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

Saraswata Chaudhuri† and Eric Renault‡

This version: March 18, 2022.

Abstract

Several recent papers propose methods to estimate regression (conditional mean) parameters at least as precisely as the ordinary least squares (OLS) and parametric weighted LS (WLS) estimators even when the parametric model for the conditional variance of the regression error is misspecified. We show that an estimation principle described in Cragg [1992], when suitably adapted, outperforms all these estimators based on the same criterion that these estimators seek to optimize. We also demonstrate the same superior performance using simulations under the same designs used in these recent papers. This principle of estimation, without our adaptation, dates back to the early research on optimal design of experiments, and has also been gainfully used in the recent literatures on doubly-robust estimation and regression discontinuity design.

JEL Classification: C12; C13; C21.

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

*We thank F. Bugni, J. Galbraith, S. Goncalves, J-M. Dufour, J. MacKinnon, P.C.B. Phillips, C. Rothe, P. Sant'Anna, A. Santos, R. Startz, Y. Shin, K. Xu, and V. Zinde-Walsh for their very useful suggestions.

†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

Let $(y_i, x_i')_{i=1}^n$ be i.i.d. copies of the random variables (y, x') from a linear regression model:

$$y = x'\beta^0 + u \quad \text{with} \quad E[u|x] = 0 \quad \text{almost surely in } x. \tag{1}$$

Let $h(\beta)$ be our scalar parameter of interest with $h^0 := h(\beta^0)$ its true value.

In principle, a semiparametric weighted least squares estimator of $h(\beta)$ based on nonparametric estimation of $V(u|x)$ delivers semiparametric efficiency; see Carroll [1982], Robinson [1987], etc. However, it is rare to see such estimation in practice because nonparametric estimation of $V(u|x)$ generally requires very large sample size for the asymptotic properties of the semiparametric weighted least squares estimator to be good approximation of its finite-sample properties. Parametric weighted least squares, where $V(u|x)$ is estimated based on some user-specified parametric model, is also not an attractive solution because its precision can be even less than that of ordinary least squares (OLS) if the user-specified parametric model is incorrect.

Starting with the paper "Resurrecting weighted least squares" by Romano and Wolf [2017], the recent literature has come up with various interesting proposals to mitigate this twin problems with semiparametric and parametric weighted least squares; see, e.g., DiCiccio, Romano, and Wolf [2019], Spady and Stouli [2019], Lu and Wooldridge [2020], etc. Taking as given a user-specified and possibly incorrect parametric model $\omega^2(x;\gamma)$, known up to a finite dimensional parameter $\gamma \in \Gamma \subseteq \mathbb{R}^{d_\gamma}$, for $V(u|x)$ this literature proposes parametric estimators that improve upon OLS and parametric weighted least squares (WLS) estimators in terms of precision.

Our paper follows this recent literature and shows that we can obtain further substantial improvement in precision by an "optimal" treatment of γ in the parametric model $\omega^2(x;\gamma)$.[1] We classify the recently proposed estimators of $h(\beta)$ into three categories and consider their infeasible (with respect to γ) versions as functions of γ, i.e., we take the estimators without plugging in the values of γ that were proposed in the literature. Our proposed estimator of $h(\beta)$ under each category then plugs in an estimator of that γ that minimizes the asymptotic variance of that category's infeasible estimator of $h(\beta)$. By construction, the asymptotic variance of our proposed estimators of $h(\beta)$ cannot exceed that of any estimator of $h(\beta)$ in their respective categories. Simulations under the designs of these recent papers demonstrate that the gain in precision due to our proposal can be substantial without much cost even in small samples.

[1] Nonparametric estimators with"fixed" tuning parameters — e.g. series estimators with the number of terms in the series fixed — can be viewed as parametric estimators since the quality of their approximation of $V(u|x)$ does not get better with the increase in n. Although not considered explicitly, such estimators are covered by our discussion.

Our proposed estimators build on Cragg [1992]'s idea of minimizing the trace or determinant of the asymptotic variance of a similar infeasible version of the WLS estimator of β (denote it by $\widehat{\beta}_n(\gamma)$) with respect to γ. While Cragg [1992] does not discuss it, such minimization leads respectively the well-known notions of A and D optimality; see Elfving [1952], Chernoff [1953] and, respectively, Wald [1943]. These notions of optimality (and others, e.g., the E-optimality of Ehrenfeld [1956]; the L-optimality due to Karlin and Studden [1966] and Federov [1971]; Kiefer [1974]'s general optimality, etc.) would be compromises for the fact that unless $\omega^2(x;\gamma)$ correctly specifies $V(u|x)$, there is no guarantee of existence of a minimized (with respect to γ) asymptotic variance matrix of $\widehat{\beta}_n(\gamma)$. Without that existence, for some of the regression coefficients, the standard errors of estimators from using Cragg [1992] may exceed that from using WLS, and it is evident in hindsight that similar notions of optimality are not attractive in empirical work.[2] This concern of nonexistence is not just academic; we found ample evidence of its adverse effect resulting in Cragg's method having much larger than WLS standard error.

We bypass this critical issue of existence of the minimized matrix by reducing the problem to minimization of a scalar function, the asymptotic variance of an estimator of $h(\beta)$. Then, continuity of this function with respect to $\gamma \in \Gamma$ and compactness of Γ in $\mathbb{R}^{d_\gamma}$ ensure the existence of the minimized variance and its minimizer by the extreme value theorem.

Of course, if $\omega^2(x;\gamma)$ correctly specifies $V(u|x)$ then the "optimal" γ exists for β itself and hence works for all $h(\beta)$'s, e.g., elements of β. Then WLS (also OLS if $V(u|x)$ is constant), the recently proposed estimators, and our proposed estimators are all asymptotically equivalent and deliver semiparametric efficiency. Otherwise, our proposed estimators under each category deliver the "second-best" solution while WLS and others cannot, and OLS does not even try.

We conclude the introduction by noting that other literatures — see e.g. Cao, Tsiatis, and Davidian [2009] for doubly-robust estimation, Noack, Olma, and Rothe [2021] for regression discontinuity design, etc. — have also gainfully used ideas similar to that in our paper.

Our paper proceeds as follows. Section 2 begins with a discussion of the recently proposed estimators to motivate the construction of the infeasible (with respect to γ) estimators. Then it presents the algorithm for implementation of our proposed estimators based on minimizing with respect to γ the estimated asymptotic variance of these infeasible estimators. Finally, it presents the asymptotic properties of the proposed estimators and inference based on them. Section 3 demonstrates the superior finite-sample precision (without much cost otherwise) of the proposed

[2] This perhaps led to Cragg [1992]'s method being unfortunately overlooked in the empirical and theoretical literature. Even among the recent papers on this topic of improvement in precision over OLS and WLS, the only mention of Cragg [1992] is rather cursory — Romano and Wolf [2017] mention in their footnote 2 : "For some even earlier related work, see Cragg (1983, 1992), though he is mainly interested in estimation as opposed to inference."

estimators using the simulation designs and empirical examples from Romano and Wolf [2017] and Lu and Wooldridge [2020]. (Additional simulation results are available from us.) Section 4 concludes. Technical discussions and proofs of results are collected in the appendix.

2 Motivation, Implementation and Asymptotic properties

We will call the user's chosen parametric model $\omega^2(x;\gamma)$ correctly specified for $V(u|x)$ if:

$$\text{there exists } \gamma^0 \in \Gamma \subseteq \mathbb{R}^{d_\gamma} \text{ such that } \omega^2(x;\gamma^0) \propto V(u|x). \tag{2}$$

We will not maintain (2), but will only consider it as an unlikely special case. On the other hand, following the related literature and resembling common empirical practice, we will maintain that the user's parametric model $\omega^2(x;\gamma^0)$ can accommodate for conditional homoskedasticity of u, i.e.,

$$\text{there exists } \bar{\gamma} \in \Gamma \text{ such that } \omega^2(x;\bar{\gamma}) \propto 1. \tag{3}$$

2.1 Motivation behind the proposed estimators:

It will be useful at the outset to define the following building blocks to fix ideas and streamline the discussion. For any $\gamma \in \Gamma$ we define an infeasible weighted-by-$\omega^2(x;\gamma)$ estimator of $h(\beta)$ as:

$$\widehat{h}(\gamma) := h(\widehat{\beta}(\gamma)) \text{ where } \widehat{\beta}(\gamma) = \left(\sum_{i=1}^{n} \frac{x_i x_i'}{\omega^2(x_i;\gamma)}\right)^{-1} \sum_{i=1}^{n} \frac{x_i y_i}{\omega^2(x_i;\gamma)}. \tag{4}$$

To relate $\widehat{h}(\gamma)$ with the classical estimators, do note from (4) that the OLS and WLS estimators of $h(\beta)$ are $\widehat{h}_{OLS} \equiv \widehat{h}(\bar{\gamma})$ and $\widehat{h}_{WLS} = \widehat{h}(\widehat{\gamma}_{WLS})$ since that of β are, respectively, $\widehat{\beta}_{OLS} \equiv \widehat{\beta}(\bar{\gamma})$ and $\widehat{\beta}_{WLS} = \widehat{\beta}(\widehat{\gamma}_{WLS})$ where $\widehat{\gamma}_{WLS} \xrightarrow{p} \gamma_{WLS} := \arg\min_{\gamma\in\Gamma} E\left[\left(u^2 - \omega^2(x;\gamma)\right)^2\right]$.

For a heuristic discussion of the motivation here, with the precise statements postponed to Section 2.3, it will help to define the following components of the sandwich variance matrices:

$$\begin{aligned} &B_1 := E[xx'], \quad B_2(\gamma) := E\left[\frac{xx'}{\omega^2(x;\gamma)}\right], \quad B(\gamma) := [B_1(\gamma), B_2(\gamma)], \quad \text{and} \\ &C(\gamma) := \begin{bmatrix} C_{11} := E[V(u|x)xx'] & C_{12}(\gamma) := E\left[\frac{V(u|x)xx'}{\omega^2(x;\gamma)}\right] \\ C_{12}(\gamma) & C_{22}(\gamma) := E\left[\frac{V(u|x)xx'}{(\omega^2(x;\gamma))^2}\right] \end{bmatrix}. \end{aligned} \tag{5}$$

Now, consider any estimator $\widehat{\gamma} \xrightarrow{p} \gamma$ for some given $\gamma \in \Gamma$. It is well known that $E[u|x] = 0$ (see (1)) gives $E\left[\frac{xu}{\omega^2(x;\gamma)} \frac{\partial}{\partial\gamma'}\omega^2(x;\gamma)\right] = 0$ if $\frac{\partial}{\partial\gamma'}\omega^2(x;\gamma)$ exists (almost surely in x). Therefore, under standard conditions with $H := H(\beta^0)$ finite where $H(\beta) := \partial h(\beta^0)/\partial\beta'$, we have:

$$\begin{aligned}\sqrt{n}\left(\widehat{h}(\widehat{\gamma}) - h^0\right) &= \sqrt{n}\left(\widehat{h}(\gamma) - h^0\right) + o_p(1) \\ &= HB_2^{-1}(\gamma)\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{x_i u_i}{\omega^2(x_i;\gamma)} + o_p(1) \\ &\xrightarrow{d} N\left(0, \sigma^2(\gamma) := HB_2^{-1}(\gamma)C_{22}(\gamma)B_2^{-1}(\gamma)H'\right).\end{aligned} \tag{6}$$

Moreover, generalizing (6) using similar steps gives the joint distribution:

$$\begin{aligned}\sqrt{n}\begin{bmatrix}\widehat{h}_{OLS} - h^0 \\ \widehat{h}(\widehat{\gamma}) - h^0\end{bmatrix} &= \sqrt{n}\begin{bmatrix}\widehat{h}_{OLS} - h^0 \\ \widehat{h}(\gamma) - h^0\end{bmatrix} + o_p(1) \\ &= \begin{bmatrix} HB_1^{-1} & 0 \\ 0 & HB_2^{-1}(\gamma)\end{bmatrix}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\begin{bmatrix} x_i u_i \\ \frac{x_i u_i}{\omega^2(x_i;\gamma)}\end{bmatrix} + o_p(1) \\ &\xrightarrow{d} N\left(0, \Sigma(\gamma) := \begin{bmatrix} HB_1^{-1} & 0 \\ 0 & HB_2^{-1}(\gamma)\end{bmatrix} C(\gamma) \begin{bmatrix} HB_1^{-1} & 0 \\ 0 & HB_2^{-1}(\gamma)\end{bmatrix}'\right).\end{aligned} \tag{7}$$

With this background in place, we will divide the recently proposed estimators that improve upon OLS and WLS into three categories that all contain OLS and WLS as special cases.

- **Category 1:** Estimators of the form $\phi_n\widehat{h}(\widehat{\gamma}) + (1-\phi_n)\widehat{h}_{OLS}$ for: (i) some $\phi_n \xrightarrow{p} 1$ or $\phi_n \xrightarrow{p} 0$ and (ii) some $\widehat{\gamma} \xrightarrow{p} \gamma$ for some $\gamma \in \Gamma$. Therefore, under standard conditions, such estimators are asymptotically equivalent to either $\widehat{h}(\gamma)$ or $\widehat{h}_{OLS}$ depending on whether $\phi_n \xrightarrow{p} 1$ or $\phi_n \xrightarrow{p} 0$. Hence its asymptotic variance cannot be smaller than $\min\{\sigma^2(\gamma), \sigma^2(\bar{\gamma})\}$; see, (6) and (3).

  Romano and Wolf [2017]'s ALS estimator takes: (i) $\phi_n = 1$ if a consistent test cannot reject the null of homoskedasticity at some level α (e.g. 10%) and $\phi_n = 0$ otherwise; and (ii) $\widehat{\gamma} = \widehat{\gamma}_{WLS}$. DiCiccio, Romano, and Wolf [2019]'s MIN estimator takes: (i) $\phi_n = 1$ if $\widehat{h}_{WLS}$ has smaller standard error than $\widehat{h}_{OLS}$ and $\phi_n = 0$ otherwise; and (ii) $\widehat{\gamma} = \widehat{\gamma}_{WLS}$. Spady and Stouli [2019]'s estimator under $E[u|x] = 0$ takes: (i) $\phi_n \equiv 1$ for all $n \geq 1$, and (ii) $\widehat{\gamma} \xrightarrow{p} \gamma_{SS}$ where γ_{SS} solves $E\left[\frac{\partial}{\partial\gamma}\omega(X;\gamma_{SS})\frac{V(u|x)-\omega^2(X;\gamma_{SS})}{\omega^2(X;\gamma_{SS})}\right] = 0$; see their equation (3.9), Corollary 2.
- **Category 2:** Estimators of the form $\widehat{\lambda}(\widehat{\gamma})\widehat{h}(\widehat{\gamma}) + (1-\widehat{\lambda}(\widehat{\gamma}))\widehat{h}_{OLS}$ for some $\widehat{\gamma} \xrightarrow{p} \gamma$ for some $\gamma \in \Gamma$, and where $\widehat{\lambda}(\widehat{\gamma}) \xrightarrow{p} \lambda(\gamma) := \arg\min_{\lambda\in[0,1]} \text{Avar}\left(\lambda\widehat{h}(\widehat{\gamma}) + (1-\lambda)\widehat{h}_{OLS}\right)$, i.e.,

$$\lambda(\gamma) = \frac{\text{Avar}(\widehat{h}_{OLS}) - \text{Acov}(\widehat{h}_{OLS}, \widehat{h}(\gamma))}{\text{Avar}(\widehat{h}_{OLS}) + \text{Avar}(\widehat{h}(\gamma)) - 2\text{Acov}(\widehat{h}_{OLS}, \widehat{h}(\gamma))} \tag{8}$$

  with Avar and Acov denoting asymptotic variance and covariance respectively. Under standard conditions, we know from (7) that such estimators are asymptotically normal, asymptotically unbiased, and have asymptotic variance equal to:

$$\sigma^2_{cat2}(\gamma) \;:=\; \begin{bmatrix} 1-\lambda(\gamma) \\ \lambda(\gamma) \end{bmatrix}' \Sigma(\gamma) \begin{bmatrix} 1-\lambda(\gamma) \\ \lambda(\gamma) \end{bmatrix}. \tag{9}$$

DiCiccio, Romano, and Wolf [2019]'s convex combination (CC) estimator takes $\widehat{\gamma} = \widehat{\gamma}_{WLS}$.

- **Category 3:** Estimators of the form $h(\widehat{\beta}_{MC}(\widehat{\gamma}))$ where $\widehat{\beta}_{MC}(\widehat{\gamma})$ is a moment combination (MC) estimator, specifically the efficient GMM estimator:

$$\arg\min_{\beta} \left\{ \frac{1}{n}\sum_{i=1}^{n} \begin{bmatrix} x_i(y_i - x_i'\beta) \\ \frac{1}{\omega^2(x_i;\widehat{\gamma})} x_i(y_i - x_i'\beta) \end{bmatrix} \right\}' \widehat{C}^{+}(\widehat{\gamma}) \left\{ \frac{1}{n}\sum_{i=1}^{n} \begin{bmatrix} x_i(y_i - x_i'\beta) \\ \frac{1}{\omega^2(x_i;\widehat{\gamma})} x_i(y_i - x_i'\beta) \end{bmatrix} \right\} \tag{10}$$

for some $\widehat{\gamma} \xrightarrow{p} \gamma$ for some $\gamma \in \Gamma$, and $\widehat{C}^{+}(\widehat{\gamma}) \xrightarrow{p} C^{+}(\gamma)$ with the superscript $+$ denoting the Moore-Penrose (MP) inverse. For any $\gamma \in \Gamma$, if we write the four $d_\beta \times d_\beta$ (d_β being dimension of β) blocks of $\widehat{C}^{+}(\gamma)$ as $\widehat{C}^{+}_{ij}(\gamma)$ for $i, j = 1, 2$ then we obtain the closed-form expression:

$$\widehat{\beta}_{MC}(\gamma) = \widehat{\delta}(\gamma)\widehat{\beta}(\gamma) + (I_{d_\beta} - \widehat{\delta}(\gamma))\widehat{\beta}_{OLS} \tag{11}$$

where $\widehat{\delta}(\gamma) := \left(\widehat{B}(\gamma)\widehat{C}^{+}(\gamma)\widehat{B}'(\gamma)\right)^{-1} \left(\widehat{B}_1\widehat{C}^{+}_{12}(\gamma) + \widehat{B}_2(\gamma)\widehat{C}^{+}_{22}(\gamma)\right)\widehat{B}_2(\gamma)$ with the $\widehat{B}$'s and $\widehat{C}$'s denoting the sample analogs of the B' and C's (and defined precisely in Section 2.2). Under standard conditions and the conditions for convergence in probability of sample MP inverse to its population counterpart (see, e.g., Puri, Russell, and Mathew [1984]):

$$\begin{aligned}
\sqrt{n}\left(h(\widehat{\beta}_{MC}(\widehat{\gamma})) - h^0\right) \;&=\; \sqrt{n}\left(h(\widehat{\beta}_{MC}(\gamma)) - h^0\right) + o_p(1) \\
&= H\left(B(\gamma)C^{+}(\gamma)B'(\gamma)\right)^{-1}\left[B_1, B_2(\gamma)\right]C^{+}(\gamma)\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\begin{bmatrix} x_i u_i \\ \frac{1}{\omega^2(x_i;\widehat{\gamma})} x_i u_i \end{bmatrix} + o_p(1) \\
&\xrightarrow{d} N\left(0,\ \sigma^2_{cat3}(\gamma) \;:=\; H\left(B(\gamma)C^{+}(\gamma)B'(\gamma)\right)^{-1} H'\right).
\end{aligned} \tag{12}$$

(Full row-rank of the Jacobian via, e.g., a nonsingular B_1 is maintained throughout; see, e.g., Bonhomme and Weidner [2021].) One can take $\widehat{\gamma} = \widehat{\gamma}_{WLS}$. Alternatively, Lu and Wooldridge [2020]'s estimator uses the Gamma/Exponential quasi maximum likelihood estimator (QMLE) for $\widehat{\gamma}$, and the standard inverse in place of the MP inverse.[3] QMLE or any converging (in

[3]The standard inverse does not exist in the limit (population) if u is conditionally homoskedastic because then the asymptotic variance of the moment vector at the truth is rank-deficient and of rank equal to the dimension of β.

probability to some $\gamma \in \Gamma$) estimator $\widehat{\gamma}$ is also a valid option for all three categories.

The above description of the categories directly provides the motivation behind our proposed estimator. Building on Cragg [1992], for each category, we will use an estimator $\widehat{\gamma} \xrightarrow{p} \gamma^*$ for some γ^* that leads to the smallest asymptotic variance for that category. More precisely:

- **Category 1:** We will take $\phi_n \equiv 1$ for $n \geq 1$ and $\widehat{\gamma} = \widehat{\gamma}_{cat1}$ for some estimator $\widehat{\gamma}_{cat1} \xrightarrow{p} \gamma^*_{cat1} := \arg\min_{\gamma \in \Gamma} \sigma^2(\gamma)$; see (6). This leads to the proposed estimator being $\widehat{h}(\widehat{\gamma}_{cat1})$ with asymptotic variance $\sigma^{*^2}_{cat1} := \min_{\gamma \in \Gamma} \sigma^2(\gamma)$.
- **Category 2:** We will take $\widehat{\gamma} = \widehat{\gamma}_{cat2}$ for some estimator $\widehat{\gamma}_{cat2} \xrightarrow{p} \gamma^*_{cat2} := \arg\min_{\gamma \in \Gamma} \sigma^2_{cat2}(\gamma)$, see, (9). This leads to the proposed estimator being $\widehat{\lambda}(\widehat{\gamma}_{cat2})\widehat{h}(\widehat{\gamma}_{cat2}) + (1 - \widehat{\lambda}(\widehat{\gamma}_{cat2}))\widehat{h}_{OLS}$ with asymptotic variance $\sigma^{*^2}_{cat2} := \min_{\gamma \in \Gamma} \sigma^2_{cat2}(\gamma)$.
- **Category 3:** We will take $\widehat{\gamma} = \widehat{\gamma}_{cat3}$ for some estimator $\widehat{\gamma}_{cat3} \xrightarrow{p} \gamma^*_{cat3} := \arg\min_{\gamma \in \Gamma} \sigma^2_{cat3}(\gamma)$, see, (12). This leads to the proposed estimator being $h(\widehat{\beta}_{MC}(\widehat{\gamma}_{cat3}))$ with asymptotic variance $\sigma^{*^2}_{cat3} := \min_{\gamma \in \Gamma} \sigma^2_{cat3}(\gamma)$.

Remarks: Three remarks are in order. First, while $\sigma^{*^2}_{cat1} \geq \sigma^{*^2}_{cat2}$, in a given application the standard error of $\widehat{\lambda}(\widehat{\gamma}_{cat2})\widehat{h}(\widehat{\gamma}_{cat2}) + (1 - \widehat{\lambda}(\widehat{\gamma}_{cat2}))\widehat{h}_{OLS}$ may exceed that of $\widehat{h}(\widehat{\gamma}_{cat1})$. This is because in each category the optimal γ is obtained minimizing a sample variance based on some preliminary estimator ($\widehat{h}_{OLS}$, in effect, $\widehat{\beta}_{OLS}$) while, following convention, the standard error is computed based on that category's final/proposed estimator of $h(\beta)$; see Section 2.2 for details.

Second, Category 3 does not generalize Category 1 or holds equivalence with Category 2 unless β is a scalar like $h(\beta)$. The non-equivalence between Categories 2 and 3 is evident from comparing $\widehat{\lambda}(\gamma)\widehat{h}(\gamma) + (1 - \widehat{\lambda}(\gamma))\widehat{h}_{OLS}$ in Category 2 with $h(\widehat{\beta}_{MC}(\gamma))$ (see (11)) in Category 3 even if $h(\beta)$ is linear in β. Since our focus is on $h(\beta)$ and not β, this non-equivalence does not contradict Chen, Jacho-Chavez, and Linton [2016]. Their result — the optimal linear combination of estimators of β that are obtained by solving their respective just-identifying-for-β moment restrictions is the same as the efficient GMM estimator of β obtained by optimally combining all those just-identifying moment restrictions for β — is for β and not $h(\beta)$.

Third, while our proposal can in principle be extended to accommodate for a weighted version of Papadopoulosa and Tsionas [2021], it will require a separate treatment of the matter. Extension to nonlinear regressions as in Lin and Chou [2018] is more immediate. We do not pursue these interesting extensions to focus on our main message and keep the exposition simple.

2.2 Implementation of the proposed estimators:

Informed by (6), (9) and (12), we define the key sample quantities for implementation by category as follows. For $g \in \mathbb{R}^{d_\gamma}$ and $b, b_1, b_2 \in \mathbb{R}^{d_\beta}$ where d_β is the dimension of β, we define:

$$\begin{aligned}
\widehat{\sigma}^2_{cat1}(b,g) &:= H(b)\widehat{B}_2^{-1}(g)\widehat{C}_{22}(b,g)\widehat{B}_2^{-1}(g)H'(b), \\
\widehat{\sigma}^2_{cat2}(b_1,b_2,g) &:= \left[1-\widehat{\lambda}(b_1,b_2,g),\ \widehat{\lambda}(b_1,b_2,g)\right]\widehat{\Sigma}(b_1,b_2,g)\left[1-\widehat{\lambda}(b_1,b_2,g),\ \widehat{\lambda}(b_1,b_2,g)\right]', \\
\widehat{\sigma}^2_{cat3}(b,g) &:= H(b)\left(\widehat{B}(g)\widehat{C}^{+}(b,g)\widehat{B}'(g)\right)^{-1}H'(b),
\end{aligned}$$

where, resembling their population analogs in (5), (7) and (8), we have defined the components:

$$\begin{aligned}
\widehat{B}_1 &:= \frac{1}{n}\sum_{i=1}^n x_i x_i', \quad \widehat{B}_2(g) := \frac{1}{n}\sum_{i=1}^n \frac{x_i x_i'}{\omega^2(x_i;g)}, \quad \widehat{B}(g) := \left[\widehat{B}_1, \widehat{B}_2(g)\right], \\
\widehat{C}(b_1,b_2,g) &:= \begin{bmatrix} \widehat{C}_{11}(b_1) := \frac{1}{n}\sum_{i=1}^n (y_i - x_i'b_1)^2 x_i x_i' & \widehat{C}_{12}(b_1,b_2,g) := \frac{1}{n}\sum_{i=1}^n \frac{(y_i - x_i'b_1)(y_i - x_i'b_2)x_i x_i'}{\omega^2(x_i;g)} \\ \widehat{C}_{12}(b_1,b_2,g) & \widehat{C}_{22}(b_2,g) := \frac{1}{n}\sum_{i=1}^n \frac{(y_i - x_i'b_2)^2 x_i x_i'}{(\omega^2(x_i;g))^2} \end{bmatrix}, \\
\widehat{\Sigma}(b_1,b_2,g) &:= \begin{bmatrix} \widehat{\Sigma}_{11}(b_1,g) := H(b_1)\widehat{B}_1^{-1}\widehat{C}_{11}(b_1)\widehat{B}_1^{-1}H'(b_1) & \widehat{\Sigma}_{12}(b_1,b_2,g) := H(b_1)\widehat{B}_1^{-1}\widehat{C}_{12}(b_1,b_2,g)\widehat{B}_2^{-1}(g)H'(b_2) \\ \widehat{\Sigma}_{12}(b_1,b_2,g) & \widehat{\Sigma}_{22}(b_2,g) : H(b_2)\widehat{B}_2^{-1}(g)\widehat{C}_{22}(b_2,g)\widehat{B}_2^{-1}(g)H'(b_2) \end{bmatrix}, \\
\widehat{\lambda}(b_1,b_2,g) &:= \frac{\widehat{\Sigma}_{11}(b_1,g) - \widehat{\Sigma}_{12}(b_1,b_2,g)}{\widehat{\Sigma}_{11}(b_1,g) + \widehat{\Sigma}_{22}(b_2,g) - 2\widehat{\Sigma}_{12}(b_1,b_2,g)}.
\end{aligned}$$

The proposed algorithm involves three steps for each category. Step 1 constructs the suitable sample objective function for γ. Step 2 estimates the optimal γ by minimizing that sample objective function. Step 3 uses the estimated optimal γ to obtain the proposed estimator of $h(\beta)$ and thereafter its standard error. To streamline notation, we only use $\widehat{h}_{OLS}$ (in effect, $\widehat{\beta}_{OLS}$) to obtain the objective function in Step 1, while we use the estimated proposed estimator (and the associated estimator for β) to compute the standard error of the proposed estimator.

Steps for the proposed estimator under Category 1:

1. Using the OLS estimator $\widehat{\beta}_{OLS}$ obtain $\widehat{\sigma}^2_{cat1}(\widehat{\beta}_{OLS}, \gamma)$ as a function of γ.
2. Obtain the minimizer $\widehat{\gamma}_{cat1} := \arg\min_{\gamma\in\Gamma} \widehat{\sigma}^2_{cat1}(\widehat{\beta}_{OLS}, \gamma)$.
3. Obtain $\widehat{h}_{cat1} := \widehat{h}(\widehat{\gamma}_{cat1})$ as in (4) and its standard error $se_{cat1,n} := \sqrt{\widehat{\sigma}^2_{cat1}(\widehat{\beta}(\widehat{\gamma}_{cat1}), \widehat{\gamma}_{cat1})\big/ n}$.

Steps for the proposed estimator under Category 2:

1. Using the OLS estimator $\widehat{\beta}_{OLS}$ obtain $\widehat{\sigma}^2_{cat2}(\widehat{\beta}_{OLS}, \widehat{\beta}_{OLS}, \gamma)$ as a function of γ.

2. Obtain the minimizer $\widehat{\gamma}_{cat2} := \arg\min_{\gamma\in\Gamma} \widehat{\sigma}^2_{cat2}(\widehat{\beta}_{OLS}, \widehat{\beta}_{OLS}, \gamma)$.

3. Obtain $\widehat{h}_{cat2} := \widehat{\lambda}(\widehat{\gamma}_{cat2})\widehat{h}(\widehat{\gamma}_{cat2}) + (1 - \widehat{\lambda}(\widehat{\gamma}_{cat2}))\widehat{h}_{OLS}$ and its standard error $se_{cat2,n} := \sqrt{\widehat{\sigma}^2_{cat2}(\widehat{\beta}_{OLS}, \widehat{\beta}(\widehat{\gamma}_{cat2}), \widehat{\gamma}_{cat2})\Big/ n}$.

Steps for the proposed estimator under Category 3:

1. Using the OLS estimator $\widehat{\beta}_{OLS}$ obtain $\widehat{\sigma}^2_{cat3}(\widehat{\beta}_{OLS}, \gamma)$ as a function of γ.

2. Obtain the minimizer $\widehat{\gamma}_{cat3} := \arg\min_{\gamma\in\Gamma} \widehat{\sigma}^2_{cat3}(\widehat{\beta}_{OLS}, \gamma)$.

3. Obtain $\widehat{h}_{cat3} := h(\widehat{\beta}_{MC}(\widehat{\gamma}_{cat3}))$ as in (10)/(11) and its standard error $se_{cat3,n} := \sqrt{\widehat{\sigma}^2_{cat3}(\widehat{\beta}_{MC}(\widehat{\gamma}_{cat3}), \widehat{\gamma}_{cat3})\Big/ n}$.

More refined implementation — e.g., iteration of steps or joint estimation of $h(\beta)$ and γ, and (in cases of concern with bias) even cross-fitting — is also possible. If so preferred, one could use the so-called HC3-robust standard errors (specifically, the HC3 version of $\widehat{C}(.)$) at least in step 3, or use bootstrap for inference; see, e.g., Romano and Wolf [2017] and DiCiccio, Romano, and Wolf [2019] respectively.[4] Nevertheless, we recommended the simple implementation above because our experience so far with simulations under the designs of the related papers suggests that it works well even in small samples under the simple framework of those papers and ours.

2.3 Asymptotic properties of the proposed estimators:

Assumptions:

A1. $\gamma^*_j = \arg\inf_{\gamma\in\Gamma} \sigma^2_j(\gamma)$ exists for $j = cat1, cat2, cat3$.

A2. For any $\delta > 0$ and $j = cat1, cat2, cat3$ there exists $\epsilon(\delta) > 0$ such that: $\inf_{\gamma\in\Gamma:\|\gamma-\gamma^*_j\|>\delta} |\sigma^2_j(\gamma) - \sigma^2_j(\gamma^*_j)| \geq \epsilon(\delta)$.

A3. For any $\delta_n \downarrow 0$ and all $\gamma \in \Gamma : \|\gamma - \gamma^*_j\| \leq \delta_n$ and $j = cat1, cat2, cat3$ there exists a constant $M > 0$ such that: $|\sigma^2_j(\gamma) - \sigma^2_j(\gamma^*_j)| \geq M\|\gamma - \gamma^*_j\|$.

A4. $H(\beta) := \partial h(\beta)/\partial\beta$ exists in an open ball around β^0 and is continuous at β^0.

A5. $\widehat{B}(\gamma) := [\widehat{B}_1, \widehat{B}_2(\gamma)] \xrightarrow{p} B(\gamma) := [B_1, B_2(\gamma)]$, $[\widehat{B}_1^{-1}, \widehat{B}_2^{-1}(\gamma)] \xrightarrow{p} [B_1^{-1}, B_2^{-1}(\gamma)]$, $\widehat{C}(b_1, b_2, \gamma) \xrightarrow{p} C(\gamma)$ and $\widehat{C}^+(b_1, b_2, \gamma) \xrightarrow{p} C^+(\gamma)$ uniformly in $\gamma \in \Gamma$ for any $\widehat{b}_1, \widehat{b}_2 \xrightarrow{p} \beta^0$.

[4] HC3 version is straightforward for the proposed estimator in Category 1; but is more challenging in Categories 2 and 3. In fact, due to the covariance terms, the HC3 version may not even be positive (semi) definite in small samples for Category 2. Also, a development similar to Lin and Chou [2018] does not guarantee positive (semi) definite HC3 version in small samples for Category 3. Nevertheless, the asymptotic results in the next subsection will remain unchanged due to the asymptotic equivalence of the various HC-robust standard errors; see, e.g., Theorem 7.6 in Hansen [2020] whose proof works in our case with minor and obvious modifications; while finite-sample inference will possibly improve due to reduced over-rejection of the truth unless the non-positive-definiteness affects the standard ordering HC1 $\geq$ HC2 $\geq$ $HC3$. The theory for validity of pairs and wild bootstrap can similarly be developed following DiCiccio, Romano, and Wolf [2019]. However, the real justification behind HC3 or bootstrap, i.e., the proof of asymptotic refinement (if any) due to them is, as usual, quite complicated and beyond the scope of our paper.

A6. $\widehat{C}(b_1, b_2, \gamma) - C(\gamma) = O_p(n^{-1/2})$, $\widehat{C}^{+}(b_1, b_2, \gamma) - C^{+}(\gamma) = O_p(n^{-1/2})$ and (as implied by A5) $[\widehat{B}_1, \widehat{B}_2(\gamma)] - [B_1, B_2(\gamma)] = o_p(1)$, $[\widehat{B}_1^{-1}, \widehat{B}_2^{-1}(\gamma)] - [B_1^{-1}, B_2^{-1}(\gamma)] = o_p(1)$ uniformly in $\gamma \in \Gamma : \|\gamma - \gamma_j^*\| \leq \delta_n$ for $j = cat1, cat2, cat3$, for any $\delta_n \downarrow 0$, and any $\widehat{b}_1, \widehat{b}_2 \xrightarrow{p} \beta^0$.

A7. $\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left[x_i u_i, x_i u_i / \omega^2(x_i; \gamma_j^*)\right] \xrightarrow{d} N\left(0, C(\gamma_j)\right)$ for $j = cat1, cat2, cat3$.

A8. There exist a $1 \times d_\gamma$ vector $\Delta_{1,j}(x)$ and a $\Delta_{2,j}(x) \geq 0$ with $E\|xu\Delta_{2,j}(x)\| \leq \infty$ such that for $j = cat1, cat2, cat3$, the following holds with probability one for large n and some $\delta > 0$: $\sup_{\gamma \in \Gamma : \|\gamma - \gamma_j^*\| \leq \delta} \left\{ \left|1/\omega^2(x;\gamma) - 1/\omega^2(x;\gamma_j^*) - \Delta_{1,j}(x)(\gamma - \gamma_j^*)\right| - \frac{1}{2}\Delta_{2,j}(x)\|\|\gamma - \gamma_j^*\|^2 \right\} \leq 0.$

Remarks: The existence condition in A1 can be ensured, e.g., by assuming $\sigma_j^2(\gamma)$ for $j = cat1, cat2, cat3$ is continuous in $\gamma \in \Gamma$ and Γ is compact in $\mathbb{R}^{d_\gamma}$ where d_γ is finite. It is typically difficult to provide primitive conditions for the global identification condition of the optimal γ in A2. The local identification condition of the optimal γ in A3 can be satisfied in various ways, e.g., $\sigma_j^2(\gamma)$ for $j = cat1, cat2, cat3$ is differentiable with non-zero derivative at $\gamma = \gamma_j^*$. A4 is a standard assumption enabling the use of the delta-method, and also in conjunction with A5 and A6 leading to the consistency of the $\widehat{\sigma}_j^2(.)$'s for the $\sigma_j^2(.)$'s. A5 is a standard uniform convergence assumption and under our setup can be satisfied if, e.g., in addition to pointwise convergence of the concerned quantities (via, e.g., continuity and existence of moments), $\omega^2(x;\gamma)$ is bounded away from 0 for $\gamma \in \Gamma$ with probability one. A6 strengthens A5 locally by imposing a rate condition that leads to the rate of convergence of $\widehat{\gamma}_j$ to γ_j^* for $j = cat1, cat2, cat3$. A7 is a standard asymptotic joint distribution assumption that follows from conventional conditions for the central limit theorem. A8 imposes standard smoothness conditions on $1/\omega^2(x;\gamma)$ locally.

Our main results below are based on A1-A8 and the various definitions heretofore.

Lemma 1

(a) *Let assumptions A1, A2, A4 and A5 hold. Then* $\widehat{\gamma}_j \xrightarrow{p} \gamma_j^*$ *for* $j = cat1, cat2, cat3$.

(b) *Let* $\widehat{\gamma}_j \xrightarrow{p} \gamma_j^*$ *for* $j = cat1, cat2, cat3$ *and assumptions A1, A3, A4 and A6 hold. Then* $\widehat{\gamma}_j - \gamma_j^* = O_p(n^{-1/2})$ *for* $j = cat1, cat2, cat3$.

Remark: The result of Lemma 1(b) is stronger than required since, as is well known in similar contexts, $\widehat{\gamma}_j - \gamma_j^* = o_p(n^{-1/4})$ for $j = cat1, cat2, cat3$ could have been made sufficient for our purpose. However, the $n^{-1/2}$ rate follows naturally since the $\widehat{\gamma}_j$'s are parametric estimators.

Using these properties of $\widehat{\gamma}_j$ for $j = cat1, cat2, cat3$ we will now establish the asymptotic properties of the proposed estimators and the standard Wald-inference based on them.

Theorem 1 *Let* $\widehat{\gamma}_j - \gamma_j^* = O_p(n^{-1/2})$ *for* $j = cat1, cat2, cat3$. *Let assumptions A4, A7, A8, and A6 (allowing a weaker form that replaces the* $O_p(n^{-1/2})$ *rates by* $o_p(1)$*) hold. Then:*

(a) $\sqrt{n}(\widehat{h}_j - h^0) \xrightarrow{d} N(0, \sigma_j^{*^2})$ *for* $j = cat1, cat2, cat3$*;*

(b) *the test that rejects the null* $K_{null} : h(\beta) = h_{null}$ *against the alternative* $K_{alt} : h(\beta) \neq h_{null}$ *if* $|(\widehat{h}_j - h_{null})/se_{j,n}| > z_{1-\alpha/2}$ *has asymptotic power* $\Phi(z_{\alpha/2} + \mu/\sigma_j^*) + \Phi(z_{\alpha/2} - \mu/\sigma_j^*)$ *for* $j = cat1, cat2, cat3$ *and* $h^0 = h_{null} + \mu/\sqrt{n}$ *where* z_c *satisfies* $\Phi(z_c) = c \in (0, 1)$*;*

(c) *the confidence interval* $\left[\widehat{h}_j - z_{1-\alpha/2}se_{j,n}, \widehat{h}_j + z_{1-\alpha/2}se_{j,n}\right]$ *for* $h(\beta)$ *has asymptotic coverage* $1 - \alpha$ *for* $j = cat1, cat2, cat3$*.*

Remark: The proposed estimators and standard Wald inference based on them have the desired asymptotic properties. One-sided inference can be done similarly. First-order asymptotically, the proposed estimators cannot perform worse than the estimators in their respective categories.

3 Simulation evidence and Empirical illustrations

We will explore the small-sample performance of the proposed estimators under all three categories using simulation experiments based on 10000 Monte Carlo trials. The estimators:

- OLS and WLS, that belong in all three categories, are put under the label classical estimators;
- ALS, MIN and the proposed estimator, named modified WLS (MWLS), under Category 1
- CC and the proposed estimator, named modified CC (MCC), under Category 2
- MC using WLS and QML, denoted respectively as MCls and MCqm, and the proposed estimator, named modified MC (MMC) under Category 3,

will be included in the study.[5] We do not include the estimator from the working paper Spady and Stouli [2019] since its stated purpose is different from that of the ones above. We will use the simulation designs in Romano and Wolf [2017] and Lu and Wooldridge [2020]; the design in DiCiccio, Romano, and Wolf [2019] is similar to that in Romano and Wolf [2017].[6] We will also revisit the empirical illustrations in Romano and Wolf [2017] and Lu and Wooldridge [2020].

The main message of the numerical results here is that if the user's model $\omega^2(x;\gamma)$ for $V(u|x)$ allows for improvement in precision over the existing estimators then the proposed estimators achieve it. Like Romano and Wolf [2017], we report the improvement in the empirical mean squared error (MSE), and find that its reduction by the proposed estimators can be huge by any conceivable standard. Under all cases there does not seem to be any major cost, in terms

[5] We got helpful suggestions for more informative names of the proposed estimators, e.g., "targeted" or "minimax" WLS, CC, MC, etc. that may have other connotations. We opted for the generic name "modified" to avoid controversy.

[6] The extensive simulation study here, of which only a subset of results is presented while the rest are available from us, complements Rilstone [1991]'s early simulations that focused on OLS, WLS and its semiparametric versions.

of empirical bias, size, etc., to using the proposed estimators. Comparison among the proposed estimators across categories does not however give a clear winner. Based on these observations and the simplicity of the estimators we recommend all three proposed estimators in practice.

3.1 Simulations under the design in Romano and Wolf [2017]:

Romano and Wolf [2017] take $y = x_{(1)}\beta_1 + x_{(2)}\beta_2 + u$ in (1), with $x_{(1)} = 1, x_{(2)} \sim U(1,4)$, $x = (x_{(1)}, x_{(2)})'$; $\beta = (\beta_1, \beta_2)'$, $\beta^0 = (0,0)'$; $u = s(x)z$ where $z \sim N(0,1)$ is independent of $x_{(2)}$ and thus $E[u|x] = 0$ and $V(u|x) = s^2(x)$. They consider 10 cases for the skedastic function:

Case 1: (a) $s^2(x) = 1$; (b) $s^2(x) = x_{(2)}$; (c) $s^2(x) = x_{(2)}^2$; (d) $s^2(x) = x_{(2)}^4$.

Case 2: (a) $s^2(x) = \big(\log(x_{(2)})\big)^2$; (b) $s^2(x) = \big(\log(x_{(2)})\big)^4$.

Case 3: (a) $s^2(x) = \exp\Big(.1(x_{(2)} + x_{(2)}^2)\Big)$; (b) $s^2(x) = \exp\Big(.15(x_{(2)} + x_{(2)}^2)\Big)$.

Case 4: (a) $s^2(x) = \begin{cases} 1 & \text{if } x_{(2)} < 2 \\ 2 & \text{if } 2 \le x_{(2)} < 3 \\ 3 & \text{if } x_{(2)} \ge 3 \end{cases}$; (b) $s^2(x) = \begin{cases} 1 & \text{if } x_{(2)} < 2 \\ 2^2 & \text{if } 2 \le x_{(2)} < 3 \\ 3^2 & \text{if } x_{(2)} \ge 3 \end{cases}$.

To emphasize the gain in precision, we will add a Case 2(c) with $s^2(x) = \big(\log(x_{(2)})\big)^6$.

Romano and Wolf [2017] consider two parametric models $\omega^2(x;\gamma)$ — Model 1: $\omega^2(x;\gamma) := \exp(\gamma_1 + \gamma_2 \log(x_{(2)}))$ and Model 2: $\omega^2(x;\gamma) := \exp(\gamma_1 + \gamma_2 x_{(2)})$ — and like them our results here are also very similar for both models. However, since there is slightly more action in terms of improved precision in case of estimators based on Model 2, for brevity we report here the results based on Model 2 only (the unreported results are available from us).[7]

Romano and Wolf [2017] report for β_2 the empirical MSE's (their ratios) of estimators, empirical coverage probability of 95% confidence intervals (1 - empirical size of 5% t tests) and ratios of the average length of these intervals. We will do the same while considering sample sizes $n = 50, 100, 200, 400$. We take the parameter of interest $h(\beta)$ as β_1 and β_2 respectively.

Tables 1 and 2 present, respectively for β_1 and β_2, the ratio of the empirical MSE of each estimator with respect to that of OLS. Besides Case 1(a) (conditional homoskedasticity), the other estimators lead to smaller, sometimes much smaller, MSE. (To compare any two non-OLS estimators, say A with respect to B, divide the ratio under A with that under B.) Importantly, the proposed estimator under each category either performs very similar to the other estimators in the category or leads to really big gain in precision as in Cases 2 (a), (b) and (c).

[7] Model 1 is correct for $V(u|x)$ in the sense of (2) under Cases 1(a)-1(d) with $\gamma_2^0 = 0, 1, 2, 4$ respectively. Model 2 is correct for $V(u|x)$ only under Case 1(a) with $\gamma_2^0 = 0$. So, all estimators are asymptotically efficient under Case 1(a), and all estimators other than OLS are asymptotically efficient under Cases 1(b)-1(d) when using Model 1.

Tables 3 and 4 present, respectively for β_1 and β_2, the empirical size (empirical rejection probability of the truth) of 5% Wald tests based on each estimator. The results look reasonable except in the case of the MC estimators with small samples. This happens because being true to Lu and Wooldridge [2020] we use HC0 standard error for the MC, i.e., Category 3, estimators and, as is well-known, that does have an adverse effect in small samples. While the size-corrected empirical power is not reported here for brevity (but is available from us), we note that the proposed estimator in each category always has either the same or substantially greater (in Cases 2) empirical size-corrected power than its competitors.

Tables 5 and 6 present, respectively for β_1 and β_2, the average length of each of the non-OLS confidence intervals with respect that of the OLS intervals. For brevity we report this for Case 2 only where, as noted above, the benefit of the proposed estimators' precision is most prominently evident. These are indeed big gains in precision of confidence intervals by any standard.

3.2 Simulations under the design in Lu and Wooldridge [2020]:

Lu and Wooldridge [2020] take $y = x_{(1)}\beta_1 + x_{(2)}\beta_2 + x_{(3)}\beta_3 + x_{(3)}\beta_4 + u$ in (1), with $x_{(1)} = 1, x_{(2)} \sim N(1,1)$, $x_{(3)} = .8 + .2x_{(2)} + e_1$, $x_{(4)} = 1(x_{(5)} > x_{(3)})$, $u = s(x)e_3$ where e_1, e_2, e_3 are independent $N(0,1)$, and $x_{(5)} = .3 + .1x_{(2)} + .1x_{(3)} + e_2$. They take $x = (x_{(1)}, x_{(2)}, x_{(3)}, x_{(4)})'$, e_3 as independent of x (giving $E[u|x] = 0$ and $V(u|x) = s^2(x)$), and $\beta = (\beta_1, \beta_2, \beta_3, \beta_4)'$ with $\beta^0 = (.5, 1, 1, 1)'$. They consider 4 cases for the skedastic function:

Case 1: $s^2(x) = (\beta_1 + \beta_2 x_{(2)} + \beta_3 x_{(3)} - 3\beta_4 x_{(4)} + .1x_{(2)}(x_{(3)} + x_{(4)}) - .1x_{(3)}x_{(4)} - .05x_{(2)}^2 + .05x_{(3)}^2))^2$

Case 2: $s^2(x) = (\beta_1 + \beta_2|x_{(2)}| + \beta_3 x_{(3)}^2 + \beta_4 x_{(4)})^2$.

Case 3: $s^2(x) = \exp(\beta_1 + \beta_2|x_{(2)}| + \beta_4 x_{(4)})$.

Case 4: $s^2(x) = \exp(\beta_1 + \beta_2 x_{(2)} + \beta_3 x_{(3)} + \beta_4 x_{(4)})$.

They consider the parametric model $\omega^2(x;\gamma) = \exp(x'\gamma)$, which is correct for $V(u|x)$ in the sense of (2) with $\gamma^0 = \beta^0$ in Case 4.

We take $h(\beta) = \beta_1, \beta_2, \beta_3, \beta_4$ respectively and sample size $n = 1000, 5000$. Lu and Wooldridge [2020] take $n = 1000, 10000$ and report Monte Carlo mean and standard deviations in their Table 1. In this case, the large sample size largely mitigates concerns with inference and, therefore, similar to Lu and Wooldridge [2020] we focus and report results here only for estimation.

Table 7 presents the ratio of the empirical MSE of each estimator with respect to that of OLS.[8] It is of interest to note that in our implementation of Cases 1 and 2, WLS based on

[8] Our results for WLS are not the same as Lu and Wooldridge [2020]'s because they use Gamma QMLE for γ in WLS whereas we use the conventional WLS. Our results for MCqm should have been the same as their GMM results

an incorrect model $\omega^2(x;\gamma)$ can be much less precise than OLS, which is a possibility that DiCiccio, Romano, and Wolf [2019] (p.2, paragraph 7) noted as motivation to their MIN and CC estimators but conjectured as "rare". ALS also suffers from the same problem in this case since ALS and WLS are very similar here because of high level of heteroskedasticity of u.

On the other hand, the MIN, CC and MCC estimators deliver big gains in precision over OLS. Additionally, when the parametric model $\omega^2(x;\gamma)$ is far from correct for $V(u|x)$, i.e., Cases 1 and 2, we see that our proposed estimators deliver further substantial gains in precision. However, when $\omega^2(x;\gamma)$ is correct for $V(u|x)$, i.e., in Case 4, there is no room for improvement since all non-OLS estimators are then asymptotically efficient (not considering the information that β's appear in both $E[y|x]$ and $V(y|x)$). Then our proposed estimators are less precise than their non-OLS competitors. This problem however diminishes with larger sample size $n = 5000$.

3.3 Empirically relevant simulations in Romano and Wolf [2017]:

Romano and Wolf [2017]'s simulation based on a real-life example revisits the well-known cross-sectional data set from 1970 containing $n = 506$ observations from communities in the Boston area (see, Wooldridge [2012]). They consider a linear regression as in (1) with:

$$E[y|x] = x'\beta = x_{(1)}\beta_1 + x_{(2)}\beta_2 + x_{(3)}\beta_3 + x_{(4)}\beta_4 + x_{(5)}\beta_5$$

where y is the log of the median housing price in a community, $x_{(1)} = 1$, $x_{(2)}$ is the log of nitrogen oxide in the air (in parts per million), $x_{(3)}$ is the log of weighted distance from five employment centers (in miles), $x_{(4)}$ is the average number of rooms per house, and $x_{(5)}$ is the average student–teacher ratio in the community's schools.

To mimic the true conditional heteroskedasticity in this data, Romano and Wolf [2017]: (i) obtain $\widehat{e}_i = (y_i - x'\widehat{\beta}_{OLS})/\sqrt{1-q_{i,i}}$ for $i = 1,\dots,n$ where $q_{i,i} = x_i'(\sum_j x_j x_j')^{-1}x_i$ is i-th diagonal element of the hat-matrix; (ii) generate artificial data (y_i^*, x_i^*) for $i = 1,\dots,n$ where $x_i^* = x_i$ and $y_i^* = x_i'\widehat{\beta}_{OLS} + \widehat{e}_i v_i$ where $v_i \sim N(0,1)$ independently of the system. Thus, the true β in this artificial data is $\widehat{\beta}_{OLS}$. Romano and Wolf [2017] then report for each element of β the empirical MSE's (their ratios) of estimators, empirical coverage probability of 95% confidence intervals (1 - empirical size of 5% t tests) and ratios of the average length of these intervals.

We will do the same, and since the improvement shown by Romano and Wolf [2017] is noticeably better with their Model 1, i.e., $\omega^2(x;\gamma) = \exp(\gamma_1 + \sum_{k=2}^5 \log(x_{(k)}))$, we will for

because both use Gamma QMLE for γ. The results were not close. To avoid a negative representation of Lu and Wooldridge [2020]'s estimator due to possible computational error on our part, we will not report MCqm hereafter.

brevity only report the further improvement provided by our proposed estimators based on Model 1. These are reported in Tables 8, 9 and 10 respectively for the ratio of the empirical MSE's with respect to OLS, the empirical size of 5% Wald test, and the ratio of the average length of confidence intervals based on other estimators to that based on OLS. As is clearly evident, the proposed estimators deliver noticeably big further gains over its competitors.

3.4 Empirical illustration in Lu and Wooldridge [2020]:

Lu and Wooldridge [2020] use a subset of the well-known cross-sectional individual-level data set ' 401ksubs' (see Wooldridge [2012]) to estimate a linear regression as in (1) with:

$$E[y|x] = x'\beta = \sum_{k=1}^{10} x_{(k)}\beta_k$$

where y is net total financial assets (in \$ 1000) and is denoted by "nettfa"; $x_{(1)} = 1$ and is denoted by "constant"; $x_{(2)}$ is annual income (in \$1000) in excess of population (data) average and is denoted by "inc_0"; $x_{(3)} = x_{(2)}^2$ and is denoted by "inc_0^2"; $x_{(4)}$ is age in excess of population (data) average and is denoted by "age_0"; $x_{(5)} = x_{(4)}^2$ and is denoted by "age_0^2"; $x_{(6)} = x_{(2)} \times x_{(4)}$ and is denoted by "$\text{inc}_0.\text{age}_0$"; $x_{(7)}$ is a dummy variable for eligibility for a 401k plan and is denoted by "e401k"; $x_{(8)}$ is a dummy variable for male and is denoted by "male"; $x_{(9)} = x_{(7)} \times x_{(2)}$ and is denoted by "e401k.inc_0"; and $x_{(10)} = x_{(7)} \times x_{(4)}$ and is denoted by "e401k.age_0".

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 11 the various estimates and standard errors (in parentheses) for the coefficients of this regression model. We use Lu and Wooldridge [2020] parametric model $\omega^2(x;\gamma) = \exp(x'\gamma)$. Lu and Wooldridge [2020] showed big gains in precision by WLS over OLS, and then further improvement over WLS by their GMM estimator. Our results in Table 11 of course confirm these findings of Lu and Wooldridge [2020]. Additionally, our results also demonstrate that even further gains, and often substantial ones, in precision over all those estimators can be obtained by our proposed estimators.

4 Conclusion

Inspired by Romano and Wolf [2017], our paper followed the recent literature that tries to improve upon the OLS and (parametric) WLS estimators. This literature takes the user's parametric model $\omega^2(x;\gamma)$ for $V(u|x)$ as given, without assuming that it is correct, and focuses on estimating the coefficients in a regression model given by $y = E[y|x] + u$ where $E[y|x] = x'\beta$. We showed that an old idea from Cragg [1992] can be suitably adapted to improve not only upon OLS and WLS, but also upon the recently proposed estimators in this literature.

Compared to Cragg [1983], that takes a more nonparametric approach to estimating $V(u|x)$ and coincides with the explosion of nonparametric estimation in theoretical econometrics, Cragg [1992] seemed to have been largely overlooked. This might have been because the optimization program of minimizing the determinant or trace of the asymptotic variance of the estimators of the regression coefficients often delivers poor (individually sub-optimal) standard errors for the individual coefficients that are typically of interest in applied research. (They may be optimal in other sense, e.g., minimized volume of the Wald joint-confidence set for all regression coefficients, an attractive criterion in the early design of experiments.) While Cragg [1992] does not discuss the motivation behind his specific optimization-proposals, the issue is that such optimizations are compromises for the fact that a minimizer of the asymptotic variance matrix itself (in a matrix sense) may not exist unless $\omega^2(x;\gamma)$ is a correct model for $V(u|x)$. Our adaptation of Cragg [1992] bypassed the issue of existence by instead focusing on scalar functions of the regression coefficients, e.g., the individual coefficients, their sums, differences, etc., that are typically the focus in applied research. We showed how this adaptation led to our proposed estimators that are conceptually very simple and based on elementary econometric theory. We also demonstrated, using a variety of simulation experiments from the recent literature, the substantial improvements that our proposed estimators can provide over the existing estimators.

References

S. Bonhomme and M. Weidner. Minimizing sensitivity to model misspecification. Forthcoming: Quantitative Economics, 2021.

W. Cao, A. Tsiatis, and M. Davidian. Improving Efficiency and Robustness of the Doubly Robust Estimator for a Population Mean with Incomplete Data. *Biometrika*, 96:723–734, 2009.

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

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

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

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

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

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

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

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

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

B. E. Hansen. Econometrics. Online Textbook, 2020.

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

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

E. S. Lin and T-S. Chou. Finite-sample refinement of GMM approach to nonlinear models under heteroskedasticity of unknown form. *Econometric Reviews*, 37: 1–28, 2018.

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

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

C. Noack, T. Olma, and C. Rothe. Flexible Covariate Adjustments in Regression Discontinuity Designs. arXiv:2107.07942 [econ.EM], 2021.

A. Papadopoulosa and M. G. Tsionas. Efficiency gains in least squares estimation: A new approach. Forthcoming: Econometric Reviews, 2021.

M. L. Puri, C. T. Russell, and T. Mathew. Convergence of Generalized Inverses with Applications to Asymptotic Hypothesis Testing. *Sankhya: Series A*, 46: 277–286, 1984.

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

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

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

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

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

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

True	Sample	Classical	Category 1			Category 2		Category 3		
$V(u\|x)$	size	WLS	ALS	MIN	MWLS	CC	MCC	MCls	MCqm	MMC
Case (1a)	50	1.0348	1.0348	1.0217	1.0592	1.0184	1.0818	1.0788	1.0787	1.1093
	100	1.0201	1.0201	1.0124	1.0409	1.0108	1.0635	1.0572	1.0569	1.0763
	200	1.0116	1.0116	1.0070	1.0201	1.0063	1.0331	1.0276	1.0273	1.0341
	400	1.0072	1.0072	1.0036	1.0082	1.0028	1.0183	1.0148	1.0148	1.0193
Case (1b)	50	.9302	.9518	.9325	.9391	.9286	1.0099	.9606	.9466	.9753
	100	.9162	.9242	.9260	.9307	.9207	.9964	.9459	.9357	.9551
	200	.9075	.9082	.9088	.9130	.9099	.9425	.9175	.9153	.9271
	400	.8884	.8885	.8887	.8864	.8892	.9000	.8909	.8891	.8985
Case (1c)	50	.6765	.6853	.6812	.6763	.6791	.7092	.7205	.6828	.7078
	100	.6674	.6677	.6688	.6718	.6714	.7051	.7006	.6781	.6885
	200	.6608	.6608	.6608	.6621	.6623	.6679	.6742	.6692	.6721
	400	.6330	.6330	.6330	.6298	.6330	.6296	.6403	.6362	.6328
Case (1d)	50	.2677	.2677	.2677	.2736	.2683	.2437	.3752	.3651	.2867
	100	.2494	.2494	.2494	.2534	.2500	.2360	.3227	.3394	.2557
	200	.2426	.2426	.2426	.2428	.2427	.2304	.2958	.3109	.2382
	400	.2230	.2230	.2230	.2207	.2223	.2126	.2506	.2798	.2126
Case (2a)	50	.4139	.4139	.4139	.3585	.4128	.2527	.3611	.4530	.3015
	100	.4251	.4251	.4251	.3547	.4247	.2385	.3608	.4892	.2566
	200	.4136	.4136	.4136	.3623	.4137	.2424	.3707	.5034	.2274
	400	.3864	.3864	.3864	.3339	.3862	.2321	.3421	.4777	.2039
Case (2b)	50	.2082	.2082	.2082	.1975	.2091	.1237	.2083	.3122	.1764
	100	.1864	.1864	.1864	.1558	.1870	.0908	.1806	.3331	.1324
	200	.1772	.1772	.1772	.1333	.1778	.0800	.1751	.3416	.1027
	400	.1591	.1591	.1591	.1079	.1590	.0756	.1540	.3153	.0780
Case (2c)	50	.1374	.1374	.1374	.1243	.1381	.0280	.1343	.2340	.1211
	100	.1008	.1008	.1008	.0823	.1010	.0200	.0957	.2348	.0801
	200	.0881	.0881	.0881	.0529	.0882	.0177	.0772	.2508	.0519
	400	.0753	.0753	.0753	.0359	.0754	.0169	.0619	.2128	.0342
Case (3a)	50	.8628	.8954	.8738	.8736	.8675	.9553	.9129	.8851	.9377
	100	.8457	.8547	.8596	.8595	.8532	.9270	.8887	.8693	.9097
	200	.8371	.8377	.8382	.8433	.8402	.8633	.8541	.8463	.8631
	400	.8100	.8100	.8102	.8101	.8109	.8161	.8174	.8116	.8231
Case (3b)	50	.6717	.6841	.6813	.6803	.6780	.7326	.7468	.6863	.7379
	100	.6547	.6553	.6602	.6643	.6610	.7048	.7278	.6688	.7047
	200	.6474	.6474	.6472	.6518	.6496	.6606	.6934	.6517	.6671
	400	.6135	.6135	.6135	.6135	.6142	.6164	.6509	.6113	.6193
Case (4a)	50	.9444	.9541	.9491	.9517	.9416	1.0404	.9782	.9625	.9999
	100	.9264	.9269	.9383	.9386	.9296	1.0112	.9611	.9473	.9636
	200	.9150	.9150	.9218	.9206	.9171	.9445	.9295	.9239	.9318
	400	.8960	.8960	.8983	.8960	.8956	.9055	.9039	.8988	.9030
Case (4b)	50	.7208	.7382	.7323	.7226	.7202	.7865	.7641	.7216	.7472
	100	.6992	.7006	.7069	.7035	.6998	.7506	.7288	.6990	.7182
	200	.6858	.6858	.6862	.6866	.6842	.6981	.6946	.6816	.6959
	400	.6606	.6606	.6609	.6561	.6557	.6607	.6587	.6500	.6543

Table 1: Ratio of MSE of estimators with respect to MSE of OLS estimator of $h(\beta) := \beta_1$ based on 10000 Monte Carlo trials under the design of Romano and Wolf [2017] and using their Model 2.

True $V(u\|x)$	Sample size	Classical WLS	Category 1 ALS	Category 1 MIN	Category 1 MWLS	Category 2 CC	Category 2 MCC	Category 3 MCls	Category 3 MCqm	Category 3 MMC
Case (1a)	50	1.0400	1.0400	1.0239	1.0492	1.0206	1.0731	1.0737	1.0732	1.0937
	100	1.0238	1.0238	1.0164	1.0385	1.0137	1.0620	1.0572	1.0570	1.0703
	200	1.0137	1.0137	1.0081	1.0199	1.0073	1.0347	1.0289	1.0287	1.0344
	400	1.0088	1.0088	1.0047	1.0091	1.0037	1.0188	1.0162	1.0161	1.0192
Case (1b)	50	.9472	.9683	.9542	.9497	.9437	.9869	.9692	.9655	.9856
	100	.9326	.9402	.9435	.9424	.9368	.9772	.9564	.9546	.9648
	200	.9226	.9232	.9270	.9267	.9249	.9388	.9320	.9333	.9424
	400	.9069	.9069	.9084	.9050	.9067	.9104	.9091	.9099	.9154
Case (1c)	50	.7556	.7624	.7665	.7592	.7578	.7613	.7769	.7736	.7921
	100	.7382	.7383	.7439	.7432	.7425	.7495	.7574	.7559	.7648
	200	.7289	.7289	.7291	.7316	.7307	.7291	.7351	.7420	.7419
	400	.7062	.7062	.7062	.7042	.7048	.7005	.7084	.7115	.7049
Case (1d)	50	.4289	.4289	.4298	.4378	.4327	.4095	.5454	.5356	.4540
	100	.3812	.3812	.3813	.3859	.3829	.3679	.4839	.4768	.3890
	200	.3658	.3658	.3658	.3684	.3659	.3531	.4619	.4392	.3604
	400	.3436	.3436	.3436	.3434	.3410	.3306	.4173	.4025	.3293
Case (2a)	50	.6218	.6218	.6250	.6289	.6234	.5451	.6153	.6870	.5790
	100	.6035	.6035	.6037	.5967	.6046	.4942	.5775	.6781	.4969
	200	.5980	.5980	.5980	.5963	.5983	.4851	.5782	.6797	.4572
	400	.5716	.5716	.5716	.5662	.5682	.4622	.5442	.6491	.4258
Case (2b)	50	.4149	.4149	.4160	.4236	.4221	.3269	.4290	.5452	.3775
	100	.3562	.3562	.3563	.3374	.3599	.2441	.3570	.5199	.2814
	200	.3384	.3384	.3384	.3080	.3403	.2057	.3406	.5128	.2292
	400	.3151	.3151	.3151	.2730	.3138	.1983	.3107	.4801	.1943
Case (2c)	50	.2743	.2743	.2744	.2267	.2777	.0774	.2533	.4050	.2233
	100	.1989	.1989	.1989	.1468	.1998	.0547	.1790	.3716	.1454
	200	.1728	.1728	.1728	.1021	.1730	.0457	.1518	.3730	.0995
	400	.1539	.1539	.1539	.0784	.1540	.0451	.1305	.3297	.0735
Case (3a)	50	.8617	.8953	.8742	.8638	.8662	.9015	.8939	.8832	.9118
	100	.8450	.8540	.8575	.8514	.8521	.8807	.8738	.8670	.8895
	200	.8344	.8349	.8359	.8364	.8375	.8459	.8471	.8441	.8572
	400	.8109	.8109	.8111	.8084	.8114	.8118	.8140	.8118	.8187
Case (3b)	50	.6925	.7036	.7014	.6921	.6990	.7082	.7339	.7074	.7369
	100	.6728	.6733	.6776	.6745	.6783	.6880	.7220	.6845	.7062
	200	.6615	.6615	.6615	.6611	.6637	.6655	.6939	.6654	.6770
	400	.6318	.6318	.6318	.6290	.6322	.6309	.6598	.6284	.6335
Case (4a)	50	.9658	.9750	.9738	.9691	.9589	1.0097	.9904	.9839	1.0041
	100	.9467	.9473	.9624	.9551	.9480	.9918	.9732	.9675	.9751
	200	.9332	.9332	.9420	.9362	.9333	.9486	.9454	.9425	.9505
	400	.9181	.9181	.9215	.9154	.9152	.9203	.9226	.9196	.9230
Case (4b)	50	.8132	.8256	.8387	.8217	.8096	.8331	.8345	.8190	.8370
	100	.7821	.7831	.7973	.7851	.7785	.7987	.7960	.7824	.7982
	200	.7642	.7642	.7665	.7623	.7583	.7657	.7662	.7587	.7709
	400	.7446	.7446	.7450	.7347	.7322	.7356	.7359	.7292	.7311

Table 2: Ratio of MSE of estimators with respect to MSE 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 2.

True $V(u\|x)$	Sample size	Classical		Category 1			Category 2		Category 3		
		OLS	WLS	ALS	MIN	MWLS	CC	MCC	MCls	MCqm	MMC
Case (1a)	50	5.42	6.06	6.06	6.11	7.13	6.04	8.22	9.57	9.66	11.82
	100	4.70	4.93	4.93	4.93	5.60	4.96	6.29	6.91	6.83	7.66
	200	4.88	5.03	5.03	5.06	5.11	5.03	5.60	5.65	5.65	5.99
	400	4.83	4.92	4.92	4.89	4.99	4.90	5.13	5.29	5.28	5.45
Case (1b)	50	5.03	5.96	6.19	6.11	7.00	6.11	9.24	9.75	9.10	11.22
	100	4.58	5.14	5.19	5.26	5.78	5.25	7.77	7.04	6.67	7.58
	200	4.79	5.09	5.10	5.12	5.40	5.17	6.29	6.01	5.85	6.39
	400	4.87	4.96	4.96	4.97	4.99	4.94	5.29	5.29	5.40	5.56
Case (1c)	50	4.46	5.52	5.66	5.56	6.34	5.64	8.61	11.73	8.33	10.44
	100	4.68	5.18	5.19	5.18	5.74	5.25	7.43	8.82	6.68	7.93
	200	4.82	4.93	4.93	4.93	5.33	5.10	5.99	6.66	5.80	6.71
	400	4.90	4.98	4.98	4.98	5.06	5.05	5.40	5.88	5.38	5.79
Case (1d)	50	4.64	5.04	5.04	5.04	5.77	5.22	6.60	14.06	6.44	10.29
	100	4.95	4.96	4.96	4.96	5.35	5.07	6.31	11.80	6.05	8.95
	200	5.01	5.10	5.10	5.10	5.34	5.30	5.59	8.08	5.53	7.05
	400	4.75	4.90	4.90	4.90	5.10	5.04	5.14	6.07	5.14	5.85
Case (2a)	50	4.04	4.31	4.31	4.31	4.86	4.38	5.40	7.19	6.79	9.65
	100	4.72	4.99	4.99	4.99	4.78	5.00	4.90	6.28	6.29	8.03
	200	4.96	5.20	5.20	5.20	5.21	5.17	4.46	5.83	5.75	6.21
	400	4.92	5.06	5.06	5.06	4.74	5.06	4.73	5.21	5.30	5.36
Case (2b)	50	4.44	5.01	5.01	5.01	5.80	5.10	8.35	8.99	8.00	10.13
	100	5.13	5.10	5.10	5.10	5.33	5.21	5.93	7.37	6.70	8.33
	200	4.93	5.05	5.05	5.05	4.97	5.19	4.82	6.10	5.97	7.03
	400	4.73	4.95	4.95	4.95	4.62	4.95	5.06	5.47	5.54	6.18
Case (2c)	50	4.86	5.29	5.29	5.29	6.27	5.44	5.80	11.87	8.46	11.16
	100	5.41	5.20	5.20	5.20	5.96	5.21	4.57	9.78	6.89	9.21
	200	4.99	5.19	5.19	5.19	5.26	5.19	4.39	6.91	5.99	7.17
	400	4.88	5.06	5.06	5.06	4.76	5.05	4.88	5.92	5.32	6.63
Case (3a)	50	4.92	6.01	6.32	6.21	6.93	6.27	9.21	10.57	8.97	12.67
	100	4.65	5.07	5.19	5.21	5.82	5.34	7.37	7.52	6.84	8.51
	200	4.82	5.04	5.06	5.10	5.40	5.29	5.84	6.22	5.82	6.60
	400	4.92	4.83	4.83	4.83	5.10	4.90	5.22	5.50	5.48	5.74
Case (3b)	50	4.57	5.87	6.05	5.96	6.66	6.06	8.88	12.10	8.53	12.33
	100	4.81	5.12	5.12	5.18	5.85	5.43	7.09	8.68	7.00	8.79
	200	4.80	5.06	5.06	5.05	5.40	5.14	5.66	6.47	5.79	6.48
	400	4.87	5.01	5.01	5.01	5.09	5.08	5.26	5.72	5.43	5.70
Case (4a)	50	4.98	5.91	5.96	6.13	6.85	6.19	9.16	9.62	9.01	11.18
	100	4.65	5.04	5.04	5.27	5.75	5.25	7.36	7.06	6.82	7.63
	200	4.75	5.13	5.13	5.23	5.44	5.30	6.03	6.00	5.80	6.31
	400	4.80	4.87	4.87	4.91	5.04	4.95	5.17	5.32	5.30	5.40
Case (4b)	50	4.53	5.76	5.95	5.91	6.24	6.14	8.72	10.79	8.57	10.29
	100	4.55	5.34	5.36	5.47	5.74	5.56	7.21	7.98	6.69	7.62
	200	4.85	5.00	5.00	5.01	5.18	5.21	5.49	6.08	5.83	6.19
	400	4.83	5.01	5.01	5.01	5.11	5.08	5.33	5.69	5.44	5.72

Table 3: Empirical size (in %) of 5% Wald test for $h(\beta) := \beta_1$ based on 10000 Monte Carlo trials under the simulation design of Romano and Wolf [2017] and using their Model 2.

True $V(u\|x)$	Sample size	Classical		Category 1			Category 2		Category 3		
		OLS	WLS	ALS	MIN	MWLS	CC	MCC	MCls	MCqm	MMC
Case (1a)	50	5.15	5.64	5.64	5.75	6.35	5.70	7.09	8.65	8.62	9.93
	100	4.75	5.12	5.12	5.15	5.69	5.14	6.16	6.76	6.75	7.41
	200	4.87	5.01	5.01	5.04	5.21	5.03	5.52	5.80	5.76	6.05
	400	5.00	5.12	5.12	5.15	5.22	5.13	5.35	5.56	5.52	5.61
Case (1b)	50	4.78	5.14	5.41	5.44	5.74	5.31	7.25	8.35	8.06	9.19
	100	4.70	5.12	5.19	5.23	5.59	5.20	6.53	6.88	6.71	7.34
	200	4.82	4.90	4.89	4.98	5.25	5.08	5.61	5.97	5.85	6.26
	400	4.84	5.03	5.03	5.04	4.95	5.02	5.25	5.32	5.45	5.65
Case (1c)	50	4.86	5.09	5.18	5.25	5.71	5.34	6.57	8.29	7.95	8.79
	100	4.92	4.98	4.99	5.07	5.38	5.27	6.15	7.08	6.89	7.37
	200	5.02	5.03	5.03	5.04	5.31	5.19	5.66	5.94	5.86	6.45
	400	4.94	5.19	5.19	5.19	5.13	5.15	5.35	5.65	5.39	5.66
Case (1d)	50	5.28	5.22	5.22	5.25	5.83	5.60	6.39	8.98	6.82	9.04
	100	5.25	5.06	5.06	5.06	5.35	5.21	5.97	8.29	6.04	7.49
	200	5.06	5.15	5.15	5.15	5.12	5.23	5.44	6.47	5.65	6.33
	400	4.86	5.01	5.01	5.01	5.10	5.13	5.07	5.54	5.27	5.64
Case (2a)	50	4.98	4.98	4.98	5.04	5.43	5.34	6.29	7.81	7.90	8.62
	100	4.99	5.03	5.03	5.03	5.43	5.17	5.62	6.61	6.83	7.02
	200	4.89	5.17	5.17	5.17	5.24	5.27	5.16	5.89	5.74	5.78
	400	4.89	5.02	5.02	5.02	5.05	5.06	4.80	5.36	5.40	5.03
Case (2b)	50	5.26	5.09	5.09	5.11	6.33	5.48	8.41	8.74	8.45	9.78
	100	5.27	5.05	5.05	5.05	5.65	5.24	6.33	6.98	6.97	7.62
	200	5.03	5.08	5.08	5.08	5.18	5.23	4.99	5.88	5.92	6.54
	400	4.87	5.02	5.02	5.02	4.76	4.98	5.18	5.38	5.45	5.63
Case (2c)	50	5.32	5.49	5.49	5.49	6.41	5.61	6.22	10.35	8.54	10.60
	100	5.46	5.25	5.25	5.25	5.73	5.32	5.04	8.29	6.97	8.56
	200	5.00	5.16	5.16	5.16	5.10	5.20	4.66	6.28	6.04	6.89
	400	4.91	5.01	5.01	5.01	4.73	5.02	4.81	5.53	5.26	6.23
Case (3a)	50	4.81	5.30	5.64	5.54	5.80	5.53	6.97	8.68	8.15	9.92
	100	4.80	5.06	5.12	5.28	5.59	5.21	6.23	7.02	6.85	7.70
	200	4.95	5.06	5.08	5.09	5.34	5.09	5.45	6.01	6.01	6.51
	400	4.86	5.08	5.08	5.08	5.01	5.05	5.15	5.40	5.39	5.48
Case (3b)	50	4.96	5.17	5.37	5.35	5.76	5.54	6.61	8.76	7.94	9.65
	100	4.94	4.95	4.96	5.01	5.53	5.18	5.93	7.21	6.79	7.83
	200	5.05	5.20	5.20	5.21	5.38	5.24	5.50	5.89	6.00	6.25
	400	4.84	5.14	5.14	5.14	5.14	5.16	5.15	5.56	5.34	5.56
Case (4a)	50	4.83	5.26	5.31	5.59	5.76	5.47	7.23	8.42	8.15	9.15
	100	4.76	5.15	5.15	5.45	5.53	5.34	6.30	6.79	6.81	7.17
	200	4.87	4.90	4.90	5.05	5.27	5.03	5.59	5.95	6.01	6.23
	400	4.82	5.05	5.05	5.09	4.92	5.01	5.15	5.30	5.34	5.51
Case (4b)	50	4.95	5.00	5.21	5.49	5.71	5.40	6.81	8.46	8.17	8.91
	100	4.96	4.98	4.99	5.20	5.48	5.24	6.11	7.02	6.72	7.47
	200	5.03	5.08	5.08	5.12	5.31	5.26	5.50	5.94	5.89	6.12
	400	4.80	5.11	5.11	5.12	5.25	5.17	5.37	5.57	5.37	5.50

Table 4: 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 2.

True $V(u\|x)$	Sample size	Classical WLS	Category 1 ALS	Category 1 MIN	Category 1 MWLS	Category 2 CC	Category 2 MCC	Category 3 MCls	Category 3 MCqm	Category 3 MMC
	50	.6128		.6128	.5676	.6114	.4864	.5221	.5983	.4353
Case	100	.6285		.6285	.5730	.6281	.4792	.5535	.6522	.4286
(2a)	200	.6309	same	.6309	.5914	.6306	.4925	.5828	.6790	.4396
	400	.6224		.6224	.5794	.6223	.4844	.5798	.6817	.4376
	50	.4317		.4317	.4023	.4287	.3118	.3775	.4871	.3233
Case	100	.4155		.4155	.3679	.4143	.2855	.3778	.5306	.2948
(2b)	200	.4132	as	.4132	.3565	.4123	.2805	.3954	.5560	.2824
	400	.3974		.3974	.3322	.3970	.2736	.3837	.5480	.2624
	50	.3415		.3415	.2917	.3407	.1627	.2676	.4163	.2452
Case	100	.3007		.3007	.2415	.3005	.1413	.2442	.4436	.2127
(2c)	200	.2891	WLS	.2891	.2092	.2891	.1336	.2501	.4723	.1881
	400	.2715		.2715	.1813	.2715	.1298	.2367	.4488	.1638

Table 5: Ratio of the average length of confidence intervals of $h(\beta) := \beta_1$ using each estimators with respect to the average length of confidence intervals using OLS. Results are based on 10000 Monte Carlo trials under the design of Romano and Wolf [2017] and using their Model 2.

True $V(u\|x)$	Sample size	Classical WLS	Category 1 ALS	Category 1 MIN	Category 1 MWLS	Category 2 CC	Category 2 MCC	Category 3 MCls	Category 3 MCqm	Category 3 MMC
	50	.7705		.7700	.7554	.7607	.6973	.6928	.7311	.6538
Case	100	.7605		.7605	.7475	.7557	.6779	.7057	.7619	.6395
(2a)	200	.7639	same	.7639	.7573	.7596	.6870	.7294	.7861	.6469
	400	.7526		.7526	.7469	.7499	.6762	.7253	.7877	.6405
	50	.6292		.6291	.6044	.6210	.5058	.5666	.6459	.5039
Case	100	.5858		.5858	.5560	.5819	.4634	.5483	.6648	.4620
(2b)	200	.5754	as	.5754	.5426	.5722	.4461	.5574	.6821	.4412
	400	.5569		.5569	.5215	.5548	.4403	.5446	.6741	.4233
	50	.4985		.4985	.4237	.4969	.2676	.4044	.5557	.3581
Case	100	.4306		.4306	.3488	.4302	.2326	.3633	.5628	.3087
(2c)	200	.4078	WLS	.4078	.3038	.4077	.2141	.3627	.5800	.2741
	400	.3875		.3875	.2727	.3874	.2110	.3479	.5582	.2479

Table 6: Ratio of the average length of confidence intervals of $h(\beta) := \beta_2$ using each estimators with respect to the average length of confidence intervals using OLS. Results are based on 10000 Monte Carlo trials under the design of Romano and Wolf [2017] and using their Model 2.

True $V(u\|x)$	$h(\beta)$	Classical WLS	Category 1 ALS	Category 1 MIN	Category 1 MWLS	Category 2 CC	Category 2 MCC	Category 3 MCls	Category 3 MCqm	Category 3 MMC
Case (1)	β_1	.8316	.8285	.7995	.5318	.7893	.5164	.6330	.7003	.4490
	β_2	1.0231	1.0416	.9488	.7593	.9039	.8829	.8509	.8464	.6861
	β_3	.9011	.8935	.8361	.6249	.8289	.6845	.7263	.7149	.5598
	β_4	1.6956	1.6735	.9817	.8987	1.0012	.9871	.8498	.8384	.7813
Case (2)	β_1	1.4923	1.4483	.8719	.5154	.7906	.5781	.4181	.5391	.3868
	β_2	1.4674	1.5110	.9761	.7178	.8530	.7431	.8093	.7541	.7708
	β_3	2.4286	2.3621	.9205	.5298	.8066	.6043	.4395	.5139	.4139
	β_4	1.4274	1.4204	.8926	.6629	.7923	.6654	.6217	.6199	.5544
Case (3)	β_1	.8672	.8521	.8536	.8623	.8666	.8714	.8267	.8078	.8041
	β_2	.7987	.8403	.8376	.7617	.7894	.7902	.7622	.7567	.6957
	β_3	.8095	.8104	.8112	.8002	.8075	.8518	.7933	.7839	.8336
	β_4	.9655	.9497	.9462	.9502	.9603	.9625	.8731	.8604	.8047
Case (4)	β_1	.1684	.1616	.1616	.1847	.1686	.1798	.1777	.3412	.1960
	β_2	.0716	.0721	.0721	.1080	.0717	.0941	.0887	.2116	.1145
	β_3	.0723	.0724	.0724	.1091	.0724	.0973	.0897	.2249	.1157
	β_4	.1183	.1158	.1158	.1435	.1185	.1392	.1331	.2843	.1470
Case (1)	β_1	.8134	.8230	.8213	.5985	.7858	.5652	.6388	.6947	.4741
	β_2	1.0198	1.0440	.9885	.7960	.9204	.7784	.8854	.8974	.6408
	β_3	.9035	.9248	.9079	.6720	.8437	.6695	.7704	.6801	.5349
	β_4	1.7835	1.7916	1.0158	.9002	1.0001	.9989	.8567	.8461	.7641
Case (2)	β_1	1.7941	1.6708	1.0041	.5623	.8552	.5869	.4394	.5470	.4094
	β_2	1.5993	1.6056	.9950	.7316	.8892	.7303	.8642	.8219	.7458
	β_3	3.2366	3.0016	1.0293	.5868	.8905	.6093	.4464	.5266	.4033
	β_4	1.7205	1.5769	.9730	.6868	.8436	.6885	.6373	.6346	.5654
Case (3)	β_1	.8745	.8658	.8659	.8704	.8737	.8741	.8346	.8249	.7828
	β_2	.7985	.8184	.8184	.7865	.7974	.7878	.7886	.7862	.6314
	β_3	.8081	.8167	.8168	.8025	.8074	.8030	.8011	.8036	.7949
	β_4	.9597	.9387	.9386	.9552	.9577	.9564	.8774	.8527	.7954
Case (4)	β_1	.1568	.1529	.1529	.1621	.1568	.1596	.1596	.3289	.1627
	β_2	.0615	.0609	.0609	.0785	.0616	.0719	.0691	.2447	.0805
	β_3	.0662	.0654	.0654	.0805	.0662	.0760	.0732	.2569	.0829
	β_4	.1114	.1094	.1094	.1192	.1115	.1164	.1161	.3015	.1191

Table 7: Ratio of MSE of estimators with respect to MSE of OLS estimator of various $h(\beta)$'s based on 10000 Monte Carlo trials under the design of Lu and Wooldridge [2020]. The top panel (above the horizontal line) corresponds to sample size $n = 1000$, and the bottom panel to $n = 5000$. The parametric model $\omega^2(x;\gamma)$ is correctly specified for $V(u|x)$ in the sense of (2) under Case 4.

$h(\beta)$	Classical	Category 1			Category 2		Category 3	
	WLS	ALS	MIN	MWLS	CC	MCC	MCls	MMC
β_1	.6063		.6064	.4910	.6064	.5018	.5452	.4732
β_2	.6681	same	.6687	.5553	.6675	.5524	.5982	.4844
β_3	.5055	as	.5056	.3422	.5056	.3403	.4141	.3329
β_4	.4963	WLS	.4963	.3396	.4963	.3521	.3936	.3155
β_5	.9330		.9228	.8893	.9118	.9063	.8250	.7762

Table 8: Ratio of MSE of estimators with respect to MSE of OLS estimator of coefficients based on 10000 Monte Carlo trials under the empirical design of Romano and Wolf [2017] [c.f. their Table C7] and using their Model 1 that, in their Table C7, performed noticeably better than Model 2.

$h(\beta)$	Classical		Category 1			Category 2		Category 3	
	OLS	WLS	ALS	MIN	MWLS	CC	MCC	MCls	MMC
β_1	4.65	5.09		5.09	6.60	5.09	7.28	5.76	7.50
β_2	4.70	4.79	same	4.82	5.77	4.82	5.91	5.49	6.27
β_3	4.99	4.90	as	4.90	6.27	4.91	6.38	5.70	6.95
β_4	4.17	4.74	WLS	4.74	7.18	4.74	8.51	6.02	7.94
β_5	4.80	5.22		5.37	5.48	5.39	5.84	5.94	6.43

Table 9: Empirical size (in %) of 5% Wald test for coefficients based on 10000 Monte Carlo trials under the empirical design of Romano and Wolf [2017] [c.f. their Table C8] and using their Model 1 that, in their Table C8, performed noticeably better than Model 2.

$h(\beta)$	Classical	Category 1			Category 2		Category 3	
	WLS	ALS	MIN	MWLS	CC	MCC	MCls	MMC
β_1	.7781		.7781	.6626	.7781	.6542	.7170	.6318
β_2	.8132	same	.8129	.7230	.8124	.7199	.7523	.6562
β_3	.7132	as	.7132	.5664	.7131	.5633	.6320	.5443
β_4	.7067	WLS	.7067	.5470	.7067	.5340	.6080	.5129
β_5	.9522		.9434	.9218	.9385	.9272	.8701	.8242

Table 10: Ratio of the average length of confidence interval for each $h(\beta)$ using each estimators with respect to the average length of confidence interval of that $h(\beta)$ using OLS. Results based on 10000 Monte Carlo trials under the empirical design of Romano and Wolf [2017][c.f. their Table C8] and using their Model 1 that, in their Table C8, performed noticeably better than Model 2.

$h(\beta)$	Classical		Category 1			Category 2		Category 3	
	OLS	WLS	ALS	MIN	MWLS	CC	MCC	MCls	MMC
constant	5.905	6.394			6.214	6.352	6.074	6.619	6.196
	(2.115)	(.977)			(.912)	(.961)	(.910)	(.906)	(.867)
inc_0	.633	.464			.478	.482	.473	.499	.457
	(.152)	(.063)			(.056)	(.061)	(.055)	(.054)	(.048)
inc_0^2	.000	.003			.001	.003	.002	.002	.002
	(.005)	(.002)			(.002)	(.002)	(.002)	(.002)	(.002)
age_0	.704	.605			.597	.608	.581	.677	.626
	(.141)	(.087)			(.076)	(.087)	(.076)	(.074)	(.071)
age_0^2	.031	.011			.007	.012	.006	.013	.009
	(.014)	(.005)	same		(.004)	(.005)	(.004)	(.004)	(.003)
$\text{inc}_0.\text{age}_0$	.044	.026			.029	.027	.028	.031	.029
	(.013)	(.006)	as		(.005)	(.006)	(.005)	(.005)	(.005)
e401k	6.346	6.760			6.451	6.641	5.174	7.477	4.362
	(2.022)	(1.842)	WLS		(1.442)	(1.806)	(1.518)	(1.510)	(1.124)
male	1.799	1.505			1.511	1.517	1.579	1.662	1.486
	(1.959)	(.757)			(.537)	(.753)	(.523)	(.719)	(.504)
e401k.inc_0	.307	.258			.232	.265	.226	.317	.204
	(.216)	(.128)			(.101)	(.125)	(.087)	(.107)	(.090)
e401k.age_0	.154	.160			.118	.159	.228	.162	.190
	(.262)	(.120)			(.105)	(.118)	(.102)	(.112)	(.100)

Table 11: 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]. Standard errors of the proposed estimators are highlighted with blue color.

A Appendix A: Proofs

Proof of Lemma 1: (1) and assumption A5 imply that $\widehat{\beta}_{OLS} \xrightarrow{p} \beta^0$.

(a) Using this and assumptions A4 and A5 we obtain that $\widehat{\sigma}^2_{cat1}(\widehat{\beta}_{OLS}, \gamma) - \sigma^2_{cat1}(\gamma) \xrightarrow{p} 0$, $\widehat{\sigma}^2_{cat2}(\widehat{\beta}_{OLS}, \widehat{\beta}_{OLS}, \gamma) - \sigma^2_{cat2}(\gamma) \xrightarrow{p} 0$ and $\widehat{\sigma}^2_{cat3}(\widehat{\beta}_{OLS}, \gamma) - \sigma^2_{cat3}(\gamma) \xrightarrow{p} 0$ uniformly in $\gamma \in \Gamma$. We show the proof for Category 2; the proof for the other two categories follows in the same way.

Take any $\delta > 0$ and note that assumption A2 implies that $P(\|\widehat{\gamma}_{cat2} - \gamma^*_{cat2}\| > \delta) \leq P\left(|\sigma^2_{cat2}(\widehat{\gamma}_{cat2}) - \sigma^2_{cat2}(\gamma^*_{cat2})| \geq \epsilon(\delta)\right)$ for some $\epsilon(\delta) > 0$. As usual, we will prove the result by showing as follows that the probability on the righthand side goes to zero as $n \to \infty$:

$$\begin{aligned}
0 &\leq \sigma^2_{cat2}(\widehat{\gamma}_{cat2}) - \sigma^2_{cat2}(\gamma^*_{cat2}) \\
&= \sigma^2_{cat2}(\widehat{\gamma}_{cat2}) - \widehat{\sigma}^2_{cat2}(\widehat{\beta}_{OLS}, \widehat{\beta}_{OLS}, \widehat{\gamma}_{cat2}) + \widehat{\sigma}^2_{cat2}(\widehat{\beta}_{OLS}, \widehat{\beta}_{OLS}, \widehat{\gamma}_{cat2}) - \sigma^2_{cat2}(\gamma^*_{cat2}) \\
&\leq \sigma^2_{cat2}(\widehat{\gamma}_{cat2}) - \widehat{\sigma}^2_{cat2}(\widehat{\beta}_{OLS}, \widehat{\beta}_{OLS}, \widehat{\gamma}_{cat2}) + \widehat{\sigma}^2_{cat2}(\widehat{\beta}_{OLS}, \widehat{\beta}_{OLS}, \gamma^*_{cat2}) - \sigma^2_{cat2}(\gamma^*_{cat2})
\end{aligned}$$

where the first line follows by the definition of γ^*_{cat2}, the second line is simply adding and subtracting the same thing, and the third line follows by the definition of $\widehat{\gamma}_{cat2}$. Therefore,

$$P\left(|\sigma^2_{cat2}(\widehat{\gamma}_{cat2}) - \sigma^2_{cat2}(\gamma^*_{cat2})| \geq \epsilon(\delta)\right) \leq P\left(\sup_{\gamma \in \Gamma} |\sigma^2_{cat2}(\gamma) - \widehat{\sigma}^2_{cat2}(\widehat{\beta}_{OLS}, \widehat{\beta}_{OLS}, \gamma)| \geq \frac{\epsilon(\delta)}{2}\right) \to 0$$

using that $\widehat{\sigma}^2_{cat2}(\widehat{\beta}_{OLS}, \widehat{\beta}_{OLS}, \gamma) - \sigma^2_{cat2}(\gamma) \xrightarrow{p} 0$ uniformly in $\gamma \in \Gamma$.

(b) As in (a), we can use $\widehat{\beta}_{OLS} \xrightarrow{p} \beta^0$, and assumptions A4 and A6 to obtain that $\widehat{\sigma}^2_{cat1}(\widehat{\beta}_{OLS}, \gamma) - \sigma^2_{cat1}(\gamma) = O_p(n^{-1/2})$, $\widehat{\sigma}^2_{cat2}(\widehat{\beta}_{OLS}, \widehat{\beta}_{OLS}, \gamma) - \sigma^2_{cat2}(\gamma) = O_p(n^{-1/2})$ and $\widehat{\sigma}^2_{cat3}(\widehat{\beta}_{OLS}, \gamma) - \sigma^2_{cat3}(\gamma) = O_p(n^{-1/2})$ uniformly in $\{\gamma \in \Gamma : \|\gamma - \gamma^*_j\| \leq \delta_n\}$ for any $\delta_n \downarrow 0$ and where $j = cat1, cat2, cat3$.

The result in (a) implies that for each $j = cat1, cat2, cat3$ we have $P(\|\widehat{\gamma}^*_j - \gamma^*_j\| \leq \delta_n) \to 1$ for any $\delta_n \downarrow 0$ as $n \to \infty$. So, as in (a), but now conditioning on the event $\{\|\widehat{\gamma}^*_j - \gamma^*_j\| \leq \delta_n\}$, we can obtain that:

$$\begin{aligned}
0 &\leq \sigma^2_{cat2}(\widehat{\gamma}_{cat2}) - \sigma^2_{cat2}(\gamma^*_{cat2}) \\
&\leq \sigma^2_{cat2}(\widehat{\gamma}_{cat2}) - \widehat{\sigma}^2_{cat2}(\widehat{\beta}_{OLS}, \widehat{\beta}_{OLS}, \widehat{\gamma}_{cat2}) + \widehat{\sigma}^2_{cat2}(\widehat{\beta}_{OLS}, \widehat{\beta}_{OLS}, \gamma^*_{cat2}) - \sigma^2_{cat2}(\gamma^*_{cat2}) \\
&\leq 2 \sup_{\gamma \in \Gamma : \|\gamma - \gamma^*_j\| \leq \delta_n} |\sigma^2_{cat2}(\gamma) - \widehat{\sigma}^2_{cat2}(\widehat{\beta}_{OLS}, \widehat{\beta}_{OLS}, \gamma)\| = O_p(n^{-1/2})
\end{aligned}$$

by the local uniform convergence established above. Therefore, $|\sigma^2_{cat2}(\widehat{\gamma}_{cat2}) - \sigma^2_{cat2}(\gamma^*_{cat2})| = O_p(n^{-1/2})$. Hence, assumption A3 now gives: $\|\widehat{\gamma}_{cat2} - \gamma^*_{cat2}\| \leq |\sigma^2_{cat2}(\widehat{\gamma}_{cat2}) - \sigma^2_{cat2}(\gamma^*_{cat2})|/M = O_p(n^{-1/2})$. Proofs for Categories 1 and 3 follow similarly. ■

Proof of Theorem 1: (a) The proof is very standard, so we simply provide the two key steps here. For any $\widehat{\gamma}_j \xrightarrow{p} \gamma_j^*$ for $j = cat1, cat2, cat3$:

$$\begin{aligned}
\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{x_i u_i}{\omega^2(x_i;\widehat{\gamma}_j)} &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{x_i u_i}{\omega^2(x_i;\gamma_j^*)} + E\left[xu\Delta_{1,j}(x)\right]\sqrt{n}(\widehat{\gamma}_j-\gamma_j^*) + R_{1,n} + R_{2,n} \\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{x_i u_i}{\omega^2(x_i;\gamma_j^*)} + o_p(1) \qquad (13)
\end{aligned}$$

since $E\left[xu\Delta_{1,j}(x)\right]=0$ by (1); $R_{1,n} := \left[\frac{1}{n}\sum_{i=1}^{n}\left(x_i u_i \Delta_{1,j}(x_i) - E\left[xu\Delta_{1,j}(x)\right]\right)\right]\sqrt{n}(\widehat{\gamma}_j-\gamma_j^*) = o_p(1)$ by the weak law of large numbers because $E\left[xu\Delta_{1,j}(x)\right]=0$; and:

$$\begin{aligned}
|R_{2,n}| &:= \left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}x_i u_i\left[\frac{1}{\omega^2(x_i;\widehat{\gamma}_j)} - \frac{1}{\omega^2(x_i;\gamma_j^*)} - \Delta_{1,j}(x_i)(\widehat{\gamma}_j-\gamma_j^*)\right]\right| \\
&\leq \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\|x_i u_i\| \times \left|\frac{1}{\omega^2(x_i;\widehat{\gamma}_j)} - \frac{1}{\omega^2(x_i;\gamma_j^*)} - \Delta_{1,j}(x_i)(\widehat{\gamma}_j-\gamma_j^*)\right| \\
&\leq \frac{1}{2\sqrt{n}}\sum_{i=1}^{n}\|x_i u_i\| \times |\Delta_{2,j}| \times \|\widehat{\gamma}_j-\gamma_j^2\|^2 \\
&\leq \left(\frac{1}{2n}\sum_{i=1}^{n}\|x_i u_i \Delta_{2,j}\|\right)\left(n^{1/4}\|\widehat{\gamma}_j-\gamma_j^*\|\right)^2 = o_p(1),
\end{aligned}$$

where the first inequality follows by the Cauchy-Schwartz inequality, the second and third inequalities by assumption A8, and the last equality follows by assumption A8 and Lemma 1(b).

(13) along with assumptions A4, A5 and A7 directly gives the results for Categories 1 and 3. The result for Category 2 follows once we additionally note that for any $b_1, b_2 \xrightarrow{p} \beta^0$ we have : (i) $\widehat{\lambda}(b_1,b_2,\widehat{\gamma}_{cat2}) \xrightarrow{p} \lambda(\gamma_{cat2})$ by assumption A6 and Lemma 1(a) (see also the expressions for $\widehat{\lambda}(b_1,b_2,\widehat{\gamma}_{cat2})$ and $\lambda(\gamma)$ in Section 2.2 and equation (8) respectively); and hence (ii)

$$\begin{aligned}
&\sqrt{n}\left[\widehat{\lambda}(b_1,b_2,\widehat{\gamma}_{cat2})\widehat{h}(\widehat{\gamma}_{cat2}) + (1-\widehat{\lambda}(b_1,b_2,\widehat{\gamma}_{cat2}))\widehat{h}_{OLS} - h^0\right] \\
=\ &\sqrt{n}\left[\lambda(\gamma_{cat2})\widehat{h}(\widehat{\gamma}_{cat2}) + (1-\lambda(\gamma_{cat2}))\widehat{h}_{OLS} - h^0\right] \\
&+\left(\widehat{\lambda}(b_1,b_2,\widehat{\gamma}_{cat2}) - \lambda(\gamma_{cat2})\right)\left[\sqrt{n}(\widehat{h}(\widehat{\gamma}_{cat2}) - h^0) - \sqrt{n}(\widehat{h}_{OLS} - h^0)\right] \\
=\ &\sqrt{n}\left[\lambda(\gamma_{cat2})\widehat{h}(\widehat{\gamma}_{cat2}) + (1-\lambda(\gamma_{cat2}))\widehat{h}_{OLS} - h^0\right] + o_p(1)
\end{aligned}$$

where the first equality follows by adding and subtracting off the same terms, and the second equality by (i) and the joint asymptotic normality of $\sqrt{n}(\widehat{h}(\widehat{\gamma}_{cat2}) - h^0)$ and $\sqrt{n}(\widehat{h}_{OLS} - h^0)$.

(b) Follows by assumptions A4 and A6, Lemma 1(a) and Theorem 1(a), that jointly with Slutsky's lemma give the asymptotic normality of the test statistic in each of the three categories.

(c) Follows by Theorem 1 (a) and (b) and by definition. ■