SIMULTANEOUS MEAN-VARIANCE REGRESSION

RICHARD H. SPADY† AND SAMI STOULI§

Abstract. We propose simultaneous mean-variance regression for the linear estimation and approximation of conditional mean functions. In the presence of heteroskedasticity of unknown form, our method accounts for varying dispersion in the regression outcome across the support of conditioning variables by using weights that are jointly determined with the mean regression parameters. Simultaneity generates outcome predictions that are guaranteed to improve over ordinary least-squares prediction error, with corresponding parameter standard errors that are automatically valid. Under shape misspecification of the conditional mean and variance functions, we establish existence and uniqueness of the resulting approximations and characterize their formal interpretation and robustness properties. In particular, we show that the corresponding mean-variance regression location-scale model weakly dominates the ordinary least-squares location model under a Kullback-Leibler measure of divergence, with strict improvement in the presence of heteroskedasticity. The simultaneous mean-variance regression loss function is globally convex and the corresponding estimator is easy to implement. We establish its consistency and asymptotic normality under misspecification, provide robust inference methods, and present numerical simulations that show large improvements over ordinary and weighted least-squares in terms of estimation and inference in finite samples. We further illustrate our method with two empirical applications to the estimation of the relationship between economic prosperity in 1500 and today, and demand for gasoline in the United States.

Keywords: Conditional mean and variance functions, simultaneous approximation, heteroskedasticity, robust inference, misspecification, convexity, influence function, ordinary least-squares, linear regression, dual regression.

Date: November 18, 2019. We are grateful to Frank Diebold, James MacKinnon, Alex Tetenov, Daniel Wilhelm, participants to the Bristol Econometric Study Group 2018 and the 13th GNYMA Econometrics Colloquium, Princeton, for useful comments and discussions, and to Richard Blundell, Joel Horowitz and Matthias Parey for sharing the data used in the demand for gasoline empirical illustration.
† Nuffield College, Oxford, and Department of Economics, Johns Hopkins University, rspady@jhu.edu.
§ Department of Economics, University of Bristol, s.stouli@bristol.ac.uk.

1. Introduction

Ordinary least-squares (OLS) is the method of choice for the linear estimation and approximation of the conditional mean function (CMF). However, in the presence of heteroskedasticity the standard errors of OLS are inconsistent, and subsequent inference is therefore unreliable. As a way of achieving valid inference, practitioners instead often use the heteroskedasticity-corrected standard errors of Eicker (1963, 1967), Huber (1967) and White (1980a). Although valid asymptotically, numerous limitations of this approach have been highlighted in the literature such as bias and sensitivity to outliers, incorrect size and low power of robust tests in finite samples (MacKinnon and White, 1985; Chesher and Jewitt, 1987; Chesher, 1989; Chesher and Austin, 1991). These findings in turn generated a large number of proposals in order to reconcile the large-sample validity of the approach and its observed finite-sample limitations, surveyed in MacKinnon (2013).

The finite-sample limitations of OLS-based inference essentially originate from the fact that OLS assigns a constant weight to each observation in fitting the best linear predictor for the regression outcome. Hence the least-squares criterion does not account for the varying accuracy of the information available about the outcome across the covariate space. This yields point estimates and linear approximations that are sensitive to high-leverage points and outliers, which in turn generate biased estimates of the residuals' second moments used in the calculation of the robust variance-covariance matrix of OLS parameters. In finite samples, uniform weighting not only compromises the validity of OLS-based statistical inference in the presence of heteroskedasticity, but also the reliability of OLS point estimates.

In this paper, we propose simultaneous mean-variance regression (MVR) as an alternative to OLS for the linear estimation and approximation of CMFs. MVR characterizes the conditional mean and variance functions jointly, thereby providing a solution to the problems of estimation, approximation and inference in the presence of heteroskedasticity of unknown form with five main features. First, it incorporates information from the second conditional moment in the determination of the first conditional moment parameters. Second, simultaneity generates approximations with improved robustness properties relative to OLS. Third, the resulting approximations have a formal interpretation under shape misspecification of the conditional mean and variance functions. Fourth, MVR solutions are well-defined, i.e., exist and are unique. Fifth, standard errors of the corresponding estimator are automatically valid in the presence of heteroskedasticity of unknown form, and reduce to those of OLS under homoskedasticity.

The MVR criterion can be interpreted as a penalized weighted least-squares (WLS) loss function. The presence and the form of the penalty ensure global convexity of the objective function, so that MVR conditional mean and variance approximations are jointly well-defined. This differs from the usual WLS approach where a sequential procedure is followed, obtaining the weights first, and then implementing a weighted regression to determine the parameters of the linear specification. Our simultaneous approach allows us to give theoretical guarantees on the relative approximation properties of MVR and OLS. We use MVR to construct and estimate a new class of approximations of the conditional mean and variance functions, with improved robustness and precision in finite samples. We establish the interpretation of MVR approximations, we derive the asymptotic properties of the corresponding MVR estimator, and we give tools for robust inference. We also illustrate the practical benefits of the MVR estimator with extensive numerical simulations, and find very large finite-sample improvements over both OLS and WLS in terms of estimation performance and heteroskedasticity-robust inference.

This paper generalizes the results of Spady and Stouli (2018) for the primal problem of the dual regression estimator of linear location-scale models. We provide a unified theory allowing for a large class of scale functions. This paper is also related to the interpretation of OLS under misspecification of the shape of the CMF. OLS gives the minimum mean squared error linear approximation to the CMF, an important motivation for its use in empirical work (White, 1980b; Chamberlain, 1984; Angrist and Krueger, 1999; Angrist and Pischke, 2008). MVR introduces a class of WLS approximations accounting for potential variation in the outcome across the support of conditioning variables, and with weights that have a clear interpretation under misspecification. Our approach thus complements the textbook WLS proposal of Cameron and Trivedi (2005) and Wooldridge (2010, 2012) (see also Romano and Wolf, 2017), who advocate the reweighting of OLS with generalized least-squares (GLS) weights and further correcting the standard errors for heteroskedasticity.

This paper makes three main contributions. First, we show existence and uniqueness of MVR solutions under general misspecification, thereby introducing a new class of location-scale models corresponding to MVR approximations. The results in Spady and Stouli (2018) did not cover the case of misspecified conditional mean and variance functions. Second, we establish favorable approximation and robustness properties of

MVR relative to OLS. We show that MVR is a minimum WLS linear approximation to the CMF, with weights determined such that the MVR approximation improves over OLS in the presence of heteroskedasticity under the MVR loss. For our main specifications of the scale function, we further show that OLS root mean squared prediction error is an upper bound for the MVR weighted mean squared prediction error. We then extend this result to show that under a Kullback-Leibler information criterion (KLIC) the proposed MVR location-scale models weakly dominate the OLS location model, with strict improvement in the presence of heteroskedasticity. These results provide theoretical guarantees motivating the use of MVR over OLS, and are not shared by alternative WLS proposals. Third, we derive the asymptotic distribution of the MVR estimator under misspecification and provide robust inference methods. In particular we propose a robust one-step heteroskedasticity test that complements existing OLS-based tests (e.g., Breusch and Pagan, 1979; White, 1980a; Koenker, 1981).

The rest of the paper is organized as follows. Section 2 introduces MVR under correct specification of conditional mean and variance functions. Section 3 establishes the main approximation properties of MVR under misspecification, including existence and uniqueness. Section 4 gives asymptotic theory. Section 5 reports the results of an empirical application to the relationship between economic prosperity in 1500 and today, and illustrates the finite-sample performance of MVR with numerical simulations. All proofs of the main results are given in the Appendix. The online Appendix Spady and Stouli (2018) contains supplemental material, including an additional empirical application to demand for gasoline in the United States.

2. Simultaneous Mean-Variance Regression

2.1. **The Mean-Variance Regression Problem.** Given a scalar random variable Y and a random $k \times 1$ vector X that includes an intercept, i.e., has first component 1, denote the mean and standard deviation functions of Y conditional on X by $\mu(X) := E[Y \mid X]$ and $\sigma(X) := E[(Y - E[Y \mid X])^2 \mid X]^{1/2}$, respectively. We start with a simplified setting where the conditional mean and variance functions take the parametric forms

$$\mu(X) = X'\beta_0, \qquad \sigma(X)^2 = s(X'\gamma_0)^2, \tag{2.1}$$

for some positive scale function $t \mapsto s(t)$, and where the parameters β_0 and γ_0 belong to the parameter space $\Theta = \mathbb{R}^k \times \Theta_\gamma$, with $\Theta_\gamma = \{\gamma \in \mathbb{R}^k : \Pr[s(X'\gamma) > 0] = 1\}$. Two leading examples for the scale function are the linear and exponential specifications $s(t) = t$ and $s(t) = \exp(t)$, with domains $(0, \infty)$ and $\mathbb{R}$, respectively.

The parameter vector $\theta_0 := (\beta_0, \gamma_0)'$ is uniquely determined as the solution to the globally convex MVR population problem

$$\min_{\theta\in\Theta} E\left[\frac{1}{2}\left\{\left(\frac{Y - X'\beta}{s(X'\gamma)}\right)^2 + 1\right\} s(X'\gamma)\right]. \tag{2.2}$$

When the functions $x \mapsto \mu(x)$ and $x \mapsto \sigma(x)^2$ satisfy model (2.1), they are simultaneously characterized by problem (2.2). As a consequence, MVR incorporates information on the dispersion of Y across the support of X in the determination of the mean parameter β. We show below that problem (2.2) is formally equivalent to an infeasible sequential least-squares estimator of the conditional mean and variance functions for model (2.1). Problem (2.2) is a generalization of the dual regression primal problem introduced in Spady and Stouli (2018), for which the scale function is linear. Considering scale functions with domain the real line, such as the exponential function, allows the transformation of the dual regression primal problem into an unconstrained convex problem over $\Theta = \mathbb{R}^{2\times k}$.

Inspection of the first-order conditions confirms that θ_0 is indeed a valid solution to problem (2.2). Denoting the derivative of the scale function by $s_1(t) := \partial s(t)/\partial t$ and letting $e(Y, X, \theta) := (Y - X'\beta)/s(X'\gamma)$, the first-order conditions of (2.2) are

$$E\left[Xe\left(Y, X, \theta\right)\right] = 0 \tag{2.3}$$

$$E\left[Xs_1(X'\gamma)\{e\left(Y, X, \theta\right)^2 - 1\}\right] = 0. \tag{2.4}$$

These conditions are satisfied by θ_0 since specification (2.1) is equivalent to the location-scale model

$$Y = X'\beta_0 + s(X'\gamma_0)\varepsilon, \quad E[\varepsilon \mid X] = 0, \quad E[\varepsilon^2 \mid X] = 1. \tag{2.5}$$

Therefore, the parameter vector θ_0 also satisfies the relations

$$E\left[e(Y, X, \theta_0) \mid X\right] = E\left[\varepsilon \mid X\right] = 0$$
$$E\left[e(Y, X, \theta_0)^2 - 1 \mid X\right] = E\left[\varepsilon^2 - 1 \mid X\right] = 0,$$

which imply that $E[h(X)e(Y, X, \theta_0)] = 0$ and $E[h(X)\{e(Y, X, \theta_0)^2 - 1\}] = 0$ hold for any measurable function $x \mapsto h(x)$, and in particular for $h(X) = X$ and $h(X) = Xs_1(X'\gamma)$.

2.2. **Formal Framework.** Let $\mathcal{X}$ denote the support of X, and for a vector $u = (u_1, \ldots, u_k)' \in \mathbb{R}^k$, let $||\cdot||$ denote the Euclidean norm, i.e., $||u|| = (u_1^2 + \ldots + u_k^2)^{1/2}$; we define a compact subset $\Theta^c \subset \Theta$ as

$$\Theta^c := \left\{ \theta \in \Theta \, : \, ||\theta|| \leq C_\theta \text{ and } \inf_{x \in \mathcal{X}} s(x'\gamma) \geq C_s \right\},$$

for some finite constant C_θ and some constant $C_s > 0$, with interior set denoted $\text{int}(\Theta^c)$. The second and third derivatives of the scale function $t \mapsto s(t)$ are denoted by $s_j(t) := \partial^j s(t)/\partial t^j$, $j = 2, 3$. We also denote the MVR objective function in (2.2) by $Q(\theta) := E[\{e\,(Y, X, \theta)^2 + 1\} s(X'\gamma)/2]$.

Our first assumption specifies the class of scale functions we consider.

Assumption 1. For $a = 0$ or $-\infty$, the scale function $s : (a, \infty) \to (0, \infty)$ is a three times differentiable strictly increasing convex function that satisfies $\lim_{t \to a} s(t) = 0$ and $\lim_{t \to \infty} s(t) = \infty$.

Assumption 1 encompasses several types of scale functions such as polynomial specifications $s(x'\gamma) = (x'\gamma)^\alpha$ with $a = 0$ and $\Pr[X'\gamma > 0] = 1$, or exponential-polynomial specifications $s(x'\gamma) = \exp(x'\gamma)^\alpha$ with $a = -\infty$, for some $\alpha > 0$. For $\alpha = 1$, we recover the linear and exponential scale leading cases.

The next assumptions complete our formal framework.

Assumption 2. The conditional variance function $x \mapsto \sigma(x)^2$ is bounded away from 0 uniformly in $\mathcal{X}$.

Assumption 3. We have (i) $E[Y^4] < \infty$ and $E||X||^4 < \infty$, and, (ii) for all $\gamma \in \Theta_\gamma$, $E[\|X\|^4 \, s_2(X'\gamma)^2] < \infty$, $E[\|X\|^6 \, s_3(X'\gamma)^2] < \infty$ and $E[\|X\|^6 \, s_1(X'\gamma)^2 s_2(X'\gamma)^2] < \infty$.

Assumption 4. For all $\gamma \in \Theta_\gamma$, $E[XX'/s(X'\gamma)]$ is nonsingular.

Assumptions 1-4 are sufficient conditions for global convexity of the MVR criterion over the parameter space Θ, and therefore for problem (2.2) to have a unique solution.

Theorem 1. *If Assumptions 1-4 hold, and the conditional mean and variance functions of Y given X satisfy model (2.1) a.s. with $\theta_0 \in \text{int}(\Theta^c)$, then θ_0 is the unique minimizer of $Q(\theta)$ over Θ.*

Theorem 1 applies when the conditional mean and variance functions are well-specified, and thus provides primitive conditions for identification of θ_0 in the location-scale model (2.5). This extends the uniqueness result in Spady and Stouli (2018) for objective (2.2) with a linear scale function to the class of scale functions defined in Assumption 1.

Remark 1. In the linear scale case, $s_1(t) = 1$ and $s_j(t) = 0$, $j = 2, 3$, so that Assumption 3 reduces to Assumption 3(i). In the exponential scale case, $s_j(t) = \exp(t)$, $j = 1, 2, 3$, so that Assumption 3(ii) reduces to the requirement that $E[\|X\|^6 \exp(X'\gamma)^4]$ be finite. This is satisfied for instance if X is bounded. □

2.3. **Simultaneous Mean-Variance Regression Interpretation.** Problem (2.2) is equivalent to an infeasible sequential least-squares estimator of conditional mean and variance functions. The first-order conditions of (2.2) can also be written as

$$E\left[\frac{X}{s(X'\gamma)}(Y - X'\beta)\right] = 0 \tag{2.6}$$

$$E\left[X\frac{s_1(X'\gamma)}{s(X'\gamma)^2}\left\{(Y - X'\beta)^2 - s(X'\gamma)^2\right\}\right] = 0. \tag{2.7}$$

Given knowledge of γ_0, WLS regression of Y on X with weights $1/s(X'\gamma_0)$ has first-order conditions (2.6), with solution β_0. Moreover, given knowledge of β_0, nonlinear WLS regression of $(Y - X'\beta_0)^2$ on X with weights $1/s(X'\gamma_0)^3$ and quadratic link function has first-order conditions (2.7), and therefore solution γ_0.

Proposition 1. *If Assumptions 1-4 hold, and (i) $E[Y^4] < \infty$, $E[\|X\|^4] < \infty$ and $E[s(X'\gamma)^4] < \infty$ for all $\gamma \in \Theta_\gamma$, and (ii) the conditional mean and variance functions of Y given X satisfy model (2.1) a.s., then the MVR population problem (2.2) is equivalent to the infeasible sequential estimator with first step*

$$\beta_0 = \arg\min_{\beta\in\Theta_\beta} E\left[\frac{1}{\sigma(X)}(Y - X'\beta)^2\right], \tag{2.8}$$

and second step

$$\gamma_0 = \arg\min_{\gamma\in\Theta_\gamma} E\left[\frac{1}{\sigma(X)^3}\left\{(Y - X'\beta_0)^2 - s(X'\gamma)^2\right\}^2\right]. \tag{2.9}$$

An immediate implication of the Law of Iterated Expectations and Proposition 1 is that MVR implements simultaneous weighted linear regression of $\mu(X)$ on X and weighted nonlinear regression of $\sigma(X)^2$ on X by solving for β and γ such that the weighted residuals $(\mu(X) - X'\beta)/s(X'\gamma)$ and $\{\sigma(X)^2 - s(X'\gamma)^2\}/s(X'\gamma)^2$ are simultaneously orthogonal to X and $Xs_1(X'\gamma)$, respectively. Proposition 1 thus establishes the simultaneous mean and variance regression interpretation of problem (2.2).

3. Approximation Properties of MVR under Misspecification

Under misspecification, OLS provides the minimum mean squared error linear approximation to the CMF. For the proposed MVR criterion, existence of an approximating solution and the nature of the approximation are nontrivial when the shapes of the conditional mean and variance functions are misspecified. In this section, we first establish existence and uniqueness of a solution to the MVR problem under misspecification, and then characterize the interpretation and properties of the corresponding MVR approximations.

3.1. Existence and Uniqueness of an MVR Solution.

Assumptions 1-4 are sufficient for characterizing the smoothness properties, shape, and behaviour on the boundaries of the parameter space of the MVR criterion $Q(\theta)$. Under these assumptions $\theta \mapsto Q(\theta)$ is continuous and its level sets are compact. Compactness of the level sets is a sufficient condition for existence of a minimizer in Θ, and is a consequence of the explosive behaviour of the objective function at the boundaries of the parameter space. The objective $Q(\theta)$ is a *coercive* function over the open set Θ, i.e., it satisfies

$$\lim_{||\theta|| \to \infty} Q(\theta) = \infty, \qquad \lim_{\theta \to \partial\Theta} Q(\theta) = \infty,$$

where $\partial\Theta$ is the boundary set of Θ. Thus the MVR criterion is infinity at infinity, and for any sequence of parameter values in Θ approaching the boundary set $\partial\Theta$, the value of the objective is also driven towards infinity. Therefore, the level sets of the objective function have no limit point on their boundary, ruling out existence of a boundary solution, and continuity of $\theta \mapsto Q(\theta)$ is then sufficient to conclude that it admits a minimizer. Continuity and coercivity of the objective function are the two properties that guarantee existence of at least one minimizer in Θ.[1] Assumptions 1-4 are also sufficient for $\theta \mapsto Q(\theta)$ to be strictly convex, and therefore further ensure that $Q(\theta)$ admits at most one minimizer in Θ.

Theorem 2. *If Assumptions 1-4 hold, then there exists a unique solution $\theta^* \in \Theta$ to the MVR population problem (2.2).*

Theorem 2 is the second main result of the paper. It establishes that the MVR problem (2.2) has a well-defined solution, and an immediate corollary is the existence

[1] The boundary set of Θ may be empty, for instance for the exponential scale specification. In that case the coercivity property reduces to $\lim_{||\theta|| \to \infty} Q(\theta) = \infty$.

and uniqueness of the MVR location-scale representation

$$Y = X'\beta^* + s(X'\gamma^*)e, \quad E[Xe] = 0, \quad E[Xs_1(X'\gamma^*)(e^2 - 1)] = 0.$$

This result clarifies further how MVR generalizes OLS by establishing the existence and the form of the MVR location-scale model when no shape restrictions are imposed on the conditional mean and variance functions. The OLS location model is a particular case with the scale function restricted to be a constant function.

Although a unique MVR approximation exists irrespective of the nature of the misspecification, the interpretation of the MVR approximating functions $x \mapsto (x'\beta^*, s(x'\gamma^*)^2)$ depends on which of the conditional moment functions is misspecified. We distinguish two types of shape misspecification:

(1) Mean misspecification: the CMF $x \mapsto \mu(x)$ is misspecified.
(2) Variance misspecification: only the conditional variance function $x \mapsto \sigma(x)^2$ is misspecified.

The case when both the conditional mean and variance functions are misspecified is a particular case of mean misspecification.

3.2. **Interpretation Under Mean Misspecification.** The location-scale representation

$$Y = \mu(X) + \sigma(X)\varepsilon, \quad E[\varepsilon \mid X] = 0, \quad E[\varepsilon^2 \mid X] = 1, \tag{3.1}$$

provides a general expression for Y in terms of its conditional mean and standard deviation functions, and is always valid, as long as first and second conditional moments exist. Substituting expression (3.1) for Y into the MVR objective function $Q(\theta)$ gives rise to a criterion for the joint approximation of $x \mapsto (\mu(x), \sigma(x)^2)$.

The criterion $Q(\theta)$ can also be appropriately restricted in order to define the corresponding OLS approximations. Letting $\Theta_{\gamma,\text{LS}} = \{\gamma \in \mathbb{R} : s(\gamma) > 0\}$, define $\Theta_{\text{LS}} = \mathbb{R}^k \times \Theta_{\gamma,\text{LS}}$. Upon setting $s(X'\gamma) = s(\gamma)$ in the MVR problem (2.2),

$$(\beta_{\text{LS}}, \gamma_{\text{LS}}) := \arg \min_{(\beta,\gamma)\in\Theta_{\text{LS}}} E\left[\frac{1}{2}\left\{\left(\frac{Y - X'\beta}{s(\gamma)}\right)^2 + 1\right\} s(\gamma)\right]$$

is a particular case of MVR. Since the OLS solution $\theta_{\text{LS}} := (\beta_{\text{LS}}, \gamma_{\text{LS}}, 0_{k-1})'$ belongs to the parameter space Θ, uniqueness of θ^* implies that the OLS approximation of the conditional moment functions $x \mapsto (\mu(x), \sigma(x)^2)$ cannot improve upon the MVR approximation, according to the MVR loss.

Theorem 3. *If Assumptions 1-4 hold, then the MVR population problem (2.2) has the following properties.*

(i) Problem (2.2) is equivalent to the infeasible problem

$$\min_{\theta \in \Theta} \frac{1}{2} E\left[\left\{\left(\frac{\mu(X) - X'\beta}{s(X'\gamma)}\right)^2 + 1\right\} s(X'\gamma)\right] + \frac{1}{2} E\left[\frac{\sigma(X)^2}{s(X'\gamma)}\right], \tag{3.2}$$

with first-order conditions

$$E\left[X\left(\frac{\mu(X) - X'\beta}{s(X'\gamma)}\right)\right] = 0 \tag{3.3}$$

$$E\left[X s_1(X'\gamma)\left\{\left(\frac{\mu(X) - X'\beta}{s(X'\gamma)}\right)^2 + \left(\frac{\sigma(X)^2}{s(X'\gamma)^2} - 1\right)\right\}\right] = 0. \tag{3.4}$$

(ii) The optimal value of problem (2.2) satisfies $Q(\theta^*) \leq Q(\theta_{LS})$, with equality if and only if $\theta^* = \theta_{LS}$.

Theorem 3(i) shows that under misspecification the function $x \mapsto x'\beta^*$ is an infeasible MVR approximation of the true CMF penalized by the mean ratio of the true variance over its standard deviation approximation. An equivalent formulation is

$$\min_{\theta \in \Theta} \frac{1}{2} E\left[\frac{1}{s(X'\gamma)}(\mu(X) - X'\beta)^2\right] + \frac{1}{2} E\left[\left\{\frac{\sigma(X)^2}{s(X'\gamma)^2} + 1\right\} s(X'\gamma)\right], \tag{3.5}$$

the penalized WLS interpretation of the MVR problem (3.2).

The penalty term in (3.5) is a functional of a weighted mean variance ratio of the true variance over its approximation. The first-order conditions (3.3)-(3.4) shed additional light on how the weights are determined as well as on the form of the penalty, by characterizing the optimality properties of MVR approximations. Because X includes an intercept, when both functions $x \mapsto \mu(x)$ and $x \mapsto \sigma(x)^2$ are misspecified, β^* and γ^* are chosen such that the sum of the weighted mean squared error for the conditional mean and the mean variance ratio error is zero, balancing the two approximation errors. When the scale function is linear the two types of approximation error are equalized. For the exponential specification, the two types of approximation error weighted by $\exp(X'\gamma)$ are equalized. The MVR solution is thus determined by minimizing the weighted mean squared error for the conditional mean, while simultaneously setting the weighted mean variance ratio as close as possible to one.

Theorem 3(ii) formalizes the approximation guarantee of MVR in terms of the MVR criterion. For the linear and exponential scale function specifications, the improvement of the MVR solution relative to the OLS solution in MVR loss further guarantees that optimal weights are selected such that the weighted mean squared MVR prediction error for Y is not larger than the root mean squared OLS prediction error.

Corollary 1. *If the scale function $t \mapsto s(t)$ is specified as $s(t) = t$ or $s(t) = \exp(t)$, then*

$$E\left[\frac{1}{s(X'\gamma^*)}(Y - X'\beta^*)^2\right] \leq E\left[(Y - X'\beta_{LS})^2\right]^{\frac{1}{2}},$$

with equality if and only if $\theta^* = \theta_{LS}$.

Compared to OLS, improvement in MVR loss is also related to a key robustness property of MVR under a Kullback-Leibler measure of divergence. Define the scaled Gaussian density function

$$f_\theta(Y, X) := \frac{1}{s(X'\gamma)}\phi\left(\frac{Y - X'\beta}{s(X'\gamma)}\right), \tag{3.6}$$

where $\phi(z) = (2\pi)^{-1/2}\exp(-z^2/2)$. The OLS solution maximizes a restricted version, $E[\log f_\theta(Y, X)]|_{\gamma_{-1}=0}$, of the expected log-likelihood $E[\log f_\theta(Y, X)]$ over Θ_{LS}, where the components of γ except the first are set to zero. The corresponding expected log-likelihood value $E[\log f_{\theta_{\text{LS}}}(Y, X)]$ is no greater than the value of the expected log-likelihood at the MVR solution:

$$E\left[\log f_{\theta_{\text{LS}}}(Y, X)\right] \leq E\left[\log f_{\theta^*}(Y, X)\right]. \tag{3.7}$$

The MVR solution θ^* formally corresponds to an improvement of the expected log-likelihood value over OLS, and therefore corresponds to a probability distribution that is KLIC closer to the true data probability distribution $f_{Y|X}(Y \mid X)$.

Define the quantity $\epsilon := E\left[\log\left(\sigma^2_{\text{LS}}/s(X'\gamma^*)^2\right)\right]$, which is positive as shown in Appendix B.6. Our next result summarises the key implication of (3.7).

Theorem 4. *Suppose that $E[|\log f_{Y|X}(Y \mid X)|] < \infty$, $E\left[e(Y, X, \theta^*)^2\right] \leq 1+\epsilon$, and the scale function $t \mapsto s(t)$ is specified as $s(t) = t$ or $s(t) = \exp(t)$. Then the probability distribution f_{θ^*} corresponding to θ^* satisfies*

$$E\left[\log\left(\frac{f_{Y|X}(Y \mid X)}{f_{\theta^*}(Y, X)}\right)\right] \leq E\left[\log\left(\frac{f_{Y|X}(Y \mid X)}{f_{\theta_{LS}}(Y, X)}\right)\right], \tag{3.8}$$

with equality if and only if $\theta^* = \theta_{LS}$.

For the linear scale specification, the MVR first-order conditions include the constraint $E\left[e(Y,X,\theta^*)^2-1\right]=0$ so that the bound on $E\left[e(Y,X,\theta^*)^2\right]$ is satisfied by construction. For the exponential specification, the corresponding constraint is $E\left[\exp(X'\gamma^*)\left\{e(Y,X,\theta^*)^2-1\right\}\right]=0$ which can result in a value for $E\left[e(Y,X,\theta^*)^2\right]$ that differs from one under misspecification. The bound then characterizes the deviations from unit variance that preserve the validity of (3.8).[2]

The approximation guarantee (3.8) is a general result that holds under misspecification of the conditional mean and/or variance functions. When the mean is misspecified, it formally establishes that the MVR approximation for the mean corresponds to a better model than the OLS location model according to the classical KLIC for model selection (e.g., Akaike, 1973; Sawa, 1978). Similarly to the classical argument motivating the use of maximum likelihood (ML) under misspecification, Theorem 4 thus provides an information-theoretic justification for the use of MVR and a formal characterization of the robustness to misspecification of MVR relative to OLS.

3.3. **Interpretation Under Variance Misspecification.** If the CMF is linear, Theorem 3 has important additional implications for the robustness and optimality properties of MVR solutions. The k orthogonality conditions (3.4) are then sufficient to determine the scale parameter γ^* since condition (3.3) is uniquely satisfied by $\beta=\beta_0$. Thus in the classical particular case of the linear conditional mean model, the MVR solution for β is fully robust to misspecification of the scale function. Consequently, when the CMF is correctly specified the OLS and MVR solutions for β coincide. In the special case of linear scale specification, $Xs_1(X'\gamma)$ reduces to X. Because X includes an intercept, the scale parameter γ^* is then chosen such that the MVR conditional variance approximation also satisfies the remarkable property of zero mean variance ratio error.

Corollary 2. *If Assumptions 1-4 hold and $\mu(X)=X'\beta_0$ a.s., then $\beta^*=\beta_0$ and γ^* is solely determined by the k orthogonality conditions*

$$E\left[Xs_1(X'\gamma)\left\{\frac{\sigma(X)^2}{s(X'\gamma)^2}-1\right\}\right]=0. \tag{3.9}$$

In particular, for the linear specification $s(t)=t$, the conditional variance approximating function $x\mapsto(x'\gamma^*)^2$ satisfies the optimality property $E[\{\sigma(X)^2/(X'\gamma^*)^2\}-1]=0$.

[2]For the numerical simulations in Section 5.2 the typical sample mean of an estimate $e(Y,X,\hat{\theta})^2$ we observe is smaller than or equal to one, when the scale function is specified as $s(t)=\exp(t)$. It is an open question whether the bound $E\left[e(Y,X,\theta^*)^2\right]\leq 1+\epsilon$ can be binding.

When the CMF is correctly specified an optimal characterization of β_0 that will lead to an efficient estimator can be formulated by GLS. Define

$$f_\beta^\dagger(Y,X) := \frac{1}{\sigma(X)}\phi\left(\frac{Y - X'\beta}{\sigma(X)}\right).$$

In the population, GLS maximizes the expected log-likelihood $E[\log f_\beta^\dagger(Y,X)]$ with respect to β, with solution β_0. Then we further have

$$E\left[\log f_{\theta^*}(Y,X)\right] \leq E\left[\log f_{\beta_0}^\dagger(Y,X)\right], \tag{3.10}$$

and inequalities (3.7) and (3.10) together imply that, compared to OLS, the MVR solution θ^* formally corresponds to a probability distribution that is KLIC closer to the reference probability distribution $f_{\beta_0}^\dagger(Y,X)$ associated to the GLS model.

Theorem 5. *Suppose that $E[|\log f_{Y|X}(Y \mid X)|] < \infty$, $E\left[e(Y,X,\theta^*)^2\right] \leq 1+\epsilon$, and the scale function $t \mapsto s(t)$ is specified as $s(t) = t$ or $s(t) = \exp(t)$. If Assumptions 1-4 hold and $\mu(X) = X'\beta_0$ a.s., then f_{θ^*} also satisfies*

$$E\left[\log\left(\frac{f_{\beta_0}^\dagger(Y,X)}{f_{\theta^*}(Y,X)}\right)\right] \leq E\left[\log\left(\frac{f_{\beta_0}^\dagger(Y,X)}{f_{\theta_{LS}}(Y,X)}\right)\right], \tag{3.11}$$

with equality if and only if $\theta^* = \theta_{LS}$.

When the mean is correctly specified, all of the likelihood improvement comes from selecting a better approximation for the standard deviation function than OLS. Relative to the efficient GLS model for the mean, inequality (3.11) formally establishes that the OLS location model is rejected against the MVR location-scale model according to a likelihood ratio criterion (e.g., Vuong, 1989; Schennach and Wilhelm, 2017). If the true conditional variance is not constant then the improvement in (3.11) is strict and the MVR model is closer to the efficient GLS model than the OLS location model.

In the presence of heteroskedasticity, MVR optimality and approximation properties (3.9) and (3.11) under correct mean specification provide a theoretical justification for the largely improved MVR-based inference relative to OLS-based inference in the numerical simulations of Section 5 and the Supplementary Material. In view of its interpretation and since it always admits a well-defined minimizer, the MVR criterion thus offers a natural generalization of OLS for the estimation of linear models.

3.4. **Connection with Gaussian Maximum Likelihood.** MVR provides one criterion for the simultaneous approximation of conditional mean and variance functions. A related criterion is the KLIC of the scaled Gaussian density $f_\theta(Y, X)$ defined in (3.6) from the true conditional density function $f_{Y|X}(Y \mid X)$, which is minimized at a ML pseudo-true value (White, 1982). Define for $\theta \in \Theta$,

$$\mathcal{L}(\theta) := -E[\log f_\theta(Y,X)] = \frac{1}{2}\log(2\pi) + E\left[\log s(X'\gamma) + \frac{1}{2}e(Y,X,\theta)^2\right], \tag{3.12}$$

with first-order conditions

$$E\left[\frac{X}{s(X'\gamma)}e(Y,X,\theta)\right] = 0, \quad E\left[X\frac{s_1(X'\gamma)}{s(X'\gamma)}\left\{e(Y,X,\theta)^2 - 1\right\}\right] = 0. \tag{3.13}$$

In general the MVR solution θ^* need not satisfy equations (3.13), and therefore cannot be interpreted as a ML pseudo-true value. Compared with the MVR criterion, an important limitation of criterion (3.12) is its lack of convexity. The second-order derivative of $\mathcal{L}(\theta)$ with respect to the first component γ_1 of γ, i.e., for fixed β, γ_{-1}, is

$$\frac{\partial^2\mathcal{L}(\theta)}{\partial\gamma_1^2} = E\left[\frac{1}{s(X'\gamma)^2}\left\{3e(Y,X,\theta)^2 - 1\right\}\right],$$

which is strictly negative for all $\theta \in \Theta$ such that $e(Y,X,\theta)^2 \leq 1/3$ a.s. The non convexity of (3.12) in γ_1 implies that $\mathcal{L}(\theta)$ is not jointly convex[3], and that a ML pseudo-true value might not exist; even if there exists one, it need not be unique. In the latter case, some solutions may only be local minima of (3.12), not endowed with a KLIC-closest interpretation and thus no longer guaranteed to improve over OLS and MVR in a meaningful way.

These observations together with Theorems 4 and 5 clarify the relationship between ML, OLS and MVR approximating properties. The ML pseudo-true value is the parameter value associated with the distribution which is KLIC closest to the true data generating process, but is not well-defined due to the objective's lack of convexity. The OLS pseudo-true value is well-defined, but it maximizes a restricted version of the Gaussian expected log-likelihood resulting in a relatively lower likelihood. The MVR loss function strikes a compromise by providing a well-defined convex alternative to Gaussian ML, and relative to OLS by selecting a pseudo-true value that corresponds to a distribution which is KLIC closer to the true data generating process, and KLIC closer to the efficient GLS model under correct mean specification.

[3] Owen (2007) also noted the lack of joint convexity of the negative Gaussian log-likelihood when the scale function is specified to a constant, i.e., for the case $s(X'\gamma) = \sigma \in (0,\infty)$ in (3.12).

4. Estimation and Inference

We use the sample analog of the MVR population problem (2.2) for estimation of its solution θ^* in finite samples. We establish existence, uniqueness and consistency of the MVR estimator. We also derive its asymptotic distribution allowing for misspecification of the shapes of the conditional mean and variance functions, and discuss the robustness properties of its influence function. Finally, we provide corresponding tools for robust inference and introduce a one-step MVR-based test for heteroskedasticity.

We assume that we observe a sample of n independent and identically distributed realizations $\{(y_i, x_i)\}_{i=1}^n$ of the random vector (Y, X). We denote the $n \times k$ matrix of explanatory variables values by X_n. We define $\Theta_n = \mathbb{R}^k \times \Theta_{\gamma,n}$, with $\Theta_{\gamma,n} = \{\gamma \in \mathbb{R}^k : s(x_i'\gamma) > 0, i = 1, \ldots, n\}$, the sample analog of the parameter space Θ. For $\gamma \in \Theta_{\gamma,n}$, we let $\Omega_n(\gamma) = \text{diag}(s(x_i'\gamma))$, an $n \times n$ diagonal matrix with diagonal elements $s(x_1'\gamma), \ldots, s(x_n'\gamma)$. We also define the MVR moment functions

$$m_1(y_i, x_i, \theta) := x_i e(y_i, x_i, \theta), \quad m_2(y_i, x_i, \theta) := \frac{1}{2} x_i s_1(x_i'\gamma)\{e(y_i, x_i, \theta)^2 - 1\},$$

and the corresponding vector $m(y_i, x_i, \theta) := (m_1(y_i, x_i, \theta), m_2(y_i, x_i, \theta))'$.

4.1. The MVR Estimator.

The solution to the finite-sample analog of problem (2.2) is the MVR estimator

$$\hat{\theta} := \arg\min_{\theta \in \Theta_n} \frac{1}{n} \sum_{i=1}^n \frac{1}{2} \left\{ \left(\frac{y_i - x_i'\beta}{s(x_i'\gamma)} \right)^2 + 1 \right\} s(x_i'\gamma). \tag{4.1}$$

For $a = 0$ in Assumption 1, the sample objective in (4.1) is minimized subject to the n inequality constraints $s(x_i'\gamma) > 0$, $i = 1, \ldots, n$. For $a = -\infty$, the parameter space simplifies to $\Theta_n = \mathbb{R}^{2 \times k}$ and problem (4.1) is unconstrained. In terms of implementation, this constitutes an attractive feature of the exponential scale specification.

We derive the asymptotic properties of $\hat{\theta}$ under the following assumptions stated for a scale function in the class defined by Assumption 1.

Assumption 5. (i) $\{(y_i, x_i)\}_{i=1}^n$ are identically and independently distributed, and (ii) for all $\gamma \in \Theta_{\gamma,n}$, the matrix $X_n'\Omega_n^{-1}(\gamma)X_n$ is finite and positive definite.

Assumption 6. $E[Y^6] < \infty$, $E[||X||^6] < \infty$, and for all $\gamma \in \Theta_\gamma$, $E[\|X\|^6 \, s_2(X'\gamma)^6] < \infty$.

Assumption 5(i) can be replaced with the condition that $\{(y_i, x_i)\}_{i=1}^n$ is stationary and ergodic (Newey and Mc Fadden, 1994). Assumption 6 is needed for asymptotic normality of estimates of θ^*. When the scale function $t \mapsto s(t)$ is specified to be linear, this assumption simplifies to the requirement that $E[Y^6]$ and $E[\|X\|^6]$ be finite.

Letting $e = e(Y, X, \theta^*)$, the variance-covariance matrix of the MVR estimator $\hat{\theta}$ is $G^{-1}SG^{-1}/n$, where

$$G := \begin{bmatrix} G_{11} & G_{12} \\ G_{21} & G_{22} \end{bmatrix} := E \begin{bmatrix} \frac{XX'}{s(X'\gamma^*)} & \frac{XX'}{s(X'\gamma^*)} s_1(X'\gamma^*)e \\ \frac{XX'}{s(X'\gamma^*)} s_1(X'\gamma^*)e & XX' \left\{ \frac{(s_1(X'\gamma^*)e)^2}{s(X'\gamma^*)} - \frac{1}{2} s_2(X'\gamma^*)(e^2-1) \right\} \end{bmatrix}$$

and

$$S := \begin{bmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{bmatrix} := E \begin{bmatrix} XX'e^2 & \frac{1}{2} XX' s_1(X'\gamma^*)e(e^2-1) \\ \frac{1}{2} XX' s_1(X'\gamma^*)e(e^2-1) & \frac{1}{4} XX' \{s_1(X'\gamma^*)(e^2-1)\}^2 \end{bmatrix}.$$

The exact form of each component of matrices G and S depends on the specification of the conditional mean and variance functions, and simplifications of the variance-covariance matrix occur according to the type of misspecification. Under mean misspecification, the form of the variance-covariance matrix of the MVR estimator is not affected by the specification of the conditional variance function.

Define estimates of G and S by $\hat{G} := n^{-1} \sum_{i=1}^n \partial m(y_i, x_i, \hat{\theta})/\partial\theta$ and $\hat{S} := n^{-1} \sum_{i=1}^n m(y_i, x_i, \hat{\theta}) m(y_i, x_i, \hat{\theta})'$, respectively. The next theorem states the asymptotic properties of the MVR estimator.

Theorem 6. *If Assumptions 1-6 hold, then (i) there exists $\hat{\theta}$ in Θ with probability approaching one; (ii) $\hat{\theta} \to^p \theta^*$; and (iii)*

$$n^{1/2}(\hat{\theta} - \theta^*) \to_d \mathcal{N}(0, G^{-1}SG^{-1}). \tag{4.2}$$

If $\mu(X) = X'\beta^*$ a.s., then the following simplifications occur

$$G_{12} = G_{21} = 0_{k\times k}, \quad S_{12} = S_{21} = \frac{1}{2} E[XX' s_1(X'\gamma^*)e^3]. \tag{4.3}$$

If $\mu(X) = X'\beta^*$ a.s. and $\sigma(X)^2 = s(X'\gamma^*)^2$ a.s., then the following additional simplifications occur

$$G_{22} = E\left[\frac{XX'}{s(X'\gamma^*)} s_1(X'\gamma^*)\right], \quad S_{11} = E[XX'], \quad S_{22} = \frac{1}{4} E[XX' s_1(X'\gamma^*)^2(e^4-1)]. \tag{4.4}$$

Moreover, $\hat{G}^{-1}\hat{S}\hat{G}^{-1} \to^p G^{-1}SG^{-1}$.

Theorem 6 allows the construction of confidence intervals and the implementation of hypothesis tests for θ under each type of model specification using standard errors constructed from the corresponding variance-covariance matrix. The various forms of the variance-covariance matrix in Theorem 6 provide a basis for the construction of a range of specification tests, similarly to the information matrix equality test in ML theory (White, 1982; Chesher and Spady, 1991). Inference using the general asymptotic variance formula in (4.2) will automatically be robust to all forms of misspecification, and therefore to the presence of heteroskedasticity of unknown form.

For the linear homoskedastic model $Y = X'\beta_0 + U$ with $E[U \mid X] = 0$ and $\mathrm{Var}(U \mid X) = \sigma_0^2$, the MVR and OLS variance-covariance matrices coincide asymptotically, and MVR is efficient. Our numerical simulations in Section 5 and the Supplementary Material illustrate that there is close to no finite-sample loss in estimating linear homoskedastic models using MVR instead of OLS. If both the conditional mean and variance functions are correctly specified, then GLS with weights $1/\sigma^2(x)$ is an efficient estimator for β_0. Letting $\check{Y} := Y/s(X'\gamma^*)$, $\check{X} := X/s(X'\gamma^*)$ and $\check{s}(X'\gamma) := s(X'\gamma)/s(X'\gamma^*)$, define the weighted MVR objective

$$Q^{\mathrm{WMVR}}(\theta) := E\left[\frac{1}{2}\left\{\left(\frac{\check{Y} - \check{X}'\beta}{\check{s}(X'\gamma)}\right)^2 + 1\right\}\check{s}(X'\gamma)\right] = E\left[\frac{1}{2}\left\{e\left(Y, X, \theta\right)^2 + 1\right\}\frac{s(X'\gamma)}{s(X'\gamma^*)}\right].$$

If $\sigma^2(X) = s(X'\gamma_0)^2$, then $\gamma^* = \gamma_0$ and $Q^{\mathrm{WMVR}}(\theta)$ has first-order conditions for β

$$\frac{\partial Q^{\mathrm{WMVR}}(\theta)}{\partial \beta} = -E\left[\frac{X}{s(X'\gamma_0)}e\left(Y, X, \theta\right)\right] = 0,$$

which are satisfied by $\theta = \theta_0$ and coincide with the GLS (and ML) first-order conditions for β at a solution. In general, the functional form of the conditional variance function is unknown, and the MVR and weighted MVR solutions will differ.

An important implication of Theorem 6 is that the influence function of the MVR estimator for β is proportional to both moment functions m_1 and m_2:

$$IF_\beta(y, x, \theta) = -(G_{11} - G_{12}G_{22}^{-1}G_{21})^{-1}[m_1(y, x, \theta) - G_{12}G_{22}^{-1}m_2(y, x, \theta)].$$

The quadratic term m_2 dominates and an influential observation is defined as having $(y_i - x_i'\beta)^2$ large enough for $e(y_i, x_i, \theta)^2$ to be large. Observations that are influential for β are influential relative to the dispersion of Y, accounting for mean misspecification.

When the CMF is well-specified, the variance-covariance matrix takes the form

$$G^{-1}SG^{-1} = \begin{bmatrix} G_{11}^{-1}S_{11}G_{11}^{-1} & G_{11}^{-1}S_{12}G_{22}^{-1} \\ G_{22}^{-1}S_{21}G_{11}^{-1} & G_{22}^{-1}S_{22}G_{22}^{-1} \end{bmatrix}.$$

The influence function of β is then proportional to m_1 only, and the influence function of γ is proportional to m_2 only, since the off-diagonal blocks of G are then $0_{k\times k}$. For the mean parameter β, an observation (y_i, x_i) with large influence will be such that y_i is large enough for the standardized residual $e(y_i, x_i, \theta)$ to be large. Because $\hat\beta$ and $\hat\gamma$ are determined simultaneously, the influence of outliers on the mean parameter is limited by the restriction that the sample second moment of $e(y_i, x_i, \theta)$ must remain close to one, and be exactly one if the scale is linear. In sharp contrast with OLS, the scale parameter will simultaneously compensate an increase in Y dispersion so as to keep the variance of $e(y_i, x_i, \theta)$ constant. The MVR influence function although unbounded for a fixed value of γ thus robustifies OLS through the simultaneous reweighting of the residuals, downweighting regions in the covariate space where the information on Y is imprecise, as measured by $s(x'\gamma)$, in the calculation of the regression fit.

In summary, the MVR estimator does not robustify OLS through the bounding of the influence function (Koenker, 2005), but by incorporating information about the dispersion of Y across the covariate space in the definition of an influential outlier.

Remark 2. (Implementation.) Under our assumptions, the MVR objective is globally convex in θ, and therefore in β for any $\gamma \in \Theta_{\gamma,n}$. This implies that for any $\gamma \in \Theta_{\gamma,n}$ there exists a unique corresponding minimizer $\hat\beta(\gamma)$. This observation forms the basis of our implementation, and letting $y = (y_1, \ldots, y_n)'$, we first obtain $\hat\gamma$ by solving

$$\min_{\gamma\in\mathbb{R}^k} \frac{1}{n}\sum_{i=1}^{n}\frac{1}{2}\left\{\left(\frac{y_i - x_i'\hat\beta(\gamma)}{s(x_i'\gamma)}\right)^2 + 1\right\} s(x_i'\gamma), \quad \hat\beta(\gamma) := [X_n'\mathit{\Omega}_n^{-1}(\gamma)X_n]^{-1}X_n'\mathit{\Omega}_n^{-1}(\gamma)y,$$

$$\text{s.t.} \quad s(x_i'\gamma) > 0, \quad i = 1, \ldots, n, \quad \text{if } s(t) \leq 0 \text{ for some } t \in \mathbb{R}.$$

Concentrating out β for each γ provides a convenient implementation of the MVR estimator $\hat\gamma$, with the final estimate for β defined as $\hat\beta := \hat\beta(\hat\gamma)$. □

4.2. **Inference.** Given the MVR estimator $\hat\theta = (\hat\beta, \hat\gamma)'$, inference is performed based on the estimated asymptotic variance-covariance matrix $\hat V := \hat G^{-1}\hat S\hat G^{-1}$, which can be partitioned into 4 blocks

$$\hat V = \begin{bmatrix} \hat V_{11} & \hat V_{12} \\ \hat V_{21} & \hat V_{22} \end{bmatrix}.$$

The specific form of $\hat{V}$ depends on the specification assumptions made on the conditional mean and variance functions. For $\hat{\beta}_j$ and $\hat{\gamma}_j$ the jth components of $\hat{\beta}$ and $\hat{\gamma}$, respectively, MVR standard errors are obtained as

$$\text{s.e.}(\hat{\beta}_j) := \left(\frac{1}{n}\left[\hat{V}_{11}\right]_{j,j}\right)^{\frac{1}{2}}, \quad \text{s.e.}(\hat{\gamma}_j) := \left(\frac{1}{n}\left[\hat{V}_{22}\right]_{j,j}\right)^{\frac{1}{2}},$$

with resulting two-sided confidence intervals with nominal level $1-\alpha$,

$$\hat{\beta}_j \pm \Phi^{-1}(1-\alpha/2) \times \text{s.e.}(\hat{\beta}_j), \quad \hat{\gamma}_j \pm \Phi^{-1}(1-\alpha/2) \times \text{s.e.}(\hat{\gamma}_j),$$

where $\Phi^{-1}(1-\alpha/2)$ denotes the $1-\alpha/2$ quantile of the Gaussian distribution. A significance test of the null $\beta_j = 0$ and $\gamma_j = 0$ can then be performed using the test statistics $\hat{\beta}_j/\text{s.e.}(\hat{\beta}_j)$ and $\hat{\gamma}_j/\text{s.e.}(\hat{\gamma}_j)$.

Simultaneous significance testing or hypothesis tests on linear combination of multiple parameters can be implemented by a Wald test. For $h \leq 2 \times k$, letting R be an $h \times (2 \times k)$ matrix of constants of full rank h and r be an $h \times 1$ vector of constants, define

$$H_0 : R\theta^* - r = 0, \quad H_1 : R\theta^* - r \neq 0,$$

the null and alternative hypotheses for a two-sided tests of linear restrictions on the location-scale model $Y = X'\beta^* + s(X'\gamma^*)e$. It follows from asymptotic normality of $\hat{\theta}$ in (4.2) that the corresponding MVR Wald statistic W_{MVR} satisfies

$$W_{\text{MVR}} := (R\hat{\theta} - r)'[R(\hat{V}/n)R']^{-1}(R\hat{\theta} - r) \sim \chi^2_{(h)},$$

under the null H_0.

The Wald statistic W_{MVR} can be specialized to formulate a one-step robust MVR-based test for heteroskedasticity. Letting

$$h = k-1, \quad R = \left[\begin{array}{cc} 0_{k-1,k+1} & I_{k-1} \end{array}\right], \quad r = 0_{k-1},$$

the statistic W_{MVR} gives a robust test of the null hypothesis $H_0 : \gamma_2^* = \ldots = \gamma_k^* = 0$.

Remark 3. When the CMF is linear, robust MVR inference on $\hat{\beta}$ uses the closed-form variance formula

$$\widehat{\text{Var}}(\hat{\beta}) = n^{-1}(X_n'\Omega_n^{-1}(\hat{\gamma})X_n)^{-1}(X_n'\hat{\Psi}_e X_n)(X_n'\Omega_n^{-1}(\hat{\gamma})X_n)^{-1},$$

where $\hat{\Psi}_e = \text{diag}(\hat{e}_i^2)$. □

5. Numerical Illustrations

All computational procedures can be implemented in the software R (R Development Core Team, 2017) using open source software packages for nonlinear optimization such as Nlopt, and its R interface Nloptr (Ypma, Borchers and Eddelbuettel (2018)).

5.1. **Empirical Application: Reversal of Fortune.** We apply our methods to the study of the effect of European colonialism on today's relative wealth of former colonies, as in Acemoglu, Johnson and Robinson (2002). They show that former colonies that were relatively rich in 1500 are now relatively poor, and provide ample empirical evidence of this reversal of fortune. In particular, they study the relationship between urbanization in 1500 and GDP per capita in 1995 (PPP basis), using OLS regression analysis. The sample size ranges from 17 to 41 former colonies, allowing the illustration of MVR properties in small samples.

We take the outcome Y to be log GDP per capita in 1995 and in the baseline specification X includes an intercept and a measure of urbanization in 1500, a proxy for economic development. We implement MVR with both linear (ℓ-MVR) and exponential (e-MVR) scale functions, and report estimated standard errors robust to mean misspecification according to (4.2). We also report OLS estimates, with heteroskedasticity-robust standard errors. In the Supplementary Material we also provide results including standard errors with finite-sample adjustments suggested by MacKinnon and White (1985), and we also report MVR standard errors calculated under correct mean misspecification. Our findings are robust to these modifications.

Table 1 reports our results for urbanization in the baseline specification across 5 different sets of countries, and for 4 additional specifications[4] including continent dummies, and controlling for latitude, colonial origin and religion[5]. A striking feature of the results is the robustness to scale specification of MVR point estimates and standard errors. They are nearly identical across all specifications, except for Panel (3). Compared to OLS, MVR point estimates are all smaller in magnitude, suggesting a negative bias of OLS away from zero while standard errors are of similar magnitude, making it more likely to find a significant relationship with OLS estimates in this empirical application. The urbanization coefficient loses significance with MVR in 4 specifications, mainly as a result of the change in point estimates.

[4]We exclude two specifications of Table III in Acemoglu, Johnson and Robinson (2002) for which not all types of OLS and MVR standard errors are well-defined.
[5]See Acemoglu, Johnson and Robinson (2002) for a detailed description of the data.

	Dependent variable is log GDP per capita (PPP) in 1995								
	OLS	ℓ-MVR	e-MVR	OLS	ℓ-MVR	e-MVR	OLS	ℓ-MVR	e-MVR
	(1) Base sample ($n = 41$)			(2) Without North Africa ($n = 37$)			(3) Without the Americas ($n = 17$)		
Urbanization in 1500	-0.078	-0.067	-0.069	-0.101	-0.099	-0.099	-0.115	-0.064	-0.077
	(0.023)	(0.028)	(0.026)	(0.032)	(0.034)	(0.034)	(0.043)	(0.127)	(0.113)
	(4) Just the Americas ($n = 24$)			(5) With the continent dummies ($n = 41$)			(6) Without neo-Europes ($n = 37$)		
Urbanization in 1500	-0.053	-0.045	-0.044	-0.082	-0.063	-0.060	-0.046	-0.036	-0.038
	(0.029)	(0.032)	(0.032)	(0.031)	(0.029)	(0.030)	(0.021)	(0.023)	(0.023)
	(7) Controlling for Latitude ($n = 41$)			(8) Controlling for colonial origin ($n = 41$)			(9) Controlling for religion ($n = 41$)		
Urbanization in 1500	-0.072	-0.069	-0.070	-0.071	-0.063	-0.062	-0.060	-0.042	-0.040
	(0.020)	(0.022)	(0.021)	(0.025)	(0.026)	(0.027)	(0.027)	(0.029)	(0.029)

TABLE 1. Reversal of fortune. Asymptotic heteroskedasticity-robust OLS standard errors and MVR standard errors are in parenthesis.

Specifically, we find that MVR provides supporting evidence of a significant statistical relationship between urbanization in 1500 and GDP per capita in 1995 in the whole sample, but also dropping North Africa, including continent dummies, and controlling for latitude and for colonial origin. However, the relationship between urbanization in 1500 and GDP per capita in 1995 is not statistically significant in the four remaining specifications. When the Americas are dropped (Panel (3)), when only former colonies from the Americas are considered (Panel (4)), and when controlling for religion (Panel (9)), the urbanization coefficient is no longer significant with MVR. These results are robust to implementing finite-sample adjustments. Specification (6) drops observations for neo-Europes (United States, Canada, New Zealand, and Australia), and only the e-MVR estimate is significant at the 10 percent level when no finite-sample adjustments are implemented, and loses significance otherwise.

MVR results provide renewed empirical support for a subset of the specifications, but overall show that the mean relationship in this empirical application is weaker and less robust than first suggested by the OLS-based analysis[6]. These findings illustrate that MVR can substantially alter the conclusions obtained using OLS in practice.

5.2. **Numerical Simulations.** We investigate the properties of MVR in small samples and compare its performance with OLS and WLS by implementing numerical simulations based on the experimental setup in MacKinnon (2013). In the Supplementary Material, we provide additional results for models featuring a nonlinear CMF and report simulation results from an experiment calibrated to a second empirical example. We find that using MVR approximations does not result in a loss in the quality of approximation of nonlinear CMFs compared to OLS, and MVR estimation and inference finite-sample properties compare favorably to both OLS and WLS.

5.2.1. *Design of Experiments.* The data generating process is

$$\begin{aligned} Y &= \beta_0 + X_1\beta_1 + X_2\beta_2 + X_3\beta_3 + X_4\beta_4 + \sigma\varepsilon, \quad \varepsilon \sim \mathcal{N}(0,1) \\ \sigma &= z(\alpha)\left(\gamma_0 + X_1\gamma_1 + X_2\gamma_2 + X_3\gamma_3 + X_4\gamma_4\right)^{\alpha}, \quad \alpha \in \{0, 0.5, 1, 1.5, 2\}, \end{aligned}$$

where all regressors are drawn from the standard log-normal distribution, and $z(\alpha)$ is chosen such that the expected variance of $\sigma\varepsilon$ is equal to 1. The log-normal regressors

[6]We also implemented WLS and found that the magnitude of most WLS coefficients is smaller than MVR point estimates. In addition to specifications (3), (4), (6) and (9), specification (5) is also found to be not statistically significant. We report the results in Section 3 of the Supplementary Material.

ensure that many samples will include high-leverage points with a few observations taking extreme values. This feature of the design distorts the distribution of test statistics based on heteroskedasticity-robust estimators of OLS standard errors. The parameter coefficient values are set to $\beta_j = \gamma_j = 1$ for $j = 0, \ldots, 4$.[7] The index α measures the degree of heteroskedasticity in the model, with $\alpha = 0$ corresponding to homoskedasticity, and $\alpha = 2$ corresponding to high heteroskedasticity. The numerical simulations are implemented for sample sizes $n = 20, 40, 80, 160, 320, 640$ and 1280.

For each α and sample size, we generate 10000 samples, and implement OLS, WLS and MVR. We implement MVR with both linear (ℓ-MVR) and exponential (e-MVR) scale functions. For WLS we follow the implementations proposed by Romano and Wolf (2017, cf. equation (3.4) and (3.5), p. 4). Denote the OLS estimator by $\hat{\beta}_{\mathrm{LS}}$ and let $\tilde{x}_i = (x_{1i}, x_{2i}, x_{3i}, x_{4i})'$. We form the OLS residuals $\hat{u}_i := y_i - x_i'\hat{\beta}_{\mathrm{LS}}$, $i = 1, \ldots, n$, and perform the OLS regressions

$$\log(\max(\delta^2, \hat{u}_i^2)) = \nu + \pi'|\tilde{x}_i| + \eta_i,$$

for WLS with linear scale (ℓ-WLS)[8], and

$$\log(\max(\delta^2, \hat{u}_i^2)) = \nu + \pi' \log(|\tilde{x}_i|) + \eta_i,$$

for WLS with exponential scale (e-WLS), where $\delta = 0.1$ as in the implementation of Romano and Wolf (2017), and with estimates $(\hat{\nu}, \hat{\pi})$. The WLS weights are formed as $\hat{w}_i^{\ell} := \hat{\nu} + \hat{\pi}'|\tilde{x}_i|$ and $\hat{w}_i^{e} := \exp(\hat{\nu} + \hat{\pi}' \log(|\tilde{x}_i|))$, and the WLS estimators are

$$\hat{\beta}_{\mathrm{WLS}}^{m} := [X_n'(W_n^m)^{-1}X_n]^{-1}X_n'(W_n^m)^{-1}y, \quad W_n^m := \mathrm{diag}(\hat{w}_i^m), \quad m = \ell, e,$$

where $y = (y_1, \ldots, y_n)'$, X_n is the $n \times 5$ matrix of explanatory variables values, and $\mathrm{diag}(\hat{w}_i^{\ell})$ and $\mathrm{diag}(\hat{w}_i^{e})$ denote the $n \times n$ diagonal matrices with diagonal elements $\hat{w}_1^{\ell}, \ldots, \hat{w}_n^{\ell}$ and $\hat{w}_1^{e}, \ldots, \hat{w}_n^{e}$, respectively.

In all experiments the results for β_1 to β_4 are similar and we thus only report the results for β_4 for brevity. Also, throughout the relative performance of MVR and WLS is assessed by comparing ℓ-MVR to ℓ-WLS, and e-MVR to e-WLS.

5.2.2. *Estimation results.* Tables 2 and 3 report the ratio of MVR root mean squared errors (RMSE) across simulations over the OLS and WLS RMSEs for the coefficient

[7] This departs slightly from the original MacKinnon (2013) design where $\beta_4 = \gamma_4 = 0$. We are grateful to James MacKinnon for suggesting this modification that preserves heteroskedasticity in X_4.

[8] This regression is performed imposing the n constraints $\nu + \pi|\tilde{x}_i| \geq \delta$, $i = 1, \ldots, n$, using the `lsei` R package (Wang, Lawson and Hanson, 2017).

parameter β_4, each sample size and value of heteroskedasticity parameter α, in percentage terms. Denoting an estimator $\tilde{\beta}_4^{(s)}$ of β_4 for the sth simulation, the RMSE is computed as $\{\frac{1}{S}\sum_{s=1}^{S}(\tilde{\beta}_4^{(s)} - \beta_4)^2\}^{1/2}$, for $S = 10000$.

Table 2 shows that the performance of both MVR estimators relative to OLS improves as n and α increase. As expected, for the homoskedastic case $\alpha = 0$ the performance of MVR and OLS estimators is very similar, and the ratios converge to 100 from above, reflecting the efficiency of the OLS estimator in that case. The performance of MVR then becomes markedly superior as n and α increase, with ratios that reach 31.7 for ℓ-MVR and 22.9 for e-MVR. The estimator ℓ-MVR dominates e-MVR slightly for $n = 20$. The performance of the estimator e-MVR then becomes superior as the degree of heteroskedasticity and sample size increase, showing higher robustness of the exponential scale specification in more extreme designs in these simulations.

In Table 3, we find that the relative performance of both MVR estimators relative to WLS also improves as n increases and as α increases from 0.5 to 2. For the homoskedastic case $\alpha = 0$, an interesting feature of the simulation results is that the relative performance of MVR and WLS estimators now converges to 100 from below. This reflects the fact that for homoskedastic designs MVR weights are better able to mitigate the cost of reweighting in small samples compared to WLS weights. For other designs with $\alpha > 0$, the relative performance of both MVR estimators dominates the performance of WLS with ratios that reach 76.2 for ℓ-MVR and 43.5 for e-MVR. For $\alpha = 1$, WLS with linear scale is efficient and dominates slightly ℓ-MVR as n increases. Compared to OLS and the results of Table 2, these results show that in this experiment WLS also improves over OLS, that ℓ-MVR improves over WLS as n increases and α deviates from 1, with little loss for $\alpha \leq 1$, and e-MVR yields substantial additional gains over WLS as both n and α increase.

5.2.3. *Inference.* In order to study the finite-sample performance of MVR inference relative to heteroskedasticity-robust OLS and WLS inference, we first compare the rejection probabilities of asymptotic t tests of the null hypothesis $\beta_4 = 1$ based on the standard normal distribution[9]. We then compare the lengths of the confidence intervals constructed for the coefficient parameter β_4. MVR standard errors are calculated under mean misspecification according to (4.2). OLS and WLS standard errors used in the construction of confidence intervals and tests statistics are the asymptotic

[9]We also performed simulations with and tested for $\beta_4 = 0$, and calculated rejection probabilities using a t_{n-k} distribution. The relative performance of the methods remains similar.

	ℓ-MVR					e-MVR				
α	0	0.5	1	1.5	2	0	0.5	1	1.5	2
$n = 20$	105.5	104.1	98.9	91.9	84.9	106.7	105.2	99.7	92.4	84.4
$n = 40$	103.7	100.1	90.9	79.6	69.2	103.4	100.0	91.1	79.5	67.2
$n = 80$	102.3	96.7	84.0	69.7	58.5	102.1	96.9	84.3	69.1	54.6
$n = 160$	101.7	94.0	77.4	60.5	49.5	101.5	94.2	77.4	58.9	43.2
$n = 320$	101.1	91.3	71.5	53.0	42.6	100.9	91.4	70.7	50.2	34.6
$n = 640$	100.7	89.5	66.8	47.1	37.3	100.6	89.6	66.1	44.1	28.7
$n = 1280$	100.5	86.6	61.1	40.8	31.7	100.4	86.6	60.1	37.3	22.9

TABLE 2. Ratio ($\times$100) of MVR RMSE for β_4 over corresponding OLS counterpart.

	ℓ-MVR					e-MVR				
α	0	0.5	1	1.5	2	0	0.5	1	1.5	2
$n = 20$	98.7	99.0	99.2	98.7	97.1	97.3	98.4	99.4	99.6	97.8
$n = 40$	97.1	97.8	98.6	97.4	93.1	94.9	96.2	97.4	96.5	90.2
$n = 80$	96.5	98.1	99.7	96.2	89.6	97.2	98.0	97.8	93.6	82.5
$n = 160$	96.7	99.2	101.7	94.4	86.7	97.9	96.9	93.1	85.0	70.1
$n = 320$	96.9	100.3	102.1	92.4	85.2	99.0	96.0	89.0	77.8	59.8
$n = 640$	97.4	101.3	103.4	91.0	80.4	99.8	95.8	85.7	71.0	50.8
$n = 1280$	98.1	102.3	104.1	88.8	76.2	99.8	94.2	81.6	64.7	43.5

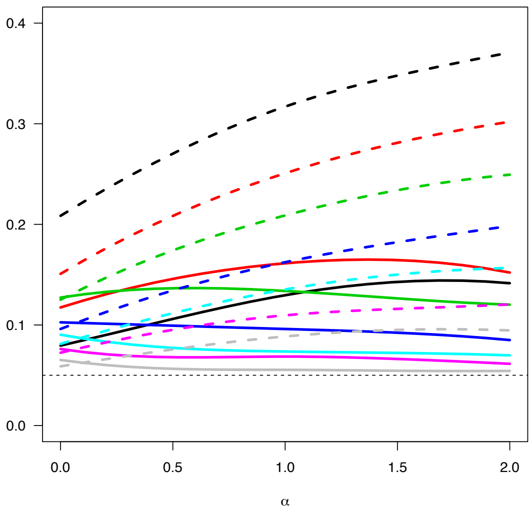
TABLE 3. Ratio ($\times$100) of MVR RMSE for β_4 over corresponding WLS counterpart.

heteroskedasticity-robust standard errors. For completeness, in the Supplementary Material we also compare rejection probabilities and confidence intervals based on standard errors with finite-sample adjustments suggested by MacKinnon and White (1985), and we also replicate all experiments using MVR standard errors calculated under correct mean misspecification with the simplifications in (4.3).

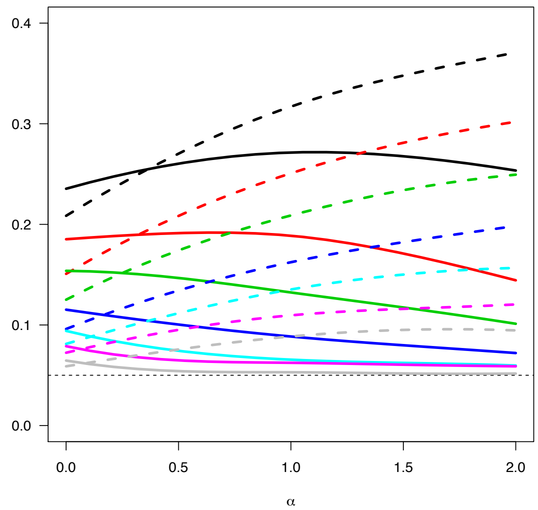
Figure 5.1 displays rejection probability curves of asymptotic t tests of the null hypothesis $\beta_4 = 1$ for each sample size and value of the heteroskedasticity parameter α. The nominal size of the tests is set to 5%. Figures 5.1(a)-(d) show that MVR addresses overrejection of the OLS- and WLS-based tests in the presence of heteroskedasticity ($\alpha > 0$). A striking feature of MVR rejection probability curves is that they flatten very quickly across α as n increases, a feature somewhat more pronounced for ℓ-MVR curves. This is in sharp contrast with OLS and e-WLS rejection probability curves that are increasing with the degree of heteroskedasticity α. For ℓ-WLS the rejection curves are distorted around $\alpha = 1$, for which it is efficient, and overall the rejection probabilities are much larger than for ℓ-MVR. The curves for ℓ-MVR and ℓ-WLS coincide only for the case where ℓ-MVR is efficient ($\alpha = 1$) at moderate sample size and above ($n \geq 320$). The ℓ-MVR rejection probability curve for $n = 20$ (black curve) is not placed above the other curves although it is above the nominal level for all values of α. This feature disappears when finite-sample corrections are implemented (Figures 2.1-2.3 in the Supplementary Material).

In order to further investigate the relative performance of MVR-based inference, Tables 4 and 5 report the ratio of average MVR confidence interval lengths across simulations over the average OLS and WLS confidence interval lengths for β_4 for each sample size and value of heteroskedasticity index α, in percentage terms. We find that in the presence of heteroskedasticity, the length of MVR confidence intervals is shorter for all designs compared to both OLS and WLS confidence intervals for some n large enough. The only exception is relative to ℓ-WLS for $\alpha = 0.5, 1$, as expected for $\alpha = 1$ from ℓ-WLS being efficient in that case. The relatively larger average length of the confidence intervals for ℓ-MVR when $n = 20$ in Tables 4 and 5 is very much reduced with finite-sample corrections (Tables 1-4 in the Supplementary Material).

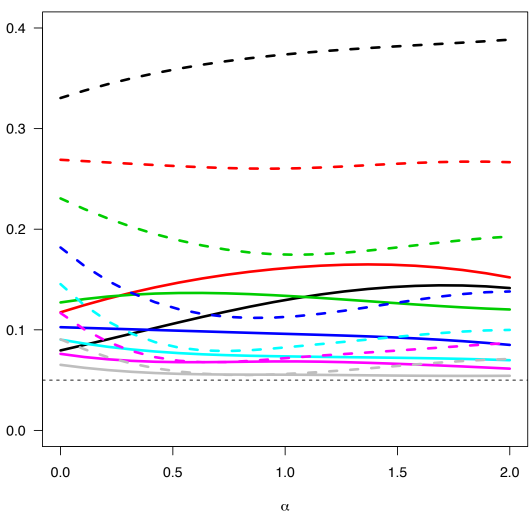
Overall these simulation results demonstrate that MVR achieves large improvements in terms of estimation and inference compared to OLS in the presence of heteroskedasticity, and compared to WLS when the conditional variance function is misspecified. Our numerical simulations confirm MVR robustness to the specification of the scale function, and both ℓ-MVR and e-MVR perform very well in finite samples. In the presence of heteroskedasticity rejection probabilities for MVR are much closer to nominal level than those for OLS and for WLS with misspecified weights. MVR achieves these improvements while simultaneously displaying tighter confidence intervals in all designs for sample sizes large enough. They are also shorter than their WLS counterpart with misspecified weights for sample sizes large enough. The precision of MVR estimates measured in RMSE is also largely superior to OLS under heteroskedasticity and to WLS with misspecified weights, with lower losses than WLS relative to OLS under homoskedasticity. These results and the simulations in the Supplementary Material illustrate the higher precision, improved finite-sample inference, and favorable approximation properties of MVR compared to classical least-squares methods.



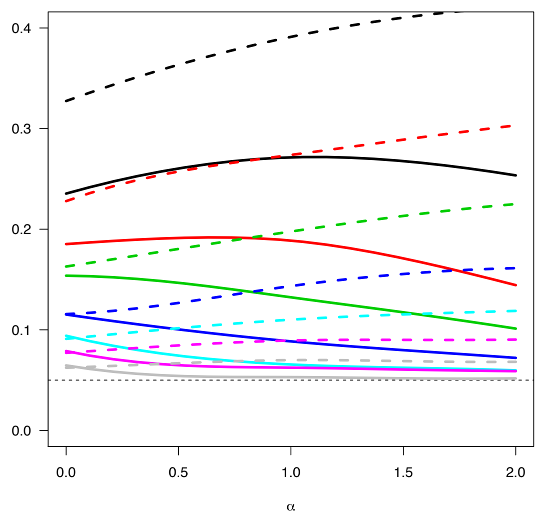
(A) ℓ-MVR vs OLS.



(B) e-MVR vs OLS.



(C) ℓ-MVR vs ℓ-WLS.



(D) e-MVR vs e-WLS.

FIGURE 5.1. Rejection frequencies for asymptotic t tests calculated with asymptotic standard errors: ℓ-MVR and e-MVR (solid lines), and OLS and WLS (dashed lines). Sample sizes: 20 (black), 40 (red), 80 (green), 160 (blue), 320 (cyan), 640 (magenta), 1280 (grey).

	ℓ-MVR					e-MVR				
α	0	0.5	1	1.5	2	0	0.5	1	1.5	2
$n = 20$	204.1	193.8	179.9	167.2	158.7	114.7	121.8	123.2	123.4	121.9
$n = 40$	121.0	117.5	109.8	100.0	91.0	104.2	109.8	108.8	103.1	94.1
$n = 80$	106.7	105.9	97.7	85.5	75.0	101.6	106.0	100.7	89.1	74.3
$n = 160$	102.9	101.8	90.1	74.8	64.0	100.9	103.3	92.9	76.4	59.1
$n = 320$	101.4	98.5	83.1	65.6	55.3	100.6	100.0	84.8	65.1	47.3
$n = 640$	100.6	95.2	76.2	57.5	47.8	100.2	96.3	77.0	55.6	38.2
$n = 1280$	100.4	92.1	70.1	50.5	41.3	100.2	93.0	70.3	47.9	31.1

Table 4. Ratio ($\times 100$) of MVR average confidence interval lengths for β_4 over corresponding OLS counterpart. Confidence intervals constructed with asymptotic standard errors.

	ℓ-MVR					e-MVR				
α	0	0.5	1	1.5	2	0	0.5	1	1.5	2
$n = 20$	245.2	226.0	210.7	201.3	197.4	136.1	143.3	149.8	158.7	165.9
$n = 40$	140.4	128.8	123.4	120.1	117.0	113.0	118.8	126.1	132.7	133.8
$n = 80$	120.5	111.5	109.8	107.5	103.2	105.7	110.7	116.1	119.4	113.5
$n = 160$	112.7	105.7	106.3	101.9	96.7	102.8	106.8	108.6	107.2	96.1
$n = 320$	108.7	103.4	105.4	98.3	92.7	101.4	103.9	102.0	96.5	81.8
$n = 640$	105.5	102.7	105.1	95.3	89.0	100.7	101.2	95.8	87.0	70.1
$n = 1280$	103.4	102.7	104.9	93.0	85.4	100.4	98.9	90.5	79.4	60.9

Table 5. Ratio ($\times 100$) of MVR average confidence interval lengths for β_4 over corresponding WLS counterpart. Confidence intervals constructed with asymptotic standard errors.

6. Conclusion

We introduce a new loss function for the linear estimation and approximation of CMFs. The proposed alternative generalises OLS, resulting in more robust approximations under misspecification and large improvements in finite samples. Given the importance of the least-squares loss in econometrics and statistics, and the common occurence of heteroskedasticity in empirical practice, the range of applications for simultaneous mean-variance regression will be vast. Examples of natural avenues for future research include the method of instrumental variables, GARCH models, and flexible specification of the conditional variance function for efficient estimation. These extensions will be considered in subsequent work.

Appendix A. Theory for the MVR Criterion

A.1. **Notation and Definitions.** We define

$$L(X,Y,\theta) := \frac{1}{2}\left\{\left(\frac{Y - X'\beta}{s(X'\gamma)}\right)^2 + 1\right\} s(X'\gamma),$$

and

$$\widetilde{L}(X,\theta) := \frac{1}{2}\left\{\frac{E[(Y - X'\beta)^2 \mid X]}{s(X'\gamma)} + s(X'\gamma)\right\}.$$

so that by iterated expectations the objective function can be expressed as

$$Q(\theta) = E[L(X,Y,\theta)] = E[\widetilde{L}(X,\theta)], \qquad \theta \in \Theta.$$

We denote the level sets of $Q(\theta)$ by $\mathcal{B}_c = \{\theta \in \Theta : \ Q(\theta) \leq c\}$, $c \in \mathbb{R}$, with boundary set $\partial\mathcal{B}_c$. We also define the compact set $\mathcal{B} = \mathcal{B}_\beta \times \mathcal{B}_\gamma \subseteq \Theta$, where $\mathcal{B}_\beta$ and $\mathcal{B}_\gamma$ are compact subsets of $\mathbb{R}^k$ and Θ_γ, respectively, and the boundary set of Θ

$$\partial\Theta = \mathbb{R}^k \times \partial\Theta_\gamma, \quad \partial\Theta_\gamma = \left\{\gamma \in \mathbb{R}^k \ : \ \Pr[s(X'\gamma) = 0] > 0\right\}.$$

For any two real numbers a and b, $a \vee b = \max(a,b)$. For two random variables U and V, $\mathcal{U}$ denotes the support of U, defined as the set of values of U such that the density $f_U(u)$ of U is bounded away from 0, and $\mathcal{U}_v$ is the conditional support of U given $V = v$, $v \in \mathcal{V}$. Throughout, C is a generic constant whose value may change from place to place.

A.2. **Preliminary Results.** This section gathers two preliminary results used in establishing the properties of $Q(\theta)$.

Lemma 1. *Let V be a random k vector such that $E[VV']$ exists and is nonsingular. Then, for every sequence (γ_n) in $\mathbb{R}^k$ such that $||\gamma_n|| \to \infty$, there exists $v^* \in \mathcal{V}$ such that $\lim_{||\gamma_n|| \to \infty} |\gamma_n' v^*| = \infty$ a.s.*

Proof. Consider a sequence (γ_n) in $\mathbb{R}^k$ such that $||\gamma_n|| \to \infty$, and define $\delta_n = \frac{\gamma_n}{||\gamma_n||}$. The sequence (δ_n) is bounded, and by application of the Bolzano-Weierstrass theorem there exists a convergent subsequence δ_{n_l}, $n_l \to \infty$ as $l \to \infty$, with limit δ_o. Moreover, $E[VV']$ nonsingular implies that it is positive definite, so that $E[(V'\delta_o)^2] = \delta_o' E[VV']\delta_o > 0$. It follows that $V'\delta_o \neq 0$ on a set of positive probability, and there exists a value $v^* \in \mathcal{V}$ such that $\delta_o' v^* \neq 0$ a.s. Therefore, $\delta_{n_l} = \frac{\gamma_{n_l}}{||\gamma_{n_l}||}$ satisfies $\delta_{n_l}' v^* \to \delta_o' v^* \neq 0$ as $l \to \infty$, which implies that $\lim_{l\to\infty} |\gamma_{n_l}' v^*| \to \infty$:

$$\lim_{l\to\infty} |\gamma_{n_l}' v^*| = \lim_{l\to\infty} \left|(\delta_{n_l}' v^*)||\gamma_{n_l}||\right| = \left|(\delta_o' v^*) \lim_{l\to\infty} ||\gamma_{n_l}||\right| = \infty.$$

The stated result follows. □

Lemma 2. *Suppose that Assumptions 1, 2 and 4 hold. Then the matrix*

$$\Psi(\theta) = E\left[\begin{array}{cc} \frac{XX'}{s(X'\gamma)} & \frac{XX'}{s(X'\gamma)} s_1(X'\gamma)e(Y,X,\theta) \\ \frac{XX'}{s(X'\gamma)} s_1(X'\gamma)e(Y,X,\theta) & \frac{XX^\top}{s(\gamma\cdot X)}\{s_1(X'\gamma)e(Y,X,\theta)\}^2 \end{array}\right],$$

defined for all $\theta \in \mathcal{B}$, is positive definite.

Proof. The proof builds on the proofs of Lemma S3 in Spady and Stouli (2018) and Theorem 1 in Newey and Stouli (2018). Positive definiteness of $E[XX'/s(X'\gamma)]$ for $\gamma \in \mathcal{B}_\gamma$ under Assumption 4 implies that $\Psi(\theta)$ is positive definite for all $\theta \in \mathcal{B}$ if and only if the Schur complement of $E[XX'/s(X'\gamma)]$ in $\Psi(\theta)$ is positive definite (Boyd and Vandenberghe, 2004, Appendix A.5.5) for all $\theta \in \mathcal{B}$, i.e., if and only if

$$\begin{aligned}\varUpsilon(\theta) := {} & E\left[\frac{XX'}{s(X'\gamma)}\{s_1(X'\gamma)e(Y,X,\theta)\}^2\right] \\ & - E\left[\frac{XX'}{s(X'\gamma)} s_1(X'\gamma)e(Y,X,\theta)\right] E\left[\frac{XX'}{s(X'\gamma)}\right]^{-1} E\left[\frac{XX'}{s(X'\gamma)} s_1(X'\gamma)e(Y,X,\theta)\right]\end{aligned}$$

satisfies $\det\{\varUpsilon(\theta)\} > 0$, for all $\theta \in \mathcal{B}$.

Letting

$$\Xi(\theta) := E\left[\frac{XX'}{s(X'\gamma)} s_1(X'\gamma)e(Y,X,\theta)\right] E\left[\frac{XX'}{s(X'\gamma)}\right]^{-1},$$

for all $\theta \in \mathcal{B}$, $\varUpsilon(\theta)$ is equal to

$$E\left[\left\{\frac{Xs_1(X'\gamma)e(Y,X,\theta)}{s(X'\gamma)^{1/2}} - \Xi(\theta)\frac{X}{s(X'\gamma)^{1/2}}\right\}\left\{\frac{Xs_1(X'\gamma)e(Y,X,\theta)}{s(X'\gamma)^{1/2}} - \Xi(\theta)\frac{X}{s(X'\gamma)^{1/2}}\right\}'\right],$$

a finite positive definite matrix, if and only if for all $\lambda \neq 0$ and all $\theta \in \mathcal{B}$ there is no d such that

$$\Pr\left[\left\{\frac{\lambda' X}{s(X'\gamma)^{1/2}}\right\} s_1(X'\gamma)e(Y,X,\theta) = d'\left\{\Xi(\theta)\frac{X}{s(X'\gamma)^{1/2}}\right\}\right] > 0; \tag{A.1}$$

this is an application of the Cauchy-Schwarz inequality for matrices stated in Tripathi (1999).

Positive definiteness of $E\left[XX'/s(X'\gamma)\right]$ for $\gamma \in \mathcal{B}_\gamma$ under Assumption 4 implies that $E\left[\{\lambda' X\}^2/s(X'\gamma)\right] > 0$ for all $\lambda \neq 0$, which implies that $\Pr[\lambda' X/s(X'\gamma)^{1/2} \neq 0] > 0$ for all $\lambda \neq 0$. Also, by Assumptions 1-2, we have $\Pr[s_1(X'\gamma) > 0] = 1$ and

$$\Pr\left[\mathrm{Var}(s_1(X'\gamma)e(Y, X, \theta) \mid X) > 0\right] = \Pr\left[\left(\frac{s_1(X'\gamma)}{s(X'\gamma)}\right)^2 \mathrm{Var}(Y \mid X) > 0\right] = 1.$$

Thus for all $\lambda \neq 0$, by $\Xi(\theta)$ being a constant matrix for all $\theta \in \mathcal{B}$, there is no d such that (A.1) holds, and the result follows. $\square$

A.3. **Main Properties of $Q(\theta)$.**

Lemma 3. *[Continuity] Suppose that Assumptions 1 and 3 hold. Then $\theta \mapsto Q(\theta)$ is continuous over $\mathcal{B}$.*

Proof. We first show that $E[\sup_{\theta\in\mathcal{B}} |L(X, Y, \theta)|] < \infty$ for all $\theta \in \mathcal{B}$. By the Triangle Inequality,

$$2\,|L(X, Y, \theta)| \leq |e(Y, X, \theta)^2 s(X'\gamma)| + |s(X'\gamma)|. \tag{A.2}$$

Compactness of $\mathcal{B}_\gamma$ implies that there exists a constant C such that $\sup_{\gamma\in\mathcal{B}_\gamma} 1/s(X'\gamma) \leq C < \infty$ a.s. Thus for $\theta \in \mathcal{B}$, the bound

$$|e(Y, X, \theta)^2 s(X'\gamma)| \leq C[2Y^2 + 2(X'\beta)^2] \leq 2C[Y^2 + \sup_{\beta\in\mathcal{B}_\beta} ||\beta||^2 ||X||^2], \tag{A.3}$$

and $\sup_{\beta\in\mathcal{B}_\beta} ||\beta|| < \infty$ together imply that $E[\sup_{\beta\in\mathcal{B}} |e(Y, X, \theta)^2 s(X'\gamma)|] < \infty$ requires $E[Y^2] < \infty$ and $E||X||^2 < \infty$, which hold under Assumption 3(i).

It remains to show that $E[\sup_{\gamma\in\mathcal{B}_\gamma} |s(X'\gamma)|] < \infty$. For $\gamma \in \mathcal{B}_\gamma$, some $0 \leq \kappa \in (a, \infty)$ and some intermediate values $\bar{\gamma}$, a mean-value expansion about $(\kappa, 0_{k-1})'$ yields

$$|s(X'\gamma)| = |s(\kappa) + s_1(X'\bar{\gamma})(X'\gamma - \kappa)| \leq s(\kappa) + \sup_{\gamma\in\mathcal{B}_\gamma} ||\gamma||\,||X|| s_1(X'\bar{\gamma}).$$

With $s(\kappa), \sup_{\gamma\in\mathcal{B}_\gamma} ||\gamma|| < \infty$, $E[\sup_{\gamma\in\mathcal{B}_\gamma} |s(X'\gamma)|] < \infty$ requires $E[||X|| s_1(X'\bar{\gamma})] < \infty$, which holds under Assumption 3(ii).

Bound (A.2) now implies that $E[\sup_{\theta\in\mathcal{B}} |L(X, Y, \theta)|] < \infty$, and continuity of $Q(\theta)$ then follows from continuity of $\theta \mapsto L(X, Y, \theta)$ and dominated convergence. $\square$

Lemma 4. *[Continuous Differentiability] If Assumptions 1 and 3 hold, then, for all $\theta \in \mathcal{B}$, $Q(\theta)$ is continuously differentiable and $\partial E[L(X, Y, \theta)]/\partial\theta = E[\partial L(X, Y, \theta)/\partial\theta]$.*

Proof. We first show that $E[\sup_{\theta\in\mathcal{B}}||\partial L(X,Y,\theta)/\partial\theta||]<\infty$. Computing

$$\partial L(X,Y,\theta)/\partial\beta=-Xe(Y,X,\theta),\quad \partial L(X,Y,\theta)/\partial\gamma=-\frac{1}{2}Xs_1(X'\gamma)\{e(Y,X,\theta)^2-1\}.$$

Compactness of $\mathcal{B}_\gamma$ implies that there exists a constant C such that $\sup_{\gamma\in\mathcal{B}_\gamma}1/s(X'\gamma)\leq C<\infty$ a.s. Thus for $\theta\in\mathcal{B}$, the bound

$$||Xe(Y,X,\theta)||\leq C||X||\,|Y-X'\beta|\leq C[|Y|\,||X||+\sup_{\beta\in\mathcal{B}_\beta}||\beta||\,||X||^2],$$

and $\sup_{\beta\in\mathcal{B}_\beta}||\beta||<\infty$, imply that $E[\sup_{\theta\in\mathcal{B}}||\partial L(X,Y,\theta)/\partial\beta||]<\infty$ requires $E[|Y|\,||X||]<\infty$ and $E||X||^2<\infty$, which hold under Assumptions 3(i).

We now show that $E[\sup_{\theta\in\mathcal{B}}||\partial L(X,Y,\theta)/\partial\gamma||]<\infty$. Since $-1\leq e(Y,X,\theta)^2-1$ a.s., for $\theta\in\mathcal{B}$,

$$\begin{aligned}\left\|Xs_1(X'\gamma)\{e(Y,X,\theta)^2-1\}\right\| &\leq ||X||\,s_1(X'\gamma)\,|e(Y,X,\theta)^2-1|\\ &\leq ||X||\,s_1(X'\gamma)\,[1\vee 2C(Y^2+\sup_{\beta\in\mathcal{B}_\beta}||\beta||^2||X||^2)].\end{aligned}\tag{A.4}$$

For $\gamma\in\mathcal{B}_\gamma$, some $0\leq\kappa\in(a,\infty)$ and some intermediate values $\bar{\gamma}$, a mean-value expansion about $(\kappa,0_{k-1})'$ yields

$$|s_1(X'\gamma)|=|s_1(\kappa)+s_2(X'\bar{\gamma})(X'\gamma-\kappa)|\leq s_1(\kappa)+\sup_{\gamma\in\mathcal{B}_\gamma}||\gamma||\,||X||s_2(X'\bar{\gamma}).\tag{A.5}$$

This bound and (A.4) together imply

$$\begin{aligned}\left\|Xs_1(X'\gamma)\{e(Y,X,\theta)^2-1\}\right\| &\leq ||X||\,[s_1(\kappa)+\sup_{\gamma\in\mathcal{B}_\gamma}||\gamma||\,||X||s_2(X'\bar{\gamma})]\\ &\quad\times[1\vee 2C(Y^2+\sup_{\beta\in\mathcal{B}_\beta}||\beta||^2||X||^2)].\end{aligned}\tag{A.6}$$

Since $\mathcal{B}$ is compact and $0<s_1(\kappa)<\infty$, $E[\sup_{\theta\in\mathcal{B}}||\nabla_\gamma L(X,Y,\theta)||]<\infty$ requires $E||X||^3<\infty$, $E[Y^2\,||X||]<\infty$ and, for all $\gamma\in\mathcal{B}_\gamma$, $E[\|X\|^4\,s_2(X'\gamma)]<\infty$ and $E[Y^2||X||^2s_2(X'\gamma)]<\infty$, which hold under Assumption 3.

We have shown that $E[\sup_{\theta\in\mathcal{B}}||\partial L(X,Y,\theta)/\partial\theta||]<\infty$ and it now follows by Lemma 3.6 in Newey and Mc Fadden (1994) that $Q(\theta)$ is continuously differentiable over $\mathcal{B}$, and that the order of differentiation and integration can be interchanged. □

Lemma 5. *[Convexity] Suppose that Assumptions 1, 3 and 4 hold. Then $\theta\mapsto Q(\theta)$ is strictly convex over $\mathcal{B}$.*

Proof. $Q(\theta)$ is differentiable for all $\theta \in \mathcal{B}$ and the order of integration and differentiation can be interchanged by Lemma 4. In order to show that $\partial Q(\theta)/\partial\theta$ is differentiable for $\theta \in \mathcal{B}$, we show that $E[\sup_{\theta\in\mathcal{B}} ||\partial^2 L(X,Y,\theta)/\partial\theta\partial\theta||] < \infty$. Direct calculations yield

$$\begin{aligned}\frac{\partial^2 L(X,Y,\theta)}{\partial\theta\partial\theta} &= \begin{bmatrix} \frac{XX'}{s(X'\gamma)} & \frac{XX'}{s(X'\gamma)} s_1(X'\gamma)e(Y,X,\theta) \\ \frac{XX'}{s(X'\gamma)} s_1(X'\gamma)e(Y,X,\theta) & \frac{XX'}{s(X'\gamma)}\{s_1(X'\gamma)e(Y,X,\theta)\}^2 \end{bmatrix} \\ &\quad + \begin{bmatrix} 0_{k\times k} & 0_{k\times k} \\ 0_{k\times k} & -\frac{1}{2}XX's_2(X'\gamma)\{e(Y,X,\theta)^2-1\} \end{bmatrix} \\ &=: \ h_1(X,Y,\theta) + h_2(X,Y,\theta).\end{aligned}$$

We first consider $E[\sup_{\theta\in\mathcal{B}} ||\partial^2 h_1(X,Y,\theta)/\partial\theta\partial\theta||]$. Steps similar to those leading to (A.3) imply that $E[\sup_{\theta\in\mathcal{B}} ||\partial^2 h_1(X,Y,\theta)/\partial\beta\partial\beta||] < \infty$ is satisfied since $E||X||^2 < \infty$ and $\mathcal{B}$ is compact. Moreover, $E[\sup_{\theta\in\mathcal{B}} ||\partial^2 h_1(X,Y,\theta)/\partial\beta\partial\gamma||] < \infty$ and $E[\sup_{\theta\in\mathcal{B}} ||\partial^2 h_1(X,Y,\theta)/\partial\gamma\partial\beta||] < \infty$ are implied by $E[\sup_{\theta\in\mathcal{B}} ||\partial^2 h_1(X,Y,\theta)/\partial\gamma\partial\gamma||] < \infty$.

Steps similar to those leading to (A.6) yield, for $\theta \in \mathcal{B}$,

$$\left\|\frac{XX'}{s(X'\gamma)}\{s_1(X'\gamma)e(Y,X,\theta)\}^2\right\| \le C||X||^2 s_1(X'\gamma)^2[Y^2 + \sup_{\beta\in\mathcal{B}_\beta}||\beta||^2||X||^2].$$

This bound and expansion (A.5) together imply, for some $0 \le \kappa \in (a,\infty)$ and some intermediate value $\bar{\gamma}$,

$$\begin{aligned}\left\|\frac{XX'}{s(X'\gamma)}\{s_1(X'\gamma)e(Y,X,\theta)\}^2\right\| &\le C||X||^2\left[s_1(\kappa)^2 + 2\sup_{\gamma\in\mathcal{B}_\gamma}||\gamma||\,||X||s_1(X'\bar{\gamma})s_2(X'\bar{\gamma})\right] \\ &\quad \times [Y^2 + \sup_{\beta\in\mathcal{B}_\beta}||\beta||^2||X||^2].\end{aligned}$$

Since $\mathcal{B}$ is compact and $0 < s_1(\kappa) < \infty$, $E[\sup_{\theta\in\mathcal{B}} ||\partial^2 h_1(X,Y,\theta)/\partial\gamma\partial\gamma||] < \infty$ requires $E||X||^4 < \infty$, $E[Y^2\,||X||^2] < \infty$ and, for all $\gamma \in \mathcal{B}_\gamma$, $E[||X||^5\, s_1(X'\gamma)s_2(X'\gamma)] < \infty$ and $E\,[Y^2||X||^3 s_1(X'\gamma)s_2(X'\gamma)] < \infty$, which hold under Assumption 3.

We now show that $E[\sup_{\theta\in\mathcal{B}} ||\partial^2 h_2(X,Y,\theta)/\partial\theta\partial\theta||] < \infty$. It suffices to show that $E[\sup_{\theta\in\mathcal{B}} ||\partial^2 h_2(X,Y,\theta)/\partial\gamma\partial\gamma||] < \infty$. Steps similar to those leading to (A.6), yield, for $\theta \in \mathcal{B}$,

(A.7)
$$\left\|XX's_2(X'\gamma)\{e(Y,X,\theta)^2-1\}\right\| \le ||X||^2 s_2(X'\gamma)[1 \vee 2C(Y^2 + \sup_{\beta\in\mathcal{B}_\beta}||\beta||^2||X||^2)]].$$

For $\gamma \in \mathcal{B}_\gamma$ a mean-value expansion about $(\kappa, 0_{k-1})'$ yields

$$|s_2(X'\gamma)| = |s_2(\kappa) + s_3(X'\bar{\gamma})(X'\gamma - \kappa)| \leq s_2(\kappa) + \sup_{\gamma \in \mathcal{B}_\gamma} ||\gamma|| \, ||X|| s_3(X'\bar{\gamma})$$

This bound and (A.7) together imply

$$\begin{aligned}\left\| XX's_2(X'\gamma)\{e(Y,X,\theta)^2 - 1\}\right\| \;\; &\leq \;\; ||X||^2 \left[s_2(\kappa) + \sup_{\gamma \in \mathcal{B}_\gamma} ||\gamma|| \, ||X|| s_3(X'\bar{\gamma})\right] \\ &\qquad \times [1 \vee 2C(Y^2 + \sup_{\beta \in \mathcal{B}_\beta} ||\beta||^2 ||X||^2)].\end{aligned}$$

Since $\mathcal{B}$ is compact and $0 \leq s_2(\kappa) < \infty$, $E[\sup_{\theta \in \mathcal{B}} ||\partial^2 h_2(X,Y,\theta)/\partial\gamma\partial\gamma||] < \infty$ requires $E||X||^4 < \infty$, $E[Y^2 \, ||X||^2] < \infty$ and, for all $\gamma \in \mathcal{B}_\gamma$, $E[\|X\|^5 \, s_3(X'\gamma)] < \infty$ and $E[Y^2||X||^3 s_3(X'\gamma)] < \infty$, which hold under Assumption 3.

We have shown that $E[\sup_{\theta \in \mathcal{B}} ||\partial^2 L(X,Y,\theta)/\partial\theta\partial\theta||] < \infty$ and it follows by Lemma 3.6 in Newey and Mc Fadden (1994) that $\partial Q(\theta)/\partial\theta$ is continuously differentiable over $\mathcal{B}$, and that the order of differentiation and integration can be interchanged.

Letting $H_1(\theta) := E[h_1(X,Y,\theta)]$ and $H_2(\theta) := E[h_2(X,Y,\theta)]$, for all $\theta \in \mathcal{B}$, the Hessian matrix of $Q(\theta)$ is $H(\theta) := H_1(\theta) + H_2(\theta)$, which is positive semidefinite if $H_1(\theta)$ and $H_2(\theta)$ are positive semidefinite (Horn and Johnson, 2012, p.398, 7.1.3. observation). And if either one of $H_1(\theta)$ and $H_2(\theta)$ is positive definite (while the other is positive semidefinite), then $H(\theta)$ is positive definite. All principal minors of $H_2(\theta)$ have determinant 0 for all $\theta \in \mathcal{B}$, and $H_2(\theta)$ is thus positive semidefinite. Applying Lemma 2 with $\Psi(\theta) = H_1(\theta)$, we have that $H_1(\theta)$ is positive definite for all $\theta \in \mathcal{B}$. We conclude that $H(\theta)$ is positive definite for all $\theta \in \mathcal{B}$, and the result follows. □

Lemma 6. *[Level Sets Compactness] If Assumptions 1, 3 and 4 hold then the level sets of $\theta \mapsto Q(\theta)$ are compact.*

Proof. We show that the level sets $\mathcal{B}_c$, $c \in \mathbb{R}$, of $\theta \mapsto Q(\theta)$ are closed and bounded. The result then follows by the Heine-Borel theorem.

Step 1. *[$\mathcal{B}_c$ is bounded].* We show that every sequence in $\mathcal{B}_c$ is bounded. Suppose the contrary. Then there exists an unbounded sequence (θ_n) in $\mathcal{B}_c$, and a subsequence (θ_{n_l}), $n_l \to \infty$ as $l \to \infty$, such that either $||\beta_{n_l}|| \to \infty$ or $||\gamma_{n_l}|| \to \infty$.

Step 1.1. If $||\gamma_{n_l}|| \to \infty$, then $E[XX']$ nonsingular implies that there exists a value $x^* \in \mathcal{X}$ such that $|\gamma_{n_l}' x^*| \to \infty$ as $l \to \infty$, a.s., by Lemma 1, which implies $s(\gamma_{n_l}' x^*) \to 0$ or ∞ by definition of $t \mapsto s(t)$ in Assumption 1.

Moreover, for $x^* \in \mathcal{X}$ such that $|\gamma_{n_l}' x^*| \to \infty$ as $l \to \infty$, we have that $E[(Y - X'\beta)^2 \mid X = x^*] < \infty$ for all $\beta \in \mathbb{R}^k$ under Assumption 3(i). It follows that for all $\beta \in \mathbb{R}^k$,

$$\text{(A.8)} \qquad \lim_{l\to\infty} \tilde{L}(x^*, \beta, \gamma_{n_l}) = \frac{1}{2} \lim_{l\to\infty} \frac{E[(Y - X'\beta)^2 \mid X = x^*]}{s(\gamma_{n_l}' x^*)} + \frac{1}{2} \lim_{l\to\infty} s(\gamma_{n_l}' x^*) = \infty.$$

Since $\tilde{L}(X, \theta)$ is positive a.s. for all $\theta \in \Theta$ and the density $f_X(x)$ is bounded away from 0 for all $x \in \mathcal{X}$ by definition of $\mathcal{X}$, (A.8) implies that $E[\lim_{l\to\infty} \tilde{L}(X, \beta, \gamma_{n_l})] = \infty$. Since $E[\tilde{L}(X, \theta)] = Q(\theta)$, Fatou's lemma then implies that $\lim_{l\to\infty} Q(\beta, \gamma_{n_l}) = \infty$, for all $\beta \in \mathbb{R}^k$. Therefore γ is bounded.

Step 1.2. If $||\beta_{n_l}|| \to \infty$, then $E[XX']$ nonsingular implies that there exists a value $x^{**} \in \mathcal{X}$ such that $|\beta_{n_l}' x^{**}| \to \infty$ a.s., by a second application of Lemma 1. Thus $(Y - \beta_{n_l}' x^{**})^2 \to \infty$ as $l \to \infty$ a.s.

Moreover, for $x^{**} \in \mathcal{X}$ such that $|\beta_{n_l}' x^{**}| \to \infty$ as $l \to \infty$, we have that $E[\lim_{l\to\infty} (Y - X'\beta_{n_l})^2 \mid X = x^{**}] = \infty$, and Fatou's lemma then implies that $\lim_{l\to\infty} E[(Y - X'\beta_{n_l})^2 \mid X = x^{**}] = \infty$. Also, $s(\gamma' x^{**})$ is finite and positive for any $\gamma \in \Theta_\gamma$. It follows that for all $\gamma \in \Theta_\gamma$,

$$\text{(A.9)} \qquad \lim_{l\to\infty} \tilde{L}(x^{**}, \beta_{n_l}, \gamma) \;\; \geq \;\; \frac{1}{2} \lim_{l\to\infty} \frac{E[(Y - X'\beta_{n_l})^2 \mid X = x^{**}]}{s(\gamma' x^{**})} = \infty.$$

Since $\tilde{L}(X, \theta)$ is positive a.s. for all $\theta \in \Theta$ and the density $f_X(x)$ is bounded away from 0 for all $x \in \mathcal{X}$ by definition of $\mathcal{X}$, (A.9) implies that $E[\lim_{l\to\infty} \tilde{L}(X, \beta_{n_l}, \gamma)] = \infty$. Fatou's lemma then implies that $\lim_{l\to\infty} E[\tilde{L}(X, \beta_{n_l}, \gamma)] = \lim_{l\to\infty} Q(\beta_{n_l}, \gamma) = \infty$ for all $\gamma \in \Theta_\gamma$. Therefore β is bounded.

Step 2. *[$\mathcal{B}_c$ is closed]*. We examine the behaviour of $\theta \mapsto Q(\theta)$ on the boundary set $\partial\Theta$ in order to determine whether $\mathcal{B}_c$ is closed. We show that for every sequence in $\mathcal{B}_c$ converging to a boundary point in $\partial\Theta$, $\theta \mapsto Q(\theta)$ is unbounded. Continuity of $\theta \mapsto Q(\theta)$ established in Lemma 3 then implies that $\mathcal{B}_c$ is closed.

Consider a sequence θ_n in $\mathcal{B}_c$ such that $\theta_n \to t_o \in \partial\Theta$ as $n \to \infty$. Then, there exists $x^* \in \mathcal{X}$ such that $\gamma_n' x^* \to 0$ as $n \to \infty$ a.s., by definition of $\partial\Theta$. Moreover, $E[(Y - X'\beta)^2 | X] > 0$ a.s. for all $\beta \in \mathbb{R}^k$ under Assumption 2. Thus

$$\text{(A.10)} \qquad \lim_{n\to\infty} \tilde{L}(x^*, \theta_n) = \frac{1}{2} \lim_{n\to\infty} \frac{E[(Y - X'\beta_n)^2 \mid X = x^*]}{\gamma_n' x^*} = \infty.$$

Since $\tilde{L}(X, \theta)$ is positive a.s. for all $\theta \in \Theta$ and the density $f_X(x)$ is bounded away from 0 for all $x \in \mathcal{X}$ by definition of $\mathcal{X}$, (A.10) implies that $E[\lim_{n\to\infty} \tilde{L}(X, \theta_n)] = \infty$. Fatou's lemma then implies that $\lim_{n\to\infty} E[\tilde{L}(X, \theta_n)] = \lim_{n\to\infty} Q(\theta_n) = \infty$. This yields a contradiction since $Q(\theta_n) \leq c$ for $\theta_n \in \mathcal{B}_c$. Moreover, continuity of $\theta \mapsto Q(\theta)$ implies $Q(t_o) = \lim_{n\to\infty} Q(\theta_n) \leq c$. Therefore, $t_o \in \mathcal{B}_c$ and $\mathcal{B}_c$ is closed. $\square$

Appendix B. Proofs for Sections 2 and 3

B.1. **Proof of Theorem 1.** Under Assumptions 1-4, the order of integration and differentiation for the MVR population problem (2.2) can be interchanged by Lemma 4. Therefore the first-order conditions of problem (2.2) are (2.3)-(2.4), which are satisfied by θ_0. Uniqueness follows from strict convexity of $Q(\theta)$ over compact subsets of Θ established in Lemma 5, and compactness of the level sets of the objective function $Q(\theta)$ established in Lemma 6. $\square$

B.2. **Proof of Proposition 1.** The first-order conditions (2.3)-(2.4) of the MVR population problem (2.2) can be written as

$$E\left[\frac{X}{s(X'\gamma)}(Y - X'\beta)\right] = 0 \tag{B.1}$$

$$E\left[X\frac{s_1(X'\gamma)}{s(X'\gamma)^2}\left\{(Y - X'\beta)^2 - s(X'\gamma)^2\right\}\right] = 0, \tag{B.2}$$

with unique solutions β_0 and γ_0, by Theorem 1.

Under the stated assumptions, the first-order conditions of problem (2.8)-(2.9) are

$$E\left[\frac{X}{\sigma(X)}(Y - X'\beta)\right] = 0 \tag{B.3}$$

$$-4E\left[X\frac{s(X'\gamma)s_1(X'\gamma)}{\sigma(X)^3}\left\{(Y - X'\beta_0)^2 - s(X'\gamma)^2\right\}\right] = 0. \tag{B.4}$$

By assumption the variance of Y conditional on X is correctly specified and $\sigma(X)^2 = s(X'\gamma_0)^2$ a.s. Conditions (B.3)-(B.4) are therefore satisfied for $(\beta, \gamma) = (\beta_0, \gamma_0)$, and are then equivalent to the MVR first-order conditions (B.1)-(B.2) evaluated at the solution $(\beta, \gamma) = (\beta_0, \gamma_0)$. $\square$

B.3. **Proof of Theorem 2.** *Step 1: Existence.* Pick $c \in \mathbb{R}$ such that the level set $\mathcal{B}_c = \{\theta \in \Theta : Q(\theta) \leq c\}$ is nonempty. By Lemma 6, $\mathcal{B}_c$ is compact. Continuity of $\theta \mapsto Q(\theta)$ over compact subsets of Θ established in Lemma 3 then implies that there exists at least one minimizer to $Q(\theta)$ in $\mathcal{B}_c$ by the Weierstrass extreme value theorem. Minimizing $Q(\theta)$ over Θ is equivalent to minimizing $Q(\theta)$ over any of its nonempty level sets, which establishes existence of a minimizer $\theta^* \in \Theta$.

Step 2: Uniqueness. By Lemma 5, we have that $\theta \mapsto Q(\theta)$ is strictly convex over compact subsets of Θ, and thus over $\mathcal{B}_c$, so that $Q(\theta)$ admits at most one minimizer in $\mathcal{B}_c$. This establishes uniqueness of a minimizer $\theta^* \in \Theta$. $\square$

B.4. **Proof of Theorem 3.** *Proof of part (i).* For $\theta \in \Theta$, define the function

$$\widetilde{Q}(\theta) := \frac{1}{2}\int \left\{ \left(\frac{\mu(x) - x'\beta}{s(x'\gamma)} \right)^2 + \frac{\sigma(x)^2}{s(x'\gamma)^2} + 1 \right\} s(x'\gamma) dF_X(x).$$

We show that $\widetilde{Q}(\theta)$ is equal to $Q(\theta)$ for all $\theta \in \Theta$.

The location-scale representation (3.1) for $Y \mid X$ implies that, for $\theta \in \Theta$,

$$\begin{aligned} \left(\frac{Y - X'\beta}{s(X'\gamma)} \right)^2 &= \left(\frac{[\mu(X) - X'\beta] + \sigma(X)\varepsilon}{s(X'\gamma)} \right)^2 \\ &= \left(\frac{\mu(X) - X'\beta}{s(X'\gamma)} \right)^2 + 2\left(\frac{\mu(X) - X'\beta}{s(X'\gamma)} \right) \frac{\sigma(X)}{s(X'\gamma)}\varepsilon + \frac{\sigma(X)^2}{s(X'\gamma)^2}\varepsilon^2, \end{aligned} \tag{B.5}$$

and the change of variable formula

$$f_{Y|X}(Y \mid X) = f_{\varepsilon|X}\left(\frac{Y - \mu(X)}{\sigma(X)} \mid X \right) \left(\frac{1}{\sigma(X)} \right), \tag{B.6}$$

hold a.s.

The definition of $Q(\theta)$ and expressions (B.5)-(B.6) together imply, for $\theta \in \Theta$,

$$\begin{aligned} Q(\theta) &= \frac{1}{2}\int \left\{ \left(\frac{y - x'\beta}{s(x'\gamma)} \right)^2 + 1 \right\} s(x'\gamma) f_{Y|X}(y \mid x) dy dF_X(x) \\ &= \frac{1}{2}\int \left\{ \left(\frac{y - x'\beta}{s(x'\gamma)} \right)^2 + 1 \right\} s(x'\gamma) f_{\varepsilon|X}\left(\frac{y - \mu(x)}{\sigma(x)} \mid x \right) \left(\frac{1}{\sigma(x)} \right) dy dF_X(x) \\ &= \frac{1}{2}\int \left\{ \left(\frac{\mu(x) - x'\beta}{s(x'\gamma)} \right)^2 + 2\left(\frac{\mu(x) - x'\beta}{s(x'\gamma)} \right) \frac{\sigma(x)}{s(x'\gamma)} e + \frac{\sigma(x)^2}{s(x'\gamma)^2} e^2 + 1 \right\} \\ &\quad \times s(x'\gamma) f_{\varepsilon|X}(e \mid x) \, de dF_X(x) \\ &= \frac{1}{2}\int \left\{ \left(\frac{\mu(x) - x'\beta}{s(x'\gamma)} \right)^2 + \frac{\sigma(x)^2}{s(x'\gamma)^2} + 1 \right\} s(x'\gamma) dF_X(x) = \widetilde{Q}(\theta), \end{aligned}$$

where the final step uses the law of iterated expectations and the mean zero and unit variance property of ε conditional on X. Since θ^* is the unique minimizer of $Q(\theta)$ in Θ, it is also the unique minimizer of $\widetilde{Q}(\theta)$ in Θ. $\square$

Proof of part (ii). Since $\Theta_{\rm LS} \subset \Theta$, and θ^* and $\theta_{\rm LS}$ are the unique minimizers of $Q(\theta)$ over Θ and $\Theta_{\rm LS}$, respectively, it follows that $Q(\theta^*) \leq Q(\theta_{\rm LS})$. □

B.5. **Proof of Corollary 1.** For the linear scale specification $s(t) = t$, conditions (2.4) imply $E[(X'\gamma^*)e(Y,X,\theta^*)^2] = E[X'\gamma^*]$. For the exponential scale specification $s(t) = \exp(t)$, because X includes an intercept, conditions (2.4) imply that $E[\exp(X'\gamma^*)e(Y,X,\theta^*)^2] = E[\exp(X'\gamma^*)]$. It follows that for the linear and exponential scale specifications, $E[s(X'\gamma^*)e(Y,X,\theta^*)^2] = E[s(X'\gamma^*)]$, and

$$Q(\theta^*) = \frac{1}{2}E[\{e(Y,X,\theta^*)^2 + 1\}s(X'\gamma^*)] = E[s(X'\gamma^*)].$$

We have shown that for the linear and exponential scale specifications, $Q(\theta^*) = E[e(Y,X,\theta^*)^2 s(X'\gamma^*)]$. For OLS, conditions (2.4) simplify to $E[e(Y,X,\theta_{\rm LS})^2 - 1] = 0$, which implies that $s(\gamma_{\rm LS}) = E[(Y - X'\beta_{\rm LS})^2]^{1/2}$, and

$$Q(\theta_{\rm LS}) = \frac{1}{2}E[\{e(Y,X,\theta_{\rm LS})^2 + 1\}s(\gamma_{\rm LS})] = s(\gamma_{\rm LS}).$$

We have shown that for the constant scale specification $Q(\theta_{\rm LS}) = E[(Y - X'\beta_{\rm LS})^2]^{1/2}$. The result then follows by Theorem 3(ii). □

B.6. **Proof of Theorem 4.** By Theorem 3(ii) we have that $E\left[s(X'\gamma^*)\right] = Q(\theta^*) \leq Q(\theta_{\rm LS}) = \sigma_{\rm LS}$. This and Jensen's inequality for concave functions imply that

$$E\left[\log s(X'\gamma^*)\right] \leq \log E\left[s(X'\gamma^*)\right] \leq \log\left(\sigma_{\rm LS}\right), \tag{B.7}$$

by monotonicity of the logarithmic function. Thus

$$\epsilon = 2\left\{\log\left(\sigma_{\rm LS}\right) - E\left[\log s(X'\gamma^*)\right]\right\} \geq 0.$$

By definition (3.6) of the Gaussian density, inequality (B.7), and $E[e\left(Y,X,\theta_{\rm LS}\right)^2] = 1$, the bound $E\left[e(Y,X,\theta^*)^2\right] \leq 1 + \epsilon$, implies

$$\begin{aligned} -E\left[\log f_{\theta^*}(Y,X)\right] &= \frac{1}{2}\log\left(2\pi\right) + E\left[\log s(X'\gamma^*) + \frac{1}{2}e(Y,X,\theta^*)^2\right] \\ &\leq \frac{1}{2}\left[\log\left(2\pi\right) + 1\right] + \log\left(\sigma_{\rm LS}\right) \\ &= -E\left[\log f_{\theta_{\rm LS}}(Y,X)\right]. \end{aligned} \tag{B.8}$$

Therefore, upon adding $E[\log f_{Y|X}(Y \mid X)]$ to each side of the inequality, we obtain inequality (3.8). □

B.7. **Proof of Corollary 2.** By assumption $\mu(X) = X'\beta_0$ a.s., and by definition θ^* satisfies conditions (3.3)-(3.4). Then the first-order conditions (3.3) are satisfied by $\beta^* = \beta_0$, and γ^* must satisfy

$$E\left[X s_1(X'\gamma^*)\left\{\frac{\sigma(X)^2}{s(X'\gamma^*)^2} - 1\right\}\right] = 0. \tag{B.9}$$

If there exists a pair $(\beta, \gamma) \in \Theta$ satisfying all $2 \times k$ conditions (3.3)-(3.4) simultaneously, then this pair is unique, by strict convexity of $\theta \mapsto Q(\theta)$. It follows from the existence proof of Theorem 2 that the restriction $q(\gamma) := Q(\theta)|_{\beta=\beta_0}$ has a minimizer in Θ_γ, i.e., there exists γ^* such that (B.9) holds. Since $\gamma \mapsto q(\gamma)$ is also strictly convex, $q(\gamma)$ admits a unique minimizer $\gamma^*(\beta_0)$ in Θ_γ. Therefore, the pair $(\beta_0, \gamma^*(\beta_0))$ is the unique minimizer of $Q(\theta)$ when $\mu(X) = X'\beta_0$ a.s. $\square$

B.8. **Proof of Theorem 5.** Letting $L^1(\mathcal{X}) = \{\psi : \int |\psi(x)|\, f_X(x)dx < \infty\}$, define the set of admissible conditional standard deviation functions as

$$\mathcal{S} = \left\{\psi \in L^1(\mathcal{X}) \,:\, \Pr[\psi(X) > 0] = 1\right\}.$$

We first show that $\sigma(X)$ is a minimizer of the MVR criterion

$$Q_0(\psi) := E\left[\frac{1}{2}\left\{\left(\frac{Y - X'\beta_0}{\psi(X)}\right)^2 + 1\right\}\psi(X)\right], \qquad \psi \in \mathcal{S}.$$

Assuming that the order of differentiation and integration can be interchanged, as in Lemma 4, and setting $\psi_\alpha(X) := \sigma(X) + \alpha\psi(X)$ and $e_\alpha(Y,X) := (Y - X'\beta_0)/\psi_\alpha(X)$, we have

$$\begin{aligned}
\left.\frac{\partial Q_0(\psi_\alpha)}{\partial \alpha}\right|_{\alpha=0} &= \frac{1}{2}\, E\left[2e_\alpha(Y,X)\frac{\partial e_\alpha(Y,X)}{\partial \alpha}\psi_\alpha(X) + \frac{\partial \psi_\alpha(X)}{\partial \alpha}e_\alpha(Y,X)^2 + \frac{\partial \psi_\alpha(X)}{\partial \alpha}\right]\Bigg|_{\alpha=0} \\
&= \frac{1}{2}\, E\left[-\frac{\partial \psi_\alpha(X)}{\partial \alpha}\left\{e_\alpha(Y,X)^2 - 1\right\}\right]\Bigg|_{\alpha=0},
\end{aligned}$$

upon using that $\partial e_\alpha(Y,X)/\partial\alpha = -e_\alpha(Y,X)\{\partial\psi_\alpha(X)/\partial\alpha\}\{\psi_\alpha(X)\}^{-1}$. Therefore

$$\left.\frac{\partial Q_0(\psi_\alpha)}{\partial \alpha}\right|_{\alpha=0} = \frac{1}{2}E\left[-\psi(X)\{e(Y,X)^2 - 1\}\right] = 0, \qquad e(Y,X) := \frac{Y - X'\beta_0}{\sigma(X)},$$

for each $\psi \in \mathcal{S}$ since $E\left[e(Y,X)^2 \mid X\right] = 1$ implies that $E\left[h(X)\left\{e(Y,X)^2 - 1\right\}\right] = 0$ for any measurable function of X, and in particular each $\psi \in \mathcal{S}$.

We next show that $Q_0(\psi)$ is strictly convex and thus admits at most one minimizer. Assuming that the order of differentiation and integration can be interchanged, as in

Lemma 5, we have

$$\left.\frac{\partial^2 Q_0(\psi_\alpha)}{\partial\alpha^2}\right|_{\alpha=0} = \frac{1}{2}\,E\left[-2\psi(X)e_\alpha(Y,X)\frac{\partial e_\alpha(Y,X)}{\partial\alpha}\right]\Bigg|_{\alpha=0}$$
$$= E\left[\frac{\psi(X)^2}{\psi_\alpha(X)}e_\alpha(Y,X)^2\right]\Bigg|_{\alpha=0}$$
$$= E\left[\frac{\psi(X)^2}{\sigma(X)}e(Y,X)^2\right]$$
$$= E\left[\frac{\psi(X)^2}{\sigma(X)}\right] > 0,$$

for each $\psi \in \mathcal{S}$. Therefore $\sigma(X)$ is the unique minimizer of $Q_0(\psi)$ in $\mathcal{S}$.

Under correct specification of the CMF, we have that $\beta^* = \beta_0$, by Corollary 2. By $s(X'\gamma^*) \in \mathcal{S}$ and $\sigma(X)$ being the unique minimizer of $Q_0(\psi)$ in $\mathcal{S}$ we have that $Q_0(\sigma) \leq Q(\theta^*)$. By $Q_0(\sigma) = E\left[\sigma(X)\right]$ and $Q(\theta^*) = E\left[s(X'\gamma^*)\right]$, we have shown that $E\left[\sigma(X)\right] \leq E\left[s(X'\gamma^*)\right]$. This and Jensen's inequality for concave functions imply

$$E\left[\log\sigma(X)\right] \leq \log E\left[\sigma(X)\right] \leq \log E\left[s(X'\gamma^*)\right], \tag{B.10}$$

by monotonicity of the logarithmic function.

By iterated expectations and using that $z \mapsto z - 1 - \log z$ is a positive-valued strictly convex function for $z \geq 0$ gives the bound

(B.11)
$$E\left[e(Y,X,\theta^*)^2\right] = E\left[\left(\frac{\sigma(X)}{s(X'\gamma^*)}\right)^2\right] \geq 1 + 2\left\{E\left[\log\sigma(X)\right] - E\left[\log s(X'\gamma^*)\right]\right\}.$$

By definition of the Gaussian density, inequality (B.10), and $E[e\left(Y,X\right)^2] = 1$, bound (B.11) implies

$$\begin{aligned}
-E\left[\log f_{\theta^*}(Y,X)\right] &= \frac{1}{2}\log\left(2\pi\right) + E\left[\log s(X'\gamma^*) + \frac{1}{2}e(Y,X,\theta^*)^2\right] \\
&\geq \frac{1}{2}\left[\log\left(2\pi\right) + 1\right] + E\left[\log\sigma(X)\right] \\
&= -E\left[\log f^\dagger_{\beta_0}(Y,X)\right]. \qquad \text{(B.12)}
\end{aligned}$$

Upon combining (B.8) and (B.12) we obtain inequality (3.11). □

Lemma 7. *Suppose that Assumptions 1, 3 and 5 holds. Then $Q_n(\theta)$ is strictly convex over $\mathcal{B}$.*

Proof. For $e_i = e(y_i, x_i, \theta)$, the Hessian matrix $H_n(\theta)$ of $Q_n(\theta)$,

$$\begin{aligned} H_n(\theta) &= \frac{1}{n}\sum_{i=1}^{n}\begin{bmatrix} \frac{x_i x_i'}{s(x_i'\gamma)} & \frac{x_i x_i'}{s(x_i'\gamma)} s_1(x_i'\gamma)e_i \\ \frac{x_i x_i'}{s(x_i'\gamma)} s_1(x_i'\gamma)e_i & \frac{x_i x_i'}{s(x_i'\gamma)}\{s_1(x_i'\gamma)e_i\}^2 \end{bmatrix} \\ &\quad + \frac{1}{n}\sum_{i=1}^{n}\begin{bmatrix} 0_{k\times k} & 0_{k\times k} \\ 0_{k\times k} & -\frac{1}{2}x_i x_i' s_2(x_i'\gamma)(e_i^2-1) \end{bmatrix} := H_{1n}(\theta) + H_{2n}(\theta), \end{aligned}$$

defined for $\theta \in \mathcal{B}$, is positive definite. Steps similar to the proof of Lemma 2 show that $H_{1n}(\theta)$ is positive definite for all $\theta \in \mathcal{B}$. Moreover, all principal minors of $H_{2n}(\theta)$ have determinant 0 for all $\theta \in \mathcal{B}$, and $H_{2n}(\theta)$ is thus positive semidefinite. Since $H_n(\theta) = H_{1n}(\theta) + H_{2n}(\theta)$, we conclude that H_n is positive definite for all $\theta \in \mathcal{B}$ (Horn and Johnson, 2012, p.398, 7.1.3. observation), and the result follows. $\square$

C.1. **Proof of Theorem 6.** *Proof of part (i)-(ii).* By Theorem 2, $\theta^* \in \Theta$ is the unique minimizer of $Q(\theta)$, and the identification condition (i) in Theorem 2.7 in Newey and Mc Fadden (1994) is thus verified. Since Θ is convex and open, existence of $\theta^* \in \Theta$ established in Theorem 2, as well as strict convexity of $Q_n(\theta)$ established in Lemma 7 imply that their condition (ii) is satisfied. Finally, since the sample is i.i.d. by Assumption 5, pointwise convergence of $Q_n(\theta)$ to $Q_0(\theta)$ follows from $Q_0(\theta)$ bounded (established in the proof of Lemma 3) and application of Khinchine's law of large numbers. Hence, all conditions of Newey and McFadden's Theorem 2.7 are satisfied. Therefore, there exists $\hat{\theta} \in \Theta$ with probability approaching one, and $\hat{\theta} \to^p \theta^*$. $\square$

Proof of part (iii). The sample MVR solution $\hat{\theta}$ can be equivalently formulated as the Method-of-Moments estimator

$$\hat{\theta} = \arg\min_{\theta\in\Theta}\left[\frac{1}{n}\sum_{i=1}^{n} m(y_i, x_i, \theta)\right]'\left[\frac{1}{n}\sum_{i=1}^{n} m(y_i, x_i, \theta)\right],$$

The asymptotic normality result $n^{1/2}(\hat{\theta} - \theta^*) \xrightarrow{d} N(0, G^{-1}S(G^{-1})')$ then follows from this characterization upon verifying the assumptions of Theorem 3.4 in Newey and Mc Fadden (1994), for instance. Block symmetry of G then implies that $V = G^{-1}SG^{-1}$.

By Theorem 2, θ^* is in the interior of Θ so that their Condition (i) is satisfied. The mapping $\theta \mapsto m(Y, X, \theta)$ is continuously differentiable, by inspection, so that their Condition (ii) is satisfied. By definition, θ^* satisfies $E[m(Y, X, \theta^*)] = 0$, hence the first part of their condition (iii) is satisfied. Moreover, bound (A.6) in the proof of Lemma 4 and the steps below show that $E[||m(Y, X, \theta^*)||^2]$ is finite under Assumption 6, verifying their Condition (iii). Finally, under our assumptions, from the proof of Lemma 5, $E[\sup_{\theta\in\Theta} ||\partial m(Y, X, \theta)/\partial\theta||] = E[\sup_{\theta\in\Theta} ||\partial^2 L(X, Y, \theta)/\partial\theta\partial\theta||]$ is finite and $G = E[\partial m(Y, X, \theta^*)/\partial\theta]$ is nonsingular. Their Conditions (iv) and (v) are satisfied.

If $\mu(X) = X'\beta^*$ a.s., then $E[(Y - X'\beta^*)/s(X'\gamma^*) \mid X] = 0$. Therefore, by iterated expectations, the off-diagonal blocks of G and S simplify to (4.3). If $\mu(X) = X'\beta^*$ a.s. and $\sigma^2(X) = s(X'\gamma^*)^2$ a.s., then $E[e^2 - 1 \mid X] = 0$. Therefore, repeated use of iterated expectations imply

$$E\left[\frac{XX'}{s(X'\gamma^*)}\{s_1(X'\gamma^*)e\}^2\right] - E\left[\frac{1}{2}XX's_2(X'\gamma^*)(e^2-1)\right] = E\left[\frac{XX'}{s(X'\gamma^*)}s_1(X'\gamma^*)^2\right],$$

which yields G_{22} in (4.4), and S_{11} and S_{22} in (4.4).

Application of Theorem 4.5 in Newey and Mc Fadden (1994) implies that $\hat{G}^{-1}\hat{S}\hat{G}^{-1} \to^p G^{-1}SG^{-1}$. Under Assumption 6, steps similar to the proof of Lemma 4 show that for a neighborhood $\mathcal{N}$ of θ^* we have $E[\sup_{\theta\in\mathcal{N}} ||m(Y, X, \theta)||^2] < \infty$: bound (A.6) implies, for all $\theta \in \mathcal{N}$,

$$\begin{aligned}\left\|Xs_1(X'\gamma)\{e(Y, X, \theta)^2 - 1\}\right\|^2 &\leq \left\{C[||X|| + ||X||s_2(X'\bar{\gamma})][Y^2 + ||X||^2]\right\}^2 \\ &\leq 2C[||X||^2 + ||X||^2 s_2(X'\bar{\gamma})^2][Y^4 + ||X||^4],\end{aligned}$$

and $E[\sup_{\theta\in\mathcal{N}} ||m(Y, X, \theta)||^2] < \infty$ requires $E[Y^4||X||^2] < \infty$, $E[||X||^6] < \infty$, and, for all $\gamma \in \Theta_\gamma$, $E[Y^4||X||^2 s_2(X'\gamma)^2] < \infty$, and $E[||X||^6 s_2(X'\gamma)^2] < \infty$. By Holder's inequality $E[Y^4||X||^2] < \infty$ and $E[Y^4||X||^2 s_2(X'\gamma)^2] < \infty$ if $E[Y^6] < \infty$, $E[||X||^6] < \infty$, and $E[||X||^6 s_2(X'\gamma)^6] < \infty$, which hold under Assumption 6. Moreover, $\theta \mapsto m(Y, X, \theta)$ is continuous at θ^* a.s. The result follows. □

References

Acemoglu, D., Johnson, S. and Robinson, J. (2002). Reversal of Fortune: Geography and Institutions in the Making of the Modern World Income Distribution. *Quarterly Journal of Economics* 117, pp. 1231–1294.

Akaike, H. (1973). Information Theory and an Extension of the Likelihood Principle. In *Proceedings of the Second International Symposium of Information Theory*, ed. by B. N. Petrov, and F. Csaki.

Angrist, J. and Krueger, A. (1999). Empirical Strategies in Labor Economics. In *Handbook of Labor Economics*, Vol. 3, pp. 1277–1366. Amsterdam: Elsevier.

Angrist, J. D. and Pischke, J. S. (2008). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University.

Blundell, R., Horowitz, J. and Parey, M. (2012). Measuring the Price Responsiveness of Gasoline Demand: Economic Shape Restrictions and Nonparametric Demand Estimation. *Quantitative Economics* 3, pp. 29–51.

Boyd, S. P. and Vandenberghe, L. (2004). *Convex Optimization*. Cambridge University Press.

Breusch, T. S. and Pagan, A. R. (1979). A Simple Test for Heteroscedasticity and Random Coefficient Variation. *Econometrica* (47, September), pp. 1287–1294.

Cameron, A. C. and Trivedi, P. K. (2005). *Microeconometrics: Methods and Applications*. Cambridge University Press.

Chamberlain, G. (1984). Panel Data. In *Handbook of Econometrics*, Vol. 2, ed. by Z. Griliches and M. Intriligator. Amsterdam: North-Holland, pp. 1247–1318.

Chesher, A. (1989). Hajek Inequalities, Measures of Leverage and the Size of Heteroskedasticity Robust Tests. *Journal of Econometrics* 57, pp. 971–977.

Chesher, A. and Austin, G. (1991). The Finite-Sample Distributions of Heteroskedasticity Robust Wald Statistics. *Journal of Econometrics* 47, pp. 153–173.

Chesher, A. and Jewitt, I. (1987). The Bias of a Heteroskedasticity Consistent Covariance Matrix Estimator. *Econometrica* (55, September), pp. 1217–1222.

Chesher, A. and Spady, R. H. (1991). Asymptotic Expansions of the Information Matrix Test Statistic. *Econometrica* (59, May), pp. 787–815.

Eicker, H. (1963). Asymptotic Normality and Consistency of the Least-Squares Estimators for Families of Linear Regressions. *The Annals of Mathematical Statistics* 34, pp. 447–456.

Eicker, H. (1967). Limit Theorems for Regressions with Unequal and Dependent Errors. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, Vol. 1, pp. 59–82.

Horn, R. A. and Johnson, C. R. (2012). *Matrix Analysis*. 2nd ed., Cambridge University Press.

Huber, P. (1967). The Behavior of Maximum Likelihood Estimates under Nonstandard Conditions. In *Proceedings of the fifth Berkeley symposium on mathematical* *statistics and probability*, Vol. 1, pp. 221–233.

Huber, P. (1981). *Robust Statistics*. Wiley, New York.

Koenker, R. (1981). A Note on Studentizing a Test for Heteroscedasticity. *Journal of Econometrics* 17, pp. 107–112.

Koenker, R. (2005). *Quantile Regression*. Econometric Society Monograph Series, Vol. 38. Cambridge University Press.

MacKinnon, J. G. (2013). Thirty Years of Heteroskedasticity-Robust Inference. In *Recent advances and future directions in causality, prediction, and specification analysis*, pp. 437-461. Springer, New York, NY.

MacKinnon, J. G. and White, H. (1985). Some Heteroskedasticity-Consistent Covariance Matrix Estimators with Improved Finite Sample Properties. *Journal of Econometrics* 29, pp. 305–325.

Newey, W. and Mc Fadden, D. (1994). Large Sample Estimation and Hypothesis Testing. In *Handbook of Econometrics*, Vol. 4, ch. 36, 1st ed., pp. 2111–2245. Amsterdam: Elsevier.

Newey, W. and Stouli, S. (2018). Control variables, discrete instruments, and identification of structural functions. *arXiv preprint arXiv:1809.05706.*

Owen, A. B. (2007). A Robust Hybrid of Lasso and Ridge Regression. *Contemporary Mathematics* 443, pp. 59-72.

R Development Core Team (2017). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

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

Sawa, T. (1978). Information Criteria for Discriminating Among Alternative Regression Models. *Econometrica* 46, pp. 1273–1291.

Schennach, S. M. and Wilhelm, D. (2017). A simple parametric model selection test. *Journal of the American Statistical Association*, 112, pp.1663–1674.

Spady, R. H. and Stouli, S. (2018). Dual Regression. *Biometrika* 105, pp. 1–18.

Tripathi, G. (2006). A Matrix Extension of the Cauchy-Schwarz Inequality. *Economics Letters* 63, pp. 1–3.

Vuong, Q. H. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica*, pp.307–333.

Wang, Y., Lawson, C. L. and Hanson, R. J. (2017). lsei: Solving Least Squares or Quadratic Programming Problems under Equality/Inequality Constraints. R package version 1.2.0.

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

White, H. (1980b). Using Least Squares to Approximate Unknown Regression Functions. *International Economic Review* 21, pp. 149–170.

White, H. (1982). Maximum Likelihood Estimation of Misspecified Models. *Econometrica* (50, January), pp. 1–25.

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

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

Ypma, J., Borchers, H. W. and Eddelbuettel, D. (2018). nloptr: R Interface to NLopt. R package version 1.2.1.

SUPPLEMENTARY MATERIAL FOR "SIMULTANEOUS MEAN-VARIANCE REGRESSION"

RICHARD H. SPADY† AND SAMI STOULI§

1. Summary

This Supplementary Material presents further simulation results and an additional empirical example for "Simultaneous Mean-Variance Regression".

In Section 2 we give additional results for the simulations based on MacKinnon (2013) in order to study further the finite-sample properties of MVR. We compare the finite-sample inference performance of MVR, OLS and WLS when finite-sample corrections are applied to the standard errors of each estimator. For completeness we also replicate all experiments under the assumption that the conditional mean function (CMF) is correctly specified and the variance is misspecified, imposing the simplifications shown in equation (4.3) in the calculation of the MVR standard errors in Theorem 6 of the main text. Overall we find that the favorable theoretical properties of MVR translate into very substantial finite-sample gains over both OLS and WLS in terms of estimation performance and largely improved inference. In summary, for the numerical simulations based on MacKinnon (2013) the main findings are:

- MVR-based inference brings very large improvements relative to inference based on the asymptotic heteroskedasticity-robust standard errors. We find that in the presence of heteroskedasticity rejection probabilities for MVR are much closer to nominal level than those for OLS and for WLS with misspecified weights. These relative gains of MVR remain large when finite-sample corrections of the standard errors are implemented for all estimators.
- MVR achieves the improvements above while simultaneously displaying tighter confidence intervals in all designs for sample sizes large enough. When MVR standard errors are calculated under the assumption that the CMF is correctly

Date: January 2, 2019.
† Nuffield College, Oxford, and Department of Economics, Johns Hopkins University, rspady@jhu.edu.
§ Department of Economics, University of Bristol, s.stouli@bristol.ac.uk.

specified, then average MVR confidence interval lengths across simulations are shorter than their OLS counterpart for all sample sizes and designs. They are also shorter than their WLS counterpart with misspecified weights for sample sizes large enough. The relative gains of MVR are then also larger when finite-sample corrections of the standard errors are implemented for all estimators.

- The precision of MVR estimates in root mean squared error (RMSE) is largely superior to OLS under heteroskedasticity and to WLS with misspecified weights, with lower losses than WLS relative to OLS under homoskedasticity.

In Section 3 we give a full set of additional results for the reversal of fortune application, reporting standard errors with finite-sample adjustments and with and without assuming correct specification of the CMF. We find that in this example MVR standard errors are robust to finite-sample corrections and standard errors assuming correct specification of the CMF tend to be slightly smaller.

In Section 4, we report an additional empirical application to demand for gasoline in the United States, and additional numerical simulations calibrated to this example. In particular, we study the finite-sample approximation properties of MVR by implementing simulations calibrated to the demand for gasoline empirical example with a nonlinear CMF. Overall, all experiments confirm the favorable finite-sample estimation, inference and approximation properties of MVR.

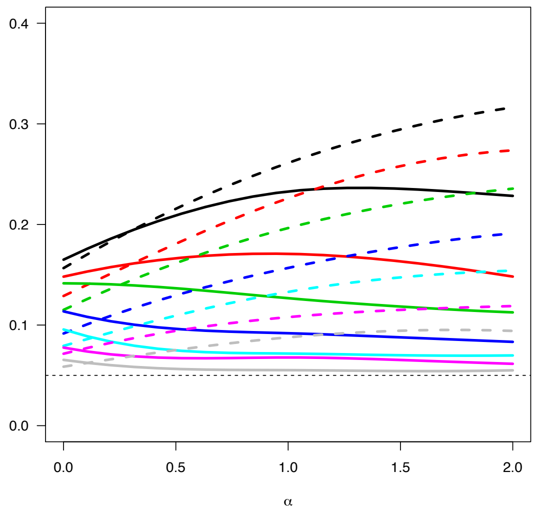
2. Additional Results for the Numerical Simulations Based on MacKinnon (2013)

2.1. **Finite-Sample Corrections.** In this Section we compare rejection probabilities and confidence intervals based on standard errors with finite-sample adjustments proposed by MacKinnon and White (1985), for the numerical simulations in Section 5 of the main text. For an estimator $\tilde{\beta}$ of β, define the sample residuals $\tilde{u}_i := y_i - x_i'\tilde{\beta}$, $i = 1, \dots, n$. The first correction (HC1) uses a degrees-of-freedom adjustment by rescaling the squared sample residuals by the factor $n/(n-k)$ in the calculation of the heteroskedasticity-robust standard errors. The second correction we consider approximates a jackknife estimator (HC3) for the robust variance-covariance matrix, as suggested for small samples by Long and Ervin (2000), for instance, where for each $i = 1, \dots, n$, the ith squared sample residual is rescaled by the factor $1/(1-h_i)$, where h_i is the ith element of the hat matrix $X_n(X_n'X_n)^{-1}X_n'$. The same corrections are applied to the standard errors of OLS, WLS and MVR.

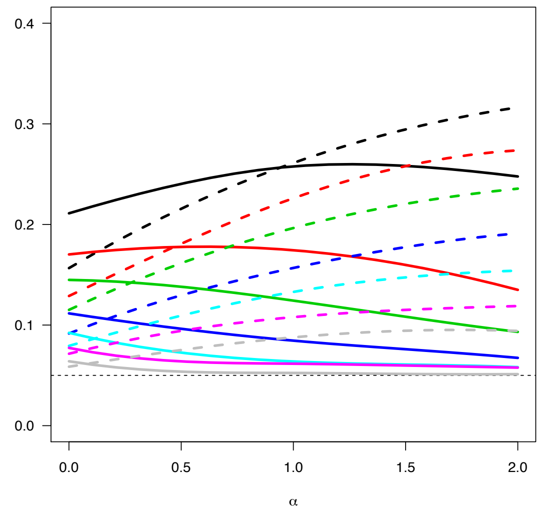
As in the main text we consider rejection probabilities for β_4. Figure 2.1 shows that when degrees-of-freedom corrections (HC1) are implemented, the OLS and WLS rejection probabilities improve but MVR-based inference continues to yield large improvements for all sample sizes in the presence of heteroskedasticity, with the exception of OLS rejection probability for $n = 20$ and $\alpha = 0.5$ which is slightly lower than its e-MVR counterpart (Figure 2.1(b)). Figures 2.2-2.3 compare rejection probability curves when the HC3 correction is implemented. For clarity of representation the curves for MVR and OLS/WLS are shown on different figures and the scale has been modified compared to Figure 2.1. Overall, the rejection probabilities are largely reduced for all estimators, and MVR-based inference continues to yield substantial improvements in the presence of heteroskedasticity relative to both OLS and WLS. The main exceptions are for $n = 20$, and partly for $n = 40$, where the OLS rejection probability curves are not placed above the other curves and show rejection probabilities closer to the nominal level than the corresponding MVR curves (Figures 2.2(b) and (d)).

In order to further investigate the relative performance of MVR-based inference with finite-sample corrections, Tables 1-4 report the ratio of average MVR confidence interval lengths across simulations over the average OLS and WLS confidence interval lengths for β_4 for each sample size and value of heteroskedasticity index α, in percentage terms. Tables 1 and 3 show that the relatively larger average length of the confidence intervals for ℓ-MVR when $n = 20$ in Tables 4-5 in the main text is very much reduced with finite-sample corrections. With HC3 corrections, Tables 2 and 4 show that in the presence of heteroskedasticity shorter average MVR confidence interval lengths across simulations are obtained for all sample sizes and all designs, except for e-MVR with $\alpha = 0.5$. These results show that finite-sample corrections substantially reduce the average MVR confidence interval lengths relative to both OLS and WLS.

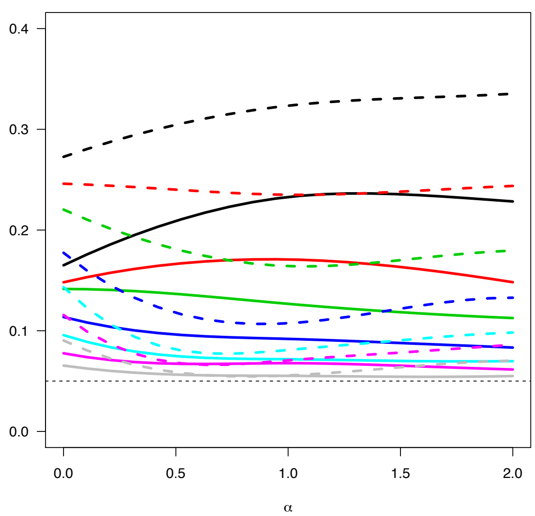
Although the finite-sample corrections for MVR should be regarded as experimental, the simulation results we report indicate that additional improvements relative to OLS and WLS can be achieved and should be explored further in future work.



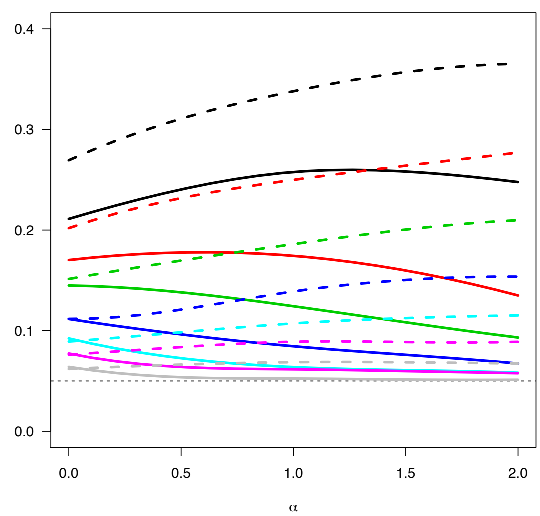
(A) ℓ-MVR vs OLS.



(B) e-MVR vs OLS.

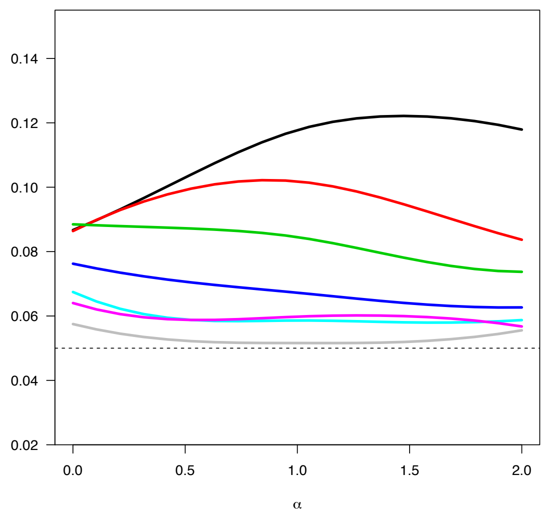


(C) ℓ-MVR vs ℓ-WLS.

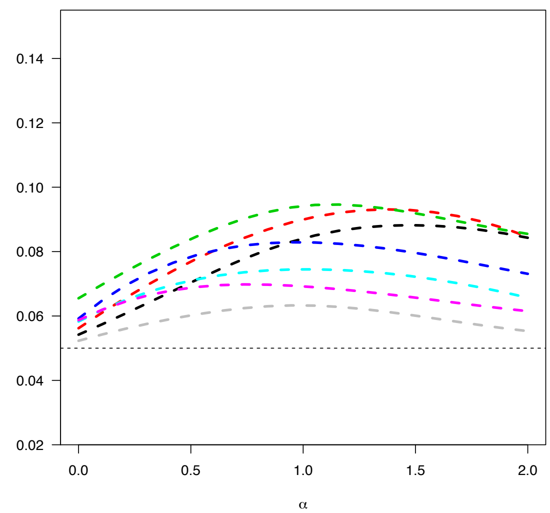


(D) e-MVR vs e-WLS.

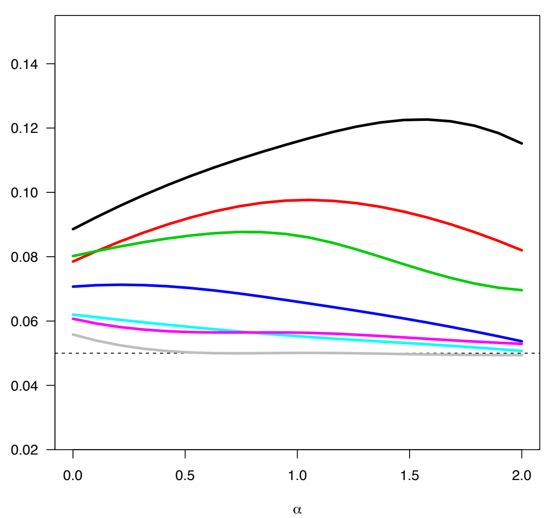
FIGURE 2.1. Rejection frequencies for asymptotic t tests calculated with standard errors with HC1 correction: MVR (solid lines), and OLS and WLS (dashed lines). Sample sizes: 20 (black), 40 (red), 80 (green), 160 (blue), 320 (cyan), 640 (magenta), 1280 (grey).



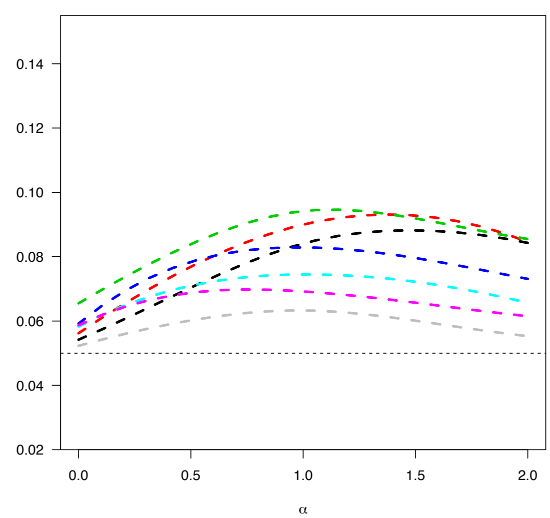
(A) ℓ-MVR.



(B) OLS.

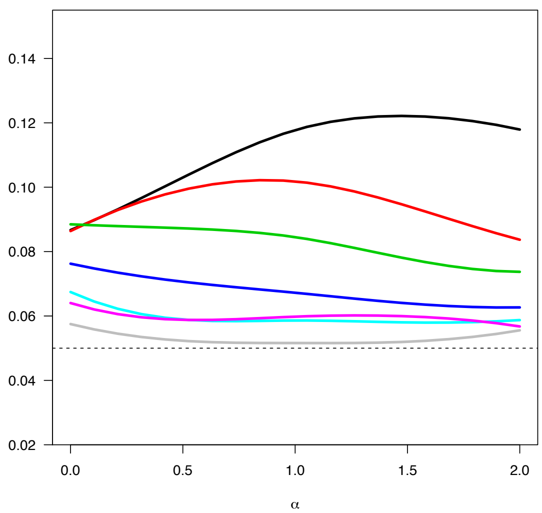


(C) e-MVR.

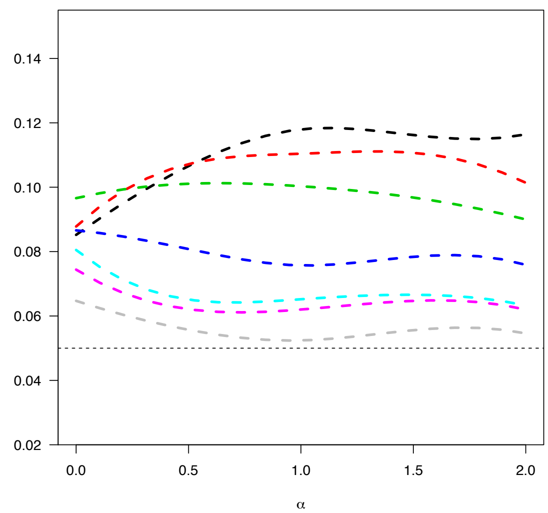


(D) OLS.

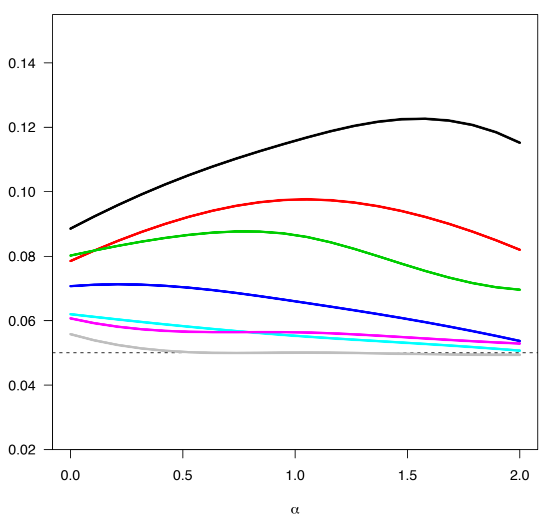
FIGURE 2.2. Rejection frequencies for asymptotic t tests calculated with standard errors with HC3 correction: MVR (solid lines), and OLS (dashed lines). Sample sizes: 20 (black), 40 (red), 80 (green), 160 (blue), 320 (cyan), 640 (magenta), 1280 (grey).



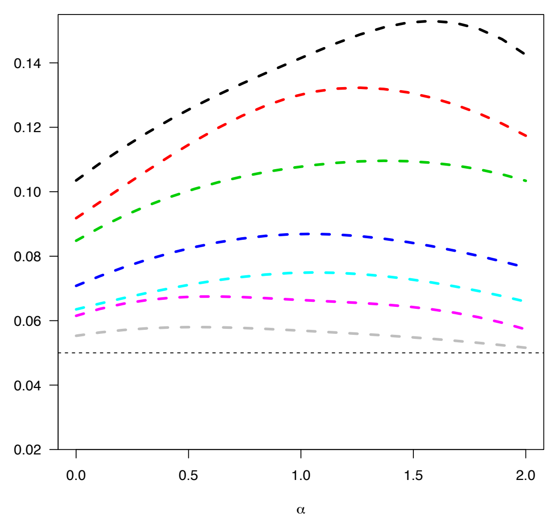
(A) ℓ-MVR.



(B) ℓ-WLS.



(C) e-MVR.



(D) e-WLS.

FIGURE 2.3. Rejection frequencies for asymptotic t tests calculated with standard errors with HC3 correction: MVR (solid lines), and WLS (dashed lines). Sample sizes: 20 (black), 40 (red), 80 (green), 160 (blue), 320 (cyan), 640 (magenta), 1280 (grey).

	ℓ-MVR					e-MVR				
α	0	0.5	1	1.5	2	0	0.5	1	1.5	2
$n = 20$	109.8	111.0	109.4	105.3	100.7	101.7	106.5	107.4	105.2	101.0
$n = 40$	106.0	107.7	103.6	95.6	87.0	101.7	107.0	105.9	99.9	90.8
$n = 80$	102.6	104.3	97.0	85.1	74.1	101.0	105.3	100.0	88.5	73.9
$n = 160$	101.4	101.5	90.0	74.8	63.3	100.7	103.0	92.7	76.3	59.1
$n = 320$	100.8	98.5	83.0	65.5	54.7	100.5	100.0	84.8	65.1	47.3
$n = 640$	100.4	95.2	76.2	57.4	47.4	100.2	96.3	77.0	55.6	38.1
$n = 1280$	100.3	92.1	70.1	50.5	41.1	100.2	93.0	70.3	47.9	31.1

Table 1. Ratio ($\times 100$) of MVR average confidence interval lengths for β_4 over corresponding OLS counterpart. Confidence intervals constructed with standard errors with HC1 correction.

	ℓ-MVR					e-MVR				
α	0	0.5	1	1.5	2	0	0.5	1	1.5	2
$n = 20$	93.5	90.5	84.8	77.9	71.3	93.4	91.1	86.0	79.0	71.5
$n = 40$	97.9	93.9	84.9	74.2	64.2	98.6	94.7	86.0	75.3	63.9
$n = 80$	99.8	94.2	82.2	68.3	56.9	100.0	94.5	82.8	68.4	54.2
$n = 160$	100.5	93.0	77.7	61.5	50.0	100.4	93.3	78.1	60.8	45.1
$n = 320$	100.5	91.3	73.0	55.1	44.3	100.5	91.7	73.2	53.6	37.5
$n = 640$	100.3	89.5	68.7	49.8	39.6	100.3	89.9	68.6	47.7	31.7
$n = 1280$	100.3	87.7	64.4	44.8	35.1	100.2	88.0	64.1	42.2	26.8

Table 2. Ratio ($\times 100$) of MVR average confidence interval lengths for β_4 over corresponding OLS counterpart. Confidence intervals constructed with standard errors with HC3 correction.

	ℓ-MVR					e-MVR				
α	0	0.5	1	1.5	2	0	0.5	1	1.5	2
$n = 20$	131.9	129.5	128.1	126.8	125.3	120.7	125.3	130.5	135.3	137.5
$n = 40$	123.1	118.1	116.3	114.9	111.9	110.3	115.8	122.6	128.6	129.2
$n = 80$	115.9	109.8	109.1	107.0	102.0	105.0	110.0	115.3	118.6	112.8
$n = 160$	111.0	105.4	106.2	101.8	95.7	102.6	106.5	108.4	107.0	96.0
$n = 320$	108.0	103.4	105.4	98.2	91.7	101.3	103.8	101.9	96.5	81.7
$n = 640$	105.3	102.7	105.1	95.3	88.2	100.6	101.1	95.7	87.0	70.1
$n = 1280$	103.4	102.7	104.9	93.0	85.0	100.4	98.9	90.5	79.4	60.9

TABLE 3. Ratio ($\times 100$) of MVR average confidence interval lengths for β_4 over corresponding WLS counterpart. Confidence intervals constructed with standard errors with HC1 correction.

	ℓ-MVR					e-MVR				
α	0	0.5	1	1.5	2	0	0.5	1	1.5	2
$n = 20$	100.7	102.5	103.0	101.6	98.3	105.8	108.6	111.5	112.2	109.0
$n = 40$	101.3	105.3	106.2	103.5	97.7	104.7	107.5	110.1	109.9	103.9
$n = 80$	101.8	105.9	106.7	101.7	93.5	103.0	104.2	105.2	103.0	93.1
$n = 160$	101.7	105.6	106.6	98.6	89.3	101.9	101.2	99.3	94.1	81.0
$n = 320$	101.2	104.9	106.1	96.0	86.6	101.1	99.2	94.6	86.6	70.7
$n = 640$	100.9	104.1	105.7	93.4	82.8	100.6	97.5	90.1	79.6	62.0
$n = 1280$	100.6	103.6	105.3	91.1	79.2	100.4	96.0	86.4	74.2	55.4

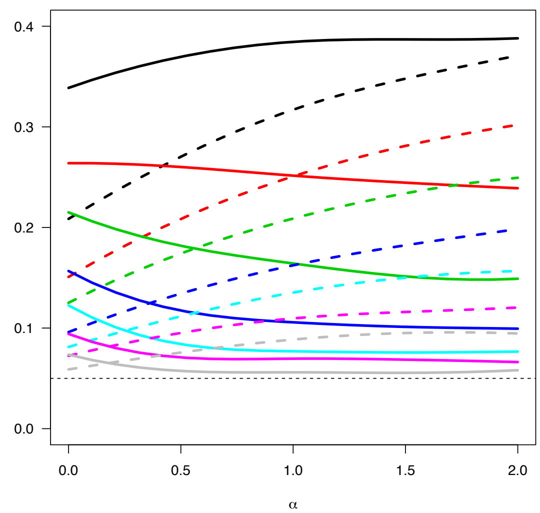
TABLE 4. Ratio ($\times 100$) of MVR average confidence interval lengths for β_4 over corresponding WLS counterpart. Confidence intervals constructed with standard errors with HC3 correction.

2.2. **Inference under Variance Misspecification.** In order to complete this study of the finite-sample performance of MVR inference relative to heteroskedasticity-robust OLS and WLS inference, we also compare the rejection probabilities and the lengths of the confidence intervals when MVR standard errors are calculated under the assumption of correct specification of the CMF, imposing the simplifications shown in equation (4.3) in the main text.

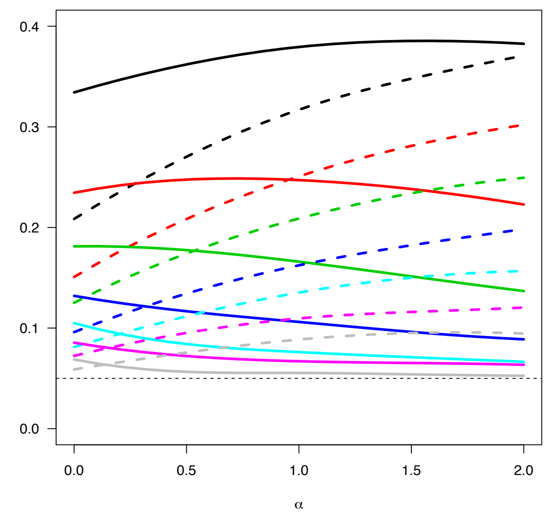
Figure 2.4 shows that the rejection probability curves for ℓ-MVR and $n = 20, 40$ are now placed well above the curves for larger sample sizes. MVR leads to smaller rejection probabilities than OLS for $\alpha \geq 1$ with $n = 40$, and in the presence of heteroskedasticity for all larger sample sizes. The MVR curves are now closer to WLS curves although the overall improvements remain substantial, in particular relative to e-WLS. The finite-sample corrections results in Figures 2.5-2.7 do not alter substantially the main conclusions.

In terms of relative confidence interval lengths, Tables 5-7 show that assuming correct specification of the CMF lead to average MVR confidence interval lengths that are shorter for all sample sizes and designs compared to OLS. The degrees-of-freedom corrections HC1 do not affect the relative length of the confidence intervals. Compared to WLS, Tables 8-10 again show shorter average MVR confidence interval lengths for n large enough and all designs where the conditional variance function is misspecified.

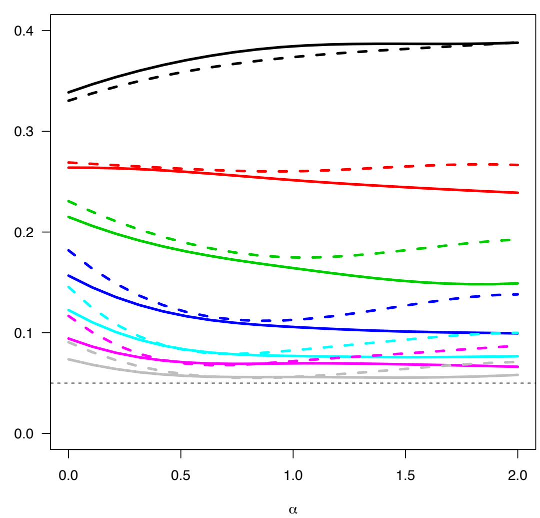
Overall these simulation results under correct specification of the CMF further illustrate the large MVR finite-sample improvements for inference in heteroskedastic designs.



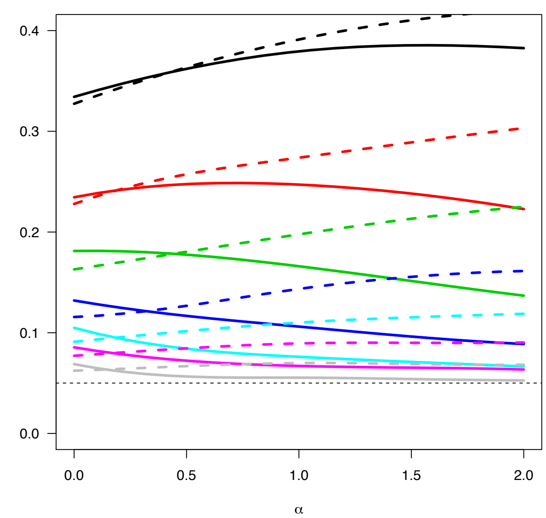
(A) ℓ-MVR vs OLS.



(B) e-MVR vs OLS.

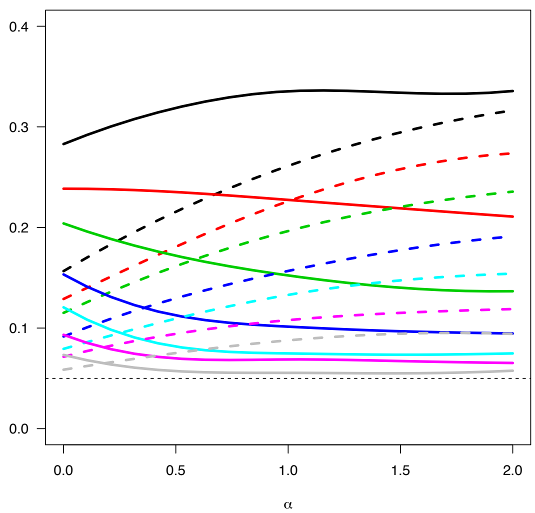


(C) ℓ-MVR vs ℓ-WLS.

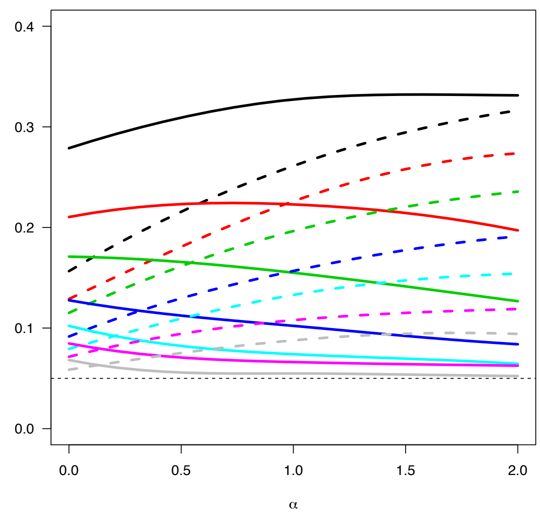


(D) e-MVR vs e-WLS.

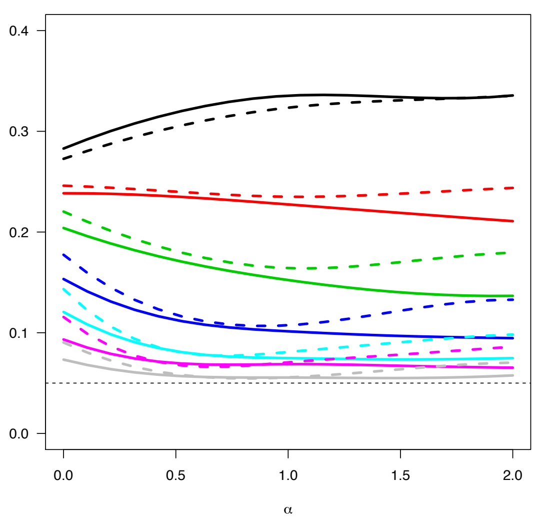
FIGURE 2.4. Rejection frequencies for asymptotic t tests calculated with asymptotic standard errors under correct specification of the CMF: MVR (solid lines), and OLS and WLS (dashed lines). Sample sizes: 20 (black), 40 (red), 80 (green), 160 (blue), 320 (cyan), 640 (magenta), 1280 (grey).



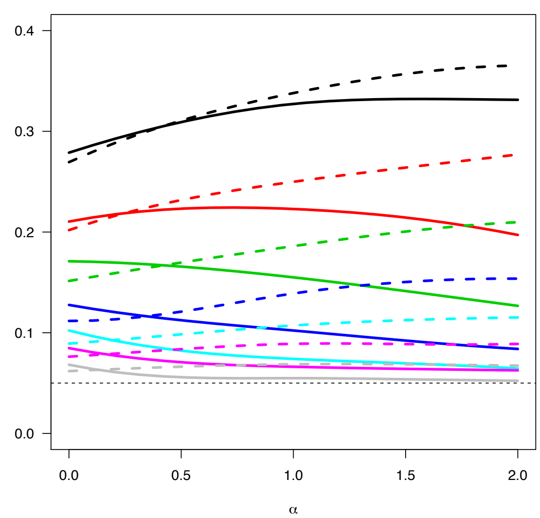
(A) ℓ-MVR vs OLS.



(B) e-MVR vs OLS.

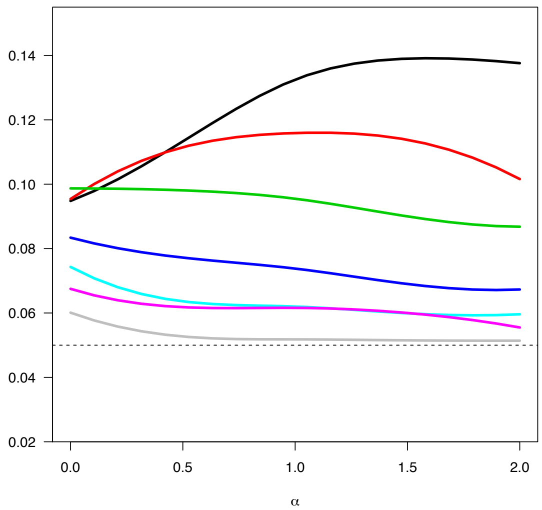


(C) ℓ-MVR vs ℓ-WLS.

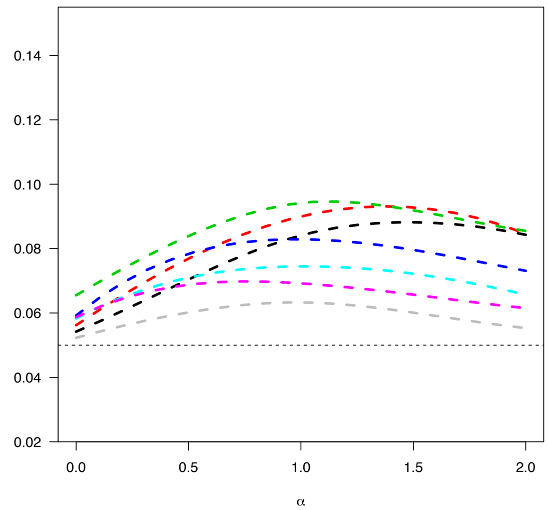


(D) e-MVR vs e-WLS.

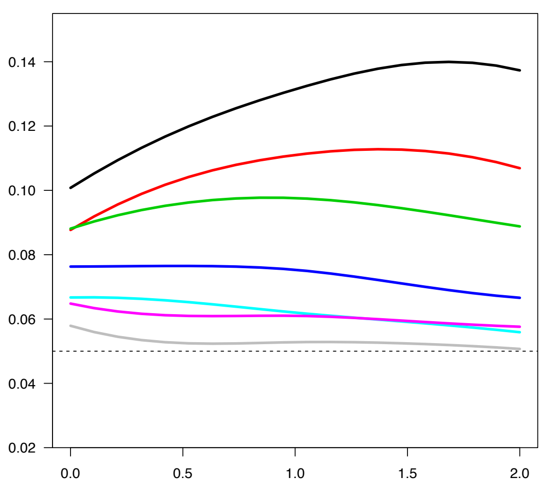
FIGURE 2.5. Rejection frequencies for asymptotic t tests calculated with standard errors with HC1 correction, under correct specification of the CMF: MVR (solid lines), and OLS and WLS (dashed lines). Sample sizes: 20 (black), 40 (red), 80 (green), 160 (blue), 320 (cyan), 640 (magenta), 1280 (grey).



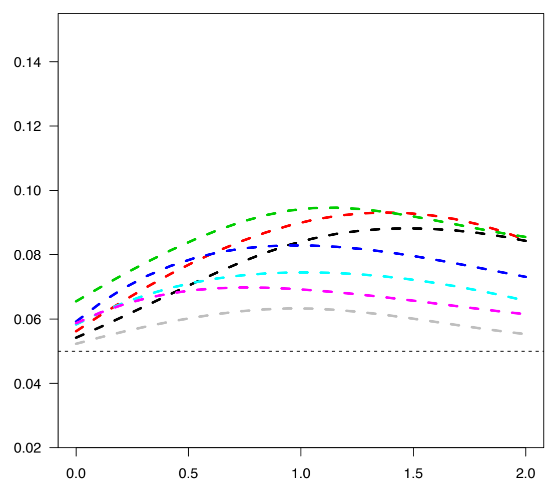
(A) ℓ-MVR.



(B) OLS.

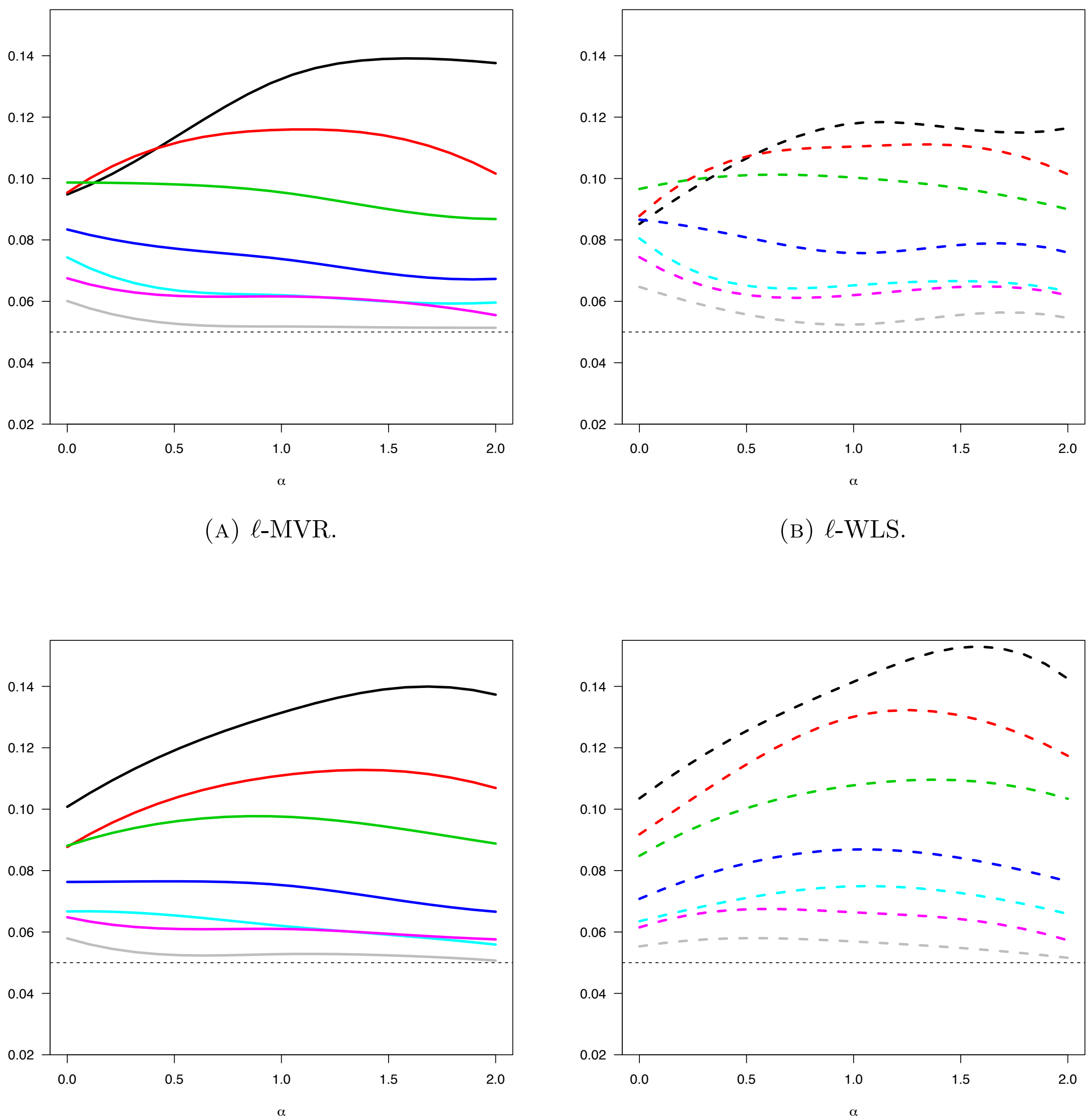


(C) e-MVR.



(D) OLS.

FIGURE 2.6. Rejection frequencies for asymptotic t tests calculated with standard errors with HC3 correction, under correct specification of the CMF: MVR (solid lines), and OLS and WLS (dashed lines). Sample sizes: 20 (black), 40 (red), 80 (green), 160 (blue), 320 (cyan), 640 (magenta), 1280 (grey).



(A) ℓ-MVR.

(B) ℓ-WLS.

(C) e-MVR.

(D) e-WLS.

FIGURE 2.7. Rejection frequencies for asymptotic t tests calculated with standard errors with HC3 correction, under correct specification of the CMF: MVR (solid lines), and OLS and WLS (dashed lines). Sample sizes: 20 (black), 40 (red), 80 (green), 160 (blue), 320 (cyan), 640 (magenta), 1280 (grey).

	ℓ-MVR					e-MVR				
α	0	0.5	1	1.5	2	0	0.5	1	1.5	2
$n = 20$	82.4	84.6	84.1	81.3	77.5	82.3	85.3	84.8	81.6	76.7
$n = 40$	86.6	91.5	90.3	84.4	76.8	89.2	93.0	91.1	85.0	76.1
$n = 80$	90.2	96.0	91.5	81.0	70.1	93.0	96.5	91.4	80.8	67.6
$n = 160$	93.3	97.5	88.0	73.4	61.6	95.4	97.4	87.9	72.7	56.7
$n = 320$	95.4	96.8	82.4	65.1	53.8	96.9	96.6	82.2	63.5	46.4
$n = 640$	97.1	94.4	76.1	57.1	46.6	98.0	94.4	75.7	54.9	37.8
$n = 1280$	98.3	91.8	70.1	50.2	40.4	98.8	91.8	69.6	47.5	30.9

TABLE 5. Ratio ($\times$100) of MVR average confidence interval lengths for β_4 over corresponding OLS counterpart. Confidence intervals constructed with asymptotic standard errors, assuming correct specification of the CMF.

	ℓ-MVR					e-MVR				
α	0	0.5	1	1.5	2	0	0.5	1	1.5	2
$n = 20$	82.4	84.6	84.1	81.3	77.5	82.3	85.3	84.8	81.6	76.7
$n = 40$	86.6	91.5	90.3	84.4	76.8	89.2	93.0	91.1	85.0	76.1
$n = 80$	90.2	96.0	91.5	81.0	70.1	93.0	96.5	91.4	80.8	67.6
$n = 160$	93.3	97.5	88.0	73.4	61.6	95.4	97.4	87.9	72.7	56.7
$n = 320$	95.4	96.8	82.4	65.1	53.8	96.9	96.6	82.2	63.5	46.4
$n = 640$	97.1	94.4	76.1	57.1	46.6	98.0	94.4	75.7	54.9	37.8
$n = 1280$	98.3	91.8	70.1	50.2	40.4	98.8	91.8	69.6	47.5	30.9

TABLE 6. Ratio ($\times$100) of MVR average confidence interval lengths for β_4 over corresponding OLS counterpart. Confidence intervals constructed with standard errors with HC1 correction, assuming correct specification of the CMF.

	ℓ-MVR					e-MVR				
α	0	0.5	1	1.5	2	0	0.5	1	1.5	2
$n = 20$	88.9	85.7	79.7	72.5	65.9	87.3	85.0	79.5	72.3	64.6
$n = 40$	93.1	88.8	80.1	69.5	59.9	93.3	88.9	80.1	69.3	58.0
$n = 80$	95.3	90.3	79.0	65.7	54.8	95.7	89.9	78.3	64.4	50.8
$n = 160$	96.7	90.5	76.1	60.4	49.3	97.1	89.9	75.1	58.5	43.6
$n = 320$	97.7	89.9	72.4	54.8	44.2	98.0	89.3	71.3	52.4	36.8
$n = 640$	98.3	88.9	68.5	49.7	39.8	98.6	88.4	67.5	47.1	31.5
$n = 1280$	98.9	87.4	64.4	44.8	35.5	99.0	87.1	63.5	41.9	26.6

TABLE 7. Ratio ($\times$100) of MVR average confidence interval lengths for β_4 over corresponding OLS counterpart. Confidence intervals constructed with standard errors with HC3 correction, assuming correct specification of the CMF.

α	ℓ-MVR 0	0.5	1	1.5	2	e-MVR 0	0.5	1	1.5	2
$n = 20$	99.0	98.7	98.5	97.9	96.4	97.7	100.3	103.1	104.9	104.5
$n = 40$	100.5	100.4	101.4	101.4	98.7	96.8	100.7	105.5	109.4	108.3
$n = 80$	101.8	101.1	102.9	101.8	96.5	96.7	100.8	105.4	108.2	103.2
$n = 160$	102.2	101.3	103.8	99.9	93.1	97.2	100.8	102.7	102.0	92.2
$n = 320$	102.2	101.5	104.5	97.5	90.2	97.7	100.3	98.8	94.1	80.2
$n = 640$	101.9	101.9	104.9	94.8	86.8	98.4	99.1	94.1	85.9	69.5
$n = 1280$	101.3	102.3	104.9	92.5	83.5	98.9	97.7	89.6	78.8	60.6

TABLE 8. Ratio ($\times 100$) of MVR average confidence interval lengths for β_4 over corresponding WLS counterpart. Confidence intervals constructed with asymptotic standard errors, assuming correct specification of the CMF.

α	ℓ-MVR 0	0.5	1	1.5	2	e-MVR 0	0.5	1	1.5	2
$n = 20$	99.0	98.7	98.5	97.9	96.4	97.7	100.3	103.1	104.9	104.5
$n = 40$	100.5	100.4	101.4	101.4	98.7	96.8	100.7	105.5	109.4	108.3
$n = 80$	101.8	101.1	102.9	101.8	96.5	96.7	100.8	105.4	108.2	103.2
$n = 160$	102.2	101.3	103.8	99.9	93.1	97.2	100.8	102.7	102.0	92.2
$n = 320$	102.2	101.5	104.5	97.5	90.2	97.7	100.3	98.8	94.1	80.2
$n = 640$	101.9	101.9	104.9	94.8	86.8	98.4	99.1	94.1	85.9	69.5
$n = 1280$	101.3	102.3	104.9	92.5	83.5	98.9	97.7	89.6	78.8	60.6

TABLE 9. Ratio ($\times 100$) of MVR average confidence interval lengths for β_4 over corresponding WLS counterpart. Confidence intervals constructed with standard errors with HC1 correction, assuming correct specification of the CMF.

α	ℓ-MVR 0	0.5	1	1.5	2	e-MVR 0	0.5	1	1.5	2
$n = 20$	95.8	97.1	96.9	94.5	90.9	98.8	101.3	103.2	102.7	98.5
$n = 40$	96.3	99.6	100.1	97.0	91.2	99.0	101.0	102.4	101.1	94.3
$n = 80$	97.2	101.5	102.7	97.9	90.1	98.6	99.2	99.5	96.9	87.3
$n = 160$	97.8	102.7	104.5	97.0	88.1	98.5	97.5	95.5	90.6	78.2
$n = 320$	98.4	103.3	105.2	95.5	86.5	98.6	96.7	92.2	84.7	69.4
$n = 640$	98.9	103.3	105.4	93.3	83.4	98.9	95.8	88.8	78.7	61.5
$n = 1280$	99.1	103.2	105.3	91.2	80.0	99.1	94.9	85.6	73.6	55.2

TABLE 10. Ratio ($\times 100$) of MVR average confidence interval lengths for β_4 over corresponding WLS counterpart. Confidence intervals constructed with standard errors with HC3 correction, assuming correct specification of the CMF.

3. Reversal of Fortune: Additional Results

We complement the analysis for the reversal of fortune empirical application in the main text by reporting standard errors with finite-sample corrections (HC1 and HC3 as described in Section 2.1) and results for WLS in Table 11. The exponential scale specification, e-WLS, cannot be used in this empirical application due to several regressors taking value zero for some observations, so that the log transformation cannot be applied to those regressors. Thus we only report the results for WLS with linear scale, ℓ-WLS.

The results in Table 11 further strengthen the main conclusions in the main text with all standard errors increasing slightly when finite-sample corrections are applied, except for MVR when the Americas are dropped (Panel (3)) where the urbanization in 1500 coefficient remains insignificant. For WLS we find that the magnitude of WLS coefficients is smaller than MVR point estimates (except for Panel (3)). In addition to specifications (3), (4), (6) and (9), specification (5) is also found to be not statistically significant with WLS, due to a large drop in the coefficient estimated value relative to both OLS and MVR.

We also report MVR standard errors assuming correct specification of the CMF with and without finite-sample corrections in Table 12. The results confirm that in this example MVR standard errors are robust to finite-sample corrections, and standard errors assuming correct specification of the CMF tend to be slightly smaller.

Overall, we find that our main qualitative conclusions are robust to implementing finite-sample corrections and assuming that the CMF is correctly specified in the calculation of standard errors. Although the numerical simulations in Section 5 of the main text and Section 2 suggest some caution in using ℓ-WLS inference in such small samples, the results in Tables 11-12 provide additional evidence that the relationship between urbanization in 1500 and GDP per capita in 1995 (PPP basis) is weaker and less robust that found using OLS.

	Dependent variable is log GDP per capita (PPP) in 1995											
	OLS	ℓ-WLS	ℓ-MVR	e-MVR	OLS	ℓ-WLS	ℓ-MVR	e-MVR	OLS	ℓ-WLS	ℓ-MVR	e-MVR
	(1) Base sample ($n = 41$)				(2) Without North Africa ($n = 37$)				(3) Without the Americas ($n = 17$)			
Urban. 1500	-0.078	-0.063	-0.067	-0.069	-0.101	-0.097	-0.099	-0.099	-0.115	-0.090	-0.064	-0.077
HC0	0.023	0.021	0.028	0.026	0.032	0.033	0.034	0.034	0.043	0.040	0.127	0.113
HC1	0.023	0.022	0.028	0.027	0.033	0.034	0.035	0.035	0.046	0.043	0.097	0.100
HC3	0.025	0.023	0.029	0.028	0.036	0.037	0.038	0.038	0.056	0.053	0.063	0.082
	(4) Just the Americas ($n = 24$)				(5) With the continent dummies ($n = 41$)				(6) Without neo-Europes ($n = 37$)			
Urban. 1500	-0.053	-0.029	-0.045	-0.044	-0.082	-0.034	-0.063	-0.060	-0.046	-0.034	-0.036	-0.038
HC0	0.029	0.032	0.032	0.032	0.031	0.033	0.029	0.030	0.021	0.020	0.023	0.023
HC1	0.030	0.034	0.033	0.033	0.033	0.035	0.028	0.031	0.022	0.021	0.024	0.024
HC3	0.033	0.044	0.038	0.037	0.035	0.051	0.030	0.029	0.023	0.022	0.025	0.025
	(7) Controlling for Latitude ($n = 41$)				(8) Controlling for colonial origin ($n = 41$)				(9) Controlling for religion ($n = 41$)			
Urban. 1500	-0.072	-0.067	-0.069	-0.070	-0.071	-0.056	-0.063	-0.062	-0.060	-0.015	-0.042	-0.040
HC0	0.020	0.017	0.022	0.021	0.025	0.022	0.026	0.027	0.027	0.029	0.029	0.029
HC1	0.021	0.018	0.022	0.021	0.026	0.023	0.027	0.028	0.029	0.031	0.031	0.030
HC3	0.022	0.019	0.023	0.022	0.028	0.026	0.028	0.029	0.032	0.042	0.034	0.034

TABLE 11. Reversal of fortune. Asymptotic heteroskedasticity-robust OLS, ℓ-WLS and MVR standard errors (HC0), and with finite-sample corrections (HC1 and HC3).

	Dependent variable is log GDP per capita (PPP) in 1995											
	OLS	ℓ-WLS	ℓ-MVR	e-MVR	OLS	ℓ-WLS	ℓ-MVR	e-MVR	OLS	ℓ-WLS	ℓ-MVR	e-MVR
	(1) Base sample ($n = 41$)				(2) Without North Africa ($n = 37$)				(3) Without the Americas ($n = 17$)			
Urban. 1500	-0.078	-0.063	-0.067	-0.069	-0.101	-0.097	-0.099	-0.099	-0.115	-0.090	-0.064	-0.077
HC0	0.023	0.021	0.022	0.022	0.032	0.033	0.033	0.033	0.043	0.040	0.035	0.039
HC1	0.023	0.022	0.022	0.023	0.033	0.034	0.034	0.034	0.046	0.043	0.038	0.041
HC3	0.025	0.023	0.024	0.024	0.036	0.037	0.037	0.037	0.056	0.053	0.051	0.051
	(4) Just the Americas ($n = 24$)				(5) With the continent dummies ($n = 41$)				(6) Without neo-Europes ($n = 37$)			
Urban. 1500	-0.053	-0.029	-0.045	-0.044	-0.082	-0.034	-0.063	-0.060	-0.046	-0.034	-0.036	-0.038
HC0	0.029	0.032	0.030	0.030	0.031	0.033	0.025	0.023	0.021	0.020	0.020	0.021
HC1	0.030	0.034	0.031	0.031	0.033	0.035	0.027	0.025	0.022	0.021	0.021	0.021
HC3	0.033	0.044	0.036	0.035	0.035	0.051	0.029	0.027	0.023	0.022	0.022	0.023
	(7) Controlling for Latitude ($n = 41$)				(8) Controlling for colonial origin ($n = 41$)				(9) Controlling for religion ($n = 41$)			
Urban. 1500	-0.072	-0.067	-0.069	-0.070	-0.071	-0.056	-0.063	-0.062	-0.060	-0.015	-0.042	-0.040
HC0	0.020	0.017	0.018	0.019	0.025	0.022	0.021	0.022	0.027	0.029	0.025	0.026
HC1	0.021	0.018	0.019	0.020	0.026	0.023	0.023	0.023	0.029	0.031	0.027	0.028
HC3	0.022	0.019	0.020	0.021	0.028	0.026	0.025	0.025	0.032	0.042	0.030	0.031

TABLE 12. Reversal of fortune. Asymptotic heteroskedasticity-robust OLS, ℓ-WLS and MVR standard errors assuming correct specification of the CMF (HC0), and with finite-sample corrections (HC1 and HC3).

4. Demand for Gasoline in the United States

4.1. **Empirical Application.** To illustrate our methods further, we consider a second empirical application to the parametric approximation of demand for gasoline in the United States. We use the same data set as in Blundell, Horowitz and Parey (2012), which comes from the 2001 National Household Travel Survey, conducted between March 2001 and May 2002[1]. Blundell, Horowitz and Parey (2012) perform both parametric and nonparametric estimation of the average demand function, and provide evidence of nonlinearities. The data set for their main specifications is large, with a sample of 5254 individual households, and contains household level variables, including gasoline price and consumption, and demographic characteristics. We use these features of the data set to compare the approximation properties of MVR and OLS , to implement our inference methods under misspecification and to calibrate our numerical simulations.

We consider an MVR approximation for the demand for gasoline function

$$Y = \beta_0 + X_1\beta_1 + X_2\beta_2 + X_3'\beta_3 + s(\gamma_0 + X_1\gamma_1 + X_2\gamma_2 + X_3'\gamma_3)e,$$

where e satisfies the orthogonality conditions $E[Xe] = 0$ and $E[Xs_1(X'\gamma)(e^2-1)] = 0$, with $X = (1, X_1, X_2, X_3')'$ and $\gamma = (\gamma_0, \gamma_1, \gamma_2, \gamma_3')'$. We take the outcome Y to be log gasoline annual consumption in gallons, X_1 is log average price in dollars per gallon in county of residence, and X_2 is log income in dollars with each household assigned to 1 of 18 income groups. Following Blundell, Horowitz and Parey (2012), the baseline specification only includes log price and log income, and further covariates are added in other specifications. The vector of additional controls X_3 includes the log of age of household respondent, household size, number of drivers and workers in the household (specification (2)), as well as a dummy for public transport availability (specification (3)), 4 urbanity dummies (specification (4)), 8 population density dummies and 9 regional dummies (specification (5)).

Table 13 reports estimates and standard errors for the average price and income elasticities obtained by OLS, ℓ-MVR and e-MVR across the 5 linear specifications. In the baseline specification, MVR price elasticities are -0.89 and exactly coincide with the average price elasticity found by Yatchew and No (2001) and West (2004), and differ slightly from the OLS point estimate -0.93 in this sample. For specifications (1)-(4), MVR price elasticities are slightly smaller than OLS estimates, and the price

[1]See Blundell, Horowitz and Parey (2012) and ONRL (2004) for a detailed description of the data.

Dependent variable is log of annual household gasoline demand in gallons					
Log price coefficient $\hat{\beta}_1$			Log income coefficient $\hat{\beta}_2$		
OLS	ℓ-MVR	e-MVR	OLS	ℓ-MVR	e-MVR
(1) Baseline specification					
-0.925	-0.892	-0.888	0.289	0.283	0.283
(0.150)	(0.144)	(0.144)	(0.0190)	(0.0173)	(0.0172)
(2) With demographics					
-0.879	-0.857	-0.854	0.246	0.244	0.244
(0.143)	(0.137)	(0.137)	(0.0183)	(0.0169)	(0.0167)
(3) With demographics and public transports					
-0.830	-0.820	-0.816	0.269	0.268	0.268
(0.143)	(0.137)	(0.137)	(0.0187)	(0.0172)	(0.0171)
(4) With demographics, public transports and urbanity					
-0.495	-0.483	-0.478	0.298	0.301	0.301
(0.141)	(0.135)	(0.134)	(0.0190)	(0.0174)	(0.0173)
(5) With demographics, public transports, urbanity and regions					
-0.358	-0.415	-0.408	0.297	0.302	0.302
(0.270)	(0.256)	(0.256)	(0.0199)	(0.0181)	(0.0181)

TABLE 13. Demand for gasoline. Asymptotic heteroskedasticity-robust OLS standard errors and MVR standard errors are in parenthesis.

elasticity drops sharply in specification (4) which adds indicators for urbanity and population density. Adding regional dummies (Panel (5)) results in a further reduction in price elasticities and a loss of significance, although to a much smaller extent for MVR estimates[2]. Given the large sample size, it is interesting to note that for all specifications MVR and OLS standard errors still differ, with MVR standard errors

[2]The p-values for price elasticities increase to 0.185 for OLS and to 0.105 and 0.111 for MVR estimates.

smaller than heteroskedasticity-corrected OLS standard errors, which is a reflection of the heteroskedasticity detected for all specifications[3].

4.2. **Numerical Simulations.** We assess and illustrate the finite-sample properties of our estimators in a Monte Carlo experiment calibrated to our second empirical example. Our models feature a linear CMF, and we implement OLS and MVR with linear and exponential scale functions.

The explanatory variables included in the simulations are chosen according to specification (4) in the demand for gasoline example, the preferred linear specification in Blundell, Horowitz and Parey (2012) (the log price coefficient is no longer significant in specification (5)). We report estimation and inference simulation results for log price and log income, but include all covariates in the simulations. All designs are calibrated to specification (4) by Gaussian maximum likelihood.

Design LOC. Our first design is the homoskedastic model

$$Y = \beta_0 + X_1\beta_1 + X_2\beta_2 + X_3'\beta_3 + \sigma\varepsilon, \quad \varepsilon \sim \mathcal{N}(0,1).$$

Design LIN. Our second design is a set of heteroskedastic models with linear-polynomial scale functions

$$Y = \beta_0 + X_1\beta_1 + X_2\beta_2 + X_3'\beta_3 + (X'\gamma)^\alpha\varepsilon, \quad \varepsilon \sim \mathcal{N}(0,1), \quad \alpha \in \{0.5, 1, 1.5, 2\}.$$

Design EXP. Our third design is a set of heteroskedastic models with exponential-polynomial scale functions

$$Y = \beta_0 + X_1\beta_1 + X_2\beta_2 + X_3'\beta_3 + \exp(X'\gamma)^\alpha\varepsilon, \quad \varepsilon \sim \mathcal{N}(0,1), \quad \alpha \in \{0.5, 1, 1.5, 2\}.$$

For all experiments, we set the sample size to $n = 500, 1000$, and 5254, the sample size in the empirical application, and 5000 simulations are performed. For $n = 5254$, we fix X to the values in the data set, whereas for the smaller sample sizes we draw X with replacement from the values in the data set and keep them fixed across replications. The location design LOC serves as a benchmark for comparing the relative performance of MVR and OLS when OLS is efficient. For $\alpha = 1$, ℓ-MVR is correctly specified for the design LIN, and e-MVR is correctly specified for design EXP. Designs with $\alpha = 0.5$ feature low heteroskedasticity, whereas $\alpha = 2$ corresponds to high heteroskedasticity.

[3]For each specification we implemented the tests of Breusch and Pagan (1979), White (1980) and Koenker (1981)) for heteroskedasticity for OLS and the MVR-based test introduced in Section 4 of the main text. All tests reject the null of homoskedasticity for all specifications.

	Design	LOC	LIN				EXP			
	α	0	0.5	1	1.5	2	0.5	1	1.5	2
			Log price coefficient $\hat{\beta}_1$							
ℓ-MVR	$n = 500$	102.1	100.5	96.0	89.0	80.6	100.6	96.1	89.0	80.3
	$n = 1000$	100.5	98.7	93.6	85.4	75.0	98.8	93.6	85.3	75.2
	$n = 5254$	100.1	98.5	93.4	85.2	74.5	98.5	93.4	84.9	73.8
e-MVR	$n = 500$	102.1	100.4	95.9	89.1	80.5	100.4	95.8	88.7	79.5
	$n = 1000$	100.5	98.7	93.6	85.9	76.4	98.7	93.5	85.3	74.8
	$n = 5254$	100.1	98.5	93.6	86.0	76.5	98.5	93.4	85.1	74.3
			Log income coefficient $\hat{\beta}_2$							
ℓ-MVR	$n = 500$	101.6	99.6	93.7	85.1	75.1	99.4	92.8	82.9	71.2
	$n = 1000$	101.4	98.2	89.4	77.0	63.5	97.4	86.0	70.3	53.9
	$n = 5254$	100.3	96.8	88.1	76.2	63.2	96.0	85.2	70.6	55.3
e-MVR	$n = 500$	101.7	99.8	93.8	84.8	73.9	99.6	92.6	81.9	68.8
	$n = 1000$	101.4	98.3	89.1	75.8	61.1	97.3	84.9	67.4	48.9
	$n = 5254$	100.3	96.7	87.6	74.9	61.0	95.8	84.0	67.7	50.4

Table 14. Ratio ($\times 100$) of MVR RMSE for β_1 and β_2 over corresponding OLS counterpart.

Table 14 reports a first set of results regarding the accuracy of our estimators. We report the ratios of RMSEs for β_1 and β_2 of ℓ-MVR and e-MVR over RMSEs of OLS, in percentage terms. The results show that MVR estimators achieve large gains relative to OLS in the presence of heteroskedasticity, with ratios that reach 73.8 for $\hat{\beta}_1$ and 50.4 for $\hat{\beta}_2$ under heteroskedasticity, with e-MVR outperforming ℓ-MVR slightly in this example. Gains in estimation precision increase with the degree of heteroskedasticity and sample size. In the homoskedastic case where OLS is efficient, there is close to no loss in precision from using MVR, with ratios ranging from 100.1 to 102.1. OLS and MVR become equivalent as sample size increases for the homoskedastic case.

Table 15 reports ratios of ℓ-MVR and e-MVR average confidence interval lengths across simulations for β_1 and β_2 over OLS average confidence interval lengths, in percentage terms. In these simulations MVR yields substantially tighter confidence

	Design	LOC	LIN				EXP			
	α	0	0.5	1	1.5	2	0.5	1	1.5	2
		Log price coefficient $\hat{\beta}_1$								
	$n = 500$	98.8	98.1	96.1	92.8	88.6	98.1	96.1	92.8	88.3
ℓ-MVR	$n = 1000$	99.2	98.3	95.8	91.6	86.0	98.3	95.8	91.6	85.7
	$n = 5254$	99.8	98.9	96.3	91.9	86.0	98.9	96.3	91.8	85.6
	$n = 500$	98.6	97.9	95.8	92.5	88.1	97.9	95.8	92.3	87.5
e-MVR	$n = 1000$	99.1	98.3	95.8	92.0	86.9	98.3	95.8	91.6	85.9
	$n = 5254$	99.8	98.9	96.4	92.4	87.2	98.9	96.3	91.9	85.9
		Log income coefficient $\hat{\beta}_2$								
	$n = 500$	98.9	98.2	95.5	91.3	86.3	98.0	95.0	90.1	84.0
ℓ-MVR	$n = 1000$	99.2	98.0	93.9	87.7	80.1	97.5	92.3	84.1	74.4
	$n = 5254$	99.8	98.3	93.9	87.5	79.9	97.9	92.4	84.3	74.7
	$n = 500$	98.7	97.8	94.9	90.4	84.6	97.7	94.3	88.8	81.6
e-MVR	$n = 1000$	99.1	97.8	93.5	86.8	78.5	97.3	91.5	82.2	70.8
	$n = 5254$	99.8	98.2	93.6	86.8	78.4	97.8	91.7	82.5	71.4

Table 15. Ratio ($\times 100$) of MVR average confidence interval lengths for β_1 and β_2 over corresponding OLS counterpart. Confidence intervals constructed with asymptotic standard errors.

intervals compared to OLS in the presence of heteroskedasticity, with confidence interval lengths ratios that reach 78.4 for $\hat{\beta}_1$ and 70.8 for $\hat{\beta}_2$, while not incurring any loss in precision for the homoskedastic data generating process. The relative performance of e-MVR improves with the degree of heteroskedasticity.

For completeness we also report results for confidence intervals constructed assuming correct specification of the CMF. Table 16 reports ratios of ℓ-MVR and e-MVR average confidence interval lengths across simulations for β_1 and β_2 over OLS average confidence interval lengths, in percentage terms. MVR confidence intervals are slightly more favorable to MVR compared to the results obtained with standard errors robust to mean misspecification reported in Table 15, while not incurring any loss in precision for the homoskedastic data generating process.

	Design	LOC	LIN				EXP			
	α	0	0.5	1	1.5	2	0.5	1	1.5	2
		Log price coefficient $\hat{\beta}_1$								
	$n = 500$	96.8	96.1	93.9	90.5	86.0	96.1	94.0	90.5	85.8
ℓ-MVR	$n = 1000$	98.4	97.6	95.0	90.7	84.7	97.6	95.0	90.6	84.5
	$n = 5254$	99.7	98.8	96.2	91.8	85.8	98.8	96.2	91.7	85.4
	$n = 500$	97.4	96.3	94.2	91.0	86.7	96.3	94.2	90.8	86.1
e-MVR	$n = 1000$	98.6	97.6	95.2	91.4	86.4	97.6	95.2	91.1	85.4
	$n = 5254$	99.7	98.8	96.3	92.4	87.1	98.8	96.2	91.9	85.8
		Log income coefficient $\hat{\beta}_2$								
	$n = 500$	96.5	95.8	93.1	88.9	83.7	95.7	92.7	87.8	81.8
ℓ-MVR	$n = 1000$	98.1	97.1	93.2	87.0	79.5	96.7	91.6	83.6	74.0
	$n = 5254$	99.7	98.2	93.8	87.4	79.8	97.8	92.3	84.2	74.7
	$n = 500$	97.1	95.8	92.9	88.4	82.7	95.6	92.3	86.9	79.9
e-MVR	$n = 1000$	98.6	97.0	92.7	86.0	77.8	96.5	90.7	81.6	70.4
	$n = 5254$	99.7	98.1	93.5	86.7	78.4	97.7	91.7	82.5	71.4

TABLE 16. Ratio ($\times 100$) of MVR average confidence interval lengths for β_1 and β_2 over corresponding OLS counterpart. Confidence intervals constructed with asymptotic standard errors assuming correct specification of the CMF.

4.3. **Additional Simulations: Nonlinear CMF.** We present the results of a second set of experiments in which we compare the approximation properties of MVR to those of OLS under misspecification of the CMF, in RMSE. The designs of our simulations are modified to incorporate a nonlinear relationship between X_1 (log price) and Y (log gasoline annual consumption). We specify the nonlinear relationship in X_1 by means of trigonometric basis functions

$$f(x_1, \delta_1) = \delta_{11} x_1 + \delta_{12} \sin(2\pi x_1) + \delta_{13} \cos(2\pi x_1) + \delta_{14} \sin(4\pi x_1) + \delta_{15} \cos(4\pi x_1).$$

All designs are calibrated to specification (4) by Gaussian maximum likelihood.

Design LOC. Our first design is the homoskedastic model

$$Y = \beta_0 + f(X_1, \beta_1) + X_2\beta_2 + X_3'\beta_3 + \sigma\varepsilon.$$

	Design	LOC	LIN				EXP			
	α	0	0.5	1	1.5	2	0.5	1	1.5	2
	$n = 500$	101.2	100.9	100.0	98.5	96.5	100.9	100.0	98.6	96.5
ℓ-MVR	$n = 1000$	100.7	100.0	97.9	94.6	90.1	100.0	97.9	94.1	88.9
	$n = 5254$	100.1	99.6	97.8	95.0	91.6	99.5	97.5	94.1	89.6
	$n = 500$	100.8	100.6	99.9	98.6	96.8	100.6	99.8	98.3	96.2
e-MVR	$n = 1000$	100.6	99.9	97.8	94.5	90.2	99.9	97.7	93.8	88.3
	$n = 5254$	100.1	99.6	97.8	95.0	91.6	99.5	97.4	93.9	89.1

TABLE 17. Ratio ($\times 100$) of average MVR RMSE for $\mu(x)$ over corresponding OLS counterpart.

Design LIN. Our second design is a set of heteroskedastic models with linear-polynomial scale functions

$$Y = \beta_0 + f(X_1, \beta_1) + X_2\beta_2 + X_3'\beta_3 + s(\gamma_0 + f(X_1, \gamma_1) + X_2\gamma_2 + X_3'\gamma_3)^\alpha \varepsilon, \quad \alpha \in \{0.5, 1, 1.5, 2\}.$$

Design EXP. Our third design is a set of heteroskedastic models with exponential-polynomial scale functions

$$Y = \beta_0 + f(X_1, \beta_1) + X_2\beta_2 + X_3'\beta_3 + s(\gamma_0 + f(X_1, \gamma_1) + X_2\gamma_2 + X_3'\gamma_3)^\alpha \varepsilon, \quad \alpha \in \{0.5, 1, 1.5, 2\}.$$

where $\varepsilon \sim \mathcal{N}(0, 1)$. For all designs we implement MVR and OLS for the same sample sizes and X values as in Section 4.2, with the number of simulations set to 5000.

Table 17 reports results regarding the accuracy of OLS and MVR linear approximations of the $\mu(x, \beta) = \beta_0 + f(x_1, \beta_1) + x_2\beta_2 + x_3'\beta_3$, evaluated at the n sample values x_{1i} of X_1, and at fixed values of the remaining variables.[4] For each data generating process we report the ratios of average estimation errors across simulations of ℓ-MVR and e-MVR relative to OLS in percentage terms. Estimation errors are measured for each simulation in RMSE, and then averaged across simulations.

In these simulations MVR yields more accurate approximation of nonlinear CMFs than OLS, measured in RMSE. Thus, in presence of heteroskedasticity the minimum

[4]The non binary variables X_2, $X_{31}, \ldots X_{34}$, are evaluated at their modal values. These variables are the log of household income, age of household respondent, household size, number of drivers and workers in the household, respectively. We fix the value of the remaining indicators for public transport availability, urbanity and population density included in X_3 to one.

mean squared error OLS property does not necessarily translate into more accurate CMF approximation in finite samples relative to MVR.

References

Blundell, R., Horowitz, J. & Parey, M. (2012). Measuring the Price Responsiveness of Gasoline Demand: Economic Shape Restrictions and Nonparametric Demand Estimation. *Quantitative Economics* 3, pp. 29–51.

Breusch, T. S. and Pagan, A. R. (1979). A Simple Test for Heteroscedasticity and Random Coefficient Variation. *Econometrica* (47, September), pp. 1287–1294.

Koenker, R. (1981). A Note on Studentizing a Test for Heteroscedasticity. *Journal of Econometrics* 17, pp. 107–112.

Long, J. S. and Ervin, L. H. (2000). Using Heteroscedasticity Consistent Standard Errors in the Linear Regression Model. *The American Statistician* 54, pp. 217–224.

MacKinnon, J. G. (2013). Thirty Years of Heteroskedasticity-Robust Inference. In *Recent advances and future directions in causality, prediction, and specification analysis*, pp. 437-461. Springer, New York, NY.

MacKinnon, J. G. and White, H. (1985). Some Heteroskedasticity-Consistent Covariance Matrix Estimators with Improved Finite Sample Properties. *Journal of Econometrics* 29, pp. 305–325.

ONRL (2004). 2001 National Household Travel Survey. User's Guide, Oak Ridge National Laboratory.

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

West, S. (2004). Distributional Effects of Alternative Vehicle Pollution Control Policies. *Journal of Public Economics* 88, pp. 735–757.

Yatchew, A. and No, J. A. (2001). Household Gasoline Demand in Canada. *Econometrica* (69, November), pp. 1697–1709.