e_{i,2}$. All the variables are i.i.d. for $i = 1, \ldots, n$. Let $\beta^0 = (.5, 1, 1, 1)'$, $x_i = (1, x_{i,2}, x_{i,3}, x_{i,4})'$ and $e_{i,3}$ as independent of x_i. Thus, $E[u_i|x_i] = 0$ and $V(u_i|x_i) \equiv \omega_0^2(x_i) = s^2(x_i)$. Lu and Wooldridge (2020) consider 4 cases for $\omega_0^2(x_i)$:

Case 1: $\omega_0^2(x_i) = (\beta_1^0 + \beta_2^0 x_{i,2} + \beta_3^0 x_{i,3} - 3\beta_4^0 x_{i,4} + .1x_{i,2}(x_{i,3} + x_{i,4}) - .1x_{i,3}x_{i,4} - .05x_{i,2}^2 + .05x_{i,3}^2))^2$.

Case 2: $\omega_0^2(x_i) = (\beta_1^0 + \beta_2^0|x_{i,2}| + \beta_3^0 x_{i,3}^2 + \beta_4^0 x_{i,4})^2$.

Case 3: $\omega_0^2(x_i) = \exp(\beta_1^0 + \beta_2^0|x_{i,2}| + \beta_4^0 x_{i,4})$.

Case 4: $\omega_0^2(x_i) = \exp(\beta_1^0 + \beta_2^0 x_{i,2} + \beta_3^0 x_{i,3} + \beta_4^0 x_{i,4})$.

Lu and Wooldridge (2020) use $\omega^2(x_i;\gamma) = \exp(x_i'\gamma) = \exp(\gamma_1 + \gamma_2 x_{i,2} + \gamma_3 x_{i,3} + \gamma_4 x_{i,4})$ as the user's parametric model. This model is correct for $\omega_0^2(x_i)$ with $\gamma^0 = \beta^0$ in Case 4 and

no refinement can improve upon WLS in Case 4 since WLS is optimal. We omit Case 4 for brevity (less congested tables) since the observations on the finite-sample properties of the refinements are similar to what we already saw under correct specification in the four cases 1(a)-1(d) in Tables 2 and 3 for Section 4.1: extra noise in the refinements of WLS, especially ours, affect both EMSE and empirical size but they recover as sample size increases.

We consider four different targets $h(\beta) = \beta_1, \beta_2, \beta_3, \beta_4$. We take the sample size $n = 1000, 2000$ (Lu and Wooldridge (2020) take $n = 1000, 10000$). Our results for WLS and GMM differ from Lu and Wooldridge (2020); they use Gamma quasi-maximum likelihood estimator for γ whereas we use $\widehat{\gamma}_{WLS}$ to maintain uniformity with the rest of the simulations.

Table 4 presents the ratio of the EMSE of each estimator with respect to that of OLS. We see that WLS based on an incorrect model $\omega^2(x_i;\gamma)$ in Cases 1 and 2 can be much less precise than OLS. This is a possibility that DiCiccio et al. (2019) noted to motivate their MIN and CC estimators but conjectured as "rare". ALS is also much less precise than OLS in this case since ALS and WLS are almost identical here because of the high level of conditional heteroskedasticity of u_i. On the other hand, the MIN, CC and GMM estimators deliver big gains in precision over OLS (and WLS and ALS). Additionally, when the parametric model $\omega^2(x_i;\gamma)$ is far from correct for $\omega_0^2(x_i)$, i.e., in Cases 1 and 2, we see that our proposed estimators TWLS, TCC and TGMM deliver even further substantial gains in precision.

Under Cases 1 and 2, i.e., when the user's model $\omega^2(x_i;\gamma)$ is "more" incorrect for $\omega_0^2(x_i)$, we again find here that simply by targeting, TWLS can be more precise than CC and GMM, and TCC than GMM — a pattern of improvement that was not predicted by general theory.

Table 5 presents the empirical size of the 5% Wald tests based on each estimator. The results look reasonable except that in some cases with the smaller sample, TCC and TGMM have empirical size that is noticeably larger than their nominal level of 5% — as high as 9.3% and 9.6% for TCC and TGMM respectively. However, it is also evident that this problem disappears with the increase in the sample size.

$V(u\|x)$	n	WLS	ALS	MIN	TWLS	CC	TCC	GMM	TGMM
	50	1.044	1.044	1.029	1.089	1.024	1.060	1.090	1.099
Case	100	1.033	1.033	1.020	1.045	1.015	1.048	1.057	1.068
(1a)	200	1.011	1.011	1.007	1.019	1.005	1.021	1.030	1.037
	400	1.005	1.005	1.003	1.006	1.002	1.009	1.014	1.016
	50	.935	.955	.941	.974	.936	.987	.980	.990
Case	100	.918	.929	.927	.933	.920	.961	.943	.954
(1b)	200	.900	.901	.907	.902	.903	.916	.909	.915
	400	.905	.905	.906	.907	.907	.914	.911	.913
	50	.756	.762	.766	.765	.760	.784	.795	.809
Case	100	.698	.698	.702	.707	.701	.722	.720	.736
(1c)	200	.687	.687	.686	.684	.688	.689	.692	.700
	400	.679	.679	.679	.678	.681	.679	.682	.686
	50	.437	.437	.438	.407	.441	.414	.447	.455
Case	100	.330	.330	.330	.311	.331	.315	.334	.341
(1d)	200	.334	.334	.334	.324	.334	.327	.336	.343
	400	.304	.304	.304	.301	.304	.302	.304	.308
	50	.579	.579	.579	.547	.583	.517	.572	.568
Case	100	.555	.555	.555	.512	.556	.452	.520	.481
(2a)	200	.499	.499	.499	.453	.500	.408	.451	.412
	400	.482	.482	.482	.472	.483	.431	.459	.412
	50	.396	.396	.397	.332	.400	.282	.381	.367
Case	100	.301	.301	.301	.205	.302	.163	.283	.261
(2b)	200	.280	.280	.280	.205	.280	.173	.243	.213
	400	.270	.270	.270	.204	.270	.162	.233	.182
	50	.289	.289	.289	.062	.290	.059	.235	.224
Case	100	.159	.159	.159	.028	.159	.027	.143	.131
(2c)	200	.152	.152	.152	.031	.152	.030	.110	.098
	400	.142	.142	.142	.028	.142	.027	.090	.067
	50	.868	.903	.882	.891	.874	.929	.896	.907
Case	100	.854	.862	.860	.858	.858	.884	.863	.873
(3a)	200	.818	.819	.821	.823	.819	.836	.813	.823
	400	.828	.828	.828	.831	.830	.833	.826	.830
	50	.701	.713	.713	.716	.711	.738	.711	.725
Case	100	.678	.681	.680	.677	.682	.689	.668	.683
(3b)	200	.634	.634	.634	.633	.636	.637	.610	.619
	400	.651	.651	.651	.652	.652	.653	.630	.633
	50	.965	.973	.972	.991	.959	1.014	1.003	1.011
Case	100	.948	.949	.959	.952	.945	.978	.969	.976
(4a)	200	.927	.927	.934	.928	.923	.953	.935	.942
	400	.929	.929	.937	.927	.927	.932	.934	.936
	50	.795	.809	.815	.825	.803	.844	.837	.838
Case	100	.752	.753	.762	.757	.755	.776	.775	.783
(4b)	200	.735	.735	.738	.730	.732	.739	.736	.736
	400	.744	.744	.745	.732	.736	.735	.735	.730

Table 2: Ratio of empirical MSEs with respect to that of OLS estimator of $h(\beta) = \beta_2$ based on 10000 Monte Carlo trials under the design of Romano and Wolf (2017) and using their Model 1.

$V(u\|x)$	n	OLS	WLS	ALS	MIN	TWLS	CC	TCC	GMM	TGMM
	50	5.1	5.7	5.7	5.9	5.8	5.8	6.9	7.4	8.0
Case	100	4.8	5.1	5.1	5.1	5.1	5.1	5.9	6.0	6.5
(1a)	200	4.8	4.8	4.8	4.9	4.8	4.9	5.3	5.4	5.6
	400	4.8	5.0	5.0	5.0	4.9	5.0	5.1	5.2	5.2
	50	4.5	4.9	5.1	5.2	5.2	5.2	6.3	6.6	7.3
Case	100	5.4	5.7	5.9	5.9	5.6	5.8	6.8	6.8	7.2
(1b)	200	4.6	4.9	4.9	5.0	4.8	5.0	5.2	5.2	5.5
	400	5.3	5.3	5.3	5.3	5.2	5.3	5.4	5.5	5.6
	50	4.8	5.3	5.4	5.5	5.5	5.6	6.3	6.7	7.5
Case	100	4.8	5.3	5.3	5.3	5.3	5.4	6.0	5.9	6.7
(1c)	200	4.9	4.9	4.9	4.9	5.0	5.0	5.4	5.5	5.8
	400	4.8	5.2	5.2	5.2	5.1	5.2	5.2	5.4	5.5
	50	5.4	5.7	5.7	5.7	5.1	5.9	5.7	6.2	7.0
Case	100	5.5	5.1	5.1	5.1	4.8	5.2	5.2	4.7	5.3
(1d)	200	5.4	5.8	5.8	5.8	5.6	5.9	5.8	5.5	6.0
	400	5.2	5.0	5.0	5.0	4.9	5.0	5.0	4.9	5.2
	50	5.0	5.0	5.0	5.0	4.7	5.2	5.4	5.9	6.1
Case	100	5.2	5.1	5.1	5.2	5.0	5.2	5.2	5.4	5.3
(2a)	200	5.2	5.5	5.5	5.5	5.1	5.5	5.4	5.1	4.9
	400	5.3	5.0	5.0	5.0	5.0	5.0	5.1	5.2	4.9
	50	5.4	5.5	5.5	5.5	6.8	5.6	6.6	5.2	5.4
Case	100	5.7	5.5	5.5	5.5	6.0	5.5	5.3	4.8	4.7
(2b)	200	5.1	5.3	5.3	5.3	5.7	5.3	5.7	4.4	3.8
	400	5.0	5.0	5.0	5.0	5.4	5.0	4.9	4.5	3.9
	50	5.3	5.7	5.7	5.7	4.6	5.7	4.4	4.1	4.2
Case	100	5.2	5.2	5.2	5.2	4.3	5.2	4.3	2.9	2.9
(2c)	200	4.8	5.5	5.5	5.5	4.8	5.5	5.0	2.9	2.3
	400	5.0	5.2	5.2	5.2	4.8	5.2	4.7	3.3	2.1
	50	4.8	5.2	5.5	5.5	5.4	5.5	6.7	7.0	7.5
Case	100	5.1	5.6	5.7	5.7	5.4	5.8	6.2	6.3	6.7
(3a)	200	5.0	5.1	5.1	5.1	5.0	5.1	5.4	5.7	6.0
	400	4.7	5.1	5.1	5.1	5.0	5.1	5.0	5.4	5.5
	50	4.9	5.1	5.2	5.4	5.2	5.4	5.8	6.2	7.0
Case	100	5.2	5.5	5.5	5.5	5.2	5.7	5.6	6.2	6.8
(3b)	200	5.3	5.2	5.2	5.2	5.1	5.2	5.2	5.5	5.8
	400	4.6	4.9	4.9	4.9	4.9	4.9	4.9	5.2	5.2
	50	4.8	5.4	5.4	5.7	5.3	5.6	6.8	7.0	7.3
Case	100	5.0	5.6	5.6	5.7	5.3	5.7	6.2	6.4	6.7
(4a)	200	4.9	5.0	5.0	5.1	5.0	5.1	5.4	5.6	5.8
	400	4.7	5.2	5.2	5.2	5.1	5.2	5.2	5.4	5.5
	50	4.7	4.8	4.9	5.3	4.9	5.3	6.0	6.3	6.8
Case	100	5.2	5.1	5.1	5.3	5.1	5.5	5.8	6.3	7.0
(4b)	200	5.3	5.0	5.0	5.0	5.0	5.3	5.5	5.6	5.9
	400	4.6	4.8	4.8	4.8	4.9	5.0	5.0	5.2	5.3

Table 3: Empirical size (in %) of 5% Wald test for $h(\beta) = \beta_2$ based on 10000 Monte Carlo trials under the simulation design of Romano and Wolf (2017) and using their Model 1.

$V(u\|x)$	$h(\beta)$	$n=1000$								$n=2000$							
		WLS	ALS	MIN	TWLS	CC	TCC	GMM	TGMM	WLS	ALS	MIN	TWLS	CC	TCC	GMM	TGMM
Case (1)	β_1	.806	.806	.788	.553	.781	.510	.631	.442	.800	.800	.791	.578	.787	.534	.637	.441
	β_2	.982	.982	.926	.817	.908	.901	.859	.697	.994	.994	.949	.810	.924	.827	.876	.673
	β_3	.874	.874	.840	.685	.827	.706	.714	.562	.888	.888	.871	.676	.852	.700	.748	.545
	β_4	1.561	1.561	1.000	.898	1.001	.988	.844	.786	1.587	1.587	1.000	.904	1.001	.993	.864	.783
Case (2)	β_1	1.422	1.422	.878	.530	.785	.562	.405	.380	1.587	1.587	.944	.556	.831	.584	.428	.394
	β_2	1.403	1.403	.946	.743	.847	.764	.813	.765	1.531	1.531	.986	.757	.890	.762	.851	.767
	β_3	2.235	2.235	.932	.792	.815	.621	.451	.421	2.590	2.590	.987	.590	.861	.615	.462	.414
	β_4	1.354	1.354	.902	.669	.784	.671	.616	.555	1.430	1.430	.959	.682	.811	.677	.635	.561
Case (3)	β_1	.870	.870	.871	.867	.869	.861	.829	.795	.865	.865	.865	.857	.866	.854	.824	.776
	β_2	.814	.814	.811	.796	.806	.817	.783	.690	.807	.807	.807	.791	.804	.795	.784	.656
	β_3	.814	.814	.815	.801	.812	.853	.794	.823	.804	.804	.805	.792	.805	.803	.792	.798
	β_4	.962	.962	.960	.944	.956	.954	.867	.801	.962	.962	.962	.951	.959	.956	.865	.795

Table 4: Ratio of empirical MSE of estimators with respect to empirical MSE of OLS estimator of various $h(\beta)$'s based on 10000 Monte Carlo trials under the design of Lu and Wooldridge (2020).

$V(u\|x)$	$h(\beta)$	$n=1000$									$n=2000$								
		OLS	WLS	ALS	MIN	TWLS	CC	TCC	GMM	TGMM	OLS	WLS	ALS	MIN	TWLS	CC	TCC	GMM	TGMM
Case 1	β_1	4.8	5.1	5.1	5.2	5.9	5.2	5.7	5.2	7.4	4.8	5.1	5.1	5.1	5.6	5.1	5.2	5.2	6.2
	β_2	5.0	5.3	5.3	5.5	7.4	5.4	9.3	5.9	9.6	5.1	5.1	5.1	5.3	5.6	5.1	6.8	5.3	7.1
	β_3	5.1	4.9	4.9	5.0	6.8	4.8	8.8	5.5	8.1	5.0	5.4	5.4	5.5	6.0	5.5	7.5	5.8	6.8
	β_4	5.1	4.7	4.7	5.1	5.1	5.1	5.2	5.1	5.8	4.8	4.7	4.7	4.8	5.1	4.8	4.9	5.2	5.4
Case 2	β_1	4.6	4.9	4.9	5.4	5.8	5.1	6.0	4.8	6.1	4.6	5.0	5.0	5.3	5.5	5.2	5.8	5.0	5.3
	β_2	5.1	4.7	4.7	5.4	6.0	5.4	6.6	5.6	8.5	4.9	5.0	5.0	5.2	5.5	5.1	5.5	5.2	6.5
	β_3	4.7	5.1	5.1	5.7	6.8	5.7	7.3	5.4	7.5	4.8	4.9	4.9	5.6	6.2	5.4	6.3	5.5	6.4
	β_4	4.8	4.1	4.1	4.6	4.8	4.7	4.8	4.8	5.2	4.8	4.5	4.5	5.1	4.9	4.8	4.9	4.7	4.8
Case 3	β_1	4.8	4.8	4.8	4.8	5.1	4.9	5.5	5.0	6.3	4.9	4.8	4.8	4.8	4.9	4.8	5.1	4.7	5.4
	β_2	5.4	5.6	5.6	5.7	6.0	5.7	6.7	6.0	7.2	5.1	5.3	5.3	5.3	5.5	5.3	5.6	5.2	6.2
	β_3	5.1	5.0	5.0	5.0	5.2	5.1	6.7	5.4	8.5	5.3	5.3	5.3	5.3	5.4	5.3	5.6	5.4	6.9
	β_4	4.8	4.8	4.8	4.8	4.7	4.9	4.9	5.1	4.9	4.8	4.9	4.9	4.9	4.7	4.8	4.8	4.8	5.0

Table 5: Empirical size (in %) of 5% Wald test for $h(\beta) := \beta_j$ for $j = 1, \ldots, 4$ based on 10000 Monte Carlo trials under the simulation design of Lu and Wooldridge (2020).

4.3 Based on the real-life example in Romano and Wolf (2017)

We already discussed the setup of this simulation experiment based on the Boston housing data with $n = 506$ in Section 2.5; also see DiCiccio et al. (2019) and Miller and Startz (2019). So, we do not repeat that here. The target $h(\beta)$'s are the regression coefficients $\beta_1, \ldots, \beta_5$.

Model 1: $\omega^2(x_i; \gamma) = \exp(\gamma_1 + \sum_{k=2}^{5} \log(|x_{i,k}|))$ is Romano and Wolf (2017)'s preferred model. Model 1 led to better relative performance of their proposed estimator with respect to OLS in their simulations. We report the further improvement provided by our proposed estimators based on Romano and Wolf (2017)'s Model 1. These are reported in Table 6 for the ratio of the EMSE's with respect to OLS, and in Table 7 for the ratio of the average length of the 95% Wald confidence intervals based on other estimators to that based on OLS (and empirical size of the 5% Wald test within parentheses). It is evident that our proposed estimators deliver big gains over their respective competitors, especially for $\beta_1, \ldots, \beta_4$.

$h(\beta)$	WLS	ALS	MIN	TWLS	CC	TCC	GMM	TGMM
β_1	.613		.613	.501	.613	.500	.492	.469
β_2	.676	same	.675	.562	.675	.559	.558	.486
β_3	.506	as	.506	.337	.506	.337	.372	.332
β_4	.500	WLS	.500	.348	.501	.350	.341	.317
β_5	.927		.917	.883	.904	.896	.814	.774

Table 6: Ratio of EMSE with respect to OLS based on 10000 Monte Carlo trials under Romano and Wolf (2017)'s design with real-life data [c.f. their Table C7] and using their Model 1.

$h(\beta)$	OLS	WLS	ALS	MIN	TWLS	CC	TCC	GMM	TGMM
β_1	1 (4.5)	.779 (5.0)		.779 (5.0)	.671 (6.2)	.779 (5.0)	.670 (6.3)	.691 (5.2)	.643 (6.4)
β_2	1 (4.6)	.812 (5.2)	same	.812 (5.2)	.736 (5.1)	.812 (5.2)	.734 (5.1)	.735 (5.4)	.670 (5.6)
β_3	1 (4.5)	.713 (4.6)	as	.713 (4.6)	.577 (5.1)	.713 (4.6)	.576 (5.2)	.609 (4.8)	.558 (5.6)
β_4	1 (4.4)	.710 (5.1)	WLS	.710 (5.1)	.555 (7.4)	.710 (5.1)	.551 (7.8)	.590 (5.2)	.531 (7.4)
β_5	1 (4.7)	.953 (4.9)		.944 (5.0)	.941 (4.6)	.939 (4.9)	.943 (4.7)	.883 (5.2)	.839 (5.7)

Table 7: Ratio of the average length of 95% Wald confidence interval with respect to that of OLS. Within parenthesis is the empirical size of 5% Wald tests. Based on 10000 Monte Carlo trials under Romano and Wolf (2017)'s design with real-life data [c.f. their Table C8], using their Model 1.

4.4 Empirical illustration in Lu and Wooldridge (2020)

Lu and Wooldridge (2020) use the well-known data set '401ksubs' (see Wooldridge (2012)) to estimate a regression with: $E[y_i|x_i] = x_i'\beta^0 = \beta_1^0 + \beta_2^0 x_{i,2} + \ldots + \beta_{10}^0 x_{i,10}$ where y_i is net total financial assets (in \$1000); $x_{i,2}$ is annual income (in \$1000) in excess of population average and is denoted by "inc$_0$"; $x_{i,3} = x_{i,2}^2$ and is denoted by "inc$_0^2$"; $x_{i,4}$ is age in excess of population average and is denoted by "age$_0$"; $x_{i,5} = x_{i,4}^2$ and is denoted by "age$_0^2$"; $x_{i,6} = x_{i,2} \times x_{i,4}$ and is denoted by "inc$_0$.age$_0$"; $x_{i,7}$ is a dummy for eligibility for a 401k plan and is denoted by "e401k"; $x_{i,8}$ is a dummy for male and is denoted by "male"; $x_{i,9} = x_{i,7} \times x_{i,2}$ and is denoted by "e401k.inc$_0$"; and $x_{i,10} = x_{i,7} \times x_{i,4}$ and is denoted by "e401k.age$_0$".

$h(\beta)$	OLS	WLS	ALS	MIN	TWLS	CC	TCC	GMM	TGMM
intercept	5.905	6.393			6.176	6.350	6.074	6.615	6.205
	(2.115)	(.978)			(.917)	(.961)	(.915)	(.922)	(.889)
inc$_0$	.633	.463			.472	.482	.474	.502	.458
	(.152)	(.063)			(.056)	(.061)	(.056)	(.056)	(.050)
inc$_0^2$	.000	.003			.002	.003	.002	.002	.002
	(.005)	(.002)			(.002)	(.002)	(.002)	(.002)	(.002)
age$_0$	.704	.605			.581	.608	.581	.676	.629
	(.141)	(.087)			(.076)	(.087)	(.076)	(.075)	(.073)
age$_0^2$	.031	.011			.005	.011	.006	.013	.009
	(.014)	(.005)	same		(.004)	(.005)	(.004)	(.004)	(.004)
inc$_0$.age$_0$	.044	.026	as		.027	.027	.028	.031	.029
	(.013)	(.006)	WLS		(.005)	(.006)	(.005)	(.005)	(.005)
e401k	6.346	6.770			6.921	6.647	6.868	7.400	5.244
	(2.022)	(1.844)			(1.454)	(1.807)	(1.447)	(1.540)	(1.177)
male	1.799	1.505			1.558	1.517	1.580	1.656	1.063
	(1.959)	(.756)			(.534)	(.752)	(.526)	(.740)	(.601)
e401k.inc$_0$	.307	.258			.216	.265	.226	.309	.263
	(.216)	(.128)			(.092)	(.125)	(.089)	(.112)	(.104)
e401k.age$_0$	.154	.160			.228	.160	.228	.161	.182
	(.262)	(.120)			(.104)	(.118)	(.103)	(.116)	(.103)

Table 8: Estimates and standard errors (in parentheses) of regression coefficients in the financial wealth equation in Lu and Wooldridge (2020)'s empirical application [c.f. their Table 3]. While not done here to adhere to standard empirical practice, use of the same regression residuals would enforce that the standard errors of TGMM estimates for coefficients of male and e401k.inc$_0$ do not exceed that of TCC and TWLS.

We use the same data set, matching the descriptive statistics and OLS coefficients in Lu and Wooldridge (2020)'s Table 2 and 3 respectively (the OLS standard errors don't match because we report the HC3 version). We report in Table 8 the various estimates and standard errors (in parentheses) for all the regression coefficients. We use Lu and Wooldridge (2020)'s parametric model $\omega^2(x_i;\gamma) = \exp(x_i'\gamma)$ for heteroskedasticity. Lu and Wooldridge (2020) showed big gains in precision over OLS by WLS, and then further improvement over WLS by their GMM estimator. Our results confirm their findings. Moreover, our results demonstrate that even further gains in precision, and often substantial ones, over all estimators including Lu and Wooldridge (2020)'s can be obtained by our proposed targeted estimators.

4.5 Classical and machine learning semiparametric WLS

Semiparametric WLS has a long history; see Carroll (1982), Robinson (1987), Rilstone (1991), Newey (1994), Fan and Yao (1998), etc. Recently Miller and Startz (2019) and Gonzales-Coya and Perron (2024) use machine learning to fit heteroskedasticity and demonstrate improvement over classical semiparametric WLS. The classical and the new (machine learning) semiparametric WLS are asymptotically efficient, and hence cannot be asymptotically less precise than TWLS, TCC or TGMM. Nevertheless, we find that our TWLS, TCC or TGMM can still give finite-sample precision gains over the classical and new semiparametric WLS estimators under the simulation designs used in Miller and Startz (2019) and Gonzales-Coya and Perron (2024). Since this precision gain is only a finite-sample phenomenon and will not hold under first-order asymptotics, we collect its description in Appendix C.

5 Conclusion

The standard empirical practice of regression is to run OLS with its robust standard errors instead of WLS with its robust standard errors even when the user cares about regression

$E[y_i|x_i]$. This practice makes sense because even when the regression assumption $E[u_i|x_i] = 0$ is true, WLS based on the user's "hard work of modelling the heteroskedasticity" (Leamer (2010)) can be less precise than OLS if this user's model for heteroskedasticity $V(y_i|x_i)$ is incorrect. This is a serious drawback of (parametric) WLS since, in economics, heteroskedasticity and the incorrectness of the user's model for it are the norm rather than the exception.

Semiparametric WLS byapsses incorrect modeling for $V(y_i|x_i)$ asymptotically and is efficient. But it is also not much used in practice, although potential applications abound.

A recent literature aims to change the empirical practice of always using OLS, by providing better estimation methods when the user cares about regression $E[y_i|x_i]$ and is willing to commit to a linear or nonlinear parametric model for $E[y_i|x_i]$. Romano and Wolf (2017), DiCiccio et al. (2019), Lu and Wooldridge (2020), etc. seek to improve WLS where the user posits, but not necessarily believes in, a parametric model for heteroskedasticity $V(y_i|x_i)$, the nuisance parameters. Miller and Startz (2019), Gonzales-Coya and Perron (2024), etc. seek to improve semiparametric WLS by using machine learning to model $V(y_i|x_i)$.

Our paper belongs to this literature. Where we differ from these other papers is the following. Others are true to the classical WLS algorithm in obtaining the WLS weights by least squares or maximum likelihood fitting of the squared OLS residuals to some user-specified parametric or nonparametric function, possibly subject to penalization. On the other hand, we modify WLS by computing the WLS weights in a target-driven way that we have shown can lead to big gains in precision over the other methods when the user's model for heteroskedasticity is incorrect. It can also give finite-sample precision gains over the classical and new semiparametric WLS that, in theory, are indeed asymptotically efficient.

We have argued that OLS and WLS along with estimators that aim at "resurrecting WLS", can suffer significant and, importantly, *avoidable* loss in precision because they resort to suboptimal criterion functions of fitting heteroskedasticity. OLS does not attempt to fit, while the others use criterion functions that may have little to do with the relevant criterion

of minimizing the variance of estimators for the regression coefficients of interest.

By contrast, when focusing our optimality criterion on the asymptotic variance of estimators of scalar targets based on the regression coefficients, we realized that there could be precision gains with respect to OLS, WLS and their refinements, to be drawn from the proper target-driven choice of weights given the user's parametric model $\omega^2(x_i;\gamma)$ for heteroskedasticity. We illustrated this through our three estimators: TWLS, TCC and TGMM.

TWLS, TCC and TGMM do not search for parametric models to obtain a significant result. Rather, they take the user's/expert's model $\omega^2(x_i;\gamma)$ as given and then search for the target-specific optimal γ for this given model. The TWLS, TCC or TGMM estimates may or may not be significant, but are nevertheless the most precise ones that could be obtained under their respective strategy/class, given the user's/expert's model for heteroskedasticity.

6 Bibliography

Andrews, D. W. K. (1987). Asymptotic results for generalized Wald tests. *Econometric Theory*, 3: 348–358.

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.

Chen, X., Jacho-Chavez, D. T., and Linton, O. (2016). Averaging of an increasing number of moment condition estimators. *Econometric Theory*, 32: 30–70.

Cragg, J. G. (1983). More efficient estimation in the presence of heteroskedasticity of unknown form. *Econometrica*, 51: 751–763.

Cragg, J. G. (1992). Quasi-aitken estimation for heteroskedasticty of unknown form. *Journal of Econometrics*, 54: 179–201.

DiCiccio, C. J., Romano, J. P., and Wolf, M. (2019). Improving weighted least squares inference. *Econometrics and Statistics*, 10:96–119.

Fan, J. and Yao, Q. (1998). Efficient estimation of conditional variance functions in stochastic regressions. *Biometrika*, 85: 645–660.

Gonzales-Coya, E. and Perron, P. (2024). Estimation in the Presence of Heteroskedasticity of Unknown Form: A Lasso-based Approach. *Journal of Econometric Methods*, 13:29–48.

Gourieroux, C., Monfort, A., and Renault, E. (1996). Two-stage generalized moment method with applications to regressions with heteroscedasticity of unknown form. *Journal of Statistical Planning and Inference*, 50: 37–63.

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

Lu, C. and Wooldridge, J. M. (2020). A GMM estimator asymptotically more efficient than OLS and WLS in the presence of heteroskedasticity of unknown form. *Applied Economics Letters*, 27: 997–1001.

Miller, S. and Startz, R. (2019). Feasible Generalized Least Sqaures Using Machine Learning. *Economics Letters*, 175: 28–31.

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.

Papadopoulos, A. and Tsionas, M. G. (2022). Efficiency gains in least squares estimation: A new approach. *Econometric Reviews*, 41: 51–74.

Rilstone, P. (1991). Some Monte Carlo Evidence on the Relative Efficiency of Parametric and Semiparametric EGLS Estimators. *Journal of Business and Economic Statistics*, 9:179–187.

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.

Stewart, G. W. (1969). On the continuity of the generalized inverse. *SIAM Journal of Applied Mathematics*, 17: 33–45.

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

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

White, H. (1986). Instrumental variables analogs of generalized least squares estimators. *Advances in Statistical Analysis and Statistical Computing*, 1: 173–227.

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

Xiao, Z. (2020). Efficient GMM estimation with singular system of moment conditions. *Statistical Theory and Related Fields*, 4: 172–178.

Supplementary Material

Efficient estimation of regression models with user -specified parametric model for heteroskedasticity

Table of Contents

The numbering of the pages and equations in the appendix is consistent with the main text. The notation, however, differs a little and is arguably less heavy here. The heavier notation in the main text was essential for introducing our proposed estimators and clarifying the various steps for their implementation, which is no longer necessary in this appendix.

A Appendix A: Auxiliary material for Section 3

A.1 Useful expressions for Section 3

We will focus on the expressions for the asymptotic variances and their estimators. We have maintained the assumption in Section 3 that:

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} g(y_i, x_i, \beta^0, \gamma) \xrightarrow{d} N\left(0, V\left(\beta^0, \gamma\right)\right)$$

where:

$$V\left(\beta^0, \gamma\right) := \begin{bmatrix} V_{11}\left(\beta^0\right) := E\left[x_i x_i' \omega_0^2\left(x_i\right)\right] & V_{12}\left(\beta^0, \gamma\right) := E\left[\frac{x_i x_i' \omega_0^2(x_i)}{\omega^2(x_i,\gamma)}\right] \\ V_{21}\left(\beta^0, \gamma\right) := V_{12}\left(\beta^0, \gamma\right) & V_{22}\left(\beta^0, \gamma\right) := E\left[\frac{x_i x_i' \omega_0^2(x_i)}{\omega^4(x_i,\gamma)}\right] \end{bmatrix}.$$

Hence, the joint asymptotic distribution of OLS and UWLS(γ) is:

$$\begin{aligned} \sqrt{n}\begin{bmatrix} \widehat{\beta}_{OLS} - \beta^0 \\ \widehat{\beta}\left(\gamma\right) - \beta^0 \end{bmatrix} &= \begin{bmatrix} B_{1,n}^{-1} & 0 \\ 0 & B_{2,n}^{-1}\left(\gamma\right) \end{bmatrix} \frac{1}{\sqrt{n}} \begin{bmatrix} \sum_{i=1}^{n} x_i u_i \\ \sum_{i=1}^{n} \frac{x_i u_i}{\omega^2\left(x_i, \gamma\right)} \end{bmatrix} \\ &\xrightarrow{d} N\left(0, \begin{bmatrix} \Sigma\left(\beta^0, \gamma^{\hom}\right) & C_{12}\left(\beta^0, \gamma\right) \\ C_{21}\left(\beta^0, \gamma\right) & \Sigma\left(\beta^0, \gamma\right) \end{bmatrix}\right) \end{aligned}$$

with:

$$B_{1,n} := \frac{1}{n}\sum_{i=1}^{n} x_i x_i' \text{ and } B_{2,n}\left(\gamma\right) := \frac{1}{n}\sum_{i=1}^{n} \frac{x_i x_i'}{\omega^2\left(x_i, \gamma\right)}.$$

The joint asymptotic variance is:

$$\begin{bmatrix} \Sigma\left(\beta^0, \gamma^{\hom}\right) & C_{12}\left(\beta^0, \gamma\right) \\ C_{21}\left(\beta^0, \gamma\right) & \Sigma\left(\beta^0, \gamma\right) \end{bmatrix} := \begin{bmatrix} B_1^{-1} & 0 \\ 0 & B_2^{-1}\left(\gamma\right) \end{bmatrix} V\left(\beta^0, \gamma\right) \begin{bmatrix} B_1^{-1} & 0 \\ 0 & B_2^{-1}\left(\gamma\right) \end{bmatrix}'$$

with:

$$B_1 := E\left[x_i x_i'\right] \quad \text{and} \quad B_2\left(\gamma\right) := E\left[\frac{x_i x_i'}{\omega^2\left(x_i,\gamma\right)}\right].$$

An example of a natural estimator of this joint asymptotic variance matrix is:

$$\begin{bmatrix} B_{1,n}^{-1} & 0 \\ 0 & B_{2,n}^{-1}\left(\gamma\right) \end{bmatrix} \widehat{V}\left(\widehat{\beta}_{OLS},\gamma\right) \begin{bmatrix} B_{1,n}^{-1} & 0 \\ 0 & B_{2,n}^{-1}\left(\gamma\right) \end{bmatrix}$$

where:

$$\widehat{V}\left(\widehat{\beta}_{OLS},\gamma\right) := \begin{bmatrix} \widehat{V}_{11}\left(\widehat{\beta}_{OLS}\right) := \frac{1}{n}\sum_{i=1}^n x_i x_i' \widehat{u}_{i,OLS}^2 & \widehat{V}_{12}\left(\widehat{\beta}_{OLS},\gamma\right) := \frac{1}{n}\sum_{i=1}^n \frac{x_i x_i' \widehat{u}_{i,OLS}^2}{\omega^2(x_i;\gamma)} \\ \widehat{V}_{21}\left(\widehat{\beta}_{OLS},\gamma\right) := \widehat{V}_{12}\left(\widehat{\beta}_{OLS},\gamma\right) & \widehat{V}_{22}\left(\widehat{\beta}_{OLS},\gamma\right) := \frac{1}{n}\sum_{i=1}^n \frac{x_i x_i' \widehat{u}_{i,OLS}^2}{\omega^4(x_i;\gamma)} \end{bmatrix}.$$

This gives the expressions for the estimators of the three key quantities for Section 3.1 as:

$$\begin{aligned} \widehat{AVar}(h(\widehat{\beta}_{OLS})) &:= \frac{\partial h(\widehat{\beta}_{OLS})}{\partial \beta'} B_{1,n}^{-1} \widehat{V}_{11}\left(\widehat{\beta}_{OLS}\right) B_{1,n}^{-1} \frac{\partial h(\widehat{\beta}_{OLS})}{\partial \beta}, \\ \widehat{AVar}(h(\widehat{\beta}(\gamma))) &:= \frac{\partial h(\widehat{\beta}_{OLS})}{\partial \beta'} B_{2,n}^{-1}(\gamma) \widehat{V}_{22}\left(\widehat{\beta}_{OLS},\gamma\right) B_{2,n}^{-1}(\gamma) \frac{\partial h(\widehat{\beta}_{OLS})}{\partial \beta}, \\ \widehat{ACov}(h(\widehat{\beta}_{OLS}), h(\widehat{\beta}(\gamma))) &:= \frac{\partial h(\widehat{\beta}_{OLS})}{\partial \beta'} B_{1,n}^{-1} \widehat{V}_{12}\left(\widehat{\beta}_{OLS},\gamma\right) B_{2,n}^{-1}(\gamma) \frac{\partial h(\widehat{\beta}_{OLS})}{\partial \beta} \end{aligned}$$

where $AVar$ and $ACov$ denote the variance and covariance of the joint asymptotic distribution. As noted in the remark at the end of Section 2.4, we are writing $\widehat{\beta}_{OLS}$ as plugin for β^0 in all the expressions only for the sake uniformity of presentation. Any consistent estimator of β^0 can be used without altering the first-order asymptotic properties of any estimator of $h(\beta^0)$ considered in our paper. Also, while all the variance estimators are written in the HC0 form for simplicity, it is well-understood that other HC forms can deliver better performance in small samples. In fact, use of other HC forms is strongly suggested in this "resurrecting WLS" literature; see, e.g., Romano and Wolf (2017), Miller and Startz (2019), etc.

A.2 Proof of Proposition 1

The result will follow if we can show that for the given target $h(\beta)$:

$$AVar\left(\widehat{h}_{UGMM}(\gamma)\right) \leq AVar\left(\widehat{h}_{UTCC}(\gamma)\right) \leq AVar\left(\widehat{h}_{UWLS}(\gamma)\right) \tag{11}$$

for any $\gamma \in \Gamma$. This is because then using the definitions of $\gamma^*_{h,TGMM}$ and $\gamma^*_{h,TCC}$, and $\gamma^*_{h,TCC}$ and $\gamma^*_{h,TWLS}$ respectively, we will get:

$$\begin{aligned}
AVar\left(\widehat{h}_{TGMM}\right) = AVar\left(\widehat{h}_{UGMM}(\gamma^*_{h,TGMM})\right) &\leq AVar\left(\widehat{h}_{UGMM}(\gamma^*_{h,TCC})\right) \\
&\underset{\text{[by (11)]}}{\leq} AVar\left(\widehat{h}_{UCC}(\gamma^*_{h,TCC})\right) = AVar\left(\widehat{h}_{TCC}\right); \\
AVar\left(\widehat{h}_{TCC}\right) = AVar\left(\widehat{h}_{UCC}(\gamma^*_{h,TCC})\right) &\leq AVar\left(\widehat{h}_{UCC}(\gamma^*_{h,TWLS})\right) \\
&\underset{\text{[by (11)]}}{\leq} AVar\left(\widehat{h}_{UWLS}(\gamma^*_{h,TWLS})\right) = AVar\left(\widehat{h}_{TWLS}\right).
\end{aligned}$$

The second inequality in (11) follows by construction. So we focus on the first equality in (11). Also, since all estimators are asymptotically equivalent under conditional homoskedasticity, we ignore that case. Ignoring that case means that the MP inverse in the GMM weighting matrix is now the standard inverse. For simplicity of notation, we will work with the orthogonalized moment vectors $\widetilde{g}(y_i, x_i, \beta, \gamma) := [\widetilde{g}_1(y_i, x_i, \beta)', \widetilde{g}_2(y_i, x_i, \beta, \gamma)']'$ where:

$$\widetilde{g}_1(y_i, x_i, \beta) := g_1(y_i, x_i, \beta) \quad \text{and} \quad \widetilde{g}_2(y_i, x_i, \beta, \gamma) := g_2(y_i, x_i, \beta, \gamma) - V_{21}(\beta^0, \gamma)V_{11}^{-1}(\beta^0)g_1(y_i, x_i, \beta)$$

since the efficient GMM estimator of β based on either $g(y_i, x_i, \beta, \gamma)$ or $\widetilde{g}(y_i, x_i, \beta, \gamma)$ has the same asymptotic variance because the pre-multiplication matrix $\begin{bmatrix} I_p & 0 \\ -V_{21}(\beta^0, \gamma)V_{11}^{-1}(\beta^0) & I_p \end{bmatrix}$ is nonsingular. The estimator of β based on the moment vector $\widetilde{g}_2(y_i, x_i, \beta, \gamma)$ can be shown to be asymptotically equivalent to $\widetilde{\beta}(\gamma) := \left[B_2 - V_{21}V_{11}^{-1}B_1\right]^{-1}\left[B_2\widehat{\beta}(\gamma) - V_{21}V_{11}^{-1}B_1\widehat{\beta}_{OLS}\right]$

where we have (unless confusing) henceforth suppressed dependence on β^0 and γ for brevity.

Recalling the $\lambda_h^*(\gamma)$ from Section 3.1, the estimator $\widehat{h}_{UCC}(\gamma)$ can be shown to satisfy:

$$\widehat{h}_{UCC}(\gamma) - h(\beta^0) \;=\; \rho_1'\widehat{\beta}_{OLS} + \rho_2'\widetilde{\beta}(\gamma) - \frac{\partial h(\beta^0)}{\partial \beta'}\beta^0 + o_p(n^{-1/2})$$

where $\rho_1' := \frac{\partial h(\beta^0)}{\partial \beta'} - \rho_2'$ and $\rho_2' := \lambda_h^* \frac{\partial h(\beta^0)}{\partial \beta'}\left[I_p - B_2^{-1}V_{21}V_{11}^{-1}\right]$. Writing $B_{2.1} := B_2 - V_{21}V_{11}^{-1}B_1$ and $V_{22.1} := V_{22} - V_{21}V_{11}^{-1}V_{12}$ for further brevity, it can be similarly shown that $\widehat{h}_{UGMM}(\gamma)$ satisfies:

$$\widehat{h}_{UGMM}(\gamma) - h(\beta^0) \;=\; \Lambda_1'\widehat{\beta}_{OLS} + \Lambda_2'\widetilde{\beta}(\gamma) - \frac{\partial h(\beta^0)}{\partial \beta'}\beta^0 + o_p(n^{-1/2})$$

where $\Lambda_1' := \frac{\partial h(\beta^0)}{\partial \beta'} - \Lambda_2'$ and $\Lambda_2' := \frac{\partial h(\beta^0)}{\partial \beta'}\left[I_p + \left(B_{2.1}^{-1}V_{22.1}B_{2.1}^{-1}\right)\left(B_1^{-1}V_{11}B_1^{-1}\right)^{-1}\right]^{-1}$.

Define a generic estimator $\widetilde{h}(\gamma|\mu_1,\mu_2)$ where $\mu_1, \mu_2 \in \mathbb{R}^p$ such that $\mu_1' + \mu_2' = \frac{\partial h(\beta^0)}{\partial \beta'}$ and:

$$\widetilde{h}(\gamma|\mu_1,\mu_2) - h(\beta^0) \;=\; \mu_1'\widehat{\beta}_{OLS} + \mu_2'\widetilde{\beta}(\gamma) - \frac{\partial h(\beta^0)}{\partial \beta'}\beta^0 + o_p(n^{-1/2}).$$

This asymptotically unbiased generic estimator nests the cases of both $\widehat{h}_{UCC}(\gamma)$ and $\widehat{h}_{UGMM}(\gamma)$ (and $\widehat{h}_{UWLS}(\gamma)$). Consider optimality of this generic estimator by choosing:

$$\mu_1^*, \mu_2^* := \arg\min_{\mu_1,\mu_2 \in \mathbb{R}^p} AVar\left(\mu_1'\widehat{\beta}_{OLS} + \mu_2'\widetilde{\beta}(\gamma)\right) \quad \text{such that} \quad \mu_1' + \mu_2' = \frac{\partial h(\beta^0)}{\partial \beta'}.$$

Since $ACov\left(\widetilde{\beta}(\gamma), \widehat{\beta}_{OLS}\right) = 0$ and $AVar\left(\widetilde{\beta}(\gamma)\right) = B_{2.1}^{-1}V_{22.1}B_{2.1}^{-1}$, the proof of the first equality of (11) (and hence the proof of Proposition 1) follows since this optimal generic estimator turns out to be $\widehat{h}_{UGMM}(\gamma)$ because:

$$\begin{aligned}
\mu_2^* &= \left[B^{-1}V_{11}B^{-1} + B_{2.1}^{-1}V_{2.1}B_{2.1}^{-1}\right]^{-1} B_1^{-1}V_{11}B_1^{-1}\frac{\partial h(\beta^0)}{\partial \beta} \\
&= \left[I_p + \left(B_1^{-1}V_{11}B_1^{-1}\right)^{-1} B_{2.1}^{-1}V_{2.1}B_{2.1}^{-1}\right]^{-1} \frac{\partial h(\beta^0)}{\partial \beta} \\
&= \Lambda_2. \;\blacksquare
\end{aligned}$$

B Appendix B: Combining Estimators

For simplicity, we will abstract from estimation of γ in this appendix since it does not affect the asymptotic distribution of the estimators of β under the condition $E[y_i - x_i'\beta^0|x_i] = 0$ that will be maintained throughout. As discussed in Section 2, for any user-specified value γ of heteroskedasticity parameters, we can define a UWLS estimator $\widehat{\beta}(\gamma)$. This estimator can be interpreted as a GMM estimator provided by the following exactly identified set of moment conditions:

$$E\left[\frac{x_i}{\omega^2(x_i,\gamma)}(y_i - x_i'\beta)\right] = 0.$$

$\widehat{\beta}(\gamma)$ is a consistent estimator of β^0 with asymptotic variance $\Sigma(\beta^0,\gamma)$. Beyond the TWLS estimator $h\left(\widehat{\beta}\left(\gamma^*_{h,TWLS}\right)\right)$ defined in Section 2, it may make sense, for the sake of asymptotic variance minimization, to build new estimators by convex combinations (CC) of plug-in UWLS estimators $h\left(\widehat{\beta}(\gamma)\right)$ for different values of $\gamma \in \Gamma$. We will consider CC of only two such estimators. We first discuss such CC of estimators in the general setting of exactly identified sets of moment conditions.

B.1 A general framework

We consider two sets of exactly identified moment conditions that both identify the true unknown value β^0 of a p dimensional parameter vector β.

- A first set of p moments conditions identifies β^0:

$$E[g_1(y_i, x_i; \beta)] = 0 \iff \beta = \beta^0.$$

  This first set of moments may for instance be orthogonality conditions for OLS, UWLS, two stage least squares (2SLS) or nonlinear least squares (NLLS). In the UWLS case,

for some given value γ^1 of the heteroskedasticity parameters:

$$g_1(y_i, x_i; \beta) = \frac{x_i}{\omega^2(x_i, \gamma^1)}(y_i - x_i'\beta). \tag{12}$$

- A second set of p moments conditions also identifies β^0:

$$E[g_2(y_i, x_i; \beta)] = 0 \iff \beta = \beta^0.$$

This second set may for instance re-weight differently orthogonality conditions through another value γ^2 of the heteroskedasticity parameters:

$$g_2(y_i, x_i; \beta) = \frac{x_i}{\omega^2(x_i, \gamma^2)}(y_i - x_i'\beta). \tag{13}$$

Since the two sets of moment conditions are exactly identified, each of them defines without ambiguity a GMM estimator of β as follows:

$$\widehat{\beta}^{(j)} = \arg\min_{\beta} \|\bar{g}_{j,n}(\beta)\|, \quad j = 1, 2$$

where:

$$\bar{g}_{j,n}(\beta) = \frac{1}{n}\sum_{i=1}^{n} g_j(y_i, x_i; \beta).$$

We maintain throughout the standard assumptions for the asymptotic theory of GMM. In particular, since the two sets of moment conditions are exactly identified, we are led to assume that the two Jacobian matrices:

$$G_j = G_j(\beta^0) \quad \text{where} \quad G_j(\beta) = E\left[\frac{\partial}{\partial \beta'} g_j(\beta)\right], \quad j = 1, 2$$

are non-singular matrices. As a consequence, we have asymptotically a one-to-one mapping

between the GMM estimators and the corresponding sample moments:

$$\sqrt{n}(\widehat{\beta}^{(j)} - \beta^0) = -\left[G_j\right]^{-1} \sqrt{n}\bar{g}_{j,n}(\beta^0) + o_p(1). \tag{14}$$

In all this section, we will run affine regressions based on the joint asymptotic normal distribution of $\left[\sqrt{n}(\widehat{\beta}^{(j)} - \beta^0)\right]_{1\leq j\leq 2}$ implied by the central-limit theorem:

$$\begin{bmatrix} \sqrt{n}\bar{g}_{1,n}(\beta^0) \\ \sqrt{n}\bar{g}_{2,n}(\beta^0) \end{bmatrix} \xrightarrow{d} N\left(0, \Upsilon = \Upsilon(\beta^0)\right) \quad \text{where} \quad \Upsilon\left(\beta\right) = \begin{bmatrix} \Upsilon_{11}\left(\beta\right) & \Upsilon_{12}\left(\beta\right) \\ \Upsilon_{21}\left(\beta\right) & \Upsilon_{22}\left(\beta\right) \end{bmatrix}.$$

B.2 Convex combinations of estimators

We dub CC estimators all estimators which, extending an initial proposal of DiCiccio et al. (2019), are based on a convex combination (CC) of the two GMM estimators and thus can be written as:

$$\widehat{h}_\lambda = (1-\lambda)h(\widehat{\beta}^{(1)}) + \lambda h(\widehat{\beta}^{(2)})$$

for some $\lambda \in \mathbb{R}$. Note that we do not introduce any sign constraint on the scalar weight λ, so that the terminology "convex combination" is an abuse of language and we should rather say "affine combination". Asymptotically:

$$\begin{aligned} \sqrt{n}\left(\widehat{h}_\lambda - h(\beta^0)\right) &= \sqrt{n}\left(h(\widehat{\beta}^{(1)}) - h(\beta^0)\right) - \lambda\sqrt{n}\left(h(\widehat{\beta}^{(1)}) - h(\widehat{\beta}^{(2)})\right) \\ &= \sqrt{n}\delta'(\widehat{\beta}^{(1)} - \beta^0) - \lambda\sqrt{n}\delta'(\widehat{\beta}^{(1)} - \widehat{\beta}^{(2)}) + o_p(1) \end{aligned} \tag{15}$$

where:

$$\delta = \delta\left(\beta^0\right) \quad \text{and} \quad \delta(\beta) = \frac{\partial}{\partial\beta}h\left(\beta\right).$$

Hence, we minimize the asymptotic variance of $\widehat{h}_\lambda$ by choosing $\lambda = \lambda^*(h)$ that is the asymptotic regression coefficient in the regression of $\sqrt{n}\delta'(\widehat{\beta}^{(1)} - \beta^0)$ on $\sqrt{n}\delta'(\widehat{\beta}_n^{(1)} - \widehat{\beta}_n^{(2)})$:

$$\lambda^*(h) = \lim_{n\to\infty} \frac{\delta' Cov\left(\widehat{\beta}^{(1)}, \widehat{\beta}^{(1)} - \widehat{\beta}^{(2)}\right)\delta}{\delta' Var\left(\widehat{\beta}^{(1)} - \widehat{\beta}^{(2)}\right)\delta}.$$

The optimal CC estimator of the target $h(\beta)$ is thus:

$$\widehat{h}_{\lambda^*(\delta)} = [1 - \lambda^*(h)]\, h(\widehat{\beta}^{(1)}) + \lambda^*(h))h(\widehat{\beta}^{(2)}).$$

It leads us to our first result.

Proposition 2 *The optimal TCC (targeted CC) estimator of $h(\beta)$ based on the couple of estimators $\left(\widehat{\beta}^{(1)}, \widehat{\beta}^{(2)}\right)$ is given by:*

$$\widehat{h}_{\lambda^*(h)} = [1 - \lambda^*(h)]\, h(\widehat{\beta}^{(1)}) + \lambda^*(h)h(\widehat{\beta}^{(2)})$$

with:

$$\lambda^*(h) = \lim_{n\to\infty} \frac{\delta' Cov\left(\widehat{\beta}^{(1)}, \widehat{\beta}^{(1)} - \widehat{\beta}^{(2)}\right)\delta}{\delta' Var\left(\widehat{\beta}^{(1)} - \widehat{\beta}^{(2)}\right)\delta}.$$

Remark 1: The asymptotic expansion in (15) shows that we can also interpret our CC estimators as follows:

$$\begin{aligned}\sqrt{n}\left(\widehat{h}_\lambda - h(\beta^0)\right) &= \sqrt{n}\delta'\left[(1-\lambda)\widehat{\beta}^{(1)} + \lambda\widehat{\beta}^{(2)} - \beta^0\right] + o_p(1) \\ &= \sqrt{n}\left\{h\left((1-\lambda)\widehat{\beta}^{(1)} + \lambda\widehat{\beta}^{(2)}\right) - h\left(\beta^0\right)\right\} + o_p(1).\end{aligned}$$

In other words, the CC estimator can also be interpreted as a plug-in estimator where it is a convex combination $\left[(1-\lambda)\widehat{\beta}^{(1)} + \lambda\widehat{\beta}^{(2)}\right]$ of the two estimators of β that is plugged in.

Remark 2: The proof of Proposition 2 (see Appendix B.4) shows that the optimal TCC estimator based on $\left(\widehat{\beta}^{(1)}, \widehat{\beta}^{(2)}\right)$ is $h\left(\widehat{\beta}^{(1)}\right)$, for all possible target $h(\beta)$, if and only if:

$$Cov\left(\widehat{\beta}^{(1)}, \widehat{\beta}^{(1)} - \widehat{\beta}^{(2)}\right) = 0$$

that is, by virtue of (14) and with obvious simplified notations:

$$Cov\left(g_1, G_1^{-1} g_1 - G_2^{-1} g_2\right) = 0.$$

Elementary calculation (see proof of Proposition 2 in Appendix B.4) shows that this property is tantamount to the identity:

$$G_2 = \Upsilon_{21} \Upsilon_{11}^{-1} G_1. \tag{16}$$

Breusch et al. (1999) have shown that the condition (16) characterizes the fact that the set of moment conditions g_2 is "redundant" with respect to g_1, meaning that the complete set (g_1, g_2) of moment conditions does not deliver a GMM estimator (asymptotically) more accurate than $\widehat{\beta}^{(1)}$. It is not surprising to find that this condition characterizes the case where CC based on the couple of estimators $\left(\widehat{\beta}^{(1)}, \widehat{\beta}^{(2)}\right)$ does not deliver better estimators (of any target) than estimators based on $\widehat{\beta}^{(1)}$ only.

Example: In the UWLS (12)/(13) example:

$$\Upsilon_{11} = E\left[\frac{x_i x_i' \omega_0^2(x_i)}{\omega^4(x_i, \gamma^1)}\right], \quad \Upsilon_{21} = E\left[\frac{x_i x_i' \omega_0^2(x_i)}{\omega^2(x_i, \gamma^1)\, \omega^2(x_i, \gamma^2)}\right].$$

Let us consider the particular case where the user-specified heteroskedasticity model matches perfectly the true skedastic function for the value γ^1 of the heteroskedasticity parameter:

$$\omega_0^2(x_i) \equiv \omega^2\left(x_i, \gamma^1\right).$$

In this case:

$$\Upsilon_{11} = E\left[\frac{x_i x_i'}{\omega^2\left(x_i, \gamma^1\right)}\right] = -G_1 \quad \text{and} \quad \Upsilon_{21} = E\left[\frac{x_i x_i'}{\omega^2\left(x_i, \gamma^2\right)}\right] = -G_2$$

so that the condition (16) is automatically fulfilled. This is relevant for our study in two cases:

1st case: The user-specified heteroskedasticity model is well–specified so that $\gamma_{WLS} = \gamma^1$ and $\widehat{\beta}^{(1)} = \widehat{\beta}\left(\gamma_{WLS}\right)$ is the optimal WLS estimator. In this case, there is no relevant additional information for estimation of β brought by any other UWLS estimator $\widehat{\beta}\left(\gamma_{UWLS}\right)$.

2nd case: $\widehat{\beta}^{(1)} = \widehat{\beta}_{OLS}$ is the OLS estimator and this estimator is optimal because the DGP is homoskedastic: $\omega_0^2\left(x_i\right) \equiv \omega_{\text{hom}}^2$. In this case, irrespective of the heteroskedasticity model, there is no relevant additional information for estimation of β brought by any other UWLS estimator $\widehat{\beta}\left(\gamma_{UWLS}\right)$. Because of a subtle technicality, the case of homoskedasticity needs to be handled a little differently following the main text when a (optimal) choice of γ is involved. The message of this second case however remains the same.

Remark 3: The concept of CC is more obvious when the two estimators $\widehat{\beta}^{(1)}$and $\widehat{\beta}^{(2)}$ are asymptotically independent. Then:

$$\lambda^*(h) = \lim_{n\to\infty} \frac{Var\left(\delta'\widehat{\beta}^{(1)}\right)}{Var\left(\delta'\widehat{\beta}^{(1)}\right) + Var\left(\delta'\widehat{\beta}^{(2)}\right)}.$$

In this case, $\lambda^*(h)$ is a weight in $[0,1]$ that gives more weight to $\widehat{\beta}^{(1)}$ (resp. to $\widehat{\beta}^{(2)}$) if and only if the plug-in variance $Var\left(\delta'\widehat{\beta}^{(1)}\right)$ is smaller (resp. larger) than $Var\left(\delta'\widehat{\beta}^{(2)}\right)$.

Note that, by virtue of (14), the asymptotic independence of $\widehat{\beta}^{(1)}$ and $\widehat{\beta}^{(2)}$ is tantamount to the asymptotic independence of the moment functions $\sqrt{n}\bar{g}_{1,n}(\beta^0)$ and $\sqrt{n}\bar{g}_{2,n}(\beta^0)$. One may even consider that this condition can be maintained without loss of generality, since one can replace the second set $\sqrt{n}\bar{g}_{2,n}(\beta^0)$ of moment conditions by a set $\sqrt{n}\bar{g}_{2/1,n}(\beta^0)$ that

has been previously orthogonalized with respect to $\sqrt{n}\bar{g}_{1,n}(\beta^0)$:

$$\sqrt{n}\bar{g}_{2/1,n}(\beta) = \sqrt{n}\bar{g}_{2,n}(\beta) - \Upsilon_{21}\Upsilon_{11}^{-1}\sqrt{n}\bar{g}_{1,n}(\beta).$$

However, for these moment conditions, the Jacobian matrix is:

$$G_{2/1} = G_2 - \Upsilon_{21}\Upsilon_{11}^{-1}G_1.$$

Of course, this Jacobian matrix is nil in the case (16) of redundant moment conditions. By contrast, in many circumstances (see, e.g., the example below), we can assume that the matrix $G_{2/1}$ is non-singular, such that our general theory of CC applies to orthogonal moment functions $\sqrt{n}\bar{g}_{1,n}(\beta^0)$ and $\sqrt{n}\bar{g}_{2/1,n}(\beta^0)$.

Example: Let us consider the UWLS (12)/(13) example in the case of a well-specified heteroskedasticity model, with $\gamma^1 = \gamma^{\text{hom}}$ and $\gamma^2 = \gamma_{WLS}$. Then:

$$G_2 = -E\left[\frac{x_i x_i'}{\omega^2\left(x_i, \gamma_{WLS}\right)}\right] = -E\left[\frac{x_i x_i'}{\omega_0^2\left(x_i\right)}\right].$$

$[-G_2]^{-1}$ is the variance matrix of the WLS estimator. On the other hand:

$$\begin{aligned}\Upsilon_{21}\Upsilon_{11}^{-1}G_1 &= -E\left[\frac{x_i x_i'\omega_0^2\left(x_i\right)}{\omega^2\left(x_i, \gamma_{WLS}\right)\omega_{\text{hom}}^2}\right]\left\{E\left[\frac{x_i x_i'\omega_0^2\left(x_i\right)}{\omega_{\text{hom}}^4}\right]\right\}^{-1}E\left[\frac{x_i x_i'\omega_0^2\left(x_i\right)}{\omega_{\text{hom}}^2}\right] \\ &= -E\left[x_i x_i'\right]\left\{E\left[x_i x_i'\omega_0^2\left(x_i\right)\right]\right\}^{-1}E\left[x_i x_i'\right].\end{aligned}$$

$\left[-\Upsilon_{21}\Upsilon_{11}^{-1}G_1\right]^{-1}$ is the variance matrix of the OLS estimator. Therefore, the WLS estimator $a'\widehat{\beta}\left(\gamma_{WLS}\right)$ is strictly more accurate than the OLS estimator $a'\widehat{\beta}\left(\gamma^{\text{hom}}\right)$ for any linear combination $a'\beta$ if and only if the matrix $G_{2/1} = \left[G_2 - \Upsilon_{21}\Upsilon_{11}^{-1}G_1\right]$ is negative definite. Hence, it is reasonable to maintain the assumption that the matrix $G_{2/1}$ is nonsingular.

Remark 4: As exemplified in Appendix B.1, we may be led to consider moment conditions

that are indexed by some nuisance parameters $\gamma \in \Gamma$. While the notations did not make explicit the dependence on γ, we can for instance revisit the second set of moment functions as:

$$g_2\left(y_i, x_i, \beta\right) = g_2\left(y_i, x_i, \beta, \gamma\right).$$

In particular, for our regression application:

$$g_2\left(y_i, x_i, \beta, \gamma\right) = \frac{x_i}{\omega^2\left(x_i, \gamma\right)}\left(y_i - x_i'\beta\right). \tag{17}$$

We then want to resort to a condition of local robustness: replacing in $g_2()$ a specific value $\bar{\gamma}$ of nuisance parameters by a $\sqrt{n}$-consistent estimator $\widehat{\gamma}$ ($\sqrt{n}\left(\widehat{\gamma} - \bar{\gamma}\right) = O_P(1)$) has no impact on the asymptotic distribution of any GMM estimator of β based on moment conditions including:

$$E\left[g_2\left(y_i, x_i, \beta\right)\right] = 0 \Longleftrightarrow \beta = \beta^0.$$

This robustness property will be necessary for defining feasible versions of optimal CC estimators by using a first step consistent estimator of γ that will have no effect.

The standard assumption to ensure this robustness is:

$$E\left[\frac{\partial}{\partial \gamma'} g_2\left(y_i, x_i, \beta^0, \gamma\right)\right] = 0, \quad \text{for all} \quad \gamma \in \Gamma. \tag{18}$$

It is the simplest case of Chernozhukov et al. (2022). It is obvious that the robustness condition (18) is valid for our regression example (17), insofar as we maintain the assumption of zero conditional expectation:

$$E[y_i - x_i'\beta^0 \,|x_i] = 0.$$

It is also the case if one wants to extend our study to 2SLS or NLLS.

B.3 Matricial combinations of estimators

Following Chen et al. (2016), we introduce MCC (Matricial CC) estimators. While Remark 1 above has shown that our CC estimators can be interpreted as plugging in the target $h(\beta)$ an estimator of β that is a convex combination of $\widehat{\beta}^{(1)}$ and $\widehat{\beta}^{(2)}$, we now consider the possibility to plug in a matrix combination of estimators by considering:

$$\widehat{\beta}_A = (I_p - A)\widehat{\beta}^{(1)} + A\widehat{\beta}^{(2)}$$

for any square matrix A of size p.

Hence, we minimize the asymptotic variance matrix of $\widehat{\beta}_A$ by choosing $A = A^*$ that is the matrix of regression coefficients in the asymptotic regression of $\sqrt{n}(\widehat{\beta}^{(1)} - \beta^0)$ on $\sqrt{n}(\widehat{\beta}_n^{(1)} - \widehat{\beta}_n^{(2)})$:

$$A^* = \lim_{n\to\infty} Cov\left(\widehat{\beta}^{(1)}, \widehat{\beta}^{(1)} - \widehat{\beta}^{(2)}\right)\left[Var\left(\widehat{\beta}^{(1)} - \widehat{\beta}^{(2)}\right)\right]^{-1}.$$

It leads us to our second result.

Proposition 3 *The optimal TMCC (targeted matricial CC) estimator of $h(\beta)$ based on the couple of estimators $\left(\widehat{\beta}^{(1)}, \widehat{\beta}^{(2)}\right)$ is given by:*

$$\widehat{h}_{A^*} = h\left((I_p - A^*)\widehat{\beta}^{(1)} + A^*\widehat{\beta}^{(2)}\right).$$

Remark 5: An asymptotic expansion shows that we can also interpret our TMCC estimator as follows:

$$\begin{aligned}\sqrt{n}\left[\widehat{h}_{A^*} - h\left(\beta^0\right)\right] &= \sqrt{n}\delta'\left(\beta^0\right)\left\{(I_p - A^*)\widehat{\beta}^{(1)} + A^*\widehat{\beta}^{(2)} - \beta^0\right\} + o_P(1) \\ &= \sqrt{n}\delta'\left(\beta^0\right)(I_p - A^*)\left(\widehat{\beta}^{(1)} - \beta^0\right) + \sqrt{n}\delta'\left(\beta^0\right)A^*\left(\widehat{\beta}^{(2)} - \beta^0\right) + o_P(1).\end{aligned}$$

This expansion does not allow to interpret the TMCC estimator $\widehat{h}_{A^*}$ as a CC estimator. It is only if the vector $\delta\left(\beta^0\right)$ is an eigenvector of the matrix $A^{*\prime}$ with an eigenvalue λ^*, that we can write:

$$\begin{aligned}\sqrt{n}\left[\widehat{h}_{A^*}-h\left(\beta^0\right)\right] &= \sqrt{n}\delta'\left(\beta^0\right)\left[(1-\lambda^*)\widehat{\beta}^{(1)}+\lambda^*\widehat{\beta}^{(2)}-\beta^0\right]+o_P\left(1\right)\\ &= \sqrt{n}\left[h\left((1-\lambda^*)\widehat{\beta}^{(1)}+\lambda^*\widehat{\beta}^{(2)}\right)-h\left(\beta^0\right)\right]+o_p(1).\end{aligned}$$

This result suggests that the set of CC estimators is a strict subset of the set of MCC estimators. Therefore, we expect in general that no CC estimator of the target $h\left(\beta\right)$ can be asymptotically as accurate as the TMCC estimator, except when the target is such that its gradient vector (computed at the true value β^0 of β) is an eigenvector of the matrix $A^{*'}$.

Remark 6: The case (in some sense without loss of generality as explained in Remark 3) of asymptotically independent estimators $\widehat{\beta}^{(1)}$ and $\widehat{\beta}^{(2)}$ is helpful to figure out the efficiency gain obtained by moving from TCC to TMCC. In this case:

$$A^*=\lim_{n\to\infty}Var\left(\widehat{\beta}^{(1)}\right)\left[Var\left(\widehat{\beta}^{(1)}\right)+Var\left(\widehat{\beta}^{(2)}\right)\right]^{-1}.$$

For the sake of notational simplicity, we will write hereafter in this remark asymptotic (co)variances without the "lim" symbol. The asymptotic expansion in Remark 5 shows that asymptotically the TMCC estimator $\widehat{h}_{A^*}$ depends on the matrix A^* only through:

$$A^{*\prime}\delta=\left[Var\left(\widehat{\beta}^{(1)}\right)+Var\left(\widehat{\beta}^{(2)}\right)\right]^{-1}Var\left(\widehat{\beta}^{(1)}\right)\delta. \tag{19}$$

Therefore, the eigenvector condition discussed in Remark 5 to make TCC and TMCC estimators asymptotically equivalent is tantamount to imposing that (19) can be rewritten as:

$$\left[Var\left(\widehat{\beta}^{(1)}\right)+Var\left(\widehat{\beta}^{(2)}\right)\right]\lambda^*\delta=Var\left(\widehat{\beta}^{(1)}\right)\delta. \tag{20}$$

When left-multiplying this condition by δ', we see that λ^* must be the weight elicited by TCC. However, this necessary condition is obviously not sufficient in general to ensure the eigenvalue conditions (20). This simply confirms that, in general, no CC estimator can be asymptotically as accurate as the TMCC.

Remark 7: An alternative estimation strategy would be to estimate β by over-identified GMM based on the two sets of moment conditions $\bar{g}_n(\beta) = [\bar{g}'_{1,n}(\beta), \bar{g}'_{2,n}(\beta)]'$ stacked together. For any given weighting matrix, we would define a GMM estimator:

$$\widehat{\beta}\,(W) = \arg\min_{\beta} \bar{g}_n(\beta)' W \bar{g}_n(\beta).$$

By standard asymptotic GMM theory:

$$\sqrt{n}\left[\widehat{\beta}\,(W) - \beta^0\right] = -\left[G'WG\right]^{-1} G'W\sqrt{n}\bar{g}_n(\beta^0) + o_P\,(1) \tag{21}$$

with:

$$G = G\left(\beta^0\right) \quad \text{and} \quad G(\beta^0) = \begin{bmatrix} G_1\,(\beta) \\ G_2\,(\beta) \end{bmatrix}.$$

Then, with obvious notations:

$$\begin{aligned} G'W\sqrt{n}\bar{g}_n(\beta^0) &= [G'_1W_{11} + G'_2W_{21}]\sqrt{n}\bar{g}_{1,n}(\beta^0) + [G'_1W_{12} + G'_2W_{22}]\sqrt{n}\bar{g}_{2,n}(\beta^0) \\ &= [G'_1W_{11} + G'_2W_{21}]G_1\sqrt{n}(\widehat{\beta}^{(1)} - \beta^0) + [G'_1W_{12} + G'_2W_{22}]G_2\sqrt{n}(\widehat{\beta}^{(2)} - \beta^0) + o_p(1). \end{aligned}$$

We define a square matrix $A(W)$ of dimension p by:

$$I_p - A(W) = [G'WG]^{-1}\,[G'_1W_{11} + G'_2W_{21}]G_1.$$

With easy calculations, we can check that:

$$A(W) = [G'WG]^{-1}[G_1'W_{12} + G_2'W_{22}]G_2$$

so that the asymptotic expansion (21) of the GMM estimator can be rewritten:

$$\sqrt{n}\left[\widehat{\beta}(W) - \beta^0\right] = [I_p - A(W)]\sqrt{n}(\widehat{\beta}^{(1)} - \beta^0) + A(W)\sqrt{n}(\widehat{\beta}^{(2)} - \beta^0) + o_p(1).$$

Therefore, for any weighting matrix W, the GMM estimator $\widehat{\beta}(W)$ associated to this matrix is an MCC estimator with a matricial weight A defined by $A(W)$ given above. Hence, the class of MCC estimators asymptotically encompasses not only the CC estimators but also all GMM estimators based on the complete set of moment conditions. Not surprisingly though, MCC does not allow us to beat efficient GMM since we can prove the following result.

Proposition 4 *The optimal TMCC estimator $\widehat{h}_{A^*}$ is asymptotically equivalent to the optimal plug in GMM estimator $h\left(\widehat{\beta}(W^*)\right)$, that is computed with the optimal weighting matrix $W^* = [\Upsilon(\beta^0)]^{-1}$.*

Proofs of all the results are presented in Appendix B.4 for the sake of self-containedness. While Chen et al. (2016) seem to suggest the contrary, our proof shows that the validity of these results heavily rests upon the fact that we are combining only just identified sets of moment conditions.

B.4 Proofs

B.4.1 Proof of Propositions 2, 3, and Remarks 1 to 5 in Appendix B.2-B.3

For the sake of notational simplicity, all computations are made in the joint asymptotic Gaussian distribution of estimators of β, without making explicit any notation of limit in distribution. We are interested in CC and MCC estimators that are asymptotically equivalent

to (i.e., with difference of the order $o_p(1/\sqrt{n})$):

$$\begin{aligned}\widehat{h}_\lambda &= h\left((1-\lambda)\,\widehat{\beta}_1 + \lambda\widehat{\beta}_2\right),\\ \widehat{h}_A &= h\left((I_p - A)\,\widehat{\beta}_1 + A\widehat{\beta}_2\right).\end{aligned}$$

After asymptotic expansion:

$$\begin{aligned}\widehat{h}_\lambda &= \delta' U(\lambda), \quad U(\lambda) = \widehat{\beta}_1 - \lambda\left(\widehat{\beta}_1 - \widehat{\beta}_2\right)\\ \widehat{h}_A &= \delta' U(A), \quad U(A) = \widehat{\beta}_1 - A\left(\widehat{\beta}_1 - \widehat{\beta}_2\right).\end{aligned}$$

If we find A^* such that:

$$Var\left(U\left(A^*\right)\right) \ll Var\left(U\left(A\right)\right) \quad \text{for all} \;\; A$$

with inequalities in the sense of positive semi-definite matrices, we can be sure to have defined a minimum for:

$$Var\left(\widehat{h}_A\right) = \delta' Var\left(U\left(A\right)\right)\delta.$$

Hence, we have an optimal MCC estimator from multivariate regression coefficients:

$$A^* = Cov\left(\widehat{\beta}_1, \widehat{\beta}_1 - \widehat{\beta}_2\right)\left[Var\left(\widehat{\beta}_1 - \widehat{\beta}_2\right)\right]^{-1}, \;\; \widehat{h}_{A^*} = \delta'\left[\widehat{\beta}_1 - A^*\left(\widehat{\beta}_1 - \widehat{\beta}_2\right)\right].$$

By contrast, there does not exist in general a real number λ^* such that:

$$Var\left[U\left(\lambda^*\right)\right] \ll Var\left[U\left(\lambda\right)\right] \quad \text{for all} \;\; \lambda. \tag{22}$$

The optimal CC is defined from an optimal number λ^* that depends on the target δ:

$$\lambda^* = \arg\min_{\lambda\in\mathbb{R}} \delta' Var\left[U\left(\lambda\right)\right]\delta = \arg\min_{\lambda\in\mathbb{R}}\left[\delta'\widehat{\beta}_1 - \lambda\delta'\left(\widehat{\beta}_1 - \widehat{\beta}_2\right)\right].$$

Hence, we have an optimal CC estimator from univariate regression coefficient of $\delta'\widehat{\beta}_1$ on $\left[\delta'\widehat{\beta}_1 - \delta'\widehat{\beta}_2\right]$:

$$\begin{aligned}\lambda^* &= Cov\left(\delta'\widehat{\beta}_1, \delta'\widehat{\beta}_1 - \delta'\widehat{\beta}_2\right)\left[Var\left(\delta'\widehat{\beta}_1 - \delta'\widehat{\beta}_2\right)\right]^{-1}, \\ \widehat{h}_{\lambda^*} &= \delta'\widehat{\beta}_1 - \lambda^*\delta'\left(\widehat{\beta}_1 - \widehat{\beta}_2\right).\end{aligned}$$

While λ^* does not solve (22) in general, it does solve if:

$$\delta' A^* = \lambda^* \delta'$$

meaning that δ is an eigenvector of $A^{*\prime}$ with eigenvalue λ^*. Otherwise, we have in general:

$$Var\left(\widehat{h}_{A^*}\right) < Var\left(\widehat{h}_{\lambda^*}\right).$$

Therefore, the optimal TMCC estimator is strictly more accurate than the optimal TCC estimator.

However, if for all $\delta \in \mathbb{R}^p$, no TCC estimator based on CC of the two estimators $\left(\widehat{\beta}_1, \widehat{\beta}_2\right)$ can improve upon $h\left(\widehat{\beta}_1\right)$, it is also the case for TMCC estimators. To see that, note that if $\widehat{h}_{A^*}$ stands for our optimal TMCC and for all eigenvectors δ of $A^{*\prime}$ if $\lambda^*(\delta)$ stands for the corresponding eigenvalue, then:

$$\begin{aligned}Var\left(\widehat{h}_{\lambda^*(\delta)}\right) &= \delta' Var\left(U\left(A^*\right)\right)\delta \;=\; \delta'\left\{Var\left(\widehat{\beta}_1\right) - Var\left(A^*\left(\widehat{\beta}_1 - \widehat{\beta}_2\right)\right)\right\}\delta \\ &= \delta' Var\left(\widehat{\beta}_1\right)\delta - \delta' A^* Var\left(\widehat{\beta}_1 - \widehat{\beta}_2\right) A^{*\prime}\delta \\ &= \delta' Var\left(\widehat{\beta}_1\right)\delta - \lambda^*\delta' Var\left(\widehat{\beta}_1 - \widehat{\beta}_2\right)\lambda^*\delta.\end{aligned}$$

The fact that no TCC estimator based on CC of the two estimators $\left(\widehat{\beta}_1, \widehat{\beta}_2\right)$ can improve

upon $h\left(\widehat{\beta}_1\right)$ means that:

$$Var\left(\widehat{h}_{\lambda^*(\delta)}\right) = \delta' Var\left(\widehat{\beta}_1\right)\delta \implies \lambda^*\delta' Var\left(\widehat{\beta}_1 - \widehat{\beta}_2\right)\lambda^*\delta = 0 \implies \lambda^*\delta = 0 \implies A^{*\prime}\delta = 0.$$

Since this must be true for all eigenvectors δ of $A^{*\prime}$, we conclude:

$$A^* = 0.$$

No TMCC can improve upon $h\left(\widehat{\beta}_1\right)$. In other words, the optimal TCC is $h\left(\widehat{\beta}_1\right)$ for all possible target $h(\beta)$ if and only if:

$$Cov\left(\widehat{\beta}_1, \widehat{\beta}_1 - \widehat{\beta}_2\right) = 0.$$

Since (see Appendix B.1) estimators are one-to-one linear functions of moment conditions, this can be rewritten with obvious simplified notations:

$$Cov\left(g_1, G_1^{-1}g_1 - G_2^{-1}g_2\right) = 0.$$

This can be rewritten:

$$\Upsilon_{11}G_1^{-1\prime} = \Upsilon_{12}G_2^{-1\prime} \iff G_2^{-1}\Upsilon_{21} = G_1^{-1}\Upsilon_{11} \iff \Upsilon_{21} = G_2G_1^{-1}\Upsilon_{11} \iff G_2 = \Upsilon_{21}\Upsilon_{11}^{-1}G_1.\blacksquare$$

B.4.2 Proof of Proposition 4

We first compute the matrix of multivariate regression coefficients:

$$A^* = Cov\left(\widehat{\beta}_1, \widehat{\beta}_1 - \widehat{\beta}_2\right)\left[Var\left(\widehat{\beta}_1 - \widehat{\beta}_2\right)\right]^{-1} = \left[\Sigma_{11} - \Sigma_{12}\right]\left[\Sigma_{11} + \Sigma_{22} - \Sigma_{12} - \Sigma_{21}\right]^{-1}$$

where Σ stands for the joint asymptotic variance matrix of the couple $\left(\widehat{\beta}_1', \widehat{\beta}_2'\right)'$ of estimators:

$$\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}.$$

For efficient GMM, with weighting matrix $W = [\Upsilon(\beta^0)]^{-1}$, we have shown that it is an MCC of the two estimators with a matrix of weights:

$$\Delta = [G'WG]^{-1}[G_1'W_{12}G_2 + G_2'W_{22}G_2] = \left[\Sigma^{-1}\right]^{-1}\left[\Sigma^{12} + \Sigma^{22}\right] = \left[\Sigma^{11} + \Sigma^{21} + \Sigma^{12} + \Sigma^{22}\right]^{-1}\left[\Sigma^{12} + \Sigma^{22}\right]$$

with the notations:

$$\Sigma^{-1} = \begin{bmatrix} \Sigma^{11} & \Sigma^{12} \\ \Sigma^{21} & \Sigma^{22} \end{bmatrix} = G'WG = G_1'W_{11}G_1 + G_1'W_{12}G_2 + G_2'W_{21}G_1 + G_2'W_{22}G_2.$$

Hence, to prove Proposition 4, we need to check that, when $W = [\Upsilon(\beta^0)]^{-1}$:

$$[\Sigma_{11} - \Sigma_{12}][\Sigma_{11} + \Sigma_{22} - \Sigma_{12} - \Sigma_{21}]^{-1} = \left[\Sigma^{11} + \Sigma^{21} + \Sigma^{12} + \Sigma^{22}\right]^{-1}\left[\Sigma^{12} + \Sigma^{22}\right].$$

We know (see discussion in Appendix B.2) that we can assume without loss of generality that $\sqrt{n}\bar{g}_{1,n}$ and $\sqrt{n}\bar{g}_{2,n}$ are asymptotically independent, or equivalently the estimators $\widehat{\beta}_1$ and $\widehat{\beta}_2$ are asymptotically independent. In this case, matrices Σ and Σ^{-1} are block diagonal and, to prove Proposition 4, we only need to check that:

$$\Sigma_{11}[\Sigma_{11} + \Sigma_{22}]^{-1} = \left[\Sigma^{11} + \Sigma^{22}\right]^{-1}\Sigma^{22}$$

that is:

$$\left[\Sigma^{11} + \Sigma^{22}\right]\Sigma_{11} = \Sigma^{22}[\Sigma_{11} + \Sigma_{22}]$$

which is obvious, since in case of block-diagonality:

$$\Sigma^{11} = (\Sigma_{11})^{-1}, \ \Sigma^{22} = (\Sigma_{22})^{-1}. \ \blacksquare$$

C Appendix C: Machine learning semiparametric WLS

While not much used in current empirical practice, nonparametric estimation of the true skedastic function $\omega_0^2(x_i)$ has a long and rich history; see, e.g., Carroll (1982), Robinson (1987), Newey (1994), Fan and Yao (1998), etc. This leads to the semiparametric WLS.

Miller and Startz (2019) and Gonzales-Coya and Perron (2024) demonstrate that machine learning strategies generally outperform classical nonparametric methods applied in this context by Carroll (1982) (with kernels), Robinson (1987) (with Nearest Neighbor) or Fan and Yao (1998) (with Local Linear smoothing). These classical methods led to the classical semiparametric WLS estimators, whereas the machine learning methods lead to the new generation of semiparametric WLS estimators. Miller and Startz (2019) and Gonzales-Coya and Perron (2024) argue that the advantage of machine learning is that it does not require a tight pre-specification of the nature and number of covariates.

Miller and Startz (2019) recommend using Support Vector Regression (SVR) to model heteroskedasticity, while Gonzales-Coya and Perron (2024) recommend using Lasso. Gonzales-Coya and Perron (2024) additionally compare their Lasso method with application of SVR, random forest and other methods, and conclude that the use of Lasso delivers best performance in their simulations.

How do our targeted methods TWLS, TCC and TGMM perform in terms of precision compared to the classical and new generation of semiparametric WLS estimators?

All the semiparametric WLS estimators (the classical and the new ones) are semiparametrically efficient. Therefore, *asymptotically*, our targeted methods cannot be more precise than any of them, and in general will be less precise if the user's parametric model for heteroskedasticity is incorrect.[3] However, the reason the classical semiparametric WLS es-

[3]If there is a promise in the sense of Ackerberg et al. (2012) to make the user's parametric model for heteroskedasticity richer with the increase in sample size (i.e., as more observations become available), then our targeted methods can also be "interpreted" as semiparametrically efficient; but that is not what we do in our paper. The precision gains that we wish to highlight in our paper are due to *targeting* based on a possibly incorrect model and not because of semiparametric efficiency.

timators are not much used in practice is because such asymptotic results are generally not materialized in small samples. Therefore, to answer the question about precision, it makes sense to compare the finite-sample precision of the classical and new semiparametric WLS estimators with our TWLS, TCC and TGMM estimators based on Monte Carlo experiments.

For a quick comparison to answer this question, we look at the Monte Carlo experiments in Miller and Startz (2019) and Gonzales-Coya and Perron (2024) — the ones that directly come from Romano and Wolf (2017), since we already performed and discussed those experiments in the main text of our paper — and then compare the performance of their estimators with ours. This comparison, while not perfect, is perhaps not too off because the results for the WLS estimators, i.e., the common estimator computed by them and us, are similar.

Comparison with Miller and Startz (2019):

The common experiment in this case is the one with the Boston housing data. Table 9 collects the ratio of EMSEs with respect to OLS as the metric for comparing precision, from Miller and Startz (2019)'s Table 2 and our Table 6 in the main text, for the respective estimators with WLS being the common estimator.[4]

	from Table 2 of Miller and Startz (2019)					from our Table 6 (in main text)			
$h(\beta)$	WLS	KNN	Kernel	SVR	SVR-CV	WLS	TWLS	TCC	TGMM
β_1	.615	.456	.432	.515	.460	.613	.501	.500	.469
β_2	.674	.524	.471	.562	.510	.676	.562	.559	.486
β_3	.510	.420	.372	.446	.402	.506	.337	.337	.332
β_4	.504	.363	.358	.388	.374	.500	.348	.350	.317
β_5	.932	.866	.716	.804	.717	.927	.883	.896	.774

Table 9: Ratio of EMSE of estimators with respect to that of OLS under Romano and Wolf (2017)'s design with real-life data. WLS uses Romano and Wolf (2017)'s Model 1. Other estimators from Miller and Startz (2019) are k-nearest neighbors (KNN), local constant kernel regression (Kernel), and SVR with fixed and cross validated tuning parameters respectively (SVR and SVR-CV). Miller and Startz (2019) use 50000 Monte Carlo trials, we use 10000.

We wish to emphasize on two observations from Table 9. First, the relatively poor

[4]Although Miller and Startz (2019) refer to these numbers as the ratio of the empirical root mean squared errors, we believe that these are ratios without taking the square root, i.e., these are ratios of (E)MSEs.

precision of WLS compared to the other reported estimators makes it very likely that the user's parametric model for heteroskedasticity, i.e., Romano and Wolf (2017)'s Model 1, is incorrect. Consequently, the classical semiparametric estimators and the machine learning estimators that are all semiparametrically efficient are performing much better than WLS. The most precise semiparametric WLS estimator from Table 9, i.e., kernel, however has poor empirical size even after using Miller and Startz (2019)'s correction to the usual HC3 standard errors in the spirit of Rothenberg (1988); see Table 2 of Miller and Startz (2019).

The second observation is that, in spite of using this possibly incorrect model for heteroskedasticity, the targeted methods seem to be roughly as precise as the semiparametric estimators. This remarkable gain in precision while using the same incorrect model for heteroskedasticity happens because of targeting in the case of TWLS and because of targeting and the combination of OLS and WLS in the case of TCC and TGMM.

Comparison with Gonzales-Coya and Perron (2024):

The common experiment in this case is another experiment from Romano and Wolf (2017). Section 4.1 of our paper considers the entire experiment, but now we will consider the subset that overlaps with Gonzales-Coya and Perron (2024). Since Table 1 of Gonzales-Coya and Perron (2024) shows that precision gain due to their proposal happens primarily under Romano and Wolf (2017)'s Case 2(a), we will for brevity only focus on that case.

	from Tables 1 and S.1 of Gonzales-Coya and Perron (2024)						from our Table 2 (in main text)			
n	WLS	Lasso	SVR	LL	KNN	RF	WLS	TWLS	TCC	TGMM
100	.54	.54	.56	.55	.71	.68	.56	.51	.45	.48
200	.47	.46	.48	.49	.60	.64	.50	.45	.41	.41
400	.48	.46	.46	.48	.50	.56	.48	.47	.43	.41

Table 10: Ratio of EMSE of estimators with respect to that of OLS under Case 2(a) of Romano and Wolf (2017). WLS uses Romano and Wolf (2017)'s Model 1. Other estimators from Gonzales-Coya and Perron (2024) are their implementation of SVR, local linear regression (LL), k-nearest neighbor (KNN) and random forest (RF) to model heteroskedasticity. All results are based on 10000 Monte Carlo trials. The formatting of numbers with two places after decimal is maintained following Gonzales-Coya and Perron (2024)'s display.

Table 10 collects the ratio of EMSEs with respect to OLS, as the metric for comparing precision, from Tables 1 and S.1 in Gonzales-Coya and Perron (2024) and our Table 2 (main text) for the respective estimators. WLS based on Romano and Wolf (2017)'s Model 1 is the common estimator. We make the following observations based on Table 10.

First, as Gonzales-Coya and Perron (2024) note, semiparametric WLS based on Lasso and SVR are preferable to the other semiparametric WLS estimators. While LL also seems almost comparably good in this case, Gonzales-Coya and Perron (2024) note other problems with LL. On the other hand, KNN and RF seem to be much less precise than even classical WLS although Model 1 of Romano and Wolf (2017) used as the user's model for heteroskedasticity is clearly incorrect in this case.

Second, we again find that simply by virtue of targeting, TWLS performs favorably when compared to all the semiparametric methods. TCC and TGMM perform even better.

Third, this good performance of targeting compared to Gonzales-Coya and Perron (2024) is arguably compelling since the latter's results are somewhat biased in favor of machine learning techniques as the true skedastic function $\omega_0^2(x_i) = [\log(x_{i,2})]^2$ belongs to the space of functions $z_i = (1, x_{i,2}, [\log(x_{i,2})]^2, x_{i,2}^2, \cos(x_{i,2}), \cos(2x_{i,2}))'$ that they considered for learning. On the other hand, Romano and Wolf (2017)'s Model 1 $\omega^2(x_i;\gamma) = \exp(\gamma_1 + \gamma_2 \log(|x_{i,2}|))$ used by our targeting estimators certainly does not contain the true skedastic function $\omega_0^2(x_i) = [\log(x_{i,2})]^d$ for $d = 2, 4$, etc. We also note that $\omega_0^2(x_i) = [\log(x_{i,2})]^4$ is Romano and Wolf (2017)'s Case 2(b) for which our targeting methods provided even more improvement (see our Table 2 in the main text), but could not be used for comparison here since it is not considered in Gonzales-Coya and Perron (2024).

Finally, we note that Gonzales-Coya and Perron (2024) also consider a modified setup of the experiment in their robustness check by including in the user's model for heteroskedasticity a large number of variables that are irrelevant (at various degree) for heteroskedasticity. Predictably, Lasso leads to better precision than classical WLS in those cases. We also share

this view that when there are many possible covariates that could be included in the user's model, Lasso or some other variable selection method could be used to shrink the covariate set as a suggestive reference for the user. (We should note that it may not be clear what Lasso will do under violations of sparsity; see, e.g., Kolesar et al. (2025).)

Where we differ, and that has been the central message of our paper, is what to do with this shrunk set of covariates. Gonzales-Coya and Perron (2024) recommend doing WLS based on this shrunk set of covariates noting that they are agnostic about the correctness of the model that they thereby use. On the other hand, while we would also be similarly agnostic about the correctness of the model, we would instead recommend doing TWLS, TCC or TGMM based on this shrunk set of covariates. While such exploration of machine learning methods is beyond the scope of our current paper (we consider similar proposals in related work in other contexts), based on the extensive simulation evidence in our current paper it seems that such targeting strategies could very likely lead to much better precision.

D Bibliography

Ackerberg, D., Chen, X., and Hahn, J. (2012). A Practical Asymptotic Variance Estimator For Two-Step Semiparametric Estimators. *The Review of Economics and Statistics*, 94:481–498.

Breusch, T., Hailong, Q., Schmidt, P., and Wyhowski, D. (1999). Redundancy of moment conditions. *Journal of Econometrics*, 91: 89–111.

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

Chen, X., Jacho-Chavez, D. T., and Linton, O. (2016). Averaging of an increasing number of moment condition estimators. *Econometric Theory*, 32: 30–70.

Chernozhukov, V., Escanciano, J.-C., Ichimura, H., Newey, W., and Robins, J. (2022). Locally Robust Ssemiparametric Estimation. *Econometrica*, 90: 1501–1535.

DiCiccio, C. J., Romano, J. P., and Wolf, M. (2019). Improving weighted least squares inference. *Econometrics and Statistics*, 10:96–119.

Fan, J. and Yao, Q. (1998). Efficient estimation of conditional variance functions in stochastic regressions. *Biometrika*, 85: 645–660.

Gonzales-Coya, E. and Perron, P. (2024). Estimation in the Presence of Heteroskedasticity of Unknown Form: A Lasso-based Approach. *Journal of Econometric Methods*, 13:29–48.

Kolesar, M., Muller, U., and Roelsgaard, S. T. (2025). The Fragility of Sparsity. Working paper.

Miller, S. and Startz, R. (2019). Feasible Generalized Least Sqaures Using Machine Learning. *Economics Letters*, 175: 28–31.

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

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.

Rothenberg, T. J. (1988). Power Functions for Some Robust Tests of Regression Coefficients. *Econometrica*, 56:997–1019.