The Strong, the Weak, and the Robust

Inference and Estimation in Weakly Identified Panel Binary Dependent Variable Models

Jack Mulqueeney

Supervised by Dr. Leandro M. Magnusson

Department of Economics

University of Western Australia

Submitted on: July 25, 2025

This dissertation is submitted in partial fulfilment of the requirements for the Degree of Bachelor of Philosophy (Honours)

Abstract

I propose a hypothesis test that is robust to the presence of weak instruments in binary outcome panel data. The test relies on a conditional maximum likelihood procedure, the conditional logit, to consistently estimate reduced-form parameters. Based on a distance function that relates reduced- and structural-form parameters, the test has the correct size regardless of instrument strength while standard Wald tests over-reject by up to 100% when instruments are weak. I investigate the findings of Nunn and Qian (2014) with the proposed test and find that claimed statistical significance vanishes at the 1% and 5% significance levels. Next, I show that quasi-maximum likelihood estimation (QMLE) ignoring second-stage heteroskedasticity yields inconsistent parameter and average marginal effect (AME) estimators in panel data. This is significant because QMLE is often employed in studies estimating heteroskedasticity or cluster robust standard errors. When instruments are weak, AME percentage bias can reach 650% in panels. Moreover, AME percentage bias in panels generally increases by a factor of 2-20 when heteroskedastic errors are assumed homoskedastic.

Declaration

I, Jack Mulqueeney, certify that:

The work presented in this dissertation is my own original work.

This dissertation does not contain material which has been accepted for the award of any other degree or diploma in my name.

This dissertation does not infringe on any copyright, trademark, patent or other rights whatsoever of any person.

The word count, excluding the abstract, the acknowledgements, the declaration, the bibliography, the figures and tables, and the appendix, is: 7,592.

Consent for the Distribution of Honours/Masters Dissertation

☐ The Economics Department (UWA) may use my Honours dissertation as an example of a dissertation which may get distributed to staff and future students.

Signature Date

Contents

List of Figures

List of Tables

1 Introduction

The core objective in much of modern applied economics research is quantifying causal relationships from noisy observational data. The doyen of what Angrist and Pischke (2010) call economics' "credibility revolution" is the method of instrumental variables (IV), which is ubiquitous in applied research for its ease of use and intuitive interpretation.

An increasingly popular setting for the IV method is binary outcome panel data, which is used to analyze questions ranging from what determines labour supply (Frijters et al., 2009; Fernández et al., 2014) to whether food aid causes civil war (Miguel et al., 2004; Nunn and Qian, 2014). However, two main problems arise in this setting. The first is that standard inference performs poorly when the instrument is weakly correlated with the endogenous variable of interest, known as the "weak instruments problem". Accordingly, determining whether and how reliable inference can be conducted in IV models of binary outcome panel data when instruments are weak is critical for future research. The second is that disregarding heteroskedasticity, now commonplace in applied research estimating so-called robust standard errors, might distort parameter and marginal effect estimates. If it does, then researchers must implement new estimation techniques in binary outcome panel data to ensure confidence in their results.

I present three contributions addressing these problems in binary outcome panel data. First, I propose a test robust to the presence of weak instruments in binary outcome panel data. Second, I show that estimating parameters by disregarding heteroskedasiticity can yield inconsistent parameter estimators even when instruments are strong. And third, I demonstrate through simulations that procedures to estimate average marginal effects (AMEs) can be severely biased in binary outcome panel data regardless of instrument strength.

The weak instruments problem for linear models is well-studied in the literature, generating non-Normal estimator distributions in environments with continuous dependent variables (Andrews et al., 2019). The literature, hence, seeks practical methods to either *detect* when instruments are weak or conduct valid inference despite, or *robust* to, the presence of weak instruments. The treatment of the problem in applied work typically relies on the former, where rules of thumb such as the first-stage F-statistic exceeding 10 (Staiger and Stock, 1997) or a two-stage least squares worst-case bias exceeding 10% of the worst-case ordinary least squares (OLS) bias (Stock and Yogo, 2005) are commonly cited. Although these assume homoskedastic errors, Olea and Pflueger (2013) provide a robust test for weak instruments with heteroskedasticity. The *robust* branch, meanwhile, seeks valid hypothesis tests regardless of instrument strength.

These generally follow the logic of Anderson-Rubin (AR) tests where the AR statistic follows a chi-squared distribution regardless of instrument strength (Anderson and Rubin, 1949; Andrews et al., 2019).

When outcomes are binary, researchers often apply nonlinear regression tools such as logit or probit to restrict fitted values, which represent the estimated probability of the outcome, between 0 and 1. However, standard asymptotic analysis employing first-order Taylor or mean-value expansions are no longer valid in nonlinear models (Frazier et al., 2021). Frazier et al. (2021) also find that the correct measure of instrument strength in nonlinear models depends on a density function. As the commonly cited rules of thumb were derived for linear models, they ignore this density multiplier. The rules of thumb, then, are not a reliable measure of instrument strength. As tests developed for continuous outcomes cannot be immediately applied to binary variables, estimation and inference methods currently used in studies with weak instruments must be reconsidered.

Magnusson (2010) develops a test robust to weak instruments in binary choice models using a minimum distance (MD) principle. The idea is to perform inference on a link function that is a metric between structural- and reduced-form parameters rather than directly on the parameter of interest. This test is of the correct size and generally dominates other tests in terms of power. Magnusson (2010)'s analysis lends itself to a cross sections environment because it does not address individual heterogeneity. My first contribution is extending Magnusson (2010)'s test to a panel data environment.

Panel data are of particular interest for both empirical and theoretical reasons. Empirically, panel data are powerful because, by observing individuals over time, the researcher can control for unobserved individual heterogeneity, called FEs. In linear models, including individual level dummy variables or implementing a first-differencing (FD) procedure controls for FEs. However, a FD approach cannot control for FEs in nonlinear models and cannot be applied to the binary choice environment. Meanwhile, introducing individual dummies to control for FEs causes the incidental parameters problem (Lancaster, 2000). The incidental parameters problem complicates current robust methods and inference techniques in panel data because it causes parameter estimators to be inconsistent. Extending the robust test to panel data is therefore a nontrivial task and constitutes a significant contribution to theoretical and applied literatures alike.

To implement the robust test in nonlinear panel data IV models, I propose an estimation

procedure that relies on a control function to resolve endogeneity. I implement a conditional logit (CL) to consistently estimate reduced-form parameters which avoids the incidental parameters problem. I show that this test has the correct size regardless of instrument strength through Monte Carlo simulations of the basic one endogenous variable, one instrument model. I also provide the functional form of the robust test in the general case. I illustrate the importance of the test by reinvestigating the central specifications in Nunn and Qian (2014). I find that confidence intervals computed with the robust test are up 6 times wider than those reported by standard inference.

Motivated by the convention in applied research of estimating heteroskedasticity or cluster robust standard errors using linear probability models (LPMs), I investigate how heteroskedastic errors affect estimator consistency for panel data IV models. I find that heteroskedasticity affects estimation differently depending on which stage it enters the IV model. If it enters purely via the first-stage, quasi-maximum likelihood estimation (QMLE) that assumes homoskedastic first-stage errors is consistent. However, heteroskedastic second-stage errors makes QMLE inconsistent regardless of instrument strength. Demonstrating that QMLE yields inconsistent estimators is my *second* contribution.

I illustrate these results via Monte Carlo simulations which show that estimators produced by logit, conditional logit, and LPMs that ignore heteroskedasticity are inconsistent. Researchers, then, must assume that heteroskedasticity exists exclusively in the first-stage while second-stage errors are conditionally homoskedastic to consistently estimate parameters. These findings directly challenge the practice of disregarding heteroskedasticity when estimating parameters in binary outcome panel data.

As applied research ultimately seeks to quantify relationships between outcome and explanatory variables, I analyse how misspecified heteroskedasticity affects average marginal effects (AMEs). This is of critical importance since AMEs are the primary object of interest in most applied work and are usually relied on to interpret results and evaluate policy interventions. For cross sections, I find that ignoring heteroskedasticity yields consistent AME estimates with two-group heteroskedasticity while they can be inconsistent with individual heteroskedasticity.

Given these findings, I consider how two-group heteroskedasticity might contaminate AMEs in panel data. I find that misspecified heteroskedasticity *generally* increases percentage bias in AME estimates by a factor of 2-20, although there are cases where AMEs estimated with misspecified errors exhibit lower percentage bias.

I also study inconsistency and bias for AMEs estimated from CL, standard logit, and LPMs under different assumptions about the FEs. When instruments are strong, percentage bias in the AMEs across all models goes to 0 as the time dimension lengthens. However, some AME estimation procedures can be 20-40% biased depending on how FEs are generated. When instruments are weak, AME percentage bias can increase up to 660% for nonlinear models like CL and standard logit. Percentage bias can be up to 2,300% when estimated from LPMs. Indeed, a CL model assuming full knowledge of true FEs and first-stage errors can feature AME percentage bias up to 12% in the presence of weak instruments.

Simulations demonstrating that AMEs estimated by ignoring heteroskedasticity can be severely biased is my *third* contribution to the literature. These results are significant since economists typically rely on AMEs to determine the success or failure of government policies such as humanitarian aid.

The outline of the thesis is as follows: Section 2 provides an overview of the estimation problem, the CL likelihood, and develops and verifies the robust test in panel data through Monte Carlo simulations. Section 3 discusses parameter estimation with heteroskedastic errors and Section 4 illustrates implications for AME estimation via Monte Carlo simulations. Section 5 provides an application of these methods to estimating the effect of food aid on civil conflict incidence as analysed in Nunn and Qian (2014). Section 6 concludes and discusses avenues for further research, while necessary proofs and further simulations are provided in Appendix A and Appendix B, respectively.

2 Robust Inference

Although well-studied in linear models, the problem of weak instruments in nonlinear models receives relatively little attention across theoretical and applied literatures. Despite this, applied economists apply traditional methods to both detect weak instruments and conduct inference on parameters of weakly identified endogenous variables. To address this, I outline a hypothesis test robust to the presence of weak instruments in binary dependent variable (BDV) models and demonstrate its correct size in simulated panel data.

2.1 Model and Setup

I wish to conduct inference on the following one endogenous variable, one instrument class of panel models with structural form

$$x_{it} = z_{it}\xi + b_i + v_{it} \tag{2.1}$$

$$y^*_{it} = x_{it}\beta + c_i + u_{it}, \text{ where } u_{it} = \rho v_{it} + \varepsilon_{it} \tag{2.2}$$

Equation (2.1) and (2.2) are called the first and second-stage equations, respectively. Here, x_{it} is the endogenous explanatory variable, β is the parameter of interest, ρ describes the degree of endogeneity, u_{it} is the second stage error-term, z_{it} is the instrument, ξ is the instrument parameter, and v_{it} is a first-stage error term, for individual i in time t. As x_{it} is endogenous it must be correlated with u_{it}. Assuming the exclusion restriction, this implies that x_{it} must be correlated with u_{it} via v_{it}. This observation motivates a control function, so that $u_{it} = \rho v_{it} + \varepsilon_{it}$, for $0 < \rho < 1$ and ε_{it} is an independent second-stage random error term. The general model with multiple instruments, endogenous variables, and control variables is in Appendix A.2. Equation (2.1) and (2.2) have reduced-form

$$x_{it} = z_{it}\xi + b_i + v_{it} \tag{2.3}$$

$$y^*_{it} = z_{it}\delta_z + v_{it}\delta_v + \kappa_i + \varepsilon_{it} \tag{2.4}$$

where $\delta_z = \xi\beta$, $\delta_v = \beta + \rho$, and $\kappa_i = c_i + \beta b_i$. I assume that $v_{it} \sim \mathcal{N}(0, \sigma_v^2)$ and ε_{it} has some distribution with homoskedastic errors. The assumption that v_{it} is Normal can be relaxed, as shown in subsection 2.4. Define

$$y_{it} = \begin{cases} 1 & \text{if } y^*_{it} > 0 \\ 0 & \text{otherwise} \end{cases} \tag{2.5}$$

and assume that ε_{it} follows a Logistic distribution with location 0 and scale parameter s; that is, $\varepsilon_{it} \sim \mathscr{L}(0, s)$. This implies $\mathbb{P}(y_{it} = 1 | x_{it}, v_{it}, c_i) = \Lambda([x_{it}\beta + \rho v_{it} + c_i]/s)$, where $\Lambda(\cdot)$ is the standard Logistic cumulative distribution function (CDF) which has location 0 and scale 1. The Logistic CDF is chosen to estimate second-stage reduced-form parameters from Equation (2.4) via conditional logit likelihood. This likelihood circumvents the incidental parameters problem and yields consistent estimates of δ_z and δ_v, denoted $\hat{\delta}_z$ and $\hat{\delta}_v$ respectively. Given this setup, I wish to conduct inference not directly on the endogenous parameter β, but on the link function $r(\delta_z, \beta) = \delta_z - \xi\beta$.

2.2 Conditional Logit

A further discussion of the likelihood to estimate the second-stage in Equation (2.1) and (2.2) is required before outlining the robust test. I wish to estimate and conduct inference about β in the model described in subsection 2.1 Equation (2.1) and (2.2). Including dummy variables for each individuals via standard logit yields an inconsistent estimator of β. Hence, a procedure providing consistent parameter estimators despite the inclusion of FEs is required. The conditional logit (CL) is one such procedure which is derived by conditioning on $n_i \equiv \sum_{t=1}^{T} y_{it}$, where $0 < n_i < T$. The cases where $n_i = 0$ and $n_i = T$ contain no information to estimate β and so do not contribute to the likelihood. To fix ideas, I derive the likelihood for the $T = 2$ case while the T-period case is derived in Appendix A.1. Assume

Assumption 2.1. *The idiosyncratic errors $\varepsilon_{it} \sim \mathscr{L}(0, s)$, meaning $\mathbb{P}(y_{it} = 1|x_{it}, v_{it}, c_i) = \Lambda([x_{it}\beta + \rho v_{it} + c_i]/s)$, where $\Lambda(\cdot)$ is the standard Logistic CDF.*

Assumption 2.2. *The ε_{it} are independent of the entire regressor history. That is, $\mathbb{P}(y_{it} = 1|x_i, v_i, c_i) = \mathbb{P}(y_{it} = 1|x_{it}, v_{it}, c_i)$ where $x_i = [x_{i1}, ..., x_{iT}]$ and $v_i = [v_{i1}, ..., v_{iT}]$.*

Assumption 2.3. *The y_{it} are independent across time conditional on explanatory variables. That is, $\mathbb{P}(y_i|x_i, v_i, c_i) = \prod_{t=1}^{T} \mathbb{P}(y_{it} = y_t|x_{it}, v_{it}, c_i)$, where $y_i = [y_{i1}, ..., y_{iT}]$.*

Suppose, without loss of generality, that $y_{i2} = 1$. Assumption 2.2 and the Law of Conditional Probability imply the conditional probability

$$p_i = \mathbb{P}(y_{i2} = 1|n_i = 1, x_i) = \frac{\mathbb{P}(y_{i2} = 1, n_i = 1|x_i)}{\mathbb{P}(n_i = 1)} = \frac{\mathbb{P}(y_{i2} = 1, y_{i1} = 0|x_i)}{\mathbb{P}(n_i = 1)}$$

where $x_i = [x_{i1}, x_{i2}]$. By Assumption 2.1 and 2.3, the numerator is

$$\mathbb{P}(y_{i2} = 1, y_{i1} = 0|x_i) = \mathbb{P}(y_{i2} = 1|x_i)\mathbb{P}(y_{i1} = 0|x_i) = \Lambda(x_{i2}\beta + c_i)[1 - \Lambda(x_{i1}\beta + c_i)]$$

and by the Law of Total Probability the denominator is

$$\begin{aligned}\mathbb{P}(n_i = 1) &= \mathbb{P}(y_{i2} = 1, y_{i1} = 0|x_i) + \mathbb{P}(y_{i2} = 0, y_{i1} = 1|x_i) \\ &= \Lambda(x_{i2}\beta + c_i)[1 - \Lambda(x_{i1}\beta + c_i)] + \Lambda(x_{i1}\beta + c_i)[1 - \Lambda(x_{i2}\beta + c_i)]\end{aligned}$$

Therefore, p_i simplifies to

$$\begin{aligned}p_i &= \frac{\Lambda(x_{i2}\beta + c_i)[1 - \Lambda(x_{i1}\beta + c_i)]}{\Lambda(x_{i2}\beta + c_i)[1 - \Lambda(x_{i1}\beta + c_i)] + \Lambda(x_{i1}\beta + c_i)[1 - \Lambda(x_{i2}\beta + c_i)]} \\ &= \frac{\exp(x_{i2}\beta + c_i)}{\exp(x_{i2}\beta + c_i) + \exp(x_{i1}\beta + c_i)} = \frac{\exp([x_{i2} - x_{i1}]\beta)}{\exp([x_{i2} - x_{i1}]\beta) + 1} = \Lambda([x_{i2} - x_{i1}]\beta)\end{aligned}$$

The other case where $y_{i1} = 1$ is simply $1 - p_i$ by the Law of Complements as the event $A = \{y_{i1} = 1\}$ is the complement of $B = \{y_{i2} = 1\}$ after conditioning on $n_i = y_{i1} + y_{i2}$. Hence, the likelihood contribution of individual i is

$$\mathscr{L}_i(\beta) = \Lambda([x_{i2} - x_{i1}]\beta)^{w_i}[1 - \Lambda([x_{i2} - x_{i1}]\beta)]^{1-w_i}$$

where $w_i = 1$ when $(y_{i1}, y_{i2}) = (0, 1)$ and $w_i = 0$ when $(y_{i1}, y_{i2}) = (1, 0)$. The log-likelihood contribution of individual i is then

$$\ell_i(\beta) = w_i \log\{\Lambda([x_{i2} - x_{i1}]\beta)\} + (1 - w_i) \log\{1 - \Lambda([x_{i2} - x_{i1}]\beta)\}$$

where $\log(\cdot)$ is the natural logarithm. The conditional log-likelihood does not depend on the FEs and thus circumvents the incidental parameters problem. Therefore, estimating the reduced-form second-stage equation (2.4) with conditional logit provides consistent estimates of δ_z and δ_v.

2.3 Test Algorithm

I outline an algorithm to conduct inference on β that is robust to weak instruments in the first-stage. The test, called the AR test, indirectly infers about β by conducting inference on the link function $r(\delta_z, \beta) = \delta_z - \xi\beta$. For the hypothesis $H_0 : \beta = \beta_0$, the link function $r(\hat{\delta}_z, \beta_0) = \hat{\delta}_z - \hat{\xi}\beta_0$ has variance

$$\hat{\Psi}_{\beta_0} = \mathbb{V}(\hat{\delta}_z - \hat{\xi}\beta_0) = \hat{\Lambda}_{\delta_z\delta_z} + (\hat{\delta}_v - \beta_0)^2\hat{\Lambda}_{\xi\xi} \tag{2.6}$$

where $\hat{\delta}_v$ and $\hat{\delta}_z$ are estimators of δ_v and δ_z, respectively; $\hat{\Lambda}_{\delta_z\delta_z}$ is the variance of $\hat{\delta}_z$, and $\hat{\Lambda}_{\xi\xi}$ is the variance of the first-stage instrument estimator $\hat{\xi}$. Given this, the AR test is the quadratic form of $r(\hat{\delta}_z, \beta_0) = \hat{\delta}_z - \hat{\xi}\beta_0$

$$\text{AR}(\beta_0) = (\hat{\delta}_z - \hat{\xi}\beta_0)^2\hat{\Psi}_{\beta_0}^{-1} \overset{a}{\sim} \chi^2(1) \tag{2.7}$$

Given (2.6) and (2.7), The AR test algorithm for the one instrument, one endogenous variable case in panel data is

1. Compute the estimator $\hat{\xi}$ of ξ and its variance $\hat{\Lambda}_{\xi\xi}$ by ordinary least squares (OLS) on Equation (2.1). Also, compute the first-stage residuals $\hat{v}_{it}$

2. Compute the estimators $\hat{\delta}_z$ and $\hat{\delta}_v$ of reduced-form parameters δ_z and δ_v, respectively, via conditional logit on Equation (2.4) with $\hat{v}_{it}$ included as a regressor. Also, compute the variance of $\hat{\delta}_v$, denoted $\hat{\Lambda}_{\delta_z\delta_z}$

3. Substitute $\hat{\xi}$, $\hat{\delta}_v$, $\hat{\delta}_z$, $\hat{\Lambda}_{\xi\xi}$, and $\hat{\Lambda}_{\delta_z\delta_z}$ into Equation (2.6) and (2.7) along with the hypothesised β_0. The critical value for $100\alpha\%$ significance level is $c_{1-\alpha}$ such that $\mathbb{P}(\chi^2(1) > c_{1-\alpha}) = \alpha$. If AR($\beta_0$) exceeds the critical value, then reject the null hypothesis that $\beta = \beta_0$

The AR test does not conduct inference directly about β but rather on the link function $r(\delta_z, \beta)$. The idea behind the test is to calculate whether the distance $r(\hat{\delta}_z, \beta_0) = \hat{\delta}_z - \hat{\xi}\beta_0$ is statistically different from 0. Here, $\hat{\delta}_z$ is the consistently estimated reduced-form parameter and $\hat{\xi}\beta_0$ is the estimated value of the reduced-form parameter assuming $\beta = \beta_0$. The link function $r(\delta_z, \beta_0)$ being statistically different from 0 is evidence against the assumption $\beta = \beta_0$. So, inferring about the link function is equivalent to indirectly inferring about the parameter of interest. In this way, the AR test is robust to the weak instrument problem.

2.4 Concentration Parameter

To measure instrument strength, I define the concentration parameter μ which fixes instrument strength across simulations with different samples. The concentration parameter μ is rearranged for the scalar instrument ξ

$$\mu = \frac{1}{k_z}\left(\frac{\xi' z' z \xi}{\sigma_v^2}\right) \text{ so that } \xi = \sigma_v \sqrt{k_z \mu (z'z)^{-1}},$$

where σ_v^2 is the variance of the first-stage errors v_{it}. I normalise $z'z = 1$ and set $\mu \in \{0.01, 3, 500\}$ to represent changing instrument strength from very weak ($\mu = 0.01$), weak ($\mu = 3$), and very strong ($\mu = 500$) (Staiger and Stock, 1997).

2.5 Simulations

I conduct 10,000 simulations of a panel consisting of $n = 100$ individuals each observed over $T = 10$ time periods of the simple one endogenous variable, one instrument model described by Equation (2.1) and (2.2). The true parameter of interest is $\beta_0 = 0.5$ and the endogeneity parameter takes values $\rho = 0.20$ and $\rho = 0.99$ to describe low and high endogeneity, respectively. I generate z_{it} and v_{it} from the standard Normal distribution while ε_{it} are drawn from a Logistic distribution with centre 0 and scale parameter $\sqrt{3}/\pi$. This scale parameter was chosen as it yields a unit variance for ε_{it}. The first-stage fixed effects b_i are drawn from a Uniform distribution, $\mathcal{U}(-0.5, 0.5)$, and the second-stage fixed effects are $c_i = \rho b_i$.

Table 1 shows that the AR test has approximately the correct small sample size regardless of instrument strength whereas standard Wald tests over-reject by up to 100% of the correct size

when instruments are weak. Standard Wald tests over-reject the null hypothesis even when very strong instruments are present ($\mu = 500$).

Table 1: Size Comparison (%) – Conditional Logit Panel Model

	$\mu = 0.01$				$\mu = 3$				$\mu = 500$			
ρ	0.2		0.99		0.2		0.99		0.2		0.99	
Size	5%	10%	5%	10%	5%	10%	5%	10%	5%	10%	5%	10%
Wald($\hat{\beta}$)	7.48	13.10	12.55	19.92	5.29	10.72	12.08	19.50	5.26	10.39	5.55	11.19
AR(β_0)	4.85	10.11	4.83	9.67	4.96	10.39	5.13	10.23	5.30	10.29	4.73	9.95

Note. The above results implement the robust inference algorithm. Values in the table are percentages describing the proportion of hypothesis tests rejected at the 5% and 10% levels. I test the hypothesis $H_0 : \beta = 0.5$ against $H_1 : \beta \neq 0.5$. For both Wald and AR tests, standard errors are corrected for the generated regressor $\hat{v}_{it}$.

That the Wald test over-rejects when $\mu = 500$ in small samples adds to the evidence against applying standard inferential methods in empirical analyses of binary outcome panel data. I interpret this result with respect to the findings of Frazier et al. (2021), who find that μ overlooks the non-linearity of the estimation procedure. To measure instrument strength, μ measures the variability of $z_{it}\xi_0$, which assumes the second-stage equation is linear. Frazier et al. (2021) show that $g(\cdot)z_{it}\xi_0$, and not $z_{it}\xi_0$, captures the true first-stage variability in nonlinear models where $g(\cdot)$ is a density function. Weighting by $g(\cdot)$ reduces the variability and thus weakens instrument strength beyond what is strictly captured by μ. Thus, μ overstates the variability of the endogenous regressor with respect to the instrument, meaning μ may indicate strong instruments when they are in fact weak. This is demonstrated in mild over-rejection rates in standard inference when $\mu = 500$ and $\rho = 0.99$, indicative of extremely strong instruments and high endogeneity in the linear case.

Results in Appendix B.1 reiterates that the AR test has approximately correct size in small samples while Wald tests over-reject in alternative specifications, such as estimating reduced-form parameters via a linear probability model (LPM) rather than CL.

2.6 Confidence Intervals

To construct $100(1-\alpha)\%$ confidence intervals for scalar β from the AR test, define a grid of parameter values $\mathscr{B} = \{\beta : a \leq \beta \leq b\}$, where $a, b \in \mathbb{R}$. For every $\beta_k \in \mathscr{B}$, compute the

$\text{AR}(\beta_k)$ statistic and decide whether to reject at the $100\alpha\%$ significance level. The $100(1-\alpha)\%$ confidence interval is then

$$\mathscr{C} \equiv \{\beta_k \in \mathscr{B} : \text{AR}(\beta_k) < c_{1-\alpha}\}$$

where $c_{1-\alpha}$ is a real number such that $\mathbb{P}(\chi^2(1) \geq c_{1-\alpha}) = \alpha$. When β is an m-dimensional vector, specify grids each component β_j of β of the form $\mathscr{B}_j \equiv \{\beta_j : a_j \leq \beta_j \leq b_j\}$ where $a_j, b_j \in \mathbb{R}$. The confidence set is constructed by calculating the AR test at all elements of the Cartesian product of the $\mathscr{B}_j$ for $j = 1, 2, ...k$, which is the set of all m-tuples whose j^{th} component is an element of $\mathscr{B}_j$ for all $j = 1, 2, ..., m$.

3 Heteroskedasticity

Economists often rely on estimating heteroskedasticity or cluster robust standard errors to accommodate a more general variance-covariance structure. However, parameters are estimated under the assumption that errors are homoskedastic. How this variance misspecification affects parameter estimation is often overlooked.

Recall the control function formulation of second-stage errors in Equation (2.1) $u_{it} = \rho v_{it} + \varepsilon_{it}$, where v_{it} are first-stage errors and ε_{it} are idionsyncratic second-stage error term. I show that quasi-maximum likelihood estimation (QMLE) that misspecifies a heteroskedastic ε_{it} as homoskedastic produces an inconsistent estimator $\tilde{\beta}$ of the true β_0. However, QMLE that misspecifies a heteroskedastic v_{it} as homoskedastic remains consistent for the true β_0. I illustrate these theoretical results with a variety of models such as CL and LPM in simulated panel data.

3.1 Two-Group Heteroskedasticity

Consider the case where observations are drawn from two groups with different error variances. To focus and simplify the discussion, I examine a panel with n individuals each observed over $T = 2$ periods with no endogenous variable. From this analysis, I infer issues with the general T-period likelihood. An explicit discussion of the IV model is provided in Appendix A.4.

Call the set of individuals in group 1 $\mathscr{O}$ and the set of individuals in group 2 $\mathscr{T}$. The two-group latent variable is

$$y_{it}^* = \begin{cases} x_{it}\beta + c_i + e_{it}^{(1)}, & \text{when } i \in \mathscr{O} \\ x_{it}\beta + c_i + e_{it}^{(2)}, & \text{when } i \in \mathscr{T} \end{cases}$$

where x_{it} is a scalar explanatory variable, β its corresponding parameter with true value β_0, c_i are individual level FEs, and $e_{it}^{(j)}$ denotes the error term for group $j = 1, 2$. Assume that $e_{it}^{(1)}$ and $e_{it}^{(2)}$ are Logistically distributed with location 0 and have variance σ_1^2 and σ_2^2, respectively. Assume that y_{it} is defined as in Equation (2.5).

Let the true $\beta_0 \in \mathrm{B}$ where B is a compact subset of $\mathbb{R}$, be the unique parameter vector which solves

$$\beta_0 = \arg\max_{\beta \in \mathrm{B}} \mathbb{E}\left[g_i \ell_i^{(1)}(\beta) + (1 - g_i)\ell_i^{(2)}(\beta)\right] \tag{3.1}$$

where $\ell_i^{(1)}(\cdot)$ and $\ell_i^{(1)}(\cdot)$ are the log-likelihood contribution of individual i in group 1 or 2, respectively, and g_i if i is in group 1 and $g_i = 0$, otherwise. Following the derivation of the 2-period log-likelihood in subsection 2.2, the log-likelihood contribution for individual i in group $j = 1, 2$ of observed sample data are

$$\begin{aligned}
\ell_i^{(1)}(y_i|x_i, c_i, \beta) &= w_i \log \Lambda_1[(x_{i2} - x_{i1})\beta] + (1 - w_i)\log\{1 - \Lambda_1[(x_{i2} - x_{i1})\beta]\} \\
\ell_i^{(2)}(y_i|x_i, c_i, \beta) &= w_i \log \Lambda_2[(x_{i2} - x_{i1})\beta] + (1 - w_i)\log\{1 - \Lambda_2[(x_{i2} - x_{i1})\beta]\}
\end{aligned}$$

where $w_i = 1$ when $(y_{i1}, y_{i2}) = (0, 1)$, $w_i = 0$ when $(y_{i1}, y_{i2}) = (1, 0)$, and

$$\Lambda_j(k_i) = \frac{\exp(k_i/q\sigma_j)}{\exp(k_i/q\sigma_j) + 1} \text{ for } j = 1, 2$$

for $q = \sqrt{3}/\pi$ and $k_i(\beta) = (x_{i2} - x_{i1})\beta$. I omit the argument β in $k_i(\beta)$ to ease expression unless it improves clarity. Therefore, β_0 has consistent estimator

$$\hat{\beta} = \arg\max_{\beta \in \mathrm{B}} \frac{1}{n} \sum_{i=1}^{n} \left\{g_i \ell_i^{(1)}(y_i|x_i, c_i, \beta) + (1 - g_i)\ell_i^{(2)}(y_i|x_i, c_i, \beta)\right\}$$

QMLE would ignore the underlying heteroskedasticity and assume $\mathbb{V}(e_{it}^{(1)}) = \mathbb{V}(e_{it}^{(1)}) = \sigma^2$. Hence, the quasi-log-likelihood contribution of individual i is

$$\ell_i^{q}(y_i|x_i, c_i, \beta) = \log \Lambda[(x_{i2} - x_{i1})\beta], \text{ for } \Lambda(k_i) = \frac{\exp(k_i/q\sigma)}{\exp(k_i/q\sigma) + 1}$$

QMLE yields the estimator of β_0

$$\tilde{\beta} = \arg\max_{\beta \in \mathrm{B}} \sum_{i=1}^{n} \ell_i^{q}(y_i|x_i, c_i, \beta)$$

The first result is

Result 3.1. *$\tilde{\beta}$ is an inconsistent estimator of β_0.*

Proof is provided in Appendix A.3.

Result 3.1 means that when heteroskedasticity in the error term e_{it} is ignored parameter estimates are inconsistent for β_0. The single equation model and Result 3.1 motivates the following

Result 3.2. *Consider the two-stage IV model described by Equation (2.1) and (2.2), where β is the parameter of the endogenous variable of interest with true value β_0. Call v_{it} and ε_{it} the idiosyncratic first- and second-stage errors, respectively. QMLE that misspecifies a heteroskedastic ε_{it} as homoskedastic is inconsistent for true parameter β_0. QMLE that misspecifies heteroskedastic v_{it} as homoskedastic while $u_{it} = \rho v_{it} + \varepsilon_{it}$ is conditionally homoskedastic, which means that ε_{it} is homoskedastic, remains consistent for true parameter β_0.*

Proof is provided in the Appendix A.4.

Result 3.2 implies that researchers must be careful about ignoring heteroskedasticity in panel binary dependent variable IV models as subsequent parameter estimators, and all inference therein, may be inconsistent. Specifically, when heteroskedasticity in v_{it} is ignored, estimators remain consistent. However, when idiosyncratic second-stage heteroskedasticity via ε_{it} is ignored, all subsequent parameter estimators are inconsistent. Hence, assuming that ε_{it} is homoskedastic is necessary for parameter estimates and subsequent inference to be consistent and underpins analyses in applied research implementing IV in binary outcome panel data. I illustrate this issue in subsection 3.2 via Monte Carlo simulations.

3.2 Simulations

I estimate linear probability models (LPMs), panel data logit with dummy individual FEs (UCL), and conditional logit (CL) in simulated panel data to illustrate Result 3.2. Simulations include one endogenous variable and one instrument. I compare estimator consistency between when a heteroskedastic ε_{it} or v_{it} is assumed homoskedastic. I simulate the one instrument, one endogenous variable model

$$\begin{aligned} x_{it} &= z_{it}\xi + b_i + v_{it} \\ y_{it}^* &= x_{it}\beta + c_i + \rho v_{it} + \varepsilon_{it} \end{aligned}$$

for $n = 100$ individuals over $T = 10$ time periods. The true parameter of interest is $\beta_0 = 0.5$, $\rho \in \{0.20,\ 0.99\}$, ξ is set such that $\mu = 500$ (strong instruments) and $\mu = 1$ (weak instruments), and second-stage fixed effects, c_i are generated from $\mathcal{U}(-0.5, 0.5)$, $b_i = 0$ for all i. The different

models are summarised in Table 2. How v_{it} and ε_{it} are generated varies and is described in subsection 3.2.1 and 3.2.2.

Table 2: Different Estimation Procedures – Panel QMLE

Model	*Estimator*	*Regression*
UCL	$\hat{\beta}_{UCL}$	Regress y_{it} on x_{it}, $\hat{v}_{it}$, and dummy c_i via logit
CL	$\hat{\beta}_{CL}$	Regress y_{it} on x_{it} and $\hat{v}_{it}$ via conditional logit
LPM	$\hat{\beta}_{LPM}$	Regress y_{it} on x_{it}, $\hat{v}_{it}$, and dummy c_i via ordinary least squares (OLS)

3.2.1 Heteroskedastic First-Stage Errors

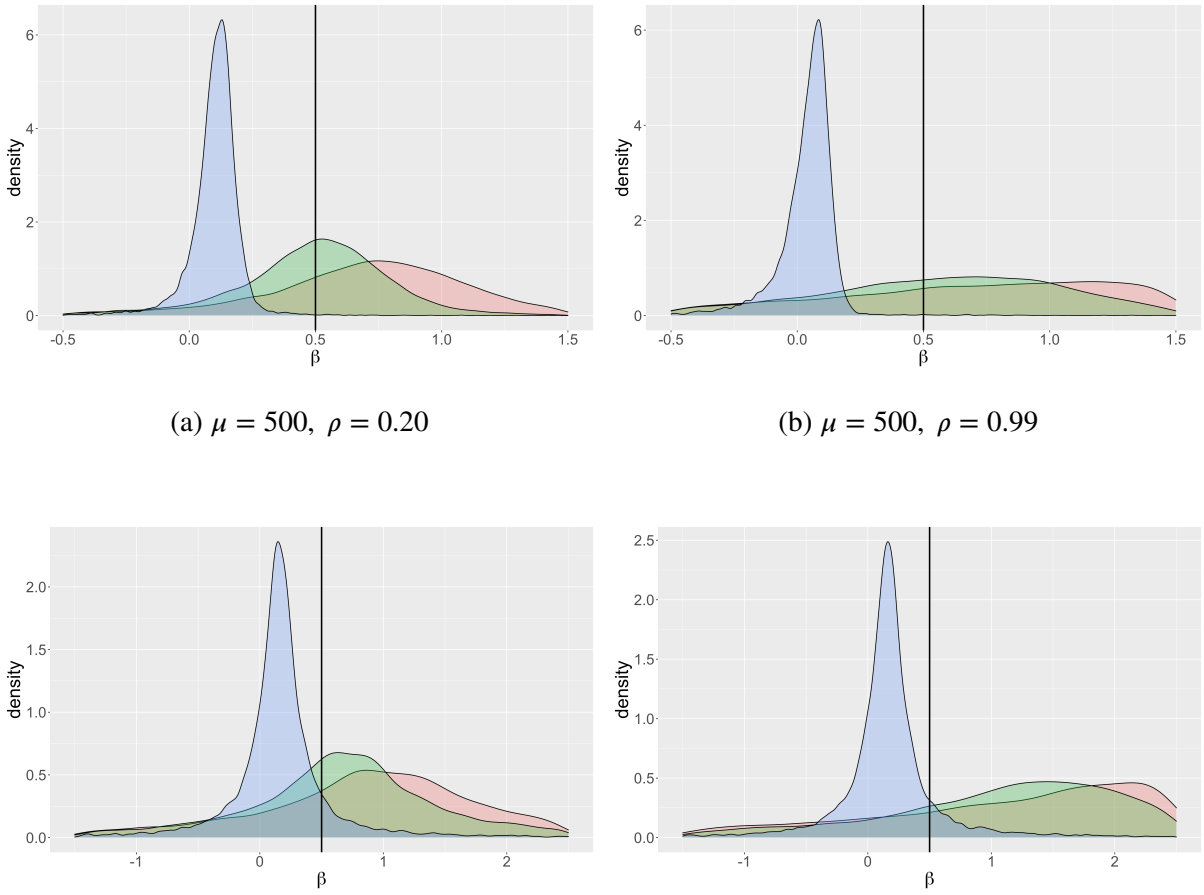
First, I illustrate that QMLE assuming heteroskedastic first-stage errors v_{it} are homoskedastic is consistent so long as ε_{it} is homoskedastic. To simulate this, I generate v_{it} from the multivariate Normal distribution

$$v_{it} \sim \mathcal{N}\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}\sigma_1^2 & 0\\0 & \sigma_2^2\end{bmatrix}\right)$$

where $\sigma_1^2 = 100$ and $\sigma_2^2 = 0.01$ represent the variance of group 1 and group 2, respectively. Each group contains 50% of the observations each. Furthermore, I generate ε_{it} from Logistic distribution with location 0 and scale parameter $s = 1$. I estimate β assuming that v_{it} are from a standard Normal distribution.

First-stage heteroskedasticity biases the LPM model towards 0 in all instrument strength and endogeneity specifications. In Figure 1(a), $\hat{\beta}_{CL}$ is Normal and approximately centred around $\beta_0 = 0.5$ for low and high endogeneity, demonstrating that QMLE ignoring first-stage heteroskedasticity remains consistent. The estimator $\hat{\beta}_{UCL}$, however, is inconsistent. When instruments are weak, estimator distributions become misshapen and inconsistently estimate β_0. Inconsistency worsens as the degree of endogeneity ρ increases. Both $\hat{\beta}_{CL}$ and $\hat{\beta}_{UCL}$ densities are slightly left skewed when $\rho = 0.99$ and $\mu = 1$.

Figure 1: Logit and LPM estimators with heteroskedastic v_{it} and homoskedastic ε_{it}



(a) $\mu = 500,\ \rho = 0.20$

(b) $\mu = 500,\ \rho = 0.99$

(c) $\mu = 1,\ \rho = 0.20$

(d) $\mu = 1,\ \rho = 0.99$

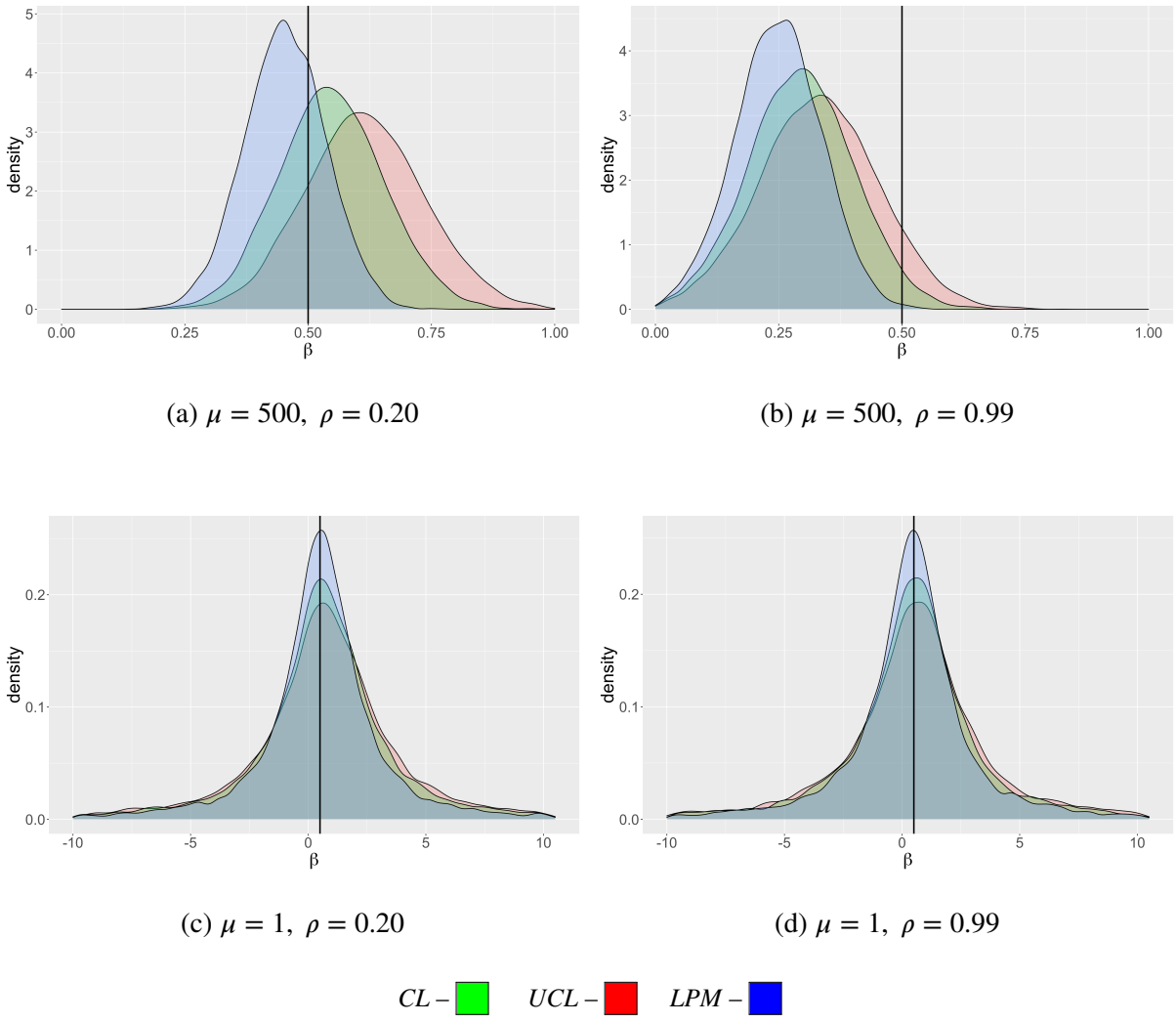
3.2.2 Heteroskedastic Second-Stage Errors

I show that QMLE assuming heteroskedastic ε_{it} are homoskedastic yields inconsistent parameter estimates. Individuals are separated into two groups each encompassing 50% of the sample, where the second-stage errors of groups 1 and 2 are Logistic distributed with location 0 and variances $\mathbb{V}[\varepsilon_{it}^{(1)}] = 100$ and $\mathbb{V}[\varepsilon_{it}^{(2)}] = 0.01$, respectively. The v_{it} are Normal with mean 0 and variance 1.

Figure 2 demonstrates that even when instruments are strong, ignoring second-stage heteroskedasticity yields inconsistent parameter estimates. When $\rho = 0.20$, the CL and UCL estimators are upwardly biased while when $\rho = 0.99$ they feature a downward bias. When instruments are weak, parameter estimates are significantly more dispersed and still feature a

bias, although this is not immediately evident from Figure 2(c) and 2(d). The LPM estimator is downwardly biased regardless of the degree of endogeneity ρ.

Figure 2: Logit and LPM estimators with homoskedastic v_{it} and heteroskedastic ε_{it}



(a) $\mu = 500,\ \rho = 0.20$

(b) $\mu = 500,\ \rho = 0.99$

(c) $\mu = 1,\ \rho = 0.20$

(d) $\mu = 1,\ \rho = 0.99$

Clearly, ignoring heteroskedasticity yields inconsistent parameter estimates even when instruments are strong as asserted by Result 3.2. This means applied research estimating heteroskedasticity robust standard errors *must assume that second-stage errors* $u_{it} = \rho v_{it} + \varepsilon_{it}$ *are conditionally homoskedastic so all subsequent parameter and standard error estimates are consistent.*

4 Marginal Effects

It is possible that parameter inconsistency from ignored heteroskedasticity translates across to the estimation of marginal effects. Investigating how ignored heteroskedasticity may contaminate marginal effect estimation is crucial since they are pervasive in applied research for interpreting nonlinear regression results. For these reasons, I extend the simulations in subsection 3.2 to investigate marginal effects between different models in panels with heteroskedastic errors. I outline the basic theory behind marginal effects, derive equations to compute them in the presence of heteroskedasticity, then progress to the simulations.

As the proposed conditional logit procedure does not estimate FEs, marginal effects cannot be directly computed without *ex post facto* assumptions about the FEs. Therefore, I first analyse the effects of different assumptions about FEs in conditional logit procedures and then discuss the effects of heteroskedasticity misspecification.

4.1 Background

Suppose $\mathbf{x}$ is an $nT\times k$ matrix of explanatory variables and $\boldsymbol{\theta}$ is a $k\times 1$ vector of parameters, where n represents the number of individuals and T the time periods. Consider the model $\mathbb{E}[y_{it}|\mathbf{x}_{it}] = \mathbb{P}(y_{it}=1|\mathbf{x}_{it}) = G(\mathbf{x}_{it}\boldsymbol{\theta})$, where $G(\cdot)$ is a CDF. Assuming that $G(\mathbf{x}\boldsymbol{\theta})$ is differentiable and $\mathbf{x}$ are continuous, the partial derivative with respect to the j^{th} explanatory variable is

$$\frac{\partial \mathbb{E}[y_{it}|\mathbf{x}_{it}]}{\partial x_j} = \frac{\partial \mathbb{P}(y_{it}=1|\mathbf{x}_{it})}{\partial x_j} = g(\mathbf{x}_{it}\boldsymbol{\theta})\theta_j \approx \frac{\Delta \mathbb{E}[y_{it}|\mathbf{x}_{it}]}{\Delta x_j}, \text{ for small } \Delta x_j \tag{4.1}$$

where $g(\cdot) = G'(\cdot)$ is the density of CDF $G(\cdot)$. Equation (4.1), called the marginal effect, represents the rate of change in the response probability with respect to the explanatory variable x_j.

As I examine panel models with fixed effects, it is natural to consider the conditional expectation to be a function of observable and unobservable explanatory variables, $\mathbf{x}$ and c respectively. This implies $\mathbb{E}[y_{it}|\mathbf{x}_{it}, c_i] = f(\mathbf{x}_{it}, c_i)$ and

$$\delta_j(\mathbf{x}, c) \equiv \frac{\partial f(\mathbf{x}, c)}{\partial x_j} = \theta_j f'(\mathbf{x}, c) \tag{4.2}$$

Equation (4.2) however is not useful in practice as δ_j requires knowledge of c, which is a contradiction since it is unobservable. To circumvent this problem, average marginal effects (AMEs) are commonly estimated (Wooldridge, 2010). A consistent estimator of AMEs in binary out-

come panel data is

$$\bar{d}_j(\mathbf{x}^o) \equiv \frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\frac{\partial \mathbb{E}[y_{it}|\mathbf{x}_{it}]}{\partial x_{i,j}}\Big|_{\mathbf{x}=\mathbf{x}^o} = \frac{\hat{\theta}_j}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T} g(\mathbf{x}_{it}^o\hat{\boldsymbol{\theta}})$$

Standard errors of $\bar{d}_j(\mathbf{x}^o)$ are obtained via the Delta method, demonstrated in Appendix A.5.

4.2 AMEs and Heteroskedasticity

Consider the model

$$\begin{aligned} x_{it} &= z_{it}\xi + b_i + v_{it} \\ y_{it}^* &= x_{it}\beta + v_{it}\rho + c_i + \varepsilon_{it} \\ \mathbb{V}[\varepsilon_{it}] &= [V(h_i\gamma)]^2 > 0 \text{ so that } \varepsilon_{it} \sim \mathscr{L}\big(0, qV(h_i\gamma)\big) \\ v_{it} &\sim \mathcal{N}(0,\sigma^2) \end{aligned}$$

where h_i is a matrix of explanatory variables determining an individual's error variance, γ the corresponding parameter vector, and $q = \sqrt{3}/\pi$ is a constant. Let the observed binary dependent variable be $y_{it} = 1$ when $y_{it}^* > 0$ and $y_{it} = 0$ when $y_{it}^* < 0$. Assume that

$$\mathbb{P}(y_{it} = 1|x_{it}, v_{it}, c_i) = \Lambda\Big\{\frac{x_{it}\beta + v_{it}\rho + c_i}{qV(h_i\gamma)}\Big\}$$

where $\Lambda(\cdot)$ is the standard Logistic CDF with location 0 and scale 1. To compute AMEs in this setting, I mirror Greene (2003) and unite second-stage variables as $\mathbf{b}_{it} = (x_{it}, v_{it})$ with corresponding parameter vector $\boldsymbol{\beta} = (\beta, \rho)'$ and all regressors $\mathbf{g}_{it} = (\mathbf{b}_{it}', h_i)'$, with corresponding parameter vector $\boldsymbol{\omega} = (\boldsymbol{\beta}', \gamma)$. I derive marginal effects with respect to all components of $\mathbf{g}$

$$\frac{\partial \mathbb{P}(y_{it} = 1|\mathbf{g}_{it}, c_i)}{\partial g_k} = \lambda\Big\{\frac{x_{it}\beta + v_{it}\rho + c_i}{qV(h_i\gamma)}\Big\}\frac{b_k V(h_{it}\gamma) - h_k V'(h_i\gamma)(x_{it}\beta + v_{it}\rho + c_i)}{q[V(h_i\gamma)]^2}$$

where

$$b_k \equiv \begin{cases} \boldsymbol{\beta}_j & \text{if } g_k \text{ is the } j^{th} \text{ element of } \mathbf{b} \\ 0, & \text{otherwise} \end{cases} \quad \text{and} \quad h_k \equiv \begin{cases} \gamma_j & \text{if } g_k \text{ is the } j^{th} \text{ element of } h \\ 0, & \text{otherwise} \end{cases}$$

Thus, the AME evaluated at fixed $\mathbf{g}^o = (x^o, v^o, h^o)'$ is

$$\bar{d}(\boldsymbol{g}^o) = \frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\lambda\Big\{\frac{x_{it}^o\beta + v_{it}^o\rho + c_i}{qV(h_i^o\gamma)}\Big\}\frac{b_k V(h_i^o\gamma) - h_k V'(h_i^o\gamma)(x_{it}^o\beta + v_{it}^o\rho + c_i)}{q[V(h_i^o\gamma)]^2} \tag{4.3}$$

which is the AME evaluated at fixed $\mathbf{g}^o = (x^o, v^o, h^o)'$ and parameter vector $\boldsymbol{\omega} = (\beta, \rho, \gamma)'$. Assuming that $\hat{\boldsymbol{\omega}} = (\hat{\theta}, \hat{\rho}, \hat{\gamma})'$ is a consistent estimator of $\boldsymbol{\omega}_0$, $\bar{d}(\mathbf{g}^o)$ evaluated with $\hat{\boldsymbol{\omega}}$ is a consistent

estimator of the true AME. The estimator $\hat{\boldsymbol{\omega}}$ is obtained via conditional logit on the second-stage where v_{it} is replaced by the first-stage estimated residuals $\hat{v}_{it}$. QMLE that falsely assumes $V(h_i\gamma) = \sigma$ yields the AME estimator

$$\bar{d}^{Q}(\mathbf{g}^{o}) = \frac{b_k}{q\sigma nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\lambda\left\{\frac{x_{it}^{o}\beta + v_{it}^{o}\rho + c_i}{q\sigma}\right\} \tag{4.4}$$

evaluated at the parameter estimate from QMLE, denoted $\boldsymbol{\omega}^{*} = (\theta^{*}, \rho^{*})'$. I expect Equation (4.4) to yield inconsistent estimates of the true marginal effects in Equation (4.3) since $\boldsymbol{\omega}^{*}$ is inconsistent for $\boldsymbol{\omega}_0$.

4.3 FEs and Panel AMEs

Computing AMEs in panel data via Equation (4.3) or (4.4) is difficult because they require a value for the FEs c_i. Estimating parameters with conditional logit does not yield estimates of the FEs. Hence, is not clear how to compute the AMEs from conditional logit in panel data. Hence, *ex post facto* assumptions about the FEs are required to compute AMEs. With this in mind, I first analyse how different *ex post facto* assumptions about FEs contaminate AME estimation and then discuss how heteroskedasticity might further complicate estimation.

4.4 Cross Section Simulations

I provide Monte Carlo simulations of binary outcome cross sectional models to illustrate how misspecified heteroskedaticity might yield inconsistent AME estimates. I compare outcomes across weak ($\mu = 1$) and strong ($\mu = 500$) instruments. I simulate the following model

$$\begin{aligned} x_i &= z_i\xi + v_i \\ y_i^{*} &= x_i\beta + \rho v_i + \varepsilon_i \end{aligned}$$

and compute true AMEs against AMEs estimated from QMLE. Here, $\beta = 0.5$, $\rho \in \{0.20, 0.99\}$, and ξ is set such that $\mu \in \{1, 500\}$. The sample size for each simulation is $n = 1,000$ and I conduct $N = 10,000$ simulations. In each simulation I compute the true AME via Equation (4.3) and the AME obtained from QMLE that assumes homoskedastic errors across individuals via Equation (4.4). The first- and second-stage error terms, v_i and ε_i, are defined in the next subsections depending on the nature of the heteroskedasticity misspecification.

4.4.1 First-Stage Misspecification

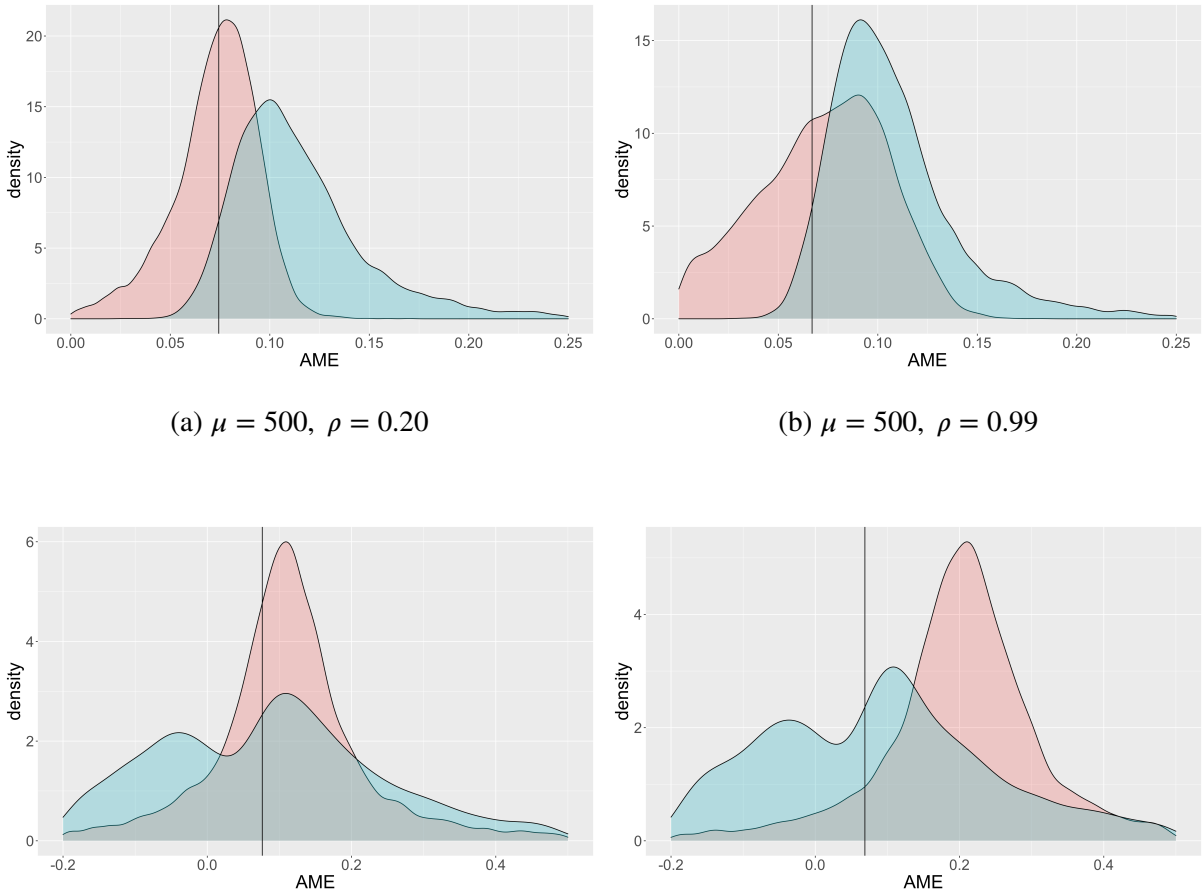
To illustrate the effect of misspecified first-stage heteroskedasticity, suppose that individuals fall into two groups each encompassing 50% of the sample. Generate the first- and second-stage errors as

$$v_i \sim \mathcal{N}\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}\sigma_1^2 & 0\\0 & \sigma_2^2\end{bmatrix}\right), \quad \varepsilon_i \sim \mathcal{L}(0, \sqrt{3}/\pi)$$

where $\sigma_1^2 = 100$ and $\sigma_2^2 = 0.01$ represent the variance of group 1 and group 2, respectively. I generate ε_i from Logistic distribution with location 0 and scale parameter $s = \sqrt{3}/\pi$. I estimate β assuming that v_i are from a standard Normal distribution via LPM (OLS) and logit.

Figure 3 plots AME estimator distributions in the above cross sectional models across weak/strong instruments and low/high endogeneity. Figure 3(a) shows that AMEs estimated from QMLE are

Figure 3: Logit and LPM AMEs with heteroskedastic v_i and homoskedastic ε_i



(a) $\mu = 500,\ \rho = 0.20$

(b) $\mu = 500,\ \rho = 0.99$

(c) $\mu = 1,\ \rho = 0.20$

(d) $\mu = 1,\ \rho = 0.99$

consistent even when first-stage heteroskedasticity is misspecified, particularly when instruments are strong and endogeneity is low. Figure 3(b) reinforces the conclusion that higher endogeneity increases the AME variance as demonstrated by the flattening density for the logit estimator. As demonstrated for parameter estimates, the quasi-LPM estimator is inconsistent about the true AME.

Figure 3(c) and 3(d) illustrate how *weak instruments severely hinder AME point estimation*, particularly in the high endogeneity case where the density's centre is far from the mean AME across the 10,000 simulations. Bimodality emerges for the quasi-LPM AME estimator in the presence of weak instruments. The bimodality is likely a significant component of the distribution rather than simply an artefact of random sampling due to the sharp peaks in density over a large number of simulations.

4.4.2 Second-Stage Misspecification

To illustrate the effect of misspecified second-stage heteroskedasticity, suppose that individuals fall into two groups each encompassing 50% of the sample. Generate the first- and second-stage errors as

$$v_i \sim \mathcal{N}(0,1),\ \varepsilon_i^{(j)} \sim \mathcal{L}(0, q\sigma_j)$$

where $\varepsilon_i^{(j)}$ represents the second-stage error distribution of individual in group $j = 1, 2$ and $q = \sqrt{3}/\pi$. I set $\sigma_1^2 = 100$ and $\sigma_2^2 = 0.01$. I compute the true AME via Equation (4.3) which in the cross sectional two-group case is

$$\bar{\delta} = \frac{\beta}{qn} \sum_{i=1}^{n} \left[\frac{g_i}{\sigma_1} \lambda \left\{ \frac{x_i\beta + v_i\rho}{q\sigma_1} \right\} + \frac{1 - g_i}{\sigma_2} \lambda \left\{ \frac{x_i\beta + v_i\rho}{q\sigma_2} \right\} \right]$$

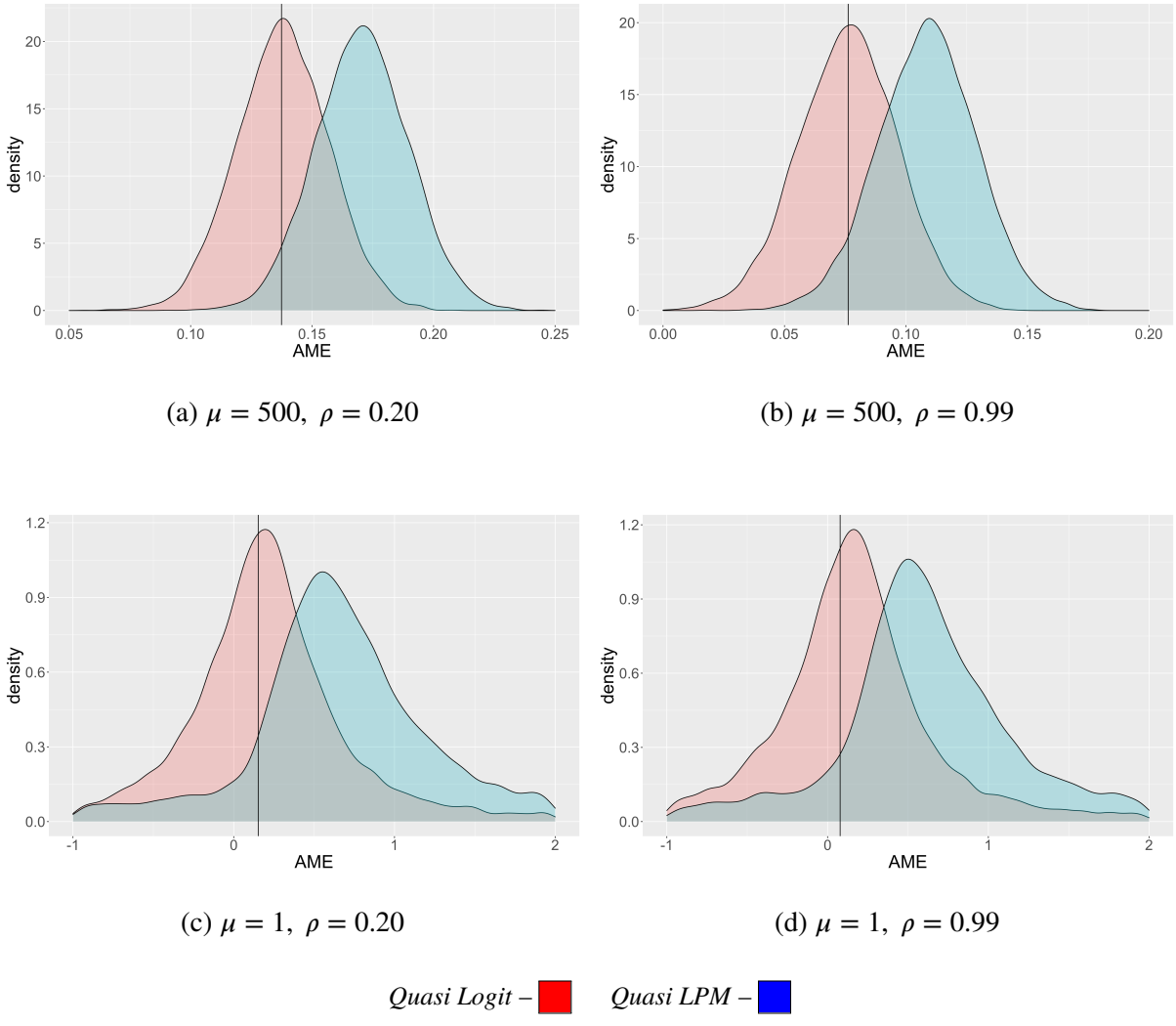
where $g_i = 1$ if individual i is in group 1 and $g_i = 0$ if i is in group 2. I compare the mean true AME over the 10,000 simulations to the AME estimated from QMLE which assumes the scale parameter for all individuals is 1; that is, the AME assuming $\varepsilon_i \sim \mathcal{L}(0, 1)$ for all individuals $i = 1, 2, ..., n$. Figure 4 plots the AME densities over the 10,000 simulations.

For strong instruments, AME estimators are consistent when estimated from quasi-logit in the two-group case. This is a surprising result given that parameter estimators are inconsistent under misspecification as presented in subsection 3.4.2. AMEs estimated from LPMs are inconsistent under the variance misspecification. This questions to practice of estimating causal effects in binary outcome settings with linear models. When instruments are weak, however,

AMEs estimated from both models become inconsistent. Distributions change little between low and high endogeneity when instruments are weak.

That AMEs are consistent under variance misspecification when instruments are strong for the two-group logit model is surprising given that parameter estimators are inconsistent. To investigate the issue further, Appendix B.2 simulates an individual heteroskedasticity model. AMEs estimated from QMLE with misspecified individual heteroskedasicity eventually become inconsistent depending on the variance of h_i, the variable determining individual i's variance. This confirms the intuition that misspecified heteroskedasticity can cause inconsistent AME estimates, and is a significant finding given how pervasive AMEs are in empirical research.

Figure 4: Logit and LPM AMEs with homoskedastic v_i and heteroskedastic ε_i



(a) $\mu = 500,\ \rho = 0.20$

(b) $\mu = 500,\ \rho = 0.99$

(c) $\mu = 1,\ \rho = 0.20$

(d) $\mu = 1,\ \rho = 0.99$

4.5 Panel Simulations

I simulate the model outlined in Section 3.2 to first investigate how *ex post facto* assumptions about FEs might contaminate AME estimation and, second, examine how misspecified heteroskedasticity might further contaminate AME estimates.

The number of individuals is $n = 100$ and the time series dimension varies from $T = 2$ to $T = 10$. I conduct $10,000$ simulations. I implement the four following AME estimation procedures: the conditional logit estimated β and assuming $c_i = 0$ for all individuals (*CL1*), the conditional logit estimated β and substituting dummy variable estimated c_i from unconditional/standard logit (*CL2*), conditional logit estimated β assuming knowledge of the true first-stage residuals, v_{it}, and true FEs c_i (*CL-true*), unconditional/standard logit estimated β with dummy variable estimated c_i (*UCL*), and a linear probability model estimated β with dummy variable estimated c_i (*LPM*). Table 3 summarises each procedure.

I draw the true fixed effects, c_i, from either $\mathcal{U}(-0.5, 0.5)$ or $\mathcal{N}(0, 1)$ and calculate the true AMEs by using the Equation (4.3). I draw from $\mu \in \{1, 500\}$ and the degree of endogeneity $\rho \in \{0.20, 0.99\}$ to simulate weak/strong instruments and low/high endogeneity. The naming conventions and direction of the following analysis are adapted from Stammann et al. (2016), although including the LPM is a novel addition to the literature.

Table 3: Panel Models for AME Estimation

Model	*Regression*	*Fixed Effects*
UCL	y_{it} on x_{it}, $\hat{v}_{it}$, and dummy c_i via standard logit MLE	$\hat{c}_i$ from *UCL*
CL1	y_{it} on x_{it}, $\hat{v}_{it}$ via conditional logit	$c_i = 0$
CL2	y_{it} on x_{it}, $\hat{v}_{it}$ via conditional logit	$\hat{c}_i$ from *UCL*
CL-true	y_{it} on x_{it}, v_{it} via conditional logit	true c_i
LPM	y_{it} on x_{it}, $\hat{v}_{it}$, and dummy c_i via OLS	$\hat{c}_i$ from *LPM*

4.5.1 FEs Assumptions

I assume both first- and second-stage errors are homoskedastic, distributed as $\varepsilon_{it} \sim \mathcal{L}(0, \sqrt{3}/\pi)$ and $v_{it} \sim \mathcal{N}(0, 1)$, respectively. Tables 4 and 5 detail AME bias for strong and weak instruments, respectively.

Table 4 shows bias in the AMEs for each model when instruments are strong. All models

feature a bias in the $T = 2$ panels across both low and high endogeneity and for Uniform and Normal FEs. Particularly surprising is that *CL-true* exhibits a small bias, between 2-5% depending on the level of endogeneity and the FE source. Bias disappears as T increases from 2 to 10 for all models. The *CL2* procedure exhibits the greatest percentage bias across low or high and Uniform or Normal FEs. *CL1*, meanwhile, features a much smaller bias than *CL2* across all panel dimensions and is mostly negligible when FEs are Uniform. However, when FEs are Normal, the bias in *CL2* increases to around 20% and 13% in low and high endogeneity, respectively, and sometimes exceeds the *CL2* bias. The *UCL* and *LPM* procedures performs best in terms of producing the smallest bias.

Table 4: AME Percentage Bias Strong Instruments ($\mu = 500$)

(a) $\rho = 0.20$

	$\mathcal{U}(-.5,.5)$					$\mathcal{N}(0,1)$				
	$CL1$	$CL2$	*CL-true*	UCL	*LPM*	$CL1$	$CL2$	*CL-true*	UCL	*LPM*
$T = 2$	3.82	48.86	2.81	17.41	1.31	21.49	49.57	4.39	18.86	0.74
	(0.004)	(-0.049)	(0.003)	(-0.017)	(0.001)	(0.018)	(-0.043)	(0.004)	(-0.016)	(-0.001)
$T = 4$	1.86	21.98	0.11	1.13	0.36	21.82	23.65	0.94	0.06	0.39
	(0.002)	(-0.023)	(0.000)	(0.000)	(0.000)	(0.019)	(-0.021)	(0.001)	(0.000)	(0.000)
$T = 6$	1.58	14.24	0.01	0.82	0.40	19.73	16.07	0.46	0.07	0.54
	(0.002)	(-0.015)	(0.000)	(0.001)	(0.000)	(0.018)	(-0.015)	(0.000)	(0.000)	(0.001)
$T = 8$	1.45	10.68	0.21	0.35	0.39	21.21	11.89	0.57	0.08	0.35
	(0.002)	(-0.012)	(0.000)	(0.000)	(0.000)	(0.019)	(-0.011)	(0.001)	(0.000)	(0.000)
$T = 10$	1.51	8.61	0.01	0.10	0.18	20.69	9.44	0.38	0.20	0.45
	(0.002)	(-0.009)	(0.000)	(0.000)	(0.000)	(0.019)	(-0.009)	(0.000)	(0.000)	(0.000)

(b) $\rho = 0.99$

	$\mathcal{U}(-.5,.5)$					$\mathcal{N}(0,1)$				
	$CL1$	$CL2$	*CL-true*	UCL	*LPM*	$CL1$	$CL2$	*CL-true*	UCL	*LPM*
$T = 2$	1.84	49.67	1.83	30.49	0.31	13.97	50.84	3.12	31.79	1.17
	(0.002)	(-0.041)	(0.002)	(-0.025)	(0.000)	(0.010)	(0.008)	(0.002)	(-0.058)	(-0.001)
$T = 4$	0.90	19.48	0.14	0.94	0.14	13.15	22.71	1.18	1.18	0.51
	(0.001)	(-0.017)	(0.000)	(0.001)	(0.000)	(0.010)	(-0.017)	(0.001)	(-0.001)	(0.000)
$T = 6$	0.70	11.73	0.14	1.42	0.39	11.46	14.89	0.25	0.02	0.13
	(0.001)	(-0.010)	(0.000)	(0.000)	(0.000)	(0.009)	(-0.011)	(0.000)	(0.000)	(0.000)
$T = 8$	0.59	8.65	0.27	0.60	0.09	12.82	10.60	0.20	0.35	0.28
	(0.001)	(-0.002)	(0.000)	(-0.185)	(0.000)	(0.010)	(-0.008)	(0.000)	(0.000)	(0.000)
$T = 10$	0.25	7.22	0.03	0.15	0.40	12.27	8.50	0.08	0.12	0.80
	(0.000)	(-0.006)	(0.000)	(0.000)	(0.000)	(0.010)	(-0.007)	(0.000)	(0.000)	(-0.001)

Note. Absolute bias in parentheses

Table 5 shows that percentage bias in AMEs significantly increases across all procedures when instruments are weak. Bias is between 140-620% and 29-660% for low and high endogeneity, respectively, for *CL1*. When FEs are Normal, the *CL1* percentage bias generally reduces for all time series dimensions in low and high endogeneity, although on the whole the bias is still between 15-300% and 76-273% for low and high endogeneity, respectively.

Table 5: AME Percentage Bias Weak Instruments ($\mu = 1$)

(a) $\rho = 0.20$

	$\mathcal{U}(-.5,.5)$					$\mathcal{N}(0,1)$				
	$CL1$	$CL2$	$CL\text{-}true$	UCL	LPM	$CL1$	$CL2$	$CL\text{-}true$	UCL	LPM
$T = 2$	620.24	238.23	17.11	676.44	182.02	360.33	197.78	0.30	285.06	320.03
	(0.686)	(0.264)	(0.019)	(0.748)	(-0.202)	(-0.332)	(-0.182)	(0.000)	(-0.263)	(-0.315)
$T = 4$	240.82	206.39	5.22	228.68	230.35	14.84	54.92	5.77	42.82	116.67
	(-0.266)	(-0.228)	(0.006)	(-0.253)	(-0.255)	(-0.014)	(-0.051)	(0.006)	(-0.039)	(-0.116)
$T = 6$	231.10	178.95	3.35	225.25	313.99	153.42	74.81	3.17	107.12	1,814.84
	(0.256)	(0.198)	(-0.004)	(0.249)	(0.349)	(0.142)	(0.069)	(0.003)	(0.099)	(-1.728)
$T = 8$	483.69	454.59	7.41	496.95	1,366.31	300.71	187.14	4.35	226.14	359.35
	(-0.535)	(-0.503)	(0.008)	(-0.550)	(-1.514)	(0.277)	(0.173)	(0.004)	(0.208)	(-0.352)
$T = 10$	142.95	118.02	0.14	141.06	60.99	213.25	147.89	7.28	173.86	94.18
	(0.158)	(0.131)	(0.000)	(0.156)	(-0.067)	(0.197)	(0.136)	(0.007)	(0.160)	(-0.091)

(b) $\rho = 0.99$

	$\mathcal{U}(-.5,.5)$					$\mathcal{N}(0,1)$				
	$CL1$	$CL2$	$CL\text{-}true$	UCL	LPM	$CL1$	$CL2$	$CL\text{-}true$	UCL	LPM
$T = 2$	28.69	305.52	11.10	477.58	334.57	145.80	121.11	2.27	98.03	279.31
	(0.025)	(0.268)	(0.010)	(0.419)	(-0.294)	(-0.114)	(-0.095)	(0.002)	(-0.077)	(-0.228)
$T = 4$	663.80	544.78	2.48	620.82	54.76	76.14	102.28	11.12	89.78	316.05
	(-0.583)	(-0.478)	(0.002)	(-0.545)	(-0.048)	(-0.060)	(-0.080)	(0.009)	(-0.070)	(-0.260)
$T = 6$	222.18	196.11	6.42	203.15	590.05	234.35	160.61	1.43	196.78	607.24
	(-0.195)	(-0.172)	(-0.006)	(-0.178)	(0.519)	(0.184)	(0.126)	(0.001)	(0.154)	(-0.486)
$T = 8$	589.90	580.59	5.31	597.15	2,265.88	199.55	132.96	7.94	159.25	379.65
	(-0.518)	(-0.510)	(0.005)	(-0.524)	(1.991)	(0.156)	(0.104)	(0.007)	(0.125)	(-0.309)
$T = 10$	268.49	237.34	2.08	269.09	8.77	273.13	246.12	5.94	258.08	286.42
	(0.236)	(0.208)	(0.002)	(0.236)	(0.008)	(-0.214)	(-0.193)	(0.005)	(-0.202)	(-0.231)

Note. Absolute bias in parentheses

This is a crucial finding because the *CL1* procedure represents a realistic option for most empirical applications using conditional logit. *CL1* computed AMEs can be either upwards or downwards biased depending on the time series dimension and so no general rule for contextualising AMEs in applied work emerges from this analysis. The control function procedure in this case fails to generate consistent AMEs given the weak instruments and lack of knowledge

about the FEs, representing a significant empirical shortcoming of the CL with control function in weak instrument environments.

Similar, but generally smaller, percentage bias levels are achieved in the *CL2* and *UCL* procedures when instruments are weak. This illustrates the lack of a good alternative to simply specifying the FEs as 0 as done in *CL1*. This is reinforced by the LPM bias, which can be up to 2300%. Again, the control function approach to removing endogeneity does not produce consistent AMEs when instruments are weak. This is a particularly important finding overall given AMEs are usually the primary object of interest in empirical evaluations of policies or treatments in economics research.

Another interesting finding is the percentage bias for the *CL-true* specification, which assumes full knowledge of true first-stage errors v_{it} and the true FEs. Expectedly, the bias for *CL-true* is much smaller across all time dimensions for low or high endogeneity and Uniform or Normal generated FEs. However, the bias fluctuates rather dramatically from approximately 0% to about 12% in the Uniform FEs case. Depending on the setting, percentage bias on the order of 12% may be significant. This demonstrates a novel finding for the binary outcome panel weak instruments setting; that is, *the control function procedure is unable to produce completely reliable estimates of AMEs when instruments are weak even in the impossible scenario of perfect knowledge.*

AME bias fails to converge to 0 for the models as the time series dimension grows. These results demonstrate that weak instruments can make AME estimates inconsistent for realistic empirical models even when errors are homoskedastic in both stages. This analysis ultimately raises questions about proper empirical conduct in the binary outcomepanel data case and how to interpret econometric analyses relying on AMEs as the primary interpretation tool, particularly when instruments are weak.

4.5.2 Heteroskedasticity

I expect misspecified heteroskedasticity to contaminate AME estimates in panel data in addition to *ex post facto* assumptions about FEs as parameter estimators are inconsistent. This expectation is also motivated by the individual heteroskedasticity cross section simulations in Appendix B.2, where AMEs estimated from QMLE were inconsistent. Whether inconsistency in AMEs follows through from inconsistent parameters is significant since AMEs are the primary interpretation and recommendation tool in applied work.

To test these expecations, I simulate the two-group heteroskedasticity model from subsection 3.2 $N = 10,000$ times and compute true AMEs from Equation (4.3) and AMEs estimated from QMLE that assumes homoskedastic errors when they are heteroskedastic. As in subsection 3.2, $\beta = 0.5$, $\rho = 0.20$ or 0.99, and ξ is set such that $\mu = 500$ (strong instruments) and $\mu = 1$ (weak instruments). I estimate the models summarised in Table 3 to investigate how misspecified heteroskedasticity might *further* contaminate AME estimates beyond incorrect *ex post facto* assumptions about FEs. The results of these simulations are summarised in Table 6 and 7 for strong and weak instruments, respectively, for $n = 100$ individuals each observed over $T = 10$ periods.

Table 6: AME Bias Strong Instruments Heteroskedasticity ($n = 100$, $T = 10$)

(a) $\rho = 0.20$

	$\mathcal{U}(-.5,.5)$					$\mathcal{N}(0,1)$				
Het. Source	$CL1$	$CL2$	CL-*true*	UCL	*LPM*	$CL1$	$CL2$	CL-*true*	UCL	*LPM*
None	1.51	8.61	0.01	0.10	0.18	20.69	9.44	0.38	0.20	0.45
	(0.002)	(-0.009)	(0.000)	(0.000)	(0.000)	(0.019)	(-0.009)	(0.000)	(0.000)	(0.000)
ε_{it}	3.19	9.69	1.76	1.32	0.64	21.25	14.07	5.21	4.90	3.70
	(0.004)	(-0.012)	(0.002)	(-0.002)	(0.001)	(0.019)	(-0.012)	(0.005)	(-0.004)	(-0.003)
v_{it}	2.08	1.60*	0.43	0.70	3.17	13.63*	16.36	0.40	16.63	14.91
	(0.001)	(0.000)	(0.000)	(0.000)	(-0.001)	(-0.004)	(-0.004)	(0.000)	(-0.005)	(-0.004)

(b) $\rho = 0.99$

Het. Source	$CL1$	$CL2$	CL-*true*	UCL	*LPM*	$CL1$	$CL2$	CL-*true*	UCL	*LPM*
None	0.25	7.22	0.03	0.15	0.40	12.27	8.50	0.08	0.12	0.80
	(0.000)	(-0.006)	(0.000)	(0.000)	(0.000)	(0.010)	(-0.007)	(0.000)	(0.000)	(-0.001)
ε_{it}	0.39	8.75	0.55	0.99	0.81	7.81*	10.97	3.95	2.40	0.67*
	(0.000)	(-0.007)	(0.000)	(-0.001)	(0.001)	(0.005)	(-0.007)	(-0.003)	(-0.002)	(0.000)
v_{it}	17.59	21.59	1.01	34.03	25.09	14.65	13.64	0.93	33.39	37.89
	(-0.002)	(-0.003)	(0.000)	(-0.005)	(-0.003)	(-0.002)	(-0.002)	(0.000)	(-0.004)	(-0.005)

Note. The above table describes percentage and absolute bias in AME calculation under different error variance misspecifications. The *None* row documents bias when both stages are homoskedastic, the ε_{it} row when the second-stage errors are heteroskedastic, and the v_{it} row when the first-stage errors are heteroskedastic. Absolute bias reported in parentheses.

Table 6 shows that when instruments are strong, misspecified heteroskedasticity in either the first- or second-stage *generally* increases AME bias across all models from when both stages are correctly specified as homoskedastic. Exceptions are highlighted in the table with an asterisk. In addition, no clear pattern holds for whether a misspecified v_{it} or ε_{it} is worse for AME

estimates. For the most part, a misspecified v_{it} yields less biased AMEs than ε_{it} misspecification when $\rho = 0.20$ and FEs are drawn from $\mathcal{U}(-0.5, 0.5)$. The opposite occurs for $\rho = 0.99$ across both FE source distributions, which is intuitive given that first-stage misspecification yields misshapen distributions while second-stage misspecification retains a fairly smooth and Normal shape, shown in subsection 4.4.

An interesting pattern emerges for AMEs estimated from the *CL-true* procedure, which assumes full knowledge of true first-stage errors and true FEs. The *CL-true* bias increases when ε_{it} is misspecified across low and high endogeneity regardless of the FE source distribution. This is because the only estimated object in the *CL-true* AME is the parameter estimate, which is inconsistent with misspecified ε_{it} and consistent with misspecified v_{it}. Hence, misspecified heteroskedasticity clearly contaminates AMEs even when instruments are strong.

Table 7 shows that when instruments are weak, misspecified heteroskedasticity in either the first- or second-stage *generally* increases AME bias from the correctly specified homososkedastic case, as before. For example, AME percentage bias in misspecified LPM procedures increases by approximately 25 times when endogeneity is low and by up to 200 times when endogeneity is high.

The exceptions to this, however, are more numerous than when instruments were strong. Indeed, sometimes AME bias is much smaller with misspecified errors. For example, when v_{it} is heteroskedastic, $\rho = 0.20$, and FEs are drawn from $\mathcal{N}(0, 1)$, the AME percentage bias decreases by between 150-210 percentage points for the *CL1*, *CL2*, and *UCL* procedures, while the *CL-true* bias decreases from 7% to 4%. The bias from LPM procedures decreases by about 70 and 160 percentage points for low and high endoegeneity, resepectively. Furthermore, *CL-true* bias decreases across both low and high endogeneity regimes when FEs are drawn from $\mathcal{N}(0, 1)$ while it increases for when FEs are drawn from $\mathcal{U}(-0.5, 0.5)$. This is surprising and contradicts the findings for when instruments are strong.

However, on average, the AME percentage bias increases on average by a factor of 2-20 across all models when there is variance misspecification, although there are exceptions to this. It might be too hasty to say that misspecified heteroskedasticity is *always* a problem for AME estimates, although this seems approximately true. Given the extremely high levels of bias across almost estimated models, however, these results raise fundamental questions about how reliable econometric analyses using AMEs to interpret results and make recommendations in the binary outcome panel data environment truly are, particularly when instruments are weak.

Table 7: AME Bias Weak Instruments Heteroskedasticity ($n = 100$, $T = 10$)

(a) $\rho = 0.20$

	$\mathcal{U}(-.5, .5)$					$\mathcal{N}(0, 1)$				
Het. Source	$CL1$	$CL2$	CL-*true*	UCL	*LPM*	$CL1$	$CL2$	CL-*true*	UCL	*LPM*
None	142.95	118.02	0.14	141.06	60.99	213.25	147.89	7.28	173.86	94.18
	(0.158)	(0.131)	(0.000)	(0.156)	(-0.067)	(0.197)	(0.136)	(0.007)	(0.160)	(-0.091)
ε_{it}	61.44*	38.95*	3.62	52.12*	1,685.52	725.60	510.12	4.77*	556.78	493.11
	(0.086)	(0.054)	(-0.005)	(0.073)	(-2.357)	(-0.576)	(-0.405)	(0.004)	(-0.442)	(-0.392)
v_{it}	218.35	214.84	9.96	218.47	1,140.71	3.77*	4.54*	4.74*	2.98*	21.56*
	(0.060)	(0.059)	(-0.030)	(0.060)	(-0.315)	(0.001)	(-0.001)	(-0.001)	(0.001)	(-0.006)

(b) $\rho = 0.99$

Het. Source	$CL1$	$CL2$	CL-*true*	UCL	*LPM*	$CL1$	$CL2$	CL-*true*	UCL	*LPM*
None	268.49	237.34	2.08	269.09	8.77	273.13	246.12	5.94	258.08	286.42
	(0.236)	(0.208)	(0.002)	(0.236)	(0.008)	(-0.214)	(-0.193)	(0.005)	(-0.202)	(-0.231)
ε_{it}	297.94	259.95	11.24	289.10	1,833.80	792.12	644.58	7.14*	695.13	602.20
	(0.228)	(0.199)	(-0.009)	(0.221)	(-1.405)	(-0.499)	(-0.406)	(-0.004)	(-0.438)	(-0.379)
v_{it}	2,931.35	3,019.33	17.84	3,439.91	208.96	298.28	280.90	3.66*	188.05*	111.03*
	(-0.390)	(-0.401)	(-0.002)	(-0.457)	(-0.028)	(0.040)	(0.037)	(0.000)	(0.025)	(0.015)

Note. The above table describes percentage and absolute bias in AME calculation under different error variance misspecifications. The *None* row documents bias when both stages are homoskedastic, the ε_{it} row when the second-stage errors are heteroskedastic, and the v_{it} row when the first-stage errors are heteroskedastic. Absolute bias reported in parentheses.

5 Empirical Application

I illustrate the AR test by estimating confidence intervals for the central specifications in Nunn and Qian (2014), who attempt to quantify the causal effect of food aid on the probability of civil conflict in the next period. My focus is on their model for civil conflict incidence, described in Equation (5.1) and (5.2). Confidence intervals are constructed as discussed in subsection 2.6.

The full sample contains data on 125 non-OECD countries over the period $1971 - 2006$. An observation is a country-year pair, and there are 4089 observations. Conflict data are sourced from the Uppsala Conflict Data Program/Peace Research Institute Oslo (UCDP/PRIO) Armed Conflict Dataset Version 4-2010 while the US food aid measure data are sourced from the Food and Agriculture Organization's (FAO) FAOSTAT database.

Countries are classified into their respective World Bank regions: South Asia, Sub-Saharan Africa, Europe and Central Asia, Middle East and North Africa, Latin America and Caribbean, and East Asia and Pacific. Conflicts are defined as the use of armed force between two parties

causing at least 25 battle deaths in a year and are categorised as intrastate (civil wars), interstate, extrasystemic, and internationalised [1]. Food aid is measured by volume of US wheat aid in metric tonnes.

The IV model is

$$C_{irt} = \beta F_{irt} + \mathbf{X}_{irt}\Gamma + \varphi_{rt} + \psi_{ir} + e_{irt} \tag{5.1}$$

$$F_{irt} = \alpha(P_{t-1} \times \bar{D}_{ir}) + \mathbf{X}_{irt}\Gamma + \varphi_{rt} + \psi_{ir} + \upsilon_{irt} \tag{5.2}$$

for country i in region r in time t. In the second-stage, C_{irt} is the existence of civil conflict (0 or 1), F_{irt} is food-aid, β is the parameter of interest, $\mathbf{X}_{irt}$ is the matrix of controls (e.g. GDP in the US, temperature and precipitation, Democrat US president), φ_{rt}, ψ_{ir} are fixed effects and e_{irt} is an error term. $P_{t-1} \times \bar{D}_{ir}$ is the instrument, where $\bar{D}_{ir}$ is the probability of receiving US wheat aid and P_{t-1} is US wheat production in the previous year. Summary statistics are provided in Appendix B.3.

The authors estimate cluster robust standard errors at the country level. Following the discussion in Section 3 and 4, I assume that second-stage errors e_{irt} are conditionally homoskedastic and have control function $e_{irt} = \rho\upsilon_{irt} + \varepsilon_{irt}$. This means that I assume ε_{irt} is homoskedastic across individuals while υ_{irt} is clustered at the country level. Hence, to estimate the AR intervals from Equation (2.6) and (2.7), I compute the cluster robust variance of the instrument and the standard variance of the reduced-form parameters.

5.1 Incidence Model

Nunn and Qian (2014)'s main goal is to quantify the causal effect of increased food aid on civil conflict incidence. I focus on the three following baseline specifications captured by Equation (5.1) and (5.2). (1) *Country FE*, whose only control variables are country and region-year level FEs; (2) *Most Controls*, which includes a range of economic (real US GDP per capita, oil price, Avg. US economic aid), political (Avg. US military aid, US Democratic President), and weather (monthly recipient temperature, weather, and precipitation) control variables; and (3) *Full Controls*, which includes the whole suite of controls including country, year, and region fixed effects.

I chose these specifications for two reasons. Firstly, these specifications demonstrate the authors central conclusion that increased food aid increases civil conflict incidence; and secondly,

[1] UCDP/PRIO (2010) defines extrasysemic conflicts as "between a state and a non-state group outside its own territory" and internationalised conflicts as"between the government of a state and one or more internal opposition group(s) with intervention from other states... on one or both sides".

the first-stage F-statistics vary between strong (> 10) and weak (< 10), adding to the analysis of standard weak instrument detection practice. Parameter estimates are available in Appendix B.4. The cluster robust interval is a t-test where standard errors are clustered at the country level. Note that the second-stage is estimated via a LPM.

Table 8 shows that the AR confidence intervals are wider than those obtained from standard inference. Generally, the increased width comes from the upper bound. Of particular note is the 99% *Country FE* AR interval which has an unbounded upper bound and is indicative of extremely weak instruments. The 95% AR intervals are approximately 1.5-3 times wider than CR and 2-6 times wider than standard t-test intervals. Similar results hold for the 99% intervals, where the AR intervals are about twice as wide as CR intervals and 3 times as wide as the t-test intervals. The significantly greater width in all confidence intervals justifies the test's importance for evaluating applied research in binary outcome panel data environments.

Estimated with standard errors clustered at the country level in the second-stage, the cluster robust (CR) intervals assume that second-stage errors are heteroskedastic. However, parameter estimates can be inconsistent when this assumption is made as shown in section 3.

Table 8: 95% and 99% Confidence Intervals – Incidence Specification (LPM)

Interval	*Statistic*	*Country FE*	*Most Controls*	*Full Controls*
95%	t-test	$(2.09, 5.19)$	$(2.07, 4.80)$	$(1.71, 4.27)$
	CR	$(0.23, 7.05)$	$(1.36, 5.51)$	$(1.11, 4.87)$
	AR	$(1.59, 20.03)$	$(1.74, 8.17)$	$(1.43, 7.45)$
99%	t-test	$(1.68, 5.68)$	$(1.64, 5.23)$	$(1.30, 4.67)$
	CR	$(-0.84, 8.13)$	$(0.70, 6.16)$	$(0.51, 5.46)$
	AR	$(-\infty, \infty)$	$(1.36, 13.16)$	$(1.08, 12.47)$
	F-statistic	5.84	12.76	12.10

Note. All bounds in the confidence intervals are multiplied by 1000 to ease presentation.

6 Conclusion

I make three distinct contributions to the inference and estimation of causal effects from binary outcome panel data. Firstly, by implementing an estimation procedure which does not estimate FEs in the second-stage, I proposed a test which performs inference on the distance between the consistently estimated structural- and reduced-form parameters. The test has the correct size regardless of instrument strength and under different error distributions in simulated panel data, while standard Wald tests can severely over-reject the true null hypothesis.

I applied the test to the central specifications in Nunn and Qian (2014) who argue that increased food aid causes a statistically significant increase in the probability of civil conflict in a panel of 125 non-OECD countries. Confidence intervals estimated by the robust test are up to 6 times as wide as those estimated by standard methods. The significant extra width in the robust confidence intervals illustrate the importance of the test in empirical work when there is doubt about instrument strength.

Secondly, I proved that quasi-maximum likelihood estimation (QMLE) that disregards heteroskedasticity can produce inconsistent parameter estimates. Recall the control function formulation of second-stage errors $u_{it} = \rho v_{it} + \varepsilon_{it}$, where v_{it} and ε_{it} are the idiosyncratic first- and second-stage error components and ρ measures the degree of endogeneity. Parameter estimators are inconsistent if a heteroskedastic ε_{it} is misspecified as homoskedastic. However, parameter estimators remain consistent when a heteroskedastic v_{it} is misspecified as homoskedastic. My conclusions question the estimation of, and inference about, causal relationships estimated from binary outcome panel data with with linear and nonlinear IV models.

Thirdly, average marginal effects (AMEs) estimated from QMLE that ignores heteroskedasticity can also be inconsistent, even when instruments are strong. For example, simulations of a two-group heteroskedasticity cross sectional model showed that AMEs estimated from QMLE are consistent. However, AMEs can be inconsistent when individual heteroskedasticity is ignored.

As the proposed conditional logit procedure does not estimate FEs, I investigated the accuracy of AMEs estimated from panel data under different assumptions about the FEs. I found that QMLE ignoring error heteroskedasticity contaminates AME estimates *beyond* these *ex post facto* assumptions about FEs. Generally, percentage bias increases by a factor of between 2-20 regardless of instrument strength, although there are exceptions that cannot be explained by the presented analysis. Conditional logit estimated AMEs that assume perfect knowledge of true

FEs and first-stage errors can still be biased by up to 12% when instruments are weak. Under different assumptions about the FEs, conditional logit estimated AMEs can be up to 660% biased in the presence of weak instruments. AMEs computed from standard logit that controls for FEs with dummy variables is between 100-670% biased depending on the FE source distribution and the panel dimension. I also found that AMEs estimated from LPMs, a popular estimation method in applied research, can be up to 2,300% biased in the presence of weak instruments. These results raise fundamental questions about how to interpret empirical analyses that report AMEs when there is doubt about instrument strength.

There are two immediate extensions of this research. Firstly, as I only explored LPMs, conditional logit, and logit, future research must evaluate if similar levels of bias in AMEs feature in other estimation models such as probit or tobit. This is the first step in possibly providing alternative estimation procedures that reduce bias and improve the credibility of applied research. At the very least, studying other nonlinear estimation procedures puts the highlighted estimation issues into perspective given the broad range of models used in modern empirical research. The second extension of this research is further analysing how misspecified heteroskedasticity complicates AME estimation, particularly in panel models. Indeed, studying the conditions when ignoring heteroskedasticity actually *improves* AME estimation is greatly needed, given the results in subsection 4.5.2.

It is my hope that the presented results provide both a better appreciation of and the tools to help guard against the weak instrument problem in binary outcome panel data. Recognising the ways in which the econometric analysis of binary outcome panel data can fail is critical for economics going forward to support its claim as an evidence based social science.

References

Anderson, T. W., & Rubin, H. (1949). Estimation of the Parameters of a Single Equation in a Complete System of Stochastic Equations. *Annals of Mathematical Statistics*, *20*(1), 46 –63.

Andrews, I., Stock, J. H., & Sun, L. (2019). Weak Instruments in Instrumental Variables Regression: Theory and Practice. *Annual Review of Economics*, *11*(1), 727–753.

Angrist, J., & Pischke, J.-S. (2010). *The Credibility Revolution in Empirical Economics: How Better Research Design is Taking the Con out of Econometrics* (Working Paper No. 15794). National Bureau of Economic Research.

Fernández, M., Ibáñez, A. M., & Peña, X. (2014). Adjusting the Labour Supply to Mitigate Violent Shocks: Evidence from Rural Colombia. *Journal of Development Studies*, *50*(8), 1135–1155.

Frazier, D. T., Renault, E., Zhang, L., & Zhao, X. (2021). Weak Identification in Discrete Choice Models.

Frijters, P., Johnston, D. W., Shah, M., & Shields, M. A. (2009). To Work or Not to Work? Child Development and Maternal Labor Supply. *American Economic Journal: Applied Economics*, *1*(3), 97–110.

Greene, W. H. (2003). *Econometric Analysis* (Fifth). Pearson Education.

Harvey, A. C. (1976). Estimating Regression Models with Multiplicative Heteroscedasticity. *Econometrica*, *44*(3), 461–465.

Lancaster, T. (2000). The Incidental Parameter Problem Since 1948. *Journal of Econometrics*, *95*(2), 391–413.

Magnusson, L. M. (2010). Inference in Limited Dependent Variable Models Robust to Weak Identification. *The Econometrics Journal*, *13*(3), S56–S79.

Martin, V., Hurn, S., & Harris, D. (2012). *Econometric Modelling with Time Series: Specification, Estimation and Testing*. Cambridge University Press.

Miguel, E., Satyanath, S., & Sergenti, E. (2004). Economic Shocks and Civil Conflict: An Instrumental Variables Approach. *Journal of Political Economy*, *112*(4), 725–753.

Nunn, N., & Qian, N. (2014). US Food Aid and Civil Conflict. *American Economic Review*, *104*(6), 1630–1666.

Olea, J. L. M., & Pflueger, C. (2013). A Robust Test for Weak Instruments. *Journal of Business & Economic Statistics*, *31*(3), 358–369.

Staiger, D., & Stock, J. H. (1997). Instrumental Variables Regression with Weak Instruments. *Econometrica*, *65*(3), 557–586.

Stammann, A., Heiß, F., & McFadden, D. Estimating Fixed Effects Logit Models with Large Panel Data. eng. In: Beiträge zur Jahrestagung des Vereins für Socialpolitik 2016: Demographischer Wandel - Session: Microeconometrics (G01-V3). Kiel und Hamburg: ZBW - Deutsche Zentralbibliothek für Wirtschaftswissenschaften, Leibniz-Informationszentrum Wirtschaft, 2016.

Stock, J., & Yogo, M. (2005). Testing for Weak Instruments in Linear IV Regression. In D. W. Andrews (Ed.), *Identification and Inference for Econometric Models* (pp. 80–108). Cambridge University Press.

UCDP/PRIO. (2010). *UCDP/PRIO Armed Conflict Dataset Codebook Version 4-2010* (Codebook). International Peace Research Institute, Oslo (PRIO).

Wooldridge, J. M. (2010). *Econometric Analysis of Cross Section and Panel Data* (Vol. 1). The MIT Press.

A Appendix A

A.1 Conditional Logit

The conditional logit (CL) is used to estimate the second-stage in Equation (2.2) and (2.4) precisely because it does not estimate the FEs. The main benefit of using CL avoids the incidental parameters problem contaminating standard logit estimated parameters. Consider a panel of i individuals each observed for T periods, where $\mathbb{P}(y_{it} = 1|\mathbf{x}_{it}, \alpha_i) = \Lambda(\mathbf{x}_{it}\boldsymbol{\beta} + \alpha_i)$, where $\Lambda(\cdot)$ is the Logistic function. I condition on $n_i \equiv \sum_{i=1}^{T} y_{it}$. Let $\mathbf{x}_i = (x_{i1}, ..., x_{iT})'$ and

$$y_{it} = \begin{cases} 1 & y_{it}^* > 0 \\ 0 & y_{it}^* \leq 0 \end{cases}$$

Individual i's contribution to the likelihood is then

$$\begin{aligned} \mathbb{P}(y_{i1} = y_1, ..., y_{iT}|\mathbf{x}_i, \alpha_i, n_i = n) &= \frac{\mathbb{P}(y_{i1} = y_1, ..., y_{iT} = y_T|\mathbf{x}_i, \alpha_i)}{\mathbb{P}(n_i = n|\mathbf{x}_i, \alpha_i)} \\ &= \frac{\prod_{t=1}^{T} \mathbb{P}(y_{it} = y_t|\mathbf{x}_i, \alpha_i)}{\mathbb{P}(n_i = n|\mathbf{x}_i, \alpha_i)} \end{aligned}$$

Assumption (2.1) and (2.3) in subsection 2.2 yield $\mathbb{P}(y_{it} = y_t|\mathbf{x}_i, \alpha_i) = \Lambda(\mathbf{x}_i\beta + \alpha_i)$ where $y_t \in \{0, 1\}$. The numerator then has form

$$\begin{aligned} \prod_{t=1}^{T} \mathbb{P}(y_{it} = y_t|\mathbf{x}_i, \alpha_i) &= \prod_{t=1}^{T} \Lambda(\mathbf{x}_{it}\beta + \alpha_i)^{y_t}[1 - \Lambda(\mathbf{x}_{it}\beta + \alpha_i)]^{1-y_t} \\ &= \prod_{t=1}^{T} \left\{ \frac{\exp(\mathbf{x}_{it}\beta + \alpha_i)}{1 + \exp(\mathbf{x}_{it}\beta + \alpha_i)} \right\}^{y_t} \left\{ 1 - \frac{\exp(\mathbf{x}_{it}\beta + \alpha_i)}{1 + \exp(\mathbf{x}_{it}\beta + \alpha_i)} \right\}^{1-y_t}, \text{ or simply} \\ &= \frac{\prod_{t=1}^{T} \exp(\mathbf{x}_{it}\beta + \alpha_i)^{y_t}}{\prod_{t=1}^{T}(1 + \exp(\mathbf{x}_{it}\beta + \alpha_i))} \end{aligned}$$

The product in the numerator will only multiply those time series observations such that $y_t = 1$. Define $\mathbf{y}_i = (y_{i1}, y_{i2}, ..., y_{iT})'$ and $\boldsymbol{\alpha}_i = (\alpha_i, ..., \alpha_i)'$. Hence, I express

$$\prod_{t=1}^{T} \exp(\mathbf{x}_{it}\boldsymbol{\beta} + \alpha_i)^{y_t} = \exp\Big(\sum_{t=1}^{T} [\mathbf{x}_{it}\boldsymbol{\beta} + \alpha_i] \Big) = \exp\left(\mathbf{y}_i[\mathbf{x}_i\boldsymbol{\beta} + \boldsymbol{\alpha}] \right)$$

To simplify expression, let

$$D(\boldsymbol{\beta}, \alpha_i) = \prod_{t=1}^{T} [1 + \exp(\mathbf{x}_{it}\boldsymbol{\beta} + \alpha_i)]$$

Then

$$\prod_{t=1}^{T} \mathbb{P}(y_{it} = y_t|\mathbf{x}_i, \alpha_i) = \frac{\exp(\mathbf{y}_i'(\mathbf{x}_i\boldsymbol{\beta} + \boldsymbol{\alpha}_i))}{D(\boldsymbol{\beta}, \alpha_i)} = \frac{\exp(\mathbf{y}_i'\mathbf{x}_i\boldsymbol{\beta} + n_i\alpha_i)}{D(\boldsymbol{\beta}, \alpha_i)} \tag{A.1}$$

The denominator, $\mathbb{P}(n_i = n|\mathbf{x}_i, \alpha_i)$ is the disjoint union of all distinct strings of 0s and 1s of length T. Let R_i be the subset of the T-dimensional Euclidean space $\mathbb{R}^T$ with elements $\mathbf{a}$ such that $a_t \in \{0, 1\}$ and $\sum_{t=1}^{T} a_t = n_i$. Simply, $\mathbf{a}$ is a T-dimensional vector of zeroes and ones where there are n_i ones. Then

$$\mathbb{P}(n_i = n|\mathbf{x}_i, \alpha_i) = \frac{\sum_{\mathbf{a}\in R_i} \exp(\mathbf{a}'(\mathbf{x}_i\boldsymbol{\beta} + \boldsymbol{\alpha}_i))}{D(\boldsymbol{\beta}, \alpha_i)} = \frac{\sum_{\mathbf{a}\in R_i} \exp(\mathbf{a}'\mathbf{x}_i\boldsymbol{\beta} + n_i\alpha_i)}{D(\boldsymbol{\beta}, \alpha_i)} \tag{A.2}$$

Taking the ratio of (A.1) and (A.2)

$$\begin{aligned}
\frac{\prod_{t=1}^{T} \mathbb{P}(y_{it} = y_t|\mathbf{x}_i, \alpha_i)}{\mathbb{P}(n_i = n|\mathbf{x}_i, \alpha_i)} &= \frac{\frac{\exp(\mathbf{y}_i'\mathbf{x}_i\boldsymbol{\beta}+n_i\alpha_i)}{D(\boldsymbol{\beta},\alpha_i)}}{\frac{\sum_{\mathbf{a}\in R_i} \exp(\mathbf{a}'(\mathbf{x}_i\boldsymbol{\beta})+n_i\alpha_i)}{D(\boldsymbol{\beta},\alpha_i)}} \\
&= \frac{\exp(n_i\alpha_i)\exp(\mathbf{y}_i'\mathbf{x}_i\boldsymbol{\beta})}{\exp(n_i\alpha_i)\sum_{\mathbf{a}\in R_i}\exp(\mathbf{a}'\mathbf{x}_i\boldsymbol{\beta})} \\
&= \frac{\exp(\mathbf{y}_i'\mathbf{x}_i\boldsymbol{\beta})}{\sum_{\mathbf{a}\in R_i}\exp(\mathbf{a}'\mathbf{x}_i\boldsymbol{\beta})}
\end{aligned}$$

So the contribution of individual i to the overall log-likelihood is

$$\ell_i = \log\left\{ \exp(\mathbf{y}_i'\mathbf{x}_i\boldsymbol{\beta})\Big[\sum_{\mathbf{a}\in R_i} \exp(\mathbf{a}'\mathbf{x}_i\boldsymbol{\beta})\Big]^{-1}\right\}$$

where

$$R_i = \{\mathbf{a} \in \mathbb{R}^T : a_t \in \{0, 1\} \text{ and } \sum_{t=1}^{T} a_t = n_i\}$$

A.2 The Test

Consider the following reduced-form representation of Equation (2.1) and (2.2) with included control variables w_{it}

$$\begin{aligned}
y_{it}^* &= z_{it}\delta_z + w_{it}\delta_w + v_{it}\delta_v + c_i + \varepsilon_{it} \\
x_{it} &= z_{it}\pi_z + w_{it}\pi_w + b_i + v_{it}
\end{aligned}$$

Define $\mathbf{z} = (z, w)$, $\boldsymbol{\delta} = (\delta_z', \delta_w', \delta_v')'$, $\boldsymbol{\pi} = (\pi_z, \pi_w)$, $\boldsymbol{\delta}_\mathbf{z} = (\delta_z', \delta_w')'$, and $\boldsymbol{\theta} = (\boldsymbol{\delta}'; \boldsymbol{\pi}')'$. This representation encompasses both the panel and cross section case. To focus the discussion on the panel case, I model the likelihood with a conditional logit first-stage and a linear second-stage.

The conditional logit is justified as fixed effects are not estimated, as demonstrated previously. As I am estimating a two-step regression, individual i's contribution to the log-likelihood has two components. The first, deriving from the first-stage which regresses the endogenous variables on the instrument, comes from a linear regression likelihood and is from the Normal distribution. The second component, deriving from the control function second-step is from the conditional logit procedure described above. Hence, the whole two-step procedure has log-likelihood $\ell_i(\boldsymbol{\theta}) = \left(\ell_i^{(2)}(\boldsymbol{\delta})', \ell_i^{(1)}(\boldsymbol{\pi})'\right)'$ with score function $\frac{\partial \ell_i(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \mathbf{g}_i(\boldsymbol{\delta};\boldsymbol{\pi}) = \left(\mathbf{s}_i(\boldsymbol{\delta};\hat{\boldsymbol{\pi}})', \mathbf{r}_i(\boldsymbol{\pi})'\right)'$, where $\ell^{(j)}$ is the log-likelihood of the j^{th} step. Note, $\mathbf{s}_i(\boldsymbol{\delta};\hat{\boldsymbol{\pi}})$ is the score of the second-stage and $\mathbf{r}_i(\boldsymbol{\pi})$ is the score of the first-stage. First, I simplify $\mathbf{s}_i(\boldsymbol{\delta})$ as follows

$$\begin{aligned}\mathbf{s}_i(\boldsymbol{\delta};\hat{\boldsymbol{\pi}}) = \nabla_{\boldsymbol{\delta}}\ell_i^{(2)}(\boldsymbol{\delta})' &= \frac{\partial}{\partial \boldsymbol{\delta}}\left[\log\left\{\exp(y_i'(\mathbf{z}_i\boldsymbol{\delta}_{\mathbf{z}} + v_i\delta_v))\Big[\sum_{\mathbf{a}\in R_i}\exp(\mathbf{a}'(\mathbf{z}_i\boldsymbol{\delta}_{\mathbf{z}} + v_i\delta_v))\Big]^{-1}\right\}\right]' \\ &= \frac{\partial}{\partial \boldsymbol{\delta}}\left[y_i'(\mathbf{z}_i\boldsymbol{\delta}_{\mathbf{z}} + v\delta_v) - \log\Big\{\sum_{\mathbf{a}\in R_i}\exp(\mathbf{a}'(\mathbf{z}_i\boldsymbol{\delta}_{\mathbf{z}} + v_i\delta_v))\Big\}\right]'\end{aligned}$$

and differentiation provides

$$\mathbf{s}_i(\boldsymbol{\delta};\hat{\boldsymbol{\pi}}) = \nabla_{\boldsymbol{\delta}}\ell_i^{(2)}(\boldsymbol{\delta})' = \begin{bmatrix}\mathbf{z}_i'\left(y_i - \dfrac{\sum_{\mathbf{a}\in R_i}\mathbf{a}\exp(\mathbf{a}'(\mathbf{z}_i\boldsymbol{\delta}_{\mathbf{z}} + v_i\delta_v))}{\sum_{\mathbf{a}\in R_i}\exp(\mathbf{a}'(\mathbf{z}_i\boldsymbol{\delta}_{\mathbf{z}} + v_i\delta_v))}\right) \\ \\ v_i'\left(y_i - \dfrac{\sum_{\mathbf{a}\in R_i}\mathbf{a}\exp(\mathbf{a}'(\mathbf{z}_i\boldsymbol{\delta}_{\mathbf{z}} + v_i\delta_v))}{\sum_{\mathbf{a}\in R_i}\exp(\mathbf{a}'(\mathbf{z}_i\boldsymbol{\delta}_{\mathbf{z}} + v_i\delta_v))}\right)\end{bmatrix} = \begin{bmatrix}\mathbf{z}_i'\tilde{\varepsilon}(h_i) \\ v_i'\tilde{\varepsilon}(h_i)\end{bmatrix}$$

where

$$\tilde{\varepsilon}(h_i) = y_i - \frac{\sum_{\mathbf{a}\in R_i}\mathbf{a}\exp(\mathbf{a}'(\mathbf{z}_i\boldsymbol{\delta}_{\mathbf{z}} + v_i\delta_v))}{\sum_{\mathbf{a}\in R_i}\exp(\mathbf{a}'(\mathbf{z}_i\boldsymbol{\delta}_{\mathbf{z}} + v_i\delta_v))}$$

is the second-stage generalised residual. As the first-stage is linear, the first-stage score function $\mathbf{r}_i(\boldsymbol{\pi}) = \nabla_{\boldsymbol{\pi}}\ell^{(1)}(\boldsymbol{\pi})'$ has representation like above, i.e., $\mathbf{r}_i(\boldsymbol{\pi}) = \text{vec}[\mathbf{z}_i' v_i]$. Furthermore, let $\mathbf{s}_n(\boldsymbol{\delta}) = \sum_{i=1}^n \mathbf{s}_i(\boldsymbol{\delta})$ and $\mathbf{r}_n(\boldsymbol{\pi}) = \sum_{i=1}^n \mathbf{r}_i(\boldsymbol{\pi})$ where $\mathbf{s}_i(\boldsymbol{\delta}) = (s_{i1}(\boldsymbol{\delta}), s_{i2}(\boldsymbol{\delta}), ..., s_{iT}(\boldsymbol{\delta}))'$ and $\mathbf{r}_i(\boldsymbol{\pi}) = (r_{i1}(\boldsymbol{\pi}), r_{i2}(\boldsymbol{\pi}), ..., r_{iT}(\boldsymbol{,}\pi))'$. From the generalised residual form of the score I can derive the Hessian by noting that $\frac{\partial \mathbf{r}}{\partial \boldsymbol{\delta}'} = 0$

$$H_n = \begin{bmatrix}\frac{\partial \mathbf{s}_n}{\partial \boldsymbol{\delta}'} & \frac{\partial \mathbf{s}_n}{\partial \boldsymbol{\pi}'} \\ \frac{\partial \mathbf{r}_n}{\partial \boldsymbol{\delta}'} & \frac{\partial \mathbf{r}_n}{\partial \boldsymbol{\pi}'}\end{bmatrix} = \begin{bmatrix}H_{\boldsymbol{\delta}\boldsymbol{\delta},n} & H_{\boldsymbol{\delta}\boldsymbol{\pi},n} \\ 0 & I \otimes (\mathbf{z}'\mathbf{z})\end{bmatrix} \tag{A.3}$$

The asymptotic distribution of the scores as:

$$\frac{1}{\sqrt{n}}\begin{pmatrix}\mathbf{s}_n(\boldsymbol{\delta},\boldsymbol{\pi}) \\ \mathbf{r}_n(\boldsymbol{\pi})\end{pmatrix} \longrightarrow^d \mathcal{N}\left(\begin{bmatrix}0 \\ 0\end{bmatrix}, \begin{bmatrix}G & 0 \\ 0 & \Gamma\end{bmatrix}\right) = \mathcal{N}(\mathbf{0}, \boldsymbol{\Omega})$$

Where

$$G = \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[\mathbf{s}_i\mathbf{s}_i'|\mathbf{z}_i, v_i], \text{ and } \Gamma = \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[\mathbf{r}_i\mathbf{r}_i'|\mathbf{z}_i, v_i]$$

The control function approach yields the block diagonal variance-covariance matrix. This is especially useful as now I can derive the asymptotic distribution of the reduced-form parameter estimator

$$\begin{aligned}\sqrt{n}\begin{pmatrix}\hat{\boldsymbol{\delta}}-\boldsymbol{\delta}\\ \hat{\boldsymbol{\pi}}-\boldsymbol{\pi}\end{pmatrix} &=_a -\Big(\frac{1}{n}H_n\Big)^{-1}\frac{1}{\sqrt{n}}\begin{pmatrix}\mathbf{s}_n(\boldsymbol{\delta},\boldsymbol{\pi})\\ \mathbf{r}_n(\boldsymbol{\pi})\end{pmatrix}\\ &= -\left(\frac{1}{n}\begin{bmatrix}H_{\delta\delta,n} & H_{\delta\boldsymbol{\pi},n}\\ 0 & I\otimes(\mathbf{z}'\mathbf{z})\end{bmatrix}\right)^{-1}\frac{1}{\sqrt{n}}\begin{pmatrix}\mathbf{s}_n(\boldsymbol{\delta},\boldsymbol{\pi})\\ \mathbf{r}_n(\boldsymbol{\pi})\end{pmatrix}\end{aligned}$$

Using the chain rule, $\frac{\tilde{\varepsilon}(h_i)}{\partial\boldsymbol{\delta}} = \sigma_i\frac{\partial h_i}{\partial\boldsymbol{\delta}}$ and $\frac{\tilde{\varepsilon}(h)}{\partial\boldsymbol{\pi}} = -\delta_v\otimes\mathbf{z}$, where $\sigma_i = \frac{\partial\tilde{\varepsilon}(h_i)}{\partial h_i}$. Define $\boldsymbol{\Sigma} = \text{diag}(\sigma_i)$, yielding the following expression for the Hessian

$$H = \begin{bmatrix}\mathbf{z}'\boldsymbol{\Sigma}\mathbf{z} & \mathbf{z}'\boldsymbol{\Sigma}v & -\delta_v'\otimes\mathbf{z}'\boldsymbol{\Sigma}\mathbf{z}\\ v'\boldsymbol{\Sigma}\mathbf{z} & v'\boldsymbol{\Sigma}v & -\delta_v'\otimes v'\boldsymbol{\Sigma}\mathbf{z}\\ 0 & 0 & I\otimes(\mathbf{z}'\mathbf{z})\end{bmatrix} \tag{A.4}$$

Thus, H_n^{-1} may be found by the properities of upper-diagonal block matrices

$$H_n^{-1} = \begin{bmatrix}H_{\delta\delta,n}^{-1} & -H_{\delta\delta,n}^{-1}H_{\delta\boldsymbol{\pi},n}(I\otimes\mathbf{z}'\mathbf{z})^{-1}\\ 0 & (I\otimes\mathbf{z}'\mathbf{z})^{-1}\end{bmatrix}$$

I define $\frac{1}{n}\mathbf{z}'\mathbf{z}\to^p Q$, $\frac{1}{n}H_{\delta\delta,n}\to^p H_{\delta\delta}$, and $\frac{1}{n}H_{\delta\boldsymbol{\pi},n}\to^p H_{\delta\boldsymbol{\pi}}$. The Continuous Mapping Theorem implies $\frac{1}{n}H_{\delta\delta,n}^{-1}\to^p H_{\delta\delta}^{-1}$ and $\frac{1}{n}H_{\delta\boldsymbol{\pi},n}^{-1}\to^p H_{\delta\boldsymbol{\pi}}^{-1}$. Hence $H_n^{-1}\to^p H^{-1}$ where

$$H^{-1} = \begin{bmatrix}H_{\delta\delta}^{-1} & -H_{\delta\delta}^{-1}H_{\delta\boldsymbol{\pi}}(I\otimes Q)^{-1}\\ 0 & (I\otimes Q)^{-1}\end{bmatrix}$$

and so the asymptotic distribution is

$$\sqrt{n}\begin{pmatrix}\hat{\boldsymbol{\delta}}-\boldsymbol{\delta}\\ \hat{\boldsymbol{\pi}}-\boldsymbol{\pi}\end{pmatrix} =_a -H^{-1}\frac{1}{\sqrt{n}}\begin{bmatrix}\mathbf{s}(\boldsymbol{\delta},\boldsymbol{\pi})\\ \mathbf{r}(\boldsymbol{\pi})\end{bmatrix}$$

Furthermore, Slutsky's theorem asserts that

$$\sqrt{n}\begin{pmatrix}\hat{\boldsymbol{\delta}}-\boldsymbol{\delta}\\ \hat{\boldsymbol{\pi}}-\boldsymbol{\pi}\end{pmatrix}\to^d \mathcal{N}\left(\mathbf{0}, H^{-1}\boldsymbol{\Omega}[H^{-1}]'\right)$$

To find $H^{-1}\mathbf{\Omega}[H^{-1}]'$, note that $\mathbf{sr}' = 0$ due to the control function procedure, meaning any cross terms reduce to 0. Hence

$$H^{-1}\mathbf{\Omega}[H^{-1}]' = \begin{bmatrix} H_{\delta\delta}^{-1} & -H_{\delta\delta}^{-1}H_{\delta\pi}(I\otimes Q)^{-1} \\ 0 & (I\otimes Q)^{-1} \end{bmatrix}\begin{bmatrix} G & 0 \\ 0 & \Gamma \end{bmatrix}\begin{bmatrix} H_{\delta\delta}^{-1} & -H_{\delta\delta}^{-1}H_{\delta\pi}(I\otimes Q)^{-1} \\ 0 & (I\otimes Q)^{-1} \end{bmatrix}'$$

This greatly simplifies the expression to

$$= \begin{bmatrix} H_{\delta\delta}^{-1}GH_{\delta\delta}^{-1} + H_{\delta\delta}^{-1}H_{\delta\pi}(I\otimes Q)^{-1}\Gamma(I\otimes Q')^{-1}H_{\delta\pi}H_{\delta\delta}^{-1} & H_{\delta\delta}^{-1}H_{\delta\pi}(I\otimes Q)^{-1}\Gamma(I\otimes Q')^{-1} \\ (I\otimes Q)^{-1}\Gamma(I\otimes Q)^{-1}H_{\delta\pi}H_{\delta\delta}^{-1} & (I\otimes Q)^{-1}\Gamma(I\otimes Q)^{-1} \end{bmatrix}$$

Let $\Lambda_{\pi\pi} = (I\otimes Q)^{-1}\Gamma(I\otimes Q)^{-1}$. Recall that $H_{\delta\pi} = \text{vec}(-\delta_v'\otimes \mathbf{z}'\mathbf{\Sigma}\mathbf{z}, -\delta_v'\otimes v'\mathbf{\Sigma}\mathbf{z})$ and $H_{\delta\delta}$ is the left-upper-square block matrix. Hence, $H_{\delta\delta}^{-1}H_{\delta\pi} = -\delta_v'\otimes I_k$. I then simplify to:

$$H^{-1}\mathbf{\Omega}[H^{-1}]' = \begin{bmatrix} H_{\delta\delta}^{-1}GH_{\delta\delta}^{-1} + (\delta_v'\otimes I_k)\Lambda_{\pi\pi}(\delta_v\otimes I_k) & (\delta_v'\otimes I_k)\Lambda_{\pi\pi} \\ \Lambda_{\pi\pi}(\delta_v\otimes I_k) & \Lambda_{\pi\pi} \end{bmatrix} \tag{A.5}$$

An initial discussion of Equation (A.5) is required. Firstly, when using the two-step estimation procedure with a conditional logit second stage, the first element is easily estimable from standard statistical software. Assuming the likelihood is correctly specified, the Information Matrix Equality (IME) states that the inverse Hessian is equal to the outer product of the gradient function. Hence, the first term, $H_{\delta\delta}^{-1}GH_{\delta\delta}^{-1} = H_{\delta\delta}^{-1}$ is just the estimated variance from standard packages. $\Lambda_{\pi\pi}$ can be derived from the variance of the first stage via the following algorithm

1. Estimate the OLS variance, called V_1. Alternatively, if estimated via a correctly specified likelihood, the (IME) states that inverse Hessian matrix is equal to the outer product of the score. Hence, I simply estimate $\Gamma = N^{-1}\sum_{i=1}^{N}\mathbf{r}_i\mathbf{r}_i'$.

2. Now, estimate $H_{\delta\pi}$. This is fairly simple as it is the gradient of the second-stage score, $\mathbf{s}(\boldsymbol{\delta}, \hat{\boldsymbol{\pi}})$, with respect to $\boldsymbol{\pi}$. Call this matrix $F = N^{-1}\sum_{i=1}^{N}\nabla_{\boldsymbol{\pi}}\mathbf{s}_i(\boldsymbol{\delta}, \hat{\boldsymbol{\pi}})$

3. Then, the correct variance for second stage estimator $\boldsymbol{\delta}$ is $H_{\delta\delta}^{-1}GH_{\delta\delta}^{-1} + H_{\delta\delta}^{-1}F\Gamma F'H_{\delta\delta}^{-1}$

This algorithm is equivalent to the corrected two-step M-estimator variance proposed in Wooldridge (2010) in section 12.4.2 and 12.5.2. Briefly, Wooldridge (2010) uses a first-order expansion to derive the following:

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathbf{s}_i(\boldsymbol{\delta};\hat{\boldsymbol{\pi}}) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathbf{s}_i(\boldsymbol{\delta};\boldsymbol{\pi}^*) + F_0\sqrt{n}(\hat{\boldsymbol{\pi}} - \boldsymbol{\pi}^*) + o_p(1)$$

Where $F_0 = \mathbb{E}\left[\nabla_{\boldsymbol{\pi}} \mathbf{s}(\boldsymbol{\delta}; \boldsymbol{\pi}^*)\right]$. I assume that for a first-stage OLS

$$\sqrt{n}(\hat{\boldsymbol{\pi}} - \boldsymbol{\pi}^*) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \mathbf{r}_i(\boldsymbol{\pi}^*) + o_p(1)$$

where $\mathbf{r}$ are first-stage scores. Then, defining $\mathbf{q}_i(\boldsymbol{\delta}; \boldsymbol{\pi}) \equiv \mathbf{s}_i(\boldsymbol{\delta}; \boldsymbol{\pi}) + F_0 \mathbf{r}_i(\boldsymbol{\pi}^*)$. Taking $\mathbb{E}[\mathbf{q}_i(\boldsymbol{\delta}; \boldsymbol{\pi}) \mathbf{q}_i(\boldsymbol{\delta}; \boldsymbol{\pi})']$, and eliminating cross-terms between $\mathbf{r}_i$ and $\mathbf{s}_i$ yields the corrected variance $\mathbb{V}[\boldsymbol{\delta}] = H_{\boldsymbol{\delta}\boldsymbol{\delta}}^{-1}\left(G + F\Gamma F'\right) H_{\boldsymbol{\delta}\boldsymbol{\delta}}^{-1}$, equivalent to above. From here, I derive the asymptotic variance of the link function $r(\delta, \beta) = \delta - \pi\beta$, denoted $\Psi(\beta_0)$ by pre-multiplying (A.5) by $[(1 \ -\beta_0') \otimes (I_{k_z} \ 0)]$ and post-multiplying by $[(1 \ -\beta_0')' \otimes (I_{k_z} \ 0)']$

$$\Psi(\beta_0) = H_{\boldsymbol{\delta}\boldsymbol{\delta}}^{-1} G H_{\boldsymbol{\delta}\boldsymbol{\delta}}^{-1} + [(\delta_v - \beta_0)' \otimes I_{k_z}] \Lambda_{\pi\pi} [(\delta_v - \beta_0) \otimes I_{k_z}] \tag{A.6}$$

Which is then used to calculate the Anderson-Rubin type test

$$AR(\beta_0) = (\hat{\delta} - \hat{\pi}\beta_0)' [\hat{\Psi}(\beta_0)]^{-1} (\hat{\delta} - \hat{\pi}\beta_0) \tag{A.7}$$

where $\hat{\Psi}(\beta_0)$ is the estimate of $\Psi(\beta_0)$, obtained from replacing $H_{\boldsymbol{\delta}\boldsymbol{\delta}}$, G, δ_v, and $\Lambda_{\boldsymbol{\pi}\boldsymbol{\pi}}$ by their respective estimates under the assumption $H_0 : \beta = \beta_0$. AR(β_0) is $\chi^2(k)$ distributed under the Null assumption $\beta = \beta_0$ because $(\hat{\delta} - \hat{\pi}\beta_0)$ is Normally distributed with variance $\hat{\Psi}(\beta_0)$.

A.3 Proof of Result 3.1

Proof. Suppose QML estimator is a consistent estimator of β_0 and that individual i is in group 1, i.e., $i \in \mathcal{O}$. Define the quasi-score and true-score

$$s_i^q(\beta) \quad = \nabla_\beta \ell_i^q(\beta)' = \quad \frac{d_i}{q\sigma}\{w_i - \Lambda(k_i)\} \tag{A.8}$$

$$s_i(\beta) \quad = \nabla_\beta \ell_i^{(1)}(\beta)' = \quad \frac{d_i}{q\sigma_j}\left\{w_i - \Lambda_2(k_i)\right\} \tag{A.9}$$

where $d_i = \partial k_i / \partial \beta = x_{i2} - x_{i1}$. By assumption, the true parameter $\beta_0 \in \mathrm{B}$, where $\mathrm{B} \subset \mathbb{R}$ is compact, uniquely maximises $\mathbb{E}[\ell^q(\beta)]$ by Equation (3.1). Thus $\mathbb{E}(s_i^q(\beta_0)|x_{i1}, x_{i2}) = 0$. By Equation (A.8) it follows that $\mathbb{E}(w_i|x_{i1}, x_{i2}) = \Lambda(k_i^0)$. However, the true maximum likelihood defined in subsection 3.1 must also produce a consistent estimator of β_0, and hence by Equation (A.8)

$$\mathbb{E}[s_i(\beta_0)|x_{i1}, x_{i2}] = \frac{d_i}{q\sigma_1}\{\mathbb{E}(w_i|x_{i1}, x_{i2}) - \Lambda_1(k_i^0)\} = 0$$

However, as $\mathbb{E}(w_i|x_{i1}, x_{i2}) = \Lambda(k_i^0)$

$$\mathbb{E}[s_i(\beta_0)|x_{i1}, x_{i2}] = \frac{d_i}{q\sigma_1}\{\Lambda(k_i^0) - \Lambda_1(k_i^0)\} = 0$$

as the true score must be valid. However, $\Lambda(k_i^0) \neq \Lambda_1(k_i^0)$ in general and so $\mathbb{E}[s_i(\beta_0)|x_{i1}, x_{i2}] \neq 0$, which is a contradiction. Hence, our assumption that $\beta_0 \in \mathrm{B}$ is the unique maximiser of $\mathbb{E}[\ell_i^q(\beta)]$ is false. A symmetric argument applies when individual i is in Group 2. Therefore, the QML estimator $\tilde{\beta}$ is inconsistent for β_0. Rather, $\tilde{\beta}$ is consistent for some $\beta^* \neq \beta_0$ where $\mathbb{E}\left[s_i^q(\beta^*)\right] = 0$, i.e., $\text{plim}(\tilde{\beta}) = \beta^* \neq \beta_0$. □

A.4 Proof of Result 3.2

A.4.1 Heteroskedastic second-stage

Proof. I consider the case where each individual has a different variance. I focus on the two-period case from which I infer issues with the general T-period likelihood. I include one endogenous regressor and one instrument. The model is

$$\begin{aligned} x_{it} &= z_{it}\xi + \phi_i + v_{it} \\ y_{it}^* &= x_{it}\beta + c_i + \rho v_{it} + \varepsilon_{it} \\ \mathbb{V}[\varepsilon_{it}] &= [V(g_i\gamma)]^2 > 0 \end{aligned}$$

where $v_{it} \sim \mathcal{N}(0, \sigma^2)$. I include a generated regressor $\hat{v}_{it}$ in place of v_{it} in the second-stage. I implement a conditional likelihood approach to gain consistent estimators of β or of reduced-form parameter $\delta_z = \xi\beta$. Therefore, I assume the $\varepsilon_{it}^{(i)}$ are Logistically distributed with location parameter 0 and scale parameter $q[V(g_i\gamma)]$, where $q = \sqrt{3}/\pi$. The log-likelihood of the two-step procedure has form $\ell_i(\cdot) = \left(\ell_i^{(2)}(\cdot), \ell_i^{(1)}(\cdot)\right)$, where $\ell_i^{(2)}$ comes from the conditional logit log-likelihood and $\ell_i^{(1)}$ comes from the log-likelihood of a Normal distribution as I consider a linear first-stage. Firstly, I consider the second-stage likelihood and how second-stage heteroskedasticity harms estimation. The second-stage standard deviation of errors is $V(g_i\gamma)$, which I use to derive $\ell_i^{(2)}$,

$$\begin{aligned} \mathbb{P}(y_{it} = 1|x_{it}, v_{it}) &= \mathbb{P}(y_{it}^* > 0|x_{it}, v_{it}) \\ &= \mathbb{P}\Big(\frac{\varepsilon_{it}}{qV(g_{it}\gamma)} > -\frac{x_{it}\beta + c_i + \rho v_{it}}{qV(g_i\gamma)}\Big| x_{it}, v_{it}\Big) \\ &= 1 - \Lambda\Big[\frac{x_{it}\beta + c_i + \rho v_{it}}{qV(g_i\gamma)}\Big] \end{aligned}$$

Where $\Lambda(\cdot)$ is the standard Logistic CDF with scale 1 and location 0. Letting $\mathbf{x}_{it} = [x_{it}, v_{it}]$ and $\boldsymbol{\beta} = (\beta, \rho)'$, I now condition on $n_i = y_{i1} + y_{i2} = 1$ to gain the following

$$\mathbb{P}(y_{it} = 1|\mathbf{x}_{it}, g_i, n_i = 1) = 1 - \Lambda\Big[\frac{(\mathbf{x}_{i1} - \mathbf{x}_{i2})\boldsymbol{\beta}}{qV(g_i\gamma)}\Big]$$

And thus the second-stage log-likelihood of individual i is

$$\ell_i^{(2)}(\boldsymbol{\beta};\gamma) = w_i \log\left\{1-\Lambda(h_i)\right\} + (1-w_i)\Lambda(h_i);\ h_i = \frac{(\mathbf{x}_{i1}-\mathbf{x}_{i2})\boldsymbol{\beta}}{V(g_i\gamma)}$$

Where $w_i = 1$ if $y_{i1} = 0$ and $y_{i1} = 1$ and $w_i = 0$ if $y_{i1} = 1$ and $y_{i1} = 0$. Let $\mathbf{d}_i = \mathbf{x}_{i1} - \mathbf{x}_{i2}$. I differentiate to gain the log-score function

$$[\nabla_{\boldsymbol{\beta}}\ell_i^{(2)}(\boldsymbol{\beta};\gamma)]' = t_i^{(2)}(\boldsymbol{\beta};\gamma) = \frac{\mathbf{d}_i'}{qV(g_i\gamma)}\frac{\lambda(h_i)[w_i-\Lambda(h_i)]}{\Lambda(h_i)[1-\Lambda(h_i)]}$$

$$\begin{aligned}[\nabla_{\gamma}\ell_i^{(2)}(\boldsymbol{\beta};\gamma)]' = r_i^{(2)}(\boldsymbol{\beta};\gamma) &= \left[\frac{\partial h_i}{\partial\gamma}\right]'\frac{\lambda(h_i)[w_i-\Lambda(h_i)]}{\Lambda(h_i)[1-\Lambda(h_i)]} \\ &= \frac{q\mathbf{d}_i'\boldsymbol{\beta}' g_i' \mathrm{v}(g_i\gamma)}{[qV(g_i\gamma)]^2}\frac{\lambda(h_i)[w_i-\Lambda(h_i)]}{\Lambda(h_i)[1-\Lambda(h_i)]}\end{aligned}$$

Where $\mathrm{v}(g_i\gamma) = \partial V(g_i\gamma)/\partial\gamma$. The full second-stage score is $\mathbf{s}_i^{(2)}(\boldsymbol{\beta};\gamma) = (t_i^{(2)}(\boldsymbol{\beta})', r_i^{(2)}(\gamma)')'$. The full second-stage score has a generalised residual formulation in $\tilde{e}(h_i) = w_i - \Lambda(h_i)$ which has 0 expectation conditional on $\mathbf{x}_i = [\mathbf{x}_{i1}, \mathbf{x}_{i2}]$ and g_i because $\Lambda(h_i)$ by definition is the probability that $y_{i1} = 0$ and $y_{i2} = 1$. It follows that $\mathbb{E}[\mathbf{s}_i(\boldsymbol{\beta};\gamma)|\mathbf{x}_i, g_i] = 0$ and so $\mathbb{E}[w_i|\mathbf{x}_i, g_i] = \Lambda(h_i)$. To complete the proof, I consider the second-stage quasi-log-likelihood, denoted $\mathscr{q}^{(2)}(\boldsymbol{\beta})$, which assumes constant variance across individuals, i.e., $V(g_i\gamma) = \sigma > 0$. Hence, the quasi-log-likelihood score is

$$\nabla_{\boldsymbol{\beta}}\mathscr{q}_i(\boldsymbol{\beta})' = s_i^{(\mathscr{q})}(\boldsymbol{\beta}) = \frac{\mathbf{d}_i'}{q\sigma}\frac{\lambda(k_i)[w_i-\Lambda(k_i)]}{\Lambda(k_i)[1-\Lambda(k_i)]}$$

Where $k_i = (\mathbf{x}_{i1} - \mathbf{x}_{i2})\boldsymbol{\beta}/q\sigma$ and $\mathbf{d}_i = \mathbf{x}_{i1} - \mathbf{x}_{i2}$. I proceed as in A.3 – assume the quasi-MLE is valid and so $\mathbb{E}[w_i|\mathbf{x}_i] = \Lambda(k_i)$, where $\mathbf{x}_i = [\mathbf{x}_{i1}, \mathbf{x}_{i2}]$. However the true score must also be valid and so this implies that $\Lambda(k_i) = \Lambda(h_i)$, which is false in general and so is a contradiction. Therefore, the QMLE does yield a consistent estimator of $\boldsymbol{\beta}_0$. □

A.4.2 Heteroskedastic first-stage

Proof. I consider the following DGP for the one endogenous regressor and one instrument case

$$\begin{aligned} y_{it}^* &= x_{it}\beta + c_i + \rho v_{it} + \varepsilon_{it} \\ x_{it} &= z_{it}\xi + \phi_i + v_{it}^{(i)} \\ \mathbb{V}[v_{it}^{(i)}] &= [V(g_i\gamma)]^2 > 0 \end{aligned}$$

where ε_{it} is a Logistic distribution with location 0 and constant scale parameter $q\sigma_2$ and $v_{it}^{(i)}$ is a Normal distribution with mean 0 and variance $[V(g_i\gamma)]^2$. As before, I consider the log-likelihood

of the two-step procedure for individual i as $\ell_i(\boldsymbol{\beta};\gamma) = (\ell_i^{(2)}(\boldsymbol{\beta})', \ell_i^{(1)}(\xi;\gamma)')'$. Here $\ell_i^{(2)}(\boldsymbol{\beta};\gamma)$ is the log-likelihood of the second-stage for individual i and comes from the conditional logit likelihood, while $\ell_i^{(1)}(\xi;\gamma)$ is the log-likelihood of the first-stage for individual i and comes from the log-likelihood of a Normal distribution. When heteroskedasticity is present, quasi-maximum likelihood estimation of the linear first-stage is consistent for ξ although standard errors are inconsistent (Martin et al., 2012). In this case, the second-stage quasi-log-likelihood coincides with the true second-stage log-likelihood. To spell this out, I have the following expressions for the quasi- and true-log-likelihood for the first-stage

$$\begin{aligned} q_i^{(1)}(\xi) &= -\ln 2\pi - \ln \sigma_1^2 - \frac{1}{\sigma_1^2}\sum_{t=1}^{2}(x_{it} - z_{it}\xi - \phi_i)^2 \\ \ell_i^{(1)}(\xi,\gamma) &= -\ln 2\pi - \ln [V(g_i\gamma)]^2 - \frac{1}{[V(g_i\gamma)]^2}\sum_{t=1}^{2}(x_{it} - z_{it}\xi - \phi_i)^2 \end{aligned}$$

Where $q_i^{(1)}$ and $\ell_i^{(1)}$ are the quasi and true log-likelihoods of the first-stage, respectively. For the second-stage,

$$\ell_i^{(2)}(\boldsymbol{\beta}) = w_i \log\big\{1 - \Lambda(h_i)\big\} + (1 - w_i)\big\{\Lambda(h_i)\big\},\ h_i = \frac{(\mathbf{x}_{i1} - \mathbf{x}_{i2})\boldsymbol{\beta}}{q\sigma_2}$$

is the true log-likelihood of the first-stage. I expand h_i as

$$h_i = \frac{(x_{i1} - x_{i2})\beta + (v_{i1} - v_{i2})\rho}{q\sigma_2} \tag{A.10}$$

I use a generated regressor $\hat{v}_{it} = x_{it} - \hat{\xi}z_{it} - \hat{\phi}_i$ in place of v_{it}. As already stated, the first-stage quasi-maximum likelihood yields consistent estimates of ξ and ϕ, and thus $\hat{v}_{it}$ is a consistent estimator of $\hat{v}_{it}$. To be precise, I compare the two log-likelihoods

$$\begin{aligned} \text{True: } \ell_i(\boldsymbol{\beta};\gamma,\xi) &= \big(\ell_i^{(2)}(\boldsymbol{\beta}), \ell_i^{(1)}(\xi,\gamma)\big) \\ \text{Quasi: } q_i(\boldsymbol{\beta};\xi) &= \big(\ell_i^{(2)}(\boldsymbol{\beta}), q_i^{(1)}(\xi)\big) \end{aligned}$$

To distinguish between true and QMLE, I call first-stage residuals computed from QMLE $\tilde{v}_{it}$ while first-stage residuals computed from the true likelihood are $\hat{v}_{it}$. To complete the proof, I show that if $\boldsymbol{\beta}_0 \in \mathrm{B}$, where B is a compact subset of $\mathbb{R}^2$, is a unique maximiser of the expected true log-likelihood then $\boldsymbol{\beta}_0$ is the unique maximiser of the expected quasi log-likelihood. This would imply that $\operatorname{plim}(\tilde{\boldsymbol{\beta}}) = \boldsymbol{\beta}_0$ where $\tilde{\boldsymbol{\beta}}$ is the estimator obtained from QMLE; that is, $\tilde{\boldsymbol{\beta}}$ is a consistent estimator of the true $\boldsymbol{\beta}_0$.

Suppose $\boldsymbol{\beta}_0 = \arg\max_{\boldsymbol{\beta}\in\mathrm{B}}\{\mathbb{E}[\ell_i(\boldsymbol{\beta};\gamma,\xi)|\gamma = \hat{\gamma}, \xi = \hat{\xi}]$, where $\hat{\xi}$ and $\hat{\gamma}$ are consistent and unbiased estimators of γ and ξ derived from the first-stage. This is equivalent to saying $\boldsymbol{\beta}_0$ is the unique maximiser of $\mathbb{E}[\ell_i^{(2)}(\boldsymbol{\beta})|\hat{\gamma},\hat{\xi}]$ after estimating the first-stage. It then follows that $\mathbb{E}[\mathbf{s}_i^{(2)}(\boldsymbol{\beta}_0)] = 0$ where $\mathbf{s}_i^{(2)}(\boldsymbol{\beta})$ is the true score of the second-stage, defined as

$$\mathbf{s}_i^{(2)}(\boldsymbol{\beta}) = \frac{\mathbf{d}_i'}{q\sigma}\frac{\lambda(h_i)(w_i - \Lambda(h_i)}{\Lambda(h_i)(1-\Lambda(h_i)}$$

where $\mathbf{d}_i = \partial h_i/\partial\boldsymbol{\beta}$. Given our assumption that $\boldsymbol{\beta}_0$ is the unique maximiser of the true log-likelihood, it follows that

$$\mathbb{E}[s_i^{(2)}(\boldsymbol{\beta}_0)|\mathbf{d}_i] = 0 \implies \mathbb{E}[w_i|\mathbf{d}_i] = \mathbb{E}[\Lambda(\hat{h}_i^0)|\mathbf{d}_i] \tag{A.11}$$

where $\hat{h}_i^0$ is equation (A.10) evaluated at $\boldsymbol{\beta} = \boldsymbol{\beta}_0$ and $v_{it} = \hat{v}_{it}$ for $t = 1, 2$. Now consider $\mathbb{E}[\mathbf{r}_i^{(2)}(\boldsymbol{\beta}_0)|\mathbf{d}_i]$, where $\mathbf{r}_i^{(2)}(\boldsymbol{\beta})$ is the quasi second-stage score defined as

$$\mathbf{r}_i^{(2)}(\boldsymbol{\beta}) = \frac{\mathbf{d}_i'}{q\sigma}\frac{\lambda(h_i)\big(w_i - \Lambda(h_i)\big)}{\Lambda(h_i)(1-\Lambda(h_i)}$$

with all variables and constants defined as before. I have

$$\mathbb{E}[\mathbf{r}_i^{(2)}(\boldsymbol{\beta}_0)|\mathbf{d}_i] = M(\tilde{k}_i^0)\big\{\mathbb{E}[w_i - \Lambda(\tilde{k}_i^0)|\mathbf{d}_i]\big\};\ M(\tilde{k}_i^0) = \frac{\mathbf{d}_i'\lambda(\tilde{h}_i^0)}{q\sigma\Lambda(\tilde{h}_i^0)[1-\Lambda(\tilde{h}_i^0)]}$$

where $\tilde{h}_i^0$ is Equation (A.10) evaluated at $\boldsymbol{\beta} = \boldsymbol{\beta}_0$ and $v_{it} = \tilde{v}_{it}$ for $t = 1, 2$, denoting the fitted residuals from QMLE on the first-stage. As stated before, $\tilde{v}_{it}$ is a consistent estimator of v_{it} and thus $\mathbb{E}[\Lambda(\tilde{h}_i^0)] = \mathbb{E}[\Lambda(\hat{k}_i^0)]$. Expression (A.11) yields

$$\mathbb{E}[\mathbf{r}_i^{(2)}(\boldsymbol{\beta}_0)|\mathbf{d}_i] = M(\tilde{k}_i^0)\big\{\mathbb{E}[w_i|\mathbf{d}_i] - \mathbb{E}[\Lambda(\tilde{k}_i^0)|\mathbf{d}_i]\big\} = M(\tilde{k}_i^0)\big\{\mathbb{E}[\Lambda(\hat{h}_i^0)|\mathbf{d}_i] - \mathbb{E}[\Lambda(\tilde{k}_i^0)|\mathbf{d}_i]\big\} = 0$$

which implies that $\boldsymbol{\beta}_0$ is a unique maximiser of the quasi log-likelihood. Therefore, the quasi-maximum likelihood estimator $\tilde{\boldsymbol{\beta}}$ is a consistent estimator of the true parameter vector $\boldsymbol{\beta}_0$, i.e., $\mathrm{plim}(\tilde{\boldsymbol{\beta}}) = \boldsymbol{\beta}_0$. □

A.5 Delta Method For AME Standard Errors

Suppose that $\sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \sim \mathcal{N}(0, \Omega)$ where $\widehat{\boldsymbol{\theta}}_n$ is a sequence of random variables and Ω is the variance-covariance matrix of $\widehat{\boldsymbol{\theta}}_n$. I derive the distribution of $\sqrt{n}(a(\widehat{\boldsymbol{\theta}}_n) - a(\boldsymbol{\theta}_0))$, where $y(\cdot)$ is some function. Consider the mean-value expansion

$$a(\hat{\boldsymbol{\theta}}_n) = a(\boldsymbol{\theta}_0) + a'(\tilde{\boldsymbol{\theta}})\big(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\big)$$

where $\widetilde{\theta}$ lies between $\widehat{\theta}_n$ and θ_0. Rearranging and multiplying by $\sqrt{n}$

$$\sqrt{n}\big(a(\widehat{\theta}_n) - a(\theta_0)\big) = a'(\widetilde{\theta})\sqrt{n}\big(\widehat{\theta}_n - \theta_0\big)$$

Furthermore, $\widetilde{\theta} \to \theta_0$ since $\widehat{\theta}_n \to \theta_0$ and $|\widetilde{\theta} - \theta_0| < |\widehat{\theta}_n - \theta_0|$. Hence

$$\sqrt{n}\big(a(\widehat{\theta}_n) - a(\theta_0)\big) \approx a'(\widehat{\theta}_n)\sqrt{n}\big(\widehat{\theta}_n - \theta_0\big)$$

in the limit $n \to \infty$. Therefore by Slutsky's theorem $\sqrt{n}\big(a(\widehat{\theta}_n) - a(\theta_0)\big) \sim \mathcal{N}(0, [a'\big(\widehat{\theta}_n\big)]^2\Omega\big)$. Applied to AMEs

$$a(\widehat{\theta}) = \bar{d}_j(\mathbf{x}) = \frac{\hat{\theta}_j}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T} g(\mathbf{x}_{it}\hat{\theta}),\ a'(\widehat{\theta}) = \nabla_{\theta_j}\bar{d}_j(\mathbf{x}) = \frac{\hat{\theta}_j}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T} x_{it,j}\, g'(\mathbf{x}_{it}\hat{\theta})$$

B Appendix B

B.1 Panel Inference

Table 9: Size Comparison (%) – $v_{it} \sim \mathcal{N}(0,2)$, $\varepsilon_{it} \sim \mathscr{L}\left(0, \frac{\sqrt{3}}{\pi}\right)$

	$\mu = 0.01$				$\mu = 3$				$\mu = 500$			
ρ	0.2		0.99		0.2		0.99		0.2		0.99	
Size	5%	10%	5%	10%	5%	10%	5%	10%	5%	10%	5%	10%
Wald($\hat{\beta}$)	9.09	15.11	18.45	26.57	9.63	15.95	18.61	26.34	5.21	10.55	7.88	14.33
$AR(\beta_0)$	4.63	9.59	4.77	9.52	5.17	10.35	4.60	9.67	5.04	10.23	4.58	9.67

Note. I test the hypothesis $H_0 : \beta_0 = 0.5$ against $H_1 : \beta_0 \neq 0.5$. The number of simulations is $N = 10,000$ and the panel dimensions are $n = 100$ and $T = 10$. AR(β_0) rejection rates are the correct size while standard Wald($\hat{\beta}$) significantly over-reject even in what is deemed extremely strong instruments, $\mu = 500$.

Table 10: Size Comparison (%) – LPM Models, $\varepsilon_{it} \sim \mathcal{N}\left(0, \frac{\sqrt{3}}{\pi}\right)$

	$\mu = 0.01$				$\mu = 3$				$\mu = 10$			
ρ	0.2		0.99		0.2		0.99		0.2		0.99	
Size	5%	10%	5%	10%	5%	10%	5%	10%	5%	10%	5%	10%
Wald($\hat{\beta}$)	91.39	92.69	91.39	92.72	91.28	92.72	91.28	92.75	91.21	92.79	91.27	92.76
$AR(\beta_0)$	5.05	10.24	5.07	10.25	5.09	10.33	5.10	10.32	5.24	10.56	5.24	10.56

Note. I test the hypothesis $H_0 : \beta_0 = 0.5$ against $H_1 : \beta_0 \neq 0.5$. The number of simulations is $N = 10,000$ and the panel dimensions are $n = 100$ and $T = 10$. AR(β_0) rejection rates are the correct size while standard Wald($\hat{\beta}$) significantly over-reject even in what is deemed strong instruments, $\mu = 10$. I choose $\mu = 10$ for strong instruments since this is commonly cited as the approximate cutoff for strong instruments in linear models.

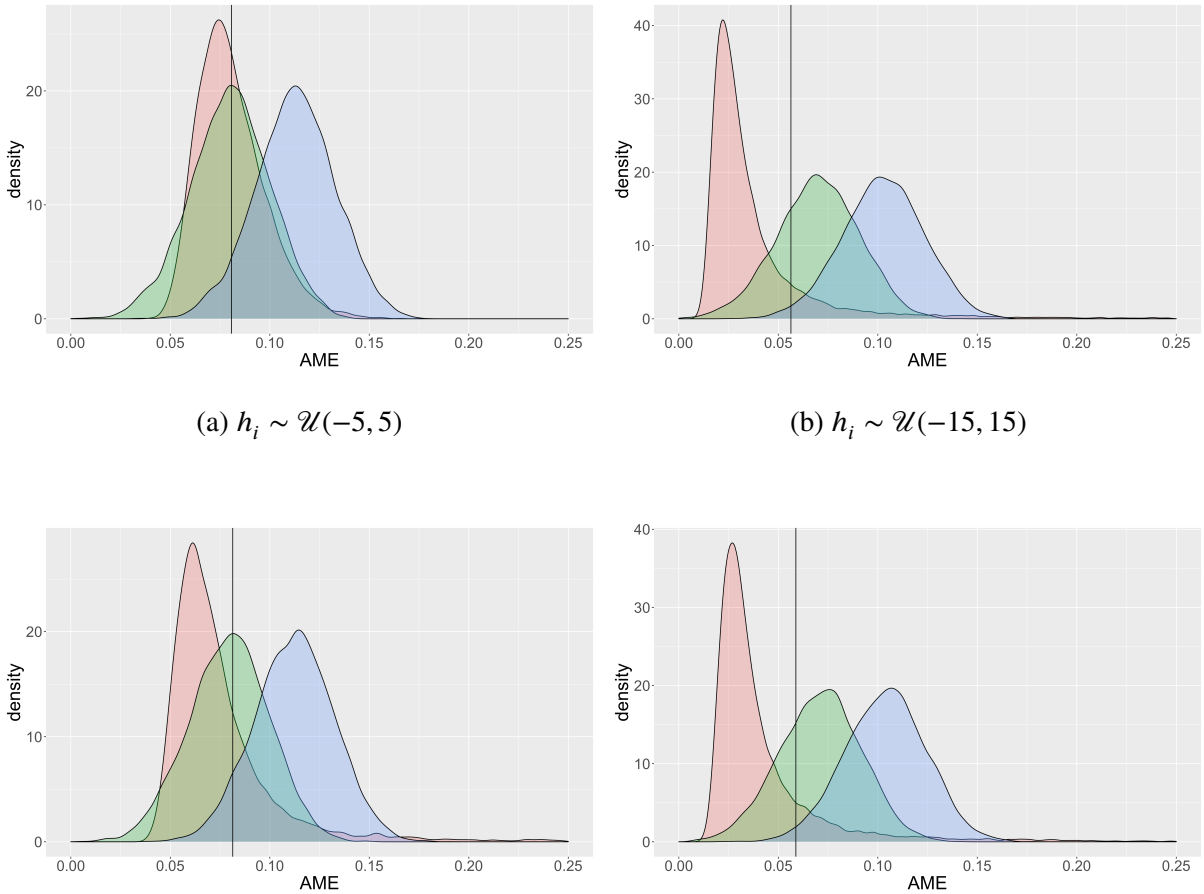
B.2 AME Individual Heteroskedasticity

I simulate the individual heteroskedasticity model from Harvey (1976)

$$\begin{aligned} y_i^* &= x_i\beta + v_i\rho + \varepsilon_i \\ x_i &= z_i\xi + v_i \\ \varepsilon_i &\sim \mathscr{L}(0, q\exp(h_i\gamma)) \\ v_i &\sim \mathcal{N}(0,1) \end{aligned} \tag{B.1}$$

where $z_i \sim \mathcal{N}(0,1)$. Equations (B.1) and (B.2) represent first- and second-stage heteroskedasticity, respectively. I set $\beta = 0.5$, $\gamma = 1$, ξ according to the concentration parameter $\mu \in \{1, 500\}$ described in Section 2, and $\rho \in \{0.2, 0.99\}$ to denote low and high endogeneity respectively. Recall that $q = \sqrt{3}/\pi$ is a constant. The sample size for each simulation is $n = 1,000$ and I conduct $N = 10,000$ simulations. In each simulation I calculate the true marginal effect via Equation (4.3) to the model. Vertical black bars represent the mean true AME across the $10,000$ simulations.

Figure 5: AME Estimator Individual Heteroskedasticity ($\mu = 500,\ \rho = 0.99$)



(a) $h_i \sim \mathcal{U}(-5, 5)$

(b) $h_i \sim \mathcal{U}(-15, 15)$

(c) $h_i \sim \mathcal{N}(0, 4^2)$

(d) $h_i \sim \mathcal{N}(0, 10^2)$

The individual heteroskedasticity simulations demonstrate that quasi-AMEs are consistent up until some critical variance in the explanatory variables determining the individual's error variance, denoted h_i. Consistency occurs despite the stark distributional differences between the true AME estimator and the quasi-AME estimator. Increasing the variance of the h_i means

larger possible observed h_i's. Marginal increases in h_i when the h_i is already large increase the individuals error variance more than the same increase when h_i is low because of the exponential function. Increasing the variance of h_i yields greater error variance separation for individuals in the higher regions of the h_i's than lower regions of the h_i's, producing a long right tail with low density, where separation is high, and left peak with high density, where the separation is low. These dynamics explain the distributional qualities of the true AMEs in Figure 5 and subsequently the eventual inconsistency of quasi-AMEs once the variance of h_i reaches some critical value.

B.3 Summary Statistics for Applications

Table 11: Civil War Onset ($n = 1454$, 1971–2006)

	Mean	Std. Dev.	Min.	Max.
Intra-state conflict onset	0.06	0.24	0	1
US Wheat Aid ('000s mt tns)	21.08	59.42	0	791.60
Lagged US Wheat Production ('000s mt tns)	59187	8754	36787	75813
Ave. US food aid probability	0.39	0.33	0	1
Peace Duration (yrs)	11.59	9.48	1	46
Instrument	22936	19924	0	75813

Note. An observation is a country-year pair. Country-year pairs without valid peace duration realisations or without possible transitions to intra-state conflict in the next period are omitted, leaving 1454 observations. The instrument is Lagged US Wheat Production interacted with the average probability of US Food Aid.

Table 12: Civil War Incidence ($n = 4089$, 1971–2006)

	Mean	Std. Dev.	Min.	Max.
Conflicts (+25 deaths)	0.22	0.41	0	1
US Wheat Aid ('000s mt tns)	27.61	116.61	0	1958
Lagged US Wheat Production ('000s mt tns)	59053	9176	36787	75813
Ave. US food aid probability	0.37	0.31	0	1
Instrument	22040	18950	0	75813
Controls				
Real US GDP per capita	3.86	3.23	0	10.67
Oil Price	16.07	17.14	0	100.54
US Democratic President	0.13	0.26	0	1

Note. An observation is a country-year pair. The instrument is Lagged US Wheat Production interacted with the average probability of US Food Aid. All controls are interacted with the average probability of receiving US food aid.

B.4 Parameter Estimates – Nunn and Qian (2014)

Table 13: Parameter Estimates – Nunn and Qian (2014)

(a) Onset – Logit Time Hazard Model

	Spec. (1)	Spec. (2)	Spec. (3)
$\hat{\alpha}$	1.21	1.38	1.34
$\hat{\beta}$	1.32	−0.81	−0.27
$\hat{\delta}$	1.59	−1.12	−0.30

(b) Incidence – LPM

	C. FE	*Most*	*Full*
$\hat{\alpha}$	2.27	3.30	3.58
$\hat{\beta}$	3.64	3.43	2.99
$\hat{\delta}$	8.29	11.33	10.71

Note. Results in the table are multiplied by 10^3 for $\hat{\alpha}$ and $\hat{\beta}$ and 10^6 for $\hat{\delta}$ and $r(\hat{\delta}, \beta_0)$ for brevity.