cemfi

working paper 2406

Information matrix tests for multinomial logit models

Dante Amengual
Gariele Fiorentini
Enrique Sentana

April 2024

CENTRO DE ESTUDIOS
MONETARIOS Y FINANCIEROS
Casado del Alisal 5, 28014 Madrid, Spain
www.cemfi.es

CEMFI Working Paper No. 2406
April 2024

Information matrix tests for multinomial logit models

Abstract

We show that the influence functions of the information matrix test for the multinomial logit model are the Kronecker product of the outer product of the generalised residuals minus their covariance matrix conditional on the explanatory variables times the outer product of those variables. Thus, it resembles a multivariate heteroskedasticity test à la White (1980), which confirms Chesher's (1984) unobserved heterogeneity interpretation. Our simulation experiments indicate that using theoretical expressions for the conditional covariance matrices involved substantially reduces size distortions, while the parametric bootstrap practically eliminates them. We also show that the test has good power against several relevant alternatives.

JEL Codes: C35, C25.

Keywords: Hessian matrix, outer product of the score, specification test, unobserved heterogeneity.

Dante Amengual
CEMFI
amengual@cemfi.es

Gariele Fiorentini
Università di Firenze and RCEA
gabriele.fiorentini@unifi.it

Enrique Sentana
CEMFI
sentana@cemfi.es

Acknowledgement

We would like to thank Manuel Arellano, as well as audiences at CEMFI, Duke and Università di Firenze for useful comments and suggestions. Of course, the usual caveat applies. The first and third authors gratefully acknowledge financial support from the Spanish Ministry of Science and Innovation through grant PID2021-128963NB-I00, while the second one is thankful to MIUR through the PRIN project “High-dimensional time series for structural macroeconomic analysis in times of pandemic”.

1 Introduction

White's (1982) information matrix (IM) test provides a general procedure for examining the correct specification of models estimated by maximum likelihood (ML). It directly assesses the IM equality, which states that the sum of the Hessian matrix and the outer product of the score vector should be zero in expected value when the estimated model is correctly specified. Chesher (1984) reinterpreted it as a score test against unobserved heterogeneity, a serious concern in microeconometric models as the parameters characterising objective functions or constraints often vary across agents. Not surprisingly, the IM test has been extensively studied for univariate probit and tobit models (see Horowitz (1994) and the references therein).

However, the IM test has not been derived for multinomial logit models in which the explanatory variables are common across categories but their effects are not. Polytomous choice models specify how the probabilities of mutually exclusive Bernoulli variables that make up a multinomial random variable $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_K)'$ vary across observations as a function of L observed characteristics $\mathbf{z}$. Typically, they are parametrised as

$$p_k = \Pr(\xi_k = 1|\mathbf{z}) = F_k(\mathbf{z}; \boldsymbol{\beta}) \quad k = 1, ..., K, \tag{1}$$

where $\boldsymbol{\beta}$ is a finite vector of parameters. Since the distribution of $\boldsymbol{\xi}$ is necessarily multinomial, correct specification of (1) is equivalent to correct specification of the functional forms for $F_k(.; \boldsymbol{\beta})$.

There are two main categories of logit-type models for polytomous unordered selection:

1. *Conditional logit models* in which the probabilities depend on the choices' characteristics (for example, travel costs for transportation mode choice), but their effects are invariant across alternatives, so that $\boldsymbol{\beta}_k = \boldsymbol{\beta}\ \forall k$.
2. *Multinomial logit models* in which the probabilities depend on the choosers' characteristics (for example, education, age and gender for occupational choice), which are invariant across choices, while their effects are captured by $\boldsymbol{\beta}'_k s$ that vary across alternatives.

We focus on the latter because they are also popular in switching regime models for time series. Thus, we complement Mai, Frejinger and Bastin (2015), who apply the IM test to a variant of the conditional logit model for transportation mode choice originally introduced by McFadden (1974).

The rest of the note is organised as follows. We derive our theoretical results in Section 2 and report the Monte Carlo exercises that look at the finite sample size and power of the test in Section 3. Finally, we conclude by discussing some avenues for further research, relegating proofs and details about our simulations to supplemental appendices.

2 Theoretical results

Consider the following parametrisation of the conditional probabilities in (1):

$$F_k(\mathbf{z};\boldsymbol{\beta}) = \frac{e^{\boldsymbol{\beta}'_k\mathbf{z}}}{\sum_{\ell=1}^{K} e^{\boldsymbol{\beta}'_\ell\mathbf{z}}}, \quad k = 1,\ldots,K, \tag{2}$$

where $\boldsymbol{\beta} = (\boldsymbol{\beta}'_1,\ldots,\boldsymbol{\beta}'_K)'$ collects the coefficient vectors. Naturally, $\sum_{k=1}^{K} p_k(\mathbf{z};\boldsymbol{\beta}) = 1$ for all $\mathbf{z}$ and $\boldsymbol{\beta}$. For identification purposes, we follow the usual practice of setting $\boldsymbol{\beta}_1 = \mathbf{0}$, so that the first category becomes the baseline one, thereby eliminating L elements of the score vector, $\mathbf{s}(\boldsymbol{\beta})$, and $L(L+1)/2$ of the Hessian matrix, $\mathbf{h}(\boldsymbol{\beta})$, without loss of generality because the ordering of the categories is arbitrary. In this respect, Lemma 1 in Amengual, Fiorentini and Sentana (2024) implies that the IM test is numerically invariant to reparametrisations.

Let $\mathbf{p}_r(\mathbf{z};\boldsymbol{\beta}) = [p_2(\mathbf{z};\boldsymbol{\beta}),\ldots,p_K(\mathbf{z};\boldsymbol{\beta})]'$ represent the vector of conditional probabilities of the non-normalised categories, and $\mathbf{u}_r(\boldsymbol{\xi}_r,\mathbf{z};\boldsymbol{\beta}) = [u_2(\xi_2,\mathbf{z};\boldsymbol{\beta}),\ldots,u_K(\xi_K,\mathbf{z};\boldsymbol{\beta})]' = \boldsymbol{\xi}_r - \mathbf{p}_r(\mathbf{z};\boldsymbol{\beta})$, with $\boldsymbol{\xi}_r = (\xi_2,\ldots,\xi_K)'$, the corresponding vector of what Gouriéroux, Monfort, Renault and Trognon (1987) called generalised residuals by analogy to OLS regressions. Finally, let $\hat{\boldsymbol{\beta}}_N = (\mathbf{0}',\hat{\boldsymbol{\beta}}'_{2N},\ldots,\hat{\boldsymbol{\beta}}'_{KN})' = (\mathbf{0}',\hat{\boldsymbol{\beta}}'_{rN})$ denote the ML estimator. Then, we can show that:

Proposition 1 *1) The score vector and Hessian matrix of model (2) are given by*

$$\mathbf{s}_r(\boldsymbol{\xi},\mathbf{z};\boldsymbol{\beta}) = \mathbf{u}_r(\boldsymbol{\xi},\mathbf{z};\boldsymbol{\beta}) \otimes \mathbf{z}, \tag{3}$$

$$\mathbf{h}_r(\boldsymbol{\xi},\mathbf{z};\boldsymbol{\beta}) = -\{diag[\mathbf{p}_r(\mathbf{z};\boldsymbol{\beta})] - \mathbf{p}_r(\mathbf{z};\boldsymbol{\beta})\mathbf{p}'_r(\mathbf{z};\boldsymbol{\beta})\} \otimes \mathbf{z}\mathbf{z}', \tag{4}$$

respectively, so that the IM influence functions are:

$$\mathbf{m}_r(\boldsymbol{\xi},\mathbf{z};\boldsymbol{\beta}) = vech[\mathbf{u}_r(\boldsymbol{\xi},\mathbf{z};\boldsymbol{\beta})\mathbf{u}'_r(\boldsymbol{\xi},\mathbf{z};\boldsymbol{\beta}) - \{diag[\mathbf{p}_r(\mathbf{z};\boldsymbol{\beta})] - \mathbf{p}_r(\mathbf{z};\boldsymbol{\beta})\mathbf{p}'_r(\mathbf{z};\boldsymbol{\beta})\}] \otimes vech(\mathbf{z}\mathbf{z}'). \tag{5}$$

2) Let $\overline{\mathbf{m}}_{rN}(\hat{\boldsymbol{\beta}}_N)$ denote the sample mean of $\mathbf{m}_r(\boldsymbol{\xi},\mathbf{z};\boldsymbol{\beta})$ evaluated at $\hat{\boldsymbol{\beta}}_N$, and define

$$\begin{bmatrix} \mathcal{R}(\boldsymbol{\beta}) & \mathcal{U}(\boldsymbol{\beta}) \\ \mathcal{U}'(\boldsymbol{\beta}) & \mathcal{I}(\boldsymbol{\beta}) \end{bmatrix} = V\left\{ \begin{array}{c} \mathbf{m}_r(\boldsymbol{\xi}_r,\mathbf{z};\boldsymbol{\beta}) \\ \mathbf{s}_r(\boldsymbol{\xi}_r,\mathbf{z};\boldsymbol{\beta}) \end{array} \right\}. \tag{6}$$

Then, under correct specification, the IM test statistic

$$N \times \overline{\mathbf{m}}'_{rN}(\hat{\boldsymbol{\beta}}_N)[\mathcal{R}(\boldsymbol{\beta}_0) - \mathcal{U}(\boldsymbol{\beta}_0)\mathcal{I}^{-1}(\boldsymbol{\beta}_0)\mathcal{U}(\boldsymbol{\beta}_0)]^{-1}\overline{\mathbf{m}}_{rN}(\hat{\boldsymbol{\beta}}_N) \xrightarrow{d} \chi^2_{K(K-1)L(L+1)/4}. \tag{7}$$

If following Newey (1985) and Tauchen (1985) we regard the IM test as a moment test of the influence functions (5), it is clear that it is effectively testing the conditional mean independence of the conditionally demeaned outer product of the generalised residuals. Thus, it resembles a multivariate version of White's (1980) test for residual conditional heteroskedasticity, which in turn confirms Chesher's (1984) reinterpretation of the IM test as a score test for neglected unobserved heterogeneity.

One feasible version of IM test statistic (7) replaces the elements of (6) by their sample counterparts evaluated at $\hat{\boldsymbol{\beta}}_N$, which Chesher (1983) and Lancaster (1984) showed is numerically identical to NR^2 in the regression of 1 on $\mathbf{m}_r(\boldsymbol{\xi}_r, \mathbf{z}; \hat{\boldsymbol{\beta}}_N)$ and $\mathbf{s}_r(\boldsymbol{\xi}_r, \mathbf{z}; \hat{\boldsymbol{\beta}}_N)$. Given that this yields very noisy estimators of (6), we propose another feasible version of the IM test that evaluates the different elements of (6) by relying on the law of iterated expectations, with $\boldsymbol{\beta}$ replaced by $\hat{\boldsymbol{\beta}}_N$ and unconditional expectations by sample averages. Our next result provides analytical expressions for the required conditional moments:

Proposition 2 *a) The relevant conditional variances and covariances required to compute $\mathcal{R}$ are:*

$$cov(\mathbf{m}_{j\ell}, \mathbf{m}_{j'\ell}) = E[E(m_{j'\ell} m_{j'\ell}|\mathbf{z}) vech(\mathbf{z}\mathbf{z}') vech'(\mathbf{z}\mathbf{z}')], \quad \text{where}$$

$$E(m_{jj}^2|\mathbf{z}) = p_j - 5p_j^2 + 8p_j^3 - 4p_j^4, \quad E(m_{j\ell}^2|\mathbf{z}) = p_j^2 p_\ell + p_j p_\ell^2 - 4p_j^2 p_\ell^2,$$

$$E(m_{jj} m_{j'j'}|\mathbf{z}) = -p_j p_{j'} + 2p_j^2 p_{j'} + 2p_j p_{j'}^2 - 4p_j^2 p_{j'}^2, \quad E(m_{jj} m_{j\ell}|\mathbf{z}) = -p_j p_\ell + 4p_j^2 p_\ell - 4p_j^3 p_\ell,$$

$$E(m_{jj} m_{j'\ell}|\mathbf{z}) = 2p_j p_{j'} p_\ell - 4p_j^2 p_{j'} p_\ell, \quad E(m_{j\ell} m_{j'\ell}|\mathbf{z}) = p_\ell p_j p_{j'} - 4p_\ell^2 p_j p_{j'}$$

and $E(m_{j\ell} m_{j'\ell'}|\mathbf{z}) = -4p_j p_\ell p_{j'} p_{\ell'}$.

b) In turn, the relevant conditional covariances required to compute $\mathcal{U}$ are:

$$E(\mathbf{m}_{j\ell}\mathbf{s}'_{j'}) = cov(\mathbf{m}_{j\ell}, \mathbf{s}_{j'}) = E[E(\mathbf{m}_{j\ell} u_j|\mathbf{z}) vech(\mathbf{z}\mathbf{z}')\mathbf{z}'], \quad \text{where}$$

$$E(m_{jj} u_j|\mathbf{z}) = p_j - 3p_j^2 + 2p_j^3, \qquad E(m_{jj} u_{j'}|\mathbf{z}) = -p_j p_{j'} + 2p_j^2 p_{j'},$$
$$E(m_{j\ell} u_j|\mathbf{z}) = -p_j p_\ell + 2p_j^2 p_\ell \quad \text{and} \quad E(m_{j\ell} u_{j'}|\mathbf{z}) = 2p_j p_\ell p_{j'}.$$

c) Finally, the information matrix is

$$\mathcal{I} = E\{[diag(\mathbf{p}_r) - \mathbf{p}_r \mathbf{p}'_r] \otimes \mathbf{z}\mathbf{z}'\}. \tag{8}$$

It is important to mention that the IM test cannot be computed when the only regressor is a constant because in that case the score simplifies to $\mathbf{u}_r$ and the influence functions underlying the IM test have zero mean in the sample when evaluated at $\hat{\boldsymbol{\beta}}_N$. The same situation arises when the explanatory variables consist of an exhaustive set of dummies that in practice generate a partition of the observations because the coefficients of those dummies effectively correspond to a model which imposes that the probabilities are constant within each category but heterogeneous across categories. In both these cases, the multinomial logit model provides a perfect fit to the data. Nevertheless, as soon as at least one of the elements of $\mathbf{z}$ is a continuous random variable, the IM test can be computed.[1]

Composite likelihood: A well-known property of multinomial logit models is that they continue to represent the relative probabilities of any subset of categories for those observations belonging to them. In particular, if we focus on the first and second categories only, we will end up with the following binary logit model:

$$p_2^b(\mathbf{z}; \boldsymbol{\beta}_2) = \Pr(\xi_2 = 1|\mathbf{z}) = \frac{e^{\boldsymbol{\beta}'_2\mathbf{z}}}{1 + e^{\boldsymbol{\beta}'_2\mathbf{z}}} = p_2(\mathbf{z}; \boldsymbol{\beta}_2) \cdot \frac{1 + \sum_{\ell=2}^{K} e^{\boldsymbol{\beta}'_\ell\mathbf{z}}}{1 + e^{\boldsymbol{\beta}'_2\mathbf{z}}}$$

[1] The number of degrees of freedom might need to be adjusted in very special circumstances. For example, in a binary logit model with a single continuous explanatory variable, the IM test statistic will generally be distributed as a χ_1^2 when the slope coefficient is actually 0.

with the identification condition $\boldsymbol{\beta}_1 = \mathbf{0}$. Since this is true for any two categories, a popular consistent estimation method for multinomial logit models obtains $\boldsymbol{\beta}_j$ from $K-1$ such binary logit models, in what is effectively a composite likelihood approach (see Lindsay (1988)). This yields computational gains at the cost of asymptotic efficiency. Nevertheless, the results in Proposition 1 apply to each of those conditional binary logit models as well, with the number of degrees of freedom becoming $L(L+1)/2$. For that reason, in Section 3 we will study these binary IM tests too.

Unfortunately, the relationship between the IM test for the full model and the $K-1$ IM tests for the binary models is not straightforward because they are based on different subsets of observations. However, they all maintain not only the same distribution for the underlying choice shocks but also the independence of irrelevant alternatives assumption, which is precisely what guarantees the validity of the binary models.

3 Monte Carlo simulations

The asymptotic distribution of the IM test might not be very reliable in small samples. For that reason, we study its size and power properties in simulated samples of length $N = 125$, $N = 500$ and $N = 2,000$. To estimate the parameters for binary and multinomial logit models, we make use of the MATLAB toolbox available at https://www.spatial-econometrics.com/ (see LeSage and Pace (2009)).

3.1 Size properties

When assessing size, we generate $10,000$ samples under the null for each data generating process (DGP) we describe below. We then compare two asymptotically equivalent versions of the infeasible IM test statistic in (7): the Outer-Product-of-the Score version proposed by Chesher (1983) and Lancaster (1984) (OPS), and one that replaces the true parameter values $\boldsymbol{\beta}_0$ with their MLEs $\hat{\boldsymbol{\beta}}_N$ in the theoretical expressions of the conditional variances and covariances in Proposition 2 (CM). In all cases, we consider not only asymptotic critical values but also a parametric bootstrap procedure in which we simulate $B = 99$ samples from the mixture model estimated under the null, as proposed by Horowitz (1994).[2]

We simulate multinomial logit models with $K = 3$ and $K = 5$ categories, always including a constant and one or two continuous regressors. Details on the specific designs can be found in Supplemental Appendix B.1. Table 1 contains the rejection rates of the multinomial IM tests at the 1%, 5% and 10% significance levels. Panels A and B refer to models with three categories,

[2] Horowitz (1994) found that increasing the number of bootstrap samples beyond 99 had little effect on the results of his experiments.

with two and three explanatory variables, respectively, while Panels C and D to models with five categories.

The rejection rates using asymptotic critical values in the left subpanels of Table 1 confirm the need for finite sample size adjustments, especially for the OPS version of the IM test.[3] Still, the quality of the asymptotic approximation is much better when we use the theoretical expressions for the weighting matrix even in samples of size $N = 500$, although there is still a systematic overrejection of the null at the 1% level.

In contrast, the bootstrap-based rejection rates in the right subpanels of Table 1 give a completely different picture: sizes are very accurate and almost all Monte Carlo rejection rates fall within the relevant 95% confidence set, with the exceptions of the OPS version for $N = 125$ and $N = 500$, and the CM version when $N = 125$ in models with five categories (Panels C and D).

In Table A1 in the supplementary material we report the same figures but for the conditional binary logits mentioned at the end of section 2.[4] Not surprisingly, there is still massive overrejection of the OPS version of the tests that rely on asymptotic critical values. Interestingly, though, the overrejections of the CM test at the 1% level are more moderate, probably due to the smaller number of degrees of freedom of their asymptotic distribution. In any event, the parametric bootstrap corrects the size distortions for all the sample sizes we consider.

3.2 Power properties

We consider four types of alternatives. Given that an important source of misspecification in many econometric models are omitted variables, we first simulate data from a model with $L + 1$ explanatory variables in which the variance of the additional one is proportional to the square of the L^{th} included variable in the estimated model. In turn, the neglected heterogeneity interpretation of the IM test provides the motivation for our next two alternatives. Specifically, we consider a model in which the coefficients for one of the z's take different values in two equally sized subgroups of the population, while remaining homogeneous within subgroups. In addition, we consider another model in which the coefficients for one of the z's are randomly distributed as a multivariate Gaussian vector across individuals. Finally, we generate data from an ordered logit model as an example of misspecification of the functional form F. In this respect, it is important to emphasise that the so-called "parallel lines" assumption of the ordered logit model implies that if we lump together all the categories below and above any given threshold, we will obtain a binary logit model, in marked contrast to the multinomial logit model (2), in which the binary

[3] Given the number of replications, the 95% asymptotic confidence intervals for the Monte Carlo rejection probabilities under the null are (0.80,1.20), (4.57,5.43) and (9.41,10.59) at the 1%, 5% and 10% levels.

[4] The corresponding results for models with five categories are available upon request.

logit models apply to any two categories after suppressing the remaining $K-2$ ones. Again, Supplemental Appendix B.1 contains the details on the specific designs.

We simulate 2,500 samples for each of these alternatives. Given our results in the previous subsection, we take an accept/reject decision by systematically relying on the bootstrap CM version of the IM test statistic, thereby ensuring that we carry out a feasible size adjustment.

In Panels A to D of Table 2 we report the results for DGP a to DGP d. As expected, power increases with the sample size N. In contrast, no clear pattern arises when increasing the number of explanatory variables. In particular, power seems to increase for DGP b and DGP d, decrease for DGP c, and present mixed patterns for DGP a. The same comment applies when we move from three to five categories.

Finally, Table A2 in Supplemental Appendix B.2 reports the same figures for the three binary logits implied by the models with three categories. As expected, the same pattern is obtained. More importantly, the IM test of the multinomial logit model is more powerful than the binary tests.

4 Extensions

The IM tests in this paper can be extended in at least three empirically relevant directions. First, we could consider discrete Markov chains in which each column of the $K \times K$ transition matrix is a multinomial logit function of the explanatory variables $\mathbf{z}$. Given that a Markov chain is a collection of K separate multinomial logit models indexed by the value taken by the preceding multinomial variable $\boldsymbol{\xi}$ with coefficients which are variation-free, the IM influence functions will be the collection of IM influence functions for each of those K multinomial models. Second, we could study mixture models and switching regression models in which the probabilities of the mixture components or regimes are determined by another multinomial logit model. Given that the multinomial variable $\boldsymbol{\xi}$ becomes latent in those circumstances, as in Amengual, Fiorentini and Sentana (2024), we would need to compute the conditional expected values of the outer product of the generalised residuals given the observable variables to obtain the IM test. Finally, we could combine the previous two extensions in a switching regime model in which the regimes follow a Markovian structure, as in Hamilton (1989), which would force us to rely on a smoother rather than a filter, as in Almuzara, Amengual and Sentana (2019). We are currently pursuing these interesting research avenues.

References

Almuzara, T., Amengual, D. and Sentana, E. (2019): "Normality tests for latent variables", *Quantitative Economics* 10, 981-1017.

Amengual, D., Fiorentini, G. and Sentana, E. (2024): "Information matrix tests for Gaussian mixtures", CEMFI Working Paper 2401.

Chesher, A. (1983): "The information matrix test: simplified calculation via a score test interpretation", *Economics Letters* 13, 45-48.

Chesher, A. (1984): "Testing for neglected heterogeneity", *Econometrica* 52, 865-872.

Gouriéroux, C., Monfort, A., Renault, E. and Trognon, A. (1987): "Generalized residuals", *Journal of Econometrics* 34, 5-32.

Hamilton, J.D. (1989), "A new approach to the economic analysis of nonstationary time series and the business cycle", *Econometrica* 57, 357-384.

Horowitz, J. (1994): "Bootstrap-based critical values for the information matrix test", *Journal of Econometrics* 61, 395-411.

Lancaster, A. (1984): "The covariance matrix of the information matrix test", *Econometrica* 52, 1051-1053.

LeSage, J. and Pace, R.K. (2009): *Introduction to spatial econometrics*, CRC Press.

Lindsay, B. (1988): "Composite likelihood methods", *Contemporary Mathematics* 80, 221-240.

Mai, T., Frejinger, E. and Bastin, F. (2015): "A misspecification test for logit based route choice models", *Economics of Transportation* 4, 215-226.

McFadden, D. (1974): "Conditional logit analysis of qualitative choice behavior", In P. Zarembka (Ed.), *Frontiers in econometrics*, pp. 104-142, Academic Press.

Newey, W.K. (1985): "Maximum likelihood specification testing and conditional moment tests", *Econometrica* 53, 1047-70.

Tauchen, G. (1985): "Diagnostic testing and evaluation of maximum likelihood models", *Journal of Econometrics* 30, 415-443.

White, H. (1980): "A heteroscedasticity-consistent covariance matrix estimator and a direct test for heteroscedasticity", *Econometrica* 48, 817-838.

White, H. (1982): "Maximum likelihood estimation of misspecified models", *Econometrica* 50, 1-25.

Table 1: Size properties: Multinomial IM tests

Panel A: Three categories, two explanatory variables: $\mathbf{z} = (1, z)'$ with $z \sim i.i.d.\ N(0,1)$

	Asymptotic critical values							Bootstraped critical values					
	OPS			CM				OPS			CM		
Sample size	10%	5%	1%	10%	5%	1%	Sample size	10%	5%	1%	10%	5%	1%
125	97.58	96.01	91.05	8.40	6.37	4.14	125	7.50	3.27	0.48	9.93	5.24	1.20
500	84.41	80.29	71.32	10.32	7.44	4.25	500	8.81	4.37	0.69	10.05	5.09	1.03
2,000	57.17	50.69	39.77	10.99	7.09	3.08	2,000	10.03	4.81	0.95	10.05	5.14	1.04

Panel B: Three categories, three explanatory variables: $\mathbf{z} = (1, z_1, z_2)'$ with $(z_1, z_2) \sim i.i.d.\ N(\mathbf{0}, \mathbf{I}_2)$

	Asymptotic critical values							Bootstraped critical values					
	OPS			CM				OPS			CM		
Sample size	10%	5%	1%	10%	5%	1%	Sample size	10%	5%	1%	10%	5%	1%
125	99.95	99.87	99.34	9.79	7.44	4.40	125	8.80	4.23	0.74	9.59	4.78	1.01
500	95.96	94.16	89.97	11.54	7.72	3.91	500	9.30	4.48	0.80	9.62	4.69	0.88
2,000	69.61	63.01	50.96	11.68	7.36	2.94	2,000	10.09	4.91	0.95	9.81	4.95	0.96

Panel C: Five categories, two explanatory variables: $\mathbf{z} = (1, z)'$ with $z \sim i.i.d.\ N(0,1)$

	Asymptotic critical values							Bootstraped critical values					
	OPS			CM				OPS			CM		
Sample size	10%	5%	1%	10%	5%	1%	Sample size	10%	5%	1%	10%	5%	1%
125	99.84	99.84	99.84	8.93	7.42	5.45	125	3.77	1.21	0.12	6.81	2.61	0.32
500	100.00	100.00	100.00	9.31	7.68	5.61	500	6.11	2.46	0.17	9.65	5.00	1.11
2,000	99.98	99.98	99.98	10.43	8.44	5.60	2,000	10.19	5.51	0.97	9.77	5.16	1.11

Panel D: Five categories, three explanatory variables: $\mathbf{z} = (1, z_1, z_2)'$ with $(z_1, z_2) \sim i.i.d.\ N(\mathbf{0}, \mathbf{I}_2)$

	Asymptotic critical values							Bootstraped critical values					
	OPS			CM				OPS			CM		
Sample size	10%	5%	1%	10%	5%	1%	Sample size	10%	5%	1%	10%	5%	1%
125	99.99	99.99	99.99	9.32	7.57	5.40	125	14.84	6.39	0.64	10.31	5.05	0.94
500	100.00	100.00	100.00	12.19	10.03	7.07	500	16.93	9.01	1.70	9.98	5.11	1.24
2,000	100.00	100.00	100.00	13.72	10.76	6.75	2,000	15.91	8.66	1.58	10.16	5.12	1.06

Notes: Monte Carlo rejection rates based on 10,000 replications. OPS refers to the version of the statistic proposed by Chesher (1983) and Lancaster (1984), while CM to the feasible version that makes use of the theoretical expressions in Proposition 2 replacing the true parameter values by their MLEs and unconditional expectations by sample averages. Rejection rates in the right subpanels are based on the asymptotic distribution in Proposition 1 while the left ones on a parametric bootstrap procedure in which we simulate $B = 99$ samples from the model estimated under the null. See Supplemental Appendix 3 for details about the DGPs.

Table 2: Multinomial IM tests: Power properties

Panel A: Alternative hypothesis: Omitted variable

	Three categories							Five categories					
	$L=2$			$L=3$				$L=2$			$L=3$		
Sample size	10%	5%	1%	10%	5%	1%	Sample size	10%	5%	1%	10%	5%	1%
125	17.12	9.60	3.00	22.32	8.72	0.56	125	18.60	6.76	0.44	15.04	5.84	0.56
500	85.40	71.96	27.08	85.40	80.20	43.64	500	66.36	45.04	13.16	87.40	57.48	6.96
2,000	99.24	97.00	64.68	100.00	100.00	96.48	2,000	96.84	83.64	25.96	100.00	100.00	94.16

Panel B: Alternative hypothesis: Group heterogeneity in β_{i2}

	Three categories							Five categories					
	$L=2$			$L=3$				$L=2$			$L=3$		
Sample size	10%	5%	1%	10%	5%	1%	Sample size	10%	5%	1%	10%	5%	1%
125	29.32	18.16	4.36	32.64	18.88	4.84	125	21.04	8.12	1.00	20.96	9.28	1.32
500	45.40	26.44	5.48	55.64	38.72	10.92	500	67.96	38.92	6.80	98.00	83.64	20.40
2,000	74.32	58.76	20.48	97.52	94.28	58.60	2,000	90.88	71.20	21.96	100.00	100.00	97.40

Panel C: Alternative hypothesis: Heterogeneous Gaussian β_{i2}

	Three categories							Five categories					
	$L=2$			$L=3$				$L=2$			$L=3$		
Sample size	10%	5%	1%	10%	5%	1%	Sample size	10%	5%	1%	10%	5%	1%
125	63.16	40.52	6.48	63.36	40.40	7.80	125	9.96	3.44	0.52	16.24	7.24	1.32
500	99.48	98.16	59.76	80.88	71.72	38.88	500	41.80	26.16	8.96	42.96	25.40	5.80
2,000	100.00	100.00	99.80	100.00	100.00	93.96	2,000	99.84	98.44	95.52	99.96	99.72	71.92

Panel D: Alternative hypothesis: Ordered logit

	Three categories							Five categories					
	$L=2$			$L=3$				$L=2$			$L=3$		
Sample size	10%	5%	1%	10%	5%	1%	Sample size	10%	5%	1%	10%	5%	1%
125	19.08	11.84	3.60	17.40	10.24	3.00	125	26.88	16.36	4.48	26.56	16.64	6.04
500	28.80	18.32	5.44	27.04	16.36	4.80	500	48.44	34.00	12.52	46.08	32.16	12.56
2,000	59.20	45.56	18.12	57.32	44.52	17.36	2,000	83.96	69.00	31.32	87.56	74.28	34.80

Notes: Monte Carlo rejection rates based on 2,500 replications. Results for the feasible version of the IM test that makes use of the theoretical expressions in Proposition 2 replacing the true parameter values by their MLEs and unconditional expectations by sample averages. Rejection rates are based on a parametric bootstrap procedure in which we simulate $B = 99$ samples from the model estimated under the null. See Supplemental Appendix 3 for details about the DGPs.

Supplemental Appendices for

Information matrix tests for multinomial logit models

Dante Amengual

CEMFI, Casado del Alisal 5, E-28014 Madrid, Spain

<amengual@cemfi.es>

Gabriele Fiorentini

Università di Firenze and RCEA, Viale Morgagni 59, I-50134 Firenze, Italy

<gabriele.fiorentini@unifi.it>

Enrique Sentana

CEMFI, Casado del Alisal 5, E-28014 Madrid, Spain

<sentana@cemfi.es>

April 2024

A Proofs

A.1 Proof or Proposition 1

Note that

$$\frac{\partial p_j(\mathbf{z};\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_j} = \frac{1}{\left(\sum_{\ell=1}^K e^{\boldsymbol{\beta}'_\ell \mathbf{z}}\right)^2}\left[e^{\boldsymbol{\beta}'_j \mathbf{z}}\left(\sum_{\ell=1}^K e^{\boldsymbol{\beta}'_\ell \mathbf{z}}\right) - e^{2\boldsymbol{\beta}'_j \mathbf{z}}\right]\mathbf{z} = p_j(\mathbf{z};\boldsymbol{\beta})[1-p_j(\mathbf{z};\boldsymbol{\beta})]\mathbf{z},$$

while

$$\frac{\partial p_k(\mathbf{z};\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_j} = \frac{-e^{\boldsymbol{\beta}'_k \mathbf{z}} e^{\boldsymbol{\beta}'_j \mathbf{z}}}{\left(\sum_{s=1}^K e^{\boldsymbol{\beta}'_s \mathbf{z}}\right)^2}\mathbf{z} = -p_k(\mathbf{z};\boldsymbol{\beta})p_j(\mathbf{z};\boldsymbol{\beta})\mathbf{z}$$

when $k \neq j$. Interestingly, these expressions coincide with $\mathbf{z}$ times the conditional variance of ξ_j given $\mathbf{z}$ and the conditional covariance between ξ_j and ξ_k given $\mathbf{z}$, respectively.

To derive the score, it is convenient to re-write both expressions together as

$$\frac{\partial p_k(\mathbf{z};\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_j} = p_k(\mathbf{z};\boldsymbol{\beta})[I(j=k) - p_j(\mathbf{z};\boldsymbol{\beta})]\mathbf{z},$$

where $I(.)$ is the usual indicator function. The contribution of a single observation to the log-likelihood function (ignoring constant terms) will be

$$\ln f(\boldsymbol{\xi},\mathbf{z};\boldsymbol{\beta}) = \textstyle\sum_{k=1}^K \xi_k \ln p_k(\mathbf{z};\boldsymbol{\beta}).$$

Hence, the score with respect to $\boldsymbol{\beta}_j$ $(k = 2,\ldots,K)$ will be given by

$$\mathbf{s}_j(\boldsymbol{\xi},\mathbf{z};\boldsymbol{\beta}) = \frac{\partial \ln f(\boldsymbol{\xi},\mathbf{z};\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_j} = \textstyle\sum_{k=1}^K \xi_k[I(j=k) - p_j(\mathbf{z};\boldsymbol{\beta})]\mathbf{z} = [\xi_j - p_j(\mathbf{z};\boldsymbol{\beta})]\mathbf{z} = u_j(\xi_j,\mathbf{z};\boldsymbol{\beta})\mathbf{z},$$

where $u_j(\xi_j,\mathbf{z};\boldsymbol{\beta}) = \xi_j - p_j(\mathbf{z};\boldsymbol{\beta})$. Thus, we can write the first-order conditions together as (3). From here, the second derivatives will be

$$\begin{aligned}\mathbf{h}_{jj}(\boldsymbol{\xi},\mathbf{z};\boldsymbol{\beta}) &= \frac{\partial^2 \ln f(\boldsymbol{\xi},\mathbf{z};\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_j \partial \boldsymbol{\beta}'_j} = -p_j(\mathbf{z};\boldsymbol{\beta})[1-p_j(\mathbf{z};\boldsymbol{\beta})]\mathbf{z}\mathbf{z}' \quad \text{and} \\ \mathbf{h}_{j\ell}(\boldsymbol{\xi},\mathbf{z};\boldsymbol{\beta}) &= \frac{\partial^2 \ln f(\boldsymbol{\xi},\mathbf{z};\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_j \partial \boldsymbol{\beta}'_\ell} = p_j(\mathbf{z};\boldsymbol{\beta})p_\ell(\mathbf{z};\boldsymbol{\beta})\mathbf{z}\mathbf{z}',\end{aligned}$$

whence (4) follows. Therefore, we will have that

$$\begin{aligned}\frac{\partial \ln f(\boldsymbol{\xi},\mathbf{z};\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_j}\frac{\partial \ln f(\boldsymbol{\xi},\mathbf{z};\boldsymbol{\beta})}{\partial \boldsymbol{\beta}'_j} + \frac{\partial \ln f(\boldsymbol{\xi},\mathbf{z};\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_j \partial \boldsymbol{\beta}'_j} &= \{u_j^2(\xi_j,\mathbf{z};\boldsymbol{\beta}) - p_j(\mathbf{z};\boldsymbol{\beta})[1-p_j(\mathbf{z};\boldsymbol{\beta})]\}\mathbf{z}\mathbf{z}' \text{ while} \\ \frac{\partial \ln f(\boldsymbol{\xi},\mathbf{z};\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_j}\frac{\partial \ln f_i(\boldsymbol{\xi},\mathbf{z};\boldsymbol{\beta})}{\partial \boldsymbol{\beta}'_\ell} + \frac{\partial \ln f(\boldsymbol{\xi},\mathbf{z};\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_j \partial \boldsymbol{\beta}'_\ell} &= [u_j(\xi_j,\mathbf{z};\boldsymbol{\beta})u_\ell(\xi_\ell,\mathbf{z};\boldsymbol{\beta}) + p_j(\mathbf{z};\boldsymbol{\beta})p_\ell(\mathbf{z};\boldsymbol{\beta})]\mathbf{z}\mathbf{z}',\end{aligned}$$

where we have used the fact that $\xi_j^2 = \xi_j$ and $\xi_j \xi_\ell = 0$. Therefore, we can write the influence functions corresponding to the information matrix equality in matrix notation as (5), the advantage of using of *vech* instead of *vec* being that we easily eliminate the duplicated influence functions that appear in (4) and the outer product of (3), thereby avoiding generalised inverses and providing the right number of degrees of freedom.

The reminder statements in the second part of the proposition follow directly from Chesher (1983) and Lancaster (1984) given the *i.i.d.* nature of the sample. □

A.2 Proof of Proposition 2

We can expand the quantities that appear in $cov(\mathbf{m}_{j\ell}, \mathbf{m}_{j'\ell})$ as

$$
\begin{aligned}
E(m_{jj}^2|\mathbf{z}) &= E\{[u_j^2 - p_{ji}(1-p_{ji})]^2|\mathbf{z}\} \;=\; E(u_j^4|\mathbf{z}) - 2p_j(1-p_j)E(u_j^2|\mathbf{z}) + p_j^2(1-p_j)^2, \\
E(m_{j\ell}^2|\mathbf{z}) &= E[(u_j u_\ell + p_j p_\ell)^2|\mathbf{z}] \;=\; E(u_j^2 u_\ell^2|\mathbf{z}) - 2p_j p_\ell E(u_j u_\ell|\mathbf{z}) + p_j^2 p_\ell^2, \\
E(m_{jj} m_{j'j'}|\mathbf{z}) &= E\{[u_j^2 - p_j(1-p_j)][u_{j'}^2 - p_{j'}(1-p_{j'})]|\mathbf{z}\} \\
&= E(u_j^2 u_{j'}^2|\mathbf{z}) - p_j(1-p_j)E(u_{j'}^2|\mathbf{z}) - p_{j'}(1-p_{j'})E(u_j^2|\mathbf{z}) + p_j(1-p_j)p_{j'}(1-p_{j'}), \\
E(m_{jj} m_{j\ell}|\mathbf{z}) &= E\{[(u_j^2 - p_j(1-p_j)](u_j u_\ell + p_j p_\ell)|\mathbf{z}\} \\
&= E(u_j^3 u_\ell|\mathbf{z}) + p_j p_\ell E(u_j^2|\mathbf{z}) - p_j(1-p_j)E(u_j u_\ell|\mathbf{z}) - p_j^2(1-p_j)p_\ell, \\
E(m_{jj} m_{j'\ell}|\mathbf{w}) &= E\{[u_j^2 - p_{ji}(1-p_{ji})](u_{j'} u_\ell + p_{j'} p_\ell)|\mathbf{z}\} \\
&= E(u_j^2 u_{j'} u_\ell|\mathbf{z}) + p_{j'} p_\ell E(u_j^2|\mathbf{z}) - p_j(1-p_j)E(u_{j'} u_\ell|\mathbf{z})] - p_j(1-p_j)p_{j'} p_\ell, \\
E(m_{j\ell} m_{j'\ell}|\mathbf{z}) &= E(u_j u_\ell + p_j p_\ell)(u_{j'} u_\ell + p_{j'} p_\ell)|\mathbf{z}] \\
&= E(u_\ell^2 u_j u_{j'}|\mathbf{z}) + p_j p_\ell p_{j'} p_\ell + p_{j'} p_\ell E(u_j u_\ell|\mathbf{z})] + p_j p_\ell E(u_{j'} u_\ell|\mathbf{z}) \quad \text{and} \\
E(m_{j\ell} m_{j'\ell'}|\mathbf{z}) &= E[(u_j u_\ell + p_j p_\ell)(u_{j'} u_{\ell'} + p_{j'} p_{\ell'})|\mathbf{z}] \\
&= E(u_j u_\ell u_{j'} u_{\ell'}|\mathbf{z}) + p_j p_\ell p_{j'} p_{\ell'} + p_j p_\ell E(u_{j'} u_{\ell'}|\mathbf{z}) + p_{j'} p_{\ell'} E(u_j u_\ell|\mathbf{z}).
\end{aligned}
$$

Then, if we use the formulae for the fourth-order centered moments of the multinomial distribution in Ouimet (2021), namely

$$
\begin{aligned}
E(u_j^4) &= (1-p_j)p_j[1 - 3(1-p_j)p_j], \\
E(u_j^3 u_\ell) &= -p_j[1 - 3(1-p_j)p_j]p_\ell, \\
E(u_j^2 u_\ell^2) &= p_j p_\ell(p_j + p_\ell - 3p_j p_\ell), \\
E(u_j^2 u_{j'} u_\ell) &= p_j(1-p_j)p_{j'} p_\ell \quad \text{and} \\
E(u_j u_\ell u_{j'} u_{\ell'}) &= -3p_j p_\ell p_{j'} p_{\ell'},
\end{aligned}
$$

we obtain the expressions in part a) of the lemma.

Doing the same with the expressions entering in $cov(\mathbf{m}_{j\ell}, \mathbf{s}_{j'})$:

$$
\begin{aligned}
E(m_{jj} u_j|\mathbf{z}) &= E\{[u_j^2 - p_j(1-p_j)]u_j|\mathbf{w}\} \;=\; E(u_j^3|\mathbf{z}) - p_j(1-p_j)E(u_j|\mathbf{z}), \\
E(m_{jj} u_{j'}|\mathbf{z}) &= E\{[u_j^2 - p_j(1-p_j)]u_{j'}|\mathbf{z}\} \;=\; E(u_j^2 u_{j'}|\mathbf{z}) - p_j(1-p_j)E(u_{j'}|\mathbf{z}), \\
E(m_{j\ell} u_j|\mathbf{z}) &= E[(u_j u_\ell + p_j p_\ell)u_j|\mathbf{z}] \;=\; E(u_j^2 u_\ell|\mathbf{z}) + p_j p_\ell E(u_j|\mathbf{z}) \quad \text{and} \\
E(m_{j\ell} u_{j'}|\mathbf{z}) &= cov(m_{j\ell}, u_{j'}|\mathbf{z}) \;=\; E[(u_j u_\ell + p_j p_\ell)u_{j'}|\mathbf{z}] \;=\; E(u_j u_\ell u_{j'}|\mathbf{z}) + p_j p_\ell E(u_{j'}|\mathbf{z}),
\end{aligned}
$$

and using the formulae for the third-order centered moments of the multinomial distribution in

Ouimet (2021),

$$\begin{aligned} E(u_j^3) &= p_j(1-p_j)(1-2p_j), \\ E(u_j^2 u_\ell) &= p_j(1-2p_j)p_\ell \\ E(u_j u_\ell u_{j'}) &= 2p_j p_\ell p_{j'}, \end{aligned}$$

we obtain the expressions in part b) of the lemma.

Finally, the expression for the information matrix follows from its definition. □

B Monte Carlo simulations: design and additional results

B.1 Design

For each DGP, we always include an intercept and either one or two standard normal uncorrelated explanatory variables. Following Horowitz (1994), we keep the explanatory variables $\mathbf{z}_i$, $i = 1, \ldots, N$ fixed in repeated samples. Nevertheless, we minimise the effects of the specific draws of these regressors by using the standard normal quantile function to generate them inverting a grid of points equally spaced over the unit interval - from $1/(2N)$ to $1 - 1/(2N)$. In the case of two non-constant regressors, we randomly permute each of them separately to ensure their independence, and additionally conduct a Cholesky decomposition to make them exactly orthogonal in the sample.

More importantly, we choose the $\boldsymbol{\beta}'s$ so that in simulated samples of five million observations they provide roughly balanced frequencies across categories and reasonable values for the pseudo-R^2's proposed by Cragg and Uhler (1970) and McFadden (1974), which we denote as R^2_{CU} and R^2_{MF}, respectively. Specifically, we consider under the null:

DGP A $K=3$, $L=2$: We pick $\boldsymbol{\beta}_2=(-1,-2)'$ and $\boldsymbol{\beta}_3=(-1,2)'$ so that the average frequencies are 0.36, 0.32 and 0.32, with $R^2_{MF}=0.34$ and $R^2_{CU}=0.14$. As the coefficient sign does not alter the explanatory power of the z's, the two binary logits have $R^2_{MF}=0.26$ and $R^2_{CU}=0.15$.

DGP B $K=3$, $L=3$: We pick $\boldsymbol{\beta}_2=(-1,-2,2)'$ and $\boldsymbol{\beta}_3=(-1,2,-1)'$ so that the average frequencies are 0.28, 0.36 and 0.36, with $R^2_{MF}=0.45$ and $R^2_{CU}=0.21$, and $R^2_{MF}=0.35$ and $R^2_{CU}=0.21$ for the binary logits.

DGP C $K=5$, $L=2$: We pick $\boldsymbol{\beta}_2=(-1,-2)'$, $\boldsymbol{\beta}_3=(-1,2)'$, $\boldsymbol{\beta}_4=(-2,-4)'$ and $\boldsymbol{\beta}_5=(-2,4)'$ so that the average frequencies are 0.24, 0.14, 0.14, 0.24 and 0.24, with $R^2_{MF}=0.37$ and $R^2_{CU}=0.10$. Once again, the sign of the coefficient does not alter the explanatory power of the z's, so that the two binary logits involving (ξ_1,ξ_2) and (ξ_1,ξ_3) are such that $R^2_{MF}=0.15$ and $R^2_{CU}=0.08$, while those for (ξ_1,ξ_4) and (ξ_1,ξ_5) have $R^2_{MF}=0.51$ and $R^2_{CU}=0.34$.

DGP D $K=5,\ L=3$: We pick $\boldsymbol{\beta}_2=(-1,-2,2)'$, $\boldsymbol{\beta}_3=(-1,2,-2)'$, $\boldsymbol{\beta}_4=(-2,-4,4)'$ and $\boldsymbol{\beta}_5=(-2,4,-4)'$ so that the average frequencies are 0.18, 0.11, 0.11, 0.30 and 0.30, with $R^2_{MF}=0.47$ and $R^2_{CU}=0.16$. In turn, the two binary logits for (ξ_1,ξ_2) and (ξ_1,ξ_3) have $R^2_{MF}=0.18$ and $R^2_{CU}=0.10$, while those for (ξ_1,ξ_4) and (ξ_1,ξ_5) have $R^2_{MF}=0.59$ and $R^2_{CU}=0.43$.

As for the alternatives, we consider:

DGP *a* We simulate the omitted variable as $z_{L+1}=\epsilon\sqrt{|z_L|}$ with ϵ obtained by applying the standard normal quantile function to an equally spaced grid of points between $1/(2N)$ and $1-1/(2N)$, choosing $\boldsymbol{\beta}_2=(-1,-2)'$ and $\boldsymbol{\beta}_3=(1,-4)'$ for $L=2$ and $\boldsymbol{\beta}_2=(1,-1,-2,-2)'$ and $\boldsymbol{\beta}_3=(-1,1,-1,4)'$ for $L=3$.

DGP *b* For the second half of the sample, we replace the slopes of z_1 by -6 and 4 when $K=3$, and -4, 6, 4 and 0 when $K=5$.

DGP *c* We perturb the $K-1$ slopes of z_1 by 3ϵ, with ϵ obtained by the standard normal quantile function to a grid of points equally spaced ranging from $1/(2N)$ to $1-1/(2N)$.

DGP *d* We draw samples from an ordered logit model with $y^*=2z+\eta$ or $y^*=\sqrt{2}z_1+\sqrt{2}z_2+\eta$ depending on whether $L=2$ or $L=3$, with η distributed as a standard logistic, and thresholds -1 and 1 for $K=3$ and -2, $-\frac{1}{2}$, $\frac{1}{2}$ and 2 for $K=5$.

B.2 Additional results for the binary logit model

In Table A1 below we report the same figures as in Table 1 but for the binary logits for models with three categories. The results for models with five categories are available upon request. Not surprisingly, the same pattern is obtained regarding the massive overrejection of the OPS version of the test when relying on asymptotic critical values. Interestingly, the overrejection of the CM test at the 1% level becomes more moderate, likely due to small number of degrees of freedom of the corresponding asymptotic distribution, namely $L(L+1)/2$. Once again, the bootstrap corrects the size distortions for all the sample sizes we consider. Similarly, in Table A2 below we report the figures but for the binary logit models when there are three categories. As expected, the power figures indicate the same pattern as in Table 2, but with less power.

B.3 Additional references

Cragg, S. G. and Uhler, R. (1970): "The demand for automobiles", *Canadian Journal of Economics*, 3, 386–406.

Ouimet, F. (2021): "General formulas for the central and non-central moments of the multinomial distribution", *Stats* 4, 18–27.

Table A1: (Binary) logit IM tests: Size properties (for three categories)

Panel A: Two explanatory variables: $\mathbf{z} = (1, z)'$ with $z \sim i.i.d.\ N(0, 1)$

	Asymptotic critical values							Bootstraped critical values					
	OPS			CM				OPS			IM		
Sample size	10%	5%	1%	10%	5%	1%	Sample size	10%	5%	1%	10%	5%	1%
	(ξ_1, ξ_2)							(ξ_1, ξ_2)					
125	63.18	55.31	38.40	7.02	4.42	2.33	125	4.81	1.61	0.09	9.46	4.67	1.12
500	41.00	34.26	24.39	8.49	5.21	2.47	500	9.63	4.55	0.76	9.66	5.02	1.12
2,000	26.20	19.85	12.38	9.57	5.75	2.14	2,000	10.37	5.25	1.06	10.05	5.28	1.19
	(ξ_1, ξ_3)							(ξ_1, ξ_3)					
125	62.37	54.16	37.90	7.24	4.70	2.51	125	4.92	1.40	0.11	9.55	4.86	1.21
500	41.51	34.73	24.97	8.47	5.13	2.36	500	9.57	4.80	0.67	9.70	4.78	0.96
2,000	24.73	18.63	11.33	9.26	5.51	1.85	2,000	9.54	4.63	1.01	9.86	4.95	0.90

Panel B: Three explanatory variables: $\mathbf{z} = (1, z_1, z_2)'$ with $(z_1, z_2) \sim i.i.d.\ N(\mathbf{0}, \mathbf{I}_2)$

	Asymptotic critical values							Bootstraped critical values					
	OPS			CM				OPS			IM		
Sample size	10%	5%	1%	10%	5%	1%	Sample size	10%	5%	1%	10%	5%	1%
	(ξ_1, ξ_2)							(ξ_1, ξ_2)					
125	77.80	70.61	53.37	7.51	4.68	2.25	125	3.37	1.00	0.11	9.44	4.62	0.96
500	58.54	51.25	39.26	9.45	5.64	2.47	500	9.63	4.37	0.63	10.11	5.18	1.04
2,000	32.10	26.48	16.46	9.44	5.48	1.88	2,000	10.01	5.31	1.01	9.29	4.67	0.99
	(ξ_1, ξ_3)							(ξ_1, ξ_3)					
125	86.94	80.87	63.79	8.47	5.54	2.58	125	4.16	1.23	0.07	10.37	5.31	1.11
500	56.83	49.86	37.42	8.84	5.34	2.05	500	9.86	4.93	0.79	9.56	4.75	0.97
2,000	32.76	25.89	16.76	10.80	6.08	1.86	2,000	10.61	5.34	0.88	10.64	5.31	0.98

Notes: Monte Carlo rejection rates based on 10,000 replications. OPS refers to the version of the statistic proposed by Chesher (1983) and Lancaster (1984), while CM to the feasible version that makes use of the theoretical expressions in Proposition 2 replacing the true parameter values by their MLEs and unconditional expectations by sample averages. Rejection rates in the right subpanels are based on the asymptotic distribution in Proposition 1 while the left ones on a parametric bootstrap procedure in which we simulate $B = 99$ samples from the model estimated under the null. See Supplemental Appendix 3 for details about the DGPs.

Table A2: (Binary) logit IM tests: Power properties (for three categories)

Panel A: Alternative hypothesis: Heterogeneous Gaussian β_{i2}

	Two regressors							Three regressors					
	(ξ_1, ξ_2)			(ξ_1, ξ_3)				(ξ_1, ξ_2)			(ξ_1, ξ_3)		
Sample size	10%	5%	1%	10%	5%	1%	Sample size	10%	5%	1%	10%	5%	1%
125	62.36	45.04	13.20	7.28	4.04	0.76	125	49.20	30.12	5.60	8.72	4.52	0.88
500	98.40	95.60	78.64	9.88	5.16	0.84	500	63.48	46.20	7.28	10.60	5.92	1.28
2,000	100.00	100.00	100.00	10.24	5.76	1.36	2,000	100.00	100.00	99.32	10.44	4.72	0.88

Panel B: Alternative hypothesis: Group heterogeneity in β_{i2}

	Two regressors							Three regressors					
	(ξ_1, ξ_2)			(ξ_1, ξ_3)				(ξ_1, ξ_2)			(ξ_1, ξ_3)		
Sample size	10%	5%	1%	10%	5%	1%	Sample size	10%	5%	1%	10%	5%	1%
125	21.48	12.80	3.16	26.64	18.60	5.20	125	17.04	8.48	1.44	35.56	21.04	3.88
500	28.28	18.20	4.48	45.36	29.76	5.96	500	29.04	15.44	2.52	52.72	38.20	11.00
2,000	34.84	24.36	6.68	73.60	61.28	25.36	2,000	40.32	28.16	7.76	94.72	89.24	54.60

Panel C: Alternative hypothesis: Omitted variable

	Two regressors							Three regressors					
	(ξ_1, ξ_2)			(ξ_1, ξ_3)				(ξ_1, ξ_2)			(ξ_1, ξ_3)		
Sample size	10%	5%	1%	10%	5%	1%	Sample size	10%	5%	1%	10%	5%	1%
125	15.84	10.12	3.32	16.08	7.72	1.56	125	9.28	5.32	1.48	34.76	16.76	1.36
500	36.24	23.88	5.04	63.40	45.80	16.80	500	17.68	9.12	2.56	43.52	25.68	5.16
2,000	54.32	41.64	16.80	99.80	99.08	88.40	2,000	41.08	27.76	8.32	99.92	99.80	96.92

Panel D: Alternative hypothesis: Ordered logit

	Two regressors							Three regressors					
	(ξ_1, ξ_2)			(ξ_1, ξ_3)				(ξ_1, ξ_2)			(ξ_1, ξ_3)		
Sample size	10%	5%	1%	10%	5%	1%	Sample size	10%	5%	1%	10%	5%	1%
125	17.16	9.64	2.68	21.84	14.84	6.92	125	15.24	8.16	2.12	24.36	16.36	9.40
500	21.20	12.96	3.52	19.08	11.84	3.64	500	19.04	10.92	3.28	16.72	9.76	2.44
2,000	42.20	31.72	10.92	26.52	18.20	5.84	2,000	39.80	27.92	10.00	25.64	16.24	5.72

Notes: Monte Carlo rejection rates based on 2,500 replications. Results for the feasible version of the IM test that makes use of the theoretical expressions in Proposition 2 replacing the true parameter values by their MLEs and unconditional expectations by sample averages. Rejection rates are based on a parametric bootstrap procedure in which we simulate $B = 99$ samples from the model estimated under the null. See Supplemental Appendix 3 for details about the DGPs.