

# Over-Identified Doubly Robust Identification and Estimation

Arthur Lewbel, Jin-Young Choi, and Zhuzhu Zhou  
Boston College, Xiamen University, and Xiamen University

Original 2018, Revised January 2022

Abstract

# 1 Introduction

Consider two different parametric models, which we will call  $G$  and  $H$ . One of these models is correctly specified, but we don't know which one (or both could be right). Both models include the same parameter vector  $\theta$ . An estimator  $\hat{\theta}$  is called Doubly Robust (DR) if  $\hat{\theta}$  is consistent no matter which model is correct. The term double robustness was coined by Robins, Rotnitzky, and van der Laan (2000), but is based on Scharfstein, Rotnitzky, and Robins (1999) and the augmented inverse probability weighting average treatment effect estimator introduced by Robins, Rotnitzky, and Zhao (1994). In their application  $\theta$  is a population

in the ATE application). However, we do not advise using our ODR for applications where DR methods

construct weights to yield the DR consistency property and for relative efficiency.

Analogous to  $g$ , let  $h$  denote the estimator of  $\theta$  based on the moments  $E[H(Z; \theta; \theta_0)] = 0$ , so  $h$  and  $g$  minimize a quadratic GMM objective function  $Q^h(\beta; \theta)$ , and are asymptotically efficient if model  $H$  is true and model  $G$  is not true. Finally, let  $f = (f_1, \dots, f_k)$ . If, let

If  $G$  is correctly specified,  $sg_0(\theta_0; \theta_0) = 0$ , then there does not exist any  $f; g^2$  with  $f; g \neq 0$

model  $H$ . In our applications, we likewise use the standard efficient two step GMM method for estimating the matrices  $\hat{g}$  and  $\hat{h}$ .

Define  $Q_0^g(\cdot; \cdot)$  and  $Q_0^h(\cdot; \cdot)$  by

$$Q_0^g(\cdot; \cdot) = g_0(\cdot; \cdot)' g g_0(\cdot; \cdot) \text{ and } Q_0^h(\cdot; \cdot) = h_0(\cdot; \cdot)' h h_0(\cdot; \cdot)$$

for given positive definite matrices  $g$  and  $h$ , where  $\hat{g} \xrightarrow{P} g$  and  $\hat{h} \xrightarrow{P} h$ .

Assumption A4: Assume there exists  $f_g(\cdot; \cdot); g(\cdot; \cdot)g^2$  such that  $Q_0^g(g(\cdot; \cdot); g(\cdot; \cdot)) < Q_0^g(\cdot; \cdot)$  for all  $f; g^2$   $\inf f_g(\cdot; \cdot); g(\cdot; \cdot)g$  and there exists  $f_h(\cdot; \cdot); h(\cdot; \cdot)h^2$  such that  $Q_0^h(h(\cdot; \cdot); h(\cdot; \cdot)) < Q_0^h(\cdot; \cdot)$  for all  $f; g^2$   $\inf f_h(\cdot; \cdot); h(\cdot; \cdot)h$ .

Assumption A4 says that, for each of the models  $G$  and  $H$ , there exists a unique value of the parameters that minimizes the limiting value of the GMM objective function. Given Assumptions A2 and A3, Assumption A4 will automatically be satisfied for model  $G$  when  $G$  is correctly specified, with  $f_g(\cdot; \cdot); g(\cdot; \cdot)g = f_{0; 0}g$ , and similarly for  $f_h(\cdot; \cdot); h(\cdot; \cdot)h$  when  $H$  is correctly specified, by Lemma 2.3 of Newey and McFadden (1994). That is, for correctly specified models, the minimizing value is the true value.

The dependence of  $\hat{g}$ ,  $\hat{h}$ , and  $\hat{Q}_0$  on the weighting matrices  $g$  and  $h$  in Assumption A4 reflects the fact that, when model  $G$  or  $H$  is incorrectly specified, the parameter values that minimize the GMM criterion functions  $Q_0^g(\cdot; \cdot)$  and  $Q_0^h(\cdot; \cdot)$  may depend on the choice of weighting matrices  $g$  and  $h$ . To save notation, we will omit this dependence when  $g$  and  $h$  are the standard efficient two step GMM weighting matrices. We have similarly dropped the dependence of  $Q_0^g(\cdot; \cdot)$  and  $Q_0^h(\cdot; \cdot)$  on  $g$  and  $h$  to save notation.

Together with our other Assumptions, Assumption A4 implies that GMM estimators of  $G$  or  $H$  will also converge to some (pseudo-true) values when they are misspecified. Consider, e.g., applying the standard

then be  $f_g(g); g(g)g$ , based on this construction of



while if  $H$  is correctly specified and  $G$  is not, then

$$\hat{W}_g \propto \frac{C_g^0}{C_g^0 + 0} = 1:$$

Before getting to our ODR estimator given by equation (1), consider the simpler estimator defined by

$$\hat{\theta} = \hat{W}_g \theta_g + (1 - \hat{W}_g) \theta_h \quad (4)$$

So  $\hat{\theta}$  is simply a weighted average of the GMM estimates  $\theta_g$  and  $\theta_h$ , where the weights are proportional to  $C_g^0$  and  $C_h^0$ . We will call  $\hat{\theta}$  the SODR (simpler ODR) estimator.

The intuition behind  $\hat{\theta}$  is straightforward (the asymptotic statements in this paragraph are proved formally in the next section). Suppose model  $H$  is **misspecified** and  $\theta_h$  is **biased**.

Although the SODR has the desired DR property, it also has two drawbacks. First, when  $G$  and  $H$  are both correct, the ratio  $\hat{W}_g$  converges to a random variable rather than a constant, which complicates the limiting distribution of  $\hat{\beta}$ . Second, when both  $G$  and  $H$  are correct,  $\hat{\beta}$  may be inefficient, relative to a GMM estimator that efficiently combines the moments from both models.

To address both of these issues, reconsider now the third model  $F$ , defined as the union of moments of the models  $G$  and  $H$ . Specifically, let  $F(Z; \beta; \gamma)$  be the vector valued function consisting of the union of elements of  $G(Z; \beta)$  and  $H(Z; \gamma)$ . Then, letting  $f(\beta; \gamma) = \frac{1}{n} \sum_{i=1}^n F(Z_i; \beta; \gamma)$ , we can define a third GMM estimator

$$\hat{\beta}_f = \arg \min_{\beta; \gamma} f(\beta; \gamma)' f(\beta; \gamma)$$

as shown earlier has the same limiting value as either  $g$  or  $h$ , depending on which is correctly speci..ed.

The estimator therefore, like , has the desired DR property. We show later that avoids the asymptotic issues has when both  $G$  and  $H$  are correctly speci..ed, and that generally performs better than in ..nite samples. This is why

consider different choices of  $\alpha$  in our applications. Overall, we found that the exponential

where  $U^0(C_t; X_t)$  denotes  $U(C_t; X_t) = \beta U_t$ . If the functional form of  $U^0$  is known, then this equation provides moments that allow  $b$  and  $\beta$  to be estimated using GMM. But suppose we have two different possible specifications of  $U^0$ , and we do not know which specification is correct. Then our ODR estimator can be immediately applied, replacing the expression in the inner parentheses in equation (7) with  $G(Z; \gamma)$  or  $H(Z; \gamma)$  to represent the two different specifications. Here  $\gamma$  would represent parameters that are the same in either specification, including the subjective rate of time preference  $\beta$ .

To give a specific example, a standard specification of utility is constant relative risk aversion with habit formation, where utility takes the form

$$U(C_t; X_t) = \frac{[C_t - M(X_t)]^{1-\gamma}}{1-\gamma}$$

where  $X_t$  is a vector of lagged values of  $C_t$ , the parameter  $\gamma$  is the coefficient of relative risk aversion, and the function  $M(X_t)$  is the habit function. See, e.g., Campbell and Cochrane (1999) or Chen and Ludvigson (2009). While this general functional form has widespread acceptance and use, there is considerable debate about the correct functional form for  $M$ , including whether  $X_t$  should include the current value of  $C_t$  or just lagged values. See, e.g., the debate about whether habits are internal or external as discussed in the above papers. Rather than take a stand on which habit model is correct, we could estimate the model by ODR.

To illustrate, suppose that with internal habits the function  $M(X_t)$  would be given by  $G(X_t; \gamma)$ , where  $G$  is the internal habits functional form. Similarly, suppose with external habits  $M(X_t)$  would be given by  $H(X_t; \gamma)$  where  $H$  is the external habits specification. Then, based on equation (7), we could define  $G(Z; \gamma)$  and  $H(Z; \gamma)$  by

$$G(Z; \gamma) = \beta R_{t+1} \frac{C_{t+1} - G(X_{t+1}; \gamma)}{1-\gamma}$$

( ; ). This would generally be the case, because the potential information set of consumers at time  $t$  is large relative to the number of parameters in the model.

### 3.2 Alternative Sets of Instruments

Consider a parametric model

$$Y = M(W; \beta) + \epsilon$$

where  $Y$  is an outcome,  $W$  is a vector of observed covariates,  $M$  is a known functional form,  $\beta$  is a vector of parameters to be estimated, and  $\epsilon$  is an unobserved error term. The errors may be correlated with  $W$ , so to estimate the model we wish to find instruments that are uncorrelated with  $\epsilon$ . Let  $R$  and  $Q$  denote

0 where  $G(Z; \cdot)$  is given by the stacked vectors

$$G(Z; \cdot) = \begin{pmatrix} X & Y & X^0 & x & S & s \\ L & Y & X^0 & x & S & s \end{pmatrix} \quad (8)$$

The main difficulty with applying this two stage least squares or GMM estimator is that one must find one or more covariates  $L$  to serve as instruments.

Lewbel (2012) proposes an alternative estimator that, rather than requiring that one find instruments  $L$ , instead constructs instruments based on assumptions regarding heteroscedasticity. This estimator consists of first linearly regressing  $S$  on  $X$ , and obtaining the residuals from that regression. Then a vector of instruments  $P$  is constructed by setting  $P$  equal to demeaned  $X$  (excluding the constant) times these residuals. This constructed vector  $P$  is then used instead of  $L$  above as instruments<sup>8</sup> As shown in Lewbel (2012), one set of conditions under which the vector  $P$  can be a valid set of instruments is when the endogeneity in  $S$  is due to classical measurement error in  $S$ .

Let  $X_c$  denote the vector  $X$  with the constant removed. Algebraically, we can write the instruments obtained in this way as  $R = fX; P g$  where  $P = (X_c \quad 1) (S$

=X

$\mathbf{X}$  is a vector of covariates that affect the consumer's tastes, and  $\mathbf{S}$  is the consumer's total consumption expenditures (i.e., their total budget, which must be allocated between food and non-food expenditures). Suppose, as is commonly the case, that  $\mathbf{S}$  is observed with some measurement error. To deal with this budget measurement error, a commonly employed set of instruments  $\mathbf{L}$  consists of functions of the consumer's income. However, validity of functions of income as instruments for total consumption in a food Engel curve assumes separability between the consumer's decisions on savings and their within period food expenditure decision, and this behavioral assumption may or may not be valid. It is therefore useful to consider the alternative set of potential instruments  $\mathbf{P}$  defined above. Use of  $\mathbf{P}$  does not require including covariates from outside the model, like income, to use as instruments, but does require that certain measurement error assumptions hold. Our later empirical application applies ODR to this application, thereby obtaining consistent estimates of  $\beta$  if either  $\mathbf{L}$  or  $\mathbf{P}$  are valid instruments.

#### 4 The ODR Estimator Asymptotics

In this section we show consistency of our ODR estimator  $\hat{\beta}$ , and then derive its limiting distribution, which is root  $n$  consistent and asymptotically normal. We make the following additional assumptions. What these assumptions mostly do is ensure that GMM estimates of model  $\mathbf{G}$ ,  $\mathbf{H}$ , and  $\mathbf{F}$  are each asymptotically normal around the true values when correctly specified, and are suitably bounded in probability around the pseudo-true values when misspecified. We do not require asymptotic normality under misspecification.

Assumption A5:  $\mathbf{G}(\mathbf{Z}; \beta)$ ,  $\mathbf{H}(\mathbf{Z}; \beta)$  and  $\mathbf{F}(\mathbf{Z}; \beta; \gamma)$  are continuous at  $\beta; \gamma$ ,  $\mathbf{f}(\mathbf{Z}; \beta; \gamma)$ , and  $\mathbf{f}(\mathbf{Z}; \beta; \gamma)$  respectively, with probability one.

Assumption A5.9 (7) (2017) 29 Tf 38( )pb ;



$f = h; h; g$ ; and  $f = f; f; f; g$ . If the models  $G$  and  $H$  are correctly specified,  $g_0 = g$ ,  $h_0 = h$ , and  $f_0 = f$ .

Assumption A7: With probability one,  $G(Z; \cdot; \cdot)$ ,  $H(Z; \cdot; \cdot)$ , and  $F(Z; \cdot; \cdot; \cdot)$  are twice continuously differentiable in a neighborhood  $\mathcal{G}$  of  $g$ ,  $\mathcal{H}$  of  $h$ , and  $\mathcal{F}$  of  $f$ , respectively.

Assumption A8:  $H_g(\frac{g}{0}) = g_0(\frac{g}{0}) = g$



## 4.1 ODR Consistency

Lemma 1 : Suppose Assumptions A1 to A15 hold. Then, for any  $\epsilon$  with  $0 < \epsilon < 1$ ,  $\hat{\theta}_n$

Case 1) Suppose both  $g_0(\sigma; \sigma) = 0$  and  $h_0(\sigma; \sigma) = 0$ . Then  $f_{g; g} \neq 0$ ,  $f_{h; h} \neq 0$ ,  $f_{f; f} \neq 0$ , and  $f_{f; f} \neq 0$ , so  $\hat{Q}^g \neq 0$ ,  $\hat{Q}^h \neq 0$ , and  $\hat{Q}^f \neq 0$ . By Lemma 1,  $\hat{W}_f$  and  $\hat{W}_g$



weights  $\hat{W}_g$  and  $\hat{W}_f$ ). It is therefore numerically desirable in finite samples to have these matrices be

estimators, the numerator of the weight on model  $H$  depends on the criterion for model  $G$  (i.e., on  $Q^g$ ) designed to put all weight on model  $H$  when model  $G$  is wrong but  $H$  is correct, and vice versa.

A difference between  $MG$  and SODR (with SODRv26.996(76onT2 .996(76tia1 10.909 4f 14.161 0 Tdd (2713)T)

that vary by correlations  $R_j$  and  $Q_j$ . The first design takes  $R_j = Q_j = 0$ , which makes both models right (both sets of instruments are valid). The second takes  $R_1 = R_2 = 0$ ,  $Q_1 = 0.4$ , and  $Q_2 = 0.6$ , which makes model **G** right (i.e., **R** are valid instruments so **G** is correctly specified) and model **H** be wrong (i.e., **Q** are not valid instruments, because they correlate with the model error). The third takes  $R_1 = 0.4$ ,  $R_2 = 0.6$  and  $Q_1 = Q_2 = 0$ , which makes model **H** right and model **G** wrong.

For the tuning function discussed in sections 2.3 and 4.4, we consider two different choices;  $n\hat{Q} = \exp(-n\hat{Q})$  and  $n\hat{Q} = (n\hat{Q})^2$  so the weighting functions  $\hat{W}_g$  and  $\hat{W}_f$  are

$$1: \hat{W}_g = \frac{\exp(-n\hat{Q}^g)g^{-1}}{\exp(-n\hat{Q}^g)g^{-1} + \exp(-n\hat{Q}^h)g^{-2}}, \hat{W}_f = 1 - \frac{1}{\exp(-n\hat{Q}^f)g^{-1}}; \quad (12)$$

$$2: \hat{W}_g = \frac{(n\hat{Q}^g)g^{-2}}{(n\hat{Q}^g)g^{-2} + (n\hat{Q}^h)g^{-2}}, \hat{W}_f = 1 - \frac{1}{(n\hat{Q}^f)g^{-2} + 1}; \quad (13)$$

For the tuning parameter  $\lambda$ , we use  $\lambda = 1 - p$ , where  $p$  is the p-value of the Wald statistic as discussed in section 2.3.

We report eight estimates of  $\beta_1$  and  $\beta_0$  for each simulation. First is **GMM** based on the model **G** moments, denoted by **GMM<sub>g</sub>** (which is only consistent if model **G** is right). Second is **GMM** based on the **H** moments, denoted by **GMM<sub>h</sub>**.



We report skewness (Skew) and kurtosis (Kurt) of these t-statistics across simulations, and the frequency (Freq) that these t-statistics are less than 2 in magnitude, corresponding to the frequency with which a 2 estimated standard error confidence interval contains the true parameter value. Also, to check the accuracy of the standard error estimates, we report the average of the estimated standard errors (SE), and standard deviation of the estimated standard errors ( $SD_{SE}$ ), across the simulations. The last five summary statistics are not reported for **SODR**, because we do not consider its limiting distribution due to the random probability limit of  $\hat{W}_g$ .

When both sets of instruments are valid, **ODR** estimates are almost as precise as **GMM<sub>f</sub>**, and when either set of instruments is invalid, **ODR** estimates are more precise than inconsistent **GMM** estimators. The **SODR** estimates are found to be less efficient than **ODR** when both **G** and **H** models are valid (as expected), but when one model is invalid, **SODR** is similar to **ODR**. In this application, the cost in efficiency of choosing the simple true 9 T. ation (as large) -45-33in-45-33aba

---

---

Table 1. Simulation Results of  $\rho_1$  (

Table 2. Simulation Results of  $\theta_0$  ( $n = 100$ )

	Bias	Mde	RMSE	MAE	SD	Skew	Kurt	Freq	SE	$SD_{SE}$
Both correct										
<b><i>GMM<sub>g</sub></i></b>	-0.0038	-0.0048	0.0112	0.0687	0.1058	0.0005	3.1738	0.9415	0.1009	0.0089
<b><i>GMM<sub>h</sub></i></b>	-0.0024	-0.0090	0.0134	0.0757	0.1157	-0.0131	2.9788	0.9490	0.1115	0.0182
<b><i>GMM<sub>f</sub></i></b>	-0.0046	-0.0073	0.0113	0.0688	0.1063	0.0212	3.1124	0.9350	0.0981	0.0085
<b><i>MG</i></b>	-0.0022	-0.0063	0.0115	0.0697	0.1071	0.0291	3.0642	0.9440	0.1022	0.0110
<b><i>ODR<sub>1</sub></i></b>	-0.0039	-0.0062	0.0113	0.0686	0.1063	0.0583	3.0524	0.9370	0.0989	0.0092
<b><i>ODR<sub>2</sub></i></b>	0.0001	-0.0017	0.0105	0.0687	0.1025	-0.0532	3.1744	0.9525	0.0990	0.0088
<b><i>SODR<sub>1</sub></i></b>	-0.0016	-0.0067	0.0120	0.0703	0.1097					
<b><i>SODR<sub>2</sub></i></b>	0.0014	0.0023	0.0108	0.0707	0.1041					
G correct										
<b><i>GMM<sub>g</sub></i></b>	-0.0038	-0.0060	0.0112	0.0683	0.1060	-0.0390	3.1287	0.9395	0.1009	0.0108
<b><i>GMM<sub>h</sub></i></b>	-0.2005	-0.1977	0.0554	0.1977	0.1234	0.1485	3.0509	0.5750	0.1103	0.0179
<b><i>GMM<sub>f</sub></i></b>	-0.0744	-0.0737	0.0219	0.0999	0.1280	-0.0354	3.1266	0.7540	0.0867	0.0074
<b><i>MG</i></b>	-0.0401	-0.0396	0.0140	0.0774	0.1115	-0.1154	3.1855	0.8885	0.0954	0.0109
<b><i>ODR<sub>1</sub></i></b>	-0.0258	-0.0198	0.0147	0.0722	0.1186	-0.2332	3.2476	0.9010	0.0996	0.0120
<b><i>ODR<sub>2</sub></i></b>	-0.0245	-0.0198	0.0136	0.0744	0.1139	-0.2004	3.0110	0.9065	0.0995	0.0114
<b><i>SODR<sub>1</sub></i></b>	-0.0258	-0.0198	0.0147	0.0722	0.1186					
<b><i>SODR<sub>2</sub></i></b>	-0.0240	-0.0194	0.0136	0.0745	0.1142					
H correct										
<b><i>GMM<sub>g</sub></i></b>	-0.1151	-0.1166	0.0230	0.1198	0.0989	0.0139	2.8983	0.6735	0.0808	0.0069
<b><i>GMM<sub>h</sub></i></b>	-0.0028	-0.0088	0.0133	0.0722	0.1153	-0.2405	2.9748	0.9530	0.1123	0.0344
<b><i>GMM<sub>f</sub></i></b>	-0.0963	-0.0966	0.0203	0.1039	0.1050	-0.0085	2.9169	0.7095	0.0791	0.0068
<b><i>MG</i></b>	-0.0035	-0.0094	0.0133	0.0720	0.1151	-0.2389	2.9660	0.9515	0.1120	0.0343
<b><i>ODR<sub>1</sub></i></b>	-0.0051	-0.0105	0.0131	0.0725	0.1146	-0.2535	2.9609	0.9475	0.1109	0.0320
<b><i>ODR<sub>2</sub></i></b>	-0.0084	-0.0187	0.0135	0.0753	0.1159	-0.1964	3.0290	0.9380	0.1095	0.0287
<b><i>SODR<sub>1</sub></i></b>	-0.0029	-0.0089	0.0133	0.0722	0.1153					
<b><i>SODR<sub>2</sub></i></b>	-0.0038	-0.0144	0.0138	0.0760	0.1176					



Table 4. Simulation Results of  $\theta_0$  ( $n = 500$ )

	Bias	Mde	RMSE	MAE	SD	Skew	Kurt	Freq	SE	$SD_{SE}$
Both correct										
<b><i>GMM</i><sub>g</sub></b>	-0.0010	-0.0002	0.0021	0.0315	0.0458	-0.1391	2.9732	0.9565	0.0459	0.0018
<b><i>GMM</i><sub>h</sub></b>	-0.0008	0.0005	0.0024	0.0328	0.0492	-0.1701	3.0631	0.9500	0.0491	0.0030
<b><i>GMM</i><sub>f</sub></b>	-0.0011	0.0000	0.0021	0.0311	0.0458	-0.1335	2.9799	0.9550	0.0454	0.0017
<b><i>MG</i></b>	-0.0007	0.0004	0.0021	0.0311	0.0463	-0.1527	3.0340	0.9570	0.0462	0.0021
<b><i>ODR</i><sub>1</sub></b>	-0.0010	0.0000	0.0021	0.0310	0.0459	-0.1327	3.0063	0.9540	0.0455	0.0018
<b><i>ODR</i><sub>2</sub></b>	0.0009	-0.0005	0.0022	0.0315	0.0471	0.0061	2.9664	0.9445	0.0455	0.0019
<b><i>SODR</i><sub>1</sub></b>	-0.0005	0.0003	0.0022	0.0321	0.0468					
<b><i>SODR</i><sub>2</sub></b>	0.0010	0.0003	0.0023	0.0334	0.0483					
G correct										
<b><i>GMM</i><sub>g</sub></b>	-0.0010	-0.0003	0.0021	0.0314	0.0458	-0.1566	2.9735	0.9570	0.0459	0.0021
<b><i>GMM</i><sub>h</sub></b>	-0.2000	-0.2000	0.0428	0.2000	0.0529	0.0663	3.1573	0.0225	0.0495	0.0033
<b><i>GMM</i><sub>f</sub></b>	-0.0732	-0.0731	0.0084	0.0739	0.0554	0.0501	2.9813	0.5400	0.0402	0.0014
<b><i>MG</i></b>	-0.0012	-0.0004	0.0021	0.0314	0.0458	-0.1553	2.9705	0.9570	0.0458	0.0021
<b><i>ODR</i><sub>1</sub></b>	-0.0010	-0.0003	0.0021	0.0314	0.0458	-0.1563	2.9744	0.9570	0.0459	0.0021
<b><i>ODR</i><sub>2</sub></b>	-0.0020	-0.0011	0.0021	0.0315	0.0459	-0.1685	2.9918	0.9550	0.0457	0.0021
<b><i>SODR</i><sub>1</sub></b>	-0.0010	-0.0003	0.0021	0.0314	0.0458					
<b><i>SODR</i><sub>2</sub></b>	-0.0020	-0.0011	0.0021	0.0315	0.0459					
H correct										
<b><i>GMM</i><sub>g</sub></b>	-0.1122	-0.1121	0.0146	0.1121	0.0448	-0.0037	3.0575	0.1945	0.0367	0.0013
<b><i>GMM</i><sub>h</sub></b>	-0.0007	-0.0007	0.0024	0.0329	0.0494	-0.2688	3.0914	0.9480	0.0492	0.0048
<b><i>GMM</i><sub>f</sub></b>	-0.0938	-0.0948	0.0111	0.0948	0.0481	-0.0661	2.9792	0.3445	0.0366	0.0013
<b><i>MG</i></b>	-0.0007	-0.0007	0.0024	0.0329	0.0494	-0.2688	3.0914	0.9480	0.0492	0.0048
<b><i>ODR</i><sub>1</sub></b>	-0.0007	-0.0007	0.0024	0.0329	0.0494	-0.2688	3.0914	0.9480	0.0492	0.0048
<b><i>ODR</i><sub>2</sub></b>	-0.0011	-0.0038	0.0025	0.0340	0.0500	-0.1804	2.9318	0.9555	0.0491	0.0049
<b><i>SODR</i><sub>1</sub></b>	-0.0007	-0.0007	0.0024	0.0329	0.0494					
<b><i>SODR</i><sub>2</sub></b>	-0.0011	-0.0037	0.0025	0.0340	0.0500					

**GMM**. This suggests a modest advantage of the exponential tuning function  $\tau_1$ .

One should expect correctly specified **GMM** estimators to be more efficient than **ODR**, and that is indeed the case. But in many of the simulations, the loss in efficiency from using **ODR** is very low. In particular, when model **G** is invalid, so only the weaker instruments are valid, the precision of **ODR** is almost identical to that of the efficient **GMM<sub>h</sub>**. So, using our **ODR**, there is little loss in efficiency from not knowing which specification is correct. In summary, we conclude that our proposed **ODR** works well, even at low sample sizes.

## 6 Empirical Application: Engel Curve Estimation

Here we empirically estimate the Engel curve example discussed in section 3.2.  $Y$  is the food budget share,  $S$  is log real total consumption expenditures, and  $X$  is a vector of other covariates that serve as controls<sup>1</sup>. The goal is estimation of the coefficient of  $S$  in a regression of  $Y$  on  $S$  and  $X$ . Total consumption  $S$  is observed with measurement error, so instrumental variables estimation is used to correct for the resulting endogeneity. The vector  $L$  consists of two candidate external instrumental variables, real total income and real total income squared. Model **G** assumes these external instruments are valid. Model **H** instead assumes that constructed instruments based on heteroscedasticity as described by Lewbel (2012) and summarized in section 3.2 above are valid. Model:

which are heteroscedasticity based constructed instruments  $GMM_f$  is the GMM estimator that uses both sets of instruments, and  $SODR$  and  $ODR$  are our new estimators given in equations (4) and (1) with the tuning functions  $\lambda_1$  and  $\lambda_2$ .

The estimated results show that the external instruments of model  $G$  are much stronger than the constructed instruments of model  $H$ . This is not surprising since the constructed instruments are based on higher moments of the data. This difference in strength can be seen in the standard errors of  $\hat{\beta}_s$ , which are much lower in model  $G$  than in model  $H$ , and also in model  $GMM_f$  which gives estimates much closer to  $GMM_g$  than  $GMM_h$ .

The point estimates of  $GMM_g$  and  $GMM_h$  are substantially different, which could be due to having one of these sets of instruments be invalid. However, this difference could also just be due to imprecision, particularly of  $GMM_h$ . This illustrates the usefulness of our  $ODR$ , which does not require resolving which set of instruments is valid, or if both are valid.

Table 5. Engel Curve Estimates

	$GMM_{g0}$	$GMM_g$	$GMM_h$	$GMM_f$	$SODR_{\lambda_1}$	$ODR_{\lambda_1}$	$SODR_{\lambda_2}$	$ODR_{\lambda_2}$
$\hat{\beta}_s$	-0.0859 (0.0198)	-0.0840 (0.0197)	-0.0521 (0.0546)	-0.0862 (0.0177)	-0.0812	-0.0862 (0.0192)	-0.0831	-0.0862 (0.0192)
$\hat{\beta}_0$	0.336 (0.0122)	0.335 (0.0120)	0.317 (0.0328)	0.337 (0.0109)	0.333	0.337 (0.0118)	0.335	0.337 (0.0118)
$\lambda_2$		0.191	12.91	15.94				
$d:f:$		1	11	13				
p-value		0.662	0.299	0.252				
$\hat{Q}$		0.0002	0.0014	0.0014				
$\hat{W}_g; \hat{W}_f; p$					0.09, 0.004, 0.86		0.03, 0.000, 0.86	

13

<sup>13</sup>Table 5 notes: We report coefficient estimates with associated standard errors in parentheses, except SODR. Also reported is  $\lambda_2$ , the Hansen (1982) test statistics for overidentified GMM, along with their degrees of freedom and p-values.  $\hat{Q}$  is the normalized minimand of the GMM estimators. The last row reports weights  $\hat{W}_g$ ,  $\hat{W}_f$ , and gives  $p$ , which is the p-value of the Wald statistic testing the null hypothesis that  $\beta_g = \beta_h$ . This  $p$  is used to construct  $\lambda_1 = 1 - p$  in  $\hat{W}_f$  in equation (5), as explained in section 2.3.

The estimated weight  $\hat{W}_g$  is 0.09 with the tuning function  $\lambda_1$  and 0.03 with  $\lambda_2$ , so **SODR** puts over ten times as much weight on model **G** as on model **H**. However, in **ODR** the weight on model **F**,  $\hat{W}_f$ , is 0.996 with  $\lambda_1$  and is one to three decimal places with  $\lambda_2$ . The very small difference in  $\hat{W}_f$  between  $\lambda_1$  and  $\lambda_2$  is why both of the **ODR** estimates appear the same in Table 5 (they actually differ in the fourth significant digit: -0.08617 vs. -0.08619 for  $\lambda_2$ ).

The very high weight on model **F** strongly suggests that both models are likely to be correctly specified. This therefore implies that the difference between  $GMM_g$  and  $GMM_h$  is likely due to imprecision of  $GMM_h$  rather than misspecification of the constructed instruments in model **H**. Further evidence that both are



all  $s > 0$ , and the third either converges to a constant or diverges depending on  $s$  (and sometimes  $\alpha$ ) as discussed below.<sup>4</sup>

First suppose model  $G$  is locally misspecified with  $s > 1/2$ . Then  $nQ^g \xrightarrow{p} \chi^2_{k_g}(0)$ , which is the same limit as when  $G$  is correctly specified, and similarly for  $H$ . As a result, in this case the SODR and ODR estimators have the same  $\sqrt{n}$  consistent, asymptotically normal limiting distribution as they have when  $G$  is correctly specified, and similarly for  $H$ . Note this means that instead of requiring that either  $G$  or  $H$  (or both) be correctly specified, it is sufficient to assume that either  $G$  or  $H$  (or both) are locally misspecified with  $s > 1/2$ , noting that correct specification is the special case  $\alpha = 1$ .

If model  $G$  is locally misspecified with  $s < 1/2$ , then  $nQ^g \xrightarrow{p} \infty$  and the SODR has the same  $\sqrt{n}$  consistent, asymptotically normal limiting distribution as when  $G$  is globally misspecified. The ODR will also have the same limiting distribution as when  $G$  is globally misspecified, as long as the tuning parameter  $\lambda$  has  $\lambda > s + 0.5$ . This then guarantees that model  $G$  will asymptotically have zero weight. Since these cases are equivalent asymptotically to  $G$  being globally misspecified, we need to assume that  $G$  is either correctly specified, or locally misspecified with  $s > 1/2$ . This generalizes our original theorems that simply assumed either  $G$  or  $H$  is correctly specified.

Finally, suppose model  $G$  is locally misspecified with  $s = 1/2$ . Then  $nQ^g$  converges to a noncentral chi-squared distribution. Specifically,  $nQ^g \xrightarrow{p} \chi^2_{k_g}(\lambda_g)$ , where the object in parentheses is the noncentrality parameter and the formula for  $\lambda_g$  is given in the Supplemental Appendix. In this case the GMM estimator of model  $G$  is consistent but not  $\sqrt{n}$  consistent, as established in, e.g., Newey and McFadden (1994). Here  $nQ^g$  is still bounded in probability, and if  $H$  is correctly specified (or locally misspecified with  $s > 1/2$ ), then  $nQ^f$  is also bounded in probability. Thus, ODR will asymptotically put weight on model  $F$ , which then is consistent but may not be  $\sqrt{n}$  consistent. As a result, in this knife edge case, ODR will be consistent, but not  $\sqrt{n}$  consistent, since  $\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{p} 0$ .

will have the same limiting distribution as efficient GMM with both  $G$  and  $H$  correctly specified. If just  $G$  is locally misspecified with  $s > 1/2$  (again including as a special case having  $G$  be correctly specified by  $s = 1$ ), and  $H$  is either misspecified or locally misspecified with  $s < 1/2$ , then (assuming  $s > s + 0.5$ ) ODR will have the same limiting distribution as efficient GMM based just on model  $G$  (and vice versa, exchanging the roles of  $G$  and  $H$ ). Equivalently we can say that our earlier Theorem 2 still holds, replacing "correctly specified model" with "locally misspecified model having any  $s > 1/2$ , including  $s = 1$ " and replacing "incorrectly specified model" with "locally misspecified model having any  $s < 1/2$ , including  $s = 0$ ."

We conclude this section with some additional Monte Carlo results (reported in Tables 6 and 7 in the Supplemental Appendix), which we find support these conclusions. We use the same designs and estimators as in section 5 but with a drift parameter  $s$  for the locally misspecified cases. Since ODR performed better with the tuning function  $\psi_1$  in section 5, to save space we only report ODR  $\psi_1$ , along with  $GMM_g$ ,  $GMM_h$ , and  $GMM_f$ . In these tables, model  $H$  is either globally misspecified, or locally misspecified with  $s$  equal to 0.25, 0.50, or 0.75. In Tables 6-1 and 6-2 model  $G$  is correctly specified, while in Tables 7-1 and 7-2  $G$  is locally misspecified with  $s = 0.75$ .

The ...nite heandav-3339 Td (tl)]TJ /TT.97 -247 -24.388 Td [7 -247 -3n

estimation of  $\hat{\theta}_l$  would then be the weighted average

$$= \frac{\hat{\theta}^g(\hat{g}; \hat{g}) \hat{\theta}^h(\hat{h}; \hat{h}) \hat{\theta}^l(\hat{l}; \hat{l}) + \hat{\theta}^l(\hat{l}; \hat{l}) \hat{\theta}^g(\hat{g}; \hat{g}) \hat{\theta}^h(\hat{h}; \hat{h})}{\hat{\theta}^g(\hat{g}; \hat{g}) \hat{\theta}^h(\hat{h}; \hat{h}) + \hat{\theta}^l(\hat{l}; \hat{l}) \hat{\theta}^h(\hat{h}; \hat{h}) + \hat{\theta}^l(\hat{l}; \hat{l}) \hat{\theta}^g(\hat{g}; \hat{g})} \quad (14)$$

$$= \frac{\frac{\hat{\theta}^l(\hat{l}; \hat{l})}{1} + \frac{\hat{\theta}^g(\hat{g}; \hat{g})}{1} + \frac{\hat{\theta}^h(\hat{h}; \hat{h})}{1}}{\frac{\hat{\theta}^l(\hat{l}; \hat{l})}{1} + \frac{\hat{\theta}^g(\hat{g}; \hat{g})}{1} + \frac{\hat{\theta}^h(\hat{h}; \hat{h})}{1}}. \quad (15)$$

In equation (14), the weight on  $\hat{\theta}^l$  is proportional to the product of objective functions for the other models,  $\hat{\theta}^g \hat{\theta}^h$ , and similarly for the weights on  $\hat{\theta}^g$  and  $\hat{\theta}^h$ .

The above estimator is a simple extension of our **SODR** estimator because the **SODR** can be rewritten as

$$= \frac{\frac{\hat{\theta}^g(\hat{g}; \hat{g})}{1} + \frac{\hat{\theta}^h(\hat{h}; \hat{h})}{1}}{\frac{\hat{\theta}^g(\hat{g}; \hat{g})}{1} + \frac{\hat{\theta}^h(\hat{h}; \hat{h})}{1}}.$$

The logic of  $\hat{\theta}_l$  is the same as for the **SODR** estimator. For example, if model **G** is right and models **L** and **H** are wrong, then only  $\hat{\theta}_g$  will get a nonzero weight asymptotically. Now suppose two but get by del 8 10.9

can suffer from well known finite sample biases when models have many more moments than parameters, and particularly when some moments might be weak. In such cases, it may be desirable to let models and  $H$  equal just a subset of the available moments for each. Existing moment selection methods such as Andrews and Lu (2001), Caner (2009), or Liao (2013) might be used prior to applying ODR, though this then introduces pretest bias that ODR is intended to avoid. A potential subject for future work could be



Lee, M.J., and Lee, S. (2019): "Double Robustness Without Weighting", *Statistics and Probability Letters*, 146, 175-180.

Average Treatment Effects", *Econometric Theory*, 34(01), 112-133.

Sueishi, M. (2013): "Generalized Empirical Likelihood-Based Focused Information Criterion and Model Averaging", *Econometrics*, 1(2), 141-156.

Wooldridge, J. (2007): "Inverse Probability Weighted Estimation for General Missing Data Problems", *Journal of Econometrics*, 141(2), 1281-1301.

# Supplemental Appendix: Over-Identified Doubly Robust Identification and Estimation

by Arthur Lewbel, Jin-Young Choi, and Zhuzhu Zhou

Original 2018, Revised November 2021

This Supplemental Appendix consists of five parts. The first is a proof of Lemma 1 and of Theorem 2, which give the asymptotic properties of ODR described in Section 4.1-2. The second part is a proof of Lemma 2 and of Theorems 3 and 4, which provide asymptotic properties of ODR



where  $k_g$  is the degrees of freedom of the chi-squared statistic that  $Q^g$  converges to if the G model is true. That is the integer  $k_g$  is the number of moments in G ( $R_g$ ) minus the number of elements in  $\beta$  ( $k_g$ ) which is positive as discussed earlier. For notational simplicity, let  $Q^g$ ,  $Q^h$ , and  $Q^f$  denote  $Q^g(\hat{\alpha}_g; \hat{\alpha}_g)$ ;  $Q^h(\hat{\alpha}_h; \hat{\alpha}_h)$ ; and  $Q^f(\hat{\alpha}_f; \hat{\alpha}_f; \hat{\alpha}_f)$ ; respectively. The population version of  $Q$  corresponding to each model is

$$Q_0^g = \frac{C_g^0}{k_g} C_g; \quad Q_0^h = \frac{C_h^0}{k_h} C_h; \quad Q_0^f = \frac{C_f^0}{k_f} C_f;$$

If the model is correctly specified, then  $Q_0^j = 0$ ,  $j = g; h; f$ .

Our proposed ODR estimator is a weighted average of  $b_g$ ,  $b_h$ , and  $b_f$ , taking the form

$$b = \hat{W}_f \hat{W}_g b_h + \hat{W}_f (1 - \hat{W}_g) b_g + (1 - \hat{W}_f) b_f \tag{1}$$

Proof of Lemma 1.

To obtain the probability limits of  $\hat{W}_g$  and  $\hat{W}_f$ , ...rst we consider without loss of generality the probability limit of  $Q^g$  when model G is correctly specified, and when it's misspecified. The asymptotics for  $Q^h$  and  $Q^f$  are obtained following the same logic. After these derivations, we then obtain the probability limits of  $\hat{W}_g$  and  $\hat{W}_f$  based on  $Q^g$ ,  $Q^h$  and  $Q^f$ . First we have

$$nQ^g = \frac{1}{k_g} \sum_{g=1}^p \bar{n} b(\hat{\alpha}_g) g^{\alpha} \hat{\alpha}_g^{1-2p} \bar{n} b(\hat{\alpha}_g) g^{\alpha} \frac{1}{k_g} \tag{2}$$

From the ...rst order condition for  $\hat{\alpha}_g$  minimizing  $Q^g(\hat{\alpha}_g)$ , we have

$$P \bar{n} b(\hat{\alpha}_g) \hat{\alpha}_g b(\hat{\alpha}_g) = 0:$$

Taylor-expanding the last term  $b(\hat{\alpha}_g)$  around  $\alpha_g$  gives

$$\begin{aligned} 0 &= P \bar{n} b(\hat{\alpha}_g) \hat{\alpha}_g b(\alpha_g) + r_0 b^{(-g)}(\hat{\alpha}_g - \alpha_g) g^{\alpha} \\ &= r_0 b(\hat{\alpha}_g) \hat{\alpha}_g P \bar{n} b(\alpha_g) + r_0 b(\hat{\alpha}_g) \hat{\alpha}_g r_0 b^{(-g)} P \bar{n}(\hat{\alpha}_g - \alpha_g) \end{aligned}$$

where  $\bar{\alpha}_g$  is a mean value between  $\alpha_g$  and  $\hat{\alpha}_g$

where  $\bar{g}$  is a mean value between  $g_0$  and  $\hat{g}$ . Plug equation (3) with  $g$  replaced by  $g_0$  into this equation to get

$$\begin{aligned} \hat{g}^{1=2P} \bar{g}(\hat{g}) &= \hat{g}^{1=2P} \bar{g}(g_0) + \hat{g}^{1=2r} \bar{g}(\bar{g})(\hat{g}^g)^{1-r} \bar{g}(\hat{g}) \hat{g}^P \bar{g}(g_0) \\ &= \text{fl}_{R_g} \hat{g}^{1=2r} \bar{g}(\bar{g})(\hat{g}^g)^{1-r} \bar{g}(\hat{g}) \hat{g}^{1=2g} \hat{g}^{1=2P} \bar{g}(g_0) = \hat{g} \hat{g}^{1=2P} \bar{g}(g_0); \end{aligned} \quad (4)$$

where  $\hat{g} \text{ fl}_{R_g} \hat{g}^{1=2r} \bar{g}(\bar{g})(\hat{g}^g)^{1-r} \bar{g}(\hat{g}) \hat{g}^{1=2g}$

and  $I_{R_g}$  is the  $R_g \times R_g$  identity matrix and  $R_g$  is the number of moments in the model G.

By Assumption A10 and the Lindberg-Levy CLT,  $\bar{g}(g_0) \xrightarrow{d} N_r(\mu, \sigma^2)$  60 hm77 -4.3355

Under Assumption A13,  $\sqrt{n}(\hat{\beta}(\hat{g}) - \beta_0(\hat{g}))$  is asymptotically normal with mean zero by the Lindeberg-Levy CLT, so it is bounded in probability. And  $\sqrt{n}(\hat{\beta}(\hat{g}) - \beta_0(\hat{g}))$  is also bounded in probability by Assumption A12. Under Assumption A7, A10, A11, A14, and the consistency of  $\hat{\beta}$

and  $\hat{W}_f \hat{W}_g \xrightarrow{P} 0$ .

Case 3). Suppose now  $g_0(\omega; \omega) \neq 0$  but  $h_0(\omega; \omega) = 0$ . Then  $f \wedge_g; \wedge_g g \xrightarrow{P} f \wedge_g; g g, f \wedge_h; \wedge_h g \xrightarrow{P} f \wedge_f; \wedge_f g \xrightarrow{P} f \wedge_f; f g$ . So  $\hat{Q}^g \xrightarrow{P} Q_0^g = c_g^0 g c_g = k_g > 0$ ,  $\hat{Q}^h \xrightarrow{P} Q_0^h = c_h^0 h c_h = k_h = 0$ , and  $\hat{Q}^f \xrightarrow{P} Q_0^f = c_f^0 f c_f = k_f > 0$ . Following the same argument as in Case 2),  $\hat{W}_g \xrightarrow{P} 1$  and  $\hat{W}_f \xrightarrow{P} 1$ . In short, the probability limits of  $\hat{W}_f$  and  $\hat{W}_g \hat{W}_f$  are categorized as follows:

- Case 1) Both G and H are correctly specified  $\Rightarrow \hat{W}_f \xrightarrow{P} 0$  and  $\hat{W}_f \hat{W}_g \xrightarrow{P} 0$ ;
- Case 2) G is correctly specified, but H is not  $\Rightarrow \hat{W}_f \xrightarrow{P} 1$  and  $\hat{W}_f \hat{W}_g \xrightarrow{P} 0$ ;
- Case 3) H is correctly specified, but G is not  $\Rightarrow \hat{W}_f \xrightarrow{P} 1$  and  $\hat{W}_f \hat{W}_g \xrightarrow{P} 1$ ;

Q.E.D.

### Proof of Theorem 2 .

Recall equation (1) and rewrite it as

$$\hat{\alpha} = \hat{W}_f \hat{W}_g (\hat{\alpha}_h \quad \omega) + \hat{W}_f (1 - \hat{W}_g) (\hat{\alpha}_g \quad \omega) + (1 - \hat{W}_f) (\hat{\alpha}_f \quad \omega):$$

From this, we have

$$\begin{aligned} P \bar{n}(\hat{\alpha} \quad \omega) &= \hat{W}_f \hat{W}_g P \bar{n}(\hat{\alpha}_h \quad \omega) + \hat{W}_f (1 - \hat{W}_g) P \bar{n}(\hat{\alpha}_g \quad \omega) + (1 - \hat{W}_f) P \bar{n}(\hat{\alpha}_f \quad \omega) \\ &= \hat{W}_f \hat{W}_g P \bar{n}(\hat{\alpha}_h \quad h) + \hat{W}_f (1 - \hat{W}_g) P \bar{n}(\hat{\alpha}_g \quad g) + (1 - \hat{W}_f) P \bar{n}(\hat{\alpha}_f \quad f) \quad (9) \end{aligned}$$

30.8 (953 d) (9) 1.920 4.2453 4.97 0.2 5.713; 0.11 (4.106 (8) 1.007 2.0396) d) 9(y.159 r(h) Tj 0

d (0)Tj 2d (0)Tj 2

Case 2). Suppose  $G$  is correct, but  $H$  is not ( $\mathcal{H}_0 \neq \emptyset$ ). In this case,  $F$  is also misspecified ( $\mathcal{F}_0 \neq \emptyset$ ). (9) can be rewritten as

$$p_{\hat{n}}(\alpha) = W_f \ln^{\wedge} \text{[()]-278(=)]TJ22(+711 3.2.383d (^)Tj /TT1 11.955 Tf -3.412 -3.022 Td (W}$$

## Appendix II: Proof of Lemma 2 and Theorems 3 and 4

Let the model  $G$  be "locally misspecified" when the parameter in the data generating process takes the form  $\theta = \theta_0 + \theta_1 n^{-s}$  for a constant  $\theta_1$  and  $s > 0$ , while  $\theta_0$  satisfies  $E f_G(Z; \theta_0) = 0$  due to Assumption A3. Ahem

Following the same steps as in Case ii) of Lemma 1, we can rewrite the last term other than  $(\hat{H}^g)^{-1}$  in (15) as

$$r \hat{g}(\hat{g})^{\wedge}_g^p$$

Case 4). Suppose that model G is correct, but H is locally misspecified with  $h = \frac{h}{0} + h n^s$ . In this case, F is also locally misspecified with  $f = \frac{f}{0} + f n^s$  for some  $f$ .

Case 4-1). If  $s = 1=2$ , as shown in Case iii),  $n\hat{Q}^h \xrightarrow{d} \frac{2}{k_h} (! \frac{0}{h} h! h) = k_h$  and  $n\hat{Q}^f \xrightarrow{d} \frac{2}{k_f} (! \frac{0}{f} f! f) = k_f$  as  $n \rightarrow \infty$ . Thus  $\hat{W}_g = n\hat{Q}^g(\hat{\Lambda}_g; \hat{\Lambda}_g) = fn\hat{Q}^g(\hat{\Lambda}_g; \hat{\Lambda}_g) + n\hat{Q}^h(\hat{\Lambda}_h; \hat{\Lambda}_h)g$  converges to a distribution on  $(0; 1)$ . For  $\hat{W}_f$ , we have

$$\hat{W}_f = 1 - \frac{1}{n\hat{Q}^f + 1} = 1 - \frac{1}{n^{-1}n\hat{Q}^f + 1} \xrightarrow{p} 0;$$

because  $n\hat{Q}^f$  is bounded in probability, and  $n^{-1} \rightarrow 0$ . Thus,  $\hat{W}_g\hat{W}_f \xrightarrow{p} 0$ .

Case 4-2). If  $s > 1=2$ ,  $n\hat{Q}^h \xrightarrow{d} \frac{2}{k_h} = k_h$ , and  $n\hat{Q}^f \xrightarrow{d} \frac{2}{k_f} = k_f$ . Therefore, it is asymptotically the same as Case 1) of Lemma 1.

Case 4-3). If  $s < 1=2$ ,  $n\hat{Q}^h$  and  $n\hat{Q}^f$  are  $O_p(n^{2(1-2s)})$ , as each is a squared version of a term analogous to (18). In this case, whereas  $\hat{W}_g \xrightarrow{p} 0$ , convergence of  $\hat{W}_f$  depends on the relationship between  $s$  and  $s$ . Because  $\hat{Q}^f = O(n^{-1})O_p(n^{2(1-2s)}) = O_p(n^{-2s})$ , when  $s > 2s$ ,  $n\hat{Q}^f$  diverges to result in  $\hat{W}_f \xrightarrow{p} 1$  and  $\hat{W}_g\hat{W}_f \xrightarrow{p} 0$ . When  $s < 2s$ ,  $n\hat{Q}^f \rightarrow 0$ , and consequently  $\hat{W}_f \xrightarrow{p} 0$  and  $\hat{W}_f\hat{W}_g \xrightarrow{p} 0$ . When  $s = 2s$ , however, (18) shows that  $n\hat{Q}^f \xrightarrow{p} ! \frac{0}{f} f! f$  because only the last term of (18) matters, so that  $\hat{W}_f \xrightarrow{p} W_f = 1 - (! \frac{0}{f} f! f + 1)^{-1}$  and  $\hat{W}_g\hat{W}_f \xrightarrow{p} 0$ .

Case 5). Suppose that model G is locally misspecified with  $g = \frac{g}{0} + g n^s$ , but model H is correct. Then essentially the same arguments as in Case 4) apply.

Case 5-1). If  $s = 1=2$ , then  $n\hat{Q}^g \xrightarrow{d} \frac{2}{k_g} (! \frac{0}{g} g! g) = k_g$  and  $n\hat{Q}^f \xrightarrow{d} \frac{2}{k_f} (! \frac{0}{f} f! f) = k_f$ . Thus,  $\hat{W}_f \xrightarrow{p} 0$  and  $\hat{W}_g\hat{W}_f \xrightarrow{p} 0$ .

Case 5-2). If  $s > 1=2$ , then  $n\hat{Q}^g \xrightarrow{d} \frac{2}{k_g} = k_g$  and  $n\hat{Q}^f \xrightarrow{d} \frac{2}{k_f} = k_f$ .



Case 4). Suppose that  $G$  is correct, but  $H$  is locally misspecified with  $h = h_0 + h_1 n^s$ . By Theorem 9.1 of Newey and McFadden (1994), still  $\hat{g} \xrightarrow{P} g$ ,  $f^h \xrightarrow{P} f$  and  $f^{\hat{g}} \xrightarrow{P} f$ . By Lemma 2, if  $s = 1/2$ , then  $\hat{W}_f \xrightarrow{P} 0$  and  $\hat{W}_g \hat{W}_f \xrightarrow{P} 0$ , and the consistency of  $\hat{g}$  in (1) follows from consistency of  $\hat{f}$ . If  $s < 1/2$ , the probability limits of  $\hat{W}_f$  and  $\hat{W}_g \hat{W}_f$  depend on the relationship between  $\sigma$  and  $s$ . If  $s < 1/2$  and  $\sigma < 2s$ , the limits are the same as in Case 1.

By Assumption A12 and A13,  $\sqrt{n}(\hat{\beta}_h - \beta_0(h))$  and  $\sqrt{n}(\hat{\beta}_h - \beta_0(h))$  are bounded in probability. Given  $\hat{\beta}_h \xrightarrow{p} \beta_0(h)$ , the last two terms in  $b_h$  converge to zero because  $\sqrt{n} \xrightarrow{p} 0$  as  $n \rightarrow \infty$  for  $s > 0$ . Therefore,

$$\sqrt{n}(\hat{\beta}_h - \beta_0(h)) = \sqrt{n}(\hat{\beta}_h - \beta_0(h)) + o_p(1)$$

By Assumption A7, A9, A10, A11, and the consistency of  $\hat{\beta}_h$  for  $\beta_0$ ,  $\sqrt{n}(\hat{\beta}_h - \beta_0(h)) \xrightarrow{d} N(0; \Sigma_h)$  where  $\Sigma_h = \text{Var}[H(Z; \beta_0(h); h)]$ ,  $\sqrt{n}(\hat{\beta}_h - \beta_0(h)) \xrightarrow{p} \beta_0(h)$ ,  $\sqrt{n}(\hat{\beta}_h - \beta_0(h)) \xrightarrow{p} \beta_0(h)$ , and  $\hat{H}^h \xrightarrow{p} H^h$  which is non-singular by Assumption A8. Thus, by the continuous mapping theorem, we get

$$\sqrt{n}(\hat{\beta}_h - \beta_0(h)) \xrightarrow{d} N(0; \Sigma_h)$$

where  $\Sigma_h$  is the same asymptotic variance as in Case 3) of Theorem 2 as if model were correct. Analogously, the same argument holds for  $\sqrt{n}(\hat{\beta}_f - \beta_0(f)) \xrightarrow{d} N(0; \Sigma_f)$ . Hence, all of  $\sqrt{n}(\hat{\beta}_g - \beta_0(g))$ ,  $\sqrt{n}(\hat{\beta}_h - \beta_0(h))$  and  $\sqrt{n}(\hat{\beta}_f - \beta_0(f))$  in the first line of (19) are asymptotically normal with mean zero and variance being that of the corresponding GMM estimator under correct specification.

Recall (19):

$$\sqrt{n}(\hat{\beta}_0 - \beta_0) = \sqrt{n}(\hat{\beta}_g - \beta_0(g)) + \sqrt{n}(\hat{\beta}_h - \beta_0(h)) + \sqrt{n}(\hat{\beta}_f - \beta_0(f)) + o_p(n^{-1/2}) + o_p(n^{-1/2})$$

Recalling (14) and its "squared version", we have

$$\sqrt{n}(\hat{\beta}_g - \beta_0(g)) = O_p(n^{-1/2}) \text{ and } \sqrt{n}(\hat{\beta}_f - \beta_0(f)) = O_p(n^{-1/2}) \Rightarrow \sqrt{n}(\hat{\beta}_0 - \beta_0) = O_p(n^{-1/2}) = O_p(n^{-1/2})$$

Consequently, for the last two terms in (19), we get

$$\begin{aligned} \sqrt{n}(\hat{\beta}_g - \beta_0(g)) + \sqrt{n}(\hat{\beta}_f - \beta_0(f)) &= \frac{1}{\sqrt{n}(\hat{\beta}_g - \beta_0(g)) + \sqrt{n}(\hat{\beta}_f - \beta_0(f))} \left( \frac{\sqrt{n}(\hat{\beta}_g - \beta_0(g))}{\sqrt{n}(\hat{\beta}_g - \beta_0(g)) + \sqrt{n}(\hat{\beta}_f - \beta_0(f))} + \frac{\sqrt{n}(\hat{\beta}_f - \beta_0(f))}{\sqrt{n}(\hat{\beta}_g - \beta_0(g)) + \sqrt{n}(\hat{\beta}_f - \beta_0(f))} \right) \\ &= \frac{1}{O_p(n^{-1/2}) + 1} \left( \frac{O_p(1)O(n^{-1/2})}{O_p(1) + O_p(n^{-1/2})} + \frac{1}{O_p(n^{-1/2}) + 1} \right) O(n) \end{aligned}$$

Case 4-2). If  $s > 1=2$ ,  $\hat{W}_f \neq 0$  and  $(1 - \hat{W}_f) f n^{1-2s} \neq 0$  as  $n \neq 1$

Under Assumption A7 and A9, the following first-order conditions hold:

$$FD^f = \frac{\partial Q^f(\lambda_f)}{\partial \lambda_f} = r \cdot \lambda_f^{\alpha_f} \lambda_f^{\beta_f} = 0; \quad FD^f = \frac{\partial Q^f(\lambda_f)}{\partial \lambda_f} = r \cdot \lambda_f^{\alpha_f} \lambda_f^{\beta_f} = 0;$$

$$FD^f = \frac{\partial Q^f(\lambda_f)}{\partial \lambda_f} = r \cdot \lambda_f^{\alpha_f} \lambda_f^{\beta_f} = 0:$$

Expand  $\lambda_f$  around the unique minimizer  $\lambda_f^*$ ;  $\lambda_f^*$ ;  $\lambda_f^*$  to get

$$\lambda_f(\lambda_f) = \lambda_f^* + r \cdot \lambda_f^{\alpha_f}(\lambda_f - \lambda_f^*) + r \cdot \lambda_f^{\alpha_f}(\lambda_f - \lambda_f^*) + r \cdot \lambda_f^{\alpha_f}(\lambda_f - \lambda_f^*);$$

where  $\lambda_f^*$  is the mean value to apply the mean value theorem. Substitute these into each  $FD^f$  to get

$$FD^f = r \cdot \lambda_f^{\alpha_f} \lambda_f^{\beta_f} + r \cdot \lambda_f^{\alpha_f}(\lambda_f - \lambda_f^*) + r \cdot \lambda_f^{\alpha_f}(\lambda_f - \lambda_f^*) + r \cdot \lambda_f^{\alpha_f}(\lambda_f - \lambda_f^*);$$

$$FD^f = r \cdot \lambda_f^{\alpha_f} \lambda_f^{\beta_f} + r \cdot \lambda_f^{\alpha_f}(\lambda_f - \lambda_f^*) + r \cdot \lambda_f^{\alpha_f}(\lambda_f - \lambda_f^*) + r \cdot \lambda_f^{\alpha_f}(\lambda_f - \lambda_f^*);$$

$$FD^f = r \cdot \lambda_f^{\alpha_f} \lambda_f^{\beta_f} + r \cdot \lambda_f^{\alpha_f}(\lambda_f - \lambda_f^*) + r \cdot \lambda_f^{\alpha_f}(\lambda_f - \lambda_f^*) + r \cdot \lambda_f^{\alpha_f}(\lambda_f - \lambda_f^*);$$

$$FD^f = fFD^f; FD^f; FD^f g = \lambda_f^* + \lambda_f^{\alpha_f}(\lambda_f - \lambda_f^*), \text{ and from these, } \lambda_f^* = \lambda_f^* + \lambda_f^{\alpha_f}(\lambda_f - \lambda_f^*);$$

$$\lambda_f^* = \lambda_f^* + \lambda_f^{\alpha_f}(\lambda_f - \lambda_f^*); \lambda_f^* = \lambda_f^* + \lambda_f^{\alpha_f}(\lambda_f - \lambda_f^*); \lambda_f^* = \lambda_f^* + \lambda_f^{\alpha_f}(\lambda_f - \lambda_f^*);$$

In this expression for  $\lambda_f^*$ , examine the part for  $\lambda_f^*$ , i.e., the first component:

$$\lambda_f^* = \lambda_f^* + \lambda_f^{\alpha_f}(\lambda_f - \lambda_f^*); \text{ where } \lambda_f^* = \lambda_f^* + \lambda_f^{\alpha_f}(\lambda_f - \lambda_f^*); \lambda_f^* = \lambda_f^* + \lambda_f^{\alpha_f}(\lambda_f - \lambda_f^*);$$

(22)

$$\lambda_f^* = \lambda_f^* + \lambda_f^{\alpha_f}(\lambda_f - \lambda_f^*); \lambda_f^* = \lambda_f^* + \lambda_f^{\alpha_f}(\lambda_f - \lambda_f^*); \lambda_f^* = \lambda_f^* + \lambda_f^{\alpha_f}(\lambda_f - \lambda_f^*);$$

Then, we have

$$\lambda_f^* = \lambda_f^* + \lambda_f^{\alpha_f}(\lambda_f - \lambda_f^*); \lambda_f^* = \lambda_f^* + \lambda_f^{\alpha_f}(\lambda_f - \lambda_f^*); \lambda_f^* = \lambda_f^* + \lambda_f^{\alpha_f}(\lambda_f - \lambda_f^*);$$

(23)

and  $\lambda_f^*$  is the influence function of  $\lambda_f^*$ .

TT d a(fff) 98 123 \$. Q TE 5 2904

Expand  $g$  around the unique minimizer  $g^*$  to get

$$g(\hat{g}) = g(g^*) + \frac{1}{2} g''(\bar{g})(\hat{g} - g^*)^2 + \frac{1}{6} g'''(\bar{g})(\hat{g} - g^*)^3$$

where  $\bar{g}$  is the value for the mean value theorem. Substitute these into each  $FD^g$  to get

$$FD^g = \frac{1}{2} g''(\bar{g})(\hat{g} - g^*)^2 + \frac{1}{6} g'''(\bar{g})(\hat{g} - g^*)^3;$$

$$FD^g = \frac{1}{2} g''(\bar{g})(\hat{g} - g^*)^2 + \frac{1}{6} g'''(\bar{g})(\hat{g} - g^*)^3;$$

$$FD^g = fFD$$

an outcome,  $T$  is a binary treatment indicator, and  $X$  is a  $J$  vector of other covariates (including

Observe that if  $h(X; \theta) = E(T|X)$ , then the first two terms in the above expectation equal equation (27) and the second two terms have mean zero. By rearranging terms, equation (30) can be rewritten as

$$= E \left[ \psi(1; X; \theta) \psi(0; X; \theta) + \frac{T}{h(X; \theta)} f_Y \psi(1; X; \theta) g - \frac{1-T}{1-h(X; \theta)} f_Y \psi(0; X; \theta) g \right] \quad (31)$$

Rewriting the equation this way, it can be seen that if  $\psi(T; X; \theta) = E(Y|T; X)$ , then the first two terms in equation (31) equal equation (26), and the second two terms have mean zero. This shows that equation (30) or equivalently (31) is doubly robust, in that it equals the average treatment effect if either  $\psi(T; X; \theta)$  or  $h(X; \theta)$  is correctly specified. The GMM estimator associated with this doubly robust estimator estimates  $\theta$ ,  $\tau$ , and  $\sigma^2$ , using the moments

$$E \left[ \frac{1}{n} \sum_{i=1}^n \begin{pmatrix} Y_i - \psi(T_i; X_i; \theta) \\ T_i - h(X_i; \theta) \end{pmatrix} \right] = \begin{pmatrix} \tau - \psi_1(\theta) \\ \sigma^2 - \psi_2(\theta) \end{pmatrix}$$

Okui, Small, Tan, and Robins (2012) propose a DR estimator for an instrumental variables (IV) additive regression model. The model is the additive regression

$$Y = M(W; \theta) + \mathbb{G}(X) + U; \tag{35}$$

$$\begin{aligned} E(Q | X) &= \mathbb{H}(X); \\ E(U | X; Q) &= 0; \end{aligned} \tag{36}$$

where  $Y$  is an observed outcome variable,  $W$  is a  $S$  vector of observed exogenous covariates,  $X$  is a  $J$  vector of observed confounders, and  $Q$  is a  $K \times S$  vector of observed instruments. Note that this model has features that are unusual for instrumental variables estimation, in particular, the assumption that  $E(U | X; Q) = 0$  is stronger than the usual  $E(U | Q) = 0$  assumption. The function  $M(W; \theta)$  is assumed to be correctly parameterized, and the goal is estimation of

Okui, Small, Tan, and Robins (2012) construct a DR estimator assuming that, in addition to the above, either  $\mathbb{G}(X) = \mathbb{G}(X; \theta)$  is correctly parameterized, or that  $\mathbb{H}(X) = \mathbb{H}(X; \theta)$  is correctly parameterized. Let  $Z = f(Y; W; X; Q)g$ , and let  $r_1(X)$  and  $r_2(X)$  be vectors of functions chosen by the user. Define  $G(\theta; Z)$  and  $H(\theta; Z)$  by

$$G(\theta; Z) = \frac{f(Y; W; X; Q) \mathbb{G}(X; \theta) r_1(X)}{f(Y; W; X; Q) \mathbb{G}(X; \theta) g} \tag{37}$$

and

$$H(\theta; Z) = \frac{f(Q; X) \mathbb{H}(X; \theta) r_2(X)}{f(Y; W; X; Q) \mathbb{H}(X; \theta) g} \tag{38}$$

Okui, Small, Tan, and Robins (2012) take  $r_1(X) = \mathbb{G}(X; \theta) = \theta$  and  $r_2(X) = \mathbb{H}(X; \theta) = \theta$ . If  $\mathbb{G}(X; \theta)$  is correctly specified, then  $EfG(\theta; Z)g = 0$ , while if  $\mathbb{H}(X; \theta)$  is correctly specified, then  $EfH(\theta; Z)g = 0$ .



Table 6-1. Model G is Correctly Specified and Model H is Misspecified (n = 500)

$\beta_1$	Bias	Mde	RMSE	MAE	SD	Skew	Kurt	Freq	SE	SD <sub>SE</sub>
s=0.25										
GMM <sub>g</sub>	0.0002	0.0006	0.0001	0.0075	0.0111	0.2310	3.1966	0.9465	0.0108	0.0011
GMM <sub>h</sub>	0.2374	0.2367	0.0566	0.2367	0.0157	0.1558	3.1392	0.0000	0.0139	0.0016
GMM <sub>f</sub>	0.1094	0.1094	0.0121	0.1094	0.0112	0.0817	3.0557	0.0000	0.0068	0.0005
ODR <sub>1</sub>	0.0002	0.0006	0.0001	0.0075	0.0111	0.2311	3.1963	0.9460	0.0108	0.0011
s=0.5										
GMM <sub>g</sub>	0.0002	0.0006	0.0001	0.0075	0.0110	0.1255	3.081			

---

---

Table 7-1. Model G is Misspeci...ed withs = 0:75 and Model H is Misspeci...edr( = 500)

---

1	Bias	Mde	RMSE	MAE	SD	Skew	Kurt	Freq	SE	SD
---	------	-----	------	-----	----	------	------	------	----	----