Weak Instruments Test in Discrete Choice Models

David T. Frazier* Donald Poskitt† Eric Renault‡ Lina Zhang§ Xueyan Zhao¶

February 5, 2019

Abstract

This paper proposes a test to consistently detect weak instruments in discrete choice models. As a by-product of our testing approach, we construct a generalized 'concentration parameter' that allows us to generalize the standard 'rule-of-thumb' for linear models to discrete choice models. This generalized concentration parameter provides insights regarding instrument strength in a host of discrete choice models. A Monte Carlo analysis compares our proposed testing approach against commonly applied weak instrument tests. The results simultaneously demonstrate the good performance of our approach and the fundamental failure of conventionally applied, i.e., linear, weak instruments tests in this context. We compare our testing approach to those commonly applied in the literature within two empirical examples: married women labor force participation, and US food aid and civil conflicts.

Keywords: Discrete Choice Models; (Nearly) Weak Instruments

Preliminary and Incomplete: Do Not Cite

*Department of Econometrics and Business Statistics, Monash University (david.frazier@monash.edu).
†Department of Econometrics and Business Statistics, Monash University (donald.poskitt@monash.edu).
‡Department of Economics, Brown University (eric.renault@brown.edu).
§Department of Econometrics and Business Statistics, Monash University (lina.zhang@monash.edu).
¶Department of Econometrics and Business Statistics, Monash University (xueyan.zhao@monash.edu).

1 Introduction

Endogeneity of regressors is a common problem faced by economists hoping to establish causal effects, and the most common solution to this issue is the use of instrumental variables (IV). However, it is now well-known that the very features that render IV useful in estimating causal effects, namely the exogeneity of the instruments with respect to the model error term, can lead to an instrument that has little covariation with the endogenous regressors. As a result, the use of so-called 'weak' instruments in the empirical literature is a commonly encountered issue.

The resulting behavior of IV parameter estimates in the presence of weak instruments has been extensively studied in the linear regression model, and we refer the reader to Stock and Yogo (2005) for a survey of results. However, we highlight here the contribution of Staiger and Stock (1997) in developing a weak instrument asymptotic regime based on a drifting data generating process (DGP) that captures the instrument strength. Under this asymptotic framework Staiger and Stock (1997) demonstrate the serious consequences of using weak IV in practice, including inconsistency of the IV estimator and non-standard behavior of test statistics calculated from this estimate.

Since the initial work on weak IV in the linear model, much progress has been made on the detection of weak instruments. The first such detection mechanism is the well-known 'rule-of-thumb' due to Staiger and Stock (1997), which characterizes the instruments as weak, in the case of one endogenous variable and multiple instruments, if the value of the first stage F-statistic is less than ten. More generally, Stock and Yogo (2005) provide a quantitative definition of weak instruments in the linear model based on the two-stage-least-squares (2SLS) bias relative to ordinary least squares, and use this definition to propose a formal test for instrument weakness. By comparing the forward and reverse 2SLS estimators, Hahn and Hausman (2002) introduce a test of the null that instruments are strong against the alternative that they are weak. Poskitt and Skeels (2009) directly test the null of a small concentration parameter and base their test on a monotonic transformation of the likelihood ratio (LR) statistic. While the above tests are formulated under conditionally homoscedastic and serially uncorrelated model errors, an extension of the Stock and Yogo (2005) testing strategy to heteroskedastic and serially correlated errors is devised in Olea and Pflueger (2013).

Subsequent to the analysis of weak IV in the linear model, the concept of a drifting DGP has been used to examine the properties of weak IV in the context of generalized method of moments (GMM) estimation, where we refer the reader to Stock et al. (2002) for a survey of results. In such cases, the weakness of instruments is measured by the speed at which the moments, or a portion thereof, become "flat", in the sense that the moments are close to zero for all values in the parameter space (Stock and Wright, 2000). In the canonical case of genuinely weak instruments and separable moments, weakness is captured by ensuring that the moments become flat at a rate that is proportional to the square root of the sample size. Subsequently, this framework has been extended to consider non-separable moments and multiple rates of weakness (see, e.g., Hahn and Kuersteiner, 2002, Caner, 2009, Antoine and Renault, 2009, Antoine and Renault, 2012 and Andrews and Cheng, 2012).

This paper builds on existing weak IV analysis by considering the behavior of GMM estimators in endogenous discrete choice models with weak instruments. Discrete choice models represent an "intermediate" between the linear model and the most general GMM context in that if we could view the latent choice utilities the resulting moments used for estimation would be linear (in the parameters), but since these utilities are not observable to the econometrician we must consider

a vector of non-separable moments to estimate the underlying structural parameters. Herein, using the drifting DGP concept in Staiger and Stock (1997) to capture instrument weakness, we develop the theoretical properties of GMM estimators for discrete choice models under instrument weakness of varying degrees. In particular, we demonstrate that if the rate of instrument weakness, as captured by the coefficient magnitude in the first-stage regression, deviates from the standard weak instrument level, i.e., the square root of the sample size, GMM estimators of discrete choice models deliver consistent estimators of the structural parameters. Using this result, we construct a consistent, size-controlled test for weak instruments that is applicable to a large class of discrete choice models.

While the existence of weak IV is a common phenomena, there is little theoretical evidence regarding the properties of GMM estimators in endogenous discrete choice models. Using Monte Carlo simulations, Dufour and Wilde (2013) demonstrate the poor behavior of Wald and LR tests in the presence of weak instrument. Magnusson (2007) considers the Wald test and the distance metric test for the probit model and finds that, with weak instruments, both tests over-reject the null hypothesis (the truth) even when the concentration parameter is larger than ten. Magnusson (2010) develops weak IV robust inference, which, however, is not straightforward compared to the widespread two-stage decision rule: a pretest for weak IV followed by standard inference procedures.

The development of a consistent weak instrument test for discrete choice models is particularly important since the similarity between the linear model and common discrete choice models, such as the probit model, have led researcher to apply tests that are appropriate for linear models to this nonlinear context. In particular, it is relatively common to apply the rule-of-thumb developed by Staiger and Stock (1997) in the linear context to detect the present of weak instruments in discrete choice models: see, e.g., Miguel et al. (2004), Arendt (2005), McKenzie and Rapoport (2011), Cawley and Meyerhoefer (2012), Block et al. (2013) and Goto and Iizuka (2016). However, the above studies give no discussion on the potential validity, or invalidity, of this rule-or-thumb in discrete choice models. That being said, others have realized the potential inappropriate use of the rule-of-thumb in this setting. In these cases, researchers often abandon the discrete choice framework in favor of the linear probability models (LPM), and assume that standard weak IV detection procedures remain valid for the LPM; see, e.g., Lochner and Moretti (2004), Powell et al. (2005), Kinda (2010), Ruseski et al. (2014). As we will see later, when the true model is an endogenous discrete choice model with weak instruments, the application of an LPM and weak instruments detection procedures will lead to misleading conclusions regarding instrument strength, and subsequently parameter inference.

This paper is generally related to the literature on identification failure within a nonlinear, and non-separable GMM context, see, e.g., Andrews and Cheng (2012), and closely relates to the sequence of papers Antoine and Renault (2009, 2012, 2017), which explore the behavior of GMM estimators under various degrees of identification failure. In the latter, the authors conceptualize identification failure using a drifting DGP that captures rank deficiency of the limit Jacobian for the moment conditions. Antoine and Renault (2009, 2012) demonstrate that if the rate at which rank deficiency occurs is slower than the square root of the sample size, any GMM estimator is consistent. Furthermore, if the rank deficiency rate is slower than the fourth root of the sample size, the asymptotic normality of the two-step GMM estimator is maintained. Using these results Antoine and Renault (2017) (AR hereafter) develop a general testing strategy for weak identification of models with nonlinear and non-separable moments. AR prove that this testing strategy can consistently detect weak moments so long as the rank deficiency of the Jacobian occurs at a

rate that is slower than the fourth root of the sample size.

Herein, we use the testing setup considered in AR to deduce a test for weak IV in discrete choice models. However, while the setup considered in this paper is similar to AR, our testing approach differs in a crucial aspect. Due to the necessary generality of the AR testing setup, the authors can only reject the null hypothesis of weak instruments if the rank deficiency in the Jacobian occurs at a rate that is slower than the fourth root of the sample size. As such, rejection of this null only means we can conduct Gaussian inference with nonstandard convergence rates. In contrast, we construct a version of the AR test that allows us to detect actual instrument weakness, i.e., the test we construct allows us to detect weakness if the strength of the IV in the first-stage regression goes to zero slower than the square root of the sample size.

In addition to presenting a test for weak IV in discrete choice models, our analysis sheds new light on the application of the popular rule-of-thumb in discrete choice models. In particular, our test can be seen as a generalization of the standard first-stage F-test and, as such, allows us to measure the genuine strength of the instruments. In contrast, due to the nonlinearity of the moments used to estimate endogenous discrete choice models, we demonstrate that the standard rule-of-thumb for linear IV models does not adequately capture the strength of instruments.

Using Monte Carlo examples, we give a rigorous comparison on the performance of our testing approach against weak IV tests commonly applied in discrete choice models. The results demonstrate that applying the linear model rule-of-thumb, or the test of Stock and Yogo (2005), to discrete choice models can yield misleading results. In addition, these results demonstrate that if we use an LPM in place of a probit/logit model, and if the moments are weak, these standard approaches to testing for weak IV in linear models also give misleading results. In contrast, we demonstrate that our weak IV test performs well and controls size across a wide range of simulation experiments.

Lastly, we apply our weak IV testing approach in two well-known empirical examples, and compare the findings of our approach against existing tests for linear models that are commonly applied in this discrete choice model setting. The first example examines the causal effects of the education level of married women on their labor force participation, using the data from the University of Michigan Panel Study of Income Dynamics in 1975. The second example considers the impact of US food aid on the incidence of civil wars in 78 non-OECD recipient countries during 1971 to 2006, as studied in Nunn and Qian (2014). Using their identification strategy, the endogeneity issue of US food aid can be overcome by using as an instrument the product of the lagged US wheat production and the average probability of receiving any US food aid. Within the confines of the first example, all testing approaches agree that the instruments are not weak. However, the second example yields mixed results: our testing approach fails to reject the null of weak instruments, while all three conventional tests reject the hypothesis of instrument weakness. Since our approach is the only test with theoretical guarantees, this is a clear empirical example where researchers would have been misled if they had applied conventional weak IV tests.

The remainder of the paper is organized as follows. Section 2 introduces our model setup and assumptions. Section 3 formulates the null hypothesis, the test statistic, and details our testing procedure. A generalization of the linear rule-of-thumb to discrete choice models is given in Section 4, and allows us to discuss the connection of weak IV tests in the linear and discrete choice models. Monte Carlo simulations in Section 5 verify the asymptotic properties of our proposed test as well as the performance of other weak IV tests. Section 6 applies our weak IV test to two empirical examples: Wooldridge (2010) married women labor force participation, and Nunn and Qian (2014) US food aid and civil conflicts.

2 Model and Assumptions

Consider the standard probit model[1]

$$y_{1,i} = \begin{cases} 1 & \text{if } y^*_{1,i} > 0 \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

with latent variable

$$y^*_{1,i} = \alpha y_{2,i} + x_i'\beta + u_i, \quad \text{for } i = 1, 2, ..., n$$

where x_i is a $k \times 1$ vector of exogenous regressors that is uncorrelated with u_i, $y_{2,i}$ is a continuous scalar regressor that is endogenous. We also assume there is a exogenous instrument z_i satisfying $\mathbb{E}[u_i|z_i, x_i] = 0$. We formalize the model (1) through the following assumptions:

(A0) $\{s_i\}_{i=1}^n = \{y_{1,i}, y_{2,i}, x_i', z_i\}_{i=1}^n$ is i.i.d. and $\mathbb{E}[\|s_i\|^{2+\kappa}] < \infty$ for some positive κ.[2]

(A1) Reduced form: $\mathbb{E}[y_{2,i}|x_i, z_i] = x_i'\pi + \xi z_i$.

(A2) "Semi-Strong" validity of instruments: $\mathbb{E}[u_i|x_i, z_i] = 0$.

(A3) Joint normality: denote $v_i = y_{2,i} - \mathbb{E}[y_{2,i}|x_i, z_i]$

$$\begin{pmatrix} u_i \\ v_i \end{pmatrix} \Bigg| \begin{pmatrix} x_i \\ z_i \end{pmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_u^2(x_i, z_i) & \sigma_v(x_i, z_i)\sigma_u(x_i, z_i)\rho(x_i, z_i) \\ \sigma_v(x_i, z_i)\sigma_u(x_i, z_i)\rho(x_i, z_i) & \sigma_v^2(x_i, z_i) \end{bmatrix} \right).$$

Thanks to the parametric structure of assumption (A3),

$$\begin{aligned} \mathbb{E}[y_{1,i}|y_{2,i}, x_i, z_i] &= \Pr\left[u_i > -\alpha y_{2,i} - x_i'\beta | y_{2,i}, x_i, z_i\right] \\ &= \Pr\left\{u_i - \mathbb{E}[u_i|y_{2,i}, x_i, z_i] > -\alpha y_{2,i} - x_i'\beta - \mathbb{E}[u_i|y_{2,i}, x_i, z_i] | y_{2,i}, x_i, z_i\right\} \\ &= \Phi\left(\frac{\alpha y_{2,i} + x_i'\beta + \mathbb{E}[u_i|y_{2,i}, x_i, z_i]}{\mathrm{Var}[u_i|y_{2,i}, x_i, z_i]^{1/2}} \right), \end{aligned} \tag{2}$$

where $\Phi(\cdot)$ denotes the normal cumulative distribution function. Based on (A1) and (A3),

$$\begin{aligned} \mathbb{E}[u_i|y_{2,i}, x_i, z_i] &= \frac{\rho(x_i, z_i)\sigma_u(x_i, z_i)}{\sigma_v(x_i, z_i)} \{y_{2,i} - \mathbb{E}[y_{2,i}|x_i, z_i]\} = \widetilde{\rho}(x_i, z_i)\{y_{2,i} - x_i'\pi - \xi z_i\} \\ \mathrm{Var}[u_i|y_{2,i}, x_i, z_i] &= \sigma_u^2(x_i, z_i)[1 - \rho^2(x_i, z_i)] \end{aligned}$$

where we denote $\widetilde{\rho}(x_i, z_i) \equiv \frac{\rho(x_i, z_i)\sigma_u(x_i, z_i)}{\sigma_v(x_i, z_i)}$. As is true in all probit models, we require a normalization to identify the parameters. Above derivations demonstrate that, in contrast to the often used normalization, i.e., that the unconditional variance of u_i is unity ($\mathbb{E}[\sigma_u^2(x_i, z_i)] = 1$), all we actually require is $\mathrm{Var}[u_i|y_{2,i}z_i, x_i] = 1$, i.e., the conditional variance should be unity. Imposing this normalization then gives

$$\sigma_u(x_i, z_i) = \frac{1}{\sqrt{1 - \rho^2(x_i, z_i)}}, \quad \text{and} \quad \widetilde{\rho}(x_i, z_i) = \frac{\rho(x_i, z_i)}{\sigma_v(x_i, z_i)\sqrt{1 - \rho^2(x_i, z_i)}}. \tag{3}$$

[1]The results of this paper are applicable to more general discrete choice models, but we focus on the probit model for clarity.

[2]The i.i.d assumption can be relaxed under additional conditions and is maintained for simplicity.

By the normalization of $\text{Var}[u_i|y_{2,i}z_i, x_i] = 1$, the conditional moment (2) becomes to

$$\mathbb{E}[y_{1,i}|y_{2,i}, x_i, z_i] = \Phi\Big([\alpha + \widetilde{\rho}(x_i, z_i)]y_{2,i} + x_i'[\beta - \widetilde{\rho}(x_i, z_i)\pi] - \widetilde{\rho}(x_i, z_i)\xi z_i\Big) \tag{4}$$

where $\widetilde{\rho}(x_i, z_i)$, expressed in (3), is a strictly increasing function of $\rho(x_i, z_i)$.

Remark 1 *Recall the definition $v_i = y_{2,i} - \mathbb{E}[y_{2,i}|x_i, z_i]$. Blundell and Powell (2004), pg 657, note that maximum likelihood estimation (MLE) can proceed if we assume joint normality of (u_i, v_i) and its independence of $(x_i', z_i)'$. However, the requirement of independence is partly redundant after imposing the normalization condition. In particular, applying the law of total conditional variance, we obtain*

$$\begin{aligned} \text{Var}\left[u_i|x_i, z_i\right] &= \mathbb{E}\left(\text{Var}\left[u_i|x_i, z_i, y_{2,i}\right]\big|x_i, z_i\right) + \text{Var}\left[\mathbb{E}\left(u_i|x_i, z_i, y_{2,i}\right)\big|x_i, z_i\right] \\ &= 1 + \widetilde{\rho}^2(x_i, z_i)\text{Var}[v_i|x_i, z_i] = 1 + \rho^2(x_i, z_i)\text{Var}\left[u_i|x_i, z_i\right]. \end{aligned}$$

It is easy to see that all we actually require is that $\sigma_u^2(x_i, z_i)$ and $\sigma_v^2(x_i, z_i)$ do not depend on $(x_i', z_i)'$. Therefore, in order to implement MLE, all we actually require are the often used normalization and homoscedasticity

$$\begin{pmatrix} u_i \\ v_i \end{pmatrix} \Bigg| \begin{pmatrix} x_i \\ z_i \end{pmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \sigma_v\rho \\ \sigma_v\rho & \sigma_v^2 \end{bmatrix}\right).$$

However, the homoscedasticity assumption is quite strong because heteroscedastic errors are of great practical relevance in empirical studies, for instance when $y_{2,i}$ has limited support, that assumption might be invalid. Thus, to relax the homoscedasticity of $y_{2,i}$, we can instead consider GMM with heteroscedastic consistent weighting matrix, which could be based on the following orthogonality conditions constructed from reduced form (A1) and conditional moment (2) at the true unknown parameter value $(\alpha^0, \widetilde{\rho}^0, \beta^0, \pi^0, \xi^0)$

$$0 = \mathbb{E}\begin{bmatrix} a(y_{2,i}, x_i, z_i)r_{1,i}(\alpha^0, \widetilde{\rho}^0, \beta^0, \pi^0, \xi^0) \\ b(x_i, z_i)r_{2,i}(\xi^0, \pi^0) \end{bmatrix} \tag{5}$$

and

$$\begin{aligned} r_{1,i}(\alpha, \widetilde{\rho}, \beta, \pi, \xi) &= y_{1,i} - \Phi\left([\alpha + \widetilde{\rho}]y_{2,i} + x_i'[\beta - \widetilde{\rho}\pi] - \widetilde{\rho}\xi z_i\right) \\ r_{2,i}(\xi, \pi) &= y_{2,i} - x_i'\pi - \xi z_i, \end{aligned}$$

where for illustration simplicity, we restrict $\widetilde{\rho}(x_i, z_i) = \widetilde{\rho}$ to be invariant with respect to $(x_i', z_i)'$.[3] We use $a(\cdot)$ and $b(\cdot)$ to denote measurable vector functions that are orthogonal to $r_{1,i}(\alpha^0, \widetilde{\rho}^0, \beta^0, \pi^0)$ and $r_{2,i}(\xi^0, \pi^0)$ respectively. Then $\sigma_v(x_i, z_i)$ can be estimated nonparametrically. For example, a

[3]This restriction means the ratio $\rho(x_i, z_i)/[\sigma_v(x_i, z_i)\sqrt{1 - \rho^2(x_i, z_i)}]$ is fixed but $\sigma_v(x_i, z_i)$ is not necessarily homoscedastic. Moreover, noticing that the conditional moment in (2) does not require the normality of v_i or even of u_i, the only thing that (A3) buys us is the structure of (2). Thus, we can further relax the joint normality assumption as long as the orthogonality conditions (5) still hold for some proper distributional assumption. The restrictions imposed here are quite general and satisfied by most error specifications that one would consider in practice, e.g., the logistic distribution $F(x) = e^x/(1 + e^x)$ as well as commonly encountered multinomial choice specifications can all be treated in a similar fashion. Therefore, the weak IV test considered in this paper is applicable to general discrete models.

possible approach is the optimal instruments GMM estimator proposed by Kawaguchi et al. (2017), which estimates the heteroscedastic variance term using a k-nearest neighbor estimation procedure.

However, for simplicity, throughout we consider that $\sigma_v(x_i, z_i) = \sigma_v$ for all $i = 1, \ldots, n$. We then consider the following simplifying assumption on the parameter space:

(A4) Let $\widetilde{\rho} = \rho/\sigma_v\sqrt{1-\rho^2}$ and $\theta = (\widetilde{\rho}, \alpha, \beta', \pi', \xi)'$. The parameter space $\Theta \subset \mathbb{R}^{d_\theta}$ is compact.

Assumptions (A1)-(A3) yield two restrictions that we will use for inference on θ:

$$-r_i(\theta) = \begin{bmatrix} -r_{1,i}(\theta) \\ -r_{2,i}(\theta_2) \end{bmatrix} = \begin{bmatrix} y_{1,i} - \Phi\left([\alpha + \widetilde{\rho}]y_{2,i} + x_i'[\beta - \widetilde{\rho}\pi] - \widetilde{\rho}\xi z_i\right) \\ y_{2,i} - x_i'\pi - \xi z_i. \end{bmatrix}. \tag{6}$$

Define $a_i = a(y_{2,i}, x_i, z_i)$ and $b_i = b(x_i, z_i)$ to be $H \times 1$ measurable vector functions that are orthogonal to $r_{1,i}(\theta)$ and $r_{2,i}(\theta_2)$, respectively. Then, we can define the moment function $g_i(\theta)$ of size $H > d_\theta$ as

$$g_i(\theta) = -[a_i, b_i]r_i(\theta) = a_i\left[y_{1,i} - \Phi\left([\alpha + \widetilde{\rho}]y_{2,i} + x_i'[\beta - \widetilde{\rho}\pi] - \widetilde{\rho}\xi z_i\right)\right] + b_i\left[y_{2,i} - x_i'\pi - \xi z_i\right]. \tag{7}$$

Using the moment function $g_i(\theta)$ and the sample mean $\bar{g}_n(\theta) = \frac{1}{n}\sum_{i=1}^n g_i(\theta)$, GMM estimators of θ can be obtained by minimizing a weighted quadratic form in $\bar{g}_n(\theta)$. In particular, for Ω an $H \times H$ dimensional positive-definite weighting matrix, a generic GMM estimator can be defined from the minimization program:

$$\min_{\theta \in \Theta} \; n \cdot \bar{g}_n(\theta)'\Omega\bar{g}_n(\theta).$$

In settings where the instruments are potentially weak, a common choice for the weighting matrix of the GMM estimator is

$$\Omega = S_n(\theta)^{-1}, \text{ where } S_n(\theta) := \left\{\frac{1}{n}\sum_{i=1}^n [g_i(\theta) - \bar{g}_n(\theta)][g_i(\theta) - \bar{g}_n(\theta)]'\right\},$$

which yields the so-called continuously updating GMM (CU-GMM) estimator of Hansen et al. (1996):

$$\hat{\theta}_n = \arg\min_{\theta \in \Theta} J_n[\theta], \text{ where } J_n[\theta] := n \cdot \bar{g}_n(\theta)' S_n^{-1}(\theta)\bar{g}_n(\theta).$$

The CU-GMM criterion function, $J_n[\theta]$, is known to have desirable properties that are useful in constructing tests that are robust to weak instruments; see, e.g., Stock and Wright (2000) Kleibergen (2005), Caner (2009) and Antoine and Renault (2017).

3 Testing Weakness: The Probit Model

In linear models, weakness of instruments arise when there is not a sufficient degree of covariation between the instruments and the endogenous regressors. Staiger and Stock (1997) characterize this weakness using a local-to-zero asymptotic regime that allows the instrument strength to degrade to zero as the sample size increases, which can be characterized as $\xi = \gamma/\sqrt{n}$ in model (1) with $\gamma \neq 0$. Using this asymptotic regime, Staiger and Stock (1997) demonstrate that general k-Class estimators, including 2SLS and LIML, are inconsistent and have nonstandard limiting distribution.

In a GMM context, Stock and Wright (2000) consider weak identification for separable moments using a drifting data generating process (DGP): for some $\lambda \geq 0$,

$$\mathbb{E}[\bar{g}_n(\theta)] = m_1(\theta)/n^{\lambda} + m_2(\theta_1), \quad \text{with} \quad \theta = (\theta_1', \theta_2')', \tag{8}$$

for some known m_1 and m_2. Under this assumption, θ_1 is identified and the identification status of θ_2 depends on λ. Stock and Wright (2000) study the case of $\lambda = 1/2$ and prove that the GMM estimator $\hat{\theta}_{2,n}$ is inconsistent. In addition, Antoine and Renault (2009, 2012) explore GMM asymptotics when $1/4 \leq \lambda < 1/2$, dubbed nearly-weak identification, and show that $n^{1/2-\lambda}\|\hat{\theta}_n - \theta^0\| = O_p(1)$ under fairly general assumptions.

For general nonlinear and non-separable moments, Antoine and Renault (2017) (hereafter, AR) test identification strength by representing the rank deficiency rate of the Jacobian as n^{λ} and formulate a test of the hypotheses $H_0: \ \lambda = 1/4$ against $H_1: \ \lambda < 1/4$. In the AR framework, it is necessary to test the range $\lambda \leq 1/4$ in order to control the size of their test statistic. Under the alternative $\lambda < 1/4$, which is exactly the case of nearly-strong identification defined in Antoine and Renault (2009), the convergence rate of the GMM estimator is faster than $n^{1/4}$ and Gaussian asymptotic inference can be conducted.

In contrast to the case of linear or separable moments, weakness of IV in discrete choice models leads to a non-linear pattern of identification strength. As a result, moment conditions in this setting cannot be partitioned according to the identification strength of parameters as (8). Therefore, we follow AR and use the rank deficiency rate of the Jacobian to construct the null hypothesis.

3.1 Null Hypothesis and Testing Framework

3.1.1 Null Hypothesis

Our goal is to construct a test that can consistently detect if instruments are too weak to guarantee consistency of the GMM estimator in discrete choice models.[4] If the proposed test rejects the null hypothesis, we can safely claim that the instruments are not weak and the corresponding GMM estimator is consistent.

To characterize potential weakness of instruments, we follow the local-to-zero approach of Staiger and Stock (1997), and consider that the instrument z_i is actually related to $y_{2,i}$ according to a drifting DGP, where the strength is captured by $\xi = \gamma/n^{\lambda}$, for some $\lambda > 0$ and $\gamma \neq 0$. Define $\theta \in \Theta$ and its partition as $\theta = (\theta_1', \theta_2')'$, where $\theta_1 = (\widetilde{\rho}, \alpha, \beta')'$ and $\theta_2 = (\pi', \xi)'$.

Recall that, under Assumptions (A1)-(A3), we have the available moment function $g_i(\theta)$, defined in equation (7), that we use to form sample moments $\bar{g}_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} g_i(\theta)$. The standard GMM identification condition regarding consistent estimation of θ^0 is generally stated as follows:

$$g(\theta) := \mathbb{E}[\bar{g}_n(\theta)] = 0 \iff \theta = \theta^0 \tag{9}$$

[4] By calculations not reported here, we know that the Jacobian of the MLE for the IV probit model, $(y_{1,i}, y_{2,i})$ given $(x_i', z_i)'$, are equivalent for identification purpose to the sample counterpart of GMM moment conditions. The log-likelihood function is $\log L(\theta) = \sum_{i=1}^{n} y_{1,i} \log \Phi_i(\theta) + (1 - y_{1,i}) \log[1 - \Phi_i(\theta)] - 1/2 \log(\sigma_v^2) - 1/2(y_{2,i} - x_i'\pi - \frac{\gamma}{n^{\lambda}} z_i)^2/\sigma_v^2$ with $\Phi_i(\theta) = \Phi\left([\alpha + \widetilde{\rho}]y_{2,i} + x_i'[\beta - \widetilde{\rho}\pi] - \widetilde{\rho}\frac{\gamma}{n^{\lambda}} z_i\right)$. The Jacobian of $\log L(\theta)$ w.r.t. $\widetilde{\rho}$ for example, is $\sum_{i=1}^{n} a_i[y_{1,i} - \Phi_i(\theta)]$, with $a_i = a(y_{2,i}, x_i, z_i) = (y_{2,i} - x_i'\pi - \frac{\gamma}{n^{\lambda}} z_i)\phi_i(\theta)/[\Phi_i(\theta)(1 - \Phi_i(\theta))]$. Thus, under the null hypothesis of weak instruments, the MLE of the IV probit model is also inconsistent. As discussed in footnote 3, a similar argument will hold for other commonly encountered discrete choice models.

However, under the local-to-zero DGP and due to the nonlinearity of the moments, it is unclear that equation (9) provides useful information regarding identification. Moreover, the nonlinearity of the moments ensures that instrument weakness contaminates the identification of all parameters in θ_1.

Nonetheless, in discrete choice models, genuine instrument weakness is captured by allowing the covarition between the instruments, z_i, and endogenous regressors, $y_{2,i}$, decays at $\sqrt{n}$, which is captured by $\lambda = 1/2$ in equation (6). Therefore, we base our test for instrument weakness on the following null and alternative hypothesis:

$$H_0 : \ \lambda = 1/2, \ \ H_1 : \lambda < 1/2. \tag{10}$$

The null hypothesis corresponds to the standard local-to-zero null that underpins testing for weak IV in the linear model. Therefore, if we reject H_0, we can conclude that the instruments are not weak and any standard GMM estimator of θ_0 will be consistent.

The null hypothesis considered in (10) differs from that considered in AR: in order to control the size of their test, AR can only consider tests of the null hypothesis $H_0 : \lambda = 1/4$. This distinction is entirely due to the generality of the AR testing approach and the necessary requirement that certain second-order terms associated with their test statistic be negligible. Interestingly, and as we discuss further in the next section, the linear structure of the discrete choice model allows us to test the broader null hypothesis $H_0 : \lambda = 1/2$.

3.1.2 Testing Framework

Due to the nonlinearity of $\mathbb{E}[\bar{g}_n(\theta)]$, the pattern of identification weakness that results under H_0 (i.e., $\lambda = 1/2$) is unclear. However, following AR, we can explore the instrument strength via the limiting behavior of the Jacobian of $\bar{g}_n(\theta)$:

$$\frac{\partial \bar{g}_n(\theta)}{\partial \theta'} = \frac{1}{n}\sum_{i=1}^{n}[a_i, b_i]\frac{\partial r_i(\theta)}{\partial \theta'},$$

and the portion of the Jacobian that determines the identification weakness is given by

$$\begin{aligned}
&\begin{bmatrix} \frac{\partial r_{1,i}(\theta)}{\partial \theta'} \\ \frac{\partial r_{2,i}(\theta)}{\partial \theta'} \end{bmatrix} = \begin{bmatrix} \frac{\partial r_{1,i}(\theta)}{\partial \theta_1'} & \frac{\partial r_{1,i}(\theta)}{\partial \theta_2'} \\ \frac{\partial r_{2,i}(\theta)}{\partial \theta_1'} & \frac{\partial r_{2,i}(\theta)}{\partial \theta_2'} \end{bmatrix} \\
= \ &\begin{bmatrix} \phi_i(\theta)(y_{2,i} - x_i'\pi - \frac{\gamma}{n^\lambda}z_i) & \phi_i(\theta)y_{2,i} & \phi_i(\theta)x_i' & -\phi_i(\theta)\widetilde{\rho}x_i' & -\phi_i(\theta)\widetilde{\rho}z_i \\ 0 & 0 & 0 & x_i' & z_i \end{bmatrix},
\end{aligned}$$

where

$$\Phi_i(\theta) = \Phi\left([\alpha + \widetilde{\rho}]y_{2,i} + x_i'[\beta - \widetilde{\rho}\pi] - \widetilde{\rho}\gamma\frac{z_i}{n^\lambda}\right), \text{ and } \phi(x) = \frac{d\Phi(x)}{dx}.$$

From these formulas above we can state the Jacobian of the moments as

$$\begin{aligned}
\frac{\partial \bar{g}_n(\theta)}{\partial \theta'} &= \frac{1}{n}\sum_{i=1}^{n}\frac{\partial g_i(\theta)}{\partial \theta'} \\
\frac{\partial g_i(\theta)}{\partial \theta'} &= \begin{bmatrix} a_i(y_{2,i} - x_i'\pi - \frac{\gamma z_i}{n^\lambda})\phi_i(\theta) & a_i y_{2,i}\phi_i(\theta) & a_i\phi_i(\theta)x_i' & (-a_i\widetilde{\rho}\phi_i(\theta) + b_i)x_i' & (-a_i\widetilde{\rho}\phi_i(\theta) + b_i)z_i \end{bmatrix}.
\end{aligned}$$

Clearly, the rank of the Jacobian $\partial \bar{g}_n(\theta)/\partial \theta'$ is deficient asymptotically at rate n^λ.

Following AR, the development of a test for weak identification requires that we are able to find a deterministic nonsingular $d_\theta \times d_\theta$ matrix A_n so that, for some $H \times d_\theta$ matrix $M(\theta^0)$ with full column rank d_θ,

$$\underset{n\to\infty}{\text{Plim}} \frac{\partial \bar{g}_n(\theta^0)}{\partial \theta'} A_n = M(\theta^0). \tag{11}$$

The identification strength of the Jacobian is then characterized by the rescaling matrix A_n[5]. One choice for A_n is

$$A_n = \begin{bmatrix} n^\lambda & 0 & 0 & 0 & 0 \\ -n^\lambda & 1 & 0 & 0 & 0 \\ n^\lambda \pi^0 & 0 & \mathbf{I}_k & 0 & 0 \\ 0 & 0 & 0 & \mathbf{I}_k & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix},$$

which yields the following matrix for $M(\theta^0)$:

$$M(\theta^0) = \mathbb{E}\left[-a_i \gamma^0 z_i \phi_i(\theta^0) \quad a_i y_{2,i} \phi_i(\theta^0) \quad a_i \phi_i(\theta^0) x_i' \quad (-a_i \widetilde{\rho}^0 \phi_i(\theta^0) + b_i) x_i' \quad (-a_i \widetilde{\rho}^0 \phi_i(\theta^0) + b_i) z_i\right].$$

The fact that A_n depends on π^0 is immaterial since this matrix will never be explicitly calculated in the testing procedure.

A sample counterpart to $M(\theta^0)$ can be computed using $\partial \bar{g}_n(\theta)/\partial \theta' A_n$ and a GMM estimator $\hat{\theta}_n$ that satisfies

$$\underset{n\to\infty}{\text{Plim}} \frac{\partial \bar{g}_n(\hat{\theta}_n)}{\partial \theta'} A_n = M(\theta^0). \tag{12}$$

The following lemma shows that (12) is satisfied so long as $\|\hat{\theta}_n - \theta^0\| = o_p(1)$.

Lemma 1 *If $\{\theta_n\}$ is such that $\|\theta_n - \theta^0\| = o_p(1)$, then $\underset{n\to\infty}{Plim} \frac{\partial \bar{g}_n(\theta_n)}{\partial \theta'} A_n = M(\theta^0)$.*

In addition, if one is willing to assume a distribution for the instruments z_i, Lemma 1 is satisfied under H_a and the primitive assumptions given in (A0)-(A4).

Lemma 2 *Assumption (A0)-(A4) are satisfied and in addition, for some n large enough, with probability approaching 1, $0 < \inf_{\theta\in\Theta} \|S_n^{-1}(\theta)\| \leq \sup_{\theta\in\Theta} \|S_n^{-1}(\theta)\| < \infty$ and $\sup_{\theta\in\Theta} \|S_n^{-1}(\theta) - \{\mathbb{E}[g_i(\theta) g_i(\theta)']\}^{-1}\| = o_p(1)$. If $z_i \sim \mathcal{N}(0,1)$, then under the alternative hypothesis, $\|\hat{\theta}_n - \theta^0\| = o_p(1)$ and $\underset{n\to\infty}{Plim} \frac{\partial \bar{g}_n(\hat{\theta}_n)}{\partial \theta'} A_n = M(\theta^0)$.*

Remark 2 *Lemma 2 demonstrates that, under particular parametric assumptions on DGP of the instruments, the CU-GMM estimator is consistent under the alternative. As a result, by Lemma 1, the required rank condition needed to implement our testing strategy is satisfied. The choice of a standard normal DGP for z_i is used only to obtain analytical formulas for $\mathbb{E}[g_i(\theta)]$, and the result of Lemma 2 remains valid under a wide range of DGPs for z_i.*

[5]It is worth mentioning that the subvector π is always strongly-identified by the reduced form of $y_{2,i}$.

3.2 Test Statistic: the Distorted J-statistic

The test statistic we employ is a distorted version of the J-statistic (Hansen, 1982) calculated from the CU-GMM estimator (Hansen et al., 1996):

$$\hat{\theta}_n = \arg\min_{\theta\in\Theta} J_n[\theta] = \arg\min_{\theta\in\Theta} n \cdot \bar{g}_n(\theta)' S_n(\theta)^{-1} \bar{g}_n(\theta), \tag{13}$$

The key intuition for our proposed test is that under the null, the curvature of $J_n[\theta]$ is insensitive to a small enough perturbation of the parameter γ. Consequently, if we perturb the CU-GMM estimator by a small positive $o(1)$ sequence, $J_n[\theta]$ remains bounded under the null. However, under the alternative, and for a well chosen perturbation term, this perturbation will give us power to detect deviations from weakness. Hence, we refer the test as a distorted J-test.

However, to ensure that our test statistic accurately controls size, we must first characterize the "directions of weakness" in the parameter space: if we distort parameter estimates that are consistent under both the null and alternative hypothesis, there is no reason to suspect that the test will have appropriate size, since the distorted objective function could display the same behavior under both H_0 and H_a.

To characterize the directions of weakness, we follow Antoine and Renault (2009, 2012) and consider a rotation in the parameter space that will allow us to partition the parameter space into two subspaces: the subspace where estimated parameters coverage at the standard $n^{1/2}$-rate, and the subspace where estimated parameters converge at $n^{1/2-\lambda}$-rate. Recall that, the asymptotic behavior of the Jacobian governs the weakness of the parameters. This means that, implicitly, the rates of convergence we are fishing for are actually mixed-up in the original Jacobian. However, due to the non-linearity of the Jacobian, we can not easily disentangle these rates.

The goal of this rotation is to transform the original parameters θ into new parameters

$$\eta = R^{-1}\theta,$$

for some full rank matrix $d_\theta \times d_\theta$ matrix R, that allows us to discern which directions in this new parameter space are weakly identified and which are strongly identified. Examining the Jacobian $\partial\bar{g}_n(\theta^0)/\partial\theta'$, it is clear that the rank deficiency, which causes the identification failure, is of order one; i.e., asymptotically, the Jacobian has rank $d_\theta - 1$ instead of d_θ. Therefore, in this rotation there will be one parameter that converges at the $n^{1/2-\lambda}$-rate, while the other parameters will converge at the standard $n^{1/2}$ rate.

Next, we associate to this parameterization a $d_\theta \times d_\theta$ diagonal matrix that we use to partition the different rates: for $p = d_\theta - 1$,

$$\Lambda_n := \begin{pmatrix} n^{1/2-\lambda} & \mathbf{O}_{p\times p} \\ \mathbf{O}_{p\times 1} & n^{1/2}\mathbf{I}_p \end{pmatrix}.$$

Note that, only the first entry has the slower rate of convergence, while the other entries have the standard $n^{1/2}$-rate. We can then obtain an explicit form for the rotation by requiring that the matrix R satisfies

$$\frac{\partial\bar{g}_n(\theta)}{\partial\theta'}\Big|_{\theta=\theta^0} A_n = \frac{\partial}{\partial\eta'}\left[\bar{g}_n(R\eta)\right]\Big|_{\theta=\theta^0}\sqrt{n}\Lambda_n^{-1} = \frac{\partial\bar{g}_n(\theta^0)}{\partial\theta'}R\sqrt{n}\Lambda_n^{-1};$$

that is, the rotation matrix R must solve

$$A_n = R\sqrt{n}\Lambda_n^{-1}.$$

In this case, it is simple to show that the matrix R is then defined as

$$R = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ -1 & 1 & 0 & 0 & 0 \\ \pi^0 & 0 & \mathbf{I}_k & 0 & 0 \\ 0 & 0 & 0 & \mathbf{I}_k & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

This rotation of the original model parameters cleanly partitions the parameter space into the direction that is weakly identified and those that are strongly identified. As such, from the definition of Λ_n, we know that η_1 is the **only** direction of weakness in this discrete choice model. Using this information, we can construct a distorted J-test by perturbing this single parameter.

For $\hat{\eta}_n := R^{-1}\hat{\theta}_n$, define the perturbed vector of parameters

$$\hat{\eta}_n^\delta := \hat{\eta}_n + \begin{pmatrix} \delta_n \\ \mathbf{O}_{d_\theta - 1} \end{pmatrix}, \quad \text{where} \quad \delta_n := \delta / r_n \quad \text{with} \quad r_n \to \infty.$$

From $\hat{\eta}_n^\delta$ we can now calculate the distorted J-statistic, J_n^δ, as

$$\hat{\theta}_n^\delta = R\hat{\eta}_n^\delta := R\hat{\eta}_n + R \begin{pmatrix} \delta_n \\ \mathbf{O}_{d_\theta - 1} \end{pmatrix} \tag{14}$$

$$J_n^\delta = J_n[\hat{\theta}_n^\delta] := n \cdot \bar{g}_n(\hat{\theta}_n^\delta)' S_n(\hat{\theta}_n^\delta)^{-1} \bar{g}_n(\hat{\theta}_n^\delta). \tag{15}$$

where $\hat{\theta}_n = (\hat{\theta}_{1,n}', \hat{\pi}_n', \hat{\xi}_n)'$ is the CU-GMM estimator defined in (13).[6] The tuning parameter (perturbation) $\delta_n = o_p(1)$ is a well-suited scalar sequence, where δ is randomly chosen from a continuous distribution.

Note that, given the definition of R and $\hat{\theta}_n^\delta$ in equation (14), even though η_1 is the only parameter that is perturbed in the rotated parameters space, it can be directly verified that the perturbation then contaminates the CU-GMM estimates of the structural parameters, i.e., θ_1, in the original parameter space. Stated another way, even though η_1 is the only weakly identified parameter in the rotated parameter space, this weakness affects all the structural parameters in the original parameter space.

To study the asymptotic properties of our distorted J-test statistic, we maintain the following high-level assumption.

(A5) For $0 < \lambda < 1/2$, any GMM estimator $\hat{\theta}_n \in \Theta$ is such that $\sqrt{n}R\Lambda_n^{-1}(\hat{\eta}_n - \eta^0) = \sqrt{n}A_n^{-1}(\hat{\theta}_n - \theta^0) = O_p(1)$.

Remark 3 *This assumption is satisfied under more primitive conditions given in Antoine and Renault (2009, 2012).*

The theorem below allows us to consistently test weak instruments in discrete choice models.

[6] By the proof of Lemma 2, we know that π is always strongly identified. In practice, we can implement the rotation using $\hat{R}_n$, where the unknown π^0 is replaced by its estimator $\hat{\pi}_n$, and $\hat{\pi}_n$ can be any $\sqrt{n}$-consistent estimator for π.

Theorem 3 (Distorted J-test)
Under assumptions (A0)-(A5), for any arbitrary deterministic sequence $r_n \to \infty$ and a random scalar δ such that, $\delta_n = \delta/r_n$, define the test by the rejection region

$$W_n^\delta = \{J_n^\delta > \chi^2_{1-\alpha}(H)\}$$

where $\chi^2_{1-\alpha}(H)$ is the $1-\alpha$-quantile of the Chi-square distribution with degree of freedom H.
(i) Under the null hypothesis $H_0 : \lambda = 1/2$, the test W_n^δ is asymptotically conservative at level α.
(ii) The test W_n^δ is consistent under the alternative so long as δ_n satisfies $n^{1/2-\lambda}\delta_n \to \infty$ as $n \to \infty$.

The key to ensure that the size of J_n^δ is asymptotically controlled is the equivalence between the usual J-statistic and the distorted J-statistic that obtains under the null hypothesis. This equivalence allows us to bound the asymptotic distribution of J_n^δ by a Chi-square distribution with H degrees of freedom, $\chi^2(H)$. However, under the alternative hypothesis, the usual J-statistic and J_n^δ are no longer asymptotically equivalent so long as $n^{1/2-\lambda}\delta_n \to \infty$, which ensures J_n^δ can detect departures from weakness. This result implies that the tuning parameter δ_n governs the power of our test: the slower δ_n goes to zero, the more powerful the test.

The parameters π, a sub-vector of θ, are strongly identified under both the null and alternative hypothesis. This allows us to consider a less conservative version of the distorted J-test, which is satisfied under the following assumption.

(A6) The vectors $a_i = a(y_{2,i}, x_i, z_i)$ and $b_i = b(x_i, z_i)$ are such that

$$G_0 := \mathbb{E}\left\{[a_i, b_i]\frac{\partial r_i(\theta^0)}{\partial \pi'}\right\} = \mathbb{E}[(-a_i\widetilde{\rho}^0\phi_i(\theta^0) + b_i)x_i']$$

has full column rank $k = \dim(\pi)$.

Corollary 4 *Under assumptions (A0)-(A6), for any arbitrary deterministic sequence $r_n \to \infty$ and a random scalar δ such that, $\delta_n = \delta/r_n$, define the test by the rejection region*

$$\widetilde{W}_n^\delta = \{J_n^\delta > \chi^2_{1-\alpha}(H-k)\},$$

where $\chi^2_{1-\alpha}(H-k)$ is the $1-\alpha$-quantile of the Chi-square distribution with degree of freedom $H-k$.
(i) Under the null hypothesis $H_0 : \lambda = 1/2$, the test $\widetilde{W}_n^\delta$ is asymptotically conservative at level α.
(ii) The test W_n^δ is consistent under the alternative so long as δ_n satisfies $n^{1/2-\lambda}\delta_n \to \infty$ as $n \to \infty$.

Remark 4 *Assumption (A6) essentially means that $[a_i, b_i]$ must be sufficiently correlated with the Jacobian of $\partial r_i(\theta^0)/\partial\pi'$; i.e., that this term is not affected by the (potential) weakness of the instruments. For example, if we set $a_i = b_i$ and let the first k elements in a_i be x_i, then we have a partition of the $H \times k$ matrix $G_0 = \left[G_1' \vdots G_2'\right]'$, where G_1 is $k \times k$ and G_2 is $(H-k) \times k$. In addition,*

$$G_1 = \mathbb{E}[(1 - \widetilde{\rho}^0\phi_i(\theta^0))x_i x_i'] = \mathbb{E}[x_i x_i'] - \widetilde{\rho}^0\mathbb{E}[\phi_i(\theta^0)x_i x_i']$$

We contend that it is reasonable to assume that the $k \times k$ matrix G_1 is full rank, since both $\mathbb{E}[x_i x_i']$ and $\mathbb{E}[\phi_i(\theta^0)x_i x_i'] = \mathbb{E}[(\phi_i(\theta^0)^{1/2}x_i)(\phi_i(\theta^0)^{1/2}x_i)']$ are positive definite matrices with full rank, and it is almost impossible that their linear combination G_1 is rank deficient. It then implies that G_0 is also full column rank. Therefore, under mild conditions, there exists a_i and b_i, under which (A6) is satisfied.

3.3 Testing Procedures

We now explain how to implement our distorted J-test in practice. The testing approach proceeds through the following four steps. When rejecting the null, empirical researchers can conclude that instruments are not weak and the CU-GMM estimator is consistent.

(1) Compute the CU-GMM estimator $\hat{\theta}_n$ defined in (13);

(2) For a sequence of tuning parameter $\delta_n = \delta/r_n$, randomly draw δ from a continuous distribution. Choose r_n such that $r_n \to \infty$, and $n^{1/2-\lambda}\delta_n \to \infty$ if $0 < \lambda < 1/2$;

(3) Compute $\hat{\theta}_n^\delta$ and the test statistic J_n^δ, as defined in (14) and (15) respectively;

(4) Rejection rule: reject the null if $J_n^\delta > \chi^2_{1-\alpha}(H-k)$.[7]

The key step in the testing procedure is to choose the tuning parameter δ_n. Theorem 2 shows that for any sequence δ_n satisfying the requirements in the testing procedure step (2), the distorted J-test is asymptotically conservative and consistent. However, the finite sample performance of the test will depend on the choice of δ_n. In particular, as will be made clear in the following section, the finite sample behavior of J_n^δ depends on the variance of the instruments. Therefore, we suggest to restrict the perturbation to be of the form $\delta_n = \delta/r_n$, and choose δ and r_n in the data-driven manner that is influenced by the generalized rule-of-thumb for discrete choice models presented in Section 4. In particular, we choose δ and r_n as follows:

$$\delta \sim \mathcal{N}(c, \sigma^2_{\delta,n}), \quad r_n = \hat{\sigma}_{\phi,z} \log(\log(n)) \tag{16}$$

where

$$c = \sqrt{d_\theta - k - \sigma^2_{\delta,n}}, \quad \sigma^2_{\delta,n} = 1/n^2, \quad \hat{\sigma}^2_{\phi,z} = \left(\frac{1}{n}\sum_{i=1}^n a_i z_i \phi_i(\hat{\theta}_n)\right)' S_n(\hat{\theta}_n)^{-1} \left(\frac{1}{n}\sum_{i=1}^n a_i z_i \phi_i(\hat{\theta}_n)\right).$$

The idea behind this data-driven choice of δ_n is twofold: firstly, by controlling for the variability of the IVs, through the term $\hat{\sigma}_{\phi,z}$, this choice of δ_n should help control the size of the test by alleviating any impact the variability of the instruments might have on the test statistic; secondly, this choice of δ_n should improve the power of the test by ensuring that the dominant term for determining the behavior of J_n^δ is directly related to the rate at which the test statistic diverges under the alternative hypothesis.

To understand this particular choice for the data-driven perturbation, it is perhaps easiest to examine the dominant term in the generalized rule-of-thumb for discrete choice models discussed in Section 4, which is proportional to

$$\left(\frac{1}{n}\sum_{i=1}^n a_i z_i \phi_i(\hat{\theta}_n)\right)' S_n(\hat{\theta}_n)^{-1} \left(\frac{1}{n}\sum_{i=1}^n a_i z_i \phi_i(\hat{\theta}_n)\right) \times \frac{n\xi^2\delta^2}{r_n^2}, \tag{17}$$

where $\xi := \gamma/n^\lambda$,$0 < \lambda \leq 1/2$, and $|\gamma| < \infty$. Recalling the definition of r_n in equation (16), and the definition of $\hat{\sigma}_{\phi,z}$, as the weighted norm on the correlation between $a_i z_i$ and $\phi_i(\hat{\theta}_n)$, we see

[7] Recalling that π is strongly identified under the null, from Corollary 4 we know that the limiting distribution of J_n^δ is dominated by $\chi^2(H-k)$ and, thus, we can use critical values from $\chi^2(H-k)$ to implement our testing approach.

that this choice of δ_n has normalized the scale of the term in equation (17) and eliminated the impact of $\text{Var}(z_i)$ on J_n^δ. Therefore, the size of J_n^δ will remain under control even for small sample sizes and even if $\text{Var}(z_i)$ is large.[8] Moreover, under the alternative, the CU-GMM estimator $\hat{\theta}_n$ is consistent and $J_n[\hat{\theta}_n] \xrightarrow{d} \chi^2(H - d_\theta)$.[9] However, since the DJ test rejects the null based on a $\chi^2(H-k)$ critical value, where $d_\theta > k$, we can improve the power of the DJ test by making up for the mean difference between the $\chi^2(H-d_\theta)$ and $\chi^2(H-k)$ critical values; this is accomplished by setting $\mathbb{E}[\delta^2] = \text{Var}[\delta] + \mathbb{E}^2[\delta] = d_\theta - k$, where $d_\theta - k$ is the difference between the means of the two Chi-square distributions.[10] The Monte Carlo simulations in Section 5 show that this data-driven choice of the perturbation performs well.

4 Generalizing the Rule-of-thumb for Probit Model

In this section, we present a relationship between the distorted J-test and the rule-of-thumb for the linear model. To simplify the presentation, we consider a simplification of the original model where there are no exogenous covariates.[11].

Recall that, in the linear model the concentration parameter μ^2 measures the strength of the instruments and is defined as

$$\mu^2 = \sum_{i=1}^{n} (\xi z_i)^2 / \sigma_v^2. \tag{18}$$

For $\xi = \gamma / n^\lambda$ and $0 < \lambda \leq 1/2$, it follows

$$\underset{n\to\infty}{\text{Plim}}\; n^{2\lambda-1}\mu^2 = \gamma^2 \mathbb{E}[z_i^2]/\sigma_v^2.$$

The following proposition underpins the behavior the F-statistic in the case of homoscedastic first-stage error terms.

Proposition 5 *Let F_n denote the first-stage F-statistic calculated via the least square estimator. Under the null of weak instruments, $H_0 : \lambda = 1/2$, we have*

$$F_n \xrightarrow{d} \frac{1}{d_z}\chi^2\left(d_z, \mu_\infty^2\right) \quad \textit{and} \quad \mu_\infty^2 \equiv \underset{n\to\infty}{\text{Plim}}\; \mu^2 = \gamma^2 \mathbb{E}[z_i^2]/\sigma_v^2,$$

[8] If we do not adjust δ_n according to the instrument variation, the size of the DJ test will be out of control and blow up as $\text{Var}(z_i)$ increases. However, if we blindly adjust δ_n by taking $r_n = \hat{\sigma}_z \log(\log(n))$, where $\hat{\sigma}_z$ is the sample standard deviation of z_i, then a nonzero power is not guaranteed for the DJ test when $\text{Var}(z_i)$ is relatively large, even if λ is small and sample size is large. From the expressions of $\hat{\sigma}_{\phi,z}$ and of the dominant term in (17), it is easy to see that by adjusting r_n using $\hat{\sigma}_{\phi,z}$, we eliminate the entire impacts of $\text{Var}(z_i)$ to the dominant term. Hence, simulation results in Table 1 and Figure 1, 2 (and results not reported here) show that the size of our DJ test is under controlled and there are always positive power when λ is not close to 0.5, say $\lambda < 0.35$, regardless of the magnitude of the IV variation.

[9] See Antoine and Renault (2012) Theorem 4.4.

[10] We note that, in the case where there is more than one instrument, say z_j, $j = 1, 2, \ldots$, the weighted norms $\hat{\sigma}_{\phi,z_j}$ can be calculated for each instrument, and then the perturbation can be defined in an analogous fashion by taking the maximum of the individual weighted norms among all the instruments.

[11] This simplification is immaterial for the validity of the general approach, however, it does simplify the interpretation.

where d_z is the dimension of z_i, $\chi^2(d_z, \mu_\infty^2)$ is the non-central Chi-square distribution with degree of freedom d_z and the noncentral parameter μ_∞^2. Under the alternative $H_1 : \lambda < 1/2$, as $n \to \infty$

$$F_n = \frac{\mu^2}{d_z} + o_p(1) \to \infty.$$

Because of the direct link between μ^2 and F_n, the standard rule-of-thumb simply utilizes this relation and declares instruments to be weak if $F_n < 10$. Similarly, the testing approaches of Stock and Yogo (2005) and Olea and Pflueger (2013) rely on the the concentration parameter in the linear model. Consequently, due to the nonlinearity of the moments in the "second-stage" equation, the conventional weak IV tests, although convenient to implement, may be invalid in discrete models. The following proposition gives a representation of the distorted J-statistic that will allow us to give a simple comparison with the standard first-stage F-statistic under the alternative hypothesis considered in this paper.

Proposition 6 *Define*

$$w_i = [a_i, b_i], \quad V_i = \left[y_{1,i} - \Phi_i(\theta^0),\ v_i\right]', \quad V_i^w = w_i V_i, \quad z_i^w = w_i z_i \quad \textit{and} \quad \Pi_i = \left[\gamma^0 \phi_i(\theta^0), 0\right]'.$$

Under the alternative hypothesis $H_1 : \lambda < 1/2$, if δ_n is such that $n^{1/2-\lambda}\delta_n \to \infty$, then, for

$$\mathcal{U}^2 := n^{1-2\lambda}\delta_n^2 Q_n(\theta^0)' S_n(\theta^0)^{-1} Q_n(\theta^0),$$
$$S_n(\theta^0) := \frac{1}{n}\sum_{i=1}^n V_i^w (V_i^w)' + o_p(1), \ \textit{and}\ Q_n(\theta^0) := \frac{1}{n}\sum_{i=1}^n z_i^w \Pi_i,$$

we have

$$\frac{J_n^\delta}{\mathcal{U}^2} = 1 + O_p\left(\frac{1}{n^{1/2-\lambda}\delta_n}\right) \quad \textit{as} \quad n \to \infty.$$

This result states that the leading term in the distorted J-statistic is $\mathcal{U}^2$. From the definition of $\mathcal{U}^2$, we can view V_i as a generalized error for the estimated probit model, which includes errors from both the structural equation and the reduced form. Π_i represents the effects of the weakly identified direction on $(y_{1,i}, y_{2,i})'$, where the zero in the second entry means the weakly identified direction has no effects on the reduced form of $y_{2,i}$. Under the alternative hypothesis,

$$\operatorname*{Plim}_{n\to\infty} (n^{1-2\lambda}\delta_n^2)^{-1}\mathcal{U}^2 = \mathbb{E}\left[z_i^w \Pi_i\right]' \operatorname{Var}\left[V_i^w\right]^{-1} \mathbb{E}\left[z_i^w \Pi_i\right]. \tag{19}$$

It is clear that $\mathbb{E}\left[z_i^w \Pi_i\right]' \operatorname{Var}\left[V_i^w\right]^{-1} \mathbb{E}\left[z_i^w \Pi_i\right]$ possesses a similar structure to the concentration parameter μ^2 in the linear IV model, and is a natural extension of the instrument strength of z_i for the probit model. Hereafter, we refer to $\mathcal{U}^2$ as the 'generalized concentration parameter' for IV probit models.

Similar to μ^2, the generalized concentration parameter is bounded asymptotically when $\lambda = 1/2$, and diverges if $\lambda < 1/2$ because $n^{1/2-\lambda}\delta_n \to \infty$, where the latter property is critical in detecting IV weakness. However, the divergence rate of $\mathcal{U}^2$ is slower than the rate of μ^2 because of the perturbation $\delta_n = o_p(1)$. Interestingly, both concentration parameters depend on the variation of the instruments and the error terms. In addition, both statistics depend on the level of endogeneity,

measured by ρ, while the impacts of ρ on the ability of the distorted J-test to detect instrument strength is complex, since ρ enters J_n^δ nonlinearly through $\phi_i(\theta)$.

While the two concentration parameters are similar in many respects, there are fundamental differences between the generalized concentration parameter and its linear counterpart. Firstly, the rule-of-thumb for the linear model depends solely on information in the reduced form. However, noting that both the matrix Π_i and the errors V_i depend on the reduced form and structural parameters, we see that the nonlinearity of the moments in the probit model ensures that the instrument strength cannot be accurately measured by the first stage F-statistic.

The generalized concentration parameter is a weighted quadratic form in the average of $z_i^w \Pi_i$. As such, the variability of z_i will have a significant impact on the finite sample behavior of our test: if H_0 is true, when the sample size is small or moderate, a large variance for z_i will result in over-rejection of the null in finite-samples. Luckily, by choosing the tuning parameter δ_n as in (16), we can mitigate the influence of the variance of z_i within J_n^δ. However, the standard rule-of-thumb cannot control the influence of the instrument variance on the magnitude of the test statistic.

Particularly, by Proposition 5 and under $H_0 : \lambda = 1/2$, F_n converges to a non-central Chi-square distribution with a non-centrality parameter depending on $\mathbb{E}[z_i^2]/\sigma_v^2$. Therefore, if σ_z^2 is large relative to σ_v^2, $F_n > 10$ can happen with high probability in finite samples even though $\lambda = 1/2$, which can generate large type I errors. As a result, the standard rule-of-thumb will falsely reject the null of weak instruments and lead to inconsistent IV estimation in both linear and probit models. Moreover, when σ_z^2 is relatively small, $F_n > 10$ may rarely happen even for $\lambda < 1/2$.

5 Monte Carlo: Conventional Weak IV Tests v.s. Distorted J-test

In this section, we verify the properties of the distorted J-test (DJ hereafter) and compare this test against three commonly used weak IV tests, which, even though they are not designed for discrete choice models, have been widely applied in the literature on discrete choice modeling: (i) Staiger and Stock (1997) standard rule-of-thumb (SS); (ii) Stock and Yogo (2005) (SY); (iii) Robust weak IV test of Olea and Pflueger (2013) (Robust).

We generate observed data according to

$$y_{1,i} = 1[\beta + \alpha y_{2,i} + u_i > 0], \quad y_{2,i} = \pi + \xi z_i + v_i, \quad i = 1, 2, ..., n \tag{20}$$

where $y_{2,i}$, z_i are univariate, $(u_i, v_i)'$ is homoscedastic and normally distributed, and $(u_i, v_i)'$ is independent of z_i. We set $\beta = 0$, $\alpha = 1$, $\pi = 1$, $\xi = \gamma/n^\lambda$ and $\gamma = 1$. In addition, we take $\rho \in \{0.5,\ 0.8\}$, $\sigma_v = 1$, $\sigma_u = 1/\sqrt{1-\rho^2}$ (to ensure normalization of $\text{Var}(u_i|y_{2,i}, z_i) = 1$). The intercept parameter π is always strongly identified, while the strength of identification for the remaining parameters depends on the value of λ. We capture instrument strength in the DGP (20) via the following grid of values for λ: $\{0.50,\ 0.45,\ 0.40,\ 0.35,\ 0.30,\ 0.25,\ 0.20\}$. In this design, H_0 corresponds to $\lambda = 0.5$, while H_a corresponds to $\lambda < 0.5$.

Since the performance of the DJ test and the standard weak IV tests may depend on $\text{Var}(z_i)$, we simulate z_i i.i.d. from one of three distributions: $z_i \sim N(0, 2)$; $z_i \sim N(0, 8)$; or $z_i \sim N(0, 36)$. For each Monte Carlo trial, we take the sample size to be one of $n = 500,\ 5000,\ 10000$ and consider 1000 Monte Carlo replications.

Across each Monte Carlo design, following the DJ testing procedures outlined in Section 3.3, θ^0 is estimated by CU-GMM with a single degree of over-identification. We choose as instrumental functions $a_i = a(y_{2,i}, z_i) = (y_{2,i}, z_i, z_i^2, z_i^3, 0, 0)'$ and $b_i = b(z_i) = (0, 0, 0, 0, z_i, z_i^2)'$. The DJ test is implemented following the test procedures and the data-driven choice of the perturbation in Section 3.3. Using a 5% significant level, we reject the null hypothesis of weak instruments in accordance to Corollary 4; i.e., we reject the null if $J_n^\delta > \chi^2_{0.95}(H-k)$, where in this case $H = 6$, $k = 1$ and $\chi^2_{0.95}(H-k) = 11.07$.

The rejection probabilities under the null hypothesis for the different testing procedures are collected in Table 1.[12] The rejection rates displayed in Table 1 confirm that the DJ test is asymptotically conservative, i.e., the size is controlled under the significant level 5%. Conversely, we see that blindly applying conventional weak instrument tests would lead to poor outcomes. In particular, under the null hypothesis and for instruments with moderate variability, i.e., $\sigma_z^2 = 8$, the rejection rates of SS and SY are about 40% and 10% respectively. Moreover, when σ_z^2 is large, i.e., $\sigma_z^2 = 36$, the rejection rates of SS and SY are near one. Therefore, we can conclude that if the instruments have relatively large variance, the conventional weak IV tests designed for linear models fail to provide reliable conclusions on the instrument strength when applied to discrete choice models.

Figures 1-2 display the power of the DJ test when $\rho = 0.5$ and $\rho = 0.8$, respectively. Due to the conservativeness of DJ test, size adjusted power (adj-DJ) is also computed[13]. Generally speaking, the resulting power curves demonstrate that the DJ-test, when size adjusted, displays non-negligible power even when instruments are very close to being weak, i.e, when $\lambda = 0.45$ or $\lambda = 0.40$, which gives convincing numerical evidence of the results in Theorem 3 and Corollary 4.

We also verify the asymptotic behavior of the CU-GMM estimator in the probit model by calculating bias, standard deviation (s.d.) and relative root mean square error (rrmse), which are constructed as follows (take $\widehat{\alpha}$ as an example):

$$\text{GMM bias} = \bar{\widehat{\alpha}} - \alpha^0, \quad \text{s.d.} = \sqrt{\frac{1}{M}\sum_{l=1}^{M}(\widehat{\alpha}_l - \bar{\widehat{\alpha}})^2}, \quad \text{rrmse} = \sqrt{\frac{1}{M}\sum_{l=1}^{M}\left(\frac{\widehat{\alpha}_l - \alpha^0}{\alpha^0}\right)^2}$$

where $\bar{\widehat{\alpha}} = 1/M\sum_{l=1}^{M}\widehat{\alpha}_l$, $\widehat{\alpha}_l$ stands for the l-th Monte Carlo estimates and α^0 is the true value. For brevity, we only report the asymptotic results for the structural parameter of interest, α, in Table 2 under the case DGP $z_i \sim \mathcal{N}(0, 8)$. Additional results for all designs can be obtained from the authors.

As mentioned in the introduction, several authors have noted the potential invalidity of conventional weak instrument tests for discrete choice models. In several of these studies, the authors have instead represented the discrete choice model by a linear probability model (LPM), in order to apply linear weak IV tests to detect weak instruments. To assess the performance of this alternative strategy, we consider the application of conventional weak IV tests in the case where the true DGP is probit but instead a LPM is estimated. To this end, suppose we fit the data generated

[12]We note that the null hypothesis of each test are slightly different: DJ- $H_0 : \lambda = 1/2$; SS- $F_n < 10$ as an informal null hypothesis; SY- the triple $\{\xi, \sigma_v^2, \sigma_z^2\}$ is such that 2SLS relative bias or Wald test size distortion is larger than a given tolerance using the Cragg-Donald statistic; the Robust test regards that the Nagar bias exceeds a fraction of the benchmark as null. Although the definitions of the weak instrument are different for each test, their null hypothesis are consistent in the sense to capture situations under which the instrument is weak.

[13]Size adjusted power of the DJ test is computed as follows: obtain the 95% quantile of the distorted J-statistic from the 1000 Monte Carlo replications when $\lambda = 1/2$ and use it as the critical value for cases when $\lambda < 1/2$.

from Model (20) to

$$y_{1,i} = \widetilde{\beta} + \widetilde{\alpha} y_{2,i} + \widetilde{u}_i, \quad y_{2,i} = \pi + \xi z_i + v_i. \tag{21}$$

The LPM is by construction heteroscedastic, which means that the testing procedures of SS and SY are invalid. However, the Robust version of the SY test can be used to mitigate the heteroscedasticity issue in the LPM.

Table 3 summarizes two measure of 2SLS performance: the 2SLS/OLS relative bias used in Stock and Yogo (2005) and the 2SLS/OLS relative absolute bias. It is noteworthy that if the true DGP is linear, as the LPM in equation (21), and if there is a single endogenous regressor, a natural relative bias measure is given by (Stock and Yogo (2005))

$$|\mathbb{E}\widehat{\widetilde{\alpha}}^{2SLS} - \widetilde{\alpha}^0| / |\mathbb{E}\widehat{\widetilde{\alpha}}^{OLS} - \widetilde{\alpha}^0|, \tag{22}$$

where $\widehat{\widetilde{\alpha}}^{2SLS}$ and $\widehat{\widetilde{\alpha}}^{OLS}$ are the 2SLS and OLS estimator of LPM (21), and $\widetilde{\alpha}^0$ is the true value of $\widetilde{\alpha}$. However, because the model is misspecified, it is possible that the usual notion of 'bias towards OLS' is no longer valid, because the impact of weakness on the IV estimators is now complicated by the models nonlinear features. In this case, there is no guarantee that the positive and negative biases will not offset each other, and lead to a spuriously small overall bias. In addition, there is no explicit mapping from the parameters θ of the true model, i.e., the probit model, to the parameter of interest in the misspecified LPM, $\widetilde{\alpha}$. Therefore, there is no such think as a 'true value' $\widetilde{\alpha}^0$ and the quantity in (22) is no longer a meaningful measure of performance

Given the above discussion, we instead choose to compare estimates using a measure of relative absolute bias defined from the pseudo-true value $\widetilde{\alpha}^* = \mathbb{E}[\phi_i(\theta^0)\alpha^0]$,[14] which is the minimizer of the limit 2SLS objective function for the LPM in equation (21). Using the pseudo-true value $\widetilde{\alpha}^*$ we can define the sample counterpart of the 2SLS/OLS relative absolute bias as:

$$\sum_{l=1}^{M} \left| \widehat{\widetilde{\alpha}}_l^{2SLS} - \widetilde{\alpha}^* \right| \Big/ \sum_{l=1}^{M} \left| \widehat{\widetilde{\alpha}}_l^{OLS} - \widetilde{\alpha}^* \right|. \tag{23}$$

Interestingly, $\widetilde{\alpha}^* = \mathbb{E}[\phi_i(\theta^0)\alpha^0]$ corresponds to the mean of the marginal effect of $y_{2,i}$ on $y_{1,i}$ under the true probit DGP. The 2SLS/OLS relative bias displayed in Table 3 is the sample counterpart of (22), but where $\widetilde{\alpha}^0$ has been replaced by $\widetilde{\alpha}^*$. Table 4 provides the first-stage F-statistic and Nagar bias relative to its benchmark (Olea and Pflueger (2013)) as extra information of the LPM estimation.

The results in Table 3 are worrying. Suppose the true DGP is a probit model with weak IV ($\lambda = 1/2$), but a LPM is used for estimation. Then, if the instruments have even moderate variability, $\sigma_z^2 = 8$ (and $\sigma_z^2 = 36$ not reported here), the rejection rates in Table 1 for the SS, SY and the Robust tests seem reasonable, since the conventional relative bias of 2SLS/OLS, calculated using (22), for the misspecified LPM are about 20% ($\rho = 0.5$) and 30% ($\rho = 0.8$) when $n = 500$ and even less when n increasing. However, the relative absolute bias (23) is extremely large: for $\sigma_z^2 = 8$, this bias is over 100% when $\rho = 0.5$ and about 70% to 90% when $\rho = 0.8$. These results indicate that if $\lambda = 0.5$ and the true DGP is probit but the LPM is used, the 2SLS estimator is extremely poor, and even worse than the OLS estimator in some circumstances. Hence, we strongly caution the use of the LPM to study discrete choice problems in the presence of endogenous regressors and weak IV.

[14] Detailed derivations can be found at Appendix A.2 Lemma 10.

6 Empirical Illustrations

In this section, we apply our distorted J-test in two well-known empirical examples to test for the presence of weak instruments. We then contrast the results of our tests with those obtained from conventional weak IV tests for linear models, namely the SS, the SY, and the Robust tests.

6.1 Labor Force Participation of Married Women

We first study married women's labor force participation (hereafter LFP) when education, measured as the women's years of schooling, is treated as an endogenous regressor. We use data from the University of Michigan Panel Study of Income Dynamics (PSID) for the year 1975[15], which have been used in several studies. Mroz (1987) provides an extensive analysis of the women's hours of labor supply, and considers a range of specifications including potential endogeneity of several regressors, the use of different instrumental variables and controls for self-selection into labor force participation. As a text book example, Wooldridge (2010) used the same dataset to study women's LFP decisions, and the potential endogeneity of education is tested after estimating an IV Probit model using Rivers and Vuong (1988) two-step conditional maximum likelihood estimator (2SCML). In what follows we use exactly the same specification as in Wooldridge (2010) (see p.468).

The PSID consists of data on 753 married, Caucasian women who are between 30 and 60 years of age at the time the sample was conducted. The dependent variable LFP is a binary response that equals unity if the respondent worked at some time during the year, and zero otherwise. Exogenous regressors include spousal income, the individuals work experience, age, the number of children less than six years old, and the number of children older than six years old. The individuals education, measured as years of schooling, is considered to be endogenous. Following the strategy in Wooldridge (2010), the individual's family education, which is recorded as the years of schooling for both the individual's father and mother, as well as the spouses education, are used as instruments for education. Table 5 contains a summary of the data for all variables present in the analysis.

Estimated coefficients and the average partial effects on the probability of LFP for all regressors are presented in Table 7 using two estimation methods: 2SCML as used in Wooldridge (2010)[16] and CU-GMM. More specifically, for the 2SCML, the first step is to regress the endogenous regressor on the instruments and all other exogenous regressors to obtain the first step residual. The second step is to run a probit maximum likelihood estimation of the binary response on the endogenous and the exogenous regressors, and the first step residual. The CU-GMM estimation with over-identification degree one is conducted using $a_i = (Y_{2,i}, X_i', Z_i', \mathbf{0}_{k+3}')'$ and $b_i = (\mathbf{0}_{k+3}', 1, X_i', Z_i')'$, where $Y_{2,i}$, X_i and Z_i denote the standardized variables corresponding to the women's education, exogenous regressors and three instruments, and k is the number of exogenous regressors and the intercept in the structural equation. The first step estimation of the 2SCML and the reduced form of the CU-GMM are listed in the first and fourth columns of Table 7 respectively. All the three IVs are highly significant based on both estimation methods. The CU-GMM estimation results are reported in columns four through six. Broadly speaking, the CU-GMM and 2SCML results are similar, with both methods providing evidence that education has a significant positive effect: one extra year of education increasing the probability of LFP by 4.04 and 3.92 percentage points for

[15]The data is public and available at Wooldridge (2010) Supplemental Content.

[16]For the LFP example, the 2SCML estimation allows for heteroscedastic standard errors.

2SCML and CU-GMM, respectively. Hansen's J-statistic is 1.60 which is less than $\chi^2_{0.95}(1) = 3.84$, therefore we fail to reject the null that all the moments are valid.

The weak IV test results are collected in Table 6 for all four tests, SS, SY, Robust and DJ. The Kleibergen-Paap F-statistic (Kleibergen and Paap, 2006) is 140.23, based on which the SS rule-of-thumb and the SY test both reject the null that IVs are weak.[17] For the Robust test, the effective F-statistic is 143.17, the critical values for a tolerance threshold $\tau \in \{5\%, 10\%, 20\%, 30\%\}$ are $[17.84, 11.19, 7.48, 6.10]$, respectively.[18] Comparing the effective F-statistic 143.17 to the critical values, the Robust test also rejects the null of weak IV. Finally, for the DJ test, the perturbation δ_n is computed as in (16) using the maximum weighted norm of the three instruments as $\hat{\sigma}_{\phi,z}$. The DJ test statistic 142.98 is much larger than its critical value $\chi^2_{0.95}(H-k) = 22.36$ with $H = 20$ and $k = 7$, leading to rejecting the null of weak IVs as well. The rejection conclusion of the DJ test is quite straightforward: when drifting the CU-GMM estimator $\hat{\theta}_n$ by a small perturbation δ_n, the value of J-statistic increases dramatically from 1.60 to 142.98, implying that the curvature of J-statistic is extremely sensitive to even very small departures. Overall, results reported in Table 6 suggest that the DJ test rejects the null of weak IVs, whilst all three conventional tests for the linear model would reach the same conclusion for this example.

6.2 US Food Aid and Civil Conflicts

In the second example we examine the impact of US food aid on the incidence of civil conflicts in recipient countries. The research in Nunn and Qian (2014) was motivated by concerns that humanitarian food aid may be ineffective and may even promote civil conflicts. The main challenge of this study is the potential endogeneity of US food aid due to reverse causality and joint determination. Their identification strategy relies on using the product of the lagged US wheat production and the average probability of receiving any US food aid for each country as the instrumental variable for wheat aid. Nunn and Qian (2014) estimate many variations of the basic binary discrete model and consider different kinds of wars, different controls and alternative specifications.

Herein, we focus on the simple cross-sectional specification considered in Nunn and Qian (2014). More specifically, we estimate the impact of US wheat aid on the probability of civil war *onset* after a period of peace (column (3), Table 7, Nunn and Qian (2014)), using precisely the same model specification as in Nunn and Qian (2014).[19] We examine the IV strength by applying our DJ test to the model, as well as the three conventional weak IV tests for linear models.

The dataset in this analysis involves observations on 78 non-OECD countries from 1971 to 2006, and the observations used for the onset analysis are those country-year observations that have no intra-state civil conflict in the previous period. The event indicator for civil war onset is

[17] The Kleibergen-Paap F-statistic is utilized when allowing for heteroscedastic standard error. The first stage F-statistic and the Cragg-Donald statistic are 155.31 when assuming homoscedastic standard error. SY rejects its null according to the 5% critical values based on the 2SLS relative bias of one endogenous regressor and three instruments case.

[18] The estimated effective degrees of freedom of the Robust test for the tolerance threshold $\tau \in \{5\%, 10\%, 20\%, 30\%\}$ are all about 2.4. See Olea and Pflueger (2013) for the definitions of the effective F-statistic, the tolerance threshold τ and the effective degrees of freedom. The Robust test statistic and the critical values are obtained using the Stata command "weakivtest" (Pflueger and Wang (2014)) under heteroscedastic-robust estimation.

[19] Data sets used to construct the incidence of conflict, US food aid, US wheat production and other variables include the UCDP/PRIO Armed Conflict Dataset Version 4-2010, the Food and Agriculture Organization's FAO-STAT database and the data from the United States Department of Agriculture. See Nunn and Qian (2014) for more detailed information.

set to be one if it is the first period of a intra-state conflict episode, and zero otherwise. Nunn and Qian (2014) estimate a Logistic discrete time hazard model for the probability of onset of war, controlling for the previous duration of peace up to a third degree of polynomial, using a control function approach with the US wheat aid in year t being instrumented by the product of US wheat production in year $t-1$ and the probability of receiving any US food aid between 1971 and 2006 for each country (column (3), Table 7, Nunn and Qian (2014)). To be consistent with the setup of the paper, we estimate a probit link model rather than a logit. Summary statistics for the data used in the onset analysis can be found in part (a) of Table 8.

Part (a) of Table 10 presents results for the estimated coefficients and average partial effects from both 2SCML[20] and CU-GMM with the degree of overidentification equal to unity. For CU-GMM, we use $a_i = (X_i', Z_i, Z_i^2, Z_i^3, Z_i x_{1,i}, \mathbf{0}_{k+1}')'$ and $b_i = (\mathbf{0}_{k+3}', 1, Z_i, X_i')'$ to construct moments. X_i, Z_i denotes the standardized variables of exogenous regressors and the instrument, $x_{1,i}$ is the non-standardized onset duration, and k is the number of exogenous regressors and the intercept in the structural equation. For comparison purposes, column (1) of Table 10 gives the estimated partial effect of US wheat aid reported by Nunn and Qian (2014) using a 2SCML logit approach, which is a key result for their analysis. Columns (2) and (5) of Table 10 demonstrate that the IV is significantly related to the endogenous regressor of wheat aid at the 1% significant level by both estimation methods. However, it is also worth noting that the estimates of interest, the effects of the US wheat aid on onset, are statistically insignificant for both approaches, as well as for the Nunn and Qian (2014) estimate in column (1), and they even differ in signs.[21] Estimates for other coefficients are quite stable and similar across the three sets of results. Finally, Hansen's J-statistic is 0.299, less than the critical value $\chi^2_{0.95}(1) = 3.84$, thus we cannot reject the null that moments are all valid. This evidence leads to the suspicion that the potential weakness of the IV could be one of the possible reasons for the unstable estimates of the US wheat aid coefficient.

This suspicion is verified by the DJ test. The perturbation for the onset analysis is chosen as (16). Panel (a) of Table 9, demonstrates that the DJ test cannot reject the null of weak instruments. Hansen's J-statistic and the distorted J-statistic both are approximately 0.30, which indicates that a small deviation has no effects on the value of the J-statistic. The distorted J-statistic is significantly less than its 5% critical value $\chi^2(H-k) = 15.51$ with $H = 12$ and $k = 4$, failing to reject the null of weak IV. However, when we apply the conventional SS, SY and Robust tests to the onset of the civil conflict model, the SS and SY tests return a rejection of the weak IV hypothesis and the Robust test also rejects the null if the tolerance threshold $\tau \geq 10\%$. As shown in Table 10, the first stage Kleibergen-Paap F-statistic for SS and SY is 26.07, much larger than 10 and the 5% critical value 16.38 of SY.[22] The Robust test effective F-statistic 26.39 is also larger than its 5% critical value 23.11.[23] In summary, for this example, the conventional weak IV tests and the DJ test suggest opposite results. This serves as a reminder that applying conventional weak IV tests for linear models to binary outcome models can lead to incorrect decisions in certain

[20]The 2SCML in this example allows intragroup correlation for standard errors, clustered by countries.

[21]The statistical insignificance of the US food aid on onset of civil conflict is also pointed out by Nunn and Qian (2014).

[22]To be consistent with Nunn and Qian (2014), standard errors (s.e.) are computed using clustered s.e. by countries. Kleibergen-Paap F-statistic (Kleibergen and Paap, 2006) is utilized when allowing for intragroup correlation s.e. The critical value 16.38 is the 5% critical value for the SY test based on 2SLS Wald test size, for the case of one endogenous variable and one IV with maximal size 10% of a 5% Wald test.

[23]The effective F-statistic and critical values are computed using the Stata command "weakivtest" (Pflueger and Wang (2014)). The critical value 23.11 is for the case of effective degrees of freedom one and a threshold tolerance $\tau = 10\%$.

circumstances, especially when IV possesses large variance.

Subsequently, we have repeated this analysis within the other 5 models considered in Nunn and Qian (2014) (columns (4)-(8) of Table 7, Nunn and Qian (2014)), which include different specifications and exogenous regressors. In all cases, our DJ test shows that we cannot rule out the possibility of weak instruments, whilst the SS and SY tests all result in a rejection of weak instrument hypothesis.[24] The Robust test also rejects the null in some cases.[25] In part (b) of Table 10, we report the estimation results for the probability of *offset* of civil war after a period of war (column (6) of Table 7, Nunn and Qian (2014)). One important result to note is that Nunn and Qian (2014) estimate a significant and negative effect for offset of war, indicating that aid prolongs civil wars with 1,000 MT extra of US wheat aid reducing the probability of civil war offset by 0.04 percentage point. Given that in the presence of weak instrument, the estimated treatment effect is no longer consistent, this finding needs to be taken with caution.

7 Conclusion

Empirical researchers often confront the issue of endogeneity when studying causal effects. In most of these situations, instrumental variable methods are used to identify and consistently estimate structural parameters. However, weak instruments is a common phenomenon, which will lead to inconsistent estimation and invalid parameter inference. This paper introduces a weak instrumental variables test for discrete choice models. Our proposed test is asymptotically conservative and consistent. While building on the general testing strategy of Antoine and Renault (2017), the test proposed in this paper is based on a null hypothesis of genuine instrument weakness, and not the nearly-strong identification null hypothesis analyzed in Antoine and Renault (2017). In particular, we demonstrate that common discrete choice models are 'close enough' to linear models to allow for such a test to have power against alternatives that represent arbitrary departures from weak instruments. A generalization of the rule-of-thumb is proposed for discrete choice models, and compared with the rule-of-thumb for linear models. This generalized rule sheds new light on how to measure the instrument strength in discrete models. Monte Carlo simulations confirm the failure of conventional linear weak instrument tests when the DGP is actually a discrete model, and also demonstrate the good performance of our proposed testing approach.

References

D. W. Andrews and X. Cheng. Estimation and inference with weak, semi-strong, and strong identification. *Econometrica*, 80(5):2153–2211, 2012.

B. Antoine and E. Renault. Efficient GMM with nearly-weak instruments. *The Econometrics Journal*, 12(s1), 2009.

[24]Results are not reported due to space limitation. SS and SY tests are based on Kleibergen-Paap F-statistic (Kleibergen and Paap, 2006). DJ test is implemented using the same a_i and b_i with those used to get the CU-GMM in Table 10, replacing X_i by the standardized exogenous regressors specified in each models. The perturbation for each model is chosen as (16).

[25]Based on the critical value 23.11 ($\tau = 10\%$), the Robust test rejects weak IV of the analysis in columns (4) and (8), but fails to reject in columns (5), (6) and (7). Results are obtained by using the Stata command "weakivtest" Pflueger and Wang (2014).

B. Antoine and E. Renault. Efficient minimum distance estimation with multiple rates of convergence. *Journal of Econometrics*, 170(2):350–367, 2012.

B. Antoine and E. Renault. Testing identification strength. 2017.

J. N. Arendt. Does education cause better health? A panel data analysis using school reforms for identification. *Economics of Education review*, 24(2):149–160, 2005.

J. H. Block, L. Hoogerheide, and R. Thurik. Education and entrepreneurial choice: An instrumental variables analysis. *International Small Business Journal*, 31(1):23–33, 2013.

R. W. Blundell and J. L. Powell. Endogeneity in semiparametric binary response models. *The Review of Economic Studies*, 71(3):655–679, 2004.

M. Caner. Testing, estimation in gmm and cue with nearly-weak identification. *Econometric Reviews*, 29(3):330–363, 2009.

J. Cawley and C. Meyerhoefer. The medical care costs of obesity: An instrumental variables approach. *Journal of health economics*, 31(1):219–230, 2012.

J.-M. Dufour and J. Wilde. Weak identification in probit models with endogenous covariates. Technical report, Working Paper, Institute of Empirical Economic Research, University of Osnabrück, 2013.

B. E. Ellison. Two theorems for inferences about the normal distribution with applications in acceptance sampling. *Journal of the American Statistical Association*, 59(305):89–95, 1964.

U. Goto and T. Iizuka. Cartel sustainability in retail markets: Evidence from a health service sector. *International Journal of Industrial Organization*, 49:36–58, 2016.

J. Hahn and J. Hausman. A new specification test for the validity of instrumental variables. *Econometrica*, 70(1):163–189, 2002.

J. Hahn and G. Kuersteiner. Discontinuities of weak instrument limiting distributions. *Economics Letters*, 75(3):325–331, 2002.

L. P. Hansen. Large sample properties of generalized method of moments estimators. *Econometrica: Journal of the Econometric Society*, pages 1029–1054, 1982.

L. P. Hansen, J. Heaton, and A. Yaron. Finite-sample properties of some alternative GMM estimators. *Journal of Business & Economic Statistics*, 14(3):262–280, 1996.

D. Kawaguchi, Y. Matsushita, and H. Naito. Moment estimation of the probit model with an endogenous continuous regressor. *The Japanese Economic Review*, 68(1):48–62, 2017.

T. Kinda. Investment climate and FDI in developing countries: firm-level evidence. *World development*, 38(4):498–513, 2010.

F. Kleibergen. Testing parameters in gmm without assuming that they are identified. *Econometrica*, 73(4):1103–1123, 2005.

F. Kleibergen and R. Paap. Generalized reduced rank tests using the singular value decomposition. *Journal of econometrics*, 133(1):97–126, 2006.

L. Lochner and E. Moretti. The effect of education on crime: Evidence from prison inmates, arrests, and self-reports. *American economic review*, 94(1):155–189, 2004.

L. M. Magnusson. Weak instruments robust tests for limited dependent variable models. Technical report, Working Paper, Brown University (RI), 2007.

L. M. Magnusson. Inference in limited dependent variable models robust to weak identification. *The Econometrics Journal*, 13(3), 2010.

D. McKenzie and H. Rapoport. Can migration reduce educational attainment? Evidence from Mexico. *Journal of Population Economics*, 24(4):1331–1358, 2011.

E. Miguel, S. Satyanath, and E. Sergenti. Economic shocks and civil conflict: An instrumental variables approach. *Journal of political Economy*, 112(4):725–753, 2004.

T. A. Mroz. The sensitivity of an empirical model of married women's hours of work to economic and statistical assumptions. *Econometrica: Journal of the Econometric Society*, pages 765–799, 1987.

N. Nunn and N. Qian. Us food aid and civil conflict. *American Economic Review*, 104(6):1630–66, 2014.

J. L. M. Olea and C. Pflueger. A robust test for weak instruments. *Journal of Business & Economic Statistics*, 31(3):358–369, 2013.

D. B. Owen. A table of normal integrals. *Communications in Statistics - Simulation and Computation*, 9(4):389–419, 1980.

C. E. Pflueger and S. Wang. A robust test for weak instruments in stata. 2014.

D. Poskitt and C. L. Skeels. Assessing the magnitude of the concentration parameter in a simultaneous equations model. *The Econometrics Journal*, 12(1):26–44, 2009.

L. M. Powell, J. A. Tauras, and H. Ross. The importance of peer effects, cigarette prices and tobacco control policies for youth smoking behavior. *Journal of health Economics*, 24(5):950–968, 2005.

D. Rivers and Q. H. Vuong. Limited information estimators and exogeneity tests for simultaneous probit models. *Journal of econometrics*, 39(3):347–366, 1988.

J. E. Ruseski, B. R. Humphreys, K. Hallman, P. Wicker, and C. Breuer. Sport participation and subjective well-being: instrumental variable results from german survey data. *Journal of Physical Activity and Health*, 11(2):396–403, 2014.

D. Staiger and J. H. Stock. Instrumental variables regression with weak instruments. *Econometrica*, 65(3):557–586, 1997.

J. H. Stock and J. H. Wright. GMM with weak identification. *Econometrica*, 68(5):1055–1096, 2000.

J. H. Stock and M. Yogo. Testing for weak instruments in linear IV regression. Chapter 5 in identification and inference in econometric models: Essays in honor of Thomas J. Rothenberg, edited by DWK Andrews and JH Stock, 2005.

J. H. Stock, J. H. Wright, and M. Yogo. A survey of weak instruments and weak identification in generalized method of moments. *Journal of Business & Economic Statistics*, 20(4):518–529, 2002.

A. W. van der Vaart and J. A. Wellner. Weak convergence. In *Weak convergence and empirical processes*, pages 16–28. Springer, 1996.

J. M. Wooldridge. *Econometric analysis of cross section and panel data*. MIT press, 2010.

A Appendix

A.1 Lemmas

In this section we collect several results that will help us to derive the asymptotic behavior of the distorted J-test.

Lemma 7 *Let $Z \sim \mathcal{N}(\mu_z, \sigma_z^2)$, then $\mathbb{E}[\Phi(Z)] = \Phi[\mu_z/\sqrt{1+\sigma_z^2}]$.*

Proof. The proof follows similar to Theorem 2 in Ellison (1964). Let $X \sim \mathcal{N}(0,1)$ be independent of Z. Then,

$$\begin{aligned}\mathbb{E}[\Phi(Z)] &= \mathbb{E}_Z\left[\mathrm{P}_x\left\{X \leq z | Z = z\right\}\right] \\ &= \mathrm{P}_{X,Z}\left[X \leq Z\right] \\ &= \mathrm{P}_{X,Z}\left[\left(X - Z\right)/\sqrt{1+\sigma_z^2} \leq 0\right].\end{aligned}$$

Noting that $(X-Z)/\sqrt{1+\sigma_z^2}$ is normal with mean $\delta = -\mu_z/\sqrt{1+\sigma_z^2}$ and unit variance,

$$\begin{aligned}\mathrm{P}_{X,Z}\left[\left(X - Z\right)/\sqrt{1+\sigma_z^2} \leq 0\right] &= \mathrm{P}_{X,Z}\left[\left(X - Z\right)/\sqrt{1+\sigma_z^2} - \delta \leq -\delta\right] \\ &= \Phi\left(\mu_z/\sqrt{1+\sigma_z^2}\right).\end{aligned}$$

■

Lemma 7 allows us to deduce the following corollary.

Lemma 8 *If $Z \sim \mathcal{N}(\mu_z, \sigma_z^2)$, and $Y|Z = z \sim \mathcal{N}[\mu_y + \xi(z - \mu_z), (1-\xi^2)]$, then for constants c, b*

$$\mathbb{E}\left[\Phi\left(cY + bZ\right)\right] = \Phi\left[\frac{D_1}{\sqrt{1+\sigma_z^2 D_2^2}} + \frac{D_2\mu_z}{\sqrt{1+\sigma_z^2 D_2^2}}\right]$$

where

$$D_1 = \frac{c(\mu_y - \xi\mu_z)}{\sqrt{1+c^2(1-\xi^2)}} \text{ and } D_2 = \frac{c\xi + b}{\sqrt{1+c^2(1-\xi^2)}}.$$

Proof. Define $X = cY + bZ$ so that $X|Z = z \sim \mathcal{N}\left[c(\mu_y - \xi\mu_z) + (c\xi + b)z, c^2(1-\xi^2)\right]$. By the law of iterated expectations:

$$\mathbb{E}_{Y,Z}[\Phi(cY + bZ)] = \mathbb{E}_Z[\mathbb{E}_X[\Phi(X)|Z = z]].$$

Now, applying Lemma 7 to the conditional expectation on the rhs of the above,

$$\mathbb{E}_X[\Phi(X)|Z = z] = \Phi\left[\frac{c(\mu_y - \xi\mu_z) + (c\xi + b)z}{\sqrt{1 + c^2(1-\xi^2)}}\right].$$

Defining $\tilde{Z} = D_1 + D_2 Z$ and again applying Lemma 7, we have

$$\begin{aligned}\mathbb{E}_{Y,Z}[\Phi(cY + bZ)] &= \mathbb{E}_Z\left[\Phi\left[\frac{c(\mu_y - \xi\mu_z) + (c\xi + b)Z}{\sqrt{1 + c^2(1-\xi^2)}}\right]\right] \\ &= \mathbb{E}_{\tilde{Z}}\left[\Phi(\tilde{Z})\right] \\ &= \Phi\left[\frac{D_1 + D_2\mu_z}{\sqrt{1 + \sigma_z^2 D_2^2}}\right].\end{aligned} \tag{24}$$

■

The following result is from Owen (1980).

Lemma 9 *Let $Z \sim \mathcal{N}(0,1)$, then $\mathbb{E}[Z\Phi(a + bZ)] = \frac{b}{\sqrt{1+b^2}}\phi(a/\sqrt{1+b^2})$, where $\phi(\cdot)$ is the standard normal pdf.*

Lemma 10 *Suppose the data is generated by the probit model* (20) *and with $z_i \sim \mathcal{N}(0, \sigma_z^2)$, and iid, but we fit it to a linear probability model* (21)*, then we have*

$$\widetilde{\alpha}^* = \mathbb{E}[\phi_i(\theta)\alpha],$$

where $\widetilde{\alpha}^*$ is the pseudo true value that minimizes the 2SLS objective function; and $\mathbb{E}[\phi_i(\theta)\alpha]$ is the mean of the marginal effect of $y_{2,i}$ on $y_{1,i}$ for the probit model (20) *with $\phi_i(\theta) = \phi(\beta + \alpha y_{2,i} + \widetilde{\rho}[y_{2,i} - \pi - \xi z_i])$.*

Proof of Lemma 10.
While this lemma is stated for $z_i \sim \mathcal{N}(0, \sigma_z^2)$, we note here that this result remains valid across a wide range of DGPs for z_i. Following Wooldridge (2010) Section 2.2.5, the marginal effect (or partial effect) of $y_{2,i}$ on $y_{1,i}$ of the probit model (20) is

$$\frac{\partial \mathbb{E}[y_{1,i}|y_{2,i}, v_i]}{\partial y_{2,i}} = \frac{\partial \Phi(\beta + \alpha y_{2,i} + \widetilde{\rho} v_i)}{\partial y_{2,i}} = \phi(\beta + \alpha y_{2,i} + \widetilde{\rho} v_i)\alpha. \tag{25}$$

If we fit the data generated from the probit model in (20) to the linear probability model $y_{1,i} = \widetilde{\beta} + \widetilde{\alpha} y_{2,i} + \widetilde{u}_i$ using 2SLS, then the pseudo true $\widetilde{\alpha}^*$ that minimizes the 2SLS objective function is

$$\widetilde{\alpha}^* = \frac{\mathbb{E}[y_{1,i} z_i]}{\mathbb{E}[y_{2,i} z_i]} = \frac{\mathbb{E}_{Y_2,Z}\{z_i \mathbb{E}[y_{1,i}|y_{2,i}, z_i]\}}{\xi\sigma_z^2} = \frac{\mathbb{E}_{Y_2,Z}\{z_i \Phi(\beta + \alpha y_{2,i} + \widetilde{\rho}[y_{2,i} - \pi - \xi z_i])\}}{\xi\sigma_z^2}. \tag{26}$$

Denote $w_i = \beta + \alpha y_{2,i} + \widetilde{\rho}[y_{2,i} - \pi - \xi z_i]$. Because $y_{2,i}|z_i \sim \mathcal{N}(\pi + \xi z_i, \sigma_v^2)$, then $w_i|z_i \sim \mathcal{N}(\beta + \alpha(\pi + \xi z_i), (\alpha + \widetilde{\rho})^2\sigma_v^2)$ and based on Lemmas 8 and 9, $\widetilde{\alpha}^*$ is given by

$$\begin{aligned}\widetilde{\alpha}^* &= \frac{1}{\xi\sigma_z^2}\mathbb{E}_{Y_2,Z}[z_i\Phi(w_i)] = \frac{1}{\xi\sigma_z^2}\mathbb{E}_Z\{z_i\mathbb{E}[\Phi(w_i)|z_i]\} = \frac{1}{\xi\sigma_z^2}\mathbb{E}_Z\left\{z_i\Phi\left(\frac{\beta + \alpha(\pi + \xi z_i)}{\sqrt{1 + (\alpha + \widetilde{\rho})^2\sigma_v^2}}\right)\right\} \\ &= \frac{\alpha}{\sqrt{1 + (\alpha + \widetilde{\rho})^2\sigma_v^2 + (\alpha\xi)^2\sigma_z^2}}\phi\left(\frac{\beta + \alpha\pi}{\sqrt{1 + (\alpha + \widetilde{\rho})^2\sigma_v^2 + (\alpha\xi)^2\sigma_z^2}}\right). \qquad (27)\end{aligned}$$

In addition, according to Owen (1980) (page 396) property of the integral of normal pdf, we have that the mean of the probit model marginal effect in equation (25) as

$$\begin{aligned}\mathbb{E}[\phi(\beta + \alpha y_{2,i} + \widetilde{\rho}v_i)\alpha] &= \mathbb{E}[\phi(w_i)\alpha] = \alpha\mathbb{E}_Z\{\mathbb{E}[\phi(w_i)|z_i]\} \\ &= \frac{\alpha}{\sqrt{1 + (\alpha + \widetilde{\rho})^2\sigma_v^2}}\mathbb{E}_Z\left[\phi\left(\frac{\beta + \alpha(\pi + \xi z_i)}{\sqrt{1 + (\alpha + \widetilde{\rho})^2\sigma_v^2}}\right)\right] \\ &= \frac{\alpha}{\sqrt{1 + (\alpha + \widetilde{\rho})^2\sigma_v^2 + (\alpha\xi)^2\sigma_z^2}}\phi\left(\frac{\beta + \alpha\pi}{\sqrt{1 + (\alpha + \widetilde{\rho})^2\sigma_v^2 + (\alpha\xi)^2\sigma_z^2}}\right). \qquad (28)\end{aligned}$$

From (27) and (28), we can conclude that $\widetilde{\alpha}^* = \mathbb{E}[\phi(\beta + \alpha y_{2,i} + \widetilde{\rho}[y_{2,i} - \pi - \xi z_i])\alpha]$. ■

A.2 Proofs

Proof of Lemma 1. Let $\bar{g}_n(\theta) = (\bar{g}_{1,n}(\theta), \bar{g}_{2,n}(\theta), ..., \bar{g}_{H,n}(\theta))'$. The mean value expansion of $\frac{\partial\bar{g}_{l,n}(\theta_n)}{\partial\theta'}$ at θ^0 yields

$$\frac{\partial\bar{g}_{l,n}(\theta_n)}{\partial\theta'} = \frac{\partial\bar{g}_{l,n}(\theta^0)}{\partial\theta'} + (\theta_n - \theta^0)'\frac{\partial^2\bar{g}_{l,n}(\widetilde{\theta}_n)}{\partial\theta'\partial\theta}, \quad l = 1, 2, ..., H$$

where $\widetilde{\theta}_n$ is component-by-component between θ^0 and θ_n. By the structure of the moment $\bar{g}_n(\theta)$ and the fact that $\phi(\cdot)$ and its derivative, a_i and b_i are all measurable, it is not hard to prove that $\|\theta_n - \theta^0\| = o_p(1)$ implies the Hessian multiplied by A_n, $\frac{\partial^2\bar{g}_{l,n}(\widetilde{\theta}_n)}{\partial\theta'\partial\theta}A_n = O_p(1)$ for $l = 1, 2, ..., H$. Therefore, $\|\theta_n - \theta^0\| = o_p(1)$ and (11) imply that Lemma 1 holds. ■

Proof of Lemma 2. We first prove two intermediate lemmas, Lemma 11 and Lemma 12, that are needed to obtain the result.

Lemma 11 *Under Assumptions (A1)-(A4), there exists a continuous map $\theta \mapsto \gamma(\theta)$ from $\Theta \subset \mathbb{R}^{d_\theta}$, compact, to $\mathbb{R}^H$, $H \geq d_\theta$, such that: $\gamma(\theta) = 0 \iff \theta = \theta^0$*

Proof. Let $r_i(\theta) = \left(y_{1i} - \Phi\left[(\alpha + \rho)\, y_{2,i} - \rho\xi z_i\right],\ \ y_{2i} - \mu_y - \xi z_i\right)'$ and let $a_i = (1, z_i)'$. We demonstrate this result for the following simplified version of $g_i(\theta)$ in equation (7), and note that the result for the moments in (7) follows as a direct consequence:

$$g_i(\theta) := a_i \otimes r_i(\theta) = \begin{pmatrix} y_{1,i} - \Phi\left[(\alpha + \rho)\, y_{2,i} - \rho\xi z_i\right] \\ z_i\left(y_{1,i} - \Phi\left[(\alpha + \rho)\, y_{2,i} - \rho\xi z_i\right]\right) \\ (y_{2,i} - \mu_y - \xi z_i) \\ z_i(y_{2,i} - \mu_y - \xi z_i) \end{pmatrix}.$$

The linear nature of the second part of the moment function in (7) ensures that if the result is satisfied for this alternative choice of g_i, the result will follow for the more complicated version in (7).

Define

$$\gamma(\theta) = \begin{pmatrix} \gamma_1(\theta) \\ \gamma_2(\theta) \\ \gamma_3(\theta) \\ \gamma_4(\theta) \end{pmatrix} = \begin{pmatrix} \mathbb{E}\left[\Phi\left(\left\{\alpha^0 + \rho^0\right\} y_{2,i} - \rho^0\xi^0 z_i\right)\right] - \mathbb{E}\left[\Phi\left(\left\{\alpha + \rho\right\} y_{2,i} - \rho\xi z_i\right)\right] \\ \mathbb{E}[z_i(y_{1,i} - \Phi\left(\left\{\alpha + \rho\right\} y_{2,i} - \rho\xi z_i\right))] \\ \mathbb{E}[y_{2,i} - \mu_y - \xi z_i] \\ \mathbb{E}[z_i(y_{2,i} - \mu_y - \xi z_i)] \end{pmatrix}.$$

First, considering $\gamma_1(\theta)$ and applying Lemma 8 we have

$$\begin{aligned} \gamma_1(\theta) =& \Phi\left[\frac{\left\{\alpha^0 + \rho^0\right\}\mu_y^0}{\sqrt{1 + \left\{\alpha^0 + \rho^0\right\}^2(1 - (\xi^0)^2) + (\alpha^0\xi^0)^2}}\right] \\ & - \Phi\left[\frac{\left\{\alpha + \rho\right\}\mu_y^0}{\sqrt{1 + \left\{\alpha + \rho\right\}^2(1 - (\xi^0)^2) + (\alpha\xi^0 + \rho\{\xi^0 - \xi\})^2}}\right]. \end{aligned} \tag{29}$$

Now, considering $\gamma_2(\theta)$ we have

$$\begin{aligned} \gamma_2(\theta) &:= Q_1(\theta^0) - Q_2(\theta) \\ &= \mathbb{E}[z_i\Phi\left(\left\{\alpha^0 + \rho^0\right\} y_{2,i} - \rho^0\xi^0 z_i\right)] - \mathbb{E}[z_i\Phi\left(\left\{\alpha + \rho\right\} y_{2,i} - \rho\xi z_i\right)]. \end{aligned}$$

First, focus on $Q_1(\theta^0)$. Under the assumption that $z_i \sim \mathcal{N}(0,1)$, we can apply the conditional expectation in equation (24), obtained as part of Lemma 8, which yields

$$\begin{aligned} Q_1(\theta^0) &= \mathbb{E}_Z\left[z_i\mathbb{E}\left[\Phi\left(\left\{\alpha^0 + \rho^0\right\} y_{2,i} - \rho^0\xi^0 z_i\right)|Z = z_i\right]\right] \\ &= \mathbb{E}_Z\left[z_i\Phi\left(a_0 + b_0 z_i\right)\right], \end{aligned}$$

where

$$a_0 = \frac{\left\{\alpha^0 + \rho^0\right\}\mu_y^0}{\sqrt{1 + \left\{\alpha^0 + \rho^0\right\}^2(1 - (\xi^0)^2)}} \text{ and } b_0 = \frac{\alpha^0\xi^0}{\sqrt{1 + \left\{\alpha^0 + \rho^0\right\}^2(1 - (\xi^0)^2)}}.$$

Now, we apply Lemma 9 to obtain, after simplification,

$$\begin{aligned} Q_1(\theta^0) =& \mathbb{E}_z\left[z_i\Phi\left(a_0 + b_0 z_i\right)\right] \\ =& \xi^0\frac{\alpha^0}{\sqrt{1 + \left\{\alpha^0 + \rho^0\right\}^2(1 - (\xi^0)^2) + (\alpha^0\xi^0)^2}}\phi\left[\frac{\left\{\alpha^0 + \rho^0\right\}\mu_y^0}{\sqrt{1 + \left\{\alpha^0 + \rho^0\right\}^2(1 - (\xi^0)^2) + (\alpha^0\xi^0)^2}}\right], \end{aligned}$$

where $\phi(x)$ denotes the standard normal density evaluated at $x \in \mathbb{R}$. Likewise, we have that

$$\begin{aligned} Q_2(\theta) =& \mathbb{E}_Z\left[z_i\mathbb{E}\left[\Phi\left(\left\{\alpha + \rho\right\} y_{2,i} - \rho\xi z_i\right)|Z = z_i\right]\right] \\ =& \mathbb{E}_Z\left[z_i\Phi\left(\frac{\left\{\alpha + \rho\right\}\mu_y^0}{\sqrt{1 + \left\{\alpha + \rho\right\}^2(1 - (\xi^0)^2)}} + \frac{\left\{\alpha + \rho\right\}\xi^0 - \rho\xi}{\sqrt{1 + \left\{\alpha + \rho\right\}^2(1 - (\xi^0)^2)}}z_i\right)\right] \\ =& \frac{\left\{\alpha + \rho\right\}\xi^0 - \rho\xi}{\sqrt{1 + \left\{\alpha + \rho\right\}^2(1 - (\xi^0)^2) + [\left\{\alpha + \rho\right\}\xi^0 - \rho\xi]^2}} \times \\ & \phi\left[\frac{\left\{\alpha + \rho\right\}\mu_y^0}{\sqrt{1 + \left\{\alpha + \rho\right\}^2(1 - (\xi^0)^2) + [\left\{\alpha + \rho\right\}\xi^0 - \rho\xi]^2}}\right] \end{aligned}$$

From $\gamma_3(\theta)$ and $\gamma_4(\theta)$, we have that, by Assumption (A1),

$$\begin{pmatrix}\gamma_3(\theta)\\ \gamma_4(\theta)\end{pmatrix} = \begin{pmatrix}\mathbb{E}_Z[\mathbb{E}[y_{2,i}|Z=z_i] - \mu_y - \xi z_i]\\ \mathbb{E}_Z[z_i\mathbb{E}[y_{2,i}|Z=z_i] - z_i\mu_y - z_i\xi z_i]\end{pmatrix} = 0 \iff \xi = \xi^0, \mu_y = \mu_y^0.$$

We can therefore take them as fixed, which allows us to re-write $\gamma_1(\theta)$ as

$$\begin{aligned}\gamma_1(\theta) =& \Phi\left[\frac{\{\alpha^0+\rho^0\}\,\mu_y^0}{\sqrt{1+\{\alpha^0+\rho^0\}^2(1-(\xi^0)^2)+(\alpha^0\xi^0)^2}}\right]\\ &- \Phi\left[\frac{\{\alpha+\rho\}\,\mu_y^0}{\sqrt{1+\{\alpha+\rho\}^2(1-(\xi^0)^2)+(\alpha\xi^0)^2}}\right].\end{aligned}$$

For Z a standard normal random variable, $\gamma_1(\theta)$ can be rewritten as

$$\gamma_1(\theta) = \Pr\left[\frac{\{\alpha+\rho\}\mu_y^0}{\sqrt{1+\{\alpha+\rho\}^2(1-(\xi^0)^2)+(\alpha\xi^0)^2}} \le Z \le \frac{\{\alpha^0+\rho^0\}\mu_y^0}{\sqrt{1+\{\alpha^0+\rho^0\}^2(1-(\xi^0)^2)+(\alpha^0\xi^0)^2}}\right]. \tag{30}$$

Applying $\xi = \xi^0$ to $\gamma_2(\theta)$, we have

$$\begin{aligned}\gamma_2(\theta) &= Q_1(\theta^0) - Q_2(\theta)\\ &= \xi^0\frac{\alpha^0}{\sqrt{1+\{\alpha^0+\rho^0\}^2(1-(\xi^0)^2)+(\alpha^0\xi^0)^2}}\times\\ &\quad\phi\left[\frac{\{\alpha^0+\rho^0\}\,\mu_y^0}{\sqrt{(1+\{\alpha^0+\rho^0\}^2(1-(\xi^0)^2)+(\alpha^0\xi^0)^2)}}\right]\\ &\quad- \xi^0\frac{\alpha}{\sqrt{1+\{\alpha+\rho\}^2(1-(\xi^0)^2)+(\alpha\xi^0)^2}}\times\\ &\quad\phi\left[\frac{\{\alpha+\rho\}\mu_y^0}{\sqrt{1+\{\alpha+\rho\}^2(1-(\xi^0)^2)+(\alpha\xi^0)^2}}\right].\end{aligned} \tag{31}$$

From the absolute continuity of Z,

$$\gamma_1(\theta) = 0 \iff \frac{\{\alpha+\rho\}\mu_y^0}{\sqrt{1+\{\alpha+\rho\}^2(1-(\xi^0)^2)+(\alpha\xi^0)^2}} = \frac{\{\alpha^0+\rho^0\}\mu_y^0}{\sqrt{1+\{\alpha^0+\rho^0\}^2(1-(\xi^0)^2)+(\alpha^0\xi^0)^2}}.$$

Applying this relationship to equation (31), we can restate $\gamma_2(\theta)$ as:

$$\begin{aligned}\gamma_2(\theta) =& \xi^0\cdot\phi\left[\frac{\{\alpha^0+\rho^0\}\,\mu_y^0}{\sqrt{(1+\{\alpha^0+\rho^0\}^2(1-(\xi^0)^2)+(\alpha^0\xi^0)^2)}}\right]\times\\ &\left[\frac{\alpha^0}{\sqrt{1+\{\alpha^0+\rho^0\}^2(1-(\xi^0)^2)+(\alpha^0\xi^0)^2}} - \frac{\alpha}{\sqrt{1+\{\alpha+\rho\}^2(1-(\xi^0)^2)+(\alpha\xi^0)^2}}\right]\end{aligned}$$

and we have

$$\gamma_2(\theta) = 0 \iff \left[\frac{\alpha^0}{\sqrt{1+\{\alpha^0+\rho^0\}^2(1-(\xi^0)^2)+(\alpha^0\xi^0)^2}} - \frac{\alpha}{\sqrt{1+\{\alpha+\rho\}^2(1-(\xi^0)^2)+(\alpha\xi^0)^2}}\right] = 0.$$

Now, we can rewrite the condition that $\gamma_1(\theta) = 0$ as

$$\{\alpha + \rho\} = \{\alpha^0 + \rho^0\} \frac{\sqrt{1 + \{\alpha + \rho\}^2(1 - (\xi^0)^2) + (\alpha\xi^0)^2}}{\sqrt{1 + \{\alpha^0 + \rho^0\}^2(1 - (\xi^0)^2) + (\alpha^0\xi^0)^2}}. \tag{32}$$

Likewise, the condition that $\gamma_2(\theta) = 0$ can be rewritten as

$$\alpha = \alpha^0 \frac{\sqrt{1 + \{\alpha + \rho\}^2(1 - (\xi^0)^2) + (\alpha\xi^0)^2}}{\sqrt{1 + \{\alpha^0 + \rho^0\}^2(1 - (\xi^0)^2) + (\alpha^0\xi^0)^2}}$$

Therefore, for $\alpha^0 \neq 0$ and $\alpha^0 + \rho^0 \neq 0$, we have

$$\alpha/\alpha^0 = \frac{\sqrt{1 + \{\alpha + \rho\}^2(1 - (\xi^0)^2) + (\alpha\xi^0)^2}}{\sqrt{1 + \{\alpha^0 + \rho^0\}^2(1 - (\xi^0)^2) + (\alpha^0\xi^0)^2}} = \{\alpha + \rho\}/\{\alpha^0 + \rho^0\}$$

and we obtain

$$\alpha = \frac{\alpha^0\rho}{\rho^0}. \tag{33}$$

Plugging in equation (33) into equation (32), we can re-arrange this condition as

$$0 = \frac{\alpha^0\rho^0}{\sqrt{1 + \{\alpha^0 + \rho^0\}^2(1 - (\xi^0)^2) + (\alpha^0\xi^0)^2}} - \frac{\alpha^0\rho}{\sqrt{1 + \frac{\rho^2}{(\rho^0)^2}\{\alpha^0 + \rho^0\}^2(1 - (\xi^0)^2) + \frac{\rho^2}{(\rho^0)^2}(\alpha^0\xi^0)^2}}$$

which indicates that ρ has the same sign with ρ^0 and can be simplified as

$$\rho/\rho^0 = \left[\frac{\frac{(\rho)^2}{(\rho^0)^2}\left[\frac{(\rho^0)^2}{(\rho)^2} + \{\alpha^0 + \rho^0\}^2(1 - (\xi^0)^2) + (\alpha^0\xi^0)^2\right]}{1 + \{\alpha^0 + \rho^0\}^2(1 - (\xi^0)^2) + (\alpha^0\xi^0)^2}\right]^{1/2}. \tag{34}$$

This equality holds only when $\rho^2 = (\rho^0)^2$, which has a unique solution $\rho = \rho^0$ due to the same sign of ρ^0 and ρ.

In addition, $\rho = \rho^0$ gives

$$\gamma_1(\theta) = 0 \iff \frac{\{\alpha + \rho^0\}\mu_y^0}{\sqrt{1 + \{\alpha + \rho^0\}^2(1 - (\xi^0)^2) + (\alpha\xi^0)^2}} = \frac{\{\alpha^0 + \rho^0\}\mu_y^0}{\sqrt{1 + \{\alpha^0 + \rho^0\}^2(1 - (\xi^0)^2) + (\alpha^0\xi^0)^2}}.$$

The function

$$f(x) = \frac{\{x + \rho^0\}\mu_y^0}{\sqrt{1 + \{x + \rho^0\}^2(1 - (\xi^0)^2) + (x\xi^0)^2}}$$

is (strictly) monotonically increasing for $x \in \mathbb{R}$, $\mu_y^0 \neq 0$, $|\rho^0| < 1$ and $|\xi^0| < 1$. As such,

$$f(x) - \frac{\{\alpha^0 + \rho^0\}\mu_y^0}{\sqrt{1 + \{\alpha^0 + \rho^0\}^2(1 - (\xi^0)^2) + (\alpha^0\xi^0)^2}} = 0 \iff x = \alpha^0.$$

Therefore, since $\alpha = \alpha^0$ is the unique solution to the equations under $\rho = \rho^0$, we can conclude that $(\alpha^0, \rho^0)'$ is the unique solution to the system, and the result follows. ■

Lemma 12 *Under Assumption (A0)-(A4), for* $\nu_n(\theta) := \frac{1}{\sqrt{n}} \sum_{i=1}^n (g_i(\theta) - \mathbb{E}[g_i(\theta)])$,

$$\nu_n(\theta) \Rightarrow \nu(\theta),$$

for $\nu(\theta)$ *a Gaussian process and* $\Rightarrow$ *denotes weak convergence in the sup-norm.*

Proof. First, recall that for $g_i(\theta) = a_i \otimes r_i(\theta)$,

$$\|g_i(\theta)\| = \|a_i \otimes r_i(\theta)\| = \|a_i\| \|r_i(\theta)\|.$$

Under Assumption (A0), a_i is iid and $\mathbb{E}[\|a_i\|^2] < \infty$. The result then follows if we can demonstrate that $r_i(\theta)$ is a Donsker class of functions.

Consider the parameterization $\vartheta = (\vartheta_1', \vartheta_2')'$, where $\vartheta_1 := (\rho\xi, \alpha + \rho)'$, and $\vartheta_2 := (\mu_y, \xi)'$. By compactness of Θ, the new parameter space $V := \{\theta \in \Theta : \vartheta := (\vartheta_1', \vartheta_2')'\}$ is compact. Rewrite $\Phi[\{\alpha + \rho\}y_{2,i} - \rho\xi z_i] = \Phi[x_i'\vartheta_1]$ for $x_i = (-z_i, y_{2,i})'$. Define the class of functions

$$\mathcal{F} := \left\{\vartheta \in V : r_1(\vartheta_1) := (y_{1,i} - \Phi[x_i'\vartheta_1]), r_2(\vartheta_2) := (y_{2,i} - a_i'\vartheta_2)\right\}.$$

from the compactness of V, $(\mathcal{F}, \|\cdot\|)$ is totally bounded with $\|\cdot\|$ the Euclidean norm.

First, focus on $r_1(\vartheta_1)$. For every x_i, for $\vartheta_1, \bar{\vartheta}_1 \in V_1$, with V_1 the subspace of V associated with ϑ_1, and $x_i'\vartheta \geq x_i'\bar{\vartheta}$

$$\|r_1(\vartheta_1) - r_1(\bar{\vartheta}_1)\|$$
$$= |\Phi(x_i'\vartheta_1) - \Phi(x_i'\bar{\vartheta}_1)| = \left| \int_{x_i'\bar{\vartheta}_1}^{x_i'\vartheta_1} \phi(t)dt \right| = \phi(c)|x_i'(\vartheta_1 - \bar{\vartheta}_1)| \leq \|x_i\| \|\vartheta_1 - \bar{\vartheta}_1\|,$$

for $c \in (x_i'\vartheta_1, x_i'\bar{\vartheta}_1)$. For P the law of x_i, by (A0),

$$\mathbb{E}_P[\|x_i\|^2] < \infty.$$

Now, consider $r_2(\vartheta_2)$ and note that, for all z,

$$\|r_2(\vartheta_2) - r_2(\vartheta_2')\| \leq \|\dot{r}_2(z)\| \|\vartheta_2 - \vartheta_2'\|, \;\; \dot{r}_2(z) = \begin{bmatrix} 1 & z_i \\ z_i & z_i^2 \end{bmatrix}.$$

It then follows from (A0) that

$$\mathbb{E}_P[\|\dot{r}_2(z)\|^2] < \infty.$$

Defining $L(z) := \max\{\|x_i\|, \dot{r}_2(z)\}$, we have that $\mathbb{E}[L(z)] < \infty$ and

$$\|r(\vartheta) - r(\bar{\vartheta})\| \leq L(z)\|\vartheta - \bar{\vartheta}\|.$$

This Lipschitz property, together with the compactness of V implies that, by Theorem 2.7.11 of van der Vaart and Wellner (1996), $\mathcal{F}$ is P-Donsker. For $g_i(\theta) = a_i \otimes r_i(\theta)$, we then have that

$$\sqrt{n}\,(\bar{g}_n(\theta) - \mathbb{E}[g_i(\theta)]) := \nu_n(\theta) \Rightarrow \nu(\theta),$$

for $\nu(\theta)$ a Gaussian process with zero mean and covariance kernel $S(\theta) := \mathbb{E}[g_i(\theta)g_i(\theta)']$. Moreover, the variance kernel is bounded: by the continuity, in θ, of $S(\theta)$, Assumption (A0), and the compactness of Θ,

$$0 < \sup_{\theta \in \Theta} \|S(\theta)\| < \infty.$$

■

We now demonstrate that under the drifting DGP:

$$\mathbb{E}[g_i(\theta)] = \frac{\Lambda_n}{\sqrt{n}}\gamma(\theta)$$

for Λ_n a deterministic diagonal matrix with minimal and maximal eigenvalues, denoted by $\lambda_{\min}[\Lambda_n]$ and $\lambda_{\max}[\Lambda_n]$, respectively, that satisfy:

$$\lim_{n\to\infty} \lambda_{\min}[\Lambda_n]) = \infty \text{ and } \lim_{n\to\infty} \lambda_{\max}[\Lambda_n]/\sqrt{n} < \infty.$$

Partition θ as $\theta = (\theta_1', \theta_2')'$, with $\theta_1 := (\alpha, \rho)'$ and $\theta_2 = (\mu_y, \xi)'$, and partition the moment function as $g_i(\theta) = (g_{1i}(\theta)', g_{2i}(\theta_2)')'$, where

$$g_{1i}(\theta) = \begin{pmatrix} y_{1,i} - \Phi\left[(\alpha+\rho)y_{2,i} - \rho\xi z_i\right] \\ z_i\left(y_{1,i} - \Phi\left[(\alpha+\rho)y_{2,i} - \rho\xi z_i\right]\right) \end{pmatrix},\ g_{2i}(\theta_2) = \begin{pmatrix} (y_{2,i} - \mu_y - \xi z_i) \\ z_i\left(y_{2,i} - \mu_y - \xi z_i\right) \end{pmatrix}$$

From the identification argument in Lemma 11, $\theta_2^0 = (\mu_y^0, \xi^0)'$ can be directly identified from g_{2i}, which would yield least squares estimators

$$\hat{\theta}_2 := \begin{pmatrix} \hat{\mu_y} \\ \hat{\xi}_n \end{pmatrix} = \begin{pmatrix} \bar{y}_2 - \hat{\xi}_n \bar{z} \\ \sum_{i=1}^n (z_i - \bar{z})(y_{2i} - \bar{y}_2) / \sum_{i=1}^n (z_i - \bar{z})^2 \end{pmatrix},$$

which are clearly consistent under Assumptions (A0)-(A3). Indeed, algebra and an application of (A1) implies

$$\hat{\xi}_n = \xi_n^0 + \frac{1}{n}\sum_{i=1}^n (z_i - \bar{z})v_i/\hat{\sigma}_z^2,\ \ \hat{\sigma}_z^2 := \left[\frac{1}{n}\sum_{i=1}^n (z_i - \bar{z})^2\right],$$

and by (A0) and (A3), an application of the Lindeberg-Levy CLT yields

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n (z_i - \bar{z})v_i/\hat{\sigma}_z^2 \Rightarrow N(0, \sigma_v^2).$$

Now, partition the stochastic process $\nu_n(\theta) = (\nu_{1,n}(\theta_1, \xi)', \nu_{2,n}(\theta_2))'$ to be conformable to $g_i(\theta) = (g_{1i}(\theta)', g_{2i}(\theta)')'$. From the $\sqrt{n}$-consistency of $\hat{\xi}_n$ and stochastic equicontinuity of $\nu_{1,n}(\theta_1, \xi)$, we can restrict our analysis on the uniform behavior of $\nu_{1,n}(\theta_1, \xi)$ to the set $\Theta_1 \times Z_{n^{-1/2}}$, where Θ_1 denotes the parameter subspace for $\theta_1 = (\alpha, \rho)'$ and where, for some finite L,

$$Z_{n^{-1/2}} := \left\{\xi_n : \|\xi_n - \xi^0\| \leq L/\sqrt{n}\right\} \text{ where } 0 < L < \infty.$$

In the remainder, take ξ_n to be any arbitrary sequence in $Z_{n^{-1/2}}$. For ξ_n as above we can rewrite the expectations given in the proof of Lemma 11 as

$$\mathbb{E}[g_{1i}(\theta_1, \xi_n)] = \begin{pmatrix} \gamma_1(\theta_1, \xi_n) \\ \gamma_2(\theta_1, \xi_n) \end{pmatrix},$$

where

$$\begin{aligned}
\gamma_1(\theta_1,\xi_n) =& \Phi\left[\frac{\{\alpha^0+\rho^0\}\,\mu_y^0}{\sqrt{1+\{\alpha^0+\rho^0\}^2(1-(\xi^0)^2)+(\alpha^0\xi^0)^2}}\right] \\
&- \Phi\left[\frac{\{\alpha+\rho\}\,\mu_y^0}{\sqrt{1+\{\alpha+\rho\}^2(1-(\xi^0)^2)+(\alpha\xi^0+O(n^{-1/2}))^2}}\right], \\
\gamma_2(\theta_1,\xi_n) =& Q_1(\theta^0)-Q_2(\theta_1,\xi_n) \\
=& \frac{\alpha^0\xi^0}{\sqrt{1+\{\alpha^0+\rho^0\}^2(1-(\xi^0)^2)+(\alpha^0\xi^0)^2}}\phi\left[\frac{\{\alpha^0+\rho^0\}\,\mu_y^0}{\sqrt{1+\{\alpha^0+\rho^0\}^2(1-(\xi^0)^2)+(\alpha^0\xi^0)^2}}\right] \\
&- \frac{\alpha\xi^0+(\xi^0-\xi_n)\rho}{\sqrt{1+\{\alpha+\rho\}^2(1-(\xi^0)^2)+[\alpha\xi^0+\rho\{\xi^0-\xi_n\}]^2}}\times \\
&\phi\left[\frac{\{\alpha+\rho\}\mu_y^0}{\sqrt{1+\{\alpha+\rho\}^2(1-(\xi^0)^2)+[\alpha\xi^0+\rho\{\xi^0-\xi_n\}]^2}}\right] \\
=& \xi^0\left[f_1(\theta^0)-f_2(\theta)\right]+O(n^{-1/2}),
\end{aligned}$$

for

$$\begin{aligned}
f_1(\theta^0) &= \frac{\alpha^0}{\sqrt{1+\{\alpha^0+\rho^0\}^2(1-(\xi^0)^2)+(\alpha^0\xi^0)^2}}\phi\left[\frac{\{\alpha^0+\rho^0\}\,\mu_y^0}{\sqrt{1+\{\alpha^0+\rho^0\}^2(1-(\xi^0)^2)+(\alpha^0\xi^0)^2}}\right] \\
f_2(\theta_1,\theta_2^0) &= \frac{\alpha}{\sqrt{1+\{\alpha+\rho\}^2(1-(\xi^0)^2)+(\alpha\xi^0)^2}}\phi\left[\frac{\{\alpha+\rho\}\mu_y^0}{\sqrt{1+\{\alpha+\rho\}^2(1-(\xi^0)^2)+(\alpha\xi^0)^2}}\right]
\end{aligned}$$

and where $O(n^{-1/2})$ follows from $\xi_n \in Z_{n^{-1/2}}$ and compactness of Θ_1. Apply $\xi^0 = \delta/n^\lambda$, for some $0<|\delta|<\infty$, to obtain

$$\gamma_2(\theta_1,\hat{\xi}_n) = \frac{\bar{\lambda}_n}{\sqrt{n}}\left[f_1(\theta^0)-f_2(\theta_1,\theta_2^0)+O(n^{-1/2+\lambda})\right], \text{ with } \bar{\lambda}_n := \delta\cdot n^{1/2-\lambda}.$$

Taking

$$\Lambda_n := \begin{bmatrix} n^{1/2} & 0 & 0 & 0 \\ 0 & \bar{\lambda}_n & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix},$$

we arrive at the result.

The remainder of the result follows a similar strategy to Theorem 2.1 in Antoine and Renault (2012). Let W be a positive-definite $H\times H$ matrix, and define $\|x\|_W^2 := x'Wx$. For $\nu_n(\theta) := \sqrt{n}\left(\bar{g}_n(\theta)-\frac{\Lambda_n}{\sqrt{n}}\gamma(\theta)\right)$ the stochastic process analyzed above, write

$$J_n[\theta] := \left\|\frac{\nu_n(\theta)}{\sqrt{n}}+\frac{\Lambda_n}{\sqrt{n}}\gamma(\theta)\right\|_{\Omega_n(\theta)}^2, \text{ for } \Omega_n(\theta) := S_n^{-1}(\theta).$$

By definition of $\hat{\theta}_n$, $J_n[\theta^0] \geq J_n[\hat{\theta}_n]$ which implies

$$\left\|\nu_n(\theta^0)/\sqrt{n}\right\|_{\Omega_n(\theta^0)}^2 \geq \left\|\nu_n(\hat{\theta}_n)/\sqrt{n}+\Lambda_n\gamma(\hat{\theta}_n)/\sqrt{n}\right\|_{\Omega_n(\hat{\theta}_n)}^2.$$

Define $\Omega_n^0 := \Omega_n(\theta^0)$, $\hat{\Omega}_n := \Omega_n(\hat{\theta}_n)$, $x_n := \nu_n(\hat{\theta}_n)$, $y_n := \Lambda_n\gamma(\hat{\theta}_n)$ and $d_n := \nu_n(\hat{\theta}_n)'\hat{\Omega}_n\nu_n(\hat{\theta}_n) - \nu_n(\theta^0)'\Omega_n^0\nu_n(\theta^0)$. Denote $\lambda_{\min}[A]$ and $\lambda_{\max}[A]$ as the smallest and the largest eigenvalue of a matrix A, respectively. Then, we obtain

$$\begin{aligned} 0 \geq J_n[\hat{\theta}_n] - J_n[\theta^0] &= d_n + \|y_n\|^2_{\Omega_n^0} + 2(\hat{\Omega}_n x_n)'y_n \\ &\geq d_n + \|y_n\|^2\lambda_{\min}\left[\hat{\Omega}_n\right] - 2\|y_n\|\|\hat{\Omega}_n x_n\|. \end{aligned} \tag{35}$$

Defining $z_n := \|\Lambda_n\gamma(\hat{\theta}_n)\|$, and for $\lambda_{\min}\left[\hat{\Omega}_n\right] > 0$, we can re-arrange equation (35) as

$$-z_n^2 + z_n\frac{\|\hat{\Omega}_n x_n\|}{\lambda_{\min}\left[\hat{\Omega}_n\right]} + \frac{d_n}{\lambda_{\min}\left[\hat{\Omega}_n\right]} \geq 0$$

Solving the above equation for z_n yields:

$$B_n - \left[B_n^2 - A_n\right]^{1/2} \leq z_n \leq B_n + \left[B_n^2 - A_n\right]^{1/2}, \;\; B_n := \frac{\|\hat{\Omega}_n x_n\|}{\lambda_{\min}\left[\hat{\Omega}_n\right]}, \;\; A_n := \frac{d_n}{\lambda_{\min}\left[\hat{\Omega}_n\right]}. \tag{36}$$

From (36), the result follows if

$$B_n = O_p(1), \text{ and } A_n = O_p(1).$$

Consider first, B_n and note that

$$B_n \leq \|x_n\|\frac{\lambda_{\max}\left[\hat{\Omega}_n\right]}{\lambda_{\min}\left[\hat{\Omega}_n\right]} \leq \sup_{\theta\in\Theta}\|\nu_n(\theta)\|\frac{\sup_{\theta\in\Theta}\lambda_{\max}\left[\Omega_n(\theta)\right]}{\inf_{\theta\in\Theta}\lambda_{\min}\left[\Omega_n(\theta)\right]}.$$

By the first part of the result, $\sup_{\theta\in\Theta}\|\nu_n(\theta)\| = O_p(1)$. It then follows that $B_n = O_p(1)$ so long as, for all n large enough, with probability approaching 1,

$$0 < \inf_{\theta\in\Theta}\lambda_{\min}\left[\Omega_n(\theta)\right] \leq \sup_{\theta\in\Theta}\lambda_{\max}\left[\Omega_n(\theta)\right] < \infty,$$

which is guaranteed under the Assumptions of the Lemma. For A_n, recalling that $d_n = \|\nu_n(\hat{\theta}_n)\|^2_{\hat{\Omega}_n} - \|\nu_n(\theta^0)\|^2_{\Omega_n}$, we obtain

$$|A_n| \leq 2\sup_{\theta\in\Theta}\|\nu_n(\theta)\|\frac{\sup_{\theta\in\Theta}\lambda_{\max}\left[\Omega_n(\theta)\right]}{\inf_{\theta\in\Theta}\lambda_{\min}\left[\Omega_n(\theta)\right]}.$$

Repeating the same argument for A_n as for B_n yields $A_n = O_p(1)$. Applying $B_n = O_p(1), A_n = O_p(1)$ to equation (36), we have $z_n = O_p(1)$. It then follows that

$$\|\gamma(\hat{\theta}_n)\| = O_p\left(1/\bar{\lambda}_n\right).$$

Consistency now follows by modifying the standard argument (see, e.g., Newey and McFadden, 1994, page 2132) as follows. By continuity of $J[\theta] := \underset{n\to\infty}{\text{Plim}} J_n[\theta]$, for any $\epsilon > 0$, there exists some δ_ϵ such that

$$\Pr\left[\|\hat{\theta} - \theta^0\| > \epsilon\right] \leq \Pr\left[J[\hat{\theta}_n] - J[\theta^0] > \delta_\epsilon\right].$$

However, $J[\theta^0] = 0$ by Lemma 11 so that

$$\delta_\epsilon < J[\hat{\theta}_n] - J[\theta^0] = \left\|\gamma(\hat{\theta}_n)\right\|^2_{\Omega(\hat{\theta}_n)} \leq \lambda_{\max}[\Omega(\hat{\theta}_n)] \left\|\gamma(\hat{\theta}_n)\right\|^2 = o_p(1),$$

where the last line follows from the fact that $\|\gamma(\hat{\theta}_n)\| = O_p\left(1/\bar{\lambda}_n\right)$.

Proof of Theorem 3. First, we recall the parameter rotation $\eta := R^{-1}\theta$, so that, for $\hat{\eta}_n = R^{-1}\hat{\theta}_n$, the perturbation defined in (14) reads:

$$\begin{aligned} \hat{\eta}_n^\delta &:= \hat{\eta}_n + \begin{pmatrix} \delta_n \\ \mathbf{O}_{d_\theta - 1} \end{pmatrix} \equiv R^{-1}\hat{\theta}_n + \begin{pmatrix} \delta_n \\ \mathbf{O}_{d_\theta - 1} \end{pmatrix}, \\ \hat{\theta}_n^\delta &:= R\hat{\eta}_n^\delta. \end{aligned} \tag{37}$$

(i) Consider the mean value expansion of the moment conditions

$$\sqrt{n}\bar{g}_n(\hat{\theta}_n^\delta) = \sqrt{n}\bar{g}_n(\hat{\theta}_n) + \sqrt{n}\frac{\partial \bar{g}_n(\bar{\theta}_n)}{\partial \theta'}(\hat{\theta}_n^\delta - \hat{\theta}_n) = \sqrt{n}\bar{g}_n(\hat{\theta}_n) + \sqrt{n}\frac{\partial \bar{g}_n(\bar{\theta}_n)}{\partial \theta'} R \begin{pmatrix} \delta/r_n \\ \mathbf{O}_{d_\theta - 1} \end{pmatrix},$$

where the last line follows from the definition in equation (37) and $\bar{\theta}_n$ is component-by-component between $\hat{\theta}_n^\delta$ and $\hat{\theta}_n$. Focusing on the Jacobian term above, and using the parameterization $A_n = \sqrt{n}R\Lambda_n^{-1}$, this term can rewritten as

$$\begin{aligned} \sqrt{n}\frac{\partial \bar{g}_n(\bar{\theta}_n)}{\partial \theta'} R \begin{pmatrix} \delta/r_n \\ \mathbf{O}_{d_\theta - 1} \end{pmatrix} &= \frac{\partial \bar{g}_n(\bar{\theta}_n)}{\partial \theta'} A_n \sqrt{n} A_n^{-1} R \begin{pmatrix} \delta/r_n \\ \mathbf{O}_{d_\theta - 1} \end{pmatrix}, \\ &= \frac{\partial \bar{g}_n(\bar{\theta}_n)}{\partial \theta'} A_n \sqrt{n} \frac{1}{\sqrt{n}} \Lambda_n R^{-1} R \begin{pmatrix} \delta/r_n \\ \mathbf{O}_{d_\theta - 1} \end{pmatrix}, \\ &= \frac{\partial \bar{g}_n(\bar{\theta}_n)}{\partial \theta'} A_n \Lambda_n \begin{pmatrix} \delta/r_n \\ \mathbf{O}_{d_\theta - 1} \end{pmatrix}. \end{aligned} \tag{38}$$

From the definition of Λ_n, $\Lambda_n(\delta_n, \mathbf{O}'_{d_\theta - 1})' = (\delta/r_n n^{1/2-\lambda}, \mathbf{O}'_{d_\theta - 1})'$. Moreover, under Assumptions (A0)-(A4), we can deduce that

$$\sup_{\theta \in \Theta} \left\| \frac{\partial \bar{g}_n(\theta)}{\partial \theta'} A_n \right\| = O_p(1). \tag{39}$$

Therefore, under $H_0 : \lambda = 1/2$, we note that $\Lambda_n(\delta_n, \mathbf{O}'_{d_\theta - 1})' = o_p(1)$ and applying this and (39) into (38), we arrive at

$$\frac{\partial \bar{g}_n(\bar{\theta}_n)}{\partial \theta'} A_n \Lambda_n \delta_n = O_p(1) o_p(1) = o_p(1). \tag{40}$$

Applying equation (40) into the mean-value expansion for $\sqrt{n}\bar{g}_n(\hat{\theta}_n^\delta)$, we deduce that

$$\sqrt{n}\bar{g}_n(\hat{\theta}_n^\delta) = \sqrt{n}\bar{g}_n(\hat{\theta}_n) + o_p(1). \tag{41}$$

In addition, (41) implies $S_n(\hat{\theta}_n^\delta) = S_n(\hat{\theta}_n) + o_p(1)$. From the definition of $\hat{\theta}_n$, $J_n[\hat{\theta}_n] \leq J_n[\theta^0]$ and we have

$$J_n^\delta = n\bar{g}_n(\hat{\theta}_n^\delta)' S_n^{-1}(\hat{\theta}_n^\delta)\bar{g}_n(\hat{\theta}_n^\delta) = J_n[\hat{\theta}_n] + o_p(1) \leq J_n[\theta^0] + o_p(1) \xrightarrow{d} \chi^2(H), \tag{42}$$

where $J_n[\theta^0] \xrightarrow{d} \chi^2(H)$ follows from the uniform convergence demonstrated in Lemma 12. Therefore, from (42) we can directly deduce that $\underset{n\to\infty}{\text{Plim}} \left| J_n^\delta - J_n[\hat{\theta}_n] \right| = 0$.

(ii) A mean value expansion of $\bar{g}_n(\hat{\theta}_n^\delta)$ yields

$$\begin{aligned}\sqrt{n}\bar{g}_n(\hat{\theta}_n^\delta) &= \sqrt{n}\bar{g}_n(\theta^0) + \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial g_i(\theta_n^*)}{\partial \theta'}(\hat{\theta}_n^\delta - \theta^0) \\ &= \sqrt{n}\bar{g}_n(\theta^0) + \frac{\partial \bar{g}_n(\theta_n^*)}{\partial \theta'}A_n\sqrt{n}A_n^{-1}\left(\hat{\theta}_n - \theta^0\right) + \frac{\partial \bar{g}_n(\theta_n^*)}{\partial \theta'}A_n\sqrt{n}A_n^{-1}R\begin{pmatrix}\delta/r_n \\ \mathbf{O}_{d_\theta-1}\end{pmatrix} \qquad (43)\end{aligned}$$

where θ_n^* is component by component between $\hat{\theta}_n$ and θ^0. We now analyze each of the terms in (43). For the first term in (43), by assumption (A0), the Lindeberg-Levy CLT leads to $\sqrt{n}\bar{g}_n(\theta^0) = O_p(1)$. For the second term, first note that, under the alternative hypothesis, $\sqrt{n}A_n^{-1}(\hat{\theta}_n - \theta^0) = O_p(1)$ implies that $\|\hat{\theta}_n - \theta^0\| = o_p(1)$, which further implies $\|\theta_n^* - \theta^0\| = o_p(1)$. From consistency of θ_n^* and Lemma 1, it follows that

$$\frac{\partial \bar{g}_n(\theta_n^*)}{\partial \theta'}A_n\sqrt{n}A_n^{-1}\left(\hat{\theta}_n - \theta^0\right) = M(\theta^0)O_p(1) = O_p(1),$$

for $M(\theta^0)$ full column rank, and therefore the second term in (43) is $O_p(1)$. For the last term in (43), applying the same type of decomposition as in equation (38), we have

$$\begin{aligned}\frac{\partial \bar{g}_n(\theta_n^*)}{\partial \theta'}A_n\sqrt{n}A_n^{-1}R\begin{pmatrix}\delta/r_n \\ \mathbf{O}_{d_\theta-1}\end{pmatrix} &= \frac{\partial \bar{g}_n\left(\theta_n^*\right)}{\partial \theta'}A_n\Lambda_n\begin{pmatrix}\delta/r_n \\ \mathbf{O}_{d_\theta-1}\end{pmatrix}, \\ &= M(\theta^0)\begin{pmatrix}\delta_n n^{1/2-\lambda} \\ \mathbf{O}_{d_\theta-1}\end{pmatrix} + o_p(1), \\ &= O_p\left(n^{1/2-\lambda}\delta_n\right) + o_p(1),\end{aligned}$$

where the last term follows from noting that $M(\theta^0)$ is full rank. Applying these order results for the three terms in (43), we obtain

$$\sqrt{n}\bar{g}_n(\hat{\theta}_n^\delta) = O_p(1) + O_p\left(n^{1/2-\lambda}\delta_n\right) + o_p(1).$$

Conclude that $\sqrt{n}\bar{g}_n(\hat{\theta}_n^\delta)$ diverges if $n^{1/2-\lambda}\delta_n \to \infty$.

Using the above result, we can now show that J_n^δ diverges under the alternative. From the proof of Lemma 12,

$$n^{1/2}\{\bar{g}_n(\theta) - \mathbb{E}[g_i(\theta)]\} \Rightarrow \nu(\theta), \qquad (44)$$

where $\nu(\theta)$ is a Gaussian stochastic process on Θ with mean zero and bounded covariance kernel $S(\theta)$. Since $\hat{\theta}_n^\delta \xrightarrow{p} \theta^0$, the uniform convergence (44) indicates that the sample covariance matrix satisfies $S_n(\hat{\theta}_n^\delta) \xrightarrow{p} S(\theta^0)$. Thus, for n large enough, $S_n(\hat{\theta}_n^\delta)$ is positive-definite with finite maximal eigenvalue. Therefore,

$$J_n^\delta \geq \lambda_{\min}\left[S_n^{-1}(\hat{\theta}_n^\delta)\right]\left\|\sqrt{n}\bar{g}_n(\hat{\theta}_n^\delta)\right\|^2 = O_p\left(\left[n^{1/2-\lambda}\delta_n\right]^2\right), \qquad (45)$$

where $\lambda_{\min}[S(\theta^0)^{-1}] > 0$ by assumption (A0) and the compactness of Θ. Therefore, if $n^{1/2-\lambda}\delta_n \to \infty$, it follows that $\underset{n\to\infty}{\text{Plim}}\ J_n^\delta \to \infty$, and the test is consistent against any alternative for which $0 < \lambda < 1/2$. ■

Proof of Corollary 4. (i) Partition and rearrange θ so that $\theta = (\varphi', \pi')'$, where $\varphi = (\widetilde{\rho}, \alpha, \beta', \xi)'$. To obtain the sharper bound for the test statistic J_n^δ, following AR Section 3.4, let $\bar{\theta}_n$ denote the infeasible CU-GMM estimator

$$\bar{\theta}_n = (\varphi^{0'}, \bar{\pi}_n')' \ \text{ where } \ \bar{\pi}_n = \arg\min_{\pi\in\Theta_\pi} J_n[\varphi^0, \pi] = \arg\min_{\pi\in\Theta_\pi} n\bar{g}_n'(\varphi^0,\pi)S_n(\varphi^0,\pi)^{-1}\bar{g}_n(\varphi^0,\pi) \tag{46}$$

and where $\varphi^0 = (\widetilde{\rho}^0, \alpha^0, \beta^{0'}, \xi^0)'$ and $\Theta_\pi = \{\pi : (\varphi^{0'}, \pi')' \in \Theta\}$. The above infeasible CU-GMM estimator is only used to obtain a upper bound with a known asymptotic distribution, and is not explicitly used in the testing procedure. Under the null hypothesis H_0, from (42) we know that the distorted J-statistic satisfies

$$J_n^\delta = n\bar{g}_n(\hat{\theta}_n^\delta)'S_n(\hat{\theta}_n^\delta)^{-1}\bar{g}_n(\hat{\theta}_n^\delta) = n\bar{g}_n(\hat{\theta}_n)'S_n(\hat{\theta}_n)^{-1}\bar{g}_n(\hat{\theta}_n) + o_p(1) = J_n[\hat{\theta}_n] + o_p(1). \tag{47}$$

Since the unrestricted CU-GMM estimator $\hat{\theta}_n$ minimizes the objective function $J_n[\theta]$, we have

$$J_n[\hat{\theta}_n] + o_p(1) \le J_n[\bar{\theta}_n] + o_p(1). \tag{48}$$

Next, we prove that $J_n[\bar{\theta}_n] \xrightarrow{d} \chi^2(H-k)$ in two parts. Firstly, we show that $\bar{\pi}_n$ is $\sqrt{n}$-consistent. Then, we establish that the Jacobian of the CU-GMM optimization problem has the desired result.

Part 1: Define function $\gamma(\pi) = \mathbb{E}[g_i(\varphi^0, \pi)]$. We have that $\gamma(\pi)$ is a continuous function from Θ_π into $\mathbb{R}^H$ such that the global identification of π and the functional CLT hold (by arguments similar to those in Lemma 11):

$$\gamma(\pi) = 0 \quad \Leftrightarrow \quad \pi = \pi^0 \ \text{ and } \ \sqrt{n}\left[\bar{g}_n(\varphi^0,\pi) - \gamma(\pi)\right] \Rightarrow \Psi(\varphi^0,\pi), \tag{49}$$

for $\Psi(\varphi^0, \pi)$ a Gaussian process on Θ_π. Then it is trivial to check the assumptions of Lemma 2 are satisfied. Thus, $\bar{\pi}_n$ is consistent.

To show that $\bar{\pi}_n$ is $n^{1/2}$-consistent, first, it is easy to see from (49) that $\bar{\pi}_n$ has the same rate of convergence for all of its k elements: $\Lambda_n = n^{1/2}\mathbf{I}_k$ in this case. Secondly, it is sufficient to verify that Antoine and Renault (2012) Assumption 3 (i), and (iii) at $\theta = \theta^0$, used to show their Theorem 3.1,[26] are satisfied in this case. We restate the two assumptions for completeness.

Assumptions for Theorem 3.1 of Antoine and Renault (2012).

1. $\gamma(\pi)$ is continuously differentiable on Θ_π.
2. The $H \times k$ matrix $\partial\gamma(\pi^0)/\partial\pi'$ has full column rank k.

The continuously differentiable in 1 is satisfied by the function form of $g_i(\varphi^0, \pi)$. Because $|\phi(\cdot)|$ is bounded, Assumption (A0), the compactness of Θ_π and the dominated convergence theorem indicate that

$$\frac{\partial\gamma(\pi^0)}{\partial\pi'} = \mathbb{E}\left[\frac{\partial g_i(\theta^0)}{\partial\pi'}\right] = \mathbb{E}\left[(-a_i\widetilde{\rho}^0\phi_i(\theta^0) + b_i)x_i'\right]. \tag{50}$$

[26] In Antoine and Renault (2012), the proof of Theorem 3.1 only needs Assumption 1, 2 and 3 (i), and (iii) the rank condition at $\theta = \theta^0$.

By Assumption (A6), we know that the Jacobian in (50) is of full column rank. Hence, $n^{1/2}(\bar{\pi}_n - \pi^0) = O_p(1)$.

Part 2: (A0) and the rank condition of (50) implies

$$\underset{n\to\infty}{\text{Plim}}\frac{\partial \bar{g}_n(\varphi^0, \pi^0)}{\partial \pi'} = \mathbb{E}[(-a_i\widetilde{\rho}^0\phi_i(\theta^0) + b_i)x_i'], \tag{51}$$

is full column rank. Moreover, the first order condition of the CU-GMM optimization problem (46) can be written as follows (Antoine and Renault (2009) proof of Proposition 4.1),

$$n\frac{\partial \bar{g}_n(\bar{\theta}_n)'}{\partial \pi}S_n(\bar{\theta}_n)^{-1}\bar{g}_n(\bar{\theta}_n) - P \cdot n\frac{\partial \bar{g}_n(\bar{\theta}_n)'}{\partial \pi}S_n(\bar{\theta}_n)^{-1}\bar{g}_n(\bar{\theta}_n) = 0 \tag{52}$$

for P the projection of the moment conditions, which satisfies

$$P \cdot \sqrt{n}\frac{\partial \bar{g}_n(\bar{\theta}_n)'}{\partial \pi} = \text{Cov}\left(\frac{\partial \bar{g}_n(\bar{\theta}_n)'}{\partial \pi}, \bar{g}_n(\bar{\theta}_n)\right)\left(\mathbf{I}_H \otimes \left[S_n(\bar{\theta}_n)^{-1}\sqrt{n}\bar{g}_n(\bar{\theta}_n)\right]\right), \tag{53}$$

and where $\text{Cov}(\cdot)$ is defined by stacking the following H matrices

$$\text{Cov}\left(\frac{\partial \bar{g}_n(\bar{\theta}_n)'}{\partial \pi}, \bar{g}_n(\bar{\theta}_n)\right) := \left[\text{Cov}\left(\frac{\partial \bar{g}_n^{(1)}(\bar{\theta}_n)}{\partial \pi}, \bar{g}_n(\bar{\theta}_n)\right), \cdots, \text{Cov}\left(\frac{\partial \bar{g}_n^{(H)}(\bar{\theta}_n)}{\partial \pi}, \bar{g}_n(\bar{\theta}_n)\right)\right]. \tag{54}$$

Plug in (53) to (52), and multiply $n^{-1/2}$ on both sides of the equation (52), we get that

$$\begin{aligned}\sqrt{n}\frac{\partial \bar{g}_n(\bar{\theta}_n)'}{\partial \pi}S_n(\bar{\theta}_n)^{-1}\bar{g}_n(\bar{\theta}_n) - \text{Cov}\left(\frac{\partial \bar{g}_n(\bar{\theta}_n)'}{\partial \pi}, \bar{g}_n(\bar{\theta}_n)\right)\\ \times \left(\mathbf{I}_H \otimes \left[S_n(\bar{\theta}_n)^{-1}\sqrt{n}\bar{g}_n(\bar{\theta}_n)\right]\right)S_n(\bar{\theta}_n)^{-1}\bar{g}_n(\bar{\theta}_n) = 0.\end{aligned} \tag{55}$$

Since $\bar{\pi}_n$ is $\sqrt{n}$-consistent, the mean value theorem, (A0) and the FCLT in (49) yield

$$\bar{g}_n(\bar{\theta}_n) = \bar{g}_n(\theta^0) + \frac{\partial \bar{g}_n(\bar{\theta}_n)}{\partial \pi'}(\bar{\pi}_n - \pi^0) = O_p(n^{-1/2}).$$

Therefore, we have $\text{Cov}\left(\frac{\partial \bar{g}_n(\bar{\theta}_n)'}{\partial \pi}, \bar{g}_n(\bar{\theta}_n)\right) = O_p(1)$ and $\mathbf{I}_H \otimes \left[S_n(\bar{\theta}_n)^{-1}\sqrt{n}\bar{g}_n(\bar{\theta}_n)\right] = O_p(1)$, which leading the second term of the left hand side of (55) to be a $o_p(1)$. Thus,

$$\sqrt{n}\frac{\partial \bar{g}_n(\bar{\theta}_n)'}{\partial \pi}S_n(\bar{\theta}_n)^{-1}\bar{g}_n(\bar{\theta}_n) = o_p(1). \tag{56}$$

Given the full column rank of (51) and (56), standard arguments can be used to derive that $J_n[\bar{\theta}_n] \xrightarrow{d} \chi^2(H-k)$, which fulfills this proof.

(ii) Under alternative hypothesis, the proof follows the proof of Theorem 3 (ii). ■

Proof of Proposition 5. F_n is calculated using the lease square estimators of the parameters in the first-stage equation, which are $\sqrt{n}$-consistent. The rest of this proof follows the proof of AR Proposition 3.2. ■

Proof of Proposition 6. In the simplified case, $\theta = (\theta_1', \theta_2)'$, $\theta_1 = (\widetilde{\rho}, \alpha)'$ and $\theta_2 = \xi$. Moment restrictions become to

$$0 = \mathbb{E}\left\{a_i \left[y_{i,1} - \Phi\left([\alpha^0 + \widetilde{\rho}^0] y_{2,i} - \widetilde{\rho}^0 \xi^0 z_i\right)\right] + b_i \left(y_{2,i} - \xi^0 z_i\right)\right\}.$$

Assumption (A5) implies θ_n^* in (43) is such that $\|\theta_n^* - \theta^0\| = o_p(1)$ under the alternative hypothesis. Thus, the first two terms in (43) are $O_p(1)$. Then, rewrite the expansion as

$$\sqrt{n}\bar{g}_n(\hat{\theta}_n^\delta) = O_p(1) - \left[\frac{1}{n}\sum_{i=1}^n a_i z_i \phi_i(\theta^0)\gamma^0\right] n^{1/2-\lambda}\delta_n.$$

Denote $Q_n(\theta) = \frac{1}{n}\sum_{i=1}^n a_i z_i \phi_i(\theta)\gamma$, then we can decompose J_n^δ as

$$\begin{aligned} J_n^\delta &= n \cdot \bar{g}_n(\hat{\theta}_n^\delta)' S_n(\hat{\theta}_n^\delta)^{-1} \bar{g}_n(\hat{\theta}_n^\delta) \\ &= O_p\left(n^{1/2-\lambda}\delta_n\right) + n^{1-2\lambda}\delta_n^2 Q_n(\theta^0)' S_n(\theta^0)^{-1} Q_n(\theta^0) \\ &= O_p\left(n^{1/2-\lambda}\delta_n\right) + \mathcal{U}^2, \end{aligned} \tag{57}$$

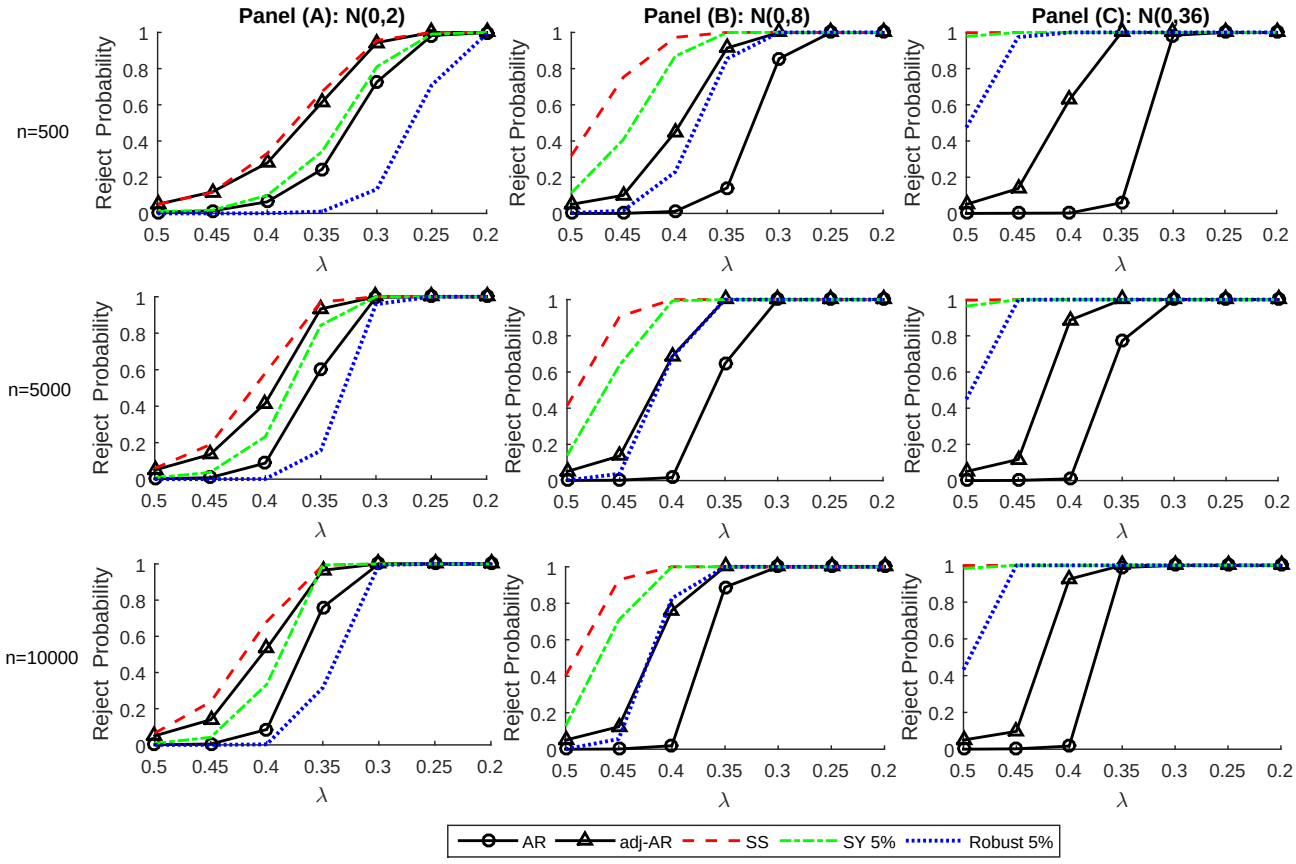
where $\mathcal{U}^2 := n^{1-2\lambda}\delta_n^2 Q_n(\theta^0)' S_n(\theta^0)^{-1} Q_n(\theta^0)$. The test will then be able to detect weak instrument so long as we choose a sequence $\delta_n = o_p(1)$ and $n^{1/2-\lambda}\delta_n \to \infty$. Rewrite $S_n(\theta^0)$ and $Q_n(\theta^0)$ as

$$\begin{aligned} S_n(\theta^0) &= \frac{1}{n}\sum_{i=1}^n [a_i, b_i] \begin{bmatrix} y_{1,i} - \Phi_i(\theta^0) \\ y_{2,i} - \xi^0 z_i \end{bmatrix} \left([a_i, b_i] \begin{bmatrix} y_{1,i} - \Phi_i(\theta^0) \\ y_{2,i} - \xi^0 z_i \end{bmatrix}\right)' + o_p(1) = \frac{1}{n}\sum_{i=1}^n w_i V_i (w_i V_i)' + o_p(1) \\ &= \frac{1}{n}\sum_{i=1}^n V_i^w (V_i^w)' + o_p(1), \\ Q_n(\theta^0) &= \frac{1}{n}\sum_{i=1}^n [a_i, b_i] z_i \begin{bmatrix} \gamma^0 \phi_i(\theta^0) \\ 0 \end{bmatrix} = \frac{1}{n}\sum_{i=1}^n (w_i z_i)\Pi_i = \frac{1}{n}\sum_{i=1}^n z_i^w \Pi_i, \end{aligned}$$

where $w_i = [a_i, b_i]$, $V_i = (y_{1,i} - \Phi_i(\theta^0),\ v_i)'$, $V_i^w = w_i V_i$, $z_i^w = w_i z_i$ and $\Pi_i = [\gamma^0 \phi_i(\theta^0), 0]'$. ■

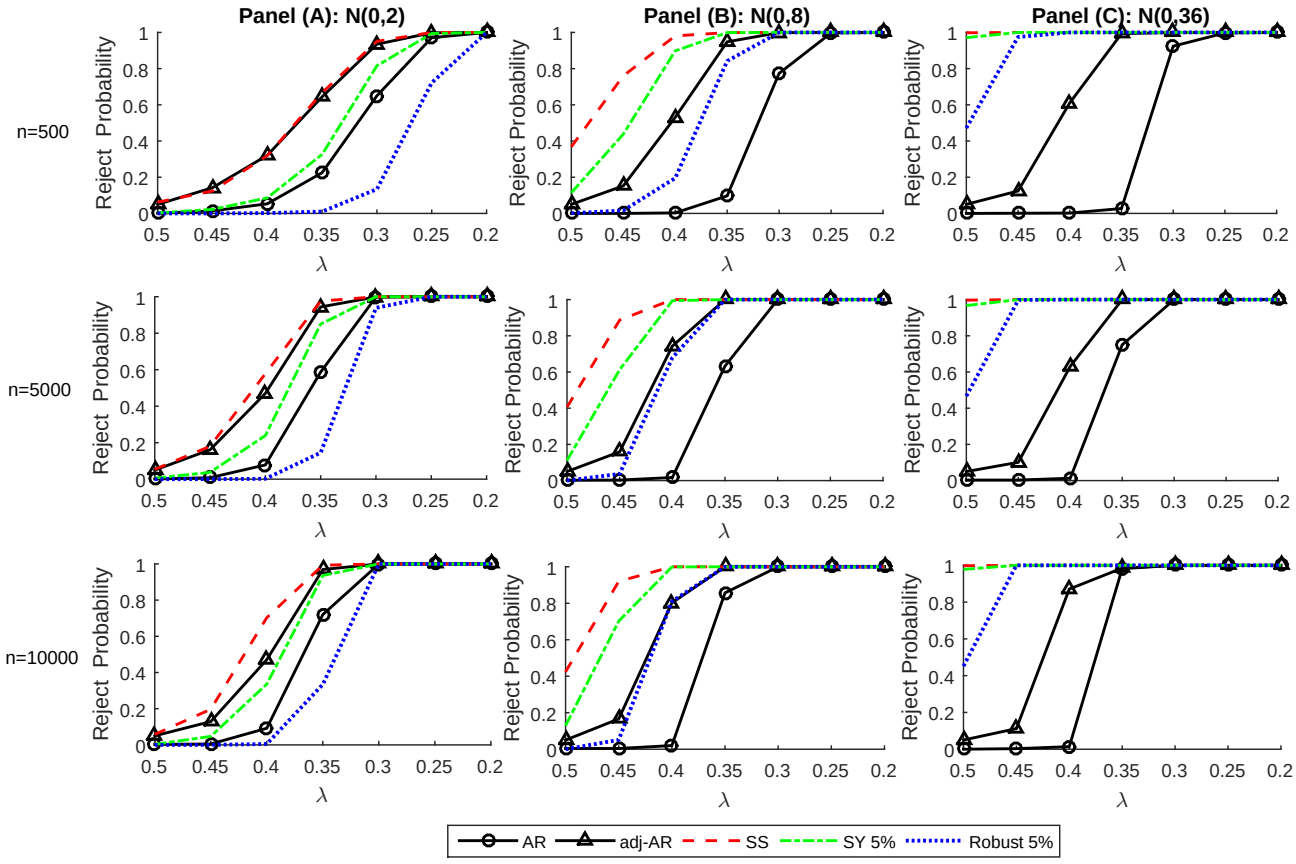
A.3 Table and Figures

Figure 1: Reject Probabilities (Power) $\lambda < 0.5$, $\rho = 0.5$



[a] x-axis is IV strength λ.
[b] First column Panel (A) $z_i \sim \mathcal{N}(0, 2)$. Second column Panel (B) $z_i \sim \mathcal{N}(0, 8)$. Third column Panel (C) $z_i \sim \mathcal{N}(0, 36)$.
[c] First row $n = 500$, second row $n = 5000$, third row $n = 10000$.
[d] The reject probabilities are computed using the 95% quantile of $\chi^2(H - 1)$.

Figure 2: Reject Probabilities (Power) $\lambda < 0.5$, $\rho = 0.8$



[a] x-axis is IV strength λ.
[b] First column Panel (A) $z_i \sim \mathcal{N}(0, 2)$. Second column Panel (B) $z_i \sim \mathcal{N}(0, 8)$. Third column Panel (C) $z_i \sim \mathcal{N}(0, 36)$.
[c] First row $n = 500$, second row $n = 5000$, third row $n = 10000$.
[d] The reject probabilities are computed using the 95% quantile of $\chi^2(H - 1)$.

Table 1: Reject Probabilities under Null Hypothesis. Significant Level 5% (Size)

		$\sigma_z^2 = 2$		$\sigma_z^2 = 8$		$\sigma_z^2 = 36$	
		$\rho = 0.5$	$\rho = 0.8$	$\rho = 0.5$	$\rho = 0.8$	$\rho = 0.5$	$\rho = 0.8$
$n = 500$	DJ	0.007	0.001	0	0	0	0
	SS	0.050	0.062	0.319	0.368	0.998	0.998
	SY	0.006	0.003	0.113	0.114	0.978	0.971
	Robust $\tau = 5\%$	0	0	0.002	0.001	0.475	0.472
	Robust $\tau = 10\%$	0.001	0.002	0.034	0.030	0.882	0.867
	Robust $\tau = 20\%$	0.008	0.008	0.151	0.137	0.980	0.977
	Robust $\tau = 30\%$	0.024	0.023	0.269	0.237	0.992	0.992
$n = 5000$	DJ	0.002	0.001	0.001	0.002	0	0.003
	SS	0.054	0.054	0.413	0.406	0.997	0.997
	SY	0.001	0.006	0.138	0.111	0.964	0.967
	Robust $\tau = 5\%$	0	0	0.001	0	0.450	0.467
	Robust $\tau = 10\%$	0	0	0.035	0.030	0.910	0.881
	Robust $\tau = 20\%$	0.004	0.006	0.149	0.147	0.986	0.982
	Robust $\tau = 30\%$	0.028	0.016	0.263	0.260	0.996	0.992
$n = 10000$	DJ	0.001	0.003	0	0.004	0	0
	SS	0.053	0.058	0.403	0.424	0.999	0.998
	SY	0.008	0.003	0.127	0.129	0.983	0.978
	Robust $\tau = 5\%$	0	0	0	0	0.433	0.452
	Robust $\tau = 10\%$	0.001	0	0.017	0.024	0.887	0.885
	Robust $\tau = 20\%$	0.010	0.005	0.143	0.130	0.986	0.984
	Robust $\tau = 30\%$	0.024	0.020	0.263	0.253	0.992	0.993

[(a)] SS rejects the null if $F_n > 10$. SY rejects the null if the Cragg-Donald statistic is larger than 16.38, where the critical value is chosen based on 5% Wald test size distortion.
[(b)] For the Robust tests, reject probabilities are computed based on critical values in Table 1 of Olea and Pflueger (2013). Critical values are $[37.42, 23.11, 15.06, 12.05]$, corresponding to the effective degree of freedom one and tolerance threshold $\tau = [5\%, 10\%, 20\%, 30\%]$, where τ is the fraction that the Nagar bias relative to the benchmark.
[(c)] The reject probabilities of DJ are computed using $\chi^2_{0.95}(5) = 11.07$ as critical value.

Table 2: GMM bias, standard deviation and rrmse of α ($z_i \sim \mathcal{N}(0, 8)$)

(a) $\rho = 0.5$

	λ	0.5	0.45	0.4	0.35	0.3	0.25	0.2
	bias	-0.023	-0.023	0.058	0.018	0.041	0.038	0.046
$n = 500$	s.d.	0.876	0.649	0.412	0.317	0.245	0.201	0.178
	rrmse	0.876	0.650	0.412	0.317	0.249	0.204	0.184
	bias	-0.046	0.005	-0.017	-0.004	-0.001	0.004	0.003
$n = 5000$	s.d.	0.755	0.458	0.302	0.188	0.127	0.087	0.059
	rrmse	0.756	0.458	0.302	0.188	0.127	0.087	0.059
	bias	-0.054	-0.022	-0.006	-0.003	0.002	0.001	0.003
$n = 10000$	s.d.	0.776	0.440	0.280	0.164	0.107	0.072	0.051
	rrmse	0.777	0.440	0.280	0.164	0.107	0.072	0.051

(b) $\rho = 0.8$

	λ	0.5	0.45	0.4	0.35	0.3	0.25	0.2
	bias	-0.064	0.073	0.134	0.120	0.179	0.114	0.218
$n = 500$	s.d.	1.448	1.722	1.377	0.689	1.095	0.601	1.006
	rrmse	1.448	1.723	1.383	0.699	1.109	0.611	1.029
	bias	-0.116	-0.069	-0.020	-0.002	-0.001	-0.001	0.007
$n = 5000$	s.d.	1.032	0.645	0.403	0.258	0.174	0.124	0.090
	rrmse	1.038	0.648	0.403	0.258	0.174	0.124	0.090
	bias	-0.083	-0.025	-0.007	-0.004	-0.010	-0.001	0.005
$n = 10000$	s.d.	0.976	0.582	0.348	0.217	0.148	0.095	0.071
	rrmse	0.979	0.582	0.348	0.217	0.149	0.095	0.071

Table 3: Relative bias of 2SLS to OLS in linear probability model ($z_i \sim \mathcal{N}(0, 8)$)

(a) $\rho = 0.5$

	λ	0.5	0.45	0.4	0.35	0.3	0.25	0.2
$n = 500$	bias 2SLS	-0.022	-0.008	-0.003	-0.002	-0.001	-0.000	-0.000
	bias OLS	0.105	0.103	0.100	0.096	0.088	0.076	0.060
	re. bias	0.205	0.076	0.027	0.016	0.007	0.006	0.005
	abs. bias 2SLS	0.134	0.083	0.058	0.041	0.030	0.021	0.015
	abs. bias OLS	0.105	0.103	0.100	0.096	0.088	0.076	0.060
	re. abs. bias	1.282	0.804	0.581	0.433	0.336	0.278	0.246
$n = 5000$	bias 2SLS	-0.021	-0.010	-0.004	-0.002	-0.001	-0.001	-0.001
	bias OLS	0.107	0.106	0.106	0.104	0.101	0.095	0.082
	re. bias	0.201	0.091	0.034	0.016	0.013	0.009	0.008
	abs. bias 2SLS	0.123	0.070	0.044	0.029	0.019	0.012	0.008
	abs. bias OLS	0.107	0.106	0.106	0.104	0.101	0.095	0.082
	re. abs. bias	1.157	0.655	0.416	0.274	0.187	0.129	0.096
$n = 10000$	bias 2SLS	-0.018	-0.003	-0.001	0.000	0.000	0.000	0.000
	bias OLS	0.107	0.107	0.106	0.106	0.103	0.098	0.087
	re. bias	0.167	0.030	0.007	0.003	0.004	0.003	0.004
	abs. bias 2SLS	0.124	0.067	0.041	0.025	0.016	0.010	0.006
	abs. bias OLS	0.107	0.107	0.106	0.106	0.103	0.098	0.087
	re. abs. bias	1.162	0.629	0.386	0.239	0.156	0.103	0.069

(b) $\rho = 0.8$

	λ	0.5	0.45	0.4	0.35	0.3	0.25	0.2
$n = 500$	bias 2SLS	-0.060	-0.010	-0.003	0.000	0.002	0.001	0.001
	bias OLS	0.191	0.188	0.183	0.175	0.161	0.140	0.112
	re. bias	0.316	0.053	0.014	0.001	0.010	0.007	0.006
	abs. bias 2SLS	0.176	0.088	0.061	0.043	0.030	0.022	0.015
	abs. bias OLS	0.191	0.188	0.183	0.175	0.161	0.140	0.112
	re. abs. bias	0.923	0.469	0.334	0.247	0.188	0.156	0.137
$n = 5000$	bias 2SLS	-0.036	-0.010	-0.004	-0.002	-0.000	-0.000	0.000
	bias OLS	0.194	0.193	0.192	0.190	0.185	0.173	0.151
	re. bias	0.185	0.052	0.020	0.008	0.002	0.002	0.000
	abs. bias 2SLS	0.137	0.078	0.050	0.032	0.020	0.013	0.008
	abs. bias OLS	0.194	0.193	0.192	0.190	0.185	0.173	0.151
	re. abs. bias	0.707	0.405	0.258	0.169	0.109	0.073	0.055
$n = 10000$	bias 2SLS	-0.039	-0.006	-0.001	-0.000	0.000	0.000	0.000
	bias OLS	0.194	0.194	0.193	0.191	0.188	0.179	0.159
	re. bias	0.202	0.031	0.006	0.002	0.000	0.001	0.001
	abs. bias 2SLS	0.137	0.070	0.043	0.027	0.017	0.011	0.007
	abs. bias OLS	0.194	0.194	0.193	0.191	0.188	0.179	0.159
	re. abs. bias	0.706	0.361	0.223	0.141	0.090	0.060	0.043

Note: (a) The value of $\widetilde{\alpha}^*$ depends on n, λ, σ_z^2 and θ^0. Here we compute $\widetilde{\alpha}^*$ by (27) in Lemma 10 for each case, e.g. for $n = 500$ and $\rho = 0.5$, $\widetilde{\alpha}^* = [0.185, 0.185, 0.184, 0.183, 0.182, 0.179, 0.174]$ for λ from 0.5 to 0.2, respectively.

(b) Bias 2SLS and bias OLS are sample counterparts of $\mathbb{E}[\widehat{\widetilde{\alpha}}^{2SLS} - \widetilde{\alpha}^*]$ and $\mathbb{E}[\widehat{\widetilde{\alpha}}^{OLS} - \widetilde{\alpha}^*]$. The re. bias is the sample counterparts of $|\mathbb{E}\widehat{\widetilde{\alpha}}^{2SLS} - \widetilde{\alpha}^*|/|\mathbb{E}\widehat{\widetilde{\alpha}}^{OLS} - \widetilde{\alpha}^*|$.

(c) Abs. bias 2SLS and abs. bias OLS are sample counterparts of $\mathbb{E}|\widehat{\widetilde{\alpha}}^{2SLS} - \widetilde{\alpha}^*|$ and $\mathbb{E}|\widehat{\widetilde{\alpha}}^{OLS} - \widetilde{\alpha}^*|$. The re. abs. bias is the ratio of abs. bias 2SLS to abs. bias OLS.

Table 4: F-statistic and Nagar relative bias of linear probability model ($z_i \sim \mathcal{N}(0, 8)$)

(a) $\rho = 0.5$

n	λ	0.5	0.45	0.4	0.35	0.3	0.25	0.2
	$\mathbb{E}[F_n]$	8.2	14.7	27.4	50.8	95.3	179.6	334.4
500	F 95% CI	(2, 20)	(5, 30)	(12, 50)	(30, 79)	(65, 135)	(135, 243)	(257, 419)
	Nagar/BM	0.116	0.063	0.033	0.018	0.010	0.005	0.003
	$\mathbb{E}[F_n]$	8.5	19.8	43.4	103.1	241.0	565.1	1327
5000	F 95% CI	(2, 20)	(8, 37)	(24, 68)	(73, 142)	(193, 294)	(483, 648)	(1196, 1462)
	Nagar/BM	0.117	0.050	0.021	0.009	0.004	0.002	0.001
	$\mathbb{E}[F_n]$	8.7	20.8	51.4	128.5	319.4	800.9	2009
10000	F 95% CI	(2, 21)	(9, 36)	(30, 78)	(94, 167)	(266, 379)	(708, 900)	(1850, 2173)
	Nagar/BM	0.117	0.047	0.019	0.007	0.003	0.001	0.000

(b) $\rho = 0.8$

n	λ	0.5	0.45	0.4	0.35	0.3	0.25	0.2
	$\mathbb{E}[F_n]$	7.8	14.9	27.9	51.7	96.4	178.3	333.3
500	F 95% CI	(2, 20)	(5, 30)	(13, 49)	(31, 79)	(62, 135)	(133, 236)	(260, 420)
	Nagar/BM	0.119	0.064	0.034	0.018	0.010	0.005	0.003
	$\mathbb{E}[F_n]$	8.7	18.7	43.5	103.5	239.3	564.8	1334
5000	F 95% CI	(2, 21)	(8, 36)	(24, 69)	(72, 140)	(190, 295)	(481, 652)	(1183, 1470)
	Nagar/BM	0.119	0.051	0.022	0.009	0.004	0.002	0.001
	$\mathbb{E}[F_n]$	8.8	21.1	51.2	127.8	317.0	798.5	2010
10000	F 95% CI	(2, 20)	(9, 39)	(31, 77)	(93, 168)	(262, 384)	(707, 896)	(1854, 2177)
	Nagar/BM	0.119	0.048	0.019	0.008	0.003	0.001	0.001

Note: (a) $\mathbb{E}[F_n]$ is the sample mean of M replications of the first stage F-statistic, F_n. F 95% CI is the 95% confidence interval of F_n.
(b) Nagar/BM is the Nagar bias relative to the benchmark defined in Olea and Pflueger (2013). Both Nagar bias and the benchmark is calculated taking the value of θ^0 and λ as known.

Table 5: Data Summary of Married Women LFP (Obs. 753)

	Mean	Std. Dev.	Min	Max
LFP	0.57	0.50	0	1
Education	12.29	2.28	5	17
Father educ.	8.81	3.57	0	17
Mother educ.	9.25	3.37	0	17
Husband educ.	12.49	3.02	3	17
Experience	10.63	8.07	0	45
Exper. square	178.04	249.63	0	2025
Nonwife income ($1000)	20.13	11.64	-0.029	96
Age	42.54	8.07	30	60
# Kids < 6 years old	0.24	0.52	0	3
# Kids > 6 years old	1.35	1.32	0	8

Note: Education, father/mother/husband education, experience and its square are measured in years.

Table 6: Tests of Weak IV. Significance level 5% (H_0: weak IV)

	SS	SY	Robust ($\tau = 5\%$)	DJ
Statistic	140.23	140.23	143.17	142.98
Critical value	10	13.91	17.84	22.36
Reject H_0	Reject	Reject	Reject	Reject

Note: (a) SS and SY test statistics 140.23 are Kleibergen-Paap F-statistic, which is heteroscedastic-robust. When assuming homoscedastic standard error, the first-stage F-statistic and the Cragg-Donald F-statistic is 155.31. SS critical value 10 is the rule-of-thumb. SY 5% significant critical value is for i,i.d. errors, based on 2SLS/OLS relative bias for the case of one endogenous regressor and three IVs, desired maximal relative bias 5%.
(b) Robust test statistics are computed using STATA command "weakivtest", see Pflueger and Wang (2014). The estimated effective degrees of freedom with $\tau \in \{5\%, 10\%, 20\%, 30\%\}$ of the Robust test are all about 2.4.
(c) DJ test critical values are $\chi^2_{0.95}(H) = 31.41$ or $\chi^2_{0.95}(H-k) = 22.36$.

Table 7: Regression Results of Labor Force Participation (LFP)

	2SCML Probit			CU-GMM		
	1st step	2nd step	margin	reduced form	structural eq.	margin
	(1)	(2)	(3)	(4)	(5)	(6)
Dependent Var.	Education	LFP		Education	LFP	
Education		0.1036**	0.0404**		0.0990**	0.0392**
		(0.0419)	(0.0164)		(0.0434)	(0.0155)
Experience	0.0578***	0.1262***	0.0493***	0.0592***	0.1253***	0.0497***
	(0.0219)	(0.0191)	(0.0075)	(0.0226)	(0.0219)	(0.0106)
Exper. squared	-0.0008	-0.0019***	-0.0008***	-0.0008	-0.0019***	-0.0008**
	(0.0007)	(0.0006)	(0.0002)	(0.0008)	(0.0007)	(0.0004)
Nonwife income	0.0157**	-0.0103*	-0.0040*	0.0152**	-0.0111*	-0.0044**
	(0.0065)	(0.0058)	(0.0023)	(0.0065)	(0.0064)	(0.0022)
Age	-0.0059	-0.0544***	-0.0212***	-0.0057	-0.0545***	-0.0216***
	(0.0094)	(0.0086)	(0.0033)	(0.0113)	(0.0090)	(0.0030)
Kids<6 years old	0.1196	-0.8631***	-0.3370***	0.1209	-0.8619***	-0.3416***
	(0.1365)	(0.1159)	(0.0455)	(0.1404)	(0.1207)	(0.0427)
Kids>6 years old	-0.0731	0.0314	0.0123	-0.0797	0.0302	0.0120
	(0.0493)	(0.0454)	(0.0177)	(0.0533)	(0.0567)	(0.0144)
Father educ.	0.0951***			0.0938***		
	(0.0207)			(0.0196)		
Mother educ.	0.1300***			0.1295***		
	(0.0217)			(0.0205)		
Husband educ.	0.3475***			0.3491***		
	(0.0273)			(0.0282)		
Correlation ρ		0.0719			0.0742	
		(0.0847)			(0.0840)	
J-statistic	–	–	–	–	1.60	
Obs.	753	753	753	753	753	753

Note: [a] Standard errors (s.e.) in parentheses. The s.e. in columns (1)-(3) are heteroscedastic-robust. The s.e. in columns (4)-(6) are computed by bootstrap. Significance *** p<0.01, ** p<0.05, * p<0.1.
[b] For CU-GMM estimation, overidentification degree is one. Hansen's J-statistic 1.60 is less than $\chi^2_{0.95}(1) = 3.84$. Overidentification test fails to reject the null hypothesis that moments are all valid.
[c] Correlation ρ is the correlation of errors (u_i, v_i) in structural equation and reduced form.
[d] Margins in columns (3) and (6) are computed based on equation (25) in Lemma 10, using 2SCML and CU-GMM estimates and plugging in sample average of explanatory variables and IVs.

Table 8: Data Summary of US Food Aid and Civil Conflict

(a) Civil Conflict Onset (obs. 1454)

	Mean	Std. Dev.	Min	Max
Onset of intra-state conflict	0.063	0.244	0	1
US wheat aid (1000 metric tons)	21.08	59.42	0	791.60
Lagged US wheat production (1000 metric tons)	59187	8754	36787	75813
Average US food aid probability 1971-2006	0.387	0.328	0	1
Peace duration (years)	11.59	9.48	1	46
Instrument	22936	19924	0	75813

(b) Civil Conflict Offset (obs. 709)

	Mean	Std. Dev.	Min	Max
Offset of intra-state conflict	0.185	0.388	0	1
US wheat aid (1000 metric tons)	56.07	123.58	0	854.7
Lagged US wheat production (1000 metric tons)	60374	8626	36787	75813
Average US food aid probability 1971-2006	0.503	0.313	0	1
Conflict duration (years)	8.70	8.45	1	42
Instrument	30413	19676	0	75813

Note: An observation is a country and year. Instrument is lag of US wheat production times average probability of receiving any US food aid during 1971 to 2006.

Table 9: Tests of Weak IV. Significance level 5% (H_0: weak IV)

(a) Civil Conflict Onset

	SS	SY	Robust ($\tau = 5\%$)	Robust ($\tau = 10\%$)	DJ
Statistic	26.07	26.07	26.39	26.39	0.301
Critical value	10	16.38	37.42	23.11	15.51
Reject H_0	Reject	Reject	Not Reject	Reject	Not Reject

(b) Civil Conflict Offset

	SS	SY	Robust ($\tau = 5\%$)	Robust ($\tau = 10\%$)	DJ
Statistic	17.29	17.29	17.49	17.49	0.369
Critical value	10	16.38	37.42	23.11	15.51
Reject H_0	Reject	Reject	Not Reject	Not Reject	Not Reject

Note: (a) For both onset and offset data, SS and SY test statistics are Kleibergen-Paap F-statistic (Kleibergen and Paap (2006)) based on clustered standard errors by countries, to be consistent with Nunn and Qian (2014). SS critical value 10 is the rule-of-thumb. SY critical value is for i.i.d. errors, based on Wald test size for the case of one endogenous regressor and one IV, desired maximal size 10% of a 5% Wald test.
(b) Robust test statistics are computed using STATA command "weakivtest", see Pflueger and Wang (2014). For both onset and offset data, the Robust test statistics are computed using clustered standard errors by countries. The estimated effective degrees of freedom with the tolerance $\tau \in \{5\%, 10\%, 20\%, 30\%\}$ of the robust test are all 1.
(c) For the offset data, the Robust test rejects H_0 when $\tau \geq 20\%$.
(d) DJ test critical values are $\chi^2_{0.95}(H) = 21.03$ or $\chi^2_{0.95}(H - k) = 15.51$.

Table 10: Regression Results of US Food Aid and Civil Conflict

(a) Civil Conflict Onset

	Nunn & Qian (2014)	2SCML Probit			CU-GMM		
	margin	1st step	2nd step	margin	reduced form	structural eq.	margin
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Dependent Var.	Onset	Wheat aid	Onset	Onset	Wheat aid	Onset	Onset
Wheat aid	0.000064		0.0011	0.000114		-0.0015	-0.000135
	(0.00026)		(0.0025)	(0.00027)		(0.0025)	(0.00031)
Peace dur.	-0.018***	-1.66	-0.18***	-0.020***	-1.66	-0.20***	-0.018***
	(0.0043)	(1.18)	(0.041)	(0.0046)	(1.14)	(0.046)	(0.0050)
Peace dur.^2	0.00087***	0.053	0.0087***	0.00093***	0.054	0.0094**	0.00084***
	(0.00028)	(0.066)	(0.0026)	(0.00029)	(0.062)	(0.0040)	(0.00029)
Peace dur.^3	-0.00001**	-0.00042	-0.00012***	-0.00001**	-0.00046	-0.00013**	-0.00001**
	(0.00000)	(0.0011)	(0.00005)	(0.00001)	(0.0010)	(0.00006)	(0.00001)
Instrument		0.0012***			0.0012***		
		(0.0002)			(0.0001)		
Correlation ρ			-0.0837			0.3692***	
			(0.1318)			(0.0893)	
J-statistic	–	–	–	–	–	0.299	–
Obs.	1454	1454	1454	1454	1454	1454	1454

(b) Civil Conflict Offset

	Nunn & Qian (2014)	2SCML Probit			CU-GMM		
	margin	1st step	2nd step	margin	reduced form	structural eq.	margin
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Dependent Var.	Offset	Wheat aid	Offset	Offset	Wheat aid	Offset	Offset
Wheat aid	-0.000428*		-0.0019*	-0.000446*		-0.0026*	-0.000516*
	(0.00025)		(0.0011)	(0.00026)		(0.0015)	(0.00027)
Conflict dur.	-0.0619***	4.97	-0.2794***	-0.0653***	4.97	-0.3318***	-0.0660***
	(0.0117)	(4.65)	(0.0525)	(0.0125)	(3.28)	(0.0733)	(0.0164)
Conflict dur.^2	0.0037***	-0.406	0.0164***	0.0038***	-0.401	0.0199***	0.0040***
	(0.0010)	(0.288)	(0.0046)	(0.0011)	(0.217)	(0.0076)	(0.0012)
Conflict dur.^3	-0.0001***	0.007	-0.0003***	-0.0001***	0.007	-0.0004**	-0.0001***
	(0.0000)	(0.005)	(0.0001)	(0.0000)	(0.004)	(0.0002)	(0.0000)
Instrument		0.003***			0.003***		
		(0.0007)			(0.0002)		
Correlation ρ			0.1277			0.5880***	
			(0.1238)			(0.1136)	
J-statistic	–	–	–	–	–	0.349	–
Obs.	709	709	709	709	709	709	709

Note: (a) Standard errors in parentheses. For both panels, the standard errors (s.e.) in columns (1)-(4) are clustered s.e. by countries. The s.e. in (1) are calculated by 2SCML using logit. The s.e. in (2)-(4) are based on the 2SCML probit estimation. The s.e. in column (5)-(7) is calculated by bootstrap. Significance *** p<0.01, ** p<0.05, * p<0.1.
(b) For CU-GMM estimation, overidentification degree is one. Hansen's J-statistics are less than $\chi^2_{0.95}(1) = 3.84$. Overidentification test fails to reject the null hypothesis that moments are all valid in both onset and offset cases.
(c) Correlation ρ is the correlation of errors (u_i, v_i) in structural equation and reduced form.
(d) Margins in columns (4) and (7) are computed based on equation (25) in Lemma 10, using 2SCML and CU-GMM estimates and plugging in sample average of explanatory variables and IVs.