cemmap
centre for microdata methods and practice

# Simultaneous mean-variance regression

Richard H. Spady
Sami Stouli

The Institute for Fiscal Studies
Department of Economics, UCL

**cemmap** working paper CWP25/18

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL
An ESRC Research Centre

# 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 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. We illustrate our method with numerical simulations and 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, linear regression, simultaneous approximation, heteroskedasticity, robust inference, misspecification, influence function, convexity, ordinary least-squares, dual regression.

## 1. INTRODUCTION

Ordinary least-squares (OLS) is the method of choice for the linear estimation and approximation of the conditional mean function. 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

*Date*: April 5, 2018. We are grateful 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.

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 (White and MacKinnon (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 in 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 conditional mean functions. 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 four main features. First, it incorporates information from the second conditional moment in the determination of the first conditional moment parameters. Second, simultaneity generates outcome predictions which are guaranteed to improve over OLS prediction error under heteroskedasticity. Third, the resulting approximations have a formal interpretation under misspecification of the shapes of the conditional mean and variance functions. Fourth, corresponding standard errors 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 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 weighted least-squares 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.

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 conditional mean function. OLS gives the minimum mean squared error linear approximation to the conditional mean function, 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 weighted least-squares approximations accounting for potential variation in the outcome across the support of conditioning variables. Our approach thus complements the textbook weighted least-squares 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 weights and further correcting the standard errors for heteroskedasticity.

This paper makes three main contributions. First, we establish 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 show that MVR is a minimum weighted least-squares linear approximation to the conditional mean function, 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. This property provides a theoretical guarantee motivating the use of MVR over OLS, and is not shared by alternative weighted least-squares proposals. Third, we derive the asymptotic distribution of our estimator under misspecification and provide robust inference methods. In particular we propose a test of heteroskedasticity that provides a one-step complement to existing OLS-based tests (Breusch and Pagan (1978), White (1980a), Koenker (1983)).

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. Section 4 gives asymptotic theory. Section 5 reports the results of two empirical applications to the relationship between economic prosperity in 1500 and today, and to demand for gasoline in the United States, and illustrates the finite-sample performance of MVR with numerical simulations calibrated to the demand analysis example. All proofs of the main results are given in the Appendix. The online Appendix Spady and Stouli (2018) contains supplemental material.

## 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 $\beta_0$ and $\gamma_0$ 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 $\beta$. 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 $\theta_0$ 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

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

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

These conditions are satisfied by $\theta_0$ since model (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 $\theta_0$ 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 $C_\theta$ 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) $E[Y^2] < \infty$, $E||X||^4 < \infty$ and $E[Y^2\, ||X||^2] < \infty$, and, (ii) for all $\gamma \in \Theta_\gamma$, $E[\|X\|^5\, s_3(X'\gamma)] < \infty$, $E[\|X\|^5\, s_1(X'\gamma)s_2(X'\gamma)] < \infty$, $E[Y^2||X||^3 s_3(X'\gamma)] < \infty$ and $E[Y^2||X||^3 s_1(X'\gamma)s_2(X'\gamma)] < \infty$.

**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 $\Theta$, 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 $\theta_0 \in int(\Theta^c)$, then $\theta_0$ is the unique minimizer of $Q(\theta)$ over $\Theta$.*

Theorem 1 applies when the conditional mean and variance functions are well-specified, and thus provides primitive conditions for identification of $\theta_0$ in the location-scale model (2.5). Theorem 1 extends the uniqueness result in Spady and Stouli (2018) for location-scale models with linear scale functions 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 $E[\|X\|^5 \exp(X'\gamma)^2]$ and $E\,[Y^2||X||^3 \exp(X'\gamma)^2]$ be finite.

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 $\gamma_0$, weighted least-squares regression of $Y$ on $X$ with weights $1/s(X'\gamma_0)$ has first-order conditions (2.6), with solution $\beta_0$. Moreover, given knowledge of $\beta_0$, nonlinear weighted least-squares 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 $\gamma_0$.

**Proposition 1.** *If Assumptions 1-4 hold, and (i) $E[Y^4] < \infty$, $E[Y^2 s(X'\gamma)^2] < \infty$, $E[\|X\|^2 s(X'\gamma)^2] < \infty$ 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 $\beta$ and $\gamma$ 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 therefore formally establishes the simultaneous mean and variance regression interpretation of the population problem (2.2).

## 3. Approximation properties of MVR under misspecification

Under misspecification, OLS provides the minimum mean squared error linear approximation to the conditional mean function. 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 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 $\Theta$, and is a consequence of the explosive behaviour of the objective function at the boundaries of the parameter space. The MVR objective $Q(\theta)$ is a *coercive* function over the open set $\Theta$, 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 $\Theta$. Thus the MVR criterion is infinity at infinity, and for any sequence of parameter values in $\Theta$ 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 $\Theta$.[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 $\Theta$.

**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 and uniqueness of the MVR location-scale representation

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

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

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 conditional mean function $x \mapsto \mu(x)$ 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 $\Theta_{\gamma,\mathrm{LS}} = \{\gamma \in \mathbb{R} : s(\gamma) > 0\}$, define $\Theta_{\mathrm{LS}} = \mathbb{R}^k \times \Theta_{\gamma,\mathrm{LS}}$. Upon setting $s(X'\gamma) = s(\gamma)$ in the MVR problem (2.2),

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

is a particular case of MVR. Since the OLS solution $\theta_{\mathrm{LS}} := (\beta_{\mathrm{LS}}, \gamma_{\mathrm{LS}}, 0_{k-1})'$ belongs to the parameter space $\Theta$, uniqueness of $\theta^*$ 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*

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

$$E\left[Xs_1(X'\gamma)\left\{\left(\frac{\mu(X)-X'\beta}{s(X'\gamma)}\right)^2+\left(\frac{\sigma(X)^2}{s(X'\gamma)^2}-1\right)\right\}\right] = 0. \tag{3.4}$$

*(ii) The optimal value of problem (2.2) satisfies $Q(\theta^*) \le Q(\theta_{LS})$, 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 conditional mean function 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 weighted least-squares 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) and (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, $\beta^*$ and $\gamma^*$ 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. 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 $\theta^* = \theta_{LS}$.*

This upper bound is a key result which provides a theoretical justification for the favorable finite-sample properties displayed by MVR in the numerical simulations of Section 5 and the Supplementary Material.

3.3. **Interpretation Under Variance Misspecification.** If the conditional mean function is linear, Theorem 2 has important implications for the robustness and optimality properties of MVR solutions. The $k$ orthogonality conditions (3.4) are sufficient to determine the scale parameter $\gamma^*$ since condition (3.3) is then uniquely satisfied by $\beta = \beta_0$. Thus in the classical particular case of the linear conditional mean model, the MVR solution for $\beta$ is fully robust to misspecification of the scale function. Consequently, when the conditional mean function is correctly specified the OLS and MVR solutions for $\beta$ coincide. In the special case of linear scale specification, $Xs_1(X'\gamma)$ reduces to $X$. Because $X$ includes an intercept, the scale parameter $\gamma^*$ 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 $\gamma^*$ is solely determined by the $k$ orthogonality conditions*

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

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

In the presence of heteroskedasticity, MVR optimality properties under correct mean specification translate into improved finite-sample properties. 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 Kullback-Leibler measure of divergence of the scaled Gaussian density

$$f_\theta (Y, X) := \frac{1}{(2\pi)^{\frac{1}{2}} s(X'\gamma)} \exp\left(-\frac{1}{2} e(Y, X, \theta)^2\right),$$

from the true conditional density function of $Y$ given $X$, which is minimized at a maximum likelihood pseudo-true value. Define for $\theta \in \Theta$,

$$\mathcal{L}(\theta) := -E\left[\log f_\theta (Y, X)\right] = \frac{1}{2}\log(2\pi) + E\left[\log s(X'\gamma) + \frac{1}{2} e(Y, X, \theta)^2\right], \tag{3.6}$$

with first-order conditions

$$E\left[\frac{X}{s(X'\gamma)} e(Y, X, \theta)\right] = 0 \tag{3.7}$$

$$E\left[X \frac{s_1(X'\gamma)}{s(X'\gamma)} \left\{e(Y, X, \theta)^2 - 1\right\}\right] = 0. \tag{3.8}$$

In general the MVR solution $\theta^*$ need not satisfy equations (3.7)-(3.8), and therefore cannot be interpreted as a maximum likelihood pseudo-true value. Compared with the MVR criterion, an important limitation of criterion (3.6) is its lack of convexity. The second-order derivative of (3.6) with respect to the first component $\gamma_1$ of $\gamma$, i.e. for fixed $\beta$, $\gamma_{-1}$, 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 (3.6) in $\gamma_1$ (for any fixed $\beta$,$\gamma_{-1}$) implies that $\mathcal{L}(\theta)$ is not jointly convex[2], and that a maximum likelihood pseudo-true value might not exist; even if there exists one, it need not be unique. In contrast, the MVR solution is well-defined, and constitutes a convex alternative to maximum likelihood approximation of conditional mean and variance functions.

[2] 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 (3.6).

## 4. Estimation and Inference

We use the sample analog of the MVR population problem (2.2) for estimation of its solution $\theta^*$ 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 $\Theta$. For $\gamma \in \Theta_{\gamma,n}$, we let $\Omega_n(\gamma) = \operatorname{diag}(s(x_i'\gamma))$, 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.** We have $E[\|X\|^4] < \infty$, $E[Y^2 \|X\|^2] < \infty$ and, for all $\gamma \in \Theta_\gamma$, $E[Y^4 \|X\|^2 s_1(X'\gamma)^2] < \infty$ and $E[\|X\|^6 s_1(X'\gamma)^2] < \infty$.

Assumption 6 is needed for asymptotic normality of estimates of $\theta^*$. When the scale function $t \mapsto s(t)$ is specified to be linear, this assumption simplifies to the requirement that $E[Y^4 \|X\|^2]$ and $E[\|X\|^6]$ be finite, as in Spady and Stouli (2018).

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 4.** *If Assumptions 1-6 hold, then (i) there exists $\hat{\theta}$ in $\Theta$ 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*

(4.4)

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

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

Theorem 4 allows the construction of confidence intervals and the implementation of hypothesis tests for $\theta$ 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 4 provide a basis for the construction of a range of specification tests, similarly to the information matrix equality test in maximum likelihood 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.

An important implication of Theorem 4 is that the influence function of the MVR estimator for $\beta$ 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 $\beta$ are observations that are influential relative to the dispersion of $Y$, accounting for mean misspecification.

When the conditional mean function 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 $\beta$ is thus proportional to $m_1$ only, and the influence function of $\gamma$ is proportional to $m_2$ only, since the off-diagonal blocks of $G$ are then $0_{k\times k}$. For the mean parameter $\beta$, 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 first and second moments of $e(y_i, x_i, \theta)$ must remain equal to zero and one, respectively. 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. Therefore, the MVR influence function although unbounded for a fixed value of $\gamma$, 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. (Linear homoskedastic model.) For the model $Y = X'\beta_0 + U$ with $E[U|X] = 0$ and $\mathrm{Var}(U|X) = \sigma_0^2$, 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.

*Remark* 3. (Generalized least-squares and weighted MVR.) If both the conditional mean and variance functions are correctly specified, then generalized least-squares with weights $1/\sigma^2(x)$ is an efficient estimator for $\beta$. 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 $\beta$

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

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

*Remark* 4. (Implementation.) Under our assumptions, the MVR objective is globally convex in $\theta$, and therefore in $\beta$ 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

$$\min_{\gamma\in\mathbb{R}^k} \frac{1}{n}\sum_{i=1}^{n}\frac{1}{2}\left\{\left(\frac{y_i - x_i'\hat{\beta}(\gamma)}{s(x_i'\gamma)}\right)^2 + 1\right\} s(x_i'\gamma), \quad \hat{\beta}(\gamma) := [X_n'\Omega_n^{-1}(\gamma)X_n]^{-1}X_n'\Omega_n^{-1}(\gamma)y,$$
$$\text{s.t.} \quad s(x_i'\gamma) > 0, \quad i = 1, \ldots, n, \quad \text{if } s(t) \leq 0 \text{ for some } t \in \mathbb{R}.$$

Concentrating out $\beta$ for each $\gamma$ provides a convenient implementation of the MVR estimator $\hat{\gamma}$, with the final estimate for $\beta$ 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 $j$th 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 via 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 $W_{\text{MVR}}$ 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 $W_{\text{MVR}}$ can be specialized to formulate a one-step robust MVR-based test for heteroskedasticity. Letting

$$h = k-1, \quad R = \begin{bmatrix} 0_{k-1,k+1} & I_{k-1} \end{bmatrix}, \quad r = 0_{k-1},$$

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

*Remark* 5. When the conditional mean function is linear, robust MVR inference on $\hat{\beta}$ uses the closed-form variance formula

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

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

## 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 et al., 2017).

### 5.1. **Empirical Applications.**

5.1.1. *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 ($\ell$-MVR) and exponential ($e$-MVR) scale functions, and we report two types of standard errors: robust to mean misspecification (MVR1) and to variance misspecification (MVR2), imposing the simplifications in (4.3). We also report OLS estimates, with naive (OLS) and heteroskedasticity-robust standard errors (HC3)[3].

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

[3]We implement the approximate jackknife estimator of MacKinnon and White (1985) for the robust variance-covariance matrix, as suggested for small samples by Long and Ervin (2000), for instance.
[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.

dummies, and controlling for latitude, colonial origin and religion[5]. A striking feature of the results displayed in Table 1 is the robustness to scale specification of MVR point estimates and standard errors. They are nearly identical across all specifications, except for Panel (3). Moreover, 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.

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, when the Americas are dropped (Panel (3)), significance of the MVR estimates relies on assuming linearity of the conditional mean function, and the change in coefficients is much less pronounced for MVR estimates. We also find that the relationship between urbanization in 1500 and GDP per capita in 1995 is not statistically significant in the three remaining specifications. When only former colonies from the Americas are considered (Panel (4)), the OLS 7 percent significance level rises to above 13 percent for MVR estimates. Specification (6) drops observations for neo-Europes (United States, Canada, New Zealand, and Australia), and MVR estimates are significant at the 7 percent level under the assumption of a linear conditional mean (MVR2), but only at the 12 ($\ell$-MVR) and 10 ($e$-MVR) percent significance level otherwise (MVR1). When controlling for religion (Panel (9)), MVR estimates are also not significant with significance levels ranging from above 9 to 17 percent, against 6 percent for OLS.

MVR results thus 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.

5.1.2. *Demand for gasoline.* To illustrate our methods in a large sample, 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[6]. Blundell, Horowitz and Parey (2012) perform

[5]See Acemoglu, Johnson and Robinson (2002) for a detailed description of the data.
[6]See Blundell, Horowitz and Parey (2012) and ONRL (2004) for a detailed description of the data.

| | Dependent variable is log GDP per capita (PPP) in 1995 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | OLS | $\ell$-MVR | $e$-MVR | OLS | $\ell$-MVR | $e$-MVR | OLS | $\ell$-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 |
| HC3 – MVR1 | 0.025 | 0.028 | 0.026 | 0.036 | 0.034 | 0.034 | 0.056 | 0.127 | 0.113 |
| OLS – MVR2 | 0.026 | 0.022 | 0.022 | 0.032 | 0.033 | 0.033 | 0.051 | 0.035 | 0.039 |
| | (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 |
| HC3 – MVR1 | 0.033 | 0.032 | 0.032 | 0.035 | 0.029 | 0.030 | 0.023 | 0.023 | 0.023 |
| OLS – MVR2 | 0.029 | 0.030 | 0.030 | 0.030 | 0.025 | 0.023 | 0.026 | 0.020 | 0.021 |
| | (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 |
| HC3 – MVR1 | 0.022 | 0.022 | 0.021 | 0.028 | 0.026 | 0.027 | 0.032 | 0.029 | 0.029 |
| OLS – MVR2 | 0.025 | 0.018 | 0.019 | 0.028 | 0.021 | 0.022 | 0.032 | 0.025 | 0.026 |

TABLE 1. Reversal of fortune. Naive (OLS) and robust (HC3) standard errors for OLS estimates, and robust to mean (MVR1) and variance (MVR2) misspecification standard errors for MVR estimates.

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 location-scale 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, as well as a dummy for public transport availability, 4 urbanity dummies, 8 population density dummies and 9 regional dummies.

Table 2 reports estimates and standard errors for the average price and income elasticities obtained by OLS, $\ell$-MVR and $e$-MVR across the 5 linear specifications considered in Blundell, Horowitz and Parey (2012). We find that MVR estimates and standard errors are robust to scale specification for both price and income elasticities. In the baseline specification, MVR price elasticities are -0.89 and coincide with the average price elasticity found by Yatchew and No (2001) and West (2004). For specifications (1)-(4), MVR price elasticities are slightly smaller than OLS estimates, and the price 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[7]. 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

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

| | Dependent variable is log of annual household gasoline demand in gallons | | | | | |
|---|---|---|---|---|---|---|
| | Log price coefficient $\hat{\beta}_1$ | | | Log income coefficient $\hat{\beta}_2$ | | |
| | OLS | $\ell$-MVR | $e$-MVR | OLS | $\ell$-MVR | $e$-MVR |
| | (1) Baseline specification | | | | | |
| Point estimate | -0.925 | -0.892 | -0.888 | 0.289 | 0.283 | 0.283 |
| HC3 – MVR1 | 0.150 | 0.144 | 0.144 | 0.0190 | 0.0173 | 0.0172 |
| OLS – MVR2 | 0.155 | 0.145 | 0.145 | 0.0145 | 0.0176 | 0.0174 |
| | (2) With demographics | | | | | |
| Point estimate | -0.879 | -0.857 | -0.854 | 0.246 | 0.244 | 0.244 |
| HC3 – MVR1 | 0.143 | 0.137 | 0.137 | 0.0183 | 0.0169 | 0.0167 |
| OLS – MVR2 | 0.149 | 0.138 | 0.138 | 0.0143 | 0.0170 | 0.0168 |
| | (3) With demographics and public transports | | | | | |
| Point estimate | -0.830 | -0.820 | -0.816 | 0.269 | 0.268 | 0.268 |
| HC3 – MVR1 | 0.143 | 0.137 | 0.137 | 0.0187 | 0.0172 | 0.0171 |
| OLS – MVR2 | 0.148 | 0.138 | 0.138 | 0.0146 | 0.0172 | 0.0171 |
| | (4) With demographics, public transports and urbanity | | | | | |
| Point estimate | -0.495 | -0.483 | -0.478 | 0.298 | 0.301 | 0.301 |
| HC3 – MVR1 | 0.141 | 0.135 | 0.134 | 0.0190 | 0.0174 | 0.0173 |
| OLS – MVR2 | 0.147 | 0.135 | 0.135 | 0.0147 | 0.0173 | 0.0172 |
| | (5) With demographics, public transports, urbanity and regions | | | | | |
| Point estimate | -0.358 | -0.415 | -0.408 | 0.297 | 0.302 | 0.302 |
| HC3 – MVR1 | 0.270 | 0.256 | 0.256 | 0.0199 | 0.0181 | 0.0181 |
| OLS – MVR2 | 0.272 | 0.257 | 0.257 | 0.0153 | 0.0180 | 0.0178 |

Table 2. Demand for gasoline. Naive (OLS) and robust (HC3) standard errors for OLS estimates, and robust to mean (MVR1) and variance (MVR2) misspecification standard errors for MVR estimates.

smaller than heteroskedasticity-corrected OLS standard errors, which is a reflection of the heteroskedasticity detected for all specifications[8].

[8]For each specification we implemented the tests of Breusch and Pagan (1978), White (1980a) and Koenker (1983)) for heteroskedasticity for OLS and the test introduced in Section 4 for MVR. All tests reject the null of homoskedasticity for all specifications.

5.2. **Simulations.** We implement Monte Carlo simulations in order to assess and illustrate the finite-sample properties of our estimators. Our models feature a linear conditional mean function, and we implement OLS and MVR with linear and exponential scale functions. In the Supplementary Material, we provide additional results for models featuring a nonlinear conditional mean function and report simulation results from an artificial experiment of MacKinnon (2013). We find that using MVR approximations does not result in a loss in the quality of approximation of nonlinear conditional mean functions compared to OLS, and MVR estimation and inference finite-sample properties compare favorably to both OLS and weighted least-squares.

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). 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 $\alpha = 1$, $\ell$-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.

| | Design | LOC | LIN | | | | EXP | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | $\alpha$ | 0 | 0.5 | 1 | 1.5 | 2 | 0.5 | 1 | 1.5 | 2 |
| | | Log price coefficient $\hat{\beta}_1$ | | | | | | | | |
| | $n = 500$ | 102.1 | 100.5 | 96.0 | 89.0 | 80.6 | 100.6 | 96.1 | 89.0 | 80.3 |
| $\ell$-MVR | $n = 1000$ | 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 |
| | $n = 500$ | 102.1 | 100.4 | 95.9 | 89.1 | 80.5 | 100.4 | 95.8 | 88.7 | 79.5 |
| $e$-MVR | $n = 1000$ | 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 $\hat{\beta}_2$ | | | | | | | | |
| | $n = 500$ | 101.6 | 99.6 | 93.7 | 85.1 | 75.1 | 99.4 | 92.8 | 82.9 | 71.2 |
| $\ell$-MVR | $n = 1000$ | 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 |
| | $n = 500$ | 101.7 | 99.8 | 93.8 | 84.8 | 73.9 | 99.6 | 92.6 | 81.9 | 68.8 |
| $e$-MVR | $n = 1000$ | 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 3. Ratio ($\times 100$) of MVR root mean squared error for $\beta_1$ and $\beta_2$ over corresponding OLS counterpart.

Table 3 reports a first set of results regarding the accuracy of our estimators. We report the ratios of root mean squared errors for $\beta_1$ and $\beta_2$ of $\ell$-MVR and $e$-MVR over root mean squared errors of OLS, in percentage terms. The results show that MVR estimators achieve large gains relative to OLS in the presence of heteroskedasticity, with ratios ranging from 73.8 to 100.6 for $\hat{\beta}_1$ and from 48.9 to 99.8 for $\hat{\beta}_2$. 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 4 reports ratios of $\ell$-MVR and $e$-MVR average confidence interval lengths across simulations for $\beta_1$ and $\beta_2$ over OLS average confidence interval lengths, in percentage terms. MVR confidence intervals are based on MVR1 standard errors, and results for MVR2 are very similar and are reported in the Supplementary Material. In these

| | Design | LOC | LIN | | | | EXP | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | $\alpha$ | 0 | 0.5 | 1 | 1.5 | 2 | 0.5 | 1 | 1.5 | 2 |
| | | Log price coefficient $\hat{\beta}_1$ | | | | | | | | |
| $\ell$-MVR | $n = 500$ | 98.8 | 98.1 | 96.1 | 92.8 | 88.6 | 98.1 | 96.1 | 92.8 | 88.3 |
| | $n = 1000$ | 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 |
| $e$-MVR | $n = 500$ | 98.6 | 97.9 | 95.8 | 92.5 | 88.1 | 97.9 | 95.8 | 92.3 | 87.5 |
| | $n = 1000$ | 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 $\hat{\beta}_2$ | | | | | | | | |
| $\ell$-MVR | $n = 500$ | 98.9 | 98.2 | 95.5 | 91.3 | 86.3 | 98.0 | 95.0 | 90.1 | 84.0 |
| | $n = 1000$ | 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 |
| $e$-MVR | $n = 500$ | 98.7 | 97.8 | 94.9 | 90.4 | 84.6 | 97.7 | 94.3 | 88.8 | 81.6 |
| | $n = 1000$ | 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 4. Ratio ($\times 100$) of MVR average confidence interval lengths for $\beta_1$ and $\beta_2$ over corresponding OLS counterpart. Confidence intervals constructed with standard errors under misspecification of the conditional mean function (MVR1).

simulations MVR yields substantially tighter confidence intervals compared to OLS in the presence of heteroskedasticity, with confidence interval lengths ratios ranging from 85.6 to 98.9 for $\hat{\beta}_1$ and from 70.8 to 98.3 for $\hat{\beta}_2$, while not incurring any loss in precision for the homoskedastic data generating process.

Overall, our numerical simulations confirm MVR robustness to the specification of the scale function, and both $\ell$-MVR and $e$-MVR perform very well in finite samples. 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.

## 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 $\mathcal{B} = \mathcal{B}_\beta \times \mathcal{B}_\gamma \subseteq \mathit{\Theta}$, where $\mathcal{B}_\beta$ and $\mathcal{B}_\gamma$ are compact subsets of $\mathbb{R}^k$ and $\Theta_\gamma$, respectively, and the boundary set of $\Theta$

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

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 $(\gamma_n)$ 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 $(\gamma_n)$ in $\mathbb{R}^k$ such that $||\gamma_n|| \to \infty$, and define $\delta_n = \frac{\gamma_n}{||\gamma_n||}$. The sequence $(\delta_n)$ is bounded, and by application of the Bolzano-Weierstrass theorem there exists a convergent subsequence $\delta_{n_l}$, $n_l \to \infty$ as $l \to \infty$, with limit $\delta_o$. 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. $\square$

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

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

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

*Proof.* Positive definiteness of $E[XX'/s(X'\gamma)]$ for $\gamma \in \mathcal{B}_\gamma$ under Assumption 4 implies that $\varPsi(\theta)$ is positive definite for all $\theta \in \mathcal{B}$ if and only if the Schur complement of $E[XX'/s(X'\gamma)]$ in $\varPsi(\theta)$ is positive definite (Boyd and Vandenberghe (2004), Appendix A.5.5) for all $\theta \in \mathcal{B}$, i.e. if and only if

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

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 semidefinite $k \times k$ matrix, and equal to zero if and only if (after multiplication by $s(X'\gamma)^{1/2}$ on both sides)

$$Xs_1(X'\gamma)e(Y,X,\theta) = \Xi(\theta)X \tag{A.1}$$

a.s.; this is an application of the Cauchy-Schwarz inequality for matrices stated in Tripathi (1999). But if (A.1) holds, then, with $\Xi_j$ denoting the $j$th row of $\Xi$,

$$X_jY = (X'\beta)X_j + s(X'\gamma)\Xi_j(\theta)X, \quad j = 1,\ldots,k,$$

a.s., which implies that

$$X_j^2 \mathrm{var}(Y|X) = 0, \quad j = 1, \ldots, k,$$

a.s. Because $X$ includes an intercept, Assumption 2 implies that these equalities cannot hold for $j = 1$. 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

$$\partial L(X, Y, \theta)/\partial\beta = -Xe(Y, X, \theta), \quad \partial L(X, Y, \theta)/\partial\gamma = -\frac{1}{2} X s_1(X'\gamma)\{e(Y, X, \theta)^2 - 1\}.$$

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

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

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 $-1 \le e(Y,X,\theta)^2 - 1$ a.s., for $\theta \in \mathcal{B}$,

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

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

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

This bound and (A.4) together imply

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

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 $E[Y^2||X||^2 s_2(X'\gamma)] < \infty$, 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 $\theta \in \mathcal{B}$,

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

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

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

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)

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

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

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

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 now 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 $(\theta_n)$ in $\mathcal{B}_c$, and a subsequence $(\theta_{n_l})$, $n_l\to\infty$ 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 $\infty$ 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 $\gamma$ 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 $\beta$ 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 $\theta_n$ 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 $\lim_{n\to\infty} E[\tilde{L}(X,\theta_n)] = \lim_{l\to\infty} Q(\theta_n) = \infty$. 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. □

## 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 $\theta_0$. Uniqueness follows from strict convexity of $Q(\theta)$ over compact subsets of $\Theta$ established in Lemma 5, and compactness of the level sets of the objective function $Q(\theta)$ established in Lemma 6. □

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 $\beta_0$ and $\gamma_0$, 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 $(\beta,\gamma) = (\beta_0,\gamma_0)$. □

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 $\Theta$ 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 $\Theta$ 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 $\Theta$, 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 $\theta \in \Theta$,

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

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

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

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 $E[e(Y, X, \theta_{\mathrm{LS}})^2 - 1] = 0$, which implies that $s(\gamma_{\mathrm{LS}}) = E[(Y - X'\beta_{\mathrm{LS}})^2]^{1/2}$, and

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

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

B.6. **Proof of Corollary 2.** By assumption $\mu(X) = X'\beta_0$ a.s., and by definition $\theta^*$ satisfies conditions (3.3)-(3.4). Then the first-order conditions (3.3) are satisfied by $\beta^* = \beta_0$, and $\gamma^*$ must satisfy

$$E\left[X s_1(X'\gamma^*)\left\{\frac{\sigma(X)^2}{s(X'\gamma^*)} - 1\right\}\right] = 0. \tag{B.7}$$

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 $\Theta_\gamma$, i.e. there exists $\gamma^*$ such that (B.7) holds. Since $\gamma \mapsto q(\gamma)$ is also strictly convex, $q(\gamma)$ admits a unique minimizer $\gamma^*(\beta_0)$ in $\Theta_\gamma$. Therefore, the pair $(\beta_0, \gamma^*(\beta_0))$ is the unique minimizer of $Q(\theta)$ when $\mu(X) = X'\beta_0$ a.s. $\square$

**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 $\theta \in \mathcal{B}$, and the result follows. □

**Proof of Theorem 4(i)-(ii).**

*Proof.* 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 $\Theta$ 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 $\hat{\theta} \to^p \theta^*$. □

**Proof of Theorem 4(iii).**

*Proof.* 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, $\theta^*$ is in the interior of $\Theta$ 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, $\theta^*$ satisfies $E[m(Y, X, \theta^*)] = 0$, hence the first part of their condition (iii) is satisfied. Moreover, bound (A.4) in the proof of Lemma 4 shows 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 $E[\sup_{\theta\in\Theta} ||m(Y, X, \theta)||^2] < \infty$. Hence, for a neighborhood $\mathcal{N}$ of $\theta^*$, we have that $E[\sup_{\theta\in\mathcal{N}} ||m(Y, X, \theta)||^2] < \infty$. Moreover, $\theta \mapsto m(Y, X, \theta)$ is continuous at $\theta^*$ 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.

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. & 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.

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.

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.

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

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.

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.

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

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

White, H. (1980a). 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.

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

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.

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

Ypma, J., Borchers, H.W. and Eddelbuettel, D. (2017). nloptr: R Interface to NLopt. R package version 1.0.4.

# SUPPLEMENTARY MATERIAL FOR "SIMULTANEOUS MEAN-VARIANCE REGRESSION"

RICHARD H. SPADY† AND SAMI STOULI§

## 1. Summary

This supplementary material presents additional simulation results for "Simultaneous Mean-Variance Regression". In Section 2, we report the results of simulations based on a set of experiments proposed by MacKinnon (2013) in order to study further the finite-sample properties of MVR. The experiments are implemented for small sample sizes and are designed to make heteroskedasticity-robust inference difficult. We compare the finite-sample estimation and inference performance of MVR, OLS and weighted least squares (WLS). In Section 3, we report additional results for the numerical experiments in the main text, and study the finite-sample approximation properties of MVR by implementing simulations calibrated to the demand for gasoline empirical example with a nonlinear conditional mean function. Overall, all experiments confirm the favorable finite-sample estimation, inference and approximation properties of MVR.

*Date*: April 5, 2018.
† Nuffield College, Oxford, and Department of Economics, Johns Hopkins University, rspady@jhu.edu.
§ Department of Economics, University of Bristol, s.stouli@bristol.ac.uk.

## 2. Numerical Simulations of MacKinnon (2013)

2.1. **Design of Experiments.** In this Section we investigate the properties of MVR in small samples and compare its performance with OLS and WLS by implementing numerical simulations proposed by MacKinnon (2013).

The data generating process is

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

where all regressors are drawn from the standard lognormal distribution, and $z(\alpha)$ is chosen such that the expected variance of $\sigma\varepsilon$ is equal to 1. The log-normal regressors 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 $\beta_j = 1$ for $j = 0, \ldots, 3$, and $\beta_4 = 0$. As in the simulations in the main text, $\alpha$ 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 $\alpha$ and sample size, we generate 10000 samples, and implement OLS, WLS, and MVR with linear and exponential scale functions. For WLS we follow the implementation suggested by Romano and Wolf (2017, cf. equation (3.4), 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 regression

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

where $\delta = 0.1$ as in the implementation of Romano and Wolf (2017), and with estimates $(\hat{\nu}, \hat{\pi})$. The WLS weights are then formed as $\hat{w}_i := \exp(\hat{\nu} + \hat{\pi}' \log(|\tilde{x}_i|))$, and the WLS estimator is

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

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

### 2.2. **Results.**

2.2.1. *Estimation.* Tables 1 and 2 report the ratio of MVR root mean square errors (RMSE) across simulations over the OLS and WLS RMSEs for the three coefficient parameters $\beta_1$, $\beta_2$ and $\beta_3$, each sample size and value of heteroskedasticity parameter $\alpha$, in percentage terms. Denoting an estimator $\tilde{\beta}_j^{(s)}$ of $\beta_j$ for the $s$th simulation, the RMSE is computed as $\{\frac{1}{S}\sum_{s=1}^{S}(\tilde{\beta}_j^{(s)} - \beta_j)^2\}^{1/2}$, for $j = 1, 2, 3$ and $S = 10000$.

Table 1 shows that the performance of both MVR estimators relative to OLS improves as $n$ and $\alpha$ 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. For all coefficients the performance of MVR then becomes markedly superior as $n$ and $\alpha$ increase, with ratios that reach 28.9 for $\ell$-MVR and 20.4 for $e$-MVR. The estimator $\ell$-MVR dominates $e$-MVR slightly for the design with low heteroskedasticity ($\alpha = 0.5$), and for moderate heteroskedasticity ($\alpha = 1$) in small samples ($n = 20, 40$ and 80). 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 2, we find that the relative performance of both MVR estimators relative to WLS also improves as $n$ increases and as $\alpha$ 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 $\alpha \neq 0$, the relative performance of both MVR estimators dominates the performance of WLS with ratios that reach 56.1 for $\ell$-MVR and 39.7 for $e$-MVR. Compared to OLS and the results of Table 1, these results show that in this experiment WLS also improves over OLS, and that MVR yields substantial additional gains over WLS.

| | $\ell$-MVR | | | | | $e$-MVR | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $\alpha$ | 0 | 0.5 | 1 | 1.5 | 2 | 0 | 0.5 | 1 | 1.5 | 2 |
| | | | | | $\beta_1$ | | | | | |
| $n = 20$ | 105.2 | 103.3 | 97.6 | 90.2 | 83.0 | 106.5 | 104.8 | 98.7 | 90.1 | 81.3 |
| $n = 40$ | 103.8 | 99.5 | 89.1 | 77.1 | 66.7 | 103.4 | 99.4 | 89.1 | 76.4 | 63.5 |
| $n = 80$ | 102.1 | 96.0 | 81.8 | 66.5 | 55.2 | 101.9 | 96.3 | 82.0 | 65.5 | 50.5 |
| $n = 160$ | 101.7 | 92.9 | 74.7 | 57.3 | 46.5 | 101.5 | 93.3 | 74.7 | 55.5 | 40.0 |
| $n = 320$ | 101.0 | 89.3 | 67.5 | 49.1 | 39.5 | 100.9 | 89.5 | 66.9 | 46.5 | 31.7 |
| $n = 640$ | 100.9 | 87.5 | 62.3 | 42.6 | 33.8 | 100.9 | 87.8 | 61.6 | 39.8 | 25.5 |
| $n = 1280$ | 100.6 | 84.2 | 56.4 | 36.6 | 28.9 | 100.5 | 84.3 | 55.6 | 33.6 | 20.4 |
| | | | | | $\beta_2$ | | | | | |
| $n = 20$ | 104.7 | 103.3 | 97.4 | 90.0 | 82.8 | 106.3 | 104.5 | 98.1 | 89.4 | 80.5 |
| $n = 40$ | 103.9 | 99.6 | 88.8 | 76.3 | 65.9 | 103.7 | 99.7 | 89.1 | 75.8 | 62.8 |
| $n = 80$ | 103.0 | 95.7 | 80.9 | 65.6 | 54.4 | 102.6 | 96.0 | 81.1 | 64.4 | 49.5 |
| $n = 160$ | 101.6 | 92.4 | 74.1 | 56.9 | 46.4 | 101.2 | 92.8 | 74.0 | 54.8 | 39.5 |
| $n = 320$ | 100.8 | 89.9 | 68.2 | 49.5 | 39.7 | 100.7 | 90.3 | 68.1 | 47.2 | 32.1 |
| $n = 640$ | 100.8 | 87.0 | 62.0 | 42.5 | 33.8 | 100.7 | 87.1 | 61.1 | 39.4 | 25.1 |
| $n = 1280$ | 100.5 | 84.7 | 56.9 | 36.9 | 29.1 | 100.5 | 84.9 | 56.2 | 34.0 | 20.6 |
| | | | | | $\beta_3$ | | | | | |
| $n = 20$ | 104.5 | 103.1 | 98.0 | 91.0 | 84.0 | 106.1 | 104.5 | 98.8 | 90.7 | 82.1 |
| $n = 40$ | 103.4 | 99.0 | 88.6 | 76.8 | 66.6 | 103.3 | 99.1 | 88.9 | 76.3 | 63.6 |
| $n = 80$ | 101.9 | 95.3 | 81.1 | 66.3 | 55.3 | 101.6 | 95.5 | 81.2 | 65.0 | 50.3 |
| $n = 160$ | 101.6 | 92.5 | 74.2 | 56.9 | 46.5 | 101.3 | 92.7 | 73.9 | 54.8 | 39.4 |
| $n = 320$ | 100.8 | 89.5 | 67.7 | 49.0 | 39.7 | 100.7 | 89.7 | 66.9 | 46.1 | 31.1 |
| $n = 640$ | 100.7 | 86.5 | 61.5 | 42.4 | 34.7 | 100.6 | 86.7 | 60.9 | 39.5 | 25.4 |
| $n = 1280$ | 100.7 | 84.4 | 56.9 | 37.2 | 31.0 | 100.6 | 84.5 | 56.1 | 34.2 | 20.8 |

TABLE 1. Ratio ($\times 100$) of MVR RMSE for $\beta_1$, $\beta_2$ and $\beta_3$ over corresponding OLS counterpart.

| | $\ell$-MVR | | | | | $e$-MVR | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $\alpha$ | 0 | 0.5 | 1 | 1.5 | 2 | 0 | 0.5 | 1 | 1.5 | 2 |
| | | | | | $\beta_1$ | | | | | |
| $n = 20$ | 96.6 | 97.6 | 98.9 | 98.7 | 97.1 | 97.8 | 99.1 | 100.0 | 98.5 | 95.1 |
| $n = 40$ | 96.0 | 97.1 | 97.6 | 95.5 | 90.4 | 95.7 | 97.0 | 97.5 | 94.7 | 86.1 |
| $n = 80$ | 96.9 | 96.6 | 95.3 | 91.2 | 83.6 | 96.8 | 96.9 | 95.6 | 89.8 | 76.5 |
| $n = 160$ | 98.5 | 96.8 | 93.0 | 85.6 | 75.6 | 98.3 | 97.2 | 93.0 | 82.9 | 64.9 |
| $n = 320$ | 99.0 | 95.9 | 90.2 | 80.4 | 68.4 | 98.8 | 96.2 | 89.4 | 76.1 | 54.9 |
| $n = 640$ | 99.7 | 95.8 | 87.9 | 76.0 | 62.5 | 99.7 | 96.0 | 87.0 | 71.0 | 47.0 |
| $n = 1280$ | 100.2 | 94.1 | 83.3 | 69.8 | 56.1 | 100.1 | 94.2 | 82.2 | 64.0 | 39.7 |
| | | | | | $\beta_2$ | | | | | |
| $n = 20$ | 96.0 | 97.3 | 98.8 | 98.9 | 97.3 | 97.4 | 98.5 | 99.5 | 98.3 | 94.7 |
| $n = 40$ | 96.4 | 97.4 | 97.5 | 95.2 | 90.0 | 96.2 | 97.5 | 97.8 | 94.6 | 85.7 |
| $n = 80$ | 97.5 | 97.1 | 95.8 | 91.3 | 83.3 | 97.1 | 97.5 | 96.1 | 89.6 | 75.8 |
| $n = 160$ | 98.6 | 96.6 | 92.7 | 85.5 | 75.8 | 98.3 | 97.0 | 92.5 | 82.3 | 64.6 |
| $n = 320$ | 99.2 | 96.2 | 90.2 | 80.8 | 69.2 | 99.0 | 96.6 | 89.9 | 77.1 | 56.0 |
| $n = 640$ | 99.8 | 94.8 | 85.7 | 73.4 | 60.9 | 99.7 | 94.9 | 84.5 | 68.0 | 45.2 |
| $n = 1280$ | 100.0 | 94.3 | 83.3 | 69.8 | 56.2 | 99.9 | 94.5 | 82.3 | 64.3 | 39.8 |
| | | | | | $\beta_3$ | | | | | |
| $n = 20$ | 95.7 | 97.3 | 98.8 | 98.9 | 97.5 | 97.2 | 98.6 | 99.5 | 98.6 | 95.3 |
| $n = 40$ | 96.2 | 97.4 | 97.8 | 95.8 | 90.6 | 96.1 | 97.5 | 98.1 | 95.2 | 86.5 |
| $n = 80$ | 97.3 | 97.3 | 96.1 | 91.9 | 84.0 | 97.1 | 97.6 | 96.2 | 90.1 | 76.4 |
| $n = 160$ | 98.6 | 97.3 | 93.8 | 86.3 | 76.2 | 98.2 | 97.4 | 93.5 | 83.0 | 64.7 |
| $n = 320$ | 98.8 | 95.8 | 89.5 | 79.6 | 68.9 | 98.7 | 95.9 | 88.5 | 74.9 | 54.1 |
| $n = 640$ | 99.6 | 95.0 | 86.1 | 74.2 | 63.2 | 99.5 | 95.3 | 85.2 | 69.0 | 46.2 |
| $n = 1280$ | 100.2 | 94.7 | 84.4 | 71.6 | 60.9 | 100.2 | 94.8 | 83.3 | 65.7 | 40.9 |

TABLE 2. Ratio ($\times 100$) of MVR RMSE for $\beta_1$, $\beta_2$ and $\beta_3$ over corresponding WLS counterpart.

2.2.2. *Inference.* In order to study the finite-sample performance of MVR inference relative to heteroskedasticity-robust OLS and WLS inference, we compare the lengths of the confidence intervals constructed for the three coefficient parameters $\beta_1$, $\beta_2$ and $\beta_3$, and the rejection probabilities of asymptotic $t$ tests of the null hypothesis $\beta_4 = 0$ based on the standard normal distribution. All OLS and WLS standard errors used in the construction of confidence intervals and tests statistics are heteroskedasticity-robust (HC3).

Tables 3-6 report the ratio of average MVR confidence interval lengths across simulations over the average OLS and WLS confidence interval lengths for $\beta_1$, $\beta_2$ and $\beta_3$, each sample size and value of heteroskedasticity index $\alpha$, in percentage terms. MVR confidence intervals in Tables 3 and 4 are constructed with standard errors under misspecification of the conditional mean function (MVR1). Tables 5 and 6 report the results for MVR confidence intervals constructed with standard errors assuming correct specification of the conditional mean function (MVR2), imposing the simplifications in (4.3) in the main text.

In Table 3 we find that that the length of MVR confidence intervals is shorter for all designs compared to OLS confidence intervals, except for $\alpha = 0$, $n = 20$ for $\ell$-MVR. For each sample size, the ratio of average confidence interval lengths is inversely related to the degree of heteroskedasticity for both MVR estimators, with the exception of $\alpha = 0, 0.5$ for $n = 20, 40$ for $e$-MVR. For no or low heteroskedasticity, the ratio of confidence interval lengths increases with sample size, but for $\alpha \geq 1$ the relative performance of MVR then improves markedly with sample size as well. The main differences between the two MVR estimators occur for $n = 20$ and for designs with high heteroskedasticity ($\alpha = 1.5, 2$), where $e$-MVR performs especially well, further illustrating its higher robustness to more extreme designs. In Table 4, we find that $\ell$-MVR-based inference without assuming linearity of the conditional mean yields longer confidence intervals than WLS for small sample size ($n = 20$). In all other cases, MVR-based inference yields shorter confidence intervals, and for $\alpha \geq 1$ the relative performance of MVR confidence intervals improves with sample size.

Table 5 shows that MVR inference assuming a linear conditional mean model yields much shorter confidence intervals than OLS for all sample sizes, including very small samples $n = 20$ and 40 and the homoskedastic case. Compared to the results in Table 3, $e$-MVR still dominates $\ell$-MVR but the performance of $\ell$-MVR is much improved.

Table 6 shows that MVR confidence intervals are shorter than WLS confidence intervals for all designs when a linear conditional mean model is assumed, and the relative performance of MVR improves with sample size and the degree of heteroskedasticity, with ratios that can be as low as 51.0 for $\ell$-MVR and 50.2 for $e$-MVR.

Figures 2.1 and 2.2 display rejection probability curves of asymptotic $t$ tests of the null hypothesis $\beta_4 = 0$ for each sample size and value of the heteroskedasticity parameter $\alpha$. The nominal size of the tests is set to 5%. Figures 2.1(A)-(D) show that MVR addresses underrejection of the OLS- and WLS-based tests for moderate heteroskedasticity and above. Rejection probabilities curves are much lower for $\ell$-MVR and WLS than for $e$-MVR, and $\ell$-MVR rejection probability curves exhibit some crossing across sample sizes, just like OLS curves. In particular, the $\ell$-MVR rejection probability curve for $n = 20$ (black curve) is not placed above the other curves although it is now above the nominal level for all values of $\alpha$.

The $\ell$-MVR and WLS rejection probability curves exhibit similar behaviour, with a more pronounced flattening across $\alpha$ for $\ell$-MVR curves as $n$ increases. For the exponential scale case, $e$-MVR rejection probability curves become close to flat as $n$ reaches 320, across values of $\alpha$, with clear convergence from above to nominal size as $n$ increases. This is a remarkable property, and addresses the key limitation noted by MacKinnon (2013) about the OLS rejection probability curves, namely, how slowly they become flatter as $n$ increases. The MVR estimator addresses this limitation, illustrating its robustness to heteroskedasticity of various degrees.

Figure 2.2 compares rejection probability curves of MVR-based asymptotic $t$ tests constructed with standard errors robust to conditional mean misspecification (MVR1) and assuming linearity of the conditional mean (MVR2). For $\ell$-MVR, rejection probability curves obtained with MVR2 standard errors are much more similar to those of $e$-MVR, although they display some crossing for larger sample sizes $n = 640, 1280$ and high heteroskedasticity ($\alpha = 2$). This reflects a higher sensitivity of $\ell$-MVR standard errors to extreme observations, since for this design the most extreme observations in each sample gets more extreme as sample size increases, as noted by MacKinnon (2013). For both MVR estimators, the rejection probability curves obtained with MVR2 standard errors are slightly flatter than those obtained with MVR1, and indicate higher rejection frequencies for small sample sizes. As expected, the discrepancy between the two sets of rejection probability curves decreases with sample size.

| | $\ell$-MVR | | | | | $e$-MVR | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $\alpha$ | 0 | 0.5 | 1 | 1.5 | 2 | 0 | 0.5 | 1 | 1.5 | 2 |
| | | | | | $\beta_1$ | | | | | |
| $n = 20$ | 109.4 | 99.1 | 89.2 | 81.2 | 76.3 | 61.4 | 62.3 | 60.5 | 58.2 | 57.3 |
| $n = 40$ | 82.9 | 76.0 | 67.1 | 58.1 | 51.0 | 71.1 | 71.2 | 66.6 | 59.5 | 51.2 |
| $n = 80$ | 83.2 | 78.0 | 67.2 | 55.2 | 46.6 | 79.1 | 78.2 | 69.4 | 57.1 | 44.5 |
| $n = 160$ | 87.9 | 81.9 | 67.2 | 52.3 | 43.1 | 85.5 | 83.1 | 69.3 | 52.8 | 38.2 |
| $n = 320$ | 91.3 | 83.7 | 65.2 | 48.2 | 39.4 | 90.3 | 85.1 | 66.4 | 47.2 | 32.2 |
| $n = 640$ | 94.0 | 84.1 | 62.1 | 43.7 | 35.5 | 93.8 | 85.3 | 62.6 | 41.9 | 27.1 |
| $n = 1280$ | 96.1 | 83.8 | 59.1 | 39.9 | 32.1 | 96.0 | 84.7 | 59.3 | 37.6 | 23.3 |
| | | | | | $\beta_2$ | | | | | |
| $n = 20$ | 110.5 | 100.1 | 90.2 | 81.7 | 76.1 | 62.2 | 63.7 | 61.9 | 59.0 | 58.1 |
| $n = 40$ | 82.5 | 75.8 | 67.1 | 58.1 | 51.3 | 71.1 | 71.2 | 66.9 | 59.8 | 51.3 |
| $n = 80$ | 83.4 | 78.1 | 67.3 | 55.4 | 46.7 | 79.3 | 78.5 | 69.6 | 57.2 | 44.6 |
| $n = 160$ | 87.9 | 82.2 | 67.4 | 52.4 | 43.3 | 85.6 | 83.3 | 69.5 | 53.0 | 38.4 |
| $n = 320$ | 91.3 | 84.0 | 65.6 | 48.5 | 39.6 | 90.1 | 85.3 | 66.8 | 47.7 | 32.6 |
| $n = 640$ | 94.0 | 84.2 | 62.2 | 43.8 | 35.6 | 93.7 | 85.3 | 62.8 | 42.0 | 27.1 |
| $n = 1280$ | 96.1 | 83.8 | 58.9 | 39.8 | 32.0 | 96.0 | 84.7 | 59.1 | 37.4 | 23.1 |
| | | | | | $\beta_3$ | | | | | |
| $n = 20$ | 109.7 | 99.5 | 89.6 | 81.8 | 76.1 | 62.3 | 64.2 | 61.4 | 58.7 | 56.9 |
| $n = 40$ | 82.7 | 76.2 | 67.5 | 58.5 | 51.5 | 71.4 | 71.6 | 67.2 | 60.1 | 51.8 |
| $n = 80$ | 83.2 | 78.0 | 67.3 | 55.4 | 46.8 | 79.0 | 78.3 | 69.7 | 57.4 | 44.8 |
| $n = 160$ | 87.7 | 81.7 | 67.1 | 52.3 | 43.3 | 85.3 | 82.8 | 69.1 | 52.7 | 38.2 |
| $n = 320$ | 91.5 | 83.9 | 65.4 | 48.3 | 39.6 | 90.4 | 85.3 | 66.6 | 47.3 | 32.3 |
| $n = 640$ | 93.9 | 84.2 | 62.3 | 43.9 | 36.1 | 93.5 | 85.4 | 62.9 | 42.1 | 27.3 |
| $n = 1280$ | 96.1 | 83.7 | 59.0 | 39.9 | 32.9 | 96.0 | 84.5 | 59.1 | 37.5 | 23.2 |

TABLE 3. Ratio ($\times 100$) of MVR average confidence interval lengths for $\beta_1$, $\beta_2$ and $\beta_3$ over corresponding OLS counterpart. Confidence intervals constructed with standard errors under misspecification of the conditional mean function (MVR1).

Overall the simulation results in this Section illustrate the large MVR finite-sample improvements for estimation and inference in heteroskedastic designs. In particular, $e$-MVR displays remarkable robustness properties for the most difficult designs.

| | $\ell$-MVR | | | | | $e$-MVR | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $\alpha$ | 0 | 0.5 | 1 | 1.5 | 2 | 0 | 0.5 | 1 | 1.5 | 2 |
| | | | | | $\beta_1$ | | | | | |
| $n = 20$ | 122.9 | 118.3 | 117.7 | 118.5 | 119.7 | 69.0 | 74.4 | 79.9 | 85.0 | 89.9 |
| $n = 40$ | 88.0 | 87.0 | 87.8 | 87.7 | 85.5 | 75.4 | 81.5 | 87.2 | 89.8 | 85.8 |
| $n = 80$ | 85.4 | 86.8 | 88.3 | 87.0 | 83.0 | 81.1 | 87.1 | 91.2 | 90.0 | 79.3 |
| $n = 160$ | 89.2 | 90.3 | 89.4 | 85.8 | 80.4 | 86.7 | 91.7 | 92.1 | 86.7 | 71.3 |
| $n = 320$ | 92.0 | 92.0 | 88.4 | 82.8 | 76.5 | 91.0 | 93.5 | 90.0 | 81.1 | 62.6 |
| $n = 640$ | 94.3 | 93.3 | 87.6 | 80.7 | 73.6 | 94.1 | 94.6 | 88.4 | 77.3 | 56.1 |
| $n = 1280$ | 96.3 | 93.6 | 85.7 | 77.5 | 70.3 | 96.1 | 94.6 | 86.0 | 73.1 | 50.9 |
| | | | | | $\beta_2$ | | | | | |
| $n = 20$ | 124.5 | 119.8 | 118.8 | 118.8 | 118.9 | 70.1 | 76.2 | 81.6 | 85.7 | 90.7 |
| $n = 40$ | 87.9 | 87.2 | 88.0 | 87.8 | 85.6 | 75.8 | 81.9 | 87.7 | 90.2 | 85.6 |
| $n = 80$ | 86.1 | 87.3 | 88.3 | 87.1 | 83.0 | 81.9 | 87.6 | 91.4 | 90.0 | 79.3 |
| $n = 160$ | 89.2 | 90.5 | 89.2 | 85.7 | 80.5 | 86.8 | 91.6 | 92.0 | 86.5 | 71.3 |
| $n = 320$ | 92.0 | 92.4 | 88.9 | 83.6 | 77.4 | 90.8 | 93.8 | 90.6 | 82.2 | 63.6 |
| $n = 640$ | 94.3 | 93.2 | 87.1 | 79.9 | 73.1 | 94.0 | 94.4 | 87.8 | 76.5 | 55.7 |
| $n = 1280$ | 96.2 | 93.3 | 85.2 | 77.1 | 69.8 | 96.1 | 94.3 | 85.4 | 72.6 | 50.4 |
| | | | | | $\beta_3$ | | | | | |
| $n = 20$ | 124.1 | 119.2 | 118.1 | 118.9 | 119.0 | 70.5 | 76.9 | 80.8 | 85.4 | 89.0 |
| $n = 40$ | 87.9 | 87.2 | 88.0 | 87.7 | 85.6 | 75.9 | 82.0 | 87.7 | 90.1 | 86.1 |
| $n = 80$ | 85.9 | 87.4 | 88.6 | 87.3 | 83.2 | 81.6 | 87.8 | 91.8 | 90.4 | 79.7 |
| $n = 160$ | 89.2 | 90.5 | 89.5 | 86.0 | 80.8 | 86.8 | 91.7 | 92.2 | 86.7 | 71.3 |
| $n = 320$ | 92.1 | 92.1 | 88.5 | 83.0 | 77.2 | 90.9 | 93.5 | 90.2 | 81.4 | 63.0 |
| $n = 640$ | 94.2 | 93.3 | 87.3 | 80.3 | 74.2 | 93.8 | 94.6 | 88.2 | 77.0 | 56.1 |
| $n = 1280$ | 96.2 | 93.5 | 85.7 | 77.9 | 72.5 | 96.1 | 94.4 | 85.8 | 73.3 | 51.0 |

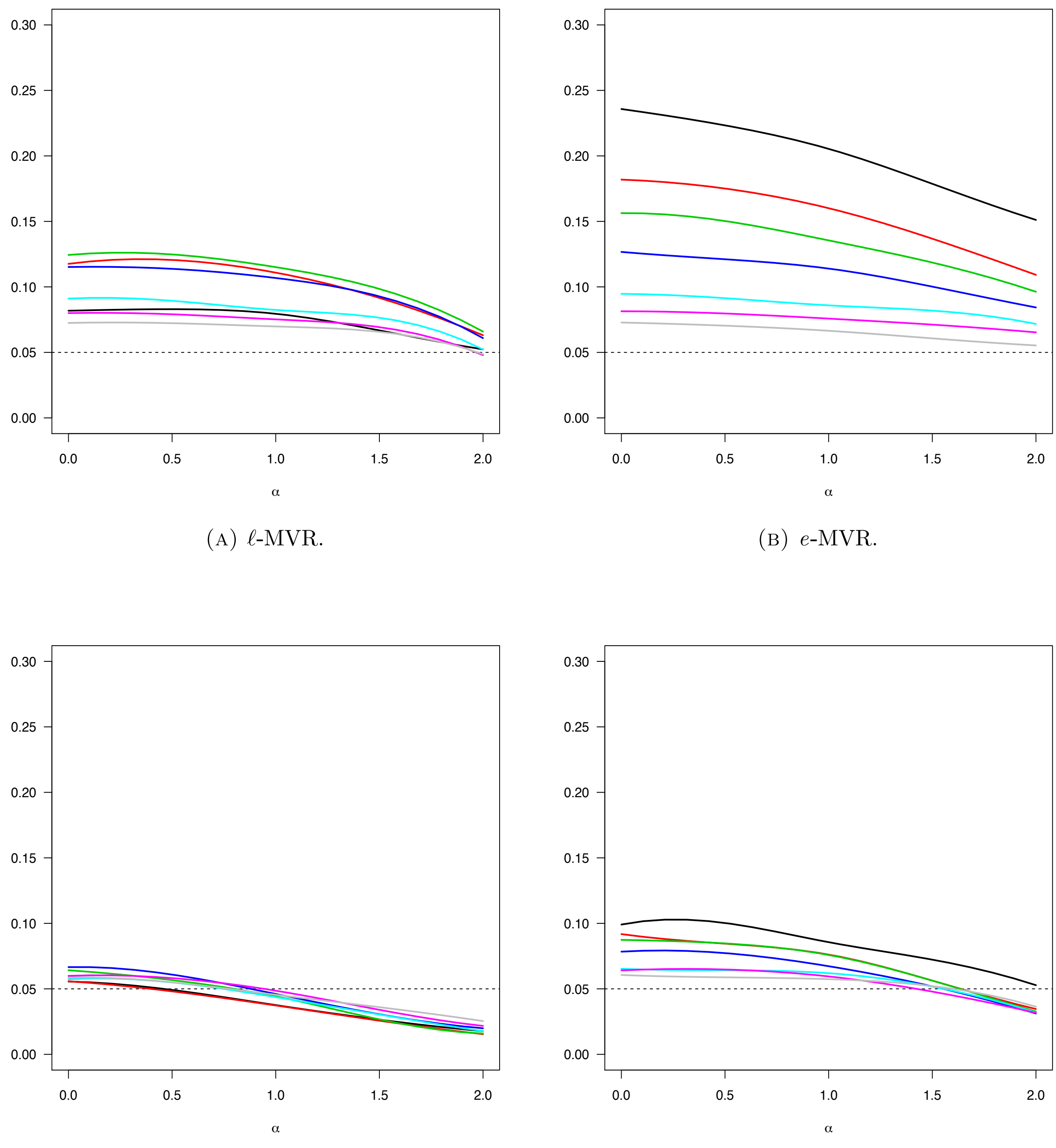
TABLE 4. Ratio ($\times 100$) of MVR average confidence interval lengths for $\beta_1$, $\beta_2$ and $\beta_3$ over corresponding WLS counterpart. Confidence intervals constructed with standard errors under misspecification of the conditional mean function (MVR1).

| | $\ell$-MVR | | | | | $e$-MVR | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $\alpha$ | 0 | 0.5 | 1 | 1.5 | 2 | 0 | 0.5 | 1 | 1.5 | 2 |
| | | | | | $\beta_1$ | | | | | |
| $n = 20$ | 45.5 | 44.4 | 42.1 | 39.0 | 36.1 | 44.2 | 43.6 | 41.3 | 38.0 | 34.3 |
| $n = 40$ | 59.3 | 59.4 | 55.4 | 49.1 | 42.7 | 60.7 | 60.0 | 55.3 | 48.5 | 41.0 |
| $n = 80$ | 70.4 | 71.0 | 63.2 | 52.4 | 43.3 | 72.3 | 71.0 | 62.7 | 51.6 | 40.5 |
| $n = 160$ | 79.2 | 78.7 | 65.8 | 51.2 | 41.2 | 80.9 | 78.4 | 65.5 | 50.1 | 36.6 |
| $n = 320$ | 85.9 | 82.3 | 64.7 | 47.7 | 37.9 | 87.1 | 82.1 | 64.4 | 46.1 | 31.6 |
| $n = 640$ | 90.7 | 83.5 | 62.0 | 43.4 | 34.4 | 91.6 | 83.6 | 61.6 | 41.3 | 26.9 |
| $n = 1280$ | 94.1 | 83.6 | 59.1 | 39.6 | 31.2 | 94.6 | 83.8 | 58.7 | 37.4 | 23.2 |
| | | | | | $\beta_2$ | | | | | |
| $n = 20$ | 45.4 | 44.4 | 42.0 | 39.0 | 36.0 | 44.7 | 44.1 | 41.7 | 38.3 | 34.5 |
| $n = 40$ | 59.2 | 59.5 | 55.6 | 49.2 | 42.9 | 60.8 | 60.0 | 55.4 | 48.7 | 41.2 |
| $n = 80$ | 70.6 | 71.3 | 63.4 | 52.5 | 43.4 | 72.6 | 71.3 | 62.9 | 51.7 | 40.5 |
| $n = 160$ | 79.2 | 78.8 | 65.9 | 51.3 | 41.4 | 81.0 | 78.5 | 65.6 | 50.3 | 36.8 |
| $n = 320$ | 85.8 | 82.7 | 65.1 | 48.0 | 38.2 | 87.1 | 82.5 | 64.9 | 46.5 | 32.0 |
| $n = 640$ | 90.7 | 83.7 | 62.1 | 43.5 | 34.5 | 91.6 | 83.7 | 61.7 | 41.4 | 26.9 |
| $n = 1280$ | 94.2 | 83.5 | 58.9 | 39.5 | 31.1 | 94.7 | 83.7 | 58.5 | 37.2 | 23.0 |
| | | | | | $\beta_3$ | | | | | |
| $n = 20$ | 45.4 | 44.5 | 42.4 | 39.3 | 36.2 | 44.8 | 44.2 | 41.9 | 38.5 | 34.7 |
| $n = 40$ | 59.2 | 59.6 | 55.7 | 49.3 | 42.9 | 60.8 | 60.2 | 55.6 | 48.9 | 41.4 |
| $n = 80$ | 70.4 | 71.2 | 63.4 | 52.6 | 43.5 | 72.4 | 71.3 | 63.0 | 51.9 | 40.7 |
| $n = 160$ | 79.0 | 78.5 | 65.7 | 51.2 | 41.3 | 80.8 | 78.2 | 65.3 | 50.1 | 36.7 |
| $n = 320$ | 85.9 | 82.6 | 64.9 | 47.8 | 38.1 | 87.2 | 82.4 | 64.6 | 46.2 | 31.7 |
| $n = 640$ | 90.5 | 83.7 | 62.2 | 43.6 | 34.9 | 91.4 | 83.7 | 61.9 | 41.6 | 27.1 |
| $n = 1280$ | 94.1 | 83.5 | 59.0 | 39.6 | 32.1 | 94.6 | 83.6 | 58.5 | 37.3 | 23.1 |

TABLE 5. Ratio ($\times 100$) of MVR average confidence interval lengths for $\beta_1$, $\beta_2$ and $\beta_3$ over corresponding OLS counterpart. Confidence intervals constructed with standard errors assuming correct specification of the conditional mean function (MVR2).

| | $\ell$-MVR | | | | | $e$-MVR | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $\alpha$ | 0 | 0.5 | 1 | 1.5 | 2 | 0 | 0.5 | 1 | 1.5 | 2 |
| | | | | | $\beta_1$ | | | | | |
| $n = 20$ | 51.0 | 53.0 | 55.5 | 57.0 | 56.5 | 49.6 | 52.1 | 54.5 | 55.4 | 53.7 |
| $n = 40$ | 62.9 | 68.0 | 72.5 | 74.1 | 71.6 | 64.4 | 68.7 | 72.3 | 73.2 | 68.7 |
| $n = 80$ | 72.2 | 79.1 | 83.1 | 82.5 | 77.1 | 74.2 | 79.1 | 82.5 | 81.3 | 72.0 |
| $n = 160$ | 80.3 | 86.8 | 87.4 | 84.0 | 76.8 | 82.0 | 86.5 | 87.0 | 82.3 | 68.3 |
| $n = 320$ | 86.5 | 90.5 | 87.8 | 82.0 | 73.7 | 87.8 | 90.3 | 87.3 | 79.1 | 61.5 |
| $n = 640$ | 91.0 | 92.6 | 87.4 | 80.2 | 71.2 | 91.9 | 92.7 | 86.9 | 76.3 | 55.6 |
| $n = 1280$ | 94.2 | 93.3 | 85.7 | 77.0 | 68.3 | 94.8 | 93.5 | 85.2 | 72.6 | 50.7 |
| | | | | | $\beta_2$ | | | | | |
| $n = 20$ | 51.2 | 53.1 | 55.4 | 56.7 | 56.2 | 50.4 | 52.7 | 54.9 | 55.7 | 53.9 |
| $n = 40$ | 63.1 | 68.4 | 72.9 | 74.4 | 71.6 | 64.8 | 69.0 | 72.6 | 73.5 | 68.8 |
| $n = 80$ | 72.9 | 79.6 | 83.2 | 82.5 | 77.0 | 74.9 | 79.6 | 82.6 | 81.3 | 72.0 |
| $n = 160$ | 80.3 | 86.7 | 87.2 | 83.9 | 76.9 | 82.1 | 86.4 | 86.9 | 82.2 | 68.4 |
| $n = 320$ | 86.4 | 90.9 | 88.3 | 82.8 | 74.5 | 87.7 | 90.7 | 88.0 | 80.2 | 62.5 |
| $n = 640$ | 91.0 | 92.6 | 86.9 | 79.4 | 70.8 | 91.9 | 92.6 | 86.4 | 75.6 | 55.2 |
| $n = 1280$ | 94.3 | 93.1 | 85.2 | 76.5 | 67.9 | 94.8 | 93.3 | 84.7 | 72.1 | 50.2 |
| | | | | | $\beta_3$ | | | | | |
| $n = 20$ | 51.3 | 53.3 | 55.8 | 57.2 | 56.6 | 50.6 | 52.9 | 55.1 | 56.0 | 54.3 |
| $n = 40$ | 62.9 | 68.2 | 72.7 | 73.9 | 71.3 | 64.6 | 68.9 | 72.5 | 73.3 | 68.8 |
| $n = 80$ | 72.7 | 79.8 | 83.5 | 82.8 | 77.3 | 74.8 | 79.9 | 83.0 | 81.7 | 72.4 |
| $n = 160$ | 80.3 | 86.9 | 87.6 | 84.2 | 77.1 | 82.2 | 86.7 | 87.2 | 82.5 | 68.5 |
| $n = 320$ | 86.4 | 90.6 | 87.9 | 82.3 | 74.4 | 87.8 | 90.4 | 87.5 | 79.5 | 61.9 |
| $n = 640$ | 90.8 | 92.6 | 87.2 | 79.8 | 71.9 | 91.7 | 92.7 | 86.7 | 76.0 | 55.7 |
| $n = 1280$ | 94.2 | 93.2 | 85.7 | 77.4 | 70.7 | 94.7 | 93.3 | 85.0 | 72.8 | 50.8 |

TABLE 6. Ratio ($\times 100$) of MVR average confidence interval lengths for $\beta_1$, $\beta_2$ and $\beta_3$ over corresponding WLS counterpart. Confidence intervals constructed with standard errors assuming correct specification of the conditional mean function (MVR2).



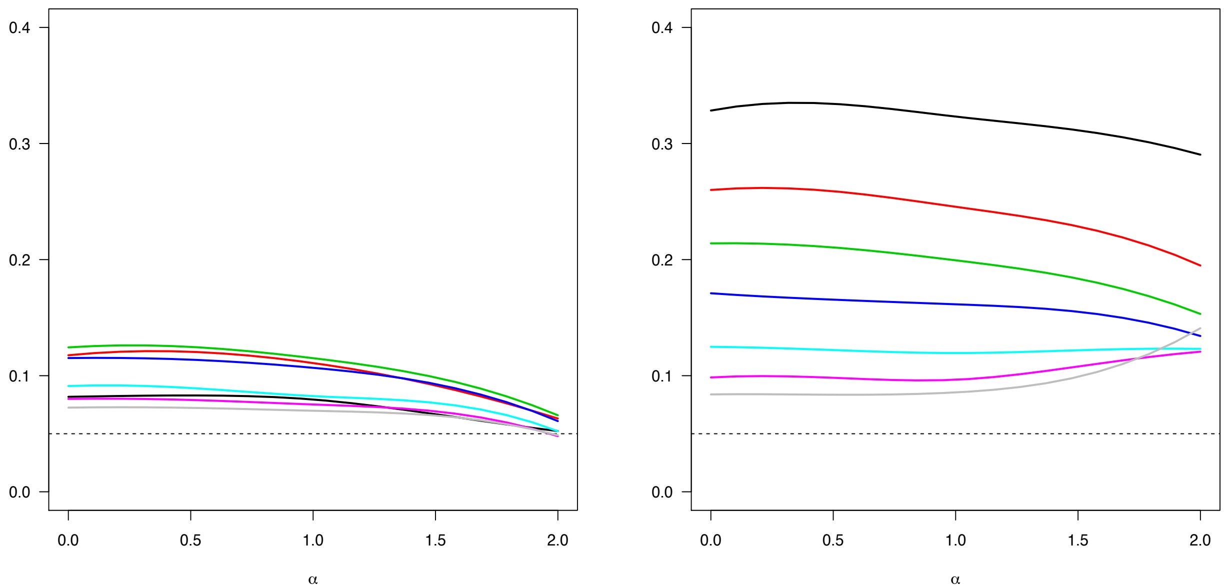

(A) $\ell$-MVR.

(B) $e$-MVR.

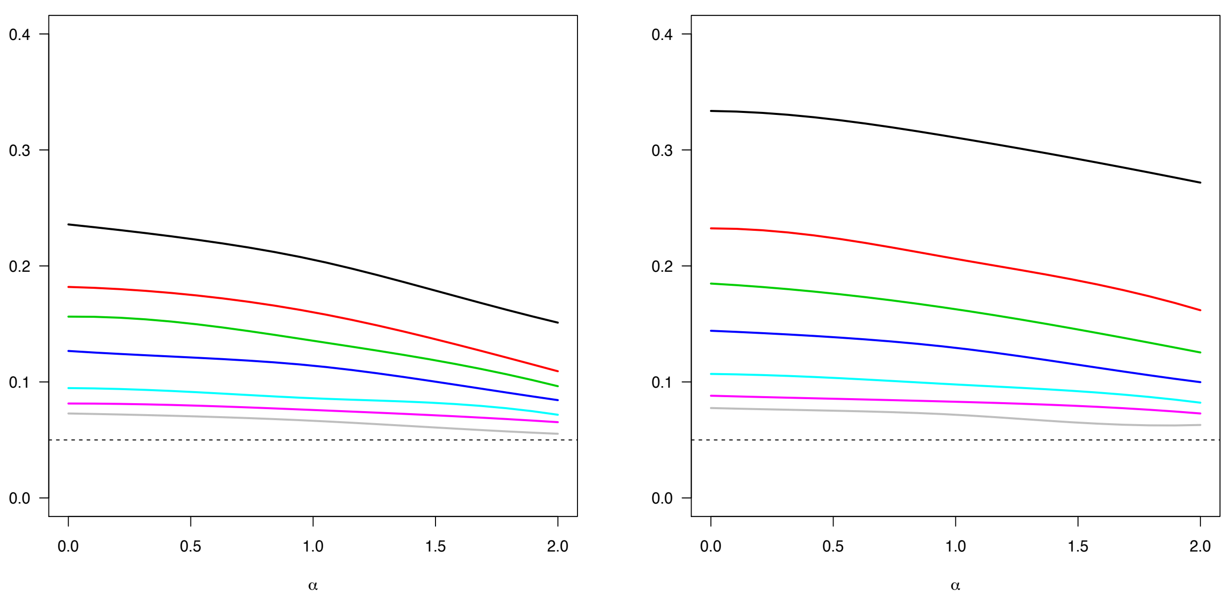
(C) OLS.

(D) WLS.

FIGURE 2.1. Rejection frequencies for asymptotic $t$ tests: $\ell$-MVR, $e$-MVR, OLS and WLS. Sample sizes: 20 (black), 40 (red), 80 (green), 160 (blue), 320 (cyan), 640 (magenta), 1280 (grey).




(A) $\ell$-MVR.




(B) $e$-MVR.

FIGURE 2.2. Rejection frequencies for asymptotic $t$ tests: $\ell$-MVR, $e$-MVR. MVR1 (left) and MVR2 (right). Sample sizes: 20 (black), 40 (red), 80 (green), 160 (blue), 320 (cyan), 640 (magenta), 1280 (grey).

## 3. Numerical Simulations Calibrated to The Demand for Gasoline Example

3.1. **Additional results: confidence intervals for MVR2.** Table 7 reports ratios of $\ell$-MVR and $e$-MVR average confidence interval lengths across simulations for $\beta_1$ and $\beta_2$ over OLS average confidence interval lengths, in percentage terms. MVR confidence intervals are based on MVR2 standard errors, and are slightly more favorable to MVR compared to the results obtained with MVR1 standard errors reported in Table 4 in the main text.

In these simulations, with standard errors calculated assuming correct specification of the conditional mean, MVR also yields substantially tighter confidence intervals compared to OLS in the presence of heteroskedasticity, with confidence interval lengths ratios ranging from 84.5 to 98.8 for $\beta_1$ and from 70.4 to 98.2 for $\beta_2$, while not incurring any loss in precision for the homoskedastic data generating process.

3.2. **Additional simulations: nonlinear conditional mean function.** We present the results of a second set of experiments based on the demand for gasoline empirical example in which we compare the approximation properties of MVR to those of OLS under misspecification of the conditional mean function, in root mean square error. The designs of our simulations in the main text 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) in the main text, 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 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 | LOC | LIN | | | | EXP | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | $\alpha$ | 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 |
| $\ell$-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 |
| $\ell$-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 7. Ratio ($\times 100$) of MVR average confidence interval lengths for $\beta_1$ and $\beta_2$ over corresponding OLS counterpart. Confidence intervals constructed standard errors assuming correct specification of the conditional mean function (MVR2).

**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 we let $\varepsilon \sim \mathcal{N}(0,1)$. For all designs we implement our estimators and OLS for the same sample sizes and $X$ values as in the main text, with the number of simulations set to 5000.

Table 8 reports results regarding the accuracy of OLS and MVR linear approximations of the conditional mean function $\mu(x, \beta) = \beta_0 + f(x_1, \beta_1) + x_2\beta_2 + x_3'\beta_3$, evaluated at the $n$ sample values $x_{1i}$ of $X_1$, and at fixed values of the remaining variables.[1] For

[1]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

| | Design | LOC | LIN | | | | EXP | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | $\alpha$ | 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 |
| $\ell$-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 8. Ratio ($\times 100$) of average MVR RMSE for $\mu(x)$ over corresponding OLS counterpart.

each data generating process we report the ratios of average estimation errors across simulations of $\ell$-MVR and $e$-MVR relative to OLS in percentage terms. Estimation errors are measured for each simulation by the root mean squared error, and then averaged across simulations.

These results confirm that in these simulations MVR does not result in finite-sample loss in the quality of approximation of nonlinear conditional mean functions relative to OLS, measured in RMSE.

## REFERENCES

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.

ROMANO, J. P. AND WOLF, M. (2017). Resurrecting Weighted Least Squares. *Journal of Econometrics* 197, pp. 1–19.

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.