Evdokimov, Kirill S.; Zeleneev, Andrei

**Working Paper**

# Simple estimation of semiparametric models with measurement errors

cemmap working paper, No. CWP08/22

**Provided in Cooperation with:**
Institute for Fiscal Studies (IFS), London

*Suggested Citation:* Evdokimov, Kirill S.; Zeleneev, Andrei (2022) : Simple estimation of semiparametric models with measurement errors, cemmap working paper, No. CWP08/22, Centre for Microdata Methods and Practice (cemmap), London, https://doi.org/10.47004/wp.cem.2022.0822

This Version is available at:
https://hdl.handle.net/10419/260389

cemmap
centre for microdata methods and practice

# Simple estimation of semiparametric models with measurement errors

Kirill S. Evdokimov
Andrei Zeleneev

The Institute for Fiscal Studies
Department of Economics, UCL

**cemmap** working paper CWP08/22

Economic and Social Research Council

# Simple Estimation of Semiparametric Models with Measurement Errors

Kirill S. Evdokimov* Andrei Zeleneev†‡§

This version: 2022/02/22

**Abstract**

We develop a practical way of addressing the Errors-In-Variables (EIV) problem in the Generalized Method of Moments (GMM) framework. We focus on the settings in which the variance of the measurement errors is a fraction of that of the mismeasured variables, which is typical for empirical applications. For any initial set of moment conditions our approach provides a "corrected" set of moment conditions that do not suffer from the EIV bias. The EIV-robust estimator is then computed as a standard GMM estimator with these corrected moment conditions. We show that our estimator is $\sqrt{n}$-consistent, and that the standard tests and confidence intervals provide valid inference. This is true even when the EIV are so large that the naive estimator (that ignores the EIV problem) may have a large bias with confidence intervals having 0% coverage. Our approach requires no nonparametric estimation, which can be particularly useful when the measurement errors are multivariate, serially correlated, or non-classical.

**Keywords:** Errors-In-Variables, measurement error robust GMM, moment conditions, nonstandard asymptotic approximation

*Universitat Pompeu Fabra and Barcelona School of Economics: kirill.evdokimov@upf.edu.

†University College London: a.zeleneev@ucl.ac.uk.

‡First version: November 7, 2016. A part of the material of this paper was previously circulated as a part of Evdokimov and Zeleneev (2018).

§We thank the participants of the numerous seminars and conferences for helpful comments and suggestions. We are also grateful to the Gregory C. Chow Econometrics Research Program at Princeton University and the Department of Economics at the Massachusetts Institute of Technology for their hospitality and support. Evdokimov also acknowledges the financial support from the National Science Foundation via grant SES-1459993, from the Spanish AEI Severo Ochoa Programme CEX2019-000915-S, and Ramon y Cajal Fellowship, RYC2020-030623-I.

# 1 Introduction

Measurement errors are a common problem for empirical studies. While the standard instrumental variables approach can be used to remove the errors in variables (EIV) bias in linear models, as pointed out by Amemiya (1985), nonlinear models require more elaborate strategies. The problems of identification, estimation, and inference in nonlinear models with EIV have received a lot of attention in the econometrics literature.[1]

The goal of this paper is to develop a simple and practical approach to estimation of general nonlinear moment condition models

$$\mathbb{E}[g(X_i^*, S_i, \theta)] = 0 \text{ iff } \theta = \theta_0, \tag{1}$$

where $g(\cdot)$ is an $m \times 1$ vector of functions, $\theta$ is a $p \times 1$ parameter vector and $\theta_0$ denotes its true value. The researcher has a random sample of $\{X_i, S_i\}_{i=1}^n$, where scalar or vector $X_i$ is a mismeasured version of unobserved $X_i^*$ with measurement error $\varepsilon_i$:

$$X_i = X_i^* + \varepsilon_i. \tag{2}$$

The measurement error can be classical or non-classical. The researcher is interested in estimating parameters $\theta_0$. A naive GMM estimator (which ignores EIV and uses $X_i$ in place of $X_i^*$) is biased because $\mathbb{E}[g(X_i, S_i, \theta_0)] \neq 0$.

To provide a practical estimation approach, we focus on the empirical settings in which the researcher believes the variability of the measurement error to be at most a fraction of the variability of the mismeasured variable, i.e., the noise-to-signal ratio $\tau \equiv \sigma_\varepsilon / \sigma_{X^*}$ to be moderate, e.g., $\tau \lesssim 0.5$. The absolute magnitude of the measurement error $\sigma_\varepsilon$ does not need to be small. Focusing on these settings allows us to develop a simple estimator, which does not require any nonparametric estimation or simulation. Such simple estimation becomes possible because in these settings we can obtain a simple approximation of the EIV bias of the moment conditions as a function of $\theta$.

We propose to bias correct the original moment conditions $g(\cdot)$, which would in turn remove the bias of the corresponding estimator of $\theta_0$. The bias correction depends on some moments of the distribution of the measurement errors that are unknown. Another difficulty is that the estimators of some components of the bias correction themselves may need to be bias corrected. To address these issues, we develop the *corrected moment*

[1] See Hausman, Ichimura, Newey, and Powell (1991); Hausman, Newey, and Powell (1995); Newey (2001); Schennach (2007); Li (2002); Schennach (2004); Chen, Hong, and Tamer (2005); Schennach (2007); Hu and Schennach (2008); Schennach (2014), among others.

*conditions*, which depend on $\theta$ and additional parameters $\gamma$ that govern the bias correction. The true parameter value $\gamma_0$ is associated with (possibly conditional) low-order moments of $\varepsilon_i$. Despite some theoretical subtleties with the construction of the corrected moment conditions, their practical implementation is straightforward and they can be computed fully automatically for any original moment function $g(\cdot)$.

We introduce the Measurement Error Robust Moments (MERM) estimator, which is a GMM estimator that uses the corrected moment conditions to jointly estimate parameters $\theta_0$ and $\gamma_0$. The estimator can be computed using any standard software for GMM estimation. Joint estimation of parameters $\theta_0$ and $\gamma_0$ using the corrected moment conditions effectively robustifies moment conditions $g(\cdot)$ against the impact of the measurement errors.

To make these ideas precise and to study the properties of the proposed estimators, we use a nonstandard asymptotic approximation that models $\tau$ as slowly shrinking with the sample size. Standard asymptotics considers $\tau$ constant, which implies that as $n \to \infty$ the bias of a naive estimator dwarfs its sampling variability: the bias is constant while the standard errors shrink proportionally to $1/\sqrt{n}$. This implication does not fit many empirical applications, in which the researcher does not expect the potential EIV bias to be several magnitudes larger than the standard errors. Thinking of bias being $\sqrt{n}$ times larger than the standard errors appears to be too pessimistic in a sufficiently wide range of applications to warrant our interest in the alternative settings with measurement errors of a moderate magnitude.[2] By considering $\tau$ as drifting towards zero with the sample size, our approach provides a better approximation of the finite sample properties of estimators when $\tau$ is small or moderate.[3]

Using this approximation, we show that the proposed estimation approach indeed addresses the EIV problem. The MERM estimator is shown to be $\sqrt{n}$-consistent and asymptotically normal and unbiased. The standard confidence intervals and tests for GMM estimators are also shown to be valid for the MERM estimator.

A large sample theory is only useful if it is able to approximate finite sample proper-

[2]Such empirical settings appear to be widespread. Although the concerns about measurement errors are often raised, the majority of applied work does not explicitly correct the EIV bias (especially in nonlinear models), and instead implicitly or explicitly argues or conjectures that the EIV bias is likely not to be too large. See also the review of Bound, Brown, and Mathiowetz (2001).

[3]Such nonstandard asymptotic approximations are often used to obtain better approximations of the finite sample behavior of estimators and tests. For example, in the instrumental variable regression settings, to consider the settings with relatively small first stage coefficients, Staiger and Stock (1997) model them as shrinking with the sample size. It is important to keep in mind that such nonstandard asymptotic approximations are merely mathematical tools. One should not take them literally and think of parameters somehow changing if more data is collected.

ties of the estimators and inference procedures. Thus, we study the MERM estimators in a variety of simulation experiments. The results confirm that the nonstandard asymptotic theory indeed provides a good approximation of the finite sample properties of the estimators in the settings even with relatively large EIV. In some of the simulation experiments, the EIV are so large that for the naive estimators standard 95% confidence intervals have actual coverages of 0% in finite samples, due to the magnitude of the EIV bias. At the same time, the MERM estimators perform well, removing the EIV bias and providing confidence intervals with correct coverage. In particular, the simulation results show that despite the simplicity of implementation, the MERM estimators can compete with and outperform semi-nonparametric estimators. We also illustrate the performance of the MERM estimators in settings where no alternative estimators exist.

General moment condition model (1) encompasses a wide variety of semiparametric models. In general nonlinear models, measurement errors in any of the variables (including outcomes) may bias naive estimators. Mismeasured variables $X_i$ do not need to be covariates. For example, in estimation of Euler equations using individual-level data, $X_i^*$ may represent consumption, which enters the moments nonlinearly and is often mismeasured. However, most of the existing methods for dealing with EIV focus on the nonlinear regression (NLR) model.

**Example 1.1** (NLR)**.** *Let $Y_i$ denote the scalar outcome, and let $X_i^*$ and $W_i$ be the vectors of mismeasured and correctly measured covariates, respectively. Suppose*

$$E\left[Y_i|X_i^*, W_i\right] = \rho\left(X_i^*, W_i, \theta_0\right)$$

*for some known function $\rho$. For example, in the logit model of binary choice, $Y_i$ is binary and $\rho\left(x, w, \theta\right) \equiv \Lambda\left(\theta_X' x + \theta_W' w\right)$, where $\theta \equiv \left(\theta_X', \theta_W'\right)'$ and $\Lambda$ is the logistic CDF.*

To estimate the NLR model, consider the moment function

$$g\left(y, x, w, z; \theta\right) \equiv \left(y - \rho\left(x, w, \theta\right)\right) h\left(x, w, q\right),$$

where $h\left(\cdot\right)$ is a vector valued function, which generally is also allowed to depend on $\theta$. To address the EIV problem in NLR, researchers typically rely on the orthogonality conditions that involve some additional variables $q$. Most often, variables $q$ are some instrumental variables. In the absence of genuine excluded variables, the literature has also advocated relying on nonlinear functions of $x$ (e.g., higher powers) as $q$. Including powers of $x$ may help identification of the model with EIV, and generally improves the

efficiency of the estimator even in the absence of measurement errors. To provide an example of function $h(\cdot)$, for simplicity of notation, suppose $x$ and an instrumental variable $z$ are scalars. One may take $h(x, w, z) = ((1, x, \ldots, x^{m_X}) \otimes (1, z, \ldots, z^{m_Z}), w')'$, where $m_X$ and $m_Z$ are some positive integers. Since moment condition (1) is stated in terms of the correctly measured $X_i^*$, determining what functions $g(\cdot)$ (or $h(\cdot)$ in the NLR model) satisfy this moment condition does not involve any consideration of the measurement errors and is straightforward.

The specific source of identification of parameters $\theta_0$ and $\gamma_0$ depends on the particular model described by the moment conditions $g(\cdot)$. One important advantage of our approach is that it accommodates a variety of sources of identification within a single unified estimation framework. Since MERM estimator is a standard GMM estimator, the analysis of identification can be conducted in familiar terms, and is not obscured by the technical complications such as Fourier transforms, completeness conditions, or ill-posed inverse problems. The bias correction approach and the corresponding structure of the corrected moment conditions dictate what features of moments $g(\cdot)$ are used to identify the nuisance parameters $\gamma_0$. Identification of $\theta_0$ is then based on the information remaining in the moment conditions that are orthogonal to the ones used to identify $\gamma_0$. Essentially, our approach first characterizes in what directions the measurement errors can bias the moment conditions $\mathbb{E}[g(X_i, S_i, \theta)]$, and then uses the moments orthogonal to those directions for estimation of $\theta_0$. Thanks to the MERM estimator being a GMM estimator, the standard arsenal of assessment tools in the GMM framework can be applied to the MERM estimator, allowing one to test model identification, conduct valid inference, and perform model specification diagnostics.

The power of the general MERM approach can be illustrated in the NLR model. The approach can make use of instrumental variables, repeated measurements, and nonlinearity/non-normality for model identification. For example, when a candidate instrumental variable is available, the conditions that an instrument needs to satisfy are much weaker than what is required by many existing approaches. Availability of a discrete instrument is sufficient for identification; and the instrument is allowed to have heterogeneous impact on covariates $X_i^*$.[4] The instrument itself can also be mismeasured, and

[4]The importance of the heterogeneity of the impacts of instruments in empirical applications has been widely recognized, e.g., see Imbens and Angrist (1994); Heckman and Vytlacil (1998); Imbens and Newey (2009). Note that such heterogeneity is ruled out by the EIV-robust methods that rely on the additive control variable assumption for identification, i.e., assume that $X_i^* = m(Z_i) + V_i$ with the control variable $V_i$ independent from $Z_i$. In contrast, in Section 4 we illustrate identification in a random coefficient first stage model.

a variable that is caused by $X_i^*$ can also be used as an instrument in the MERM approach. In Section 4, we study identification of the NLR model in the MERM framework, and show that this model is globally nonparametrically identified using such a discrete instrument.

The simplicity and practicality of the MERM approach do come at cost: there is a limit on the magnitude of the measurement errors it can handle. For example, one generally should not expect the MERM approach to work well when $\tau > 1$, i.e., when the noise dominates the signal; in this case the researcher should seek an alternative estimation method. We view the MERM approach as providing a bridge between the settings in which the measurement errors are guaranteed to be absent or negligible, and the settings where the measurement errors are so large that one has to use the relatively more complicated estimators from the earlier literature (if they exist at all).

**Related Literature** Chen, Hong, and Nekipelov (2011), Schennach (2016), and Schennach (2020) provide excellent overviews of the measurement error literature.

The existing semiparametric approaches to estimation and inference in models with EIV involve nonparametric estimation of infinite-dimensional nuisance parameters (e.g, Chesher, 2000; Li, 2002; Schennach, 2004, 2007; Hu and Schennach, 2008; Schennach and Hu, 2013), simulation (e.g., Schennach, 2014), or both (e.g., Newey, 2001; Wang and Hsiao, 2011). The exceptions include models with linear and polynomial regression functions (see Hausman et al., 1991, 1995), and models where the distributions of unobservables have Gaussian distribution such as Probit and Tobit models with Gaussian covariates and measurement errors (see Smith and Blundell, 1986; Rivers and Vuong, 1988).

To the best of our knowledge, this paper is the first to provide an approach for $\sqrt{n}$-consistent and asymptotically normal and unbiased estimation of general GMM models with EIV, which in addition requires neither nonparametric estimation nor simulation or numerical integration.

We are able to provide such an estimator because we focus on models with moderate measurement errors. Modeling variance of the measurement error as shrinking to zero with the sample size is a popular approach in Statistics. The method has been proposed by Wolter and Fuller (1982), who use it to construct an approximate MLE estimator of nonlinear regression with Gaussian errors. Following their approach, the Statistics literature has mainly focused on the settings where the moments needed to bias correct the estimators are either known or can be readily estimated from the available repeated

measurements. In Economics, such data is relatively rare. The use of approximations with shrinking variance of measurement errors in Econometrics literature has been pioneered by Kadane (1971), Amemiya (1985), and Chesher (1991). These approximations allow one to derive approximate bias formulas of (naive) estimators. These formulas depend on the moments of the measurement errors that are usually not known. One way the Econometrics literature proposed to use such bias calculations is to use them to check the sensitivity of the estimates to EIV by considering how the estimates vary as the unknown moments of the measurement variables take values in some set of plausible values, e.g., see Chesher and Schluter (2002), Chesher, Dumangane, and Smith (2002), Battistin and Chesher (2014), Chesher (2017), and Hong and Tamer (2003).

This paper differs from the earlier literature in several ways. First, it presents a way to estimate the unknown nuisance parameters (moments of the measurement errors) jointly with the parameters of interest. As a result, the approach can, for example, use instrumental variables as a source of parameter identification. Second, the method applies to a very general class of semiparametric models specified by moment conditions. Third, MERM approach allows the measurement errors to have larger magnitudes than most of the papers in the earlier literature; this is achieved by the MERM approach recursively bias correcting the bias correction terms. Fourth, our approach allows non-classical measurement errors.

The most widespread approach to identification of the EIV models in economic applications is to use instrumental variables, e.g., see Hausman et al. (1991); Newey (2001); Schennach (2007); Wang and Hsiao (2011). In a recent paper, Hahn, Hausman, and Kim (2021) reconsider the regression model in Amemiya (1990) using a bias correction similar to ours. When proper excluded variables are not available, researchers have considered using higher moments of $X_i$ as instruments, e.g., see Reiersøl (1950); Lewbel (1997); Erickson and Whited (2002); Schennach and Hu (2013); Ben-Moshe, D'Haultfœuille, and Lewbel (2017). MERM estimator accommodates these identification approaches within a unified estimation framework.

Kitamura, Otsu, and Evdokimov (2013); Andrews, Gentzkow, and Shapiro (2017); Armstrong and Kolesár (2021); Bonhomme and Weidner (2021), among others, develop tools for estimation and inference in the GMM settings that are robust to general perturbation or misspecification of the true data generating process. They focus on the settings in which these perturbations are sufficiently small, so that the naive estimators remain $\sqrt{n}$-consistent, and their biases are of the same order of magnitude as their standard errors. In contrast, we focus on the more specific forms of data contamination due to

the EIV. This allows the MERM approach to remain valid even in the settings with larger measurement errors, in which naive estimators may have slower than $\sqrt{n}$ rates of convergence.

The MERM approach also provides a useful foundation for dealing with EIV in the more complicated settings. Evdokimov and Zeleneev (2018) utilize the MERM framework to address an issue of nonstandard inference that turns out to arise generally in the measurement error settings. Evdokimov and Zeleneev (2019) extend the analysis of this paper to long panel and network settings.

**Organization of the paper** The rest of the paper is organized as follows. Section 2 introduces the Moderate Measurement Error framework and the proposed MERM estimator. Section 3 presents several Monte Carlo experiments that illustrate finite sample properties of the MERM estimators. Section 4 studies nonparametric identification in our settings. Section 5 considers several extensions of the framework.

Several appendices include more technical details. Appendix A provides the regularity conditions used to establish the asymptotic theory for the MERM estimator. Appendices C and F include some details on implementation of the estimator and on the empirical illustration. Appendices D and E provide proofs of the results in Sections 2 and 4, respectively.

# 2 Moderate Measurement Errors Framework

To present the main ideas we first consider the case of univariate $X_i^*$ and classical measurement error $\varepsilon_i$. Later we consider multivariate $X_i^*$ and non-classical measurement errors $\varepsilon_i$.

## 2.1 Moment Conditions and Estimator

To develop a practical estimation approach for general moment condition models we focus on the settings in which $\tau \equiv \sigma_\varepsilon / \sigma_{X^*}$ is small or moderate. We consider an asymptotic approximation with $\tau_n \equiv \tau \to 0$ as $n \to \infty$. Note that economically meaningful parameters are usually invariant to rescaling of $X_i^*$. Likewise, the extent of the EIV problem does not change with such rescaling.

The magnitude of the EIV bias of such parameters is also invariant to such scaling; changing the units of measurement of $X_i$ does not meaningfully change the EIV problem. For simplicity of exposition, it is convenient to assume that $X_i^*$ is scaled so that $\sigma_{X^*}$ is of order one and, correspondingly, moments $\mathbb{E}\left[|\varepsilon_i|^k\right] \propto \tau_n^k$ decrease with $k$ when $\tau_n < 1$. For example, this could be ensured by normalizing observed $X_i$ to have $\sigma_X = 1$. Let us stress that this normalization is used only to simplify the exposition; as we show in Appendix B, the proposed MERM estimator does not require any normalizations in practice. Following the rest of the literature, we assume that $\mathbb{E}[\varepsilon_i] = 0$.[5]

**Special Case: Quadratic Expansion** For clarity, we first consider a simple special case of the general approach. Let us denote $g_x^{(k)}(x, s, \theta) \equiv \partial^k g(x, s, \theta)/\partial x^k$. Since $\mathbb{E}\left[|\varepsilon_i|^k\right] \propto \tau_n^k \to 0$ as $n \to \infty$, under some regularity conditions, we can write the quadratic Taylor expansion of function $g(X_i, S_i, \theta) = g(X_i^* + \varepsilon_i, S_i, \theta)$ around $\varepsilon_i = 0$ as

$$\begin{aligned}\mathbb{E}[g(X_i, S_i, \theta)] &= \mathbb{E}\left[g(X_i^*, S_i, \theta) + g_x^{(1)}(X_i^*, S_i, \theta)\varepsilon_i + \frac{1}{2}g_x^{(2)}(X_i^*, S_i, \theta)\varepsilon_i^2\right] + O(\mathbb{E}[|\varepsilon_i|^3]) \\ &= \mathbb{E}[g(X_i^*, S_i, \theta)] + \frac{\mathbb{E}[\varepsilon_i^2]}{2}\mathbb{E}[g_x^{(2)}(X_i^*, S_i, \theta)] + O(\tau_n^3), \end{aligned} \tag{3}$$

where the second equality holds because because $\varepsilon_i$ and $(X_i^*, S_i)$ are independent, and $\mathbb{E}[\varepsilon_i] = 0$.

This expansion implies that $\mathbb{E}[g(X_i, S_i, \theta_0)] = O(\sigma_\varepsilon^2) = O(\tau_n^2)$. As a result, a naive estimator that ignores the EIV and uses $X_i$ in place of $X_i^*$ has EIV bias of order $\tau_n^2$.[6] Bias of the naive estimator should be compared with its standard error, which is of order $n^{-1/2}$. Bias of the naive estimator is not negligible, unless the measurement error is rather small (theoretically, unless $\tau_n^2 = o\left(n^{-1/2}\right)$). In particular, tests and confidence intervals based on the naive estimator are invalid and can provide highly misleading results. Moreover, if $\tau_n^2$ shrinks at a rate slower than $O\left(n^{-1/2}\right)$, the rate of convergence of the naive estimator is slower than $\sqrt{n}$.

Suppose $\tau_n = o\left(n^{-1/6}\right)$. Then, $O(\tau_n^3) = o\left(n^{-1/2}\right)$ and we can rearrange equation (3) as

$$\mathbb{E}[g(X_i^*, S_i, \theta)] = \mathbb{E}[g(X_i, S_i, \theta)] - \frac{\mathbb{E}[\varepsilon_i^2]}{2}\mathbb{E}\left[g_x^{(2)}(X_i^*, S_i, \theta)\right] + o(n^{-1/2}). \tag{4}$$

The left-hand side of this equation is exactly the moment condition (1) that we would like

[5] A location normalization such as $\mathbb{E}[\varepsilon_i] = 0$ is usually necessary because it is not possible to separately identify the means $\mathbb{E}[X_i^*]$ and $\mathbb{E}[\varepsilon_i]$.

[6] For example, consider a linear regression with a scalar mismeasured regressor. The bias of the naive OLS estimator of the slope parameter $\theta_{01}$ is $-\theta_{01}\frac{\tau_n^2}{1+\tau_n^2} = -\theta_{01}\tau_n^2 + O\left(\tau_n^4\right)$.

to use for estimation of $\theta_0$. The first term on the right-hand side involves only observed variables, and can be estimated by the sample average $\overline{g}(\theta) \equiv n^{-1} \sum_{i=1}^{n} g(X_i, S_i, \theta)$. The second term on the right-hand side can be thought of as a bias correction that removes the EIV-bias from the expected moment function $\mathbb{E}[g(X_i, S_i, \theta)]$.

The idea of the MERM estimator we propose is to make use of expansions such as (4) to bias correct the moment condition $\mathbb{E}[g(X_i, S_i, \theta)]$, which in turn removes the bias of the estimator of the parameters of interest $\theta_0$. To perform the bias correction we need to estimate two quantities: $\mathbb{E}[\varepsilon_i^2]$ and $\mathbb{E}\left[g_x^{(2)}(X_i^*, S_i, \theta)\right]$.

First, we show that in equation (4) we can substitute $\mathbb{E}\left[g_x^{(2)}(X_i^*, S_i, \theta)\right]$ with $\mathbb{E}\left[g_x^{(2)}(X_i, S_i, \theta)\right]$, which in turn can be estimated by $\overline{g}_x^{(2)}(\theta) \equiv n^{-1} \sum_{i=1}^{n} g_x^{(2)}(X_i, S_i, \theta)$. By the Taylor expansion around $\varepsilon_i = 0$ similar to equation (3), we can show that $\mathbb{E}\left[g_x^{(2)}(X_i^*, S_i, \theta)\right] = \mathbb{E}\left[g_x^{(2)}(X_i, S_i, \theta)\right] + O(\tau_n^2)$ and hence

$$\frac{1}{2}\mathbb{E}[\varepsilon_i^2]\left(\mathbb{E}\left[g_x^{(2)}(X_i^*, S_i, \theta)\right] - \mathbb{E}\left[g_x^{(2)}(X_i, S_i, \theta)\right]\right) = \mathbb{E}[\varepsilon_i^2] O\left(\tau_n^2\right) = O\left(\tau_n^4\right). \tag{5}$$

Here $O\left(\tau_n^4\right) = o\left(n^{-1/2}\right)$ because we assume that $\tau_n = o\left(n^{-1/6}\right)$. The idea behind this substitution is that the bias of order $O(\tau_n^2)$ in $\mathbb{E}\left[g_x^{(2)}(X_i, S_i, \theta)\right]$ can be ignored because it is multiplied by $E\left[\varepsilon_i^2\right] = O\left(\tau_n^2\right)$.[7] With the substitution, we can rearrange equation (4) and write it as

$$\mathbb{E}[g(X_i^*, S_i, \theta)] = \mathbb{E}\left[g(X_i, S_i, \theta) - \frac{\mathbb{E}[\varepsilon_i^2]}{2} g_x^{(2)}(X_i, S_i, \theta)\right] + o(n^{-1/2}). \tag{6}$$

Second, we propose estimating the unknown $\mathbb{E}[\varepsilon_i^2]$ together with the parameter of interest $\theta$. Specifically, let $\gamma_{02} \equiv \mathbb{E}[\varepsilon_i^2]/2$ denote the true value of parameter $\gamma_2$, and consider the following *corrected moment function*:

$$\psi(X_i, S_i, \theta, \gamma) \equiv g(X_i, S_i, \theta) - \gamma_2 g_x^{(2)}(X_i, S_i, \theta). \tag{7}$$

Function $\psi$ is a moment function parameterized by $\theta$ and $\gamma$, and

$$\mathbb{E}[\psi(X_i, S_i, \theta_0, \gamma_{02})] = \mathbb{E}[g(X_i^*, S_i, \theta_0)] + o\left(n^{-1/2}\right) = o\left(n^{-1/2}\right), \tag{8}$$

where the first equality follows from equation (6) and the definition of $\gamma_{02}$, and the second equality follows from equation (1). Hence, the corrected moment conditions $\psi$ can be

[7] Such substitutions of $X^*$ with $X$ have been used in other contexts, e.g., Chesher and Schluter (2002).

used to jointly estimate the true parameters $\theta_0$ and $\gamma_{02}$ by a GMM estimator.[8]

**Remark 1.** *If $\mathbb{E}[\varepsilon_i^3] = 0$ (e.g., if the distribution of $\varepsilon_i$ is symmetric), the remainder in equation (3) is of a smaller order $O(\tau_n^4)$. Hence, the corrected moments (8) remain valid for larger values of $\tau_n$, requiring only the weaker condition $\tau_n = o(n^{-1/8})$. The bias of the naive estimators in this case can be as large as $o(n^{-1/4})$.*

**General Case: Expansion of order $K$** The quadratic expansion of equation (3) can be extended to general order $K \geq 2$. Considering larger $K$ theoretically allows $\tau_n$ converging to zero at a slower rate. In finite samples this corresponds to the asymptotics providing good approximations for larger values of $\tau_n$, i.e., large measurement errors. Expanding $g(X_i^* + \varepsilon_i, S_i, \theta)$ around $\varepsilon_i = 0$ we have,

$$\mathbb{E}[g(X_i, S_i, \theta)] = \mathbb{E}\left[g(X_i^*, S_i, \theta) + \sum_{k=1}^{K} \frac{\varepsilon_i^k}{k!} g_x^{(k)}(X_i^*, S_i, \theta)\right] + O\left(\mathbb{E}\left[|\varepsilon_i|^{K+1}\right]\right). \tag{9}$$

The above special case of quadratic expansion corresponds to $K = 2$.

The approximation we consider is formalized by the following assumption.

**Assumption MME.** (Moderate Measurement Errors) *(i) $\tau_n = o(n^{-1/(2K+2)})$ for some integer $K \geq 2$; and (ii) $\mathbb{E}[|\varepsilon_i|^L] \leqslant C\sigma_\varepsilon^L$ for some $L \geq K+1$ and $C > 0$.*

Assumption MME (i) limits the magnitude of the measurement errors and implies that $\tau_n^{K+1} = o\left(n^{-1/2}\right)$. Assumption MME (ii) implies that $\mathbb{E}\left[|\varepsilon_i|^k\right] = O\left(\sigma_\varepsilon^k\right)$, and requires the tails of $\varepsilon_i/\sigma_\varepsilon$ to be sufficiently thin. Together, parts (i) and (ii) imply that $\mathbb{E}\left[|\varepsilon_i|^{K+1}\right] = O\left(\tau_n^{K+1}\right) = o\left(n^{-1/2}\right)$, and hence ensure that the remainder in equation (9) is negligible. Using $\mathbb{E}\left[\varepsilon_i | X_i^*, S_i\right] = 0$ to further simplify this expansion and rearranging the terms we obtain

$$\mathbb{E}[g(X_i^*, S_i, \theta)] = \mathbb{E}[g(X_i, S_i, \theta)] - \sum_{k=2}^{K} \frac{\mathbb{E}[\varepsilon_i^k]}{k!} \mathbb{E}\left[g_x^{(k)}(X_i^*, S_i, \theta)\right] + o(n^{-1/2}). \tag{10}$$

This equation is the general expansion analog of equation (4). The summation on the right hand side is the bias correction term, which we use to construct the MERM estimator.

[8] In the moment condition settings, having $o\left(n^{-1/2}\right)$ is equivalent to having 0 on the right-hand side of equation (8).

It turns out that for $K \geq 4$, estimation of $\mathbb{E}\left[g_x^{(k)}(X_i^*, S_i, \theta)\right]$ is more intricate than in the case of $K = 2$, and the substitution we made in equation (6) no longer works. Larger values of $K$ allow for larger values of $\tau_n$ and hence larger EIV biases of naive estimators $n^{-1}\sum_{i=1}^n g_x^{(k)}(X_i, S_i, \theta)$. The expansion of order $K$ includes terms up to the order $\tau_n^K$, with the asymptotically negligible remainder of order $O\left(\tau_n^{K+1}\right)$. For $K \geq 4$, terms of order $\tau_n^4$ are not negligible. This implies that we cannot ignore the EIV bias that would arise from substituting $\mathbb{E}\left[g_x^{(2)}(X_i^*, S_i, \theta)\right]$ with $\mathbb{E}\left[g_x^{(2)}(X_i, S_i, \theta)\right]$ in equation (10), because this bias is of order $O\left(\tau_n^4\right)$ according to equation (5). To address this problem, we instead replace $\mathbb{E}\left[g_x^{(2)}(X_i^*, S_i, \theta)\right]$ with the bias corrected expression $\mathbb{E}\left[g_x^{(2)}(X_i, S_i, \theta)\right] - \left(\mathbb{E}[\varepsilon_i^2]/2\right)\mathbb{E}\left[g_x^{(4)}(X_i, S_i, \theta)\right]$. Thus, for $K \geq 4$, one needs to bias correct the estimator of the bias correction term. Moreover, for larger $K$ one needs to bias correct the bias correction of the bias correction term and so on.

Fortunately, we show that these bias corrections can be constructed as linear combinations of the expectations of the higher order derivatives of $g_x^{(k)}(X_i, S_i, \theta)$. Let us define the following *corrected moment function*:

$$\psi(X_i, S_i, \theta, \gamma) \equiv g(X_i, S_i, \theta) - \sum_{k=2}^{K} \gamma_k g_x^{(k)}(X_i, S_i, \theta), \tag{11}$$

where $\gamma = (\gamma_2, \ldots, \gamma_K)'$ is a $K-1$ dimensional vector of parameters. Let $\gamma_0 \equiv (\gamma_{02}, \ldots, \gamma_{0K})'$ denote the vector of true parameters $\gamma_{0k}$, defined as

$$\gamma_{02} \equiv \frac{\mathbb{E}\left[\varepsilon_i^2\right]}{2},\ \gamma_{03} \equiv \frac{\mathbb{E}\left[\varepsilon_i^3\right]}{6},\ \text{and}\ \gamma_{0k} \equiv \frac{\mathbb{E}\left[\varepsilon_i^k\right]}{k!} - \sum_{\ell=2}^{k-2} \frac{\mathbb{E}\left[\varepsilon_i^{k-\ell}\right]}{(k-\ell)!}\gamma_{0\ell}\ \text{for}\ k \geq 4.$$

We formalize this discussion below.

**Assumption CME.** *$\varepsilon_i$ is independent from $(X_i^*, S_i)$ and $\mathbb{E}[\varepsilon_i] = 0$.*

Assumption CME is the classical measurement error assumption. We relax this assumption later in Section 5.2. The following lemma establishes validity of the corrected moment conditions under Assumptions MME, CME, and some mild regularity conditions provided in Appendix A.

**Lemma 1.** *Under Assumptions MME, CME and A.1 in Appendix A,*

$$\mathbb{E}[\psi(X_i, S_i, \theta_0, \gamma_0)] = \mathbb{E}[g(X_i^*, S_i, \theta_0)] + o\left(n^{-1/2}\right) = o\left(n^{-1/2}\right).$$

Lemma 1 implies that the corrected moment conditions $\psi$ can be used to jointly estimate parameters $\theta_0$ and $\gamma_0$. The total number of parameters to be estimated is now $\dim(\theta) + K - 1$. Thus, joint estimation of $\theta_0$ and $\gamma_0$ requires that $\dim(\psi) = \dim(g) \geq \dim(\theta) + K - 1$, i.e., that the original moment conditions $g$ include sufficiently many overidentifying restrictions.

**Measurement Error Robust Moments (MERM) estimator** The MERM estimator jointly estimates the parameters $\theta_0$ and $\gamma_0$ using moment conditions $\psi$. It is convenient to define the joint vector of parameters

$$\beta \equiv (\theta', \gamma')', \ \beta_0 \equiv (\theta_0', \gamma_0')', \ \hat{\beta} \equiv (\hat{\theta}', \hat{\gamma}')',$$

and the parameter space $\mathcal{B} \equiv \Theta \times \Gamma$, where $\Theta$ and $\Gamma$ are the parameter spaces for $\theta$ and $\gamma$. Then, MERM estimator is the GMM estimator (Hansen, 1982):

$$\hat{\beta} \equiv \underset{\beta \in \mathcal{B}}{\operatorname{argmin}} \hat{Q}(\beta), \qquad \hat{Q}(\beta) \equiv \overline{\psi}(\beta)' \hat{\Xi} \overline{\psi}(\beta), \tag{12}$$

where $\overline{\psi}(\beta) \equiv n^{-1} \sum_{i=1}^{n} \psi_i(\beta)$, $\psi_i(\beta) \equiv \psi(X_i, S_i, \beta)$, $\hat{\Xi}$ is a weighting matrix, and $\hat{Q}(\beta)$ is the standard GMM objective function.

Under some regularity conditions, estimator $\hat{\beta}$ behaves as a standard GMM-type estimator: it is $\sqrt{n}$-consistent and asymptotically normal and unbiased. This result is formalized by the following theorem.

**Theorem 2** (Asymptotic Normality)**.** *Suppose that* $\{(X_i^*, S_i', \varepsilon_i)\}_{i=1}^n$ *are i.i.d.. Then, under Assumptions MME, CME, and A.1-A.4 in Appendix A,*

$$n^{1/2} \Sigma^{-1/2} (\hat{\beta} - \beta_0) \xrightarrow{d} N(0, I_{p+K-1}), \ \textit{where} \tag{13}$$

$$\Sigma \equiv (\Psi' \Xi \Psi)^{-1} \Psi' \Xi \Omega_{\psi\psi} \Xi \Psi (\Psi' \Xi \Psi)^{-1}. \tag{14}$$

Theorem 2 shows that the MERM approach addresses the EIV bias problem, and in particular provides a $\sqrt{n}$-consistent asymptotically normal and unbiased estimator $\hat{\theta}$, which can be used to conduct inference about the true parameters $\theta_0$. The asymptotic variance $\Sigma$ takes the standard sandwich form, with $\Psi \equiv \mathbb{E}[\nabla_\beta \psi_i(\beta_0)]$, $\Omega_{\psi\psi} \equiv \mathbb{E}[\psi_i(\beta_0) \psi_i'(\beta_0)]$, and $\hat{\Xi} \to_p \Xi$.

**Remark 2.** *Notice that the bias of naive estimators (such as a GMM estimator based on the original moment conditions) is* $O(\tau_n^2)$*, so their rate of convergence is* $O_p(\tau_n^2 + n^{-1/2})$*.*

*The bias dominates sampling variability and naive estimators are not $\sqrt{n}$-consistent unless $\tau_n = O(n^{-1/4})$, i.e., unless the magnitude of the measurement error is rather small. At the same time, the MERM estimator remains $\sqrt{n}$-consistent for much larger values of $\tau_n$, up to $\tau_n = O(n^{-1/(2K+2)})$, whereas the rate of convergence of naive estimators is only $O_p(n^{-1/(K+1)})$ in this case.*

Once the corrected moment condition $\psi$ is constructed, estimation of and inference about parameters $\beta_0$ can be performed using any standard software package for GMM estimation. In other words, the proposed estimator can be simply treated as a standard GMM estimator based on the corrected moment conditions $\psi$, and the conventional standard errors, tests, and confidence intervals are valid.

The validity of the MERM estimator and equation (13) require that parameters $\beta_0$ are identified and the Jacobian matrix $\Psi$ has full rank. One advantage of the approach is that the estimation method remains the same regardless of the specific source of identification is used. The identifying information is incorporated in the moment functions. In particular, Section 4 we study identification of the nonlinear regression model, and show that it is nonparametrically identified using a (possibly discrete) instrument. In Section 5.1 we consider multivariate measurement errors, with repeated measurements as a special case. Importantly, identification can be tested by testing that $\Psi$ has full rank, see also Remark 10 below.

**Remark 3.** *Researchers may be interested in average effects of the form $\mu_0 \equiv E\left[h\left(X_i^*, S_i, \theta_0\right)\right]$. In the NLR, one may be interested in the average partial effect $x$ (i.e., $\mu_0 \equiv E\left[\nabla_x \rho\left(X_i^*, S_i, \theta_0\right)\right]$) or another covariate. The naive average partial effect estimator $\hat{\mu}_{Naive} \equiv \frac{1}{n}\sum_{i=1}^n h\left(X_i, S_i, \hat{\theta}\right)$ suffers from the EIV bias, unless function $h$ is linear in $X_i^*$. Instead, one can use estimates $\hat{\gamma}$ to construct the bias-corrected estimator $\hat{\mu}_{MERM} \equiv \frac{1}{n}\sum_{i=1}^n \left\{h\left(X_i, S_i, \hat{\theta}\right) - \sum_{k=2}^K \hat{\gamma}_k h_x^{(k)}(X_i, S_i, \hat{\theta})\right\}$.*

**Remark 4.** *It is useful to get a sense of the magnitudes of the coefficients $\mathbb{E}\left[\varepsilon_i^k\right]/k!$ in equation (10). Suppose $\varepsilon_i \sim N\left(0, \sigma_\varepsilon^2\right)$, $\sigma_\varepsilon = 0.5$, and $\sigma_{X^*} = 1$, so $\tau = \sigma_\varepsilon = 0.5$. Then the coefficients in front of $g_x^{(2)}$, $g_x^{(4)}$, and $g_x^{(6)}$ are $\mathbb{E}\left[\varepsilon_i^2\right]/2! = 0.125$, $\mathbb{E}\left[\varepsilon_i^4\right]/4! \approx 0.008$, and $\mathbb{E}\left[\varepsilon_i^6\right]/6! \approx 0.0003$.*

**Remark 5.** *It is important to note that $\gamma_{0k} \neq \mathbb{E}\left[\varepsilon_i^k\right]/k!$ for $k \geq 4$, contrary to what equation (10) might suggest. For example, $\gamma_{04} = \left(\mathbb{E}\left[\varepsilon_i^4\right] - 6\sigma_\varepsilon^4\right)/24$ is negative for many distributions, including normal. For instance, in the example of Remark 4, $\gamma_{04} \approx -0.0026$. The reason that generally $\gamma_{0k} \neq \mathbb{E}\left[\varepsilon_i^k\right]/k!$ is that the estimators of the correction terms*

*themselves need a correction, which is accounted for by the form of $\gamma_{0k}$. Since there is a one-to-one relationship between $\gamma_0$ and the moments $\mathbb{E}\left[\varepsilon_i^\ell\right]$, parameter space $\Gamma$ for $\gamma_0$ can incorporate restrictions that the moments must satisfy (e.g., $\sigma_\varepsilon^2 \geq 0$ and $\mathbb{E}\left[\varepsilon_i^4\right] \geq \sigma_\varepsilon^4$). Such restrictions can increase the efficiency of the estimator and the power of tests.*

**Remark 6.** *No parametric assumptions are imposed on the distribution of $\varepsilon_i$, i.e. the distribution of $\varepsilon_i$ is treated nonparametrically. The regularity conditions restrict only the magnitude of the moments of $\varepsilon_i$. The approach imposes no restrictions on the smoothness of the distributions of $X_i^*$ and $\varepsilon_i$, which are not even required to be continuous. Examples in which this can be useful include individual wages (whose distribution tends to have point masses at round numbers), and allowing the measurement error $\varepsilon_i$ to have a point mass at zero (a fraction of the population may have zero measurement or recall error).*

**Remark 7.** *Considering larger $K$ allows $\tau_n$ converging to zero at a slower rate, which in finite samples corresponds to the asymptotics providing better approximations for larger magnitudes of measurement errors. On the other hand, taking a larger $K$ increases the dimension of the nuisance parameter $\gamma_0$ and thus typically increases the variance of $\hat{\theta}$.*

**Remark 8.** *The usual $J$-test remains valid in the MERM settings, and can be used to check model specification. Rejection of the $J$-test may be caused by* (i) *$K$ being too small, and hence failing to entirely correct the EIV bias;* (ii) *assumptions on the EIV being invalid; or* (iii) *misspecification of the moment conditions $g$, i.e., equation (1) failing to hold. Issue* (i) *can be addressed by taking a larger $K$. In Section 5 we extend the framework to obtain corrected moments that are valid under weaker assumptions on the EIV, which can help addressing issue* (ii).

**Remark 9.** *The formulas of derivatives $g_x^{(k)}(\cdot)$ are typically easy to compute analytically or using symbolic algebra software. Alternatively, derivatives $g_x^{(k)}$ can be computed using numerical differentiation. In either way, the corrected set of moments can be automatically produced for a generic moment function $g(\cdot)$ provided by the user.*

**Remark 10.** *The framework encompasses many possible sources of identification, including instrumental variables, repeated measurements, and nonlinearities of the functional form. Importantly, one can test the strength of identification of the model parameters or even conduct identification-robust inference (e.g., Stock and Wright (2000); Kleibergen (2005); Guggenberger and Smith (2005); Guggenberger, Ramalho, and Smith (2012); Andrews and Mikusheva (2016); Andrews (2016); Andrews and Guggenberger (2019)).*

# 3 Numerical Evidence

## 3.1 Comparison with a Semi-Nonparametric Estimation Approach

We compare MERM estimator with the state-of-the-art semiparametric estimator of Schennach (2007, henceforth S07) for nonlinear regression models. The Monte Carlo designs are taken from S07, and include a polynomial, rational fraction, and Probit nonlinear regression models. Identification of the model is ensured by the availability of an instrument.

$$Y_i = \rho(X_i^*, \theta_0) + U_i, \quad X_i^* = Z_i + V_i, \quad X_i = X_i^* + \varepsilon_i, \tag{15}$$

$(Z_i, V_i, \varepsilon_i)' \sim N\left((0,0,0)', \text{Diag}(1, 1/4, 1/4)\right)$ and $n = 1000$. The conditional expectation function $\rho$, the true value of the parameter of interest $\theta_0$, and the conditional distribution of the regression error $U_i$ are design-specific and reported in Tables 1-3 below. In all designs, $\tau = \sigma_\varepsilon / \sigma_X^* \approx 0.45$, so the measurement error is "fairly large" (Schennach, 2007).

We report simulation results for the MERM estimator considering correction schemes with $K = 2$ and $K = 4$. The original moment function is

$$g(x, y, z, \theta) = (y - \rho(x, \theta))h(x, z),$$

where we use $h(x,z) = (1, x, z, x^2, z^2, x^3, z^3)'$ for $K = 2$ and $h(x,z) = (1, x, z, x^2, xz, z^2, x^3, x^2z, xz^2, z^3)'$ for $K = 4$.

The finite sample properties of the MERM estimators (evaluated based on 5,000 replications) are reported in Tables 1-3 below. For comparison, we also provide the same statistics for naive estimators (OLS/NLLS) and for the benchmark estimator of S07 (as reported in the original paper). For the polynomial model (Table 1), both $K = 2$ and $K = 4$ MERM estimators effectively remove the EIV bias. Component-wise, the MERM estimators perform similarly (for $\theta_2$ and $\theta_4$) or better (for $\theta_1$ and $\theta_3$) compared to the benchmark estimator of S07. For the rational fraction model (Table 2), both the MERM estimators are vastly superior to the benchmark estimator both in terms of the bias and the standard deviation. For the probit model (Table 3), the MERM estimator with $K = 2$ removes a large fraction of the EIV bias compared to the NLLS estimator. However, the EIV bias remains non-negligible when this simplest correction scheme is used. Employing a higher order correction scheme with $K = 4$ completely eliminates the remaining EIV

bias, while at the same time having smaller standard deviations (than the benchmark estimator of S07) . Overall, in the considered designs, the MERM estimator with $K = 4$ consistently outperforms the benchmark estimator. It also proves to be more effective in removing the EIV bias compared to the $K = 2$ estimator, especially in the highly nonlinear settings of the considered probit design.

Table 1: Simulation results for the polynomial model of S07

| | Bias | | | | Std. Dev. | | | | RMSE | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $\theta_1$ | $\theta_2$ | $\theta_3$ | $\theta_4$ | $\theta_1$ | $\theta_2$ | $\theta_3$ | $\theta_4$ | $\theta_1$ | $\theta_2$ | $\theta_3$ | $\theta_4$ | All |
| OLS | -0.00 | -0.43 | 0.00 | 0.21 | 0.07 | 0.13 | 0.06 | 0.04 | 0.07 | 0.45 | 0.06 | 0.22 | 0.51 |
| S07 | -0.05 | -0.07 | -0.02 | 0.05 | 0.17 | 0.19 | 0.24 | 0.05 | 0.17 | 0.20 | 0.24 | 0.07 | 0.36 |
| $K = 2$ | -0.00 | 0.10 | 0.00 | 0.00 | 0.10 | 0.23 | 0.10 | 0.08 | 0.10 | 0.25 | 0.10 | 0.08 | 0.29 |
| $K = 4$ | -0.00 | 0.00 | 0.00 | 0.02 | 0.09 | 0.21 | 0.10 | 0.08 | 0.09 | 0.21 | 0.10 | 0.08 | 0.27 |

The DGP is as in (15) with $\rho(x, \theta) = \theta_1 + \theta_2 x + \theta_3 x^2 + \theta_4 x^3$, $\theta_0 = (1, 1, 0, -0.5)'$, and $U_i \sim N(0, 1/4)$. The results are based on 5,000 replications.

Table 2: Simulation results for the rational fraction model of S07

| | Bias | | | Std. Dev. | | | RMSE | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | $\theta_1$ | $\theta_2$ | $\theta_3$ | $\theta_1$ | $\theta_2$ | $\theta_3$ | $\theta_1$ | $\theta_2$ | $\theta_3$ | All |
| OLS | 0.339 | -0.167 | -0.644 | 0.040 | 0.020 | 0.076 | 0.341 | 0.168 | 0.648 | 0.752 |
| S07 | 0.107 | 0.117 | -0.150 | 0.146 | 0.139 | 0.328 | 0.181 | 0.182 | 0.361 | 0.443 |
| $K = 2$ | -0.004 | -0.018 | 0.014 | 0.062 | 0.026 | 0.139 | 0.062 | 0.032 | 0.139 | 0.156 |
| $K = 4$ | 0.014 | -0.002 | -0.024 | 0.062 | 0.031 | 0.154 | 0.063 | 0.031 | 0.156 | 0.171 |

The DGP is as in (15) with $\rho(x, \theta) = \theta_1 + \theta_2 x + \frac{\theta_3}{(1+x^2)^2}$, $\theta_0 = (1, 1, 2)'$, and $U_i \sim N(0, 1/4)$. The results are based on 5,000 replications.

## 3.2 Estimation and Inference in a Multinomial Choice Model

Consider the standard multinomial logit model, in which an agent chooses between 3 available options. For an agent $i$ with characteristics $(X_i^*, W_i)$, the utility of option $j$ is given by

$$U_{ij} = \theta_{0j1} X_i^* + \theta_{0j2} W_{ij} + \theta_{0j3} + \epsilon_{ij} \quad \text{for } j \in \{1, 2\},$$

and $U_{i0} = \epsilon_{i0}$ for the outside option $j = 0$, where $\epsilon_{ij}$ are i.i.d. (across $i$ and $j$) draws from a standard type-1 extreme value distribution. Let $Y_{ij}$ be the observed binary outcome

Table 3: Simulation results for the Probit model of S07

| | Bias | | Std. Dev. | | RMSE | | |
|---|---|---|---|---|---|---|---|
| | $\theta_1$ | $\theta_2$ | $\theta_1$ | $\theta_2$ | $\theta_1$ | $\theta_2$ | All |
| NLLS | 0.38 | -0.97 | 0.06 | 0.08 | 0.39 | 0.98 | 1.05 |
| S07 | 0.05 | -0.06 | 0.39 | 0.53 | 0.39 | 0.53 | 0.69 |
| $K = 2$ | 0.11 | -0.31 | 0.18 | 0.34 | 0.21 | 0.46 | 0.51 |
| $K = 4$ | -0.01 | -0.01 | 0.23 | 0.42 | 0.23 | 0.42 | 0.48 |

The DGP is as in (15) with $\rho(x,\theta) = \frac{1}{2}(1 + \text{erf}(\theta_1 + \theta_2 x))$, $\theta_0 = (-1, 2)'$, and $U_i = 1 - \rho(X_i^*, \theta_0)$ with probability $\rho(X_i^*, \theta_0)$ and $-\rho(X_i^*, \theta_0)$ otherwise. The results are based on 5,000 replications.

variable indicating whether agent $i$ chooses option $j$, i.e. $Y_{ij} = 1$ if and only if $j = \text{argmax}_{j' \in \{0,1,2\}} U_{ij'}$. In addition,

$$X_i^* = q(Z_i, V_i), \quad X_i = X_i^* + \varepsilon_i, \quad W_{ij} = \rho X_i^* / \sigma_X^* + \sqrt{1 - \rho^2}\nu_{ij},$$

and $(Z_i, \varepsilon_i, \nu_{i1}, \nu_{i2})' \sim N\left((0,0,0,0)', \text{Diag}(\sigma_Z^2, \sigma_\varepsilon^2, \sigma_\nu^2, \sigma_\nu^2)\right) \perp V_i$. The researcher observes $\{(X_i, W_i, Y_{i1}, Y_{i2}, Y_{i0})\}_{i=1}^n$. Note that the form of function $q(z, v)$ is not known to the researcher. We consider two specifications for function $q(z, v)$ and the distribution of $V_i$. In the first case, $v$ is scalar and $q(z, v)$ is additive,

$$\text{Case A:} \quad q(Z_i, V_i) = Z_i + V_i, \quad V_i \sim N(0, \sigma_V^2). \tag{16}$$

In the second case, $q(z, v)$ is a random coefficient model and $v$ is bivariate:

$$\text{Case RC:} \quad q(Z_i, V_i) = V_{i1} Z_i + V_{i0}, \quad V_i \equiv (V_{i0}, V_{i1})' \sim N\left((0,1)', \text{Diag}(\sigma_{V0}^2, \sigma_{V1}^2)\right), \tag{17}$$

In all of the designs, we fix $(\theta_{011}, \theta_{012}, \theta_{013}, \theta_{021}, \theta_{022}, \theta_{023}, \rho, \sigma_Z^2, \sigma_\nu^2) = (1, 0, 0, 0, 0, 0, 0.7, 1, 1)$ and $n = 2000$. We also fix $\sigma_V^2 = 1$ in Case A and $(\sigma_{V0}^2, \sigma_{V1}^2) = (1/2, 1/2)$ in Case RC; in both cases $\sigma_{X^*}^2 = 2$. We consider $\tau = \sigma_\varepsilon / \sigma_{X^*} \in \{1/4, 1/2, 3/4\}$.

Similarly to Section 3.1, we report results for the MERM estimators applying the same correction scheme with $K = 2$ and $K = 4$ to the following original moment function

$$g(x, w, y, z, \theta) = \left((y_1 - p_1(x, w, \theta))\, h_1(x, z, w)', (y_2 - p_2(x, w, \theta))\, h_2(x, z, w)'\right)',$$

where

$$p_j(x,w,\theta) \equiv \mathbb{P}(Y_{ij}=1|X_i^*=x,W_i=w;\theta) = \frac{\exp(\theta_{j1}x+\theta_{j2}w_j+\theta_{j3})}{1+\exp(\theta_{11}x+\theta_{12}w_1+\theta_{13})+\exp(\theta_{21}x+\theta_{22}w_2+\theta_{23})},$$

and $h_j(x,z,w) = (1,x,z,x^2,z^2,x^3,z^3,w_j)'$ for $K=2$ and $h_j(x,z,w) = (1,x,z,x^2,xz,z^2,x^3,x^2z,xz^2,z^3,w_j)'$ for $K=4$.

We report the results on estimation and inference on the partial derivatives of the conditional choice probabilities $p_j(x,w_1,w_2)$ with respect to $x$, $w_1$, and $w_2$, evaluated at the population means.

Tables 4 and 5 report the finite sample biases, standard deviations, and RMSE of the MERM estimators, as well as the sizes of the corresponding t-tests with nominal size of 5%. To illustrate the importance of dealing with EIV, we also report the same statistics for the standard (naive) MLE estimator that ignores the presence of the measurement errors.

In all designs, the MLE estimator is biased, and the corresponding t-tests over-reject. Note that failing to account for the EIV in the mismeasured variable $X_i^*$ generally biases estimators of all of the parameter, including those corresponding to the correctly measured variables $W_{i1}$ and $W_{i2}$. In particular, the t-tests may falsely reject true null hypotheses $\partial p_j/\partial w_\ell = 0$ up to nearly 100% of the time.

As for the MERM estimators, the $K=2$ estimator removes a large fraction of the EIV bias in all of the designs. While this proves to be enough to achieve accurate size control when the magnitude of the measurement error is moderate ($\tau = 1/4$), the remaining EIV bias may still result in size distortions of the t-tests with larger measurement errors, especially $\tau = 3/4$. Using the higher order correction scheme with $K=4$ effectively removes the EIV bias in all of the simulation designs for all of the parameters. Remarkably, the corresponding finite sample null rejection probabilities remain close to the nominal 5% rate even when the standard deviation of the measurement error is as large as 75% of the standard deviation of the mismeasured $X^*$.

Table 4: Simulation results for the multinomial logit model, Case A

| | MLE | | | | $K=2$ | | | | $K=4$ | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | bias, $10^{-2}$ | std, $10^{-2}$ | rmse, $10^{-2}$ | size | bias, $10^{-2}$ | std, $10^{-2}$ | rmse, $10^{-2}$ | size | bias, $10^{-2}$ | std, $10^{-2}$ | rmse, $10^{-2}$ | size |
| | | | | | | $\tau = 1/4$ | | | | | | |
| $\partial p_1/\partial x$ | -2.93 | 1.33 | 3.22 | 60.98 | 0.48 | 2.50 | 2.55 | 4.38 | 0.77 | 2.67 | 2.78 | 6.16 |
| $\partial p_1/\partial w_1$ | 2.30 | 1.65 | 2.83 | 29.38 | -0.04 | 2.35 | 2.35 | 5.00 | -0.18 | 2.41 | 2.42 | 5.54 |
| $\partial p_1/\partial w_2$ | 0.50 | 0.76 | 0.91 | 10.00 | -0.00 | 0.87 | 0.87 | 5.16 | -0.04 | 0.89 | 0.89 | 5.86 |
| $\partial p_2/\partial x$ | 1.80 | 1.18 | 2.15 | 35.04 | -0.27 | 1.85 | 1.87 | 5.24 | -0.44 | 1.95 | 2.00 | 6.40 |
| $\partial p_2/\partial w_1$ | -1.15 | 0.83 | 1.42 | 29.28 | 0.02 | 1.18 | 1.18 | 4.92 | 0.09 | 1.21 | 1.21 | 5.48 |
| $\partial p_2/\partial w_2$ | -0.99 | 1.50 | 1.80 | 10.06 | 0.01 | 1.75 | 1.75 | 5.20 | 0.08 | 1.79 | 1.79 | 5.88 |
| $\partial p_0/\partial x$ | 1.13 | 1.00 | 1.51 | 21.82 | -0.21 | 1.35 | 1.37 | 4.88 | -0.33 | 1.41 | 1.44 | 5.94 |
| $\partial p_0/\partial w_1$ | -1.15 | 0.82 | 1.42 | 29.30 | 0.02 | 1.18 | 1.18 | 4.98 | 0.09 | 1.21 | 1.21 | 5.50 |
| $\partial p_0/\partial w_2$ | 0.49 | 0.75 | 0.89 | 10.04 | -0.01 | 0.88 | 0.88 | 5.08 | -0.04 | 0.90 | 0.90 | 5.84 |
| | | | | | | $\tau = 1/2$ | | | | | | |
| $\partial p_1/\partial x$ | -8.43 | 1.09 | 8.50 | 100.00 | -1.48 | 2.49 | 2.90 | 11.14 | 0.57 | 2.81 | 2.87 | 5.28 |
| $\partial p_1/\partial w_1$ | 6.66 | 1.55 | 6.84 | 99.32 | 1.28 | 2.43 | 2.74 | 11.06 | -0.05 | 2.51 | 2.51 | 5.46 |
| $\partial p_1/\partial w_2$ | 1.33 | 0.72 | 1.51 | 44.26 | 0.27 | 0.90 | 0.94 | 6.96 | -0.01 | 0.94 | 0.94 | 5.52 |
| $\partial p_2/\partial x$ | 4.96 | 0.98 | 5.06 | 99.88 | 0.91 | 1.87 | 2.08 | 10.90 | -0.33 | 2.10 | 2.12 | 6.42 |
| $\partial p_2/\partial w_1$ | -3.32 | 0.78 | 3.41 | 99.32 | -0.64 | 1.22 | 1.37 | 11.04 | 0.02 | 1.26 | 1.26 | 5.38 |
| $\partial p_2/\partial w_2$ | -2.61 | 1.41 | 2.96 | 44.58 | -0.54 | 1.79 | 1.87 | 7.02 | 0.03 | 1.88 | 1.88 | 5.60 |
| $\partial p_0/\partial x$ | 3.46 | 0.87 | 3.57 | 97.42 | 0.58 | 1.40 | 1.51 | 8.16 | -0.24 | 1.53 | 1.55 | 5.60 |
| $\partial p_0/\partial w_1$ | -3.34 | 0.78 | 3.43 | 99.32 | -0.64 | 1.22 | 1.37 | 10.98 | 0.02 | 1.26 | 1.26 | 5.40 |
| $\partial p_0/\partial w_2$ | 1.28 | 0.69 | 1.45 | 44.68 | 0.27 | 0.89 | 0.93 | 7.02 | -0.02 | 0.94 | 0.94 | 5.62 |
| | | | | | | $\tau = 3/4$ | | | | | | |
| $\partial p_1/\partial x$ | -12.85 | 0.87 | 12.88 | 100.00 | -5.64 | 2.56 | 6.19 | 71.12 | 0.37 | 3.09 | 3.12 | 5.06 |
| $\partial p_1/\partial w_1$ | 10.20 | 1.47 | 10.31 | 100.00 | 4.33 | 2.60 | 5.05 | 54.56 | 0.17 | 2.70 | 2.70 | 5.82 |
| $\partial p_1/\partial w_2$ | 1.90 | 0.69 | 2.02 | 78.98 | 0.88 | 0.90 | 1.26 | 21.40 | 0.04 | 1.00 | 1.01 | 5.90 |
| $\partial p_2/\partial x$ | 7.26 | 0.80 | 7.31 | 100.00 | 3.35 | 1.86 | 3.83 | 56.54 | -0.20 | 2.32 | 2.33 | 6.34 |
| $\partial p_2/\partial w_1$ | -5.08 | 0.75 | 5.13 | 100.00 | -2.17 | 1.30 | 2.53 | 54.46 | -0.08 | 1.35 | 1.35 | 5.76 |
| $\partial p_2/\partial w_2$ | -3.68 | 1.32 | 3.91 | 79.26 | -1.74 | 1.78 | 2.49 | 22.06 | -0.06 | 2.01 | 2.01 | 6.02 |
| $\partial p_0/\partial x$ | 5.59 | 0.74 | 5.63 | 100.00 | 2.29 | 1.43 | 2.70 | 45.98 | -0.17 | 1.71 | 1.72 | 5.50 |
| $\partial p_0/\partial w_1$ | -5.12 | 0.75 | 5.18 | 100.00 | -2.17 | 1.30 | 2.53 | 54.50 | -0.08 | 1.35 | 1.35 | 5.74 |
| $\partial p_0/\partial w_2$ | 1.78 | 0.64 | 1.89 | 79.40 | 0.86 | 0.88 | 1.23 | 22.22 | 0.03 | 1.01 | 1.01 | 5.88 |

This table reports the simulated finite sample bias, standard deviation, RMSE, and size of the MLE and the MERM estimators and the corresponding t-tests for the partial derivatives $\partial p_j(x, w, \theta_0)/\partial x$, $\partial p_j(x, w, \theta_0)/\partial w_1$, $\partial p_j(x, w, \theta_0)/\partial w_2$ for $j \in \{1, 2, 0\}$ evaluated at the population mean. The true values of the marginal effects are $(\partial p_1/\partial x, \partial p_2/\partial x, \partial p_0/\partial x) = (0.222, -0.111, -0.111)$ and zeros for the rest. The results are based on 5,000 replications.

Table 5: Simulation results for the multinomial logit model, Case RC

| | MLE | | | | $K=2$ | | | | $K=4$ | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | bias, $10^{-2}$ | std, $10^{-2}$ | rmse, $10^{-2}$ | size | bias, $10^{-2}$ | std, $10^{-2}$ | rmse, $10^{-2}$ | size | bias, $10^{-2}$ | std, $10^{-2}$ | rmse, $10^{-2}$ | size |
| | | | | | | $\tau = 1/4$ | | | | | | |
| $\partial p_1/\partial x$ | -3.24 | 1.36 | 3.51 | 66.98 | 0.74 | 2.63 | 2.74 | 4.30 | 1.13 | 2.73 | 2.95 | 7.86 |
| $\partial p_1/\partial w_1$ | 2.32 | 1.64 | 2.84 | 30.74 | -0.11 | 2.30 | 2.30 | 4.82 | -0.31 | 2.29 | 2.31 | 6.54 |
| $\partial p_1/\partial w_2$ | 0.48 | 0.75 | 0.90 | 9.40 | -0.04 | 0.87 | 0.87 | 4.82 | -0.08 | 0.87 | 0.87 | 5.36 |
| $\partial p_2/\partial x$ | 1.96 | 1.17 | 2.28 | 39.44 | -0.40 | 1.88 | 1.92 | 4.72 | -0.63 | 1.93 | 2.03 | 6.74 |
| $\partial p_2/\partial w_1$ | -1.16 | 0.82 | 1.42 | 30.66 | 0.06 | 1.15 | 1.15 | 4.84 | 0.15 | 1.15 | 1.16 | 6.48 |
| $\partial p_2/\partial w_2$ | -0.96 | 1.50 | 1.78 | 9.48 | 0.09 | 1.74 | 1.74 | 4.98 | 0.17 | 1.74 | 1.75 | 5.44 |
| $\partial p_0/\partial x$ | 1.28 | 1.01 | 1.63 | 25.28 | -0.34 | 1.43 | 1.47 | 5.08 | -0.50 | 1.48 | 1.56 | 7.36 |
| $\partial p_0/\partial w_1$ | -1.16 | 0.82 | 1.42 | 30.60 | 0.06 | 1.15 | 1.15 | 4.82 | 0.15 | 1.15 | 1.16 | 6.46 |
| $\partial p_0/\partial w_2$ | 0.48 | 0.74 | 0.88 | 9.46 | -0.05 | 0.87 | 0.87 | 4.90 | -0.09 | 0.87 | 0.88 | 5.32 |
| | | | | | | $\tau = 1/2$ | | | | | | |
| $\partial p_1/\partial x$ | -8.97 | 1.09 | 9.04 | 100.00 | -1.69 | 2.60 | 3.10 | 9.84 | 0.97 | 2.89 | 3.05 | 6.04 |
| $\partial p_1/\partial w_1$ | 6.44 | 1.53 | 6.62 | 98.96 | 1.39 | 2.44 | 2.81 | 12.14 | -0.21 | 2.41 | 2.42 | 5.54 |
| $\partial p_1/\partial w_2$ | 1.28 | 0.72 | 1.47 | 42.54 | 0.29 | 0.90 | 0.94 | 7.00 | -0.06 | 0.92 | 0.92 | 5.00 |
| $\partial p_2/\partial x$ | 5.22 | 0.97 | 5.31 | 99.98 | 1.05 | 1.89 | 2.16 | 10.16 | -0.53 | 2.08 | 2.15 | 6.14 |
| $\partial p_2/\partial w_1$ | -3.21 | 0.77 | 3.30 | 98.96 | -0.69 | 1.22 | 1.40 | 12.10 | 0.10 | 1.21 | 1.21 | 5.52 |
| $\partial p_2/\partial w_2$ | -2.52 | 1.41 | 2.88 | 42.82 | -0.56 | 1.79 | 1.87 | 7.20 | 0.13 | 1.84 | 1.84 | 4.98 |
| $\partial p_0/\partial x$ | 3.75 | 0.86 | 3.85 | 98.78 | 0.64 | 1.45 | 1.59 | 8.18 | -0.44 | 1.59 | 1.65 | 6.50 |
| $\partial p_0/\partial w_1$ | -3.23 | 0.78 | 3.32 | 98.96 | -0.70 | 1.22 | 1.41 | 12.06 | 0.10 | 1.21 | 1.21 | 5.48 |
| $\partial p_0/\partial w_2$ | 1.23 | 0.69 | 1.41 | 42.90 | 0.28 | 0.89 | 0.93 | 7.20 | -0.07 | 0.92 | 0.92 | 4.94 |
| | | | | | | $\tau = 3/4$ | | | | | | |
| $\partial p_1/\partial x$ | -13.35 | 0.86 | 13.38 | 100.00 | -6.83 | 2.64 | 7.32 | 80.32 | 0.71 | 3.22 | 3.29 | 4.74 |
| $\partial p_1/\partial w_1$ | 9.69 | 1.45 | 9.80 | 100.00 | 4.95 | 2.65 | 5.61 | 65.52 | 0.01 | 2.62 | 2.62 | 5.34 |
| $\partial p_1/\partial w_2$ | 1.81 | 0.69 | 1.94 | 75.30 | 1.01 | 0.89 | 1.35 | 26.08 | -0.01 | 0.98 | 0.98 | 5.24 |
| $\partial p_2/\partial x$ | 7.48 | 0.79 | 7.52 | 100.00 | 4.06 | 1.83 | 4.45 | 68.82 | -0.37 | 2.32 | 2.35 | 5.74 |
| $\partial p_2/\partial w_1$ | -4.83 | 0.73 | 4.88 | 100.00 | -2.47 | 1.32 | 2.81 | 65.46 | -0.01 | 1.31 | 1.31 | 5.28 |
| $\partial p_2/\partial w_2$ | -3.51 | 1.33 | 3.76 | 75.60 | -1.99 | 1.76 | 2.66 | 26.38 | 0.03 | 1.97 | 1.97 | 5.32 |
| $\partial p_0/\partial x$ | 5.87 | 0.73 | 5.92 | 100.00 | 2.77 | 1.47 | 3.14 | 56.28 | -0.34 | 1.77 | 1.80 | 5.82 |
| $\partial p_0/\partial w_1$ | -4.87 | 0.75 | 4.93 | 100.00 | -2.48 | 1.33 | 2.81 | 65.40 | -0.01 | 1.31 | 1.31 | 5.32 |
| $\partial p_0/\partial w_2$ | 1.70 | 0.64 | 1.82 | 75.76 | 0.98 | 0.87 | 1.31 | 26.50 | -0.02 | 0.99 | 0.99 | 5.30 |

This table reports the simulated finite sample bias, standard deviation, RMSE, and size of the MLE and the MERM estimators and the corresponding t-tests for the partial derivatives $\partial p_j(x,w,\theta_0)/\partial x$, $\partial p_j(x,w,\theta_0)/\partial w_1$, $\partial p_j(x,w,\theta_0)/\partial w_2$ for $j \in \{1,2,0\}$ evaluated at the population mean. The true values of the marginal effects are $(\partial p_1/\partial x, \partial p_2/\partial x, \partial p_0/\partial x) = (0.222, -0.111, -0.111)$ and zeros for the rest. The results are based on 5,000 replications.

## 3.3 Empirical Illustration: Choice of Transportation Mode

In this section, we illustrate the finite sample properties of the MERM estimator in the context of a classical multinomial choice application: choice of transportation mode (e.g., McFadden, 1974).

To calibrate the numerical experiment, we use the ModeCanada dataset, a survey of business travelers for the Montreal-Toronto corridor. We focus on the subset of travelers choosing between train, air, and car ($n = 2769$), and estimate the following specification of the conditional logit model with traveler $i$'s utilities given in the table below.

| Mode | Utility |
|---|---|
| Air | $U_{i1} = \theta_{01}\, Income_i^* + \theta_{02}\, Urban_i + \theta_{03} + \theta_{07}\, Price_{i1} + \theta_{08}\, InTime_{i1} + \epsilon_{i1}$ |
| Car | $U_{i2} = \theta_{04}\, Income_i^* + \theta_{05}\, Urban_i + \theta_{06} + \theta_{07}\, Price_{i2} + \theta_{08}\, InTime_{i2} + \epsilon_{i2}$ |
| Train | $U_{i0} = \theta_{07}\, Price_{i0} + \theta_{08}\, InTime_{i0} + \epsilon_{i0}$ |

To generate the simulated samples, we randomly draw covariates from their joint empirical distribution. To generate the simulated outcomes, we draw $\epsilon_{ij}$ from the standard type-I extreme value distribution. The true value of $\theta_0$ is set to be the MLE estimate based on the original dataset. More details about this numerical experiment are given in Appendix F.

To evaluate the performance of the MERM estimator in these settings, we generate mismeasured $Income_i = Income_i^* + \varepsilon_i$. We focus on the individual income because it is often mismeasured.We report the results for $\tau = \sigma_\varepsilon / \sigma_{Income^*} \in \{1/4, 1/2, 3/4\}$.

Table 6 reports the simulation results for the (naive) MLE estimator and for the MERM estimators with $K = 2$ and $K = 4$. We focus on estimation of and inference on the income elasticities (evaluated at the population mean of the covariates). The MLE estimator is considerably biased for $\tau \in \{1/2, 3/4\}$, which results in substantial size distortions of the MLE based t-tests. The MERM estimator with $K = 4$ effectively eliminates the EIV bias and the corresponding t-tests provide accurate size control in all of the considered designs. The estimator with $K = 2$ is more precise, while successfully removing the EIV bias for $\tau \leq 1/2$.

Overall, the MERM estimators perform well in the considered empirical context, providing a basis for estimation and inference even for quite large values of $\tau$.

Table 6: Simulation results for the empirically calibrated conditional logit model

| | MLE | | | | $K=2$ | | | | $K=4$ | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | bias | std | rmse | size | bias | std | rmse | size | bias | std | rmse | size |
| $\tau = 1/4$ | | | | | | | | | | | | |
| $\partial \ln p_1/\partial \ln I$ | -0.07 | 0.12 | 0.14 | 9.00 | 0.01 | 0.14 | 0.14 | 5.68 | 0.02 | 0.19 | 0.19 | 7.02 |
| $\partial \ln p_2/\partial \ln I$ | 0.03 | 0.07 | 0.08 | 5.84 | -0.00 | 0.08 | 0.08 | 5.68 | -0.01 | 0.10 | 0.10 | 6.40 |
| $\partial \ln p_0/\partial \ln I$ | 0.05 | 0.13 | 0.13 | 6.10 | 0.00 | 0.14 | 0.14 | 5.42 | -0.00 | 0.17 | 0.17 | 7.66 |
| $\tau = 1/2$ | | | | | | | | | | | | |
| $\partial \ln p_1/\partial \ln I$ | -0.24 | 0.11 | 0.27 | 61.84 | -0.05 | 0.14 | 0.15 | 6.96 | 0.02 | 0.21 | 0.21 | 6.16 |
| $\partial \ln p_2/\partial \ln I$ | 0.09 | 0.07 | 0.11 | 24.76 | 0.02 | 0.09 | 0.09 | 5.96 | -0.01 | 0.10 | 0.10 | 6.16 |
| $\partial \ln p_0/\partial \ln I$ | 0.16 | 0.12 | 0.20 | 25.36 | 0.04 | 0.15 | 0.15 | 6.06 | -0.00 | 0.18 | 0.18 | 6.86 |
| $\tau = 3/4$ | | | | | | | | | | | | |
| $\partial \ln p_1/\partial \ln I$ | -0.43 | 0.09 | 0.44 | 99.50 | -0.19 | 0.14 | 0.24 | 27.46 | 0.02 | 0.22 | 0.22 | 5.84 |
| $\partial \ln p_2/\partial \ln I$ | 0.16 | 0.06 | 0.17 | 71.88 | 0.07 | 0.08 | 0.11 | 13.78 | -0.01 | 0.11 | 0.11 | 6.32 |
| $\partial \ln p_0/\partial \ln I$ | 0.29 | 0.11 | 0.31 | 73.20 | 0.12 | 0.15 | 0.19 | 13.40 | 0.00 | 0.19 | 0.19 | 6.32 |

This table reports the simulated finite sample bias, standard deviation, RMSE, and size of the MLE and the MERM estimators and the corresponding t-tests for the income elasticities $\partial \ln p_j(I, w, \theta_0)/\partial \ln I$, $j \in \{1, 2, 0\}$, evaluated at the population mean. The true values of the income elasticities are $(\partial \ln p_1/\partial \ln I, \partial \ln p_2/\partial \ln I, \partial \ln p_0/\partial \ln I) = (1.11, -0.39, -0.82)$. The results are based on 5,000 replications.

# 4 Identification of the Nonlinear Regression Model

It is important to understand under what conditions the parameters of interest are identified in our settings. This section considers identification and estimation of the nonlinear regression function $\rho(x) \equiv E[Y_i|X_i^* = x]$. To address the problem of EIV, the researcher has an instrument $Z_i$, which can be discrete or continuous. In this section, we do not impose any functional form assumptions on the true regression function $\rho(\cdot)$, i.e., the analysis is nonparametric.

Consider the model

$$Y_i = \rho(X_i^*) + U_i, \qquad X_i = X_i^* + \varepsilon_i. \tag{18}$$

The joint distribution of observables $(Y_i, X_i, Z_i)$ satisfies the following assumptions.

**Assumption 4.1.** $E[Y_i|X_i^*, Z_i] = E[Y_i|X_i^*]$.

**Assumption 4.2.** *Functions $\rho(\cdot)$ and $f_{X^*|Z}(\cdot|z)$ have at least $p \geq 3$ bounded continuous derivatives, and $E\left[|\varepsilon_i|^4\right] \leq C\sigma^4$ for some constant $C$.*

Assumption 4.1 is a standard exclusion restriction on the instrument $Z_i$. Assump-

tion 4.2 collects some weak regularity conditions. We will study the properties of this model using the approximation $\tau = o(1)$. The identification analysis views the unknown function $\rho$ and the joint distribution of $(Y_i, X_i^*, Z_i)$ as fixed. Thus, the joint distribution of the observables $(Y_i, X_i, Z_i)$ is implicitly indexed by $\tau$, but varies with $\tau$ only due to the changes in the distribution of $\varepsilon_i$.

The naive population regression $q(x) \equiv E[Y_i|X_i = x]$ suffers from the EIV bias

$$E[Y_i|X_i = x] = \rho(x) + O(\tau^2). \tag{19}$$

Similar to the semiparametric case of Section 2, identifying $\sigma^2$ or $\tau^2$ allows dealing with this bias.[9]

Let us define

$$q(x,z) \equiv E[Y_i|X_i = x, Z_i = z], \quad s_{X|Z}(x|z) \equiv \frac{f'_{X|Z}(x|z)}{f_{X|Z}(x|z)}, \quad s_{X^*|Z}(x|z) \equiv \frac{f'_{X^*|Z}(x|z)}{f_{X^*|Z}(x|z)}. \tag{20}$$

Let $\mathcal{S}_{X^*}(z) \equiv \{x : f_{X^*|Z}(x|z) > 0\}$, so that $s_{X^*|Z}(x|z)$ is well-defined for all $x \in \mathcal{S}_{X^*}(z)$.

The following functions can be identified directly from the joint distribution of the observed $(Y_i, X_i, Z_i)$:

$$\widetilde{\sigma}^2(x) \equiv \frac{q(x,z_1) - q(x,z_2)}{q'(x)\left[s_{X|Z}(x|z_1) - s_{X|Z}(x|z_2)\right]}, \tag{21}$$

$$\widetilde{\rho}(x,z,v) \equiv q(x,z) - v\left[q'(x)\, s_{X|Z}(x|z) + \tfrac{1}{2}q''(x)\right]. \tag{22}$$

**Theorem 3.** *Suppose Assumptions CME, 4.1, 4.2 hold and either (i) $p = 3$ or (ii) $E[\varepsilon_i^3] = 0$ and $p = 4$. Suppose also that $\tau = o(1)$, and there exist $z_1$, $z_2$, and a point $x \in S_{X^*}(z_1) \cap S_{X^*}(z_2)$, such that $\rho'(x)\left[s_{X^*|Z}(x|z_1) - s_{X^*|Z}(x|z_2)\right] \neq 0$. Then*

$$\widetilde{\sigma}^2(x) \equiv \sigma^2 + O(\tau^p). \tag{23}$$

*Moreover, for any $\widetilde{\sigma}^2 = \sigma^2 + O(\tau^p)$ (including $\widetilde{\sigma}^2 = \widetilde{\sigma}^2(x)$), any $z$, and any $x \in S_{X^*}(z)$,*

$$\widetilde{\rho}\left(x, z, \widetilde{\sigma}^2\right) \equiv \rho(x) + O(\tau^p) \quad \text{for all } x \in \mathcal{S}_{X^*}(z). \tag{24}$$

Equation (23) shows that we can identify $\sigma^2$ up to an error of smaller order $O(\sigma^p) = O(\tau^p)$. As a consequence, $\rho(\cdot)$ is identified up to the same order $O(\tau^p)$. Below we will

[9]Note that $\sigma^2$ can be approximated by 0. However, to be meaningful, identification of $\sigma^2$ needs to characterize it up to an error of an order smaller than $\sigma^2$ itself.

consider the implications of this theorem for estimation.

Theorem 3 requires the rank condition $\rho'(x)\left[s_{X^*|Z}(x|z_1) - s_{X^*|Z}(x|z_2)\right] \neq 0$ to hold for some $x$. The key here is the instrument relevance condition that requires $s_{X^*|Z}(x|z)$ to vary with $z$. The proof of the theorem shows that

$$q(x,z) = \rho(x) + \sigma^2\rho'(x)\, s_{X^*|Z}(x|z) + \tfrac{1}{2}\sigma^2\rho''(x) + O(\tau^p). \tag{25}$$

This equation is used to identify $\sigma^2$ by varying $z$, since only the second term on the right-hand side depends on $z$. Note that for $q(x,z)$ to vary with $z$ we need the additional rank condition $\rho'(x) \neq 0$. Requiring that there exists a point $x$ with $\rho'(x) \neq 0$ is a weak condition, since $\rho'(x) = 0$ for all $x$ means that $\rho(x)$ is constant, so EIV do not bias the regression estimator and $q(x) = \rho(x)$.[10,11]

The rank condition $\rho'(x)\left[s_{X^*|Z}(x|z_1) - s_{X^*|Z}(x|z_2)\right] \neq 0$ in Theorem 3 can be replaced with the condition that $q'(x)\left[s_{X|Z}(x|z_1) - s_{X|Z}(x|z_2)\right]$ is bounded away from zero, which is stated in terms of observables. Likewise, set $\mathcal{S}_{X^*}(z)$ above can be replaced with $\mathcal{S}_X(z) \equiv \left\{x : f_{X^*|Z}(x|z) > C_f\right\}$ for any positive constant $C_f$.

The relevance condition imposed on the instrumental variable is weak. The equality $s_{X^*|Z}(x|z_1) = s_{X^*|Z}(x|z_2)$ can only hold for all $x$, $z_1$, and $z_2$ when $X^*$ and $Z$ are independent. Moreover, consider the following example:

**Example 4.1.** *Suppose $X_i^*$ follows a Gaussian random coefficient model:*

$$X_i^* = \Pi_{1i} + \Pi_{2i}Z_i, \qquad \textit{where} \quad \begin{pmatrix} \Pi_{1i} \\ \Pi_{2i} \end{pmatrix} \sim N\left(\begin{pmatrix} \pi_1 \\ \pi_2 \end{pmatrix}, \begin{pmatrix} \upsilon_{\pi 1}^2 & 0 \\ 0 & \upsilon_{\pi 2}^2 \end{pmatrix}\right),$$

*and $Z_i \in \{0,1\}$ is a binary instrument. In many applications the instruments are likely to have heterogeneous effects on the covariate, e.g., Angrist, Graddy, and Imbens (2000) and Heckman and Vytlacil (1998), which corresponds to $\upsilon_{\pi 2}^2 > 0$. Then, $X_i^*|Z_i = z \sim N(\mu(z), \upsilon^2(z))$ with $\mu(z) = \pi_1 + \pi_2 z$ and $\upsilon^2(z) = \upsilon_{\pi 1}^2 + \upsilon_{\pi 2}^2 z$. Thus, the instrument is relevant unless $\pi_2 = \upsilon_{\pi 2} = 0$. Notably, the instrument is relevant even if $\pi_2 = 0$ (so $\operatorname{corr}(X_i^*, Z_i) = 0$), as long as $Z_i$ has a heterogeneous effect on $X_i^*$.*

[10] See Evdokimov and Zeleneev (2018) for more details on the role of $\rho'(x)$ in the literature on measurement errors.

[11] Equation (25) extends the calculation of the conditional expectation in equation (4.6) in Chesher (1991) by introducing instrumental variables and obtaining more precise bounds on the approximation error.

**Remark 11.** *Note that $\sigma^2$ is overidentified, which can be used to test the model assumptions.*

To clarify the implications of Theorem 3 it is useful to consider nonparametric estimation of the regression function $\rho(x)$. The approach of Theorem 3 is constructive, and using equations (20)-(22) we can construct a nonparametric measurement error robust analog estimator $\widehat{\rho}^{\text{MER}}(x)$. We compare such $\widehat{\rho}^{\text{MER}}(x)$ with the naive nonparametric regression estimator $\widehat{\rho}^{\text{Naive}}(x)$ of $E[Y|X_i = x]$, which ignores the presence of EIV in the data. Both estimators can be implemented using standard nonparametric estimation methods (e.g., kernel or series estimators). In the following discussion assume that the tuning parameters are chosen optimally for each of the estimators. For brevity we focus only on the case (ii) in Theorem 3.

**Proposition 4.** *Suppose the conditions of Theorem 3 hold, $E[\varepsilon_i^3] = 0$, functions $\rho(\cdot)$ and $f_{X^*|Z}(\cdot|z)$ have $m \geq 4$ continuous derivatives, and $Z_i$ is discrete. Suppose $\tau_n = O\left(n^{-\frac{1}{4}\frac{m}{2m+1}}\right)$, then*

$$\begin{aligned}\widehat{\rho}^{MER}(x) - \rho(x) &= O_p\left(n^{-\frac{m}{2m+1}}\right), \\ \widehat{\rho}^{Naive}(x) - \rho(x) &= O_p\left(n^{-\frac{1}{2}\frac{m}{2m+1}}\right).\end{aligned}$$

The proposition provides a nonparametric analog of the semiparametric results in Section 2. In particular, $\widehat{\rho}^{\text{MER}}$ generally has a faster rate of convergence than $\widehat{\rho}^{\text{Naive}}$. For example, if $m = 4$ the rates of convergence of $\widehat{\rho}^{\text{MER}}(x)$ and $\widehat{\rho}^{\text{Naive}}(x)$ are $O_p\left(n^{-4/9}\right)$ and $O_p\left(n^{-2/9}\right)$, respectively. Note that the rate of convergence of $\widehat{\rho}^{\text{MER}}(x)$ in Proposition 4 is optimal and cannot be improved: even if had data on $(Y_i, X_i^*)$ without EIV, under the smoothness assumptions of the proposition, function $\rho(x)$ cannot be estimated nonparametrically at a rate faster than $O_p\left(n^{-\frac{m}{2m+1}}\right)$, see Stone (1980). Note also that for models with large $m$, the rates of convergence in the Proposition approach those in Remark 1.

# 5 Extensions

## 5.1 Multiple Mismeasured Variables

It is easy to use the MERM framework to deal with multiple mismeasured variables. This is useful in many applications, including not only settings with multiple mismeasured covariates, but also settings with serially correlated measurement errors, settings

where repeated measurements are available, and panel data models. Using the MERM approach is particularly advantageous in such applications, since it avoids nonparametric estimation of multivariate unobserved distributions.

Suppose $X_i^*$, $\varepsilon_i$, and $X_i$ are $d\times 1$ vectors. Let $\tau_n \equiv \max_{j\leq d} \sigma_{\varepsilon_j}/\sigma_{X_j^*}$, where $\sigma_{\varepsilon_j}$ and $\sigma_{X_j^*}$ denote the standard deviations of the $j$-th components of $\varepsilon_i$ and $X_i^*$, so $\mathbb{E}\left[|\varepsilon_{ij}|^k\right] = O(\tau_n^k)$ for $k \in \{1,\ldots,K\}$.

For a $d \times 1$ vector of non-negative integers $\kappa = (\kappa_1,\ldots,\kappa_d) \in \mathbb{Z}_+^d$, let

$$\partial_\kappa \equiv \frac{\partial^{|\kappa|}}{\partial {x_1}^{\kappa_1} \ldots \partial {x_d}^{\kappa_d}}, \quad \text{where } |\kappa| \equiv \sum_{j=1}^d \kappa_j.$$

Also, for a positive integer $k$, let $\mathcal{K}_k = \{\kappa \in \mathbb{Z}_+^d : |\kappa| = k\}$. Then, we consider the following corrected moment function

$$\psi(x,s,\theta,\gamma) = g(x,s,\theta) - \sum_{k=2}^K \sum_{\kappa\in\mathcal{K}_k} \gamma_\kappa \partial_\kappa g(x,s,\theta),$$

where, with some abuse of notation, $\gamma$ is a collection of all $\gamma_\kappa$ with $\kappa \in \mathcal{K}_k$ and $k \in \{2,\ldots,K\}$.

Under mild smoothness conditions

$$\mathbb{E}\left[\psi(X_i,S_i,\theta_0,\gamma_0)\right] = \mathbb{E}\left[g(X_i^*,S_i,\theta_0)\right] + O(\tau_n^{K+1}) = o(n^{-1/2}),$$

where the second equality holds provided that $O(\tau_n^{K+1}) = o(n^{-1/2})$. Similarly to the scalar case, components of $\gamma_0$ are determined by the moments of $\varepsilon_i$. Specifically, let $\mu_\kappa \equiv \mathbb{E}\left[\varepsilon_{i1}^{\kappa_1} \ldots \varepsilon_{id}^{\kappa_d}\right]$, then

$$\gamma_{0\kappa} = \frac{\mu_\kappa}{\kappa!}, \quad \text{for } \kappa \in \{\mathcal{K}_2, \mathcal{K}_3\}, \tag{26}$$

where $\kappa! \equiv \kappa_1! \ldots \kappa_d!$. For $|\kappa| \geqslant 4$, the coefficients can be computed by the following formulas. For example, for $\kappa \in \mathcal{K}_4$, let $\mathcal{K}_{2,\kappa} = \{\tilde\kappa \in \mathcal{K}_2 : \kappa - \tilde\kappa \in \mathcal{K}_2\}$. Then,

$$\gamma_{0\kappa} = \frac{\mu_\kappa}{\kappa!} - \sum_{\tilde\kappa\in\mathcal{K}_{2,\kappa}} \frac{\mu_{\kappa-\tilde\kappa}}{(\kappa-\tilde\kappa)!}\gamma_{0\tilde\kappa}, \quad \text{for } \kappa \in \mathcal{K}_4.$$

More generally, for $\kappa \in \mathcal{K}_k$ with $k \geqslant 4$, let $\mathcal{K}_{\ell,\kappa} = \{\tilde\kappa \in \mathcal{K}_\ell, \kappa - \tilde\kappa \in \mathcal{K}_{|\kappa|-\ell}\}$ for $\ell \leqslant |\kappa| - 2$.

Then,

$$\gamma_{0\kappa} = \frac{\mu_\kappa}{\kappa!} - \sum_{\ell=2}^{k-2} \sum_{\tilde{\kappa} \in \mathcal{K}_{\ell,\kappa}} \frac{\mu_{\kappa-\tilde{\kappa}}}{(\kappa-\tilde{\kappa})!} \gamma_{0\tilde{\kappa}}.$$

**Example** (Bivariate $X$, $K = 4$)**.**
*Suppose $X$ is bivariate (i.e., $d = 2$) and $K = 4$. For $\kappa \in \mathcal{K}_2 = \{(2,0),(1,1),(0,2)\}$ and $\kappa \in \mathcal{K}_3 = \{(3,0),(2,1),(1,2),(0,3)\}$, $\gamma_{0\kappa}$ is given by (26). For $\kappa \in \mathcal{K}_4$, $\gamma_{0\kappa}$ is given by*

| $\kappa$ | $\gamma_{0\kappa}$ |
|---|---|
| *(4,0)* | $\frac{\mathbb{E}[\varepsilon_{i1}^4] - 6\mathbb{E}[\varepsilon_{i1}^2]^2}{24}$ |
| *(3,1)* | $\frac{\mathbb{E}[\varepsilon_{i1}^3 \varepsilon_{i2}] - 6\mathbb{E}[\varepsilon_{i1}^2]\mathbb{E}[\varepsilon_{i1}\varepsilon_{i2}]}{6}$ |
| *(2,2)* | $\frac{\mathbb{E}[\varepsilon_{i1}^2 \varepsilon_{i2}^2] - 2\mathbb{E}[\varepsilon_{i1}^2]\mathbb{E}[\varepsilon_{i2}^2] - 4\mathbb{E}[\varepsilon_{i1}\varepsilon_{i2}]^2}{4}$ |
| *(1,3)* | $\frac{\mathbb{E}[\varepsilon_{i1}\varepsilon_{i2}^3] - 6\mathbb{E}[\varepsilon_{i2}^2]\mathbb{E}[\varepsilon_{i1}\varepsilon_{i2}]}{6}$ |
| *(0,4)* | $\frac{\mathbb{E}[\varepsilon_{i2}^4] - 6\mathbb{E}[\varepsilon_{i2}^2]^2}{24}$ |

*If in addition measurement errors $\varepsilon_{i1}$ and $\varepsilon_{i2}$ are independent, $\gamma_{0\kappa} = 0$ for $\kappa \in \{(1,1),(2,1),(1,2),(3,1),(1,3)\}$. In this case, the total number of the nuisance parameters to be estimated is 6.*

## 5.2 Non-Classical Measurement Errors

In some applications, Assumption CME of classical measurement errors may be restrictive. For example, the variance of the measurement error $\mathbb{E}[\varepsilon_i^2 | X_i^*, S_i]$ (and other moments) could depend on $X_i^*$ and $S_i$, i.e., the measurement error can be non-classical. In this section, we demonstrate how MERM framework can be used to address this issue.

Suppose that $\mathbb{E}\left[|\varepsilon_i|^k | X_i^*, S_i\right] = O(\tau_n^k)$. Conditional on $X_i^*$ and $S_i$, the moment function can be expanded as

$$\mathbb{E}[g(X_i, S_i, \theta) | X_i^*, S_i] = g(X_i^*, S_i, \theta) + \sum_{k=2}^{K} \frac{\mathbb{E}\left[\varepsilon_i^k | X_i^*, S_i\right]}{k!} g_x^{(k)}(X_i^*, S_i, \theta) + O(\tau_n^{K+1}).$$

Suppose $\mathbb{E}\left[\varepsilon_i^k | X_i^*, S_i\right] = v_k(X_i^*, S_i, \omega_0)$, $k \in \{2, \ldots, K\}$ where $\omega_0 \in \mathbb{R}^{\dim(\omega)}$ are unknown

parameters, i.e.,[12]

$$\mathbb{E}[g(X_i, S_i, \theta)|X_i^*, S_i] = g(X_i^*, S_i, \theta) + \sum_{k=2}^{K} \frac{v_k(X_i^*, S_i, \omega_0)}{k!} g_x^{(k)}(X_i^*, S_i, \theta) + O(\tau_n^{K+1}),$$

which implies

$$\mathbb{E}\left[g(X_i^*, S_i, \theta)\right] = \mathbb{E}\left[g(X_i, S_i, \theta) - \sum_{k=2}^{K} \frac{v_k(X_i^*, S_i, \omega_0)}{k!} g_x^{(k)}(X_i^*, S_i, \theta)\right] + O(\tau_n^{K+1}). \quad (27)$$

Equation (27) is an analog of equation (10), which motivates correcting the moment conditions by estimating and subtracting the terms $\frac{v_k(X_i^*, S_i, \omega_0)}{k!} g_x^{(k)}(X_i^*, S_i, \theta)$. The corrected moment function takes the form

$$\psi(X_i, S_i, \theta, \omega) = g(X_i, S_i, \theta) - \sum_{k=2}^{K} f_k(X_i, S_i, \theta, \omega), \quad (28)$$

where

$$\begin{aligned}
f_2(x, s, \theta, \omega) =& \frac{v_2(x, s, \omega)}{2} g_x^{(2)}(x, s, \theta), \\
f_3(x, s, \theta, \omega) =& \frac{v_3(x, s, \omega)}{6} g_x^{(3)}(x, s, \theta), \\
f_4(x, s, \theta, \omega) =& \frac{v_4(x, s, \omega) - 6v_2(x, s, \omega)^2}{24} g_x^{(4)}(x, s, \theta) \\
& - \frac{v_2(x, s, \omega)}{4} \left(2v_{2,x}^{(1)}(x, s, \omega) g_x^{(3)}(x, s, \theta) + v_{2,x}^{(2)}(x, s, \omega) g_x^{(2)}(x, s, \theta)\right).
\end{aligned}$$

Expressions for $f_k(x, s, \theta, \omega)$ for $k > 4$ are provided below.

Notice that for $K \leqslant 3$ the corrected moment functions are similar to the classical EIV case except for $v_2(x, s, \omega)$ and $v_3(x, s, \omega)$ taking the places of $\mathbb{E}\left[\varepsilon_i^2\right]$ and $\mathbb{E}\left[\varepsilon_i^3\right]$ respectively. For $K \geq 4$, we need to bias correct the bias correction terms such as $f_2(x, s, \theta, \omega)$ accounting for $v_2(X_i, S_i, \theta, \omega)$ being evaluated at $X_i$ instead of $X_i^*$.

Under the regularity conditions

$$\mathbb{E}\left[\psi(X_i, S_i, \theta_0, \omega_0)\right] = \mathbb{E}\left[g(X_i^*, S_i, \theta_0)\right] + O(\tau_n^{K+1}) = o(n^{-1/2}),$$

provided that $O(\tau_n^{K+1}) = o(n^{-1/2})$.

[12]The expansion makes the typical assumption $\mathbb{E}\left[\varepsilon_i|X_i^*, S_i\right] = 0$ that $\varepsilon_i$ is exogenous. In principle, it is straightforward to introduce $v_1(X_i^*, S_i, \omega_0)$.

Thus, the model can be estimated using GMM estimator with the corrected moment function (28), where parameters $\theta_0$ and $\omega_0$ are estimated together, i.e., the estimator in equation (12) with $\beta \equiv (\theta', \omega')'$.

Consider the following example.

**Example 5.1** (Exponential Specification, $K = 4$)**.**
*Suppose $\varepsilon_i = \exp(\omega_{01} X_i^*)\zeta_i$, where $\zeta_i$ is mean zero and independent of $X_i^*$ and $S_i$. In this case,*

$$\begin{aligned}
\mathbb{E}\left[\varepsilon_i^2 | X_i^*, S\right] &= \mathbb{E}\left[\zeta_i^2\right] \exp(2\omega_{01} X_i^*), \\
\mathbb{E}\left[\varepsilon_i^3 | X_i^*, S\right] &= \mathbb{E}\left[\zeta_i^3\right] \exp(3\omega_{01} X_i^*), \\
\mathbb{E}\left[\varepsilon_i^4 | X_i^*, S\right] &= \mathbb{E}\left[\zeta_i^4\right] \exp(4\omega_{01} X_i^*).
\end{aligned}$$

*Hence, we can take*

$$\begin{aligned}
v_2(x, s, \omega) &= \omega_2 \exp(2\omega_1 x) \\
v_3(x, s, \omega) &= \omega_3 \exp(3\omega_1 x) \\
v_4(x, s, \omega) &= \omega_4 \exp(4\omega_1 x),
\end{aligned}$$

*and $\omega_0 = (\omega_{01}, \mathbb{E}\left[\zeta_i^2\right], \mathbb{E}\left[\zeta_i^3\right], \mathbb{E}\left[\zeta_i^4\right])'$. Functions $f_k$ in equation (28) are*

$$\begin{aligned}
f_2(x, s, \theta, \omega) &= \frac{\omega_2 \exp(2\omega_1 x)}{2} g_x^{(2)}(x, s, \theta), \\
f_3(x, s, \theta, \omega) &= \frac{\omega_3 \exp(3\omega_1 x)}{6} g_x^{(3)}(x, s, \theta), \\
f_4(x, s, \theta, \omega) &= \exp(4\omega_1 x) \left( \frac{\omega_4 - 6\omega_2^2}{24} g_x^{(4)}(x, s, \theta) - \omega_1 \omega_2^2 g_x^{(3)}(x, s, \theta) - \omega_1^2 \omega_2^2 g_x^{(2)}(x, s, \theta) \right).
\end{aligned}$$

**General formula for $f_k$ in equation (28).** For $k \geqslant 4$, $f_k(x, s, \theta, \omega)$ takes the form

$$f_k(x, s, \theta, \omega) = \frac{v_k(x, s, \omega)}{k!} g_x^{(k)}(x, s, \theta) - \sum_{\ell=2}^{k-2} \frac{v_{k-\ell}(x, s, \omega)}{(k-\ell)!} \frac{\partial^{k-\ell}}{\partial x^{k-\ell}} f_\ell(x, s, \theta, \omega).$$

For $k = 4$, this coincides with the formula provided earlier. Also, if the conditional moments of $\varepsilon_i$ do not depend on $X_i^*$, the corrected moment function becomes similar to the classical EIV case except for $v_k(s, \omega)$ taking the place of $\mathbb{E}\left[\varepsilon_i^k\right]$.

Finally, we illustrate the finite sample properties of the MERM estimator with non-

classical measurement error. Specifically, we consider the following logit model

$$Y_i = \mathbb{1}\{\theta_{01} X_i^* + \theta_{02} W_i + \theta_{03} - \eta_i \geqslant 0\},$$
$$X_i^* = Z_i + V_i, \quad X_i = X_i^* + \varepsilon_i, \quad \varepsilon_i = \exp(\omega_{01} X_i^*)\zeta_i, \quad W_i = \rho X_i^*/\sigma_{X^*} + \sqrt{1-\rho^2}\nu_i,$$

where $\eta_i \sim \mathit{Logistic}$ and $(Z_i, V_i, \zeta_i, \nu_i)' \sim N\left((0,0,0,0)', \mathrm{Diag}(\sigma_Z^2, \sigma_V^2, \sigma_\zeta^2, \sigma_\nu^2)\right)$ are independent from each other.

We fix $(\theta_{01}, \theta_{02}, \theta_{03}, \omega_{01}, \rho, \sigma_Z^2, \sigma_V^2, \sigma_\nu^2) = (1, 0, 2, 0.3, 0.7, 1, 1, 1)$ and $n = 2000$. By adjusting $\sigma_\zeta^2$ accordingly, we consider $\tau = \sigma_\varepsilon/\sigma_{X^*} \in \{1/4, 1/2, 3/4\}$, where, as before, $\sigma_\varepsilon$ denotes the (unconditional) standard deviation of $\varepsilon_i$.

We report results for the MERM estimator based on the corrected moment function (28) with $K = 2$ and $K = 4$. As in Section 3.1, the original moment function is

$$g(x, w, y, z, \theta) = (y - \Lambda(\theta_1 x + \theta_2 w + \theta_3))\, h(x, z, w),$$

where we use $h(x, z, w) = (1, x, z, x^2, z^2, x^3, z^3, w)'$ for $K = 2$ and $h(x, z, w) = (1, x, z, x^2, xz, z^2, x^3, x^2 z, xz^2, z^3, w)'$ for $K = 4$. The corrected moment function is as in Example 5.1. For simplicity we set $\omega_3 = 0$ (using $\mathbb{E}\left[\zeta_i^3\right] = 0$).

Table 7 reports the simulation results. Both of the correction schemes effectively remove the EIV bias for $\tau \in \{1/4, 1/2\}$. However, employing the higher order correction scheme with $K = 4$ is needed to achieve accurate size control for larger values of $\tau$ ($\tau = 3/4$).

Table 7: Simulation results for the logit model with non-classical measurement error

| | MLE | | | | $K=2$ | | | | $K=4$ | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | bias | std | rmse | size | bias | std | rmse | size | bias | std | rmse | size |
| | | | | | | $\tau = 1/4$ | | | | | | |
| $\theta_1$ | -0.062 | 0.073 | 0.096 | 14.20 | 0.041 | 0.113 | 0.120 | 3.50 | 0.065 | 0.120 | 0.137 | 5.82 |
| $\theta_2$ | 0.063 | 0.090 | 0.110 | 10.86 | -0.014 | 0.119 | 0.120 | 2.70 | -0.027 | 0.125 | 0.128 | 3.96 |
| $\partial_x$ | -0.006 | 0.007 | 0.010 | 16.10 | 0.001 | 0.010 | 0.010 | 2.92 | 0.002 | 0.010 | 0.011 | 4.08 |
| $\partial_w$ | 0.007 | 0.009 | 0.012 | 10.74 | -0.001 | 0.012 | 0.012 | 3.06 | -0.003 | 0.012 | 0.013 | 4.36 |
| | | | | | | $\tau = 1/2$ | | | | | | |
| $\theta_1$ | -0.224 | 0.068 | 0.234 | 89.84 | 0.019 | 0.134 | 0.136 | 3.50 | 0.043 | 0.141 | 0.147 | 5.78 |
| $\theta_2$ | 0.219 | 0.083 | 0.235 | 75.32 | 0.019 | 0.138 | 0.139 | 3.82 | -0.003 | 0.141 | 0.141 | 4.10 |
| $\partial_x$ | -0.022 | 0.006 | 0.023 | 91.66 | -0.003 | 0.012 | 0.012 | 4.14 | -0.001 | 0.012 | 0.012 | 4.52 |
| $\partial_w$ | 0.024 | 0.009 | 0.025 | 74.86 | 0.002 | 0.014 | 0.014 | 3.82 | -0.000 | 0.014 | 0.014 | 4.52 |
| | | | | | | $\tau = 3/4$ | | | | | | |
| $\theta_1$ | -0.419 | 0.058 | 0.423 | 100.00 | -0.116 | 0.129 | 0.173 | 12.18 | 0.031 | 0.154 | 0.157 | 5.96 |
| $\theta_2$ | 0.396 | 0.076 | 0.403 | 99.98 | 0.172 | 0.131 | 0.217 | 31.82 | 0.033 | 0.142 | 0.146 | 5.86 |
| $\partial_x$ | -0.040 | 0.006 | 0.041 | 100.00 | -0.017 | 0.011 | 0.020 | 40.72 | -0.005 | 0.012 | 0.013 | 7.64 |
| $\partial_w$ | 0.044 | 0.009 | 0.045 | 99.96 | 0.017 | 0.013 | 0.022 | 28.20 | 0.003 | 0.014 | 0.014 | 5.74 |

This table reports the simulated finite sample bias, standard deviation, RMSE, and size of the MLE and the MERM estimators and the corresponding t-tests for $\theta_{01}$ and $\theta_{02}$, and the marginal effects associated with $X^*$ and $W$ evaluated at the population mean. The true values of the considered parameters are $(\theta_{01}, \theta_{02}, \partial_x, \partial_w) = (1, 0, 0.105, 0)$. The results are based on 5,000 replications.

# References


Amemiya, Y. (1985): "Instrumental variable estimator for the nonlinear errors-in-variables model," *Journal of Econometrics*, 28, 273–289.

——— (1990): "Two-stage instrumental variables estimators for the nonlinear errors-in-variables model," *Journal of Econometrics*, 44, 311–332.

Andrews, D. W. K. and P. Guggenberger (2019): "Identification-and singularity-robust inference for moment condition models," *Quantitative Economics*, 10, 1703–1746.

Andrews, I. (2016): "Conditional Linear Combination Tests for Weakly Identified Models," *Econometrica*, 84, 2155–2182.

Andrews, I., M. Gentzkow, and J. M. Shapiro (2017): "Measuring the Sensitivity of Parameter Estimates to Estimation Moments," *Quarterly Journal of Economics*, 132, 1553–1592.

Andrews, I. and A. Mikusheva (2016): "Conditional Inference With a Functional Nuisance Parameter," *Econometrica*, 84, 1571–1612.

Angrist, J. D., K. Graddy, and G. W. Imbens (2000): "The Interpretation of Instrumental Variables Estimators in Simultaneous Equations Models with an Application to the Demand for Fish," *Review of Economic Studies*, 67, 499–527.

Armstrong, T. B. and M. Kolesár (2021): "Sensitivity analysis using approximate moment condition models," *Quantitative Economics*, 12, 77–108.

Battistin, E. and A. Chesher (2014): "Treatment effect estimation with covariate measurement error," *Journal of Econometrics*, 178, 707–715.

Ben-Moshe, D., X. D'Haultfœuille, and A. Lewbel (2017): "Identification of additive and polynomial models of mismeasured regressors without instruments," *Journal of Econometrics*, 200, 207–222.

Bonhomme, S. and M. Weidner (2021): "Minimizing Sensitivity to Model Misspecification," *arXiv:1807.02161*.

Bound, J., C. Brown, and N. Mathiowetz (2001): "Measurement Error in Survey Data," in *Handbook of Econometrics*, Elsevier, 3705–3843.

Chen, X., H. Hong, and D. Nekipelov (2011): "Nonlinear Models of Measurement Errors," *Journal of Economic Literature*, 49, 901–937.

Chen, X., H. Hong, and E. Tamer (2005): "Measurement Error Models with Auxiliary Data," *The Review of Economic Studies*, 72, 343–366.

Chesher, A. (1991): "The effect of measurement error," *Biometrika*, 78, 451–462.

——— (2000): "Measurement Error Bias Reduction," Working paper.

——— (2017): "Understanding the effect of measurement error on quantile regressions," *Journal of Econometrics*, 200, 223–237.

Chesher, A., M. Dumangane, and R. J. Smith (2002): "Duration response measurement error," *Journal of Econometrics*, 111, 169–194.

Chesher, A. and C. Schluter (2002): "Welfare measurement and measurement error," *The Review of Economic Studies*, 69, 357–378.

Erickson, T. and T. M. Whited (2002): "Two-Step Gmm Estimation Of The Errors-In-Variables Model Using High-Order Moments," *Econometric Theory*, 18.

Evdokimov, K. S. and A. Zeleneev (2018): "Issues of Nonstandard Inference in Measurement Error Models," .

——— (2019): "Errors-In-Variables in Large Nonlinear Panel and Network Models," .

Guggenberger, P., J. J. Ramalho, and R. J. Smith (2012): "GEL statistics under weak identification," *Journal of Econometrics*, 170, 331–349.

Guggenberger, P. and R. J. Smith (2005): "Generalized empirical likelihood estimators and tests under partial, weak, and strong identification," *Econometric Theory*, 21, 667–709.

Hahn, J., J. Hausman, and J. Kim (2021): "A small sigma approach to certain problems in errors-in-variables models," *Economics Letters*, 208, 110094.

Hansen, B. E. (2022): *Econometrics*, Princeton University Press (forthcoming).

Hansen, L. P. (1982): "Large Sample Properties of Generalized Method of Moments Estimators," *Econometrica*, 50, 1029–1054.

Hausman, J. A., H. Ichimura, W. K. Newey, and J. L. Powell (1991): "Identification and estimation of polynomial errors-in-variables models," *Journal of Econometrics*, 50, 273–295.

Hausman, J. A., W. K. Newey, and J. L. Powell (1995): "Nonlinear errors in variables Estimation of some Engel curves," *Journal of Econometrics*, 65, 205–233.

Heckman, J. and E. Vytlacil (1998): "Instrumental Variables Methods for the Correlated Random Coefficient Model: Estimating the Average Rate of Return to Schooling When the Return is Correlated with Schooling," *The Journal of Human Resources*, 33, 974.

Hong, H. and E. Tamer (2003): "A simple estimator for nonlinear error in variable models," *Journal of Econometrics*, 117, 1–19.

Hu, Y. and S. M. Schennach (2008): "Instrumental Variable Treatment of Nonclassical Measurement Error Models," *Econometrica*, 76, 195–216.

Imbens, G. and W. Newey (2009): "Identification and Estimation of Triangular Simultaneous Equations Models without Additivity," *Econometrica*, 77, 1481–1512.

Imbens, G. W. and J. D. Angrist (1994): "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, 62, 467–475.

Kadane, J. B. (1971): "Comparison of k-Class Estimators When the Disturbances Are Small," *Econometrica*, 39, 723.

Kitamura, Y., T. Otsu, and K. Evdokimov (2013): "Robustness, infinitesimal neighborhoods, and moment restrictions," *Econometrica*, 81, 1185–1201.

Kleibergen, F. (2005): "Testing Parameters in GMM Without Assuming that They are Identified," *Econometrica*, 73, 1103–1123.

Koppelman, F. S. and C.-H. Wen (2000): "The paired combinatorial logit model: properties, estimation and application," *Transportation Research Part B: Methodological*, 34, 75–89.

Lewbel, A. (1997): "Constructing Instruments for Regressions With Measurement Error When no Additional Data are Available, with An Application to Patents and R&D," *Econometrica*, 65, 1201–1213.

Li, T. (2002): "Robust and consistent estimation of nonlinear errors-in-variables models," *Journal of Econometrics*, 110, 1–26.

McFadden, D. (1974): "The measurement of urban travel demand," *Journal of public economics*, 3, 303–328.

Newey, W. K. (2001): "Flexible Simulated Moment Estimation of Nonlinear Errors-in-Variables Models," *The Review of Economics and Statistics*, 83, 616–627.

Newey, W. K. and D. McFadden (1994): "Large sample estimation and hypothesis testing," *Handbook of Econometrics*, IV.

Reiersøl, O. (1950): "Identifiability of a Linear Relation between Variables Which Are Subject to Error," *Econometrica*, 18, 375–389.

Rivers, D. and Q. H. Vuong (1988): "Limited Information Estimators And Exogeneity Tests For Simultaneous Probit Models," *Journal of Econometrics*, 39, 347–366.

Schennach, S. M. (2004): "Estimation of Nonlinear Models with Measurement Error," *Econometrica*, 72, 33–75.

——— (2007): "Instrumental Variable Estimation of Nonlinear Errors-in-Variables Models," *Econometrica*, 75, 201–239.

——— (2014): "Entropic Latent Variable Integration via Simulation," *Econometrica*, 82, 345–385.

——— (2016): "Recent Advances in the Measurement Error Literature," *Annual Review of Economics*, 8, 341–377.

——— (2020): "Mismeasured and unobserved variables," in *Handbook of Econometrics*, Elsevier, 487–565.

Schennach, S. M. and Y. Hu (2013): "Nonparametric Identification and Semiparametric Estimation of Classical Measurement Error Models Without Side Information," *Journal of the American Statistical Association*, 108, 177–186.

Smith, R. J. and R. W. Blundell (1986): "An Exogeneity Test for a Simultaneous Equation Tobit Model with an Application to Labor Supply," *Econometrica*, 54, 679.

Staiger, D. and J. H. Stock (1997): "Instrumental Variables Regression with Weak Instruments," *Econometrica*, 65, 557–586.

Stock, J. H. and J. H. Wright (2000): "GMM with Weak Identification," *Econometrica*, 68, 1055–1096.

Stone, C. J. (1980): "Optimal Rates of Convergence for Nonparametric Estimators," *Annals of Statistics*, 8, 1348–1360.

Wang, L. and C. Hsiao (2011): "Method of moments estimation and identifiability of semiparametric nonlinear errors-in-variables models," *Journal of Econometrics*, 165, 30–44.

Wen, C.-H. and F. S. Koppelman (2001): "The generalized nested logit model," *Transportation Research Part B: Methodological*, 35, 627–641.

Wolter, K. M. and W. A. Fuller (1982): "Estimation of Nonlinear Errors-in-Variables Models," *The Annals of Statistics*, 10, 539–548.

# A Regularity Conditions

**Notation.** Let $\mathcal{X} \subseteq \mathbb{R}$ be some closed convex set containing the union of the supports of $X_i^*$ and $X_i$, and $\mathcal{S} = \text{supp}\,(S_i)$.

**Assumption A.1.** (Moment function) *Suppose that the moment restrictions (1) are satisfied and the following conditions hold:*

*(i) For all $s \in \mathcal{S}$ and $\theta \in \Theta$, $g_x^{(K)}(x, s, \theta)$ exists and is continuous on $\mathcal{X}$. Moreover, there exist functions $b_1, b_2 : \mathcal{X} \times \mathcal{S} \times \Theta \to \mathbb{R}_+$ and integer $M \geqslant K+1$ such that for all $x, x' \in \mathcal{X}$, $s \in \mathcal{S}$, and $\theta \in \Theta$,*

$$\left\|g_x^{(K)}(x', s, \theta) - g_x^{(K)}(x, s, \theta)\right\| \leqslant b_1(x, s, \theta)|x' - x| + b_2(x, s, \theta)|x' - x|^{M-K}; \quad \text{(A.1)}$$

*(ii) Assumption MME holds with $L \geqslant M$;*

*(iii) $\mathbb{E}\left[g_x^{(k)}(X_i^*, S_i, \theta_0)\right]$, $k \in \{1, \ldots, K\}$, and $\mathbb{E}\left[b_j(X_i^*, S_i, \theta_0)\right]$, $j \in \{1, 2\}$, exist and are bounded.*

Assumption A.1 allows us to bound the remainder of the Taylor expansion of $g(X_i, S_i, \theta)$ around $X_i^*$ by a polynomial in $|X_i - X_i^*| = |\varepsilon_i|$. Combined with Assumption MME (which bounds the moments of $\varepsilon_i$), it ensures that this remainder is $o(n^{-1/2})$, which is crucial for establishing validity of the corrected moment function $\psi$ (Lemma 1).

Notice that if $\mathcal{X}$ is compact, condition (A.1) is satisfied if $g_x^{(K+1)}(x, s, \theta)$ is continuous on $\mathcal{X}$ (for all $s \in \mathcal{S}$ and $\theta \in \Theta$). If $\mathcal{X}$ is unbounded, condition (A.1) is satisfied if for some $J$, such that $K < J \leqslant M$, $\sup_{x \in \mathcal{X}} \left\|g_x^{(J)}(x, s, \theta)\right\| \leqslant B(s, \theta)$ for some function $B(s, \theta)$. Also notice that condition (A.1) is stronger than the standard Lipschitz continuity because in many applications $\left\|g_x^{(K)}(x, s, \theta)\right\|$ behaves like a polynomial in $x$ for large $x$.

**Assumption A.2.** (Parameter space)

*(i) $\Theta \subset \mathbb{R}^p$ and $\Gamma \subset \mathbb{R}^{K-1}$ are compact, $\theta_0 \in \text{int}\,(\Theta)$ and $\gamma_{0n} \in \Gamma$;*

*(ii) $0_{K-1} \in \text{int}\,(\Gamma)$.*

**Assumption A.3.** (Regularity and smoothness conditions)

*(i) For all $s \in \mathcal{S}$, $G_x^{(K)}(x, s, \theta)$ exists and is continuous on $\mathcal{X} \times \Theta$; moreover, there exist functions $b_{G1}, b_{G2} : \mathcal{X} \times \mathcal{S} \times \Theta \to \mathbb{R}_+$ and $\delta > 0$ and for all $x, x' \in \mathcal{X}$, $s \in \mathcal{S}$, and $\theta \in B_\delta(\theta_0)$*

$$\left\|G_x^{(K)}(x', s, \theta) - G_x^{(K)}(x, s, \theta)\right\| \leqslant b_{G1}(x, s, \theta)|x' - x| + b_{G2}(x, s, \theta)|x' - x|^{M-K}$$

*(ii)* $\mathbb{E}\left[\left\|g_x^{(k)}(X_i^*, S_i, \theta_0)\right\|^2\right]$, $\mathbb{E}\left[\sup_{\theta\in\Theta}\left\|g_x^{(k)}(X_i^*, S_i, \theta)\right\|\right]$, *for* $k \in \{0, \ldots, K\}$, *and* $\mathbb{E}\left[b_j(X_i^*, S_i, \theta_0)^2\right]$, $\mathbb{E}\left[\sup_{\theta\in\Theta} b_j(X_i^*, S_i, \theta)\right]$, *for* $j \in \{1, 2\}$, *are bounded;*

*(iii) for some* $\delta > 0$, $\mathbb{E}\left[\sup_{\theta\in B_\delta(\theta_0)}\left\|G_x^{(k)}(X_i^*, S_i, \theta)\right\|\right]$, *for* $k \in \{0, \ldots, K\}$, *and* $\mathbb{E}\left[\sup_{\theta\in B_\delta(\theta_0)} b_{Gj}(X_i^*, S_i, \theta)\right]$, *for* $j \in \{1, 2\}$, *are bounded;*

*(iv)* $\hat{\Xi} \xrightarrow{p} \Xi$, *where* $\Xi$ *is a symmetric positive definite matrix;*

*(v) Assumption MME holds with* $L \geqslant 2M$.

**Assumption A.4.** (Global and local identification)

*(i)* $\mathbb{E}\left[\psi(X_i^*, S_i, \theta, \gamma)\right] = 0$ *iff* $\theta = \theta_0$ *and* $\gamma = 0$;

*(ii)* $\Psi^{*\prime}\Xi\Psi^*$ *is invertible, where*

$$\Psi^* = \mathbb{E}\left[\Psi(X_i^*, S_i, \theta_0, 0)\right] = \mathbb{E}\left[G(X_i^*, S_i, \theta_0), -g_x^{(2)}(X_i^*, S_i, \theta_0), \ldots, -g_x^{(K)}(X_i^*, S_i, \theta_0)\right].$$

Assumption A.2-A.4 is a collection of basic regularity conditions, which help to ensure $\sqrt{n}$-consistency and asymptotic normality of the suggested estimator $\hat{\theta}$. Specifically, Assumption A.3 (i) is a counterpart of Assumption A.1 (i) applied to the Jacobian function. It ensures that the effect of the measurement error on the Jacobian is localized and allows us to establish $G \to G^*$, so $\Psi \to \Psi^*$. As a result, the asymptotic properties of $\hat{\theta}$ are controlled by $G^*$ (and $\Psi^*$), the Jacobian associated with the correctly measured variables. Assumptions A.4 (i) and (ii) are the standard GMM global and local identification conditions applied to the "limiting" moment function $\psi(X_i^*, S_i, \theta, \gamma)$.

**Remark A.1.** *Assumption CME requires* $\varepsilon_i$ *to be independent from* $X_i^*$ *and* $S_i$. *This requirement is stronger than needed and can be weakened to* $\mathbb{E}\left[\varepsilon_i^k | X_i^*, S_i\right] = \mathbb{E}\left[\varepsilon_i^k\right]$ *for* $k \in \{1, \ldots, K\}$, *i.e., the first* $K$ *conditional moments of* $\varepsilon_i$ *need to be independent of* $X_i^*$ *and* $S_i$. *The higher moments of* $\varepsilon_i$ *could depend on* $X_i^*$ *and* $S_i$ *but Assumption MME (ii) needs to be adjusted as* $\mathbb{E}\left[|\varepsilon_i|^L | X_i^*, S_i\right] \leqslant C\sigma_\varepsilon^L$ *for some* $L \geqslant K + 1$ *and* $C > 0$ *a.s. In this case, the statements of Lemma 1 and Theorem 2 remain correct, and all the proofs provided in Appendix D remain nearly identical. For example, the argument provided in the proof of Lemma 1 combined with conditioning on* $X_i^*$ *and* $S_i$ *implies* $\mathbb{E}\left[\psi(X_i, S_i, \theta_0, \gamma_0) | X_i^*, S_i\right] = \mathbb{E}\left[g(X_i^*, S_i, \theta_0) | X_i^*, S_i\right] + r(X_i^*, S_i)$, *where the expected value of the remainder can be explicitly bounded as* $\mathbb{E}\left[\|r(X_i^*, S_i)\|\right] = O(\tau_n^{K+1}) = o(n^{-1/2})$. *Then, the statement of Lemma 1 follows from an application of the law of iterated expectations.*

**Remark A.2.** *Assumption A.2 (ii) ensures that $\gamma_0$ (which approaches 0 as $n$ increases under the considered asymptotics) is sufficiently far from the boundary of the parameter space $\Gamma$. Without imposing this requirement, $\hat{\theta}$ still remains $\sqrt{n}$-consistent but is not necessarily asymptotically normal.*

*At the same time, since $\gamma_0$ is determined by the moments of $\varepsilon_i$, it might be desirable to incorporate that relationship into construction of the parameter space $\Gamma$. For example, since $\gamma_{02} = \mathbb{E}[\varepsilon_i^2]/2$, it is natural to restrict $\gamma_2 \geqslant 0$. Incorporation of such additional constraints could improve efficiency of the estimators and informativeness of inference (see Evdokimov and Zeleneev (2018) for a more detailed discussion).*

# B MERM derivation when $\sigma_\varepsilon$ is not small

Note that $\tau$ can be small without $\sigma_\varepsilon$ being small in absolute magnitude. For example, suppose $\sigma_\varepsilon = 10$ and $\sigma_{X^*} = 100$. Then $\tau = 0.1$, so the measurement error is quite small relative to $\sigma_{X^*}$, and relying on the approximation $\tau \to 0$ is reasonable. At the same time, approximation $\sigma_\varepsilon \to 0$ may not be suitable for this example.

In this Appendix we show that the corrected moment conditions and the MERM estimator are valid without assuming that $\sigma_\varepsilon$ is small in absolute magnitude. In Section 2.1 we used Taylor expansions in $\varepsilon_i$ around $\varepsilon_i = 0$, with the remainder of order $\mathbb{E}\left[|\varepsilon_i|^{K+1}\right]$. When $\sigma_\varepsilon > 1$, term $O\left(\mathbb{E}\left[|\varepsilon_i|^{K+1}\right]\right)$ in equation (9) cannot be viewed as a negligible remainder, because $\mathbb{E}\left[|\varepsilon_i|^{K+1}\right] > 1$ and, moreover, terms $\mathbb{E}\left[|\varepsilon_i|^{k}\right]$ increase rather than decrease with $k$.

In Section 2.1, for to simplify the exposition, we have assumed that $X^*$ is scaled so that $\sigma_{X^*}$ is of order one. This in particular ensures that $\mathbb{E}\left[|\varepsilon_i|^{k}\right]$ decrease with $k$. We will now show that this assumption about the scale of $X^*$ is not necessary, and that the procedure remains valid without any such scaling.

We will show that by rescaling the Taylor expansions in Section 2.1 can be written in terms of powers of $\tau^k$, which necessarily decrease with $k$ when $\tau < 1$.

Remember the model of Section 2.1:

$$\mathbb{E}[g(X_i^*, S_i, \theta_0)] = 0, \quad X_i = X_i^* + \varepsilon_i, \quad \mathbb{E}[\varepsilon_i] = 0. \tag{B.1}$$

Let $\xi_i$ denote a random variable with $\mathbb{E}[\xi_i] = 0$ and $\mathbb{E}\left[\xi_i^2\right] = 1$, $\mathbb{E}\left[|\xi_i|^{L+1}\right]$ is bounded,

and $\varepsilon_i \equiv \sigma_\varepsilon \xi_i$. Also, let us denote

$$\tau \equiv \sigma_\varepsilon / \sigma_{X^*}, \quad \widetilde{X}_i \equiv X_i / \sigma_{X^*}, \quad \widetilde{X}_i^* \equiv X_i^* / \sigma_{X^*}, \quad \widetilde{g}(\widetilde{x}, s, \theta) \equiv g(\sigma_{X^*} \widetilde{x}, s, \theta).$$

Then, we can rewrite equation (B.1) as

$$\mathbb{E}[\widetilde{g}(\widetilde{X}_i^*, S_i, \theta_0)] = 0, \quad \widetilde{X}_i = \widetilde{X}_i^* + \tau \xi_i, \quad \mathbb{E}[\xi_i] = 0.$$

Expand $\widetilde{g}(\widetilde{X}_i, S_i, \theta) = \widetilde{g}(\widetilde{X}_i^* + \tau \xi_i, S_i, \theta)$ around $\tau = 0$ to obtain

$$\mathbb{E}[\widetilde{g}(\widetilde{X}_i, S_i, \theta)] = \mathbb{E}[\widetilde{g}(\widetilde{X}_i^*, S_i, \theta)] + \sum_{k=2}^{K} \frac{\tau^k \mathbb{E}\left[\xi_i^k\right]}{k!} \mathbb{E}\left[\widetilde{g}_x^{(k)}(\widetilde{X}_i^*, S_i, \theta)\right] + O(\tau^{K+1}),$$

which is similar to equation (9), except $\mathbb{E}\left[\varepsilon_i^k\right]$ is replaced by $\tau^k \mathbb{E}\left[\xi_i^k\right]$, and $\widetilde{X}_i$, $\widetilde{X}_i^*$, $\widetilde{g}$ are replaced by $X_i$, $X_i^*$, $g$. Then, the corrected moment condition has the form

$$\widetilde{\psi}(\widetilde{X}_i, S_i, \theta, \widetilde{\gamma}) = \widetilde{g}(\widetilde{X}_i, S_i, \theta) - \sum_{k=2}^{K} \widetilde{\gamma}_k \widetilde{g}_x^{(k)}(\widetilde{X}_i, S_i, \theta), \tag{B.2}$$

where true parameter values $\widetilde{\gamma}_0$ are $\widetilde{\gamma}_{02} = \tau^2 \mathbb{E}\left[\xi_i^2\right]/2 = \tau^2/2$, $\widetilde{\gamma}_{03} = \tau^3 \mathbb{E}\left[\xi_i^3\right]/6$, and $\widetilde{\gamma}_{0k} = \frac{\tau^k \mathbb{E}\left[\xi_i^k\right]}{k!} - \sum_{\ell=2}^{k-2} \frac{\tau^{k-\ell} \mathbb{E}\left[\xi_i^{k-\ell}\right]}{(k-\ell)!} \widetilde{\gamma}_{0\ell}$ for $k \geq 4$.

We will now show that

$$\gamma_{0k} = \sigma_{X^*}^k \widetilde{\gamma}_{0k} \text{ for all } k \geq 2.$$

First, $\gamma_{02} = \mathbb{E}\left[\varepsilon_i^2\right]/2 = \mathbb{E}\left[(\sigma_\varepsilon \xi_i)^2\right]/2 = \sigma_{X^*}^2 \widetilde{\gamma}_{02}$, $\gamma_{03} = \mathbb{E}\left[\varepsilon_i^3\right]/6 = \sigma_{X^*}^3 \widetilde{\gamma}_{03}$ by definition. Then, for $k \geq 4$, by induction we have

$$\begin{aligned}
\gamma_{0k} &= \frac{\mathbb{E}\left[\varepsilon_i^k\right]}{k!} - \sum_{\ell=2}^{k-2} \frac{\mathbb{E}\left[\varepsilon_i^{k-\ell}\right]}{(k-\ell)!} \gamma_{0\ell} = \sigma_{X^*}^k \left( \frac{\left(\sigma_\varepsilon / \sigma_{X^*}\right)^k \mathbb{E}\left[\xi_i^k\right]}{k!} - \sum_{\ell=2}^{k-2} \frac{\left(\sigma_\varepsilon / \sigma_{X^*}\right)^{k-\ell} \mathbb{E}\left[\xi_i^{k-\ell}\right]}{(k-\ell)!} \frac{\gamma_{0\ell}}{\sigma_{X^*}^\ell} \right) \\
&= \sigma_{X^*}^k \left( \frac{\tau^k \mathbb{E}\left[\xi_i^k\right]}{k!} - \sum_{\ell=2}^{k-2} \frac{\tau^{k-\ell} \mathbb{E}\left[\xi_i^{k-\ell}\right]}{(k-\ell)!} \widetilde{\gamma}_{0\ell} \right) = \sigma_{X^*}^k \widetilde{\gamma}_{0k}.
\end{aligned}$$

Finally, let us now show that moment condition $\widetilde{\psi}$ in equation (B.2) is numerically identical to $\psi$ in equation (11) with $\gamma_k = \sigma_{X^*}^k \widetilde{\gamma}_k$. Note that for $\widetilde{x} = x / \sigma_{X^*}$ we have

$\widetilde{g}_{\widetilde{x}}^{(k)}(\widetilde{x}, s, \theta) \equiv \nabla_{\widetilde{x}}^{k} g(\sigma_{X^*}\widetilde{x}, s, \theta) = \sigma_{X^*}^{k} g_a^{(k)}(a, s, \theta)|_{a=\sigma_{X^*}\widetilde{x}} = \sigma_{X^*}^{k} g_x^{(k)}(x, s, \theta)$, and hence

$$\begin{aligned}
\widetilde{\psi}(\widetilde{X}_i, S_i, \theta, \widetilde{\gamma}) &= g(\sigma_{X^*}\widetilde{X}_i, S_i, \theta) - \sum_{k=2}^{K} \left(\widetilde{\gamma}_k \sigma_{X^*}^{k}\right) g_x^{(k)}(\sigma_{X^*}\widetilde{X}_i, S_i, \theta) \\
&= g(X_i, S_i, \theta) - \sum_{k=2}^{K} \left(\widetilde{\gamma}_k \sigma_{X^*}^{k}\right) g_x^{(k)}(X_i, S_i, \theta) \\
&= \psi(X_i, S_i, \theta, \gamma).
\end{aligned}$$

# C Some Implementation Details

**Numerical Optimization.** Since $\overline{\psi}(\theta, \gamma)$ is a linear function of $\gamma$ it can be profiled out of the quadratic form $\hat{Q}(\theta, \gamma)$. Thus, the criterion function only needs to be minimized numerically over $\theta$.

**Choice of the weighting matrix $\hat{\Xi}$.** As for the standard GMM estimator, the optimal weighting matrix can be estimated by

$$\hat{\Xi}_{\text{eff}} \equiv \hat{\Omega}_{\psi\psi}^{-1}(\tilde{\theta}, \tilde{\gamma}),$$

where $\tilde{\theta}$ and $\tilde{\gamma}$ are some preliminary estimators of $\theta_0$ and $\gamma_0$, and $\hat{\Omega}_{\psi\psi}(\theta, \gamma) \equiv n^{-1}\sum_{i=1}^{n} \psi_i(\theta, \gamma)\psi_i(\theta, \gamma)'$. One example of such a preliminary estimator would be the 1-step (GMM-)MERM estimator using $\hat{\Xi}_{\text{GMM1}} \equiv \hat{\Omega}_{\psi\psi}^{-1}(\hat{\theta}_{\text{Naive}}, 0)$ as the first-step GMM weighting matrix, where $\hat{\theta}_{\text{Naive}}$ is a naive estimator of $\theta_0$ that ignores EIV. Note that $\hat{\Omega}_{\psi\psi}(\hat{\theta}_{\text{Naive}}, 0) = \hat{\Omega}_{gg}\left(\hat{\theta}_{\text{Naive}}\right)$, where $\hat{\Omega}_{gg}(\theta) \equiv n^{-1}\sum_{i=1}^{n} g_i(\theta)g_i(\theta)'$.

One may also consider the regularized version of the efficient weighting matrix estimator $\hat{\Xi}_{\text{eff,R}} \equiv \hat{\Omega}_{\psi\psi}^{-1}(\tilde{\theta}, 0)$. Since $\gamma_0 \to 0$, using the regularized version $\hat{\Xi}_{\text{eff,R}}$ does not lead to a loss of efficiency. Moreover, our simulation studies suggest that using the regularized weighting matrix $\hat{\Xi}_{\text{eff,R}}$ results in better finite sample performance of the MERM estimator and, hence, is recommended in practice.

Although not indicated by the notation in equation (12), the weighting matrix $\hat{\Xi} \equiv \hat{\Xi}(\theta, \gamma)$ is allowed to be a function of $\theta$ and $\gamma$. For example, Continuously Updating GMM Estimator (CUE) corresponds to taking $\hat{\Xi}_{\text{CUE}}(\theta, \gamma) \equiv \hat{\Omega}_{\psi\psi}^{-1}(\theta, \gamma)$. Similarly to $\hat{\Xi}_{\text{eff,R}}$, one may also consider $\hat{\Xi}_{\text{CUE,R}}(\theta, \gamma) \equiv \hat{\Omega}_{\psi\psi}^{-1}(\theta, 0)$ without introducing any loss of efficiency. In contrast to the criterion function of the CUE estimator, criterion function of $\hat{Q}_{\text{CUE,R}}(\theta, \gamma)$

is quadratic in $\gamma$. This implies that $\gamma$ can be profiled out analytically. This simplifies the numerical optimization problem reducing it to minimizing $\hat{Q}_{\text{CUE,R}}(\theta, \hat{\gamma}(\theta))$ over $\theta \in \Theta$. Then, the dimension of the optimization parameter $\theta$ for the corrected moment condition problem remains the same as for the original (naive) estimation problem without the EIV correction.

**Estimation of the asymptotic variance $\Sigma$.** Theorem 2 shows that the MERM estimator $\hat{\beta} = (\hat{\theta}', \hat{\gamma}')'$ behaves like a standard GMM estimator based on the corrected moment function $\psi(\theta, \gamma)$. The researchers can rely on the standard GMM inference procedures. The asymptotic variance of $\hat{\beta}$ can be consistently estimated by

$$\hat{\Sigma} \equiv (\hat{\Psi}'\hat{\Xi}\hat{\Psi})^{-1}\hat{\Psi}'\hat{\Xi}\hat{\Omega}_{\psi\psi}\hat{\Xi}\hat{\Psi}(\hat{\Psi}'\hat{\Xi}\hat{\Psi})^{-1},$$

where, $\hat{\Xi}$ is the chosen weighting matrix, and $\hat{\Psi} \equiv \overline{\Psi}(\hat{\theta}, \hat{\gamma}) = n^{-1}\sum_{i=1}^{n}\Psi_i(\hat{\theta}, \hat{\gamma})$ and $\hat{\Omega}_{\psi\psi} = \hat{\Omega}_{\psi\psi}(\hat{\theta}, \hat{\gamma})$ are estimators of $\Psi$ and $\Omega_{\psi\psi}$.

# D Proofs of Results in Appendix A

**Notation.** To stress that in our asymptotic approximation the variance and the higher moments of $\varepsilon_i$ depend on $n$, we will use $\sigma_n^2 \equiv \mathbb{E}[\varepsilon_i^2]$, $\gamma_{0n} \equiv \gamma_0$, and $\beta_{0n} \equiv \beta_0 \equiv (\theta_0', \gamma_{0n}')'$.

All vectors are columns. For some generic parameter vector $\alpha$ and a vector (or matrix) valued function $a(x, s, \alpha)$ and , let $a_i(\beta) \equiv a(X_i, S_i, \alpha)$, $\overline{a}(\alpha) \equiv n^{-1}\sum_{i=1}^{n} a_i(\alpha)$, $a(\alpha) \equiv \mathbb{E}[a_i(\alpha)]$. Similarly, we let $a_i^*(\alpha) \equiv a(X_i^*, S_i, \alpha)$, $\overline{a}^*(\alpha) \equiv n^{-1}\sum_{i=1}^{n} a_i^*(\alpha)$, $a^*(\alpha) \equiv \mathbb{E}[a_i^*(\alpha)]$.

For the true value of the parameter $\alpha_0$, we often write $a_i \equiv a(\alpha_0)$, $\overline{a} \equiv \overline{a}(\alpha_0)$, $a \equiv a(\alpha_0)$, $a_i^* \equiv a(\alpha_0)$, $\overline{a}^* \equiv \overline{a}^*(\alpha_0)$, $a^* \equiv a^*(\alpha_0)$.

## D.1 Proof of Lemma 1

*Proof of Lemma 1.* Making use of Assumption A.1 (i) , we expand $g(X_i, S_i, \theta_0)$ around $X_i^*$ as

$$\begin{aligned} g(X_i, S_i, \theta_0) =& g(X_i^*, S_i, \theta_0) + g_x^{(1)}(X_i, S_i, \theta_0)\varepsilon_i + \sum_{k=2}^{K}\frac{1}{k!}g_x^{(k)}(X_i^*, S_i, \theta_0)\varepsilon_i^k \\ &+ \frac{1}{K!}\left(g_x^{(K)}(\tilde{X}_i, S_i, \theta_0) - g_x^{(K)}(X_i^*, S_i, \theta_0)\right)\varepsilon_i^K, \end{aligned} \tag{D.1}$$

where $\tilde{X}_i$ lies between $X_i^*$ and $X_i$ (and hereafter $\tilde{X}_i$ is allowed to be component specific). Similarly, for $k' \in \{2, \ldots, K\}$, we have

$$\begin{aligned} g_x^{(k)}(X_i, S_i, \theta_0) =& g_x^{(k)}(X_i^*, S_i, \theta_0) + \sum_{\ell=k+1}^{K} \frac{1}{(\ell-k)!} g_x^{(\ell)}(X_i^*, S_i, \theta_0)\varepsilon_i^{\ell} \\ &+ \frac{1}{(K-k)!}\left(g_x^{(K)}(\tilde{X}_{ki}, S_i, \theta_0) - g_x^{(K)}(X_i^*, S_i, \theta_0)\right)\varepsilon_i^{K-k}, \end{aligned} \tag{D.2}$$

where $\tilde{X}_{ki}$ lies between $X_i^*$ and $X_i$. Hence, combining these expressions and rearranging the terms, we obtain

$$\begin{aligned} \psi(X_i, S_i, \theta_0, \gamma) =& g(X_i, S_i, \theta_0) - \sum_{k=2}^{K} \gamma_k g_x^{(k)}(X_i, S_i, \theta_0) \\ =& g(X_i^*, S_i, \theta_0) + g_x^{(1)}(X_i^*, S_i, \theta_0)\varepsilon_i + \sum_{k=2}^{K} g_x^{(k)}(X_i^*, S_i, \theta_0)\left(\frac{1}{k!}\varepsilon_i^k - \sum_{\ell=2}^{k}\frac{1}{(k-\ell)!}\varepsilon_i^{k-\ell}\gamma_\ell\right) \\ &+ \frac{1}{K!}\left(g_x^{(K)}(\tilde{X}_i, S_i, \theta_0) - g_x^{(K)}(X_i^*, S_i, \theta_0)\right)\varepsilon_i^K \\ &- \sum_{k=2}^{K}\frac{\gamma_k}{(K-k)!}\left(g_x^{(K)}(\tilde{X}_{ki}, S_i, \theta_0) - g_x^{(K)}(X_i^*, S_i, \theta_0)\right)\varepsilon_i^{K-k}. \end{aligned} \tag{D.3}$$

We want to show that for a properly chosen $\gamma = \gamma_{0n}$, $\mathbb{E}\left[\psi(X_i, S_i, \theta_0, \gamma_{0n})\right] = o(n^{-1/2})$. Note that the first two terms in (D.3) are mean zero, i.e. we have

$$\mathbb{E}\left[g(X_i^*, S_i, \theta_0)\right] = 0, \tag{D.4}$$

$$\mathbb{E}\left[g_x^{(1)}(X_i, S_i, \theta_0)\varepsilon_i\right] = 0, \tag{D.5}$$

where the latter is guaranteed by Assumptions CME.

Second, we argue that for a properly chosen $\gamma$, we have

$$\mathbb{E}\left[\frac{1}{k!}\varepsilon_i^k - \sum_{\ell=2}^{k}\frac{1}{(k-\ell)!}\varepsilon_i^{k-\ell}\gamma_\ell\right] = 0, \tag{D.6}$$

for all $k \in \{2, \ldots, K\}$. Let us reparameterize $\gamma = (\gamma_2, \ldots, \gamma_K)'$ using $\gamma_k = \sigma_n^k a_k$. Then, (D.6) can be rewritten as

$$\mathbb{E}\left[\frac{1}{k!}(\varepsilon_i/\sigma_n)^k - \sum_{\ell=2}^{k}\frac{1}{(k-\ell)!}(\varepsilon_i/\sigma_n)^{k-\ell}a_\ell\right] = 0, \tag{D.7}$$

which can also be represented as

$$B_n a = c_n \tag{D.8}$$

where $a = (a_2, \ldots, a_K)'$, and

$$B_n = \begin{bmatrix} 1 & 0 & \ldots & 0 & 0 \\ \mathbb{E}[\varepsilon_i/\sigma_n] & 1 & \ldots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \frac{\mathbb{E}[(\varepsilon_i/\sigma_n)^{K-3}]}{(K-3)!} & \frac{\mathbb{E}[(\varepsilon_i/\sigma_n)^{K-4}]}{(K-4)!} & \ldots & 1 & 0 \\ \frac{\mathbb{E}[(\varepsilon_i/\sigma_n)^{K-2}]}{(K-2)!} & \frac{\mathbb{E}[(\varepsilon_i/\sigma_n)^{K-3}]}{(K-3)!} & \ldots & \mathbb{E}[(\varepsilon_i/\sigma_n)] & 1 \end{bmatrix}, \quad c_n = \begin{bmatrix} \mathbb{E}[(\varepsilon_i/\sigma_n)^2]/2! \\ \mathbb{E}[(\varepsilon_i/\sigma_n)^3]/3! \\ \vdots \\ \mathbb{E}[(\varepsilon_i/\sigma_n)^{K-1}]/(K-1)! \\ \mathbb{E}[(\varepsilon_i/\sigma_n)^K]/K! \end{bmatrix}.$$

Since $B_n$ is invertible, (D.8) has a unique solution $a_n = B_n^{-1} c_n$. Moreover, $a_n$ is bounded since both $B_n^{-1}$ and $c_n$ are bounded (Assumption MME). Hence, we conclude that (D.6) has a unique solution $\gamma_{0n} \equiv \left(\sigma_n^2 a_{2n}, \ldots, \sigma_n^K a_{Kn}\right)'$. Since (D.6) is satisfied, using Assumption CME, we also conclude that

$$\mathbb{E}\left[\sum_{k=2}^{K} g_x^{(k)}(X_i^*, S_i, \theta_0)\left(\frac{1}{k!}\varepsilon_i^k - \sum_{\ell=2}^{k} \frac{1}{(k-\ell)!}\varepsilon_i^{k-\ell}\gamma_{0\ell n}\right)\right] = 0. \tag{D.9}$$

To complete the proof of $\mathbb{E}\left[\psi(X_i, S_i, \theta_0, \gamma_{0n})\right] = o_n(n^{-1/2})$, it is sufficient to show that

$$\mathbb{E}\left[\left(g_x^{(K)}(\tilde{X}_i, S_i, \theta_0) - g_x^{(K)}(X_i^*, S_i, \theta_0)\right)\varepsilon_i^K\right] = o(n^{-1/2}), \tag{D.10}$$

$$\gamma_{0kn}\mathbb{E}\left[\left(g_x^{(K)}(\tilde{X}_{ki}, S_i, \theta_0) - g_x^{(K)}(X_i^*, S_i, \theta_0)\right)\varepsilon_i^{K-k}\right] = o(n^{-1/2}) \quad \text{for } k \in \{2, \ldots, K\}. \tag{D.11}$$

We start with (D.10). Using Assumption A.1 (i), we obtain

$$\left\|\left(g_x^{(K)}(\tilde{X}_i, S_i, \theta_0) - g_x^{(K)}(X_i^*, S_i, \theta_0)\right)\varepsilon_i^K\right\| \leqslant b_1(X_i^*, S_i, \theta_0)\left|\varepsilon_i\right|^{K+1} + b_2(X_i^*, S_i, \theta_0)\left|\varepsilon_i\right|^M. \tag{D.12}$$

Hence, using Assumption CME, and the fact $\left|\tilde{X}_i - X_i^*\right| \leqslant \varepsilon_i$, we get

$$\begin{aligned} &\mathbb{E}\left[\left(g_x^{(K)}(\tilde{X}_i, S_i, \theta_0) - g_x^{(K)}(X_i^*, S_i, \theta_0)\right)\varepsilon_i^K\right] \\ &\leqslant \sigma_n^{K+1}\mathbb{E}\left[b_1(X_i^*, S_i, \theta_0)\right]\mathbb{E}\left[\left|\varepsilon_i/\sigma_n\right|^{K+1}\right] + \sigma_n^M\mathbb{E}\left[b_2(X_i^*, S_i, \theta_0)\right]\mathbb{E}\left[\left|\varepsilon_i/\sigma_n\right|^M\right]. \end{aligned}$$

Since (i) the expectations above are bounded (Assumptions MME, A.1 (ii), and A.1 (iii)) and (ii) $\sigma_n^{K+1} = o(n^{-1/2})$ and $\sigma_n^M = o(n^{-1/2})$ (Assumption MME), this implies that (D.10) holds. To inspect (D.11), recall that $\gamma_{0kn} = \sigma_n^k a_{kn}$. As a result, using Assumptions A.1 (i) and CME, and $\left|\tilde{X}_{ki} - X_i^*\right| \leqslant \varepsilon_i$ again, we also have

$$\begin{aligned}
&\gamma_{0kn}\mathbb{E}\left[\left(g_x^{(K)}(\tilde{X}_{ki}, S_i, \theta_0) - g_x^{(K)}(X_i^*, S_i, \theta_0)\right)\varepsilon_i^{K-k}\right] \\
&\leqslant a_{kn}\left(\sigma_n^{K+1}\mathbb{E}\left[b_1(X_i^*, S_i, \theta_0)\right]\mathbb{E}\left[\left|\varepsilon_i/\sigma_n\right|^{K+1-k}\right] + \sigma_n^M\mathbb{E}\left[b_2(X_i^*, S_i, \theta_0)\right]\mathbb{E}\left[\left|\varepsilon_i/\sigma_n\right|^{M-k}\right]\right).
\end{aligned}$$

Since $a_{kn}$ is bounded, we conclude that (D.11) holds analogously to (D.10).

Combining (D.3) with (D.4), (D.5), and (D.9)-(D.11), we conclude that $\mathbb{E}\left[\psi(X_i, S_i, \theta_0, \gamma_{0n})\right] = o(n^{-1/2})$.

Finally, we want to verify the recursive expressions for the components of $\gamma_{0n}$. Note that (D.6) can be represented as

$$B_n^\gamma \gamma_{0n} = c_n^\gamma, \tag{D.13}$$

where $\gamma_{0n} = (\gamma_{02n}, \ldots, \gamma_{0Kn})'$, and

$$B_n^\gamma = \begin{bmatrix} 1 & 0 & \ldots & 0 & 0 \\ \mathbb{E}[\varepsilon_i] & 1 & \ldots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \frac{\mathbb{E}[\varepsilon_i^{K-3}]}{(K-3)!} & \frac{\mathbb{E}[\varepsilon_i^{K-4}]}{(K-4)!} & \ldots & 1 & 0 \\ \frac{\mathbb{E}[\varepsilon_i^{K-2}]}{(K-2)!} & \frac{\mathbb{E}[\varepsilon_i^{K-3}]}{(K-3)!} & \ldots & \mathbb{E}[\varepsilon_i] & 1 \end{bmatrix}, \quad c_n^\gamma = \begin{bmatrix} \mathbb{E}[\varepsilon_i^2]/2! \\ \mathbb{E}[\varepsilon_i^3]/3! \\ \vdots \\ \mathbb{E}[\varepsilon_i^{K-1}]/(K-1)! \\ \mathbb{E}[\varepsilon_i^K]/K! \end{bmatrix}.$$

Notice that $\gamma_{02n} = \mathbb{E}\left[\varepsilon_i^2\right]/2$ and $\gamma_{03n} = \mathbb{E}\left[\varepsilon_i^3\right]/6$ (since $\mathbb{E}\left[\varepsilon_i\right] = 0$). To recursively compute $\gamma_{0kn}$ for $k \geqslant 4$, suppose that $\gamma_{0\ell n}$ are known for $\ell \in \{2, \ldots, k-1\}$. Then $\gamma_{0kn}$ can be recursively computed from the $k-1$-th equation in (D.13), which takes the form of

$$\sum_{\ell=2}^{k} \frac{\mathbb{E}\left[\varepsilon_i^{k-\ell}\right]}{(k-\ell)!}\gamma_{0\ell n} = \frac{\mathbb{E}\left[\varepsilon_i^k\right]}{k!}.$$

Plugging $\mathbb{E}\left[\varepsilon_i\right] = 0$ and rearranging the terms gives

$$\gamma_{0kn} = \frac{\mathbb{E}\left[\varepsilon_i^k\right]}{k!} - \sum_{\ell=2}^{k-2} \frac{\mathbb{E}\left[\varepsilon_i^{k-\ell}\right]}{(k-\ell)!}\gamma_{0\ell n},$$

which completes the proof. Q.E.D.

## D.2 Proof of Theorem 2

### D.2.1 Auxiliary lemmas

**Lemma D.1.** *Suppose that* $\{(X_i^*, S_i', \varepsilon_i)\}_{i=1}^n$ *are i.i.d.. Then, under Assumptions MME, CME, A.1, A.2* (i)*, and A.3* (i)-(iii)*, we have*

*(i)*

$$\sup_{\theta\in\Theta} \left\| \overline{g}_x^{(k)}(\theta) - g_x^{(k)*}(\theta) \right\| = o_p(1)$$

*and* $g_x^{(k)*}(\theta)$ *is continuous on* $\Theta$ *for* $k \in \{0, \ldots, K\}$*;*

*(ii) for some* $\delta > 0$*,*

$$\sup_{\theta\in B_\delta(\theta_0)} \left\| \overline{G}_x^{(k)}(\theta) - G_x^{(k)*}(\theta) \right\| = o_p(1),$$

*and* $G_x^{(k)*}(\theta)$ *is continuous on* $B_\delta(\theta_0)$ *for* $k \in \{0, \ldots, K\}$*.*

*Proof of Lemma D.1.* First, we show

$$\sup_{\theta\in\Theta} \|\overline{g}(\theta) - g^*(\theta)\| = o_p(1).$$

By the triangle inequality,

$$\sup_{\theta\in\Theta} \|\overline{g}(\theta) - g^*(\theta)\| \leqslant \sup_{\theta\in\Theta} \|\overline{g}(\theta) - \overline{g}^*(\theta)\| + \sup_{\theta\in\Theta} \|\overline{g}^*(\theta) - g^*(\theta)\|.$$

Then, it is sufficient to show that both terms on the right hand side of the inequality above are $o_p(1)$. Expanding $g(X_i, S_i, \theta_0)$ around $X_i^*$ as in (D.1) and invoking Assumption

A.1 (i),

$$\begin{aligned}\sup_{\theta\in\Theta}\|\overline{g}(\theta)-\overline{g}^*(\theta)\| =& \sup_{\theta\in\Theta}\left\|\sum_{k=1}^{K-1}\frac{1}{k!}\frac{1}{n}\sum_{i=1}^{n}g_x^{(k)}(X_i^*,S_i,\theta)\varepsilon_i^k+\frac{1}{K!}\frac{1}{n}\sum_{i=1}^{n}g_x^{(K)}(\tilde{X}_i^*,S_i,\theta)\varepsilon_i^K\right\| \\ \leqslant& \sum_{k=1}^{K}\frac{1}{k!}\underbrace{\frac{1}{n}\sum_{i=1}^{n}\sup_{\theta\in\Theta}\left\|g_x^{(k)}(X_i^*,S_i,\theta)\right\||\varepsilon_i|^k}_{o_p(1)}+\underbrace{\frac{1}{K!}\frac{1}{n}\sum_{i=1}^{n}\sup_{\theta\in\Theta}b_1(X_i^*,S_i,\theta)|\varepsilon_i|^{K+1}}_{o_p(1)} \\ &+\frac{1}{K!}\underbrace{\frac{1}{n}\sum_{i=1}^{n}\sup_{\theta\in\Theta}b_2(X_i^*,S_i,\theta)|\varepsilon_i|^M}_{o_p(1)},\end{aligned}$$

where $\tilde{X}_i$ lies in between of $X_i^*$ and $X_i$. Now observe that all the terms following the inequality sign are $o_p(1)$. Indeed, this is guaranteed by Markov's inequality paired with Assumptions MME, CME, and A.3 (ii). Hence, $\sup_{\theta\in\Theta}\|\overline{g}(\theta)-\overline{g}^*(\theta)\|=o_p(1)$, and we are left to show $\sup_{\theta\in\Theta}\|\overline{g}^*(\theta)-g^*(\theta)\|=o_p(1)$. This, in turn, follows from the standard ULLN (e.g., Lemma 2.4 in Newey and McFadden, 1994), which also ensures continuity of $g^*(\theta)$ on $\Theta$. Hence, we conclude that the assertion of the lemma holds for $g$.

Applying nearly identical arguments, one can also establish the desired results for $g_x^{(k)}$ for $k\in\{1,\ldots,K\}$ and for $G_x^{(k)}$ for $k\in\{0,\ldots,K\}$ (for the latter, Assumptions A.3 (i) and (iii) take the places of Assumptions A.1 (i) and A.3 (ii), respectively). Q.E.D.

**Lemma D.2.** *Suppose that the hypotheses of Lemma D.1 are satisfied. Then, $g_x^{(k)}\to g_x^{(k)*}$ and $G_x^{(k)}\to G_x^{(k)*}$ for $k\in\{0,\ldots,K\}$. Suppose also $\hat{\theta}\xrightarrow{p}\theta_0$. Then, $\overline{g}_x^{(k)}(\hat{\theta})\xrightarrow{p}g_x^{(k)*}$ and $\overline{G}_x^{(k)}(\hat{\theta})\xrightarrow{p}G_x^{(k)*}$ for $k\in\{0,\ldots,K\}$.*

*Proof of Lemma D.2.* First, we prove the assertions of the lemma for $g_x^{(k)}$. Note that, by the standard expansion of $g_x^{(k)}(X_i,S_i,\theta_0)$ around $X_i^*$ (see Eq. (D.2) above), we have

$$\begin{aligned}\left\|g_x^{(k)}-g_x^{(k)^*}\right\| \leqslant& \mathbb{E}\left[\left\|g(X_i,S_i,\theta_0)-g_x^{(k)}(X_i^*,S_i,\theta)\right\|\right] \\ \leqslant& \sum_{\ell=k+1}^{K}\frac{1}{(\ell-k)!}\mathbb{E}\left[\left\|g_x^{(\ell)}(X_i^*,S_i,\theta_0)\right\||\varepsilon_i|^\ell\right] \\ &+\frac{1}{(K-k)!}\mathbb{E}\left[\left\|\left(g_x^{(K)}(\tilde{X}_i,S_i,\theta_0)-g_x^{(K)}(X_i^*,S_i,\theta_0)\right)\right\||\varepsilon_i|^{K-k}\right].\end{aligned}$$

By Assumptions MME, CME, and A.3 (ii), $\mathbb{E}\left[\left\|g_x^{(\ell)}(X_i^*,S_i,\theta_0)\right\||\varepsilon_i|^\ell\right]\to 0$ for all $\ell\in$

$\{1,\dots,K\}$. Next, using Assumptions A.1 (i) and CME,

$$\begin{aligned}&\mathbb{E}\left[\left\|\left(g_x^{(K)}(\tilde{X}_i,S_i,\theta_0)-g_x^{(K)}(X_i^*,S_i,\theta_0)\right)\right\||\varepsilon_i|^{K-k}\right]\\&\leqslant\mathbb{E}\left[b_1(X_i^*,S_i,\theta_0)\right]\mathbb{E}\left[|\varepsilon_i|^{K+1-k}\right]+\mathbb{E}\left[b_2(X_i^*,S_i,\theta_0)\right]\mathbb{E}\left[|\varepsilon_i|^{M-k}\right]\to 0,\end{aligned}$$

where the convergence follows from Assumptions MME, A.1 (ii) and A.3 (ii). Hence, we conclude $g_x^{(k)}\to g_x^{(k)*}$.

Next, we show $\overline{g}_x^{(k)}(\hat{\theta})\xrightarrow{p} g_x^{(k)*}$. By the triangle inequality,

$$\left\|\overline{g}_x^{(k)}(\hat{\theta})-g_x^{(k)*}\right\|\leqslant\sup_{\theta\in B_\delta(\theta_0)}\left\|\overline{g}_x^{(k)}(\theta)-g_x^{(k)*}(\theta)\right\|+\left\|g_x^{(k)*}(\hat{\theta})-g_x^{(k)*}(\theta_0)\right\|,$$

where the inequality holds with probability approaching one since $\hat{\theta}\in B_\delta(\theta_0)$ with probability approaching one. Note that, By Lemma D.1, $\sup_{\theta\in B_\delta(\theta_0)}\left\|\overline{g}_x^{(k)}(\theta)-g_x^{(k)*}(\theta)\right\|=o_p(1)$ and $\left\|g_x^{(k)*}(\hat{\theta})-g_x^{(k)*}(\theta_0)\right\|=o_p(1)$, where the second result follows from consistency of $\hat{\theta}$ and continuity of $g_x^{(k)*}(\theta)$. Hence, $\overline{g}_x^{(k)}(\hat{\theta})\xrightarrow{p} g_x^{(k)*}$, which completes the proof of the results for $g_x^{(k)}$ for all $k\in\{0,\dots,K\}$.

A nearly identical argument, can be invoked to establish the same results for $G_x^{(k)}$ for $k\in\{0,\dots,K\}$, with Assumptions A.3 (i) and (iii) taking the places of Assumptions A.1 (i) and A.3 (ii), respectively. Q.E.D.

**Lemma D.3.** *Suppose that the hypotheses of Lemma D.1 are satisfied. Then, under additional Assumptions A.3 (iv) and A.4 (i), we have $\hat{\theta}\xrightarrow{p}\theta_0$, $\hat{\gamma}\xrightarrow{p}0$ and $\hat{\gamma}\xrightarrow{p}\gamma_{0n}$.*

*Proof of Lemma D.3.* First, we argue that $\sup_{\beta\in\mathcal{B}}\left\|\overline{\psi}(\beta)-\psi^*(\beta)\right\|=o_p(1)$. Notice that, by the triangle inequality,

$$\sup_{\beta\in\mathcal{B}}\left\|\overline{\psi}(\beta)-\psi^*(\beta)\right\|\leqslant\sup_{\theta\in\Theta}\|\overline{g}(\theta)-g^*(\theta)\|+\sum_{k=2}^{K}|\gamma_k|\sup_{\theta\in\Theta}\left\|\overline{g}_x^{(k)}(\theta)-g_x^{(k)*}(\theta)\right\|=o_p(1),\tag{D.14}$$

where the equality follows from Lemma D.1 (i) and boundedness of $\gamma$ (Assumption A.2 (i)). Moreover, Lemma D.1 (i) also ensures that $\psi^*(\beta)$ is continuous on compact $\mathcal{B}$ and, consequently, is bounded.

Let $\hat{Q}(\beta)=\overline{\psi}(\beta)'\hat{\Xi}\overline{\psi}(\beta)$ and $Q^*(\beta)=\psi^*(\beta)'\Xi\psi^*(\beta)$. Notice that (D.14), boundedness of $\psi^*(\beta)$, and Assumption A.3 (iv) together guarantee that $\sup_{\beta\in\mathcal{B}}\left|\hat{Q}(\beta)-Q^*(\beta)\right|=o_p(1)$. Next, recall that $\gamma_{0n}\to 0_{K-1}$ (Lemma 1). Since $\Gamma$ is compact and $\gamma_{0n}\in\Gamma$

(Assumption A.2 (i)), $0_{K-1} \in \Gamma$. Consequently, Assumptions A.4 (i) and A.3 (iv) together guarantee that $\hat{Q}^*(\beta)$ is uniquely minimized at $\theta = \theta_0$ and $\gamma = 0_{K-1}$. Consequently, applying the standard consistency argument (e.g., Theorem 2.1 of Newey and McFadden, 1994), we conclude that $\hat{\theta} \to \theta_0$ and $\hat{\gamma} \to 0_{K-1}$. Finally, since $\gamma_{0n} \to 0$ (Lemma 1), we also have $\hat{\gamma} \xrightarrow{p} \gamma_{0n}$. Q.E.D.

**Lemma D.4.** *Suppose that* $\{(X_i^*, S_i', \varepsilon_i)\}_{i=1}^n$ *are i.i.d.. Then, under Assumptions MME, CME, A.1, and A.3* (ii) *and* (v)*, we have*

$$n^{1/2}\overline{\psi}(\beta_{0n}) \xrightarrow{d} N(0, \Omega_{gg}^*),$$

*where* $\Omega_{gg}^* \equiv \mathbb{E}\left[g(X_i, S_i, \theta_0) g(X_i, S_i, \theta_0)'\right]$.

*Proof of Lemma D.4.* Using expansion (D.3), we obtain

$$\begin{aligned}
n^{1/2}\overline{\psi}(\beta_{0n}) =& n^{-1/2}\sum_{i=1}^n g(X_i^*, S_i, \theta_0) + n^{-1/2}\sum_{i=1}^n g_x^{(1)}(X_i, S_i, \theta_0)\varepsilon_i \\
&+ \sum_{k=2}^K n^{-1/2}\sum_{i=1}^n g_x^{(k)}(X_i^*, S_i, \theta_0)\left(\frac{1}{k!}\varepsilon_i^k - \sum_{\ell=2}^k \frac{1}{(k-\ell)!}\varepsilon_i^{k-\ell}\gamma_{0kn}\right) \\
&+ \frac{1}{K!}n^{-1/2}\sum_{i=1}^n \left(g_x^{(K)}(\tilde{X}_i, S_i, \theta_0) - g_x^{(K)}(X_i^*, S_i, \theta_0)\right)\varepsilon_i^K \\
&- \sum_{k=2}^K \frac{\gamma_{0kn}}{(K-k)!}n^{-1/2}\sum_{i=1}^n \left(g_x^{(K)}(\tilde{X}_{ki}, S_i, \theta_0) - g_x^{(K)}(X_i^*, S_i, \theta_0)\right)\varepsilon_i^{K-k}.
\end{aligned} \tag{D.15}$$

First, note that, by the standard CLT, $n^{-1/2}\sum_{i=1}^n g(X_i^*, S_i, \theta_0) \xrightarrow{d} N(0, \Omega_{gg}^*)$. The rest of the proof is to show that the remaining terms are $o_p(1)$. By Assumptions MME, CME, A.3 (ii), Chebyshev's inequality guarantees

$$n^{-1/2}\sum_{i=1}^n g_x^{(1)}(X_i^*, S_i, \theta_{0n})\varepsilon_i = o_p(1)$$

Next, (D.9) ensures that we can similarly apply Chebyshev's inequality (combined with Assumptions MME, CME, A.3 (ii) and (v)) to ensure that for $k \in \{2, \ldots, K\}$

$$n^{-1/2}\sum_{i=1}^n g_x^{(k)}(X_i^*, S_i, \theta_0)\left(\frac{1}{k!}\varepsilon_i^k - \sum_{\ell=2}^k \frac{1}{(k-\ell)!}\varepsilon_i^{k-\ell}\gamma_{0kn}\right) = o_p(1).$$

Next, using (D.12),

$$
\begin{aligned}
&\left\| n^{-1/2} \sum_{i=1}^{n} \left( g_x^{(K)}(\tilde{X}_i, S_i, \theta_0) - g_x^{(K)}(X_i^*, S_i, \theta_0) \right) \varepsilon_i^K \right\| \\
&\leqslant n^{-1/2} \sum_{i=1}^{n} b_1(X_i^*, S_i, \theta_0) \left| \varepsilon_i \right|^{K+1} + n^{-1/2} \sum_{i=1}^{n} b_2(X_i^*, S_i, \theta_0) \left| \varepsilon_i \right|^{M} \\
&\leqslant \underbrace{n^{1/2} \sigma_n^{K+1}}_{\to 0} \underbrace{\left( n^{-1} \sum_{i=1}^{n} b_1(X_i^*, S_i, \theta_0) \left| \varepsilon_i / \sigma_n \right|^{K+1} \right)}_{O_p(1)} + \underbrace{n^{1/2} \sigma_n^{M}}_{\to 0} \underbrace{\left( n^{-1} \sum_{i=1}^{n} b_2(X_i^*, S_i, \theta_0) \left| \varepsilon_i / \sigma_n \right|^{M} \right)}_{O_p(1)} \\
&= o_p(1),
\end{aligned}
$$

where both $n^{1/2}\sigma_n^{K+1}$ and $n^{1/2}\sigma_n^{M}$ converge to zero by Assumption MME, and the terms in the brackets are $O_p(1)$ by Markov's inequality (ensured by Assumptions MME, CME, A.1 (ii) and A.3 (ii)). Recall that in the proof of Lemma 1, we have demonstrated that $\gamma_{0kn} = \sigma_n^k a_{kn}$, where $a_{kn}$ are bounded, for $k \in \{2, \ldots, K\}$. Hence, similarly, we have

$$
\begin{aligned}
&\left\| \gamma_{0kn} n^{-1/2} \sum_{i=1}^{n} \left( g_x^{(K)}(\tilde{X}_i, S_i, \theta_0) - g_x^{(K)}(X_i^*, S_i, \theta_0) \right) \varepsilon_i^{K-k} \right\| \\
&\leqslant a_{kn} \sigma_n^k \left[ n^{-1/2} \sum_{i=1}^{n} b_1(X_i^*, S_i, \theta_0) \left| \varepsilon_i \right|^{K-k+1} + n^{-1/2} \sum_{i=1}^{n} b_2(X_i^*, S_i, \theta_0) \left| \varepsilon_i \right|^{M-k} \right] \\
&\leqslant a_{kn} \left[ \underbrace{n^{1/2} \sigma_n^{K+1}}_{\to 0} \underbrace{\left( n^{-1} \sum_{i=1}^{n} b_1(X_i^*, S_i, \theta_0) \left| \varepsilon_i / \sigma_n \right|^{K-k+1} \right)}_{O_p(1)} + \underbrace{n^{1/2} \sigma_n^{M}}_{\to 0} \underbrace{\left( n^{-1} \sum_{i=1}^{n} b_2(X_i^*, S_i, \theta_0) \left| \varepsilon_i / \sigma_n \right|^{M-k} \right)}_{O_p(1)} \right] \\
&= o_p(1).
\end{aligned}
$$

Hence, we have demonstrated that all the remaining terms in (D.15) are $o_p(1)$, i.e. we have

$$
\begin{aligned}
n^{1/2} \overline{\psi}(\beta_{0n}) =& n^{-1/2} \sum_{i=1}^{n} g(X_i^*, S_i, \theta_0) + o_p(1) \\
&\xrightarrow{d} N(0, \Omega_{gg}^*),
\end{aligned}
$$

which completes the proof. Q.E.D.

### D.2.2 Proof of Theorem 2

Equipped with Lemmas D.1-D.4, we are ready to prove Theorem 2.

*Proof of Theorem 2.* Since (i) $\hat{\theta}$ and $\hat{\gamma}$ are consistent for $\theta_0$ and $\gamma_{0n}$, respectively (Lemma D.3) and (ii) both $\theta_0$ and $\gamma_{0n} \to 0$ (Assumption MME) are bounded away from the boundaries of $\Theta$ and $\Gamma$ respectively (Assumption A.2), the standard GMM FOC is satisfied with probability approaching one, i.e., we have (with probability approaching one)

$$\overline{\Psi}(\hat{\beta})'\hat{\Xi}\overline{\psi}(\hat{\beta}) = 0.$$

Expanding $\overline{\psi}(\hat{\beta})$ around $\overline{\psi}(\beta_{0n})$ gives

$$\overline{\Psi}(\hat{\beta})'\hat{\Xi}\left(\overline{\psi}(\beta_{0n}) + \overline{\Psi}(\tilde{\beta})(\hat{\beta} - \beta_0)\right) = 0, \tag{D.16}$$

where $\tilde{\beta}$ lies between $\beta_{0n}$ and $\hat{\beta}$ (and, consequently, $\tilde{\theta} \xrightarrow{p} \theta_0$ and $\tilde{\gamma} \xrightarrow{p} 0$). Next, we argue that $\overline{\Psi}(\hat{\beta}) = \Psi^* + o_p(1)$. Observe

$$\overline{\Psi}(\hat{\beta}) = \left[\overline{G}(\hat{\theta}) - \sum_{k=2}^{K} \hat{\gamma}_k \overline{G}_x^{(k)}(\hat{\theta}), -\overline{g}_x^{(2)}(\hat{\theta}), \ldots, -\overline{g}_x^{(K)}(\hat{\theta})\right].$$

Since $\hat{\theta} \xrightarrow{p} \theta_0$ (Lemma D.3), we can invoke the result of Lemma D.2 to argue that $\overline{g}_x^{(k)}(\hat{\theta}) \xrightarrow{p} g_x^{(k)*}$ and $\overline{G}_x^{(k)}(\hat{\theta}) \xrightarrow{p} G_x^{(k)*}$ for all $k \in \{0, \ldots, K\}$. This, combined with $\hat{\gamma} \to 0$ (Lemma D.3), ensures that $\overline{\Psi}(\hat{\beta}) = \Psi^* + o_p(1)$ and, analogously, $\overline{\Psi}(\tilde{\beta}) = \Psi^* + o_p(1)$. Coupling these result with Assumption A.3 (iv), we conclude that $\overline{\Psi}(\hat{\beta})'\hat{\Xi}\overline{\Psi}(\tilde{\beta}) \xrightarrow{p} \Psi^{*\prime}\Xi\Psi^*$, which is invertible by Assumption A.4 (ii). Hence, (D.16) can be rearranged as (with probability approaching one)

$$\begin{aligned} n^{1/2}(\hat{\beta} - \beta_{0n}) &= -\left(\overline{\Psi}(\hat{\beta})'\hat{\Xi}\overline{\Psi}(\tilde{\beta})\right)^{-1}\overline{\Psi}(\hat{\beta})'\hat{\Xi}n^{1/2}\overline{\psi}(\beta_{0n}) \\ &= -\left(\Psi^{*\prime}\Xi\Psi^*\right)^{-1}\Psi^{*\prime}\Xi n^{1/2}\overline{\psi}(\beta_{0n}) + o_p(1), \end{aligned}$$

where, by Lemma D.4, $n^{1/2}\overline{\psi}(\beta_{0n}) \xrightarrow{d} N(0, \Omega_{gg}^*)$. Hence, we conclude

$$n^{1/2}(\hat{\beta} - \beta_{0n}) \xrightarrow{d} N(0, \Sigma^*),$$

where

$$\Sigma^* = \left(\Psi^{*\prime}\Xi\Psi^*\right)^{-1}\Psi^{*\prime}\Xi\Omega^*_{gg}\Psi^*\Xi\left(\Psi^{*\prime}\Xi\Psi^*\right)^{-1}.$$

To complete the proof, we need to show that $\Sigma \to \Sigma^*$. First, note that, by Lemma D.2 and $\gamma_{0n} \to 0$ (Assumption MME)

$$\Psi = \left[G - \sum_{k=2}^{K}\gamma_{0kn}G_x^{(k)}, -g_x^{(2)}, \ldots, -g_x^{(K)}\right] \to \left[G^*, -g_x^{(2)*}, \ldots, -g_x^{(K)*}\right] = \Psi^*.$$

Next, we want to argue that $\Omega_{\psi\psi} \to \Omega^*_{gg}$. Observe that

$$\Omega_{\psi\psi} = \mathbb{E}\left[\left(g_i - \sum_{k=2}^{K}\gamma_{0kn}g_{xi}^{(k)}\right)\left(g_i - \sum_{k=2}^{K}\gamma_{0kn}g_{xi}^{(k)}\right)'\right] = \mathbb{E}\left[g_i g_i'\right] + o(1),$$

where the equality follows since (i) $\gamma_{0kn} \to 0$ for all $k \in \{2, \ldots, K\}$ (Assumption MME) and (ii) $\mathbb{E}\left[g_{xi}^{(k)}\left(g_{xi}^{(k')}\right)'\right]$ is bounded for all $k, k' \in \{0, \ldots, K\}$. In particular, (ii) can be inspected by expanding $g_x^{(k)}(X_i, S_i, \theta_0)$ and $g_x^{(k')}(X_i, S_i, \theta_0)$ around $X_i^*$ as in (D.2) and bounding the expectations as in the proof of Lemma D.2 (using Assumptions MME, CME, A.1 (i), A.3 (ii), and A.3 (v)). Similarly, by expanding $g(X_i, S_i, \theta_0)$ around $X_i^*$ and bounding the residual terms as in the proof of Lemma D.2 (again, using Assumptions MME, CME, A.1 (i), A.3 (ii), and A.3 (v)), we verify that $\mathbb{E}\left[g_i g_i'\right] \to \mathbb{E}\left[g_i^* g_i^{*\prime}\right] = \Omega^*_{gg}$. Hence, $\Omega_{\psi\psi} \to \Omega^*_{gg}$ and, consequently, we verified that $\Sigma \to \Sigma^*$. Finally, we conclude

$$n^{1/2}\Sigma^{-1/2}(\hat{\beta} - \beta_{0n}) \to N(0, I_{p+K-1}),$$

which completes the proof. Q.E.D.

# E Proofs of Results in Section 4

## E.1 Proof of Theorem 3:

**1.** First, in parts 1-4 below, we prove the theorem in the case $p = 3$. Then, in part 5 of this proof, we consider the case $p = 4$. The proof in the latter case is identical, except some of the remainder terms are of smaller orders. In parts 1-4 of the proof, it will be convenient to state the resulting bounds that depend on $p$ in the general form using $p$,

but to avoid confusion, until reaching part 5 of the proof the reader should only consider the case $p = 3$.

Let us obtain some preliminary results about $f_{X|Z}(x|z)$ and $f_{\varepsilon|XZ}(\varepsilon|x,z)$. Using Assumption CME, we obtain

$$\begin{aligned} f_{\varepsilon X|Z}(\varepsilon, x|z) &= f_{\varepsilon}(\varepsilon) f_{X^*|Z}(x-\varepsilon|z), \\ f_{\varepsilon|XZ}(\varepsilon|x,z) &= \frac{f_{\varepsilon X|Z}(\varepsilon, x|z)}{f_{X|Z}(x|z)} = \frac{f_{\varepsilon}(\varepsilon) f_{X^*|Z}(x-\varepsilon|z)}{f_{X|Z}(x|z)}, \\ f_{X|Z}(x|z) &= \int f_{\varepsilon}(\varepsilon) f_{X^*|Z}(x-\varepsilon|z)\, d\varepsilon. \end{aligned}$$

Since $f_{X^*|Z}(x|z)$ has 3 bounded continuous derivatives in $x$, $f_{X|Z}(x|z)$ also has 3 bounded continuous derivatives in $x$. Moreover,

$$f_{X|Z}(x|z) = \int \left[ f_{X^*|Z}(x|z) - \varepsilon f'_{X^*|Z}(x|z) + \varepsilon^2 \tfrac{1}{2} f''_{X^*|Z}(x|z) \right] f_{\varepsilon}(\varepsilon)\, d\varepsilon + R_{f|Z}(x|z), \quad \text{(E.1)}$$

where $R_{f|Z} \equiv -(1/6) E\left[ \varepsilon_i^3 f'''_{X^*|Z}(x - \widetilde{\varepsilon}(\varepsilon_i)) | Z_i = z \right]$, $\widetilde{\varepsilon}(\varepsilon_i)$ is a point between 0 and $\varepsilon_i$, and

$$\left| R_{f|Z}(x|z) \right| \leq E\left[ |\varepsilon_i|^3 \left| f'''_{X^*|Z}(x - \widetilde{\varepsilon}(\varepsilon_i)) \right| \middle| Z_i = z \right] = O\left( E\left[ |\varepsilon_i|^3 \right] \right) = O\left( \sigma^3 \right).$$

Since $E[\varepsilon_i] = 0$ and $E[\varepsilon_i^2] = \sigma^2$, we have

$$\begin{aligned} f_{X|Z}(x|z) &= f_{X^*|Z}(x|z) - E[\varepsilon_i] f'_{X^*|Z}(x|z) + E\left[\varepsilon_i^2\right] \tfrac{1}{2} f''_{X^*|Z}(x|z) + O(\sigma^p) \\ &= f_{X^*|Z}(x|z) + \left(\sigma^2/2\right) f''_{X^*|Z}(x|z) + O(\sigma^p). \end{aligned} \quad \text{(E.2)}$$

Since $f_{X^*|Z}(x|z) > C$ for some $C > 0$, and $f''_{X^*|Z}(x|z)$ is bounded, $f_{X|Z}(x|z) > C/2$ for small enough $\sigma^2$. Thus,

$$f_{\varepsilon|XZ}(\varepsilon|x,z) = \frac{f_{\varepsilon}(\varepsilon) f_{X^*|Z}(x-\varepsilon|z)}{f_{X^*|Z}(x|z) + (\sigma^2/2) f''_{X^*|Z}(x|z) + O(\sigma^p)}. \quad \text{(E.3)}$$

Similarly,

$$f'_{X|Z}(x|z) = \int f_{\varepsilon}(\varepsilon) f'_{X^*|Z}(x-\varepsilon|z)\, d\varepsilon = f'_{X^*|Z}(x|z) + O\left(\sigma^2\right), \quad \text{(E.4)}$$

$$f''_{X|Z}(x|z) = \int f_{\varepsilon}(\varepsilon) f''_{X^*|Z}(x-\varepsilon|z)\, d\varepsilon = f''_{X^*|Z}(x|z) + O\left(\sigma^{p-2}\right). \quad \text{(E.5)}$$

**2.** Consider any function $a(x)$ that has 3 bounded continuous derivatives. Since $X_i^* = X_i - \varepsilon_i$,

$$\begin{aligned} & E\left[a\left(X_i^*\right)|X_i = x, Z_i = z\right] \\ = \; & E\left[a\left(x - \varepsilon_i\right)|X_i = x, Z_i = z\right] \\ = \; & a(x) - a'(x) E\left[\varepsilon_i|X_i = x, Z_i = z\right] + \tfrac{1}{2}a''(x) E\left[\varepsilon_i^2|X_i = x, Z_i = z\right] + R_{a|XZ} \end{aligned} \tag{E.6}$$

where $R_{a|XZ} \equiv -(1/6) E\left[\varepsilon_i^3 a'''\left(x - \widetilde{\varepsilon}(\varepsilon_i)\right)|X_i = x, Z_i = z\right]$.

We now consider $E\left[\varepsilon_i^\ell|X_i = x, Z_i = z\right]$ for $\ell \in \{1, 2\}$:

$$\begin{aligned} E\left[\varepsilon_i|X_i = x, Z_i = z\right] &= \int \varepsilon f_{\varepsilon|XZ}(\varepsilon|x, z)\, d\varepsilon = \int \varepsilon \frac{f_\varepsilon(\varepsilon) f_{X^*|Z}(x - \varepsilon|z)}{f_{X|Z}(x|z)} d\varepsilon \\ &= \frac{\int \varepsilon \left\{ f_{X^*|Z}(x|z) - \varepsilon f'_{X^*|Z}(x|z) + \varepsilon^2 \frac{1}{2} f''_{X^*|Z}\left(x - \widetilde{\varepsilon}(\varepsilon)|z\right) \right\} f_\varepsilon(\varepsilon)\, d\varepsilon}{f_{X^*|Z}(x|z) + O(\sigma^2)} \\ &= 0 - \sigma^2 \frac{f'_{X^*|Z}(x|z)}{f_{X^*|Z}(x|z) + O(\sigma^2)} + O(\sigma^p) = -\sigma^2 s_{X^*|Z}(x|z) + O(\sigma^p) \end{aligned}$$

$$\begin{aligned} E\left[\varepsilon_i^2|X_i = x, Z_i = z\right] &= \frac{\int \varepsilon^2 \left\{ f_{X^*|Z}(x|z) - \varepsilon f'_{X^*|Z}\left(x - \widetilde{\varepsilon}(\varepsilon)|z\right) \right\} f_\varepsilon(\varepsilon)\, d\varepsilon}{f_{X^*|Z}(x|z) + O(\sigma^2)} \\ &= \sigma^2 + O(\sigma^p). \end{aligned}$$

$$E\left[|\varepsilon_i^p|\,|X_i = x, Z_i = z\right] = \frac{\int |\varepsilon^p| f_{X^*|Z}\left(x - \widetilde{\varepsilon}(\varepsilon)|z\right) f_\varepsilon(\varepsilon)\, d\varepsilon}{f_{X^*|Z}(x|z) + O(\sigma^2)} = O(\sigma^p) \tag{E.7}$$

and hence $\left|R_{a|XZ}\right| \leq E\left[|\varepsilon_i^3|\,|a'''\left(x - \widetilde{\varepsilon}(\varepsilon_i)\right)|\,|X_i = x, Z_i = z\right] = O(\sigma^3)$.

Combining these with equation (E.6), we obtain

$$E\left[a\left(X_i^*\right)|X_i = x, Z_i = z\right] = a(x) + \sigma^2 a'(x) s_{X^*|Z}(x|z) + \left(\sigma^2/2\right) a''(x) + O(\sigma^p). \tag{E.8}$$

**3.** Next, consider $\nabla_x^\ell E\left[a\left(X_i^*\right)|X_i = x\right]$ for $\ell \in \{1, 2\}$. We have

$$E\left[a\left(X_i^*\right)|X_i = x\right] = \frac{1}{f_X(x)} \int a(x - \varepsilon) f_{X^*}(x - \varepsilon) f_\varepsilon(\varepsilon)\, d\varepsilon.$$

Let $\varphi(x)$ and $\eta(x)$ be any functions, possibly changing with $\sigma$, with 3 bounded

continuous derivatives. Suppose $\eta(x) \geq C$ for some $C > 0$ for all small enough $\sigma$, and let $\zeta(x) \equiv \frac{1}{\eta(x)} \int \varphi(x-\varepsilon) f_\varepsilon(\varepsilon) d\varepsilon$. Then,

$$\begin{aligned}\zeta'(x) &= \frac{1}{\eta(x)} \int \varphi'(x-\varepsilon) f_\varepsilon(\varepsilon) d\varepsilon - \frac{\eta'(x)}{(\eta(x))^2} \int \varphi(x-\varepsilon) f_\varepsilon(\varepsilon) d\varepsilon \\ &= \frac{\varphi'(x) + O(\sigma^2)}{\eta(x)} - \frac{\eta'(x)(\varphi(x) + O(\sigma^2))}{(\eta(x))^2} = \frac{\varphi'(x)}{\eta(x)} - \frac{\eta'(x)\varphi(x)}{(\eta(x))^2} + O(\sigma^2).\end{aligned}$$

Taking $\varphi(t) \equiv a(t) f_{X^*}(t)$ and $\eta(x) \equiv f_X(x)$, and using equations (E.2) and (E.4), we obtain

$$\nabla_x E[a(X_i^*)|X_i = x] = a'(x) + O(\sigma^2).$$

Similarly,

$$\zeta''(x) = \frac{\varphi''(x)}{\eta(x)} - 2\frac{\eta'(x)\varphi'(x)}{(\eta(x))^2} - \left(\frac{\eta''(x)}{(\eta(x))^2} - \frac{2(\eta'(x))^2}{(\eta(x))^3}\right)\varphi(x) + O(\sigma^{p-2}),$$

and using identical substitutions, and equations (E.2), (E.4), and (E.5), we obtain

$$\nabla_x^2 E[a(X_i^*)|X_i = x] = a''(x) + O(\sigma^{p-2}).$$

**4.** Consider

$$q(x,z) = E[\rho(X_i^*) + U_i | X_i = x, Z_i = z] = E[\rho(X_i^*)|X_i = x, Z_i = z].$$

Part 2 of the proof shows that

$$q(x,z) = \rho(x) + \sigma^2 \rho'(x) s_{X^*|Z}(x|z) + \tfrac{1}{2}\sigma^2 \rho''(x) + O(\sigma^p). \tag{E.9}$$

Therefore,

$$q(x,z_1) - q(x,z_2) = \sigma^2 \rho'(x)\left[s_{X^*|Z}(x|z_1) - s_{X^*|Z}(x|z_2)\right] + O(\sigma^p). \tag{E.10}$$

Part 1 implies that $s_{X^*|Z}(x|z) = s_{X|Z}(x|z) + O(\sigma^2)$. Let $q(x) \equiv E[Y_i|X_i = x]$. Applying part 3 with $a(x) = q(x)$ we obtain $q'(x) = \rho'(x) + O(\sigma^2)$. Substituting these into equation (E.10), we obtain

$$q(x,z_1) - q(x,z_2) = \sigma^2 q'(x)\left[s_{X|Z}(x|z_1) - s_{X|Z}(x|z_2)\right] + O(\sigma^p), \tag{E.11}$$

and hence, for any $x$ with $q'(x)\left[s_{X|Z}(x|z_1) - s_{X|Z}(x|z_2)\right] \neq 0$,

$$\widetilde{\sigma}^2(x) = \sigma^2 + O(\sigma^p), \qquad \text{where } \widetilde{\sigma}^2(x) \equiv \frac{q(x,z_1) - q(x,z_2)}{q'(x)\left[s_{X|Z}(x|z_1) - s_{X|Z}(x|z_2)\right]},$$

which identifies $\sigma^2$ up to $O(\sigma^p)$.

Next, from part 3 we also have $q''(x) = \rho''(x) + O(\sigma^{p-2})$. Thus, equation (E.9) implies that

$$q(x,z) = \rho(x) + \sigma^2 q'(x)\, s_{X|Z}(x,z) + \tfrac{1}{2}\sigma^2 q''(x) + O(\sigma^p),$$

and hence we obtain

$$\widetilde{\rho}(x,z_1) = \rho(x) + O(\sigma^p), \qquad \text{where } \widetilde{\rho}(x,z) \equiv q(x,z) - \widetilde{\sigma}^2(x)\left[q'(x)\, s_{X|Z}(x|z) + \tfrac{1}{2}q''(x)\right],$$

which identifies $\rho(x)$ up to $O(\sigma^p)$. This completes the proof for the case $p = 3$.

Note that $\widetilde{\rho}(x,z_1) - \widetilde{\rho}(x,z_2) = q(x,z_1) - q(x,z_2) - \widetilde{\sigma}^2(x)\left[q'(x)\left(s_{X|Z}(x|z_1) - s_{X|Z}(x|z_2)\right)\right] = 0$, i.e., $\widetilde{\rho}(x,z_1) = \widetilde{\rho}(x,z_2)$.

**5.** When $p = 4$ and $E[\varepsilon_i^3] = 0$, the above Taylor expansions can be extended to the next order, providing the corresponding improvements in the remainder terms.

For example, in part 1 of the proof, expansion (E.1) becomes

$$f_{X|Z}(x|z) = \int \left[f_{X^*|Z}(x|z) - \varepsilon f'_{X^*|Z}(x|z) + \varepsilon^2 \tfrac{1}{2} f''_{X^*|Z}(x|z) - \varepsilon^3 \tfrac{1}{6} f'''_{X^*|Z}(x|z)\right] f_\varepsilon(\varepsilon)\, d\varepsilon + R_{f|Z}(x|z),$$

where $R_{f|Z}(x|z) \equiv (1/24)\, E\left[\varepsilon_i^4 f''''_{X^*|Z}(x - \widetilde{\varepsilon}(\varepsilon_i))\,|Z_i = z\right]$, so $\left|R_{f|Z}(x|z)\right| = O(\sigma^4)$. Combining the expansion above with $E[\varepsilon_i^3] = 0$, we verify that (E.2) also holds with $p = 4$.

In addition to the calculations in part 2 of the proof, we also use

$$\begin{aligned} E\left[\varepsilon_i^3 | X_i = x, Z_i = z\right] &= \frac{\int \varepsilon^3 \left\{f_{X^*|Z}(x|z) - \varepsilon f'_{X^*|Z}(x - \widetilde{\varepsilon}(\varepsilon)\,|z)\right\} f_\varepsilon(\varepsilon)\, d\varepsilon}{f_{X^*|Z}(x|z) + O(\sigma^2)} \\ &= E\left[\varepsilon_i^3\right] + O\left(\sigma^4\right) = O\left(\sigma^4\right). \end{aligned}$$

In parts 2 and 3 of the proof, we also require functions $a(x)$, $\varphi(x)$, and $\eta(x)$ to have 4 continuous derivatives.

Then, the previous steps of the proof and the conclusions of the theorem hold with $p = 4$. ■

## E.2 Proof of Proposition 4

The proof of Theorem 3 shows that $q(x,z)$ and $f_{X|Z}(x|z)$ have $m$ bounded continuous derivatives. Construct $\widehat{\rho}^{\text{MER}}(x)$ using equations (21)-(22) nonparametrically estimating $q(x,z)$, $q(x)$, $f_{X|Z}(x|z)$, and their derivatives, e.g., using standard kernel or sieve estimators. If the tuning parameters are chosen optimally, under the usual regularity conditions, the rates of convergence of these estimators are $\widehat{q}(x,z)-q(x,z)=O_p\left(n^{-\frac{m}{2m+1}}\right)$, $\widehat{q}(x)-q(x)=O_p\left(n^{-\frac{m}{2m+1}}\right)$, $\widehat{q}'(x)-q'(x)=O_p\left(n^{-\frac{m-1}{2m-1}}\right)$, $\widehat{q}''(x)-q''(x)=O_p\left(n^{-\frac{m-2}{2m-3}}\right)$, and $\widehat{s}_{X|Z}(x|z)-s_{X|Z}(x|z)=O_p\left(n^{-\frac{m-1}{2m-1}}\right)$ for $x\in S_{X^*}(z)$, where $\widehat{s}_{X|Z}(x|z)\equiv \widehat{f}'_{X|Z}(x|z)\Big/\widehat{f}_{X|Z}(x|z)$. Note also that by equation (E.9), $q(x,z_1)-q(x,z_2)=O(\tau_n^2)$.

Then, since for the analog estimator $\widehat{\sigma}^2(x)$ of $\widetilde{\sigma}^2(x)$ we have

$$\begin{aligned}\widehat{\sigma}^2(x)&=\widetilde{\sigma}^2(x)+O_p\left(n^{-\frac{m}{2m+1}}+\tau_n^2 n^{-\frac{m-1}{2m-1}}\right)=\widetilde{\sigma}^2(x)+O_p\left(n^{-\frac{m}{2m+1}}\right)\\&=\sigma^2+O_p\left(\tau_n^4+n^{-\frac{m}{2m+1}}\right)=\sigma^2+O_p\left(n^{-\frac{m}{2m+1}}\right),\end{aligned}\tag{E.12}$$

where the first equality follows from equation (21), using $\widehat{a}/\widehat{b}-a/b=(\widehat{a}-a)/\widehat{b}+a(1/\widehat{b}-1/b)$, equation (E.11), and the rates of convergence listed above, the second equality holds because $\frac{m}{2m+1}<\frac{1}{2}\frac{m}{2m+1}+\frac{m-1}{2m-1}$ for $m\geqslant 2$, and the third equality holds by equation (23) in Theorem 3.

Next, consider the analog estimator $\widehat{\rho}^{\text{MER}}(x)$ of $\rho(x)$ based on equation (22),

$$\begin{aligned}\widehat{\rho}^{\text{MER}}(x)&=\widehat{q}(x,z)-\widehat{\sigma}^2(x)\left[\widehat{q}'(x)\,\widehat{s}(x,z)+\tfrac{1}{2}\widehat{q}''(x)\right]\\&=\widetilde{\rho}(x)+O_p\left(n^{-\frac{m}{2m+1}}+\tau_n^2\left(n^{-\frac{m-1}{2m-1}}+n^{-\frac{m-2}{2m-3}}\right)\right)\\&=\rho(x)+O_p\left(n^{-\frac{m}{2m+1}}+\tau_n^2 n^{-\frac{m-2}{2m-3}}+\tau_n^4\right)\\&=\rho(x)+O_p\left(n^{-\frac{m}{2m+1}}\right),\end{aligned}$$

where the first equality is the definition of the analog estimator, the second equality follows from the rates of convergence listed above and equation (E.12), the third equality holds by equation (24) in Theorem 3, and the fourth equality holds because $\frac{m}{2(2m+1)}+\frac{m-2}{2m-3}>\frac{m}{2m+1}$ for $m\geq 3$.

The naive estimator is

$$\widehat{\rho}^{\text{Naive}}(x)=\widehat{q}(x)=\rho(x)+O_p\left(\tau_n^2+n^{-\frac{m}{2m+1}}\right)=\rho(x)+O_p\left(n^{-\frac{1}{2}\frac{m}{2m+1}}\right),$$

where the second equality follows from equation (E.9). ■

# F Implementation Details of the Empirical Illustration

In this section, we provide additional details on the implementation of the numerical experiment in Section 3.3.

**Data.** The original dataset is the ModeCanada dataset supplied with the R package `mlogit`. This dataset has been extensively used in transportation research. For a detailed description of the dataset see, for example, Koppelman and Wen (2000), Wen and Koppelman (2001), and Hansen (2022). As in Koppelman and Wen (2000), we use only the subset of travelers who chose train, air, or car (and had all of those alternatives available for them), which leaves $n = 2769$ observations.

**Monte-Carlo design.** We choose $\theta_0$ to be the MLE estimates using the considered dataset, which are reported in the table bellow.

| | $\theta_1$ | $\theta_2$ | $\theta_3$ | $\theta_4$ | $\theta_5$ | $\theta_6$ | $\theta_7$ | $\theta_8$ |
|---|---|---|---|---|---|---|---|---|
| Estimates | 0.0355 | 0.2976 | -2.0891 | 0.0079 | -0.9900 | 1.8794 | -0.0223 | -0.0149 |
| Std. Err. | 0.0036 | 0.0844 | 0.4674 | 0.0036 | 0.0876 | 0.2037 | 0.0038 | 0.0008 |

To generate the simulated samples, we randomly draw the covariates (with replacement) from their joint empirical distribution. To ensure identification of the model, we also generate an instrumental variable $Z_i$ as

$$Z_i = \kappa \, Income_i^* / \sigma_{Income^*} + \sqrt{1 - \kappa^2} \zeta_i,$$

where $\sigma_{Income^*} \approx 17.5$ is the standard deviation of $Income^*$, $\kappa = 0.5$, and $\zeta_i$ are i.i.d. draws from $N(0, 1)$ (which are also independent from all the over variables). Note that the instrument $Z_i$ is "caused by $X_i^*$". For example, $Z_i$ can be some (noisy) measure of individual consumption.

**Moments.** To simplify the notation, let $X_i^* \equiv Income_i^*$, $X_i \equiv Income_i$, $R_i \equiv Urban_i$, $R_{ij} \equiv (Price_{ij}, InTime_{ij})'$ for $j \in \{0, 1, 2\}$, and $W_i \equiv (R_i, R_{i1}', R_{i2}', R_{i0}')'$. Also let $Y_{ij} \equiv$

$\mathbb{1}\{j = \text{argmax}_{j' \in \{0,1,2\}} U_{ij'}\}$ for $j \in \{0,1,2\}$, $Y_i \equiv (Y_{i1}, Y_{i2}, Y_{i0})'$, and $p_j(x,w,\theta) \equiv \mathbb{P}(Y_{ij} = 1 | X_i^* = x, W_i = w; \theta)$ with $w \equiv (r, r_1', r_2', r_0')$, so

$$p_1(x,w,\theta) = \frac{e^{\theta_1 x + \theta_2 r + \theta_3 + (\theta_7,\theta_8) r_1}}{e^{\theta_1 x + \theta_2 r + \theta_3 + (\theta_7,\theta_8) r_1} + e^{\theta_4 x + \theta_5 r + \theta_6 + (\theta_7,\theta_8) r_2} + e^{(\theta_7,\theta_8) r_0}},$$
$$p_2(x,w,\theta) = \frac{e^{\theta_4 x + \theta_5 r + \theta_6 + (\theta_7,\theta_8) r_2}}{e^{\theta_1 x + \theta_2 r + \theta_3 + (\theta_7,\theta_8) r_1} + e^{\theta_4 x + \theta_5 r + \theta_6 + (\theta_7,\theta_8) r_2} + e^{(\theta_7,\theta_8) r_0}},$$

and $p_0(x,w,\theta) = 1 - p_1(x,w,\theta) - p_2(x,w,\theta)$. Then, the original moment function takes the form of

$$g(x,w,y,z,\theta) = \left((y_1 - p_1(x,w,\theta))\, h_1(x,z,w)', (y_2 - p_2(x,w,\theta))\, h_2(x,z,w)'\right)'.$$

and $h_j(x,z,w) = (1, x, z, x^2, z^2, x^3, z^3, r, (r_j - r_0)')'$ for $K = 2$ and $h_j(x,z,w) = (1, x, z, x^2, xz, z^2, x^3, x^2 z, x z^2, z^3, r, (r_j - r_0)')'$ for $K = 4$.

**Income Elasticities.** In Section 3.3, we focus on estimation of and inference on the income elasticities

$$\frac{\partial \ln p_j}{\partial \ln x}(x,w,\theta) = \frac{x}{p_j(x,w,\theta)} \frac{\partial p_j(x,w,\theta)}{\partial x}.$$

We report the results are for the income elasticities evaluated at the sample mean of $X^*$ and $W$ in the original sample.

**Estimation of and Inference on the $\theta_0$.** In Table 8 below, we also report the estimation and inference results for the vector of parameters $\theta_0$ underlying the reported results about elasticities.

Table 8: Simulation results for the empirically calibrated conditional logit model

| | MLE | | | | $K=2$ | | | | $K=4$ | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | bias | std | rmse | size | bias | std | rmse | size | bias | std | rmse | size |
| | | | | | | $\tau = 1/4$ | | | | | | |
| $\theta_1$ | -0.0021 | 0.0035 | 0.0041 | 8.70 | 0.0001 | 0.0042 | 0.0042 | 5.48 | 0.0005 | 0.0057 | 0.0058 | 7.38 |
| $\theta_2$ | 0.0047 | 0.0932 | 0.0933 | 5.10 | 0.0028 | 0.0957 | 0.0957 | 5.36 | 0.0022 | 0.0960 | 0.0960 | 5.40 |
| $\theta_3$ | 0.1152 | 0.4452 | 0.4599 | 6.00 | -0.0048 | 0.4821 | 0.4821 | 5.94 | -0.0251 | 0.5336 | 0.5342 | 6.68 |
| $\theta_4$ | -0.0004 | 0.0031 | 0.0031 | 4.52 | -0.0001 | 0.0034 | 0.0034 | 5.32 | -0.0001 | 0.0036 | 0.0036 | 6.86 |
| $\theta_5$ | -0.0023 | 0.0894 | 0.0895 | 5.18 | -0.0088 | 0.0918 | 0.0922 | 5.54 | -0.0113 | 0.0922 | 0.0929 | 5.96 |
| $\theta_6$ | 0.0232 | 0.1821 | 0.1836 | 4.64 | 0.0250 | 0.1982 | 0.1998 | 5.72 | 0.0329 | 0.2089 | 0.2115 | 6.74 |
| $\theta_7$ | -0.0001 | 0.0035 | 0.0035 | 5.58 | -0.0002 | 0.0036 | 0.0036 | 6.24 | -0.0003 | 0.0036 | 0.0037 | 6.04 |
| $\theta_8$ | -0.0001 | 0.0007 | 0.0007 | 4.82 | -0.0001 | 0.0007 | 0.0007 | 5.48 | -0.0002 | 0.0007 | 0.0007 | 5.58 |
| | | | | | | $\tau = 1/2$ | | | | | | |
| $\theta_1$ | -0.0073 | 0.0032 | 0.0080 | 60.08 | -0.0016 | 0.0043 | 0.0046 | 6.86 | 0.0005 | 0.0061 | 0.0061 | 6.60 |
| $\theta_2$ | 0.0109 | 0.0930 | 0.0936 | 5.18 | 0.0050 | 0.0959 | 0.0960 | 5.54 | 0.0026 | 0.0964 | 0.0965 | 5.36 |
| $\theta_3$ | 0.4080 | 0.4452 | 0.6039 | 17.12 | 0.0936 | 0.4874 | 0.4963 | 6.58 | -0.0263 | 0.5475 | 0.5481 | 6.38 |
| $\theta_4$ | -0.0012 | 0.0029 | 0.0031 | 6.46 | -0.0003 | 0.0035 | 0.0035 | 5.22 | -0.0002 | 0.0038 | 0.0038 | 6.52 |
| $\theta_5$ | -0.0006 | 0.0894 | 0.0894 | 5.16 | -0.0083 | 0.0919 | 0.0923 | 5.52 | -0.0110 | 0.0924 | 0.0930 | 5.92 |
| $\theta_6$ | 0.0655 | 0.1752 | 0.1870 | 6.22 | 0.0348 | 0.2035 | 0.2064 | 5.86 | 0.0326 | 0.2158 | 0.2183 | 6.42 |
| $\theta_7$ | -0.0003 | 0.0035 | 0.0036 | 5.64 | -0.0003 | 0.0036 | 0.0037 | 6.34 | -0.0003 | 0.0037 | 0.0037 | 6.06 |
| $\theta_8$ | -0.0001 | 0.0007 | 0.0007 | 5.06 | -0.0001 | 0.0007 | 0.0007 | 5.54 | -0.0002 | 0.0007 | 0.0007 | 5.48 |
| | | | | | | $\tau = 3/4$ | | | | | | |
| $\theta_1$ | -0.0132 | 0.0029 | 0.0135 | 99.34 | -0.0056 | 0.0043 | 0.0071 | 25.12 | 0.0003 | 0.0065 | 0.0065 | 6.06 |
| $\theta_2$ | 0.0180 | 0.0923 | 0.0940 | 5.36 | 0.0102 | 0.0961 | 0.0966 | 5.76 | 0.0033 | 0.0973 | 0.0973 | 5.44 |
| $\theta_3$ | 0.7336 | 0.4496 | 0.8604 | 41.66 | 0.3203 | 0.4859 | 0.5820 | 12.00 | -0.0130 | 0.5666 | 0.5667 | 6.20 |
| $\theta_4$ | -0.0024 | 0.0026 | 0.0035 | 14.00 | -0.0009 | 0.0035 | 0.0036 | 5.94 | -0.0002 | 0.0041 | 0.0041 | 5.76 |
| $\theta_5$ | 0.0021 | 0.0890 | 0.0891 | 5.08 | -0.0071 | 0.0921 | 0.0924 | 5.68 | -0.0109 | 0.0926 | 0.0932 | 5.82 |
| $\theta_6$ | 0.1204 | 0.1654 | 0.2046 | 9.76 | 0.0648 | 0.2048 | 0.2148 | 6.56 | 0.0334 | 0.2294 | 0.2318 | 5.98 |
| $\theta_7$ | -0.0004 | 0.0036 | 0.0036 | 6.00 | -0.0004 | 0.0036 | 0.0037 | 6.34 | -0.0003 | 0.0037 | 0.0037 | 6.06 |
| $\theta_8$ | -0.0001 | 0.0007 | 0.0007 | 5.54 | -0.0002 | 0.0007 | 0.0008 | 6.00 | -0.0002 | 0.0007 | 0.0008 | 5.42 |

This table reports the simulated finite sample bias, standard deviation, RMSE, and size of the MLE and the MERM estimators and the corresponding t-tests for the components of $\theta_0$. The true value of the parameters of interest are $\theta_0 = (0.0355, 0.2976, -2.0891, 0.0079, -0.9900, 1.8794, -0.0223, -0.0149)'$. The results are based on 5,000 replications.