Tests For Independence Between Categorical Variables*

Juan Sentana
University of Essex, Colchester, CO4 3SQ, U.K.
<jsenta@essex.ac.uk>

February 2019

Abstract

I study independence tests between two categorical variables. I prove the numerical equivalence between Pearson's contingency table test and the Lagrange Multiplier and overidentifying restrictions tests in several popular linear and non-linear regression models. This equivalence holds after exchanging regressors and regressands. I also prove the heteroskedasticity-robust Wald test in a multivariate linear probability model is numerically identical to the Wald test in the conditional multinomial model. I study in detail the exact finite sample size and power of all those tests. Finally, I use them to analyze if professional soccer players follow optimal mixed strategies in penalty kicks.

Keywords: Independence, Linear Probability Model, Logit, Overidentifying restrictions, Probit.

JEL: C12, C25, C35, C57

*Earlier versions of the theoretical results in this paper appeared in Sentana (2016). I would like to thank Dante Amengual, Abhimanyu Gupta, Gordon Kemp and Simon Weidenholzer for their comments. I have also benefited from many useful conversations with Enrique Sentana. Any remaining errors will obviously be my own. I gratefully acknowledge the financial support from Fundacion Ramon Areces and the University of Essex.

1 Introduction

Economic theories are usually confronted with data to assess their validity. This is often done by deriving hypothesis implied by the theory and testing them by means of standard econometric procedures. In some important cases, those hypothesis imply the independence between two categorical variables, which only take a finite number of values H and K, respectively.

A relevant example arises in testing the implications of mixed strategy equilibrium. In particular, in games with no pure strategy Nash equilibria, such as the matching pennies game, two fundamental implications of the theory are first that the probability of winning should be the same regardless of the strategy chosen, and second that the actions of the players should be serially independent.

In both cases, those implications imply independence between categorical variables. Specifically, the first implication requires independence between a winning dummy ($H = 2$) and a variable that describes the K strategies of the player, while the second implication means serial independence in the player's strategies ($H = K$).

Different researchers have used different econometric procedures to empirically test those hypothesis. For example, Palacios-Huerta (2003) tested the first implication by means of Pearson's independence test in a contingency table, while Chiappori et al (2002) used an F-test in a Linear Probability Model (LPM). A third possibility would be to test that the winning probabilities implied by a probit or logit model do not depend on the action taken by the player (see Brown and Rosenthal (1990)). Similarly, for the second implication, one could use Wilks' lambda, Pillai trace or the Lawley-Hotelling trace tests frequently employed in multivariate analysis of variance (see Stewart (1995) for more details).

Anatolyev and Kosenok (2009) showed the asymptotic equivalence between Pearson's independence test and the usual Wald test in a multivariate version of the LPM in the general $H \times K$ case. However, this equivalence does not prevent that those tests lead to different conclusions in practice, as highlighted by Berndt and Savin (1977). In fact, it is even possible that different researchers could report contradictory results with the same dataset.

In this paper, I prove the numerical equivalence for general H and K between Pearson's independence test in a contingency table, the Lagrange Multiplier (LM) test in several popular regression models: the multivariate version of the LPM, the conditional multinomial model, the multinomial logit and probit models; and the corresponding J-test for overidentifying restrictions. It is worth mentioning that this equivalence also applies to a Markov chain, which can be regarded as an analog to the multinomial model when $H = K$, although in a serially dependent context.

My results apply to a more general setting than the analysis of mixed strategies equilibrium. For example, Pesaran and Timmermann (1994), who were interested in testing the efficient market hypothesis in financial markets, showed that the lack of directional predictability of asset returns can be interpreted as the stochastic independence between the sign of the actual returns and the sign of the predictions made by asset managers who want to time the market (see also Henriksson and Merton (1981)). Therefore, one could also use my results to test in a unified manner the ability of some popular market timing strategies to predict the direction of the movements in some well-known asset prices, such as stock indices or exchange rates.

In contrast, the Likelihood Ratio (LR) and Wald tests of independence in those linear and non-linear regression models will usually differ. However, I also prove that the heteroskedasticity-robust version of the Wald test in the multivariate LPM is numerically identical to the Wald test in the conditional multinomial model.

In addition, I study the size and power in finite samples of all the different tests of independence that have been proposed. In that context, I explain how to obtain the exact p-value of the different tests in finite samples, as well as how to calculate the exact bootstrap distribution without resorting to simulations in the 2×2 case.

Finally, I make use of the lessons learned in the Monte Carlo exercises in an empirical application to penalty kicks. To do so, I collected a dataset of 549 penalty kicks in professional soccer games that include very detailed information on many relevant aspects of the play, and specifically actions and outcomes.

As a brief summary of the main results, I find that the behavior of professional soccer players is indeed consistent with the mixed strategy equilibrium predictions, in the sense that winning probabilities are identical across strategies and player's actions are serially independent across plays.

The paper is organized as follows. Section 2 explains the different econometric methods and present my numerical equivalence results. In section 3, I explain in more detail the first and second testable implications of the Minimax theorem. Section 4 studies the size of the tests in finite samples, while in section 5 I include an analysis of the power of the tests by means of simulations. Section 6 contains the empirical results. This is followed by the conclusions and several appendices where proofs and additional details can be found.

2 Econometric Methodology

Let x be a $K \times 1$ categorical variable that takes values $(A_1, ..., A_K)$, where $A_1, ..., A_K$ are K exhaustive and mutually exclusive dummy variables which fully characterize the categorical

variable. Similarly, let $\tilde{y}$ be another $H \times 1$ categorical variable that takes values $(B_1, ..., B_H)$. Both A_k and B_h, for $k = 1, .., K$ and $h = 1, ..., H$, are dummy variables equal to 1 if its corresponding categorical value is equal to its k^{th} or h^{th} value respectively.

2.1 *Contingency table*

Let π_{hk} denote the parameters of the underlying joint bivariate probability distribution; for example, π_{12} is the probability that B_1 and A_2 simultaneously take the value 1. Similarly, let $\pi_{h\diamond}$ and π_{*k} denote the parameters of the marginal probability distributions for $\tilde{y}$ and x respectively. For example, $\pi_{1\diamond}$ is the marginal probability that $B_1 = 1$.

The null hypothesis states that there is independence between $\tilde{y}$ and x, so the joint probability simply equals the product of their marginal probabilities:

$$H_0 : \pi_{hk} = \pi_{h\diamond} \times \pi_{*k} \ , \ h = 1, ..., H \text{ and } k = 1, ..., K.$$

A contingency table summarizes the sample information. In this case, it can be written as:

$\tilde{y} \backslash x$	A_1	...	A_K	Sum
B_1	n_{11}	...	n_{1K}	$n_{1\diamond}$
...	...	...	...	...
B_H	n_{H1}	...	n_{HK}	$n_{H\diamond}$
Sum	n_{*1}	...	n_{*K}	n

where n_{hk}, for $h = 1, .., H$ and $k = 1, ..., K$, denotes the observed joint frequency; for example, n_{12} is the number of times that B_1 and A_2 are simultaneously 1 in the sample. Also, $n_{h\diamond} = \sum_{k=1}^{K} n_{hk}$ denotes the number of times that B_h is 1 and $n_{*k} = \sum_{h=1}^{H} n_{hk}$ the number of times A_k is 1. Finally, $n = \sum_{k=1}^{K} n_{*k} = \sum_{h=1}^{H} n_{h\diamond}$ yields the total number of observations.

2.1.1 *Pearson's test*

The original and best known test for independence is Pearson's contingency table statistic, which is given by:

$$Pearson = \sum_{k=1}^{K} \sum_{h=1}^{H} \left(n_{hk} - \frac{n_{*k} n_{h\diamond}}{n} \right)^2 \left(\frac{n}{n_{*k} n_{h\diamond}} \right). \quad (1)$$

This statistic follows a $\mathcal{X}^2$ distribution with $(H-1)\times(K-1)$ degrees of freedom in large samples under appropriate regularity conditions (see Mood et al (1974)).[1] Unlike most other statistics, the $\mathcal{X}^2$ can provide information not only on the significance of any observed differences, but also on exactly which categories account for any differences found.

[1] Specifically, in addition to random sampling, it requires all the joint frequences π_{hk}, $h = 1, ..., H$ and $k = 1, ..., K$, to be strictly positive, so that the observed joint frequencies n_{hk} can be expected to increase at the same rate as the sample size n.

2.2 *Multivariate regression*

A multivariate regression is a technique that combines several regression models with the same regressors, one for each dependent variable. In the case of H equations and K explanatory variables, it can be written as:

$$\left.\begin{array}{c} B_{1i} = \delta_{11}A_{1i} + ... + \delta_{1K}A_{Ki} + u_{1i} \\ \vdots \\ B_{Hi} = \delta_{H1}A_{1i} + ... + \delta_{HK}A_{Ki} + u_{Hi} \end{array}\right\}.$$

Given that both regressors and regressands are dummy variables, the coefficients of the explanatory variables are the probability of the different values of the multinomial variable $\tilde{y}$ given the other multinomial variable x. For instance,

$$\delta_{hk} = E(B_h|A_1 = 0, ..., A_k = 1, ..., A_K = 0) = \Pr(B_h = 1|A_1 = 0, ..., A_k = 1, ..., A_K = 0).$$

Hence, the sum of δ_{h1} for $h = 1, ..., H$ is equal to 1 for all the columns in the matrix of regression coefficients. Therefore, the coefficients in the equation for B_{Hi} can be obtained from the other $H - 1$ equations because $B_{Hi} = 1 - \sum_{h=1}^{H-1} B_{hi}$. For that reason, I can cross out the last equation without loss of generality to avoid the singularity (see Judge et al (1985) chapter 12, section 5 for more details).

Define the vectors $B_h' = \begin{pmatrix} B_{h1} & ... & B_{hn} \end{pmatrix}, u_h' = \begin{pmatrix} u_{h1} & ... & u_{hn} \end{pmatrix}$ and $\delta_h = \begin{pmatrix} \delta_{h1} & ... & \delta_{hK} \end{pmatrix}$ for $h = 1, ..., H - 1$.

Also, define the matrices $X = \begin{pmatrix} A_{11} & ... & A_{K1} \\ A_{12} & ... & A_{K2} \\ ... & ... & ... \\ A_{1n} & ... & A_{Kn} \end{pmatrix}$ and $Y = \begin{pmatrix} B_1 & ... & B_{H-1} \end{pmatrix}$.

Finally, define the matrix of regression coefficients $\Pi = \begin{pmatrix} \delta_{11} & ... & \delta_{1K} \\ ... & ... & ... \\ \delta_{H-1,1} & ... & \delta_{H-1,K} \end{pmatrix}$, with $\delta = vec\,(\Pi')$ and $\Sigma = V\,(u_i)$.

In this way, the multivariate regression model can be written in matrix notation as:

$$Y = X\Pi' + u.$$

Under the assumption that u is homoskedastic and serially uncorrelated, the parameters of the model can be efficiently estimated by OLS equation by equation. The reason is that $\hat{\delta}^{GLS} = \hat{\delta}^{OLS}$ because the regressors in the $H - 1$ equations are identical.

The OLS estimator of the parameters of the h^{th} equation are:

$$\hat{\delta}_h^{OLS} = (X'X)^{-1}X'B_h,$$

where $X'X = \begin{pmatrix} n_{*1} & 0 & ... & 0 \\ 0 & ... & 0 & ... \\ ... & ... & ... & 0 \\ 0 & ... & 0 & n_{*K} \end{pmatrix}$ and $X'B_h = \begin{pmatrix} n_{h1} & ... & n_{hK} \end{pmatrix}'$ because $\Sigma_{i=1}^{n} A_{ki} = n_{*k}$ and $\Sigma_{i=1}^{n} A_{ki} B_{hi} = n_{hk}$ for $k = 1, ..., K$ and $h = 1, ..., H-1$.

This yields the following result:

$$\hat{\delta}_h^{OLS} = \begin{pmatrix} \frac{n_{h1}}{n_{*1}} & ... & \frac{n_{hK}}{n_{*K}} \end{pmatrix}',$$

so that equation by equation OLS yields the natural estimator of δ_{hk}. This means that the estimated probabilities are always non-negative and they add up to 1, which avoids a common criticism of the LPM (see Wooldridge (2002)).

The null hypothesis of independence implies that $\delta_{h1} = ... = \delta_{hK} = \delta_h$, for $h = 1, ..., H-1$, so that the conditional probability of $B_h = 1$ does not depend on the value of the conditional variable x. Again, the restricted model can be estimated efficiently by OLS equation by equation because GLS is once more equal to OLS. The restricted OLS estimators are trivially:

$$\tilde{\delta}_h^{OLS} = \frac{n_{h\diamond}}{n}, \ h = 1, ..., H-1.$$

The multivariate regression has one potentially important disadvantage. Under the alternative, it violates the homoskedasticity assumption because the conditional variance of the error term u will change depending on the values of the explanatory variables rather than being the assumed constant matrix Σ_U (see Wooldridge (2002)). However, the covariance matrix of u given the dummy regressors is constant under the null hypothesis of independence, say Σ_R. This implies that the homoskedasticity assumption holds and all the usual regression tests are valid. I will return to this issue in section 2.2.1.

In practice, it is easier to test the independence hypothesis by estimating the following model:

$$\left.\begin{array}{c} B_{1i} = \beta_{10} + \beta_{11} A_{1i} + ... + \beta_{1K-1} A_{K-1i} + u_{1i} \\ \vdots \\ B_{H-1i} = \beta_{H-1,0} + \beta_{H-1,1} A_{1i} + ... + \beta_{H-1,K-1} A_{K-1i} + u_{H-1i} \end{array}\right\},$$

where $\beta_{h0} = \delta_{hK}$ and $\beta_{hk} = \delta_{hk} - \delta_{hK}$, for $h = 1, ..., H-1$ and $k = 1, ..., K-1$. In these regressions with $K-1$ explanatory variables and a constant, the β coefficients are the differences between the probabilities of the corresponding variable and the baseline, which I have chosen to be the variable A_{Ki} without loss of generality. The adjustment of these regressions is identical to the adjustment of the regressions written in terms of δ's, but they have the advantage that the null hypothesis of independence can be expressed as $\beta_{hk} = 0$, for all $h = 1, ..., H-1$ and $k = 1, .., K-1$.

2.2.1 *Tests*

The three classical multivariate regression tests are the Wald (W), Likelihood Ratio (LR) and Lagrange Multiplier (LM). Note that for any dataset, the relationship between these tests is:

$$W \geq LR \geq LM,$$

despite being asymptotically equivalent (see Berndt and Savin (1977) and Engle (1983) for more details). Moreover, they are monotonic transformations of the regression F-test in the $H = 2$ case but not in general (see Appendix A.3 for the case of $H = 3$).

These three tests can be easily transformed into the Pillai trace, Wilks' lambda and Lawley-Hotelling trace tests used in multivariate analysis of variance (see Stewart (1995) for more details). More precisely, the Pillai trace test can be written as:

$$V = \frac{LM}{n}$$

while Wilks' lambda is

$$\Lambda = \exp\left(-\frac{LR}{n}\right)$$

and the Lawley-Hotelling trace test

$$LH = \frac{W}{n}.$$

When studying the size properties in finite samples in section 4, I will use some popular F-approximations to those tests, which are supposed to be more reliable in finite samples (see Appendix A.4 for more details).

Finally, I also consider a robust test which would still be valid if the assumption of homoskedasticity was violated. Specifically, I use the heteroskedasticity-robust version of the Wald test in a multivariate regression which I derive using the results in Hansen (1982).[2]

2.3 *Multinomial model*

Define $P_{hk} = \Pr(B_h = 1 | A_1 = 0, ..., A_k = 1, ..., A_K = 0)$ for $k = 1, ..., K$ and $h = 1, ..., H-1$, so that the joint probability is $\pi_{hk} = P_{hk} \times \pi_{*k}$, where $\pi_{*k} = \Pr(A_k = 1)$. Hence, the likelihood for observation i is:

$$\mathcal{L}_i = \prod_{k=1}^{K} (1 - \sum_{h=1}^{H-1} P_{hk})^{A_{ki}\left(1-\sum_{h=1}^{H-1} B_{hi}\right)} \prod_{h=1}^{H-1} P_{hk}^{A_{ki}B_{hi}}$$
$$\times (1 - \sum_{k=1}^{K} \pi_{*k})^{A_{Ki}} \prod_{k=1}^{K-1} \pi_{*k}^{A_{ki}}.$$

[2] STATA uses a degrees of freedom correction $\frac{n}{n-K}$ in the univariate case (see STATA (2012), Robust, entry for more details).

Note that I have expressed the joint likelihood function as the product of the likelihood of the second categorical variable given the first one (conditional model) times the likelihood function of the first categorical variable (marginal model). This decomposition is convenient because the independence hypothesis relates to the conditional model, not the marginal one.

Hence, the log-likelihood of the sample becomes:

$$\ln \mathcal{L} = \sum_{k=1}^{K} \left[\left(n_{*k} - \sum_{h=1}^{H-1} n_{hk} \right) \ln(1 - \sum_{h=1}^{H-1} P_{hk}) + \sum_{h=1}^{H-1} (n_{hk} \ln P_{hk}) \right] + n_{*K} \ln(1 - \sum_{k=1}^{K-1} \pi_{*k}) + \sum_{k=1}^{K-1} n_{*k} \ln \pi_{*k} \quad (2)$$

because the number of times when $A_k = 1$ is n_{*k} while the number of times $A_{ki}B_{hi} = 1$ is n_{hk}.

Maximizing the log-likelihood with respect to P_{hk} and π_{*k} yields:

$$\hat{P}_{hk} = \frac{n_{hk}}{n_{*k}} \text{ and } \hat{\pi}_{*k} = \frac{n_{*k}}{n_{*K}},$$

so that $\hat{P}_{hk} = \hat{\delta}_{hk}$ for $h = 1, ..., H-1$ and $k = 1, ..., K$. Note that $\hat{\pi}_{*k}$ will coincide under the null and alternative, so I can ignore these parameters.

Under the null, which states that $P_{hk} = P_{h\diamond}$ for $k = 1, ..., K$, then

$$\ln \mathcal{L} = \left(n - \sum_{h=1}^{H-1} n_{h\diamond} \right) \ln(1 - \sum_{h=1}^{H-1} P_{h\diamond}) + \sum_{h=1}^{H-1} n_{h\diamond} \ln P_{h\diamond} + n_{*K} \ln(1 - \sum_{k=1}^{K-1} \pi_{*k}) + \sum_{k=1}^{K-1} n_{*k} \ln \pi_{*k},$$

which yields

$$\tilde{P}_{h\diamond} = \frac{n_{h\diamond}}{n},$$

so that $\tilde{P}_{h\diamond} = \tilde{\delta}_h$ for $h = 1, ..., H-1$.

As in the multivariate regression, I consider the LM, LR and Wald tests, which have standard expressions (see Appendix A.1.3).

2.4 *Multinomial logit model*

Define the conditional probability matrix P as:

$$P = \begin{pmatrix} P_{11} & ... & P_{1K} \\ ... & ... & ... \\ P_{H1} & ... & P_{HK} \end{pmatrix}$$

with P_{hk} being the same as in section 2.3. The multinomial logit model ensures the non-negativity of P_{hk} for all h and k, as well as the adding up constraint $\sum_{h=1}^{H} P_{hk} = 1$, by assuming that

$$\left. \begin{array}{c} P(B_h = 1 \mid A_1, ..., A_K) = \frac{1}{D} \exp\left(\sum_{k=1}^{K} \gamma_{hk} A_{ki} \right), \ h = 1, ..., H-1 \\ P(B_H = 1 \mid A_1, ..., A_K) = \frac{1}{D} \end{array} \right\},$$

where $D = 1 + \sum_{k=1}^{K} \exp\left(\sum_{h=1}^{H-1} \gamma_{hk} A_{ki} \right)$.

Therefore, the multinomial logit is simply a reparametrization of the matrix P which ensures non-negative probabilities that add up to 1 by rows.

To estimate this model, I use Maximum Likelihood (see Stata (2012), Multinomial logit, entry for more details). The log-likelihood function of this conditional model is:

$$\mathcal{L}(\gamma) = \sum_{k=1}^{K} \left[\sum_{h=1}^{H-1} n_{hk} \ln P_{hk} + n_{Hk} \ln(1 - \sum_{h=1}^{H-1} P_{hk}) \right],$$

which is like the conditional component of the multinomial model but expressed in terms of γ_{hk} instead of P_{hk}'s.

Not surprisingly, the relationship between $\hat{P}_{hk}$ and $\hat{\gamma}_{hk}$ is:

$$\hat{P}_{hk} = \frac{\exp(\hat{\gamma}_{hk} A_{ki})}{\sum_{m=1}^{H} \exp(\hat{\gamma}_{mk} A_{ki})} \; h = 1, ..., H-1,$$

with an analogous expression for P_{Hk} (see Cameron and Trivedi (2005) chapter 15, section 4 for more details).

The multinomial logit log-likelihood function under the null hypothesis $H_0 : P_{hk} = P_h$ for $k = 1, ..., K$ and $h = 1, ..., H-1$, is:

$$\mathcal{L}(\gamma) = \sum_{k=1}^{K} \left[\sum_{h=1}^{H-1} (n_{hk} \ln P_h) + n_{Hk} \ln(1 - \sum_{h=1}^{H-1} P_h) + n_{*k} \ln \pi_k \right],$$

which yields

$$\tilde{P}_h = \frac{n_{h\diamond}}{n},$$

given that $\Sigma_{h=1}^{K} n_{h\diamond} = n$.

As expected, the relationship between $\tilde{P}_h$ and $\tilde{\gamma}_h$ is:

$$\tilde{P}_h = \frac{\exp(\tilde{\gamma}_h)}{\sum_{m=1}^{H} \exp(\tilde{\gamma}_m)} \; h = 1, ..., H-1.$$

As before, I will consider the LM, LR and Wald tests.

2.5 *Multinomial probit model*

Following section 27.3 of Ruud (2000), let $B_h^* \equiv (B_{hi}^*, \; i = 1, ..., n)'$ denote a column vector of n latent dependent variables whose conditional distribution follows the multivariate regression model

$$B_h^* = x\delta_h + u_h,$$

for $h = 1, ..., H$, where $u|x \sim N(0, \Omega)$ and Ω is constant.

Let the observation rule be

$$B_{hi} = 1 \left\{ B_{hi}^* = \max_{j=1,...,n} B_{hj}^* \right\},$$

where $1\{\}$ is the indicator function. Therefore, the i^{th} element of $B_h \equiv (B_{hi},\ i = 1, .., n)'$ equals 1 if the i^{th} value of the categorical variable $\tilde{y}$ is observed; otherwise, B_{hi} equals zero. Therefore, the log-likelihood function is

$$L(\theta) = \sum_{i=1}^{n} B_{hi} \ln \Pr(B_{hi} = 1|x),$$

where θ are the model parameters.

Once again, this log-likelihood function coincides with the conditional component of the log-likelihood function of the multinomial model. Therefore, the independent probit model in which Ω is a scalar, is the most flexible model that I can identify. It is worth mentioning that if u_h, instead of being normal, comes from a Weibull distribution, then we will go back to the multinomial logit model.

The multinomial probit model under the null is entirely analogous to the multinomial logit one.

2.6 *GMM*

Given that we have seen that

$$P_{hk} = \Pr(B_h = 1|A_1 = 0, ..., A_k = 1, ..., A_K = 0) = \frac{E(B_h A_k)}{E(A_k^2)} = \delta_{hk},$$

we can express all those parameters in terms of the following set of moment conditions

$$E\left[(y_i - \Pi x_i) \otimes x_i\right] = 0,$$

which coincide with the normal equations of the multivariate LPM as well as with the scores of the conditional multinomial model. Under H_1, Π is unrestricted while under H_0, $\Pi = \upsilon l_k'$, where l_k is a vector of ones. Note that under H_0, I can write $\Pi'(\upsilon) = l_k \upsilon' I_{H-1}$, which implies that $\delta(\upsilon) = vec(\Pi'(\upsilon)) = (I_{H-1} \otimes l_k)\upsilon$.

The GMM estimator is defined as:

$$\hat{\upsilon} = \arg\min_{\upsilon} \left[\frac{1}{n}\sum_{i=1}^{n}\left((y_i - \Pi(\upsilon)x_i) \otimes x_i\right)\right]' \Upsilon^{-1} \left[\frac{1}{n}\sum_{i=1}^{n}\left((y_i - \Pi(\upsilon)x_i) \otimes x_i\right)\right],$$

where Υ is a symmetric positive definite $(K \times (H-1)) \times (K \times (H-1))$ weight matrix.

The optimal GMM estimator is the one which minimizes the GMM criterion function when Υ is efficient and equal to $\left(\Sigma_R \otimes \sum (x_i x_i')\right)$.

The J-test for overidentifying restrictions is just the value of the GMM objective function evaluated at the efficient GMM estimator (see Hansen (1982) for more details). Algebraically:

$$J = n \times \bar{g}(\hat{\upsilon})' \Upsilon^{-1} \bar{g}(\hat{\upsilon}),$$

where $\bar{g}(\hat{\upsilon}) = \frac{1}{n}\Sigma\left[(y_i - \Pi(\hat{\upsilon})x_i) \otimes x_i\right]$.

2.7 *Numerical equivalence results*

Having many different procedures to test independence may lead to different conclusions. However, in this environment three of those models are essentially the same. Specifically, the values of the log-likelihood function under the null and alternative of the multinomial logit and probit models are equal to the corresponding conditional component of the log-likelihood of the multinomial model.

Therefore, the results in section 17.4 of Ruud (2000) imply that the LR test of the multinomial model, multinomial logit and probit models must coincide because the LR test is numerically invariant to non-linear transformations of parameters and restrictions. Ruud (2000) results also imply that the LM test is numerically invariant to non-linear transformations of the restrictions when the information matrix is used for its calculation instead of the Hessian. In contrast, the Wald tests will not coincide.

The main theoretical result in this paper is that these three LM tests also coincide with the LM test in the multivariate linear probability model, and more importantly, with Pearson's test for independence as well as with the J-test for overidentifying restrictions. The following proposition, which I prove in the Appendix, contains the precise result:

Proposition 1 *The Lagrange Multiplier versions of the test for independence in a multivariate linear probability model, multinomial logit, multinomial probit and multinomial model are numerically identical to Pearson's contingency table test for independence and the J-test for overidentifying restrictions. Additionally, the same numerical equivalence result holds if one exchanges the regressors and regressands in all those models.*

This means that different researchers using different econometric procedures will reach exactly the same conclusions if they use LM tests. From the computational point of view, of course, the simplest test is Pearson's statistic, which has a very simple closed-form. In contrast, the multinomial logit and especially probit models should be avoided because they require numerical optimization.

This result also has the advantage that there will only be one bootstrap version for the different tests. Additionally, the Monte Carlo experiments previously reported in the literature on contingency table tests also apply to all the other different tests, so they could be combined.

Proposition 1 also says that if we change x and $\tilde{y}$ so that x now takes values $B_1, ..., B_H$ and $\tilde{y}$ takes values $A_1, ..., A_K$, then the contingency table in section 2.1 will be flipped but it will not change the test statistics obtained in the aforementioned models. While this is obvious for the contingency table test (1), it is far from obvious for all the other conditional models whose tests remain numerically identical. For example, one obtains numerically the exact same LM statistic

if one regresses $\tilde{y}_i$ on x_i or x_i on $\tilde{y}_i$, while in the multinomial logit, imposing independence on $P(B_h = 1 \mid A_1, ..., A_K)$ yields the same LM statistic as imposing it on $P(A_k = 1 \mid B_1, ..., B_H)$.

As stated in section 2.2, the multivariate regression under the alternative violates the homoskedasticity assumption, although the mean is correctly specified. However, under the null, both the mean and variance of the model are correctly specified as the covariance matrix of u given the dummy regressors is constant. But obviously, the conditional distribution is not normal, so the likelihood function of the multivariate regression model is different from the likelihood of the multinomial model even under the null. Consequently,

$$LR_{LPM} \neq LR_{Multinomial}$$

(see Appendix A.1 for more details).

Although the Wald tests are generally different, the numerical equivalence between the OLS estimator of the regression coefficients and the ML estimators of the conditional probabilities suggest a close relationship. It turns out that the crucial difference is the homoskedasticity assumption in the standard Wald test of the multivariate regression. Specifically, if one decided to carry out a robust test which would remain valid when the homoskedasticity assumption is violated, the following numerical equality, which I prove in the Appendix, will hold:

Proposition 2 *The heteroskedasticity-robust version of the Wald test for independence in the multivariate LPM is numerically identical to the Wald test of the conditional multinomial model.*

This implies that one could use either model and get the same conclusions and implications. However, the Wald version of the multinomial logit model is different from the multinomial Wald test in Proposition 2, and the same applies to the multinomial probit model because Wald tests are not invariant to non-linear transformations of the restrictions.

3 Applications to Mixed Strategies

3.1 *Penalty kicks as zero-sum games*

The Minimax theorem is a decision rule used in game theory for minimizing the possible loss in the worst case scenario. It can be regarded as a special case of the more general theory of Nash equilibrium, but it only applies to two-person zero-sum games. As I explained in the introduction, the first testable implication of the Minimax theorem is that the expected payoffs for the players should be the same regardless of the strategy chosen.

Similarly, the second testable implication of the Minimax theorem is that a player's strategy is the same at each play regardless of his previous actions, so that his actions should be independent draws from a multinomial random variable. In that regard, note that the players' strategies will

not be serially independent if they switch actions too often (negative serial correlation) or if they choose not to switch their actions regularly (positive serial correlation).

Due to the clarity of the rules and the detailed structure of the simultaneous one-shot play, a penalty kick in soccer captures the theoretical setting of a two-person zero-sum game extremely well. A formal model of the penalty kick can be written as follows. One goalkeeper and one kicker are facing each other at a penalty kick. The kicker preferences are to score while the goalkeeper has the opposite preferences. Specifically, the kicker's payoff is the probability of scoring while the goalkeeper's payoff is the complementary probability. The kicker may choose to kick to the goalkeeper's right (R) or to his left (L). Similarly, the goalkeeper may choose to jump to his left or to his right. When the kicker and the goalie choose the same side (L or R) the outcome is less likely to be a goal. In addition, there is usually a natural side for a kicker to shoot, so that the probability of scoring, if kicking to that side, is higher than when kicking to the opposite side, both when the goalkeeper guesses it and when he does not.

Following Sentana (2019), suppose without loss of generality that the natural side of the kicker is to shoot to the left because he is right-footed. The payoff matrix, which consists of scoring probabilities, is then:

Table 1

		Goalkeeper	
		Left	Right
Kicker	Left	$a, 1-a$	$b, 1-b$
	Right	$c, 1-c$	$d, 1-d$

where the first payoff corresponds to the kicker and the second payoff to the goalkeeper. Here a, b, c and d are the probabilities that a goal is scored so that this is a zero-sum game. We can assume that $b > a$, $c > a$, $c > d$ and $b > d$ because the goalkeeper is more likely to save when both players choose the same side and the kicker is more likely to score when he chooses his natural side. Since it is easy to see that there is no pure Nash equilibrium in this game, players who behave optimally must play mixed strategies.

Let's now consider a more realistic situation when both players have a third additional strategy: Center (C). Although in general the game will have no pure strategy Nash equilibrium, the number of parameters of the payoff matrix increases from four to nine, and it is complicated to find out the conditions that guarantee a single mixed strategy solution. For that reason,

consider the following simplified model with only three parameters:

Table 2

		Goalkeeper		
		Left	Center	Right
Kicker	Left	$1-c, c$	$a, 1-a$	$a, 1-a$
	Center	$b, 1-b$	$1-a, a$	$a, 1-a$
	Right	$b, 1-b$	$a, 1-a$	$1-a, a$

where the first payoff corresponds to the kicker and the second payoff to the goalkeeper. As in Sentana (2019), I assume that $b > 1-c$, $a > \frac{1}{2}$, $b > a > c$ and $a > 1-c$ because the goalkeeper is more likely to save when both players choose the same side and the kicker is more likely to score when he chooses his natural side.[3] Under these conditions, there is no pure strategy Nash equilibrium in this game, but it has a unique mixed strategy Nash equilibrium involving all three strategies.

Similar results can be tediously derived for the case in which the kicker and the goalkeeper have $K > 3$ actions at their disposal. For example, they could distinguish between the direction (L, C and R) and the height (high and low), which will give rise to six possible actions.

Next, I study the application of the methods studied in section 2 to this problem.

3.2 *First implication Minimax theorem in 2×K contingency tables*

3.2.1 *LPM*

The LPM under the alternative is defined as:

$$B = \delta_1 A_1 + \delta_2 A_2 + ... + \delta_K A_K + u,$$

where B and u are an $n \times 1$ vectors. I can estimate the model by OLS. In turn, the LPM under the null hypothesis states that $\delta_1 = ... = \delta_K = \delta$.

3.2.2 *Logit model*

The Logit model is described as follows:

$$\left.\begin{array}{l} B = 1 \text{ if } B^* \geq 0 \\ B^* = X\gamma + u \end{array}\right\},$$

where $X = (A_1, ..., A_K)$, $\gamma = \begin{pmatrix} \gamma_1 & ... & \gamma_K \end{pmatrix}'$ and u is logistically distributed.

Hence, the Logit model implies:

$$P(B = 1|X) = \frac{\exp(\gamma_1 A_{1i} + \gamma_2 A_{2i} + ... + \gamma_K A_{Ki})}{1 + \exp(\gamma_1 A_{1i} + \gamma_2 A_{2i} + ... + \gamma_K A_{Ki})}.$$

[3] Given that $b > a > c$, a right footed kicker will kick more frequently to his natural side than to any of the others. However, the goalkeeper may prefer to jump less than $\frac{1}{3}$ of the time to that side if $b + c > 2a$.

Therefore, in this model

$$P_k : P(B = 1|A_1 = 0, ..., A_k = 1, ..., A_K = 0) = \frac{\exp(\gamma_k)}{1 + \exp(\gamma_k)},$$

so

$$\gamma_k = \ln\left(\frac{P_k}{1 - P_k}\right), \ k = 1, ..., K.$$

The Logit model under the null hypothesis states that $\gamma_1 = ... = \gamma_K = \gamma$.

3.2.3 *Probit model*

The Probit model is defined as follows:

$$\left.\begin{array}{c} B = 1 \text{ if } B^* \geq 0 \\ B^* = \psi_1 A_1 + \psi_1 A_2 + ... + \psi_K A_K + u \end{array}\right\},$$

where $X = (A_1, ..., A_K)$, $\psi = \begin{pmatrix} \psi_1 & ... & \psi_K \end{pmatrix}'$ and u is standard normal.

Hence, the Probit model implies:

$$P(B = 1|X) = \Phi\left(\psi_1 A_1 + ... + \psi_K A_K\right),$$

where $\Phi()$ is the cumulative distribution function (cdf) of the standard normal.

I can define

$$P_k : P(B = 1|A_1 = 0, ..., A_k = 1, ..., A_K = 0) = \Phi\left(\psi_k\right),$$

so

$$\psi_k = \Phi^{-1}\left(P_k\right), \ k = 1, ..., K.$$

Under the null, the Probit model states that $\psi_1 = ... = \psi_K = \psi$.

3.2.4 *Multinomial model*

The contingency table is as follows:

$\tilde{y}\backslash x$	A_1	...	A_K	Sum
B_1	n_{11}	...	n_{1K}	$n_{1\diamond}$
B_2	n_{21}	...	n_{2K}	$n_{2\diamond}$
Sum	n_{*1}	...	n_{*K}	n

The conditional log-likelihood function under the alternative is:

$$\ln \mathcal{L} = \sum_{k=1}^{K} \left[\left(n_{*k} - n_{1k}\right)\left(1 - P_k\right) + \left(n_{1k} \ln P_k\right)\right].$$

Under the null, which states that $H_0 : P_k = P$, for $k = 1, ..., K$, it becomes

$$\ln \mathcal{L} = n_{1\diamond} \ln P + (n - n_{1\diamond}) \ln(1 - P).$$

3.3 *Second implication Minimax theorem in $K \times K$ contingency tables*

This requires a different treatment because the Markov chain is similar but not exactly a special case of the $H \times K$ case due to serial dependence across observations. However, all the other models (multivariate regression, multinomial probit and logit) can also be used with $\tilde{y}_t$ as regressand and $\tilde{y}_{t-1}$ as regressor. The same applies to the contingency table test and GMM.

3.3.1 *Markov chain $K \times K$: the unrestricted model*

Let us summarize the K strategies for each player $(A_1, ..., A_K)$ at time t by means of the vector $\tilde{y}_t$. The stochastic process $\tilde{y}_t$ has the Markov property if for all $k \geq 1$ and all t

$$Prob(\tilde{y}_{t+1}|\tilde{y}_t, \tilde{y}_{t-1}, \tilde{y}_{t-2}, ...\tilde{y}_{t-k}) = Prob(\tilde{y}_{t+1}|\tilde{y}_t).$$

Let

$$P_{hk} = Prob(\tilde{y}_{t+1} = x_h|\tilde{y}_t = x_k).$$

The conditional multinomial model is defined by the $K \times K$ transition matrix

$$P = \begin{pmatrix} P_{11} & ... & P_{1K} \\ ... & ... & ... \\ P_{K1} & ... & P_{KK} \end{pmatrix},$$

with states $k = 1, ..., K$, where $P_{Kk} = 1 - \sum_{h=1}^{K-1} P_{hk}$, $\forall k$ and $h = 1, ..., K-1$.

Let n_{hk} be the number of times that there occurs a one period transition from state k to state h, with $A_{k1} = 1$ if the first observation belongs to state k and zero otherwise. The likelihood function of the Markov chain can be written as:

$$L(\theta) = P(\tilde{y}_1) \prod_{h=1}^{H} \prod_{k=1}^{K} P_{hk}^{n_{hk}},$$

where $P(\tilde{y}_1) = \prod_{k=1}^{K} \pi_k^{A_{k1}}$ and $\theta = (P_{11}, ..., P_{KK})$ (see Lee et al (1968) for more details).

The log-likelihood function is:

$$\mathcal{L}(\theta) = \sum_{k=1}^{K} \left[n_{Kk} \ln(1 - \sum_{h=1}^{K-1} P_{hk}) + \sum_{h=1}^{K-1} (n_{hk} \ln P_{hk}) \right] + A_{K1} \ln(1 - \sum_{h=1}^{K-1} \pi_k) + \sum_{k=1}^{K-1} A_{k1} \ln \pi_k.$$

Note that this log-likelihood function is different from the multinomial log-likelihood function in (2) because the marginal model is based on a single observation while the conditional model is recursive.

3.3.2 *Markov chain: the restricted model*

If the Markov chain is serially independent, the matrix P will be:

$$P = \begin{pmatrix} 1 \\ ... \\ 1 \end{pmatrix} \times \begin{pmatrix} \pi_1, & ... & , \pi_{K-1}, & 1 - \sum_{k=1}^{K-1} \pi_k \end{pmatrix}.$$

I can achieve this by imposing the null hypothesis $H_0 : P_{hk} = P_h$, for $k = 1, ..., K$ and $h = 1, ..., K-1$ because recall that $P_K = 1 - \sum_{h=1}^{K-1} P_h$. I can then obtain the restricted estimators from the following log-likelihood:

$$\mathcal{L}(\phi) = \sum_{k=1}^{K} \{ \sum_{h=1}^{K-1} (n_{hk} \ln P_h) + n_{Kk} \ln(1 - \sum_{h=1}^{K-1} P_h)] + n_{*k} \ln \pi_k \},$$

where $\phi = (P_1, ..., P_{K-1})$ and $n_{h\#} = \Sigma_{k=1}^{K} n_{hk}$.

Serial independence can then be assessed by means of the usual Wald, LR and LM tests.

4 Size Experiments

In order to guide the empirical application in section 6, I will use the penalty kick game in section 3 as my experimental design.

4.1 *Exact distribution of tests in 2×2 contingency tables*

Even in experimental studies, few observations for each player are likely to be the rule rather than the exception. Therefore, it is important to investigate the behavior of the tests described above in small samples because their asymptotic $\mathcal{X}^2$ distribution may be unreliable when the number of observations is small. Part of the problem is that given that all the variables used are discrete, the number of states of the world is finite (2 possible actions per player $\times$ 2 possible outcomes per player's actions). In addition, the number of values of the estimators and test statistics will be repeated in many of those states of the world. For example, for $n = 5$ and two actions per player, there are $(2^3)^5$ possible states, but only 286 different contingency tables, while for $n = 20$ there are $(2^3)^{20}$ states, but only 1771 contingency tables. For that reason, I simulate the contingency tables directly, which contain all the information. Note that tests of the first and second hypothesis with two actions are equivalent in this context because, although the variables involved are different, they are all based on 2×2 contingency tables. For that reason, I focus on the first hypothesis only.

Recall that in the contingency table in section 2 applied to the penalty kick case, where now $\tilde{y}$ is the outcome (success (S) or failure (F)) and x is the action of the player (L and R), both $n_L = n_{SL} + n_{FL}$ and $n_S = n_{SL} + n_{SR}$ have values that go from 0 to n. Given those values, I only need to choose an additional element to complete the contingency table. Without loss of generality, I choose n_{SL}. For fixed n_L and n_S, n_{SL} fluctuates between a maximum and a minimum. It is easy to see that the minimum value n_{SL} can take is the maximum of 0 and $n_L + n_S - n$, while the maximum value it takes is the minimum of n_L and n_S.

To find the exact probability of each of those contingency tables and therefore of the corresponding test statistics, first note that under the null hypothesis the number of kicks to the left (n_L) and the number of goals scored (n_S) are independent random variables. So,

$$\Pr(n_L, n_S, n_{SL}) = \Pr(n_S) \times \Pr(n_L) \times \Pr(n_{SL} \mid n_L, n_S),$$

where $\Pr(n_j)$, for $j = S, L$, is binomial, whose values depend on the values of n and $E(L) = \pi_L$ or $E(S) = \pi_S$. In contrast, Fisher (1922) showed in the context of a well known tea cup classification example that $\Pr(n_{SL} \mid n_L, n_S)$ is hypergeometric, with values that depend on the values of n, n_{SL} and n_S.[4]

The parameters values that I use in my simulations are as follows. After collecting a dataset of 9,017 penalty kicks of professional soccer games in the main European leagues during September 1995-June 2012, Palacios-Huerta (2017) obtained the following payoff matrix:

Table 3

		Goalkeeper	
		Left	Right
Kicker	Left	0.591, 0.409	0.941, 0.059
	Right	0.931, 0.069	0.712, 0.288

I will use these values to see how the tests for two actions perform under the null for different sample sizes.

As I showed in section 2, there are only seven possible tests: the LM in the multivariate regression, which coincide with Pearson's independence test and the LM test in a multinomial model, multinomial logit and multinomial probit models as well as the J-test for overidentifying restrictions; the LR and Wald tests in the multivariate regression, Wald's heteroskedasticity-robust version, which coincides with the Wald test in the multinomial model, and the Wald and LR tests in the multinomial logit model, the last one being equal to the LR test in the probit and multinomial model. I will also consider the F-test of the univariate regression in this case because $H = 2$.

Nevertheless, given that the Monte Carlo simulations are not useful for inferences in a given sample because we do not know the true values of the parameters in practice, I explain in Appendix A.5 how to calculate the exact bootstrap p-values of the tests.

4.2 *Problematic cases*

Sometimes the calculations for some of the tests mentioned in the previous section breakdown. Although this is unlikely to happen with real data, I discuss here those situations because they occur in the simulations.

[4] The binomial distribution gives the probability of k successes in n trials with replacement, while the hypergeometric distribution does the same thing, but without replacement.

4.2.1 *Perfect classification*

As an example suppose that when the kicker shoots to the left, he never scores, whereas when he shoots in any other direction he may score or not. Hence, $\hat{\delta}_L = 0$ but $0 < \hat{\delta}_R < 1$. In this case, $\hat{\gamma}_L \to -\infty$ and the computation of the logit model breaks down (see Ruud (2000) section 27.1 for more details). The logit LR test is well defined although the unrestricted likelihood may also lead to numerical errors. For instance, when $\hat{\delta}_L \to 0$ the limit of the logit log-likelihood function becomes:

$$\lim_{\hat{\delta}_L \to 0} \log \left[\hat{\delta}_L^{\frac{n_{SL}}{n}} (1 - \hat{\delta}_L)^{\frac{n_{FL}}{n}} \hat{\delta}_R^{\frac{n_{SR}}{n}} (1 - \hat{\delta}_R)^{\frac{n_{FR}}{n}} \right]^n = \log \left[\hat{\delta}_R^{n_{SR}} (1 - \hat{\delta}_R)^{n_{FR}} \right].$$

In this context, Stata removes the perfectly classified observations and computes again the MLE from the remaining ones. However, it fails to provide a Wald test. In that regard, I can prove that the limit of the logit Wald test goes to zero when one of the $\hat{\gamma}_i$, for $i = L, R$, goes to plus or minus infinity.

4.2.2 *Perfect fit*

In this case, the variables L and R explain perfectly the model, i.e. $R^2 = 1$. This requires $\hat{\delta}_L = 1$ and $\hat{\delta}_R = 0$ or vice versa. In this context, I can show that the LM test in the LPM is exactly equal to the number of observations, while the usual Wald, F and LR test as well as the heteroskedasticity-robust version of the Wald test of this regression model diverge to infinity. In the logit model, the LR can still be computed and it is not generally infinity, but the limit of the Wald test is surprisingly equal to 0.

The fact that the logit Wald test is 0 while the robust and non-robust versions of the Wald test in the LPM diverge to infinity confirms that this type of test is not numerically invariant to non-linear transformation of the restrictions (see again Ruud (2000) section 17.4 for another example in which two Wald tests based on transformation of the restrictions diverge).

4.2.3 *Single outcome*

This case arises when the estimated probability of scoring ($\hat{\pi}_S$) is either 0 or 1. This implies that the residual sum of squares of both the restricted and unrestricted model (SSR_R and SSR_U) are 0, which in turn implies that $\hat{\delta}_L = \hat{\delta}_R = 0$ or $\hat{\delta}_L = \hat{\delta}_R = 1$ depending on the value of $\hat{\pi}_S$.

When this occurs, I set all the tests for the LPM and logit models to 0, so that their p-values are 1.

4.2.4 *Single choice*

This occurs when the estimated probability of choosing left ($\hat{\pi}_L$) is either zero or one, which means that the player is only employing one strategy. When this case arises, I again set all the tests to 0 because the single choice situation is like the single outcome situation in the regression of L on a constant and $\tilde{y}$. Although, the theoretical results in section 3 show that $\pi_L = 0$ would not be optimal, $\hat{\pi}_L = 0$ can happen despite $\pi_L > 0$ if n is small.

Finally, there exists also the possibility that both the Single Choice and Single Outcome cases occur simultaneously, in which case I again set all the tests to 0.

4.3 *Comparison between exact, asymptotic and Monte Carlo size in 2×2 contingency tables*

I use the procedures in section 4.1 to obtain the exact distribution of the different test statistics in samples of size 5 and 20. I also generate $10,000$ Monte Carlo replications using the payoff matrix in Table 3 to check that the simulated p-values closely agree with my exact results.

The simulations under the null are extremely simple. First, I randomly draw the probabilities of the actions of the kicker and the goalkeeper using the mixed strategy Nash equilibrium probabilities (see Sentana (2019) section 2.1 for more details). Then, I use the success probability from the relevant element of the payoff matrix to simulate whether or not the goal is scored.

All the graphs plotted in this section are cross plots of exact p-values in the horizontal axis and Monte Carlo and asymptotic p-values in the vertical axis. As usual, small p-values correspond to large test statistics and vice versa. Recall that the tests only take a small number of values, which are plotted as diamonds (Monte Carlo) or dashes (Asymptotic). If the asymptotic p-values were reliable in finite samples, the crosses should lie along the 45° degree line. In contrast, asymptotic p-values above the 45° degree line imply under rejection while those below mean over rejection.

4.3.1 *Five observations*

The graphs are presented in Figures 1.a to 1.g. The Monte Carlo and exact p-values are aligned, which means $10,000$ simulations provide a good approximation to the exact distribution, although at a considerable larger computational cost.

Empirical researchers that rely on asymptotic p-values should probably use the logit LR or the LR test in a LPM, although the latter performs worse than the former. In contrast, they should avoid the F and LM tests from the LPM and the three Wald tests, especially the one from the logit model.

4.3.2 *Twenty observations*

The graphs for this other sample size are presented in Figures 2.a to 2.g. As with $n = 5$, I found that the Monte Carlo p-values are extremely close to the exact p-values, although they are even more costly to compute in this case.

Given that there are many more possible values for the tests when $n = 20$, I focus on p-values below 15%, which are the most relevant ones in practice.

For 20 observations, researchers could still use the LR test of the LPM, but also the LM and F-test in a LPM and avoid the rest. In particular, the Wald test in the logit model and the heteroskedastic version of the Wald test in the LPM show considerable over rejections.

4.4 *Comparison between asymptotic and Monte Carlo size in 2×3 and 3×3 contingency tables*

In the case of three actions, the number of possible contingency tables is very large, and finding their exact distribution is very tedious. For that reason, I only compare Monte Carlo and asymptotic p-values using p-value plots (see Davidson and MacKinnon (1998)), which display the empirical cdf of the asymptotic p-values in the Monte Carlo simulations. The simulations are carried out analogously to the case with 2 actions.

I have simulated $10,000$ replications of the 3×3 model explained in section 3 with $n = 20$ and parameter values $a = 0.98$, $b = 0.99$ and $c = 0.97$ to check if the p-value plots are close to the 45° degree line for those below the 15% level, which are the most relevant ones.

4.4.1 *Scoring equality*

The graphs for the null hypotheses of equal scoring probability are presented in Figures 3.a to 3.g. Empirical researchers that rely on asymptotic p-values should probably use the LR, LM and F-test of the LPM and avoid the rest of the tests. In particular, the Wald test in the LPM over rejects the null considerably.

4.4.2 *Serial independence*

The p-value plots corresponding to the null hypothesis of lack of serial independence are presented in Figures 4.a to 4.h. Researchers that rely on asymptotic p-values should probably use F-versions of the Lawley-Hotelling (LH) and Wilks' test in the multivariate LPM, although the former is slightly better. At the same time, they should avoid the remaining tests. Again, the Wald test in the LPM and the multinomial logit LR test show considerable over rejections, while the Wald test in the multinomial logit hardly ever rejects.

5 Power Experiments

When choosing the significance level of a test, one sets the probability of rejecting the null when in fact it is true (Type 1 error). In the previous section, I studied which tests are more reliable when one chooses this level to be small, say 5%. At the same time, one would like to reject the null with high probability when the null is false. This is known as the power of the test. In what follows, I use Monte Carlo simulations to investigate the power of the different tests in some reasonable designs which do not satisfy the null.

As in the case of size, for all the hypothesis I have considered $10,000$ replications so that the confidence intervals for the Monte Carlo rejection rates at the 1, 5 and 10% levels under the null are $(0.80, 1.20)$, $(4.57, 5.43)$ and $(9.41, 10.6)$, respectively.[5]

In the case of 5 observations, all tests have very little power. For that reason, I am only going to discuss the results with $n = 20$.

5.1 *Alternatives to equal scoring probabilities in 2×2 contingency tables*

The alternative of the first implication of the Minimax theorem is that the player's probability of scoring depends on the strategy chosen. Here, I consider two alternatives because in reality, players do not necessarily know how to solve for the mixed strategy Nash equilibrium. At the same time, I have assumed serial independence to concentrate only in these two alternatives.

5.1.1 *Alternative 1*

In the case of the specific payoff matrix in Palacios-Huerta (2017) that I reproduce in Table 3 at the end of section 4.1, it is easy to show that under the null, the kicker and goalkeeper play left with probability 0.386 and 0.403, respectively. Instead, I assume that the kicker and goalkeeper play left with probability 0.7 and 0.3, respectively. I have chosen these probabilities because, under the assumption that the kicker is right-footed, his natural side is to shoot to the left hand side of the goalkeeper. Therefore, one could think that a naive kicker is more likely to shoot to the left, which justifies the kicker's probability of 0.7. However, the goalkeeper may believe that the obvious reasoning of a right-footed kicker is that the goalkeeper will jump to the left and therefore the probability of scoring will be low, so he will change the direction and kick to the right. That is why the chosen goalie's probability is 0.3.

[5] The formula for calculating the confidence interval is $\alpha \pm 1.96\sqrt{\alpha(1-\alpha)/\left(N^0 \ Replications\right)}$, where α is the significance level.

5.1.2 *Alternative 2*

For this alternative, I have assumed that both kicker and goalkeeper play left with probability 0.5. Therefore, it is as if at each penalty kick, both the kicker and the goalkeeper toss a fair coin to determine the direction they are going to choose. Brown and Rosenthal (1990) also considered this type of simple alternative.

5.1.3 *Power of tests*

I have only looked at the power of the LM, LR and F tests of the null hypotheses of equal scoring probabilities in the LPM because they are the only ones whose asymptotic p-values are reliable for $n = 20$ under the null.

The following tables show the percentage of times that these tests reject the null at the 1, 5 and 10% significance level under alternatives 1 and 2.

	F-test		LM test		LR test	
%	Alternative 1	Alternative 2	Alternative 1	Alternative 2	Alternative 1	Alternative 2
1	0.88	2.41	0.78	2.21	1.36	2.97
5	5.01	7.12	5.01	7.12	6.91	8.63
10	10.2	11.15	11.94	13.16	13.64	14.33

These results suggest that empirical researchers should use the LR test of the LPM because it is slightly more powerful under the two alternatives. However, it does not have much power, which implies that even very simple-minded alternatives like tossing a fair coin will be difficult to detect in practice.

5.2 *Alternatives to serial independence in 2×2 contingency tables*

The alternative to the second implication of the Minimax theorem is that a player's actions at time t depend on the action he chose at time $t-1$. To propose a specific alternative, I use a Markov Chain similar to the one in section 3.2.

Suppose there are two strategies for each player (Left and Right). In this case, the 2×2 transition matrix will be:

$$P = \begin{pmatrix} P_{LL} & 1 - P_{LL} \\ 1 - P_{RR} & P_{RR} \end{pmatrix}.$$

The stationary distribution in a Markov Chain is defined by the vector $\pi = \begin{pmatrix} \pi_L & \pi_R \end{pmatrix}$ such that $\pi P = \pi$, which yields:

$$\pi_L = \frac{1 - P_{RR}}{2 - (P_{LL} + P_{RR})} \text{ and } \pi_R = 1 - \pi_L = \frac{1 - P_{LL}}{2 - (P_{LL} + P_{RR})},$$

where π_L and π_R are the unconditional probabilities of kicking to the left and right respectively.

In this context, the serial correlation of the chain is measured by $\rho = (P_{LL} + P_{RR}) - 1$, with $-1 \leq \rho \leq 1$. When ρ is positive (negative), then we say that there is a positive (negative) association between the variables. Under the null hypotheses, ρ is equal to 0.

I considered four different alternatives depending on the value of ρ (0.5, -0.5, 0.05 and -0.05). But at the same time, I have assumed simultaneous moves and fixed the stationary probabilities of the Markov chain equal to the probabilities of the optimal strategy in the Nash equilibrium, so that the only discrepancy from the null is serial dependence.

5.2.1 *Power of tests*

Once again, I have only looked at the power of the LM, LR and F tests of the null of serial independence in the LPM because they are the only ones whose asymptotic p-values are reliable when $n = 20$.

The following table shows the percentage of times that these tests reject the null at the 1, 5 and 10% significance level under the following alternatives: $\rho = 0.5$ (1), $\rho = -0.5$ (2), $\rho = 0.05$ (3) and $\rho = -0.05$ (4).

	F-test				LM test				LR test			
%	(1)	(2)	(3)	(4)	(1)	(2)	(3)	(4)	(1)	(2)	(3)	(4)
1	24.07	45.26	1.29	2.20	21.29	36.51	0.82	1.42	32.60	45.26	1.86	2.37
5	45.54	67.30	4.87	6.46	49.79	67.30	5.61	6.77	50.32	74.51	6.72	9.10
10	54.79	79.65	8.74	11.19	54.79	82.08	9.95	13.80	62.54	82.58	12.93	16.23

These results suggest that empirical researchers should use the LR test of the LPM because it is the most powerful test under the four alternatives, although as expected, it does not have much power under alternatives 3 and 4.

5.3 *Alternatives to Equal Scoring Probabilities in 2×3 Contingency Tables*

As in the case of two actions, the null hypothesis of equal scoring probabilities will be violated when a player's probability of scoring depends on the strategy chosen. Here, I proposed two different alternatives.

5.3.1 *Alternative 1*

In the case of the model with three actions, I assume that the kicker plays left with probability 0.7 and center and right with 0.15, while the goalkeeper plays left and center with probability 0.15 and right with 0.7 (see section 5.1 for a motivation of these choices).

5.3.2 *Alternative 2*

This alternative is similar to the previous one but now I assume that the kicker plays left and center with probability 0.15 and right with 0.7 while the goalkeeper plays left with probability 0.7

and center and right with 0.15. I have chosen these probabilities because, under the assumption that the kicker is right-footed, his natural side is to shoot to the left hand side of the goalkeeper. Therefore, one could think that a naive goalkeeper is more likely to jump to the left, which justifies the probability of 0.7. However, the kicker may believe that the obvious reasoning of goalkeeper, given that he is facing a a right-footed kicker, is to jump to the left and therefore the probability of scoring will be low, so he will change the direction and kick to the right. That is why the chosen kicker's probability is 0.7.

5.3.3 *Power of tests*

I have only looked at the power of the LM and F tests of the null hypothesis of equal scoring probabilities in the LPM because they are the only ones whose asymptotic p-values are reliable under the null.

The following table shows the percentage of times that these tests reject the null at the 1, 5 and 10% significance level under alternatives 1 and 2.

	F-test		LM test	
%	Alternative 1	Alternative 2	Alternative 1	Alternative 2
1	22.42	24	18.25	19.64
5	36.56	38.63	36.28	38.2
10	44.89	47.27	45.47	47.86

These results suggest that empirical researchers should use the F-test of the LPM because it is the most powerful test under the two alternatives. The power here is higher than with only two actions because the alternatives are further away from the null.

5.4 *Alternatives to serial independence in 3×3 contingency tables*

The alternative of the hypothesis of serial independence is that the players actions at time t depend on the action chosen at time $t-1$. The transition matrix P of the Markov chain in this case is:

$$P = \begin{pmatrix} P_{LL} & P_{LC} & P_{LR} \\ P_{CL} & P_{CC} & P_{CR} \\ P_{RL} & P_{RC} & P_{RR} \end{pmatrix},$$

with states $i = L, C, R$, where $P_{iC} = 1 - P_{iL} - P_{iR}$ (see section 3.2).

Note that I can write the multivariate LPM in section 2 to detect serial dependence as the following vector autoregression:

$$\left.\begin{aligned} L_t - \pi_L &= b_{LL}\left[L_{t-1} - \pi_L\right] + b_{RL}\left[R_{t-1} - \pi_R\right] + u_{Lt} \\ R_t - \pi_R &= b_{LR}\left[L_{t-1} - \pi_L\right] + b_{RR}\left[R_{t-1} - \pi_R\right] + u_{Rt} \end{aligned}\right\},$$

where $E(L_t) = \pi_L$ and $E(R_t) = \pi_R$ are the average probabilities of kicking left and right respectively. Assume for simplicity that $b_{RL} = b_{LR} = 0$ under the alternative. Hence I get that

$P_{RL} = P_{CL} = P_{LL} - b_{LL}$ and $P_{LR} = P_{CR} = P_{RR} - b_{RR}$. These assumptions imply that the transition matrix is:

$$P = \begin{pmatrix} P_{LL} & 1 - P_{LL} - P_{RR} + b_{RR} & P_{RR} - b_{RR} \\ P_{LL} - b_{LL} & 1 - P_{LL} - P_{RR} + b_{LL} + b_{RR} & P_{RR} - b_{RR} \\ P_{LL} - b_{LL} & 1 - P_{LL} - P_{RR} + b_{LL} & P_{RR} \end{pmatrix}.$$

The stationary distribution in this Markov Chain is defined by the vector $\pi = \begin{pmatrix} \pi_L & \pi_C & \pi_R \end{pmatrix}$, where $\pi_C = 1 - \pi_L - \pi_R$, such that $\pi P = \pi$.

This yields:

$$\pi_L = \frac{P_{LL} - b_{LL}}{1 - b_{LL}} \text{ and } \pi_R = \frac{P_{RR} - b_{RR}}{1 - b_{RR}}.$$

Under the null hypotheses, b_{LL} and b_{RR} are equal to zero. Here, I propose different alternatives depending on the values of b_{LL} and b_{RR}. However, as in the 2×2 case, I have maintained that the stationary probabilities of the Markov chain equal the probabilities of the optimal strategy in the NE, so that the only discrepancy from the null is serial dependence.

5.4.1 *Power of tests*

I have only looked at the power of the F-version of the Lawley-Hotelling and Wilks' tests of the null hypothesis of serial independence in the multivariate LPM because they are the only ones whose asymptotic p-values are reliable under the null.

The following tables show the percentage of times that these tests reject the null at the 1, 5 and 10% significance level under the following alternatives: $b_{LL} = b_{RR} = 0.5$ (1), $b_{LL} = b_{RR} = -0.5$ (2), $b_{LL} = b_{RR} = 0.05$ (3) and $b_{LL} = b_{RR} = -0.05$ (4).

	Wilks' F-test				Lawley-Hotelling F-test			
%	(1)	(2)	(3)	(4)	(1)	(2)	(3)	(4)
1	23.30	50.67	1.08	1.49	24.55	50.82	1.37	1.84
5	40.37	92.31	4.42	5.64	40.92	86.03	4.79	6.03
10	50.65	97.48	8.48	10.60	51.25	97.09	8.74	10.92

These results suggest that empirical researchers should use the F-version of the Lawley-Hotelling test in the multivariate LPM because it is the most powerful test under the four alternatives, although it does not have much power under alternatives 3 and 4, as expected.

6 Empirical Application

In this section, I use the econometric methods described above to test if the empirical results obtained by Palacios-Huerta (2003) are still valid using a novel dataset I have constructed, which contains 549 penalty kicks. Moreover, I have expanded the actions of the players to Left, Center and Right for a presumably tougher test of the predictions of von Neumann's Minimax theorem.

In my dataset, there are 12 kickers with more than 20 penalty kicks each and another 11 kickers with at least 10 penalties. Similarly, there are 10 goalkeepers with more than 10 observations. The identities of goalies and kickers are shown in Appendix A.6.

The penalty data I have collected covers the period 2005-2015 from professional games in Spain, Italy, England and other European countries. The information comes from the following Spanish TV programs and internet pages: Estudio Estadio (TVE), GOL TV, Canal + Liga, El Dia Después (Movistar Plus), Deportes Cuatro, As.com and Marca.com. These TV programs and internet pages systematically review the best games played during the weekend, including all penalty kicks that take place in those games.

The data include the names of the teams involved in the match, the date of the match, the names of the kicker and goalkeeper for each penalty kick, the choices taken: Left (L), Center (C) and Right (R), the time within the match at which the penalty takes place, the score at the time of the penalty, the final score of the game, the foot used by the kicker (left or right) and the outcome of the kick (goal or miss).

The following table offers a basic description of the data with three actions.

Table 4
Distribution of strategies and scoring rates

	#Obs.	LL	LC	LR	CL	CC	CR	RL	RC	RR
All penalties	549	20.58	2.55	26.78	3.64	1.09	2.37	20.95	0.91	21.13
Scoring rate	86.34	69.91	92.86	97.96	100	0	92.31	95.65	100.00	78.45

In particular, it shows the relative proportions of different choices made by both kickers and goalkeepers (L, C or R). The first letter refers to the choice made by the kicker and the second to the choice made by the goalkeeper, both from the point of view of the goalkeeper. For instance, "LL" means that the kicker chooses to kick to the left hand side of the goalie and the goalie chooses to jump to his left.

The strategy followed by goalkeepers coincides with that followed by kickers in 42.8% of all penalties in the dataset. Kickers do not usually kick to the center (7.1% of all kicks), whereas goalies remain in the middle less often (4.55%). The percentage of kicks where the actions of the players do not coincide is mostly divided between LR (26.78%) and RL (20.95%). A goal is scored in 86.34% of all penalty kicks. The scoring rate is over 90% when the kicker choice differs from the goalie, and it is just over 65% when it coincides. These results confirm the empirical relevance of the theoretical model in section 3.

6.1 *Test of equal scoring probabilities*

To compare my results with the results obtained by Palacios-Huerta (2003), I initially consider only the actions he took into account (Left and Right) to test if the results he obtained are still

consistent with my more recent dataset. To do so, I eliminate the center strategy for all players. The results of the tests are described in Table 5.

(Table 5)

Of the 22 players in the sample, the null hypothesis of equal scoring probabilities across those two strategies is not rejected for any of the players.

Next, I carry out a stronger test on the first implication of the Minimax theorem by expanding the actions of the players to Left, Center and Right. The empirical description in Table 4 suggest that a model with three strategies is empirically more relevant. The results are shown in Table 6.

(Table 6)

These results show that the null hypothesis is rejected for one kicker and one goalkeeper at the 1% and 5% level, respectively. However, taking into account the large number of simultaneous tests that I calculate, the overall results still seem consistent with the null.

Hence, the empirical evidence on professional penalty kicks seems consistent with the first implication of the Minimax Theorem for most players but it is inconsistent with the theory for a couple of them when I include the additional action C.[6]

6.2 *Test for serial independence*

As I mentioned earlier, the second testable implication of the Minimax theorem states that the actions taken by the players must not be serially dependent. The results with two actions are shown in Table 7.

(Table 7)

This table shows the null hypotheses of serial independence with two actions is rejected for one kicker and one goalkeeper at the 10% significance level, and one additional goalkeeper at the 5% level.

As in section 6.1, I also expanded the actions of the players to Left, Center and Right. The results are shown in Table 8.

(Table 8)

In this case, the null hypothesis is only rejected for one goalkeeper at the 5% significance level. Nevertheless, if I consider the number of simultaneous tests that I compute, the overall results seem again consistent with the theory.

[6] The qualitative results obtained are similar when I rely on bootstrap p-values.

Therefore, the evidence that I find on penalty kicks is consistent with the second implication of the Minimax theorem, which is perhaps not surprising because actual penalty kicks usually take place weeks if not months apart. These findings suggest that professional soccer players seem truly able to generate random sequences; they do not appear to switch strategies too often or too seldom.

7 Conclusions

In this paper I study independence tests between two categorical variables, which only take a finite number of values H and K, respectively. The results I find can be applied to different scenarios, such as testing the ability of some popular market timing strategies to predict the direction of the movements in some well-known asset prices, or in the analysis of mixed strategies equilibrium in games.

From the econometric point of view, I prove the numerical equivalence for general H and K between Pearson's independence test in contingency tables, the Lagrange Multiplier test in several popular regression models: the multivariate version of the LPM, the conditional multinomial model, the multinomial logit and probit models; and the corresponding J-test for overidentifying restrictions. Additionally, the same results holds if one exchanges regressors and regressands in all these models. This result also has the advantage that there will only be one bootstrap version for all these different tests. Similarly, the Monte Carlo experiments previously reported in the literature on contingency table tests also apply to all the other different tests, so they could be combined.

I also prove that the heteroskedasticity-robust version of the Wald test in the multivariate LPM is numerically identical to the Wald test in the conditional multinomial model. In addition, the values of the log-likelihood function of the multinomial logit and probit models under the null and the alternative are equal to the corresponding log-likelihoods of the conditional multinomial model, so the LM and LR tests coincide, but the Wald tests can be very different.

Additionally, I study the size and power in finite samples for all the different tests of independence that have been proposed. In this regard, I explain how to obtain the exact p-value in finite samples, as well as how to calculate the exact bootstrap distribution without resorting to simulations in the 2×2 case. Also, I take into consideration that the usual formulas of some of the tests may breakdown in the simulations due to: perfect classification, perfect fit, single outcome and single choice. For that reason, I obtain alternative expressions, which remain valid in those cases.

When testing the Minimax theorem implications in penalty kick games when players only

have two actions, I find that the Monte Carlo p-values are extremely close to the exact p-values for 5 and 20 observations. In contrast, some of the asymptotic p-values are unreliable. In the case of three actions, I only compare Monte Carlo and asymptotic p-values using p-value plots because finding the exact distribution of the tests is very tedious.

My results indicate that the LPM usually offers more reliable inferences than the logit model, especially if one uses the F-versions of the different tests. The F-tests also tend to have the largest power. In contrast, all the Wald tests are very unreliable. Unfortunately, tests of equal scoring probabilities with two actions do not have as much power as one would desire.

From the empirical point of view, I find that most professional soccer players behave consistently with the first implication of the Minimax theorem (equal scoring probabilities across strategies) with two and three actions. Additionally, I find that the second implication of the Minimax theorem (player's actions are serially independent) is consistent for most of the players in the sample.

Although the analysis of this paper indicates the most reliable econometric methodology to test the implications of the Minimax theorem in finite samples for empirically relevant values of the payoff matrix, there is still much to learn about testing those implications in experimental situations using independence tests.

Finally, Anatolyev and Kosenok (2009) showed that the Pearson test for goodness of fit is also asymptotically equivalent to a Wald test in a multivariate regression. This suggests that the numerical results I have derived in this paper carry over to this context. Although proving this conjecture is beyond the scope of the current paper, it is an interesting avenue for subsequent research.

References

Anatolyev, S. and G. Kosenok (2009): "Tests in Contingency Tables as Regression Tests", Economics Letters, 105, pp. 189–192.

Berndt, E.R. and Savin, N.E. (1977): "Conflict Among Criteria for Testing Hypotheses in the Multivariate Linear Regression Model", Econometrica, 45, pp. 1263-1277.

Brown, J. and R. Rosenthal (1990): "Testing the Minimax Hypothesis: A Reexamination of O'Neill's Experiment", Econometrica, 58, pp. 1065-1081.

Cameron, A. C. and P. K. Trivedi (2005): Microeconometrics: Methods and Applications, Cambridge University Press.

Chiappori, P-A, S. Levitt and T. Groseclose (2002): "Testing Mixed-Strategy Equilibria When Players Are Heterogeneous: The Case of Penalty Kicks in Soccer", American Economic Review, 92, pp. 1138-1151.

Davidson, R. and J. MacKinnon (1998): "Graphical Methods for Investigating the Size and Power of Hypothesis Tests", The Manchester School, 66, pp. 1-26.

Engle, R.F (1983): "Wald, Likelihood Ratio and Lagrange Multiplier Tests in Econometrics", The Handbook of Econometrics, 2, eds. Z. Griliches and M.D. Intriligator, chapter 13, pp. 775-826.

Fisher, R. A (1922): "On the Interpretation of $\mathcal{X}^2$ from Contingency Tables, and the Calculation of P", Journal of the Royal Statistical Society, 85, pp. 87-94.

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

Henriksson, R.D. and R.C. Merton (1981): "On Market Timing and Investment Performance. II. Statistical Procedures for Evaluating Forecasting Skills", Journal of Business, 54, pp. 513-533.

Judge, G. , W. Griffiths, H. Lütkepohl, T-C. Lee and R. C. Hill (1985): The Theory and Practice of Econometrics, Wiley, Second edition.

Lee, T., G. Judge and A. Zellner (1968): "Maximum Likelihood and Bayesian Estimation of Transition Probabilities", Journal of the American, Statistical Association, 63, pp. 1162-1179.

Magnus, J. (2007): "The Asymptotic Variance of the Pseudo Maximum Likelihood Estimator", Econometric Theory, 23, pp. 1022–1032.

Magnus, J. and H. Neudecker (1988): Matrix Differential Calculus with Applications in Statistics and Econometrics, Wiley.

Mood, A., F-A. Graybill and D. Boes (1974): Introduction to the Theory of Statistics, Third edition, McGraw-Hill.

Newey, W.K. and K.D. West (1987):. "Hypothesis Testing with Efficient Method of Moments Estimation", International Economic Review, 28, pp. 777-787.

Osborne, J. (2003): An Introduction to Game Theory, Oxford University Press.

Palacios-Huerta, I. (2003): "Professionals Play Minimax", Review of Economic Studies 70, pp. 395-415.

Palacios-Huerta, I. (2017): "Strictly Competitive Strategic Situations", Economia Industrial, 403, pp. 19-28.

Pesaran, M.H and A. Timmermann (1994): "A Generalization of the Non-Parametric Henriksson-Merton Test of Market Timing", Economics Letters, 44, pp. 1-7.

Ruud, P. (2000): An Introduction to Classical Econometric Theory, Oxford University Press.

Sentana, J. (2016): "Mixed Strategy Equilibrium: A Field Experiment with Penalty Kicks", Unpublished Master dissertation, University College London

Sentana, J. (2019): "Mixed Strategy Equilibrium: A Field Experiment with Penalty Kicks", Working Paper, University of Essex.

Sherman, J. and W.J. Morrison (1950): "Adjustment of an Inverse Matrix Corresponding to a Change in One Element of a Given Matrix", Annals of Mathematical Statistics, 21, pp. 124-127.

StataCorp. (2011): Stata 12 Base Reference Manual, Stata Press, College Station, TX.

Stewart, K.G. (1995): "The Functional Equivalence of the W, LR and LM statistics", Economic Letters, 49, pp. 109-112.

Wooldridge, J. (2002): Introductory Econometrics: A Modern Approach, Second Edition, South-Western.

Appendix

A Proofs and Auxiliary Results

A.1 *Proof of Proposition 1*

As mentioned in section 2, to test the hypothesis of independence, one can use a multivariate regression, a multinomial logit and probit model, a conditional multinomial model as well as GMM and the classical contingency test.

A.1.1 *Contingency table test*

There are two type of contingency tables:

1) the expected one, which would satisfy the null hypothesis,

$\tilde{y}\backslash x$	A_1	...	A_K	Sum	$\hat{\pi}_x$
B_1	$\frac{n_{*1}\times n_{\diamond 1}}{n}$	...	$\frac{n_{*1}\times n_{H\diamond}}{n}$	$n_{1\diamond}$	$\frac{n_{1\diamond}}{n}$
...	...	...	...	...	...
B_H	$\frac{n_{*K}\times n_{1\diamond}}{n}$	...	$\frac{n_{*K}\times n_{H\diamond}}{n}$	$n_{H\diamond}$	$\frac{n_{H\diamond}}{n}$
Sum	n_{*1}	...	n_{*K}	n	
$\hat{\pi}_{\tilde{y}}$	$\frac{n_{*1}}{n}$	...	$\frac{n_{*K}}{n}$		1

and 2) the actual one

$\tilde{y}\backslash x$	A_1	...	A_K	Sum	$\hat{\pi}_x$
B_1	n_{11}	...	n_{1K}	$n_{1\diamond}$	$\frac{n_{1\diamond}}{n}$
...	...	...	...	...	...
B_H	n_{H1}	...	n_{HK}	$n_{H\diamond}$	$\frac{n_{H\diamond}}{n}$
Sum	n_{*1}	...	n_{*K}	n	
$\hat{\pi}_{\tilde{y}}$	$\frac{n_{*1}}{n}$	...	$\frac{n_{*K}}{n}$		1

Pearson test For my purposes, the test statistic (1) can be conveniently written as:

$$Pearson = \sum_{k=1}^{K}\sum_{h=1}^{H-1}\frac{\left(n_{hk}-\frac{n_{*k}n_{h\diamond}}{n}\right)^2}{\frac{n_{*k}n_{h\diamond}}{n}} + \sum_{k=1}^{K}\frac{\left(n_{Hk}-\frac{n_{*k}n_{H\diamond}}{n}\right)^2}{\frac{n_{*k}n_{H\diamond}}{n}}.$$

Note that $n_{H\diamond} = n - \sum_{h=1}^{H-1} n_{h\diamond}$ and $n_{Hk} = n_{*k} - \sum_{h=1}^{H-1} n_{hk}$. Therefore,

$$\begin{aligned} Pearson &= n\sum_{k=1}^{K}\sum_{h=1}^{H-1}\frac{\left(n_{hk}-\frac{n_{*k}n_{h\diamond}}{n}\right)^2}{n_{*k}n_{h\diamond}} + n\sum_{k=1}^{K}\frac{\left(\sum_{h=1}^{H-1}\left(n_{hk}-\frac{n_{*k}n_{h\diamond}}{n}\right)\right)^2}{n_{*k}n_{H\diamond}} \\ &= n\sum_{k=1}^{K}\sum_{h=1}^{H-1}\frac{\left(n_{hk}-\frac{n_{*k}n_{h\diamond}}{n}\right)^2}{n_{*k}n_{h\diamond}} + n\sum_{k=1}^{K}\sum_{h=1}^{H-1}\frac{\left(n_{hk}-\frac{n_{*k}n_{h\diamond}}{n}\right)^2}{n_{*k}n_{H\diamond}} \\ &\quad +2n\sum_{k=1}^{K}\sum_{h=1}^{H-1}\sum_{m=h+1}^{H-1}\left(\frac{\left(n_{hk}-\frac{n_{*k}n_{h\diamond}}{n}\right)\left(n_{mk}-\frac{n_{*k}n_{m\diamond}}{n}\right)}{n_{*k}n_{H\diamond}}\right), \end{aligned} \tag{A1}$$

for all $k = 1, ..., K$, $h, m = 1, ..., H-1$ and $h \neq m$.

It is important to mention that if I change x and $\tilde{y}$ so that x now takes values $B_1, ..., B_H$ and $\tilde{y}$ takes values $A_1, ..., A_K$, then the contingency table will be transposed but the independence test will not change.

A.1.2 *Multivariate regression*

Recall from section 2.2 that the model to be estimated is:

$$\left.\begin{array}{c} B_{1i} = \delta_{11} A_{1i} + ... + \delta_{1K} A_{Ki} + u_{1i} \\ \vdots \\ B_{H-1i} = \delta_{H-1,1} A_{1i} + ... + \delta_{H-1,K} A_{Ki} + u_{H-1i} \end{array}\right\}.$$

The residual variance-covariance matrix is defined as $\Sigma = E(u_i u_i')$. It can be estimated as

$$\hat{\Sigma}_U = \frac{1}{n}\begin{pmatrix} \hat{u}_1'\hat{u}_1 & ... & \hat{u}_1'\hat{u}_{H-1} \\ ... & ... & ... \\ \hat{u}_{H-1}'\hat{u}_1 & ... & \hat{u}_{H-1}'\hat{u}_{H-1} \end{pmatrix} = \frac{1}{n}\begin{pmatrix} \Sigma_{i=1}^n \hat{u}_{1i}^2 & ... & \Sigma_{i=1}^n \hat{u}_{1i}\hat{u}_{H-1i} \\ ... & ... & ... \\ \Sigma_{i=1}^n \hat{u}_{1i}\hat{u}_{H-1i} & ... & \Sigma_{i=1}^n \hat{u}_{H-1i}^2 \end{pmatrix},$$

where $\hat{u}_h'\hat{u}_m = B_h' B_m - B_h' X (X'X)^{-1} X' B_m$, $\forall h, m = 1, ..., H-1$ and $\hat{u}_h = B_h - X\hat{\delta}_h$.

Using the previous expression, the unrestricted covariance matrix estimator is:

$$\hat{\Sigma}_U = \frac{1}{n}\begin{pmatrix} n_{1\diamond} - \sum_{k=1}^K \frac{n_{1k}^2}{n_{*k}} & -\sum_{k=1}^K \frac{n_{1k}n_{2k}}{n_{*k}} & ... & -\sum_{k=1}^K \frac{n_{1k}n_{H-1k}}{n_{*k}} \\ -\sum_{k=1}^K \frac{n_{1k}n_{2k}}{n_{*k}} & ... & ... & ... \\ ... & ... & ... & ... \\ -\sum_{k=1}^K \frac{n_{1k}n_{H-1k}}{n_{*k}} & ... & ... & n_{H-1\diamond} - \sum_{k=1}^K \frac{n_{H-1k}^2}{n_{*k}} \end{pmatrix},$$

where $n_{h\diamond} = \Sigma_{i=1}^n B_{hi}$, $h = 1, ..., H-1$.

The null hypothesis of independence implies that $\delta_{h1} = ... = \delta_{hK} = \delta_h$. Using that $\sum_{k=1}^K A_{ki} = 1$, the model becomes:

$$\left.\begin{array}{c} B_1 = \delta_1 + u_1 \\ \vdots \\ B_{H-1} = \delta_{H-1} + u_{H-1} \end{array}\right\}.$$

The estimated variance covariance matrix under the null is:

$$\tilde{\Sigma}_R = \frac{1}{n}\begin{pmatrix} n_{1\diamond}(1 - \frac{n_{1\diamond}}{n}) & -\frac{n_{1\diamond}n_{2\diamond}}{n} & ... & -\frac{n_{1\diamond}n_{H-1\diamond}}{n} \\ -\frac{n_{1\diamond}n_{2\diamond}}{n} & ... & ... & ... \\ ... & ... & ... & ... \\ -\frac{n_{1\diamond}n_{H-1\diamond}}{n} & ... & ... & n_{H-1\diamond}(1 - \frac{n_{H-1\diamond}}{n}) \end{pmatrix}.$$

Test procedures The contribution from observation i to the log-likelihood function of the multivariate regression model is:

$$\mathcal{L}_i = -\frac{1}{2}\ln 2\pi - \frac{1}{2}\ln|\Sigma| - \frac{1}{2}(y_i - \Pi x_i)'\Sigma^{-1}(y_i - \Pi x_i).$$

In matrix notation, the score of the full sample can be written as:

$$S_\delta = \left(X'Y - X'X\Pi'\right)\Sigma^{-1},$$

where $\Sigma = \frac{Y'MY}{n}$ with $M = I - X(X'X)^{-1}X'$ (see Magnus (2007) for details).

Note that $Y'Y = \begin{pmatrix} n_{1\diamond} & 0 & ... & 0 \\ 0 & ... & 0 & ... \\ ... & ... & ... & 0 \\ 0 & ... & 0 & n_{H-1\diamond} \end{pmatrix}$ and $X'Y = \begin{pmatrix} n_{11} & ... & n_{H-1,1} \\ ... & ... & ... \\ n_{1K} & ... & n_{H-1,K} \end{pmatrix}$.

Under the null,

$$\tilde{\Sigma}_R = \frac{1}{n}\left(Y'Y - Y'l_n(l_n'l_n)\,l_n'Y\right),$$

where l_n is an $(H-1)\times 1$ vector of ones.

Note that $Y'l_n = \begin{pmatrix} n_{1\diamond} \\ ... \\ n_{H-1\diamond} \end{pmatrix}$, $l_n'Y = (Y'l_n)'$ and $l_n'l_n = n$.

Hence, $\tilde{\Sigma}_R$ can also be written as:

$$\tilde{\Sigma}_R = \frac{G + wr'}{n},$$

where $G = Y'Y$, $w = -Y'l_n$ and $r' = \frac{1}{n}l_n'Y$.

Using the Sherman-Morrison (1950) formula:

$$\tilde{\Sigma}_R^{-1} = n\left(G^{-1} - \frac{G^{-1}wr'G^{-1}}{1 + r'G^{-1}w}\right),$$

which yields

$$\tilde{\Sigma}_R^{-1} = n\left(\left(Y'Y\right)^{-1} + \frac{l_nl_n'}{n_{H\diamond}}\right) = \begin{pmatrix} \frac{n}{n_{1\diamond}} + \frac{n}{n_{H\diamond}} & \frac{n}{n_{H\diamond}} & ... \\ \frac{n}{n_{H\diamond}} & ... & ... \\ ... & ... & \frac{n}{n_{H-1\diamond}} + \frac{n}{n_{H\diamond}} \end{pmatrix}.$$

Recall that under the null, $\delta_{h1} = ... = \delta_{hK} = \delta_h$, $h = 1, ..., H-1$, and since $\tilde{\delta}_h^{OLS} = \frac{n_{h\diamond}}{n}$, then

$$\tilde{\Pi}_R' = l_k(l_n'l_n)^{-1}l_n'Y,$$

with l_k being a $K\times 1$ vector of ones.

Hence,

$$\left(X'Y - X'X\tilde{\Pi}_R'\right) = \begin{pmatrix} n_{11} - \frac{n_{*1}n_{1\diamond}}{n} & ... & n_{H-1,1} - \frac{n_{*1}n_{H-1\diamond}}{n} \\ ... & ... & ... \\ n_{1K} - \frac{n_{*K}n_{1\#}}{n} & ... & n_{H-1K} - \frac{n_{*K}n_{H-1\diamond}}{n} \end{pmatrix}.$$

Recall that $S_\delta = \left(X'Y - X'X\Pi'\right)\Sigma^{-1}$. The element k, h of S_δ, $k = 1, ..., K$, $h, z = 1, ..., H-1$ and $h \neq z$, is:

$$\left(n_{hk} - \frac{n_{*k}n_{h\#}}{n}\right)\left(\frac{n}{n_{h\#}} + \frac{n}{n_{H\#}}\right) + \sum_{z=1}^{H-1}\left(\left(n_{zk} - \frac{n_{*k}n_{z\#}}{n}\right)\left(\frac{n}{n_{H\#}}\right)\right)$$
$$= \frac{n_{hk}n}{n_{h\diamond}} + \frac{n_{hk}n}{n_{H\diamond}} - n_{*k} - \frac{n_{*k}n_{h\diamond}}{n_{H\diamond}} + \frac{n}{n_{H\diamond}}\sum_{z=1}^{H-1}n_{zk} - \frac{n_{*k}}{n_{H\diamond}}\sum_{z=1}^{H-1}n_{z\diamond}.$$

Note that $\sum_{z=1}^{H-1}n_{zk} = n_{*k} - n_{hk} - n_{Hk}$ and $\sum_{z=1}^{H-1}n_{z\diamond} = n - n_{h\diamond} - n_{H\diamond}$. Therefore,

$$S_{\delta_{hk}} = n\left(\frac{n_{hk}}{n_{h\diamond}} - \frac{n_{Hk}}{n_{H\diamond}}\right).$$

Note that in the multivariate regression, $S_\delta = I\left(X'Y - X'Xl_k(l_n'l_n)^{-1}l_n'Y\right)\Sigma^{-1}$and $\mathcal{I} = \left(\Sigma^{-1}\otimes(X'X)\right)$.

Given that $vec(ABC) = (C'\otimes A)vec(B)$, then

$$vec(S_\delta) = \left(\Sigma^{-1}\otimes I\right)vec\left(X'Y - X'Xl_k(l_n'l_n)^{-1}l_n'Y\right).$$

The LM test is defined as:

$$LM = S_\delta'\mathcal{I}^{-1}S_\delta,$$

so

$$LM = vec'\left(X'Y - X'Xl_k(l_n'l_n)^{-1}l_n'Y\right)\left(\Sigma^{-1}\otimes I\right)\left(\Sigma\otimes(X'X)^{-1}\right)$$
$$\left(\Sigma^{-1}\otimes I\right)vec\left(X'Y - X'Xl_k(l_n'l_n)^{-1}l_n'Y\right).$$

Hence,

$$LM = vec'\left(X'Y - X'Xl_k(l_n'l_n)^{-1}l_n'Y\right)\left(\Sigma^{-1}\otimes(X'X)^{-1}\right)vec\left(X'Y - X'Xl_k(l_n'l_n)^{-1}l_n'Y\right)$$

due to the properties of the Kronecker product.

Define $F = \begin{pmatrix} F_1 & ... & F_{H-1}\end{pmatrix} = X'Y - X'Xl_k(l_n'l_n)^{-1}l_n'Y$, so

$$LM = vec'(F)\left(\Sigma^{-1}\otimes(X'X)^{-1}\right)vec(F). \tag{A2}$$

This yields

$$LM = n\sum_{k=1}^{K}\sum_{h=1}^{H-1}\left(\frac{\left(n_{hk} - \frac{n_{*k}n_{h\diamond}}{n}\right)^2}{n_{*k}n_{h\diamond}}\right) + n\sum_{k=1}^{K}\sum_{h=1}^{H-1}\left(\frac{\left(n_{hk} - \frac{n_{*k}n_{h\diamond}}{n}\right)^2}{n_{*k}n_{H\diamond}}\right)$$
$$+2n\sum_{k=1}^{K}\sum_{h=1}^{H-1}\sum_{m=h+1}^{H-1}\left(\frac{\left(n_{hk} - \frac{n_{*k}n_{h\diamond}}{n}\right)\left(n_{mk} - \frac{n_{*k}n_{m\diamond}}{n}\right)}{n_{*k}n_{H\diamond}}\right),$$

for all $k = 1, ..., K$, $h, m = 1, ..., H-1$ and $h \neq m$.

Therefore, this LM test is exactly the same as (A1).

A.1.3 *Multinomial model*

Recall the likelihood for the sample is:

$$\mathcal{L} = \prod_{k=1}^{K}\left[(1 - \sum_{h=1}^{H-1}P_{hk})^{\left(n_{*k} - \sum_{h=1}^{H-1}n_{hk}\right)}\prod_{h=1}^{H-1}P_{hk}^{n_{hk}}\right](1 - \sum_{k=1}^{K-1}\pi_{*k})^{n_{*K}}\prod_{k=1}^{K-1}\pi_{*k}^{n_{*k}},$$

so under the alternative we get:

$$\begin{aligned}\ln\mathcal{L} \;=\; & \sum_{k=1}^{K}\left[\left(n_{*k} - \sum_{h=1}^{H-1}n_{hk}\right)\ln(1 - \sum_{h=1}^{H-1}P_{hk}) + \sum_{h=1}^{H-1}(n_{hk}\ln P_{hk})\right] \\ & + n_{*K}\ln(1 - \sum_{k=1}^{K-1}\pi_{*k}) + \sum_{k=1}^{K-1}n_{*k}\ln\pi_{*k}.\end{aligned}$$

The derivatives of the log-likelihood with respect to P_{hk} and π_{*k} are:

$$\frac{\partial \ln \mathcal{L}}{\partial P_{hk}} = \frac{n_{hk}}{P_{hk}} - \frac{\left(n_{*k} - \sum_{h=1}^{H-1} n_{hk}\right)}{(1 - \sum_{h=1}^{H-1} P_{hk})}$$
$$\frac{\partial \ln \mathcal{L}}{\partial \pi_{*k}} = \frac{n_{*k}}{\pi_{*k}} - \frac{n_{*K}}{(1 - \sum_{k=1}^{K-1} \pi_{*k})}$$

for $k = 1, ..., K$ and $h = 1, ..., H-1$.

Therefore, the FOC are:

$$\frac{n_{hk}}{\hat{P}_{hk}} - \frac{\left(n_{*k} - \sum_{h=1}^{H-1} n_{hk}\right)}{(1 - \sum_{h=1}^{H-1} \hat{P}_{hk})} = 0$$
$$\frac{n_{*k}}{\hat{\pi}_{*k}} - \frac{n_{*K}}{(1 - \sum_{k=1}^{K-1} \hat{\pi}_{*k})} = 0$$

which yields

$$\hat{P}_{hk} = \frac{n_{hk}}{n_{*k}} \text{ and } \hat{\pi}_{*k} = \frac{n_{*k}}{n_{*K}}.$$

In contrast. under the null, which states that $H_0 : P_{hk} = P_{h\diamond},$, for $k = 1, ..., K$, the log-likelihood function is

$$\ln \mathcal{L} = \sum_{h=1}^{H-1} n_{h\diamond} \ln P_{h\diamond} + \left(n - \sum_{h=1}^{H-1} n_{h\diamond}\right) \ln(1 - \sum_{h=1}^{H-1} P_{h\diamond}).$$

Taking first derivatives with respect to $P_{h\diamond}$ yields:

$$\frac{\partial \ln \mathcal{L}}{\partial P_{h\diamond}} = \frac{n_{h\diamond}}{P_{h\diamond}} - \frac{\left(n - \sum_{h=1}^{H-1} n_{h\diamond}\right)}{(1 - \sum_{h=1}^{H-1} P_{h\diamond})},$$

for $k = 1, ..., K$ and $h = 1, ..., H-1$.

Therefore, the FOC is:

$$\frac{n_{h\diamond}}{\tilde{P}_{h\diamond}} - \frac{\left(n - \sum_{h=1}^{H-1} n_{h\diamond}\right)}{(1 - \sum_{h=1}^{H-1} \tilde{P}_{h\diamond})} = 0,$$

which yields

$$\tilde{P}_{h\diamond} = \frac{n_{h\diamond}}{n},$$

so $\tilde{P}_{h\diamond} = \tilde{\delta}_h$ for $h = 1, ..., H-1$.

The Hessian of this multinomial model is:

$$H(\theta) = \frac{\partial^2 \mathcal{L}(\theta)}{\partial \theta \theta'},$$

for $\theta = (P_{11}, ..., P_{H-1,K})$, so

$$H(\theta) = \begin{pmatrix} \frac{\partial^2 \mathcal{L}}{\partial P_{11}^2} & ... & \frac{\partial^2 \mathcal{L}}{\partial P_{11} \partial P_{H-1,1}} & 0 & ... & ... & ... & ... \\ ... & ... & ... & 0 & ... & ... & ... & ... \\ \frac{\partial^2 \mathcal{L}}{\partial P_{11} \partial P_{K-1,1}} & ... & \frac{\partial^2 \mathcal{L}}{\partial P_{H-1,1}^2} & 0 & ... & ... & ... & ... \\ 0 & 0 & 0 & ... & ... & 0 & ... & ... \\ ... & ... & ... & ... & ... & 0 & 0 & 0 \\ ... & ... & ... & & & \frac{\partial^2 \mathcal{L}}{\partial P_{1K}^2} & ... & \frac{\partial^2 \mathcal{L}}{\partial P_{1K} \partial P_{H-1,K}} \\ ... & ... & ... & ... & 0 & ... & ... & ... \\ 0 & ... & ... & ... & 0 & \frac{\partial^2 \mathcal{L}}{\partial P_{1K} \partial P_{H-1,K}} & ... & \frac{\partial^2 \mathcal{L}}{\partial P_{H-1,K}^2} \end{pmatrix}$$

where

$$\frac{\partial^2 \mathcal{L}}{\partial P_{hk}^2} = -\frac{n_{hk}}{P_{hk}^2} - \frac{n_{Hk}}{1 - \sum_{h=1}^{H-1} P_{hk}}$$
$$\frac{\partial^2 \mathcal{L}}{\partial P_{1k} \partial P_{H-1,k}} = -\frac{n_{Hk}}{(1 - \sum_{h=1}^{H-1} P_{hk})^2}$$

for $k = 1, ..., K$ and $h = 1, ..., H-1$.

For a correctly specified likelihood, we have the information equality:

$$Var[s(\theta)] = -E[H(\theta)] = \mathcal{I}(\theta).$$

The matrix $\mathcal{I}(\theta)$ is defined as follows:

$$\mathcal{I}(\theta) = \begin{pmatrix} \mathcal{I}_1(\theta) & 0 & 0 \\ 0 & ... & 0 \\ 0 & 0 & \mathcal{I}_K(\theta) \end{pmatrix},$$

where

$$\mathcal{I}_k(\theta) = nE(A_{ki}) \begin{pmatrix} \frac{1}{P_{1k}} + \frac{1}{(1 - \sum_{h=1}^{H-1} P_{hk})} & \frac{1}{(1 - \sum_{h=1}^{2} P_{hk})} & ... & \frac{1}{(1 - \sum_{h=1}^{H-1} P_{hk})} \\ \frac{1}{(1 - \sum_{h=1}^{2} P_{hk})} & ... & ... & ... \\ ... & ... & ... & ... \\ \frac{1}{(1 - \sum_{h=1}^{H-1} P_{hk})} & ... & ... & \frac{1}{P_{H-1,k}} + \frac{1}{(1 - \sum_{h=1}^{H-1} P_{hk})} \end{pmatrix}.$$

Hence, its inverse will be given by:

$$\mathcal{I}_k(\theta)^{-1} = \frac{1}{nE(A_{ki})} \begin{pmatrix} P_{1k}(1 - P_{1k}) & -P_{1k}P_{2k} & ... & -P_{1k}P_{H-1,k} \\ -P_{2k}P_{1k} & ... & ... & ... \\ ... & ... & ... & ... \\ -P_{H-1,k}P_{1k} & ... & ... & P_{H-1,k}(1 - P_{H-1,k}) \end{pmatrix}.$$

Note that the score under the null is

$$s(\tilde{\theta}) = n \begin{pmatrix} \frac{n_{11}}{n_{1\diamond}} - \frac{n_{H1}}{n_{H\diamond}} & ... & \frac{n_{H-1.K}}{n_{H=1\diamond}} - \frac{n_{HK}}{n_{H\diamond}} \end{pmatrix}',$$

which is the same as the element h, k of the score under the null of the multivariate regression model, except that these scores are calculated by vectorizing the matrix P by columns while the ones in the multivariate regression are vectorized by rows.

Thus, one can go from one to another using the commutation matrix (see Magnus and Neudecker (1988) for more details). The most useful property of such matrix is that it allows the Kronecker products to commute. For that reason, the information matrix of the multivariate regression and the one in the multinomial model look as a mirror image of one another, i.e. $\mathcal{I} = \left((X'X) \otimes \Sigma^{-1}\right)$ instead of $\left(\Sigma^{-1} \otimes (X'X)\right)$. Given that after considering the re-ordering, the score and information matrix under the null are identical to the score and information matrix of the multivariate regression, the LM tests will also be numerically equal.

A.1.4 *Multinomial logit model*

Recall from section 2.4 that the log-likelihood function of this model is:

$$\mathcal{L}(\gamma) = \sum_{k=1}^{K}\{[\sum_{h=1}^{H-1} n_{hk} \ln P_{hk} + n_{Hk} \ln(1 - \sum_{h=1}^{H-1} P_{hk})]\}.$$

The score is defined as:

$$s(\gamma) = \frac{\partial \mathcal{L}(\gamma)}{\partial \gamma}$$

with

$$\frac{\partial \mathcal{L}}{\partial P_{hk}} = \left(\frac{n_{hk}}{P_{hk}} - \frac{n_{Hk}}{1 - \sum_{h=1}^{H-1} P_{hk}} \right)$$

and

$$\frac{\partial P_{hk}}{\partial \gamma_{hk}} = \frac{\exp(\gamma_{hk} A_{ki}) A_{ki}}{[1 + \exp(\gamma_{hk} A_{ki}) + \exp(\gamma_{kh} A_{ki})]^2},$$

for $k = 1, ..., K$ and $h = 1, ..., H-1$.

Solving for $\hat{P}_{hk}$ yields:

$$\hat{P}_{hk} = \frac{n_{hk}}{n_{*k}},$$

which is again the same as the estimate in the multivariate regression.

Moreover, the Hessian of the log-likelihood function is defined as:

$$H(\gamma) = \frac{\partial^2 \mathcal{L}(\gamma)}{\partial \gamma \gamma'},$$

where

$$\frac{\partial^2 \mathcal{L}}{\partial P_{hk}^2} = -\frac{n_{hk}}{P_{hk}^2} - \frac{n_{Hk}}{(1 - \sum_{h=1}^{H-1} P_{hk})^2}$$

$$\frac{\partial^2 \mathcal{L}}{\partial P_{1k} P_{H-1k}} = -\frac{n_{Hk}}{(1 - \sum_{h=1}^{H-1} P_{hk})^2}$$

with

$$\frac{\partial P_{hk}^2}{\partial \gamma_{hk}^2} = \frac{\exp(\gamma_{hk}A_{ki})A_{ki}[A_{ki}(1+\exp(\gamma_{kh}A_{ki}))(1+\exp(\gamma_{hk}A_{ki})+\exp(\gamma_{kh}A_{ki}))^2 - \exp(\gamma_{kh}A_{ki})]}{[1+\exp(\gamma_{hk}A_{ki})+\exp(\gamma_{kh}A_{ki})]^4},$$

for $k = 1, ..., K$ and $h = 1, ..., H-1$. (see Cameron and Trivedi (2005) chapter 15, section 4 for more details).

The matrix $\mathcal{I}(\gamma)$ is defined as follows:

$$\mathcal{I}(\gamma) = \begin{pmatrix} \mathcal{I}_1(\gamma) & 0 & 0 \\ 0 & ... & 0 \\ 0 & 0 & \mathcal{I}_K(\gamma) \end{pmatrix}.$$

We can again use this matrix to obtain a Wald test of independence.

The multinomial logit log-likelihood function under the null hypothesis $H_0 : P_{hk} = P_h$, for $k = 1, ..., K$ and $h = 1, ..., H-1$, is:

$$\mathcal{L}(\phi) = \sum_{k=1}^{K}\{\sum_{h=1}^{H-1}(n_{hk}\ln P_h) + n_{Hk}\ln(1-\sum_{h=1}^{H-1}P_h)] + n_{*k}\ln\pi_k\},$$

which yields

$$\tilde{P}_h = \frac{n_{h\diamond}}{n},$$

given that $\Sigma_{h=1}^{K} n_{h\diamond} = n$.

Here, the relationship between $\tilde{P}_h$ and $\tilde{\gamma}_h$ is:

$$\tilde{P}_h = \frac{\exp(\tilde{\gamma}_h)}{\sum\exp(\tilde{\gamma}_h)}.$$

It is worth mentioning that the values of the log-likelihood function under the null and alternative of the multinomial logit are equal to the corresponding log-likelihoods of the conditional multinomial model, which implies that the LM and LR test will be the same (see section 17.4 of Ruud (2000)).

A.1.5 *Multinomial probit model*

Recall from section 2.5 that the observation rule is

$$B_{hi} = 1\left\{B_{hi}^* = \max_{j=1,...,n} B_{hj}^*\right\},$$

where $1\{\}$ is the indicator function. Hence, the i^{th} element of $B_h \equiv (B_{hi},\ i = 1, .., n)'$ equals 1 if the i^{th} value of the categorical variable $\tilde{y}$ is observed. Otherwise, B_{hi} equals zero. Therefore, the log-likelihood function is

$$L(\theta) = \sum_{i=1}^{n} B_{hi}\ln\Pr(B_{hi} = 1|x).$$

Once again, this log-likelihood function coincides with the conditional component of the log-likelihood function of the multinomial model. As a result, the same derivation as in the multinomial logit apply.

Similarly, the multinomial probit model under the null is entirely analogous to the multinomial logit one, so the same numerical equalities hold.

A.1.6 *GMM*

As I explained in section 2.6,

$$E\left[g\left(z_i;\Pi\right)\right] = E\left[\left(y_i - \Pi x_i\right) \otimes x_i\right] = 0.$$

Under H_1, Π is unrestricted while under H_0, $\Pi(\upsilon) = \upsilon l_k'$.

The GMM estimator is defined as:

$$\tilde{\upsilon} = \arg\min_{\upsilon} \left[\frac{1}{n}\sum_{i=1}^{n}\left(\left(y_i - \Pi(\upsilon)x_i\right) \otimes x_i\right)\right]' \Upsilon^{-1} \left[\frac{1}{n}\sum_{i=1}^{n}\left(\left(y_i - \Pi(\upsilon)x_i\right) \otimes x_i\right)\right],$$

where Υ is a symmetric positive definite $(K \times (H-1)) \times (K \times (H-1))$ weight matrix. To obtain the unrestricted estimator of $\delta = vec(\Pi')$, one could do analogous calculations.

Given that GMM estimator can also be written as

$$\tilde{\upsilon} = \arg\min_{\upsilon} \bar{g}'\Upsilon^{-1}\bar{g},$$

where $\bar{g} = \frac{1}{n}\Sigma\left(g\left(z_i;\Pi(\upsilon)\right)\right)$, then the FOC is

$$2\bar{g}'\Upsilon^{-1}\frac{\partial \bar{g}}{\partial \tilde{\upsilon}'} = 0,$$

with $\Upsilon = \Sigma_R \otimes \sum_{i=1}^{n}\left(x_i x_i'\right)$ being optimal under the null.

Given that the moment conditions are linear, one can rewrite $\bar{g}$ as $\bar{g} = \bar{m}_n - \bar{M}_n\upsilon$, with $\bar{M}_n = \frac{1}{n}\sum_{i=1}^{n}\left(I_{H-1} \otimes x_i\right)$ and $\bar{m}_n = \frac{1}{n}\sum_{i=1}^{n}\left(y_i \otimes x_i\right)$, which implies that $\tilde{\upsilon} = \left(\bar{M}_n\Upsilon^{-1}\bar{M}_n\right)'\left(\bar{M}_n\Upsilon^{-1}\bar{m}_n\right)$. Specifically,

$$\tilde{\upsilon} = \left[\left(\frac{1}{n}\sum_{i=1}^{n}\left(I_{H-1} \otimes x_i\right)\right)'\left(\Sigma_R \otimes \frac{1}{n}\sum_{i=1}^{n}\left(x_i x_i'\right)\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\left(I_{H-1} \otimes x_i\right)\right)\right]^{-1}$$
$$\left[\left(\frac{1}{n}\sum_{i=1}^{n}\left(I_{H-1} \otimes x_i\right)\right)'\left(\Sigma_R \otimes \frac{1}{n}\sum_{i=1}^{n}\left(x_i x_i'\right)\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\left(y_i \otimes x_i\right)\right)\right],$$

which can be simplified to

$$\begin{aligned}\tilde{\upsilon} &= \left[\left(I_{H-1}\otimes X^{'}l_n\right)^{\prime}\left(\Sigma_R\otimes X'X\right)^{-1}\left(I_{H-1}\otimes X^{'}l_n\right)\right]^{-1}\\ &\times\left[\left(I_{H-1}\otimes X^{'}l_n\right)^{\prime}\left(\Sigma_R\otimes X'X\right)^{-1}\left(\sum_{i=1}^{n}\left(y_i\otimes x_i\right)\right)\right]\\ &= \left[\left(I_{H-1}\otimes l_n'X\right)\left(\Sigma_R^{-1}\otimes\left(X'X\right)^{-1}\right)\left(I_{H-1}\otimes X^{'}l_n\right)\right]^{-1}\\ &\times\left[\left(I_{H-1}\otimes l_n'X\right)\left(\Sigma_R^{-1}\otimes\left(X'X\right)^{-1}\right)\left(\sum_{i=1}^{n}\left(y_i\otimes x_i\right)\right)\right],\end{aligned}$$

and then to

$$\begin{aligned}\tilde{\upsilon} &= \left[\left(\Sigma_R^{-1}\otimes l_n'X\left(X'X\right)^{-1}X^{'}l_n\right)^{-1}\left(\Sigma_R^{-1}\otimes l_n'X\left(X'X\right)^{-1}\right)\left(\sum_{i=1}^{n}\left(y_i\otimes x_i\right)\right)\right]\\ &= \left[\left(I_{H-1}\otimes\left(l_n'X\left(X'X\right)^{-1}X^{'}l_n\right)^{-1}l_n'X\left(X'X\right)^{-1}\right)\left(\sum_{i=1}^{n}\left(y_i\otimes x_i\right)\right)\right]\\ &= \left[\left(I_{H-1}\otimes\frac{1}{n}l_k'\right)\left(\sum_{i=1}^{n}\left(y_i\otimes x_i\right)\right)\right]=\sum_{i=1}^{n}\left(y_i\otimes\frac{1}{n}l_k'x_i\right)=\sum_{i=1}^{n}\frac{y_i}{n}=\frac{n_{h\diamond}}{n},\end{aligned}$$

which is exactly the same as the restricted estimator in the multivariate regression.

The J-test for overidentifying restrictions is:

$$J = n\times\bar{g}(\tilde{\upsilon})'\Upsilon^{-1}\bar{g}(\tilde{\upsilon}),$$

Note that $\Upsilon^{-1}=\Sigma_R^{-1}\otimes(X'X)^{-1}$ is exactly the same as the information matrix of the multivariate regression. Additionally,

$$\begin{aligned}g_i &= \left(y_i-\Pi x_i\right)\otimes x_i=\left(y_i\otimes x_i\right)l_k-\left(\Pi x_i\otimes x_i\right)l_k\\ &= vec(x_il_ky_i')-vec(x_il_kx_i'\Pi')\end{aligned}$$

and since $\Pi_R(\upsilon)=\upsilon l_k^{'}=\left(l_k\upsilon' I_{H-1}\right)^{\prime}$, then

$$g_i = vec(x_iy_i')-\left(I_{H-1}\otimes x_ix_i'\right)\upsilon$$

with $\delta=\left(I_{H-1}\otimes l_k\right)vec(\upsilon)=\left(I_{H-1}\otimes l_k\right)\upsilon$, so

$$g_i = vec(x_iy_i')-\left(I_{H-1}\otimes l_k\right)\upsilon.$$

This implies that

$$\frac{1}{n}\sum_{i=1}^{n}g_i=\bar{g}(z;\tilde{\upsilon})=vec(X'Y-X'Xl_k(l_n'l_n)^{-1}l_nY)$$

which is exactly the same as $vec(F)$ from the multivariate regression (A2). Therefore, the J-test for overidentifying restrictions is numerically equivalent to the LM of the multivariate regression.

Finally, it is worth mentioning that the model under the alternative is exactly identified, so the Distance Difference test (see Newey West (1987) for more details) is exactly the same as the J-test. Hence, following the results in chapter 22 of Ruud (2000), the minimum chi-square test that compares $\hat{\Pi}$ with $\Pi(\hat{\upsilon})$ and the GMM version of the LM test will also be numerically identical to the J-test.

A.2 *Proof of Proposition 2*

A.2.1 *Multivariate model*

The weighting matrix of the heteroskedasticity-robust version of the Wald test in the multivariate regression model is defined as:

$$Q = (I_{H-1} \otimes X'X)^{-1} \hat{\Psi} (I_{H-1} \otimes X'X)^{-1},$$

where $\hat{\Psi} = \sum_i \left[(u_i \otimes x_i) \left(u_i' \otimes x_i' \right) \right]$. Note that both $\hat{\Psi}$ and $(I_{H-1} \otimes X'X)^{-1} = I_{H-1} \otimes (X'X)^{-1}$ are $(H-1)K \times (H-1)K$ matrices, with $(X'X)^{-1} = \begin{pmatrix} \frac{1}{n_{*1}} & 0 & ... \\ 0 & ... & 0 \\ ... & 0 & \frac{1}{n_{*K}} \end{pmatrix}$.

Specifically,

$$I_{H-1} \otimes (X'X)^{-1} = \begin{pmatrix} \frac{1}{n_{*1}} & 0 & ... & ... & ... & ... & ... \\ 0 & ... & 0 & ... & ... & ... & ... \\ ... & 0 & \frac{1}{n_{*K}} & 0 & ... & ... & ... \\ ... & ... & 0 & ... & ... & ... & ... \\ ... & ... & ... & 0 & \frac{1}{n_{*1}} & 0 & ... \\ ... & ... & ... & ... & 0 & ... & 0 \\ ... & ... & ... & ... & ... & 0 & \frac{1}{n_{*K}} \end{pmatrix}.$$

Similarly,

$$\hat{\Psi} = \sum_i \left[(u_i \otimes x_i) \left(u_i' \otimes x_i' \right) \right] = \sum_i \left[\left(u_i u_i' \right) \otimes \left(x_i x_i' \right) \right],$$

with $u_i u_i' = \begin{pmatrix} u_{1i}^2 & u_{1i}u_{2i} & ... \\ u_{1i}u_{2i} & ... & .... \\ ... & .... & u_{H-1i}^2 \end{pmatrix}$ and $x_i x_i' = \begin{pmatrix} A_{1i}^2 & A_{1i}A_{2i} & ... \\ A_{1i}A_{2i} & ... & .... \\ ... & .... & A_{Ki}^2 \end{pmatrix}$, where A_{ki}, for $k = 1, ..., K$, are mutually exclusive dummy variables.

Hence, when $A_{ki} = 1$,

$$u_{hi}^2 = \left(B_{hi} - \frac{n_{hk}}{n_{*k}} \right)^2 = B_{hi} \left(1 - \frac{2n_{hk}}{n_{*k}} \right) + \left(\frac{n_{hk}}{n_{*k}} \right)^2,$$

$$u_{hi}u_{mi} = \left(B_{hi} - \frac{n_{hk}}{n_{*k}} \right) \left(B_{mi} - \frac{n_{mk}}{n_{*k}} \right) = \frac{n_{hk}n_{mk}}{n_{*k}^2} - B_{hi}\frac{n_{mk}}{n_{*k}} - B_{mi}\frac{n_{hk}}{n_{*k}}$$

and

$$x_i x_i' = \begin{pmatrix} 0 & ... & ... & ... \\ ... & ... & 0 & ... \\ ... & 0 & 1 & 0 \\ ... & ... & 0 & 0 \end{pmatrix}$$

because B_{hi} and B_{mi} are also dummy variables for $h, m = 1, ..., H-1$ and $h \neq m$.

Therefore,

$$\hat{\Psi} = \sum_{A_1=1} \begin{pmatrix} B_{hi}\left(1-\frac{2n_{h1}}{n_{*1}}\right)+\left(\frac{n_{h1}}{n_{*1}}\right)^2 & 0 & ... & \frac{n_{h1}n_{m1}}{n_{*1}^2}-B_{hi}\frac{n_{m1}}{n_{*1}}-B_{mi}\frac{n_{h1}}{n_{*1}} & 0 \\ 0 & 0 & ... & 0 & 0 \\ ... & ... & ... & ... & ... \\ 0 & ... & ... & ... & ... \\ \frac{n_{h1}n_{m1}}{n_{*1}^2}-B_{hi}\frac{n_{m1}}{n_{*1}}-B_{mi}\frac{n_{h1}}{n_{*1}} & 0 & ... & B_{mi}\left(1-\frac{2n_{m1}}{n_{*1}}\right)+\left(\frac{n_{m1}}{n_{*1}}\right)^2 & 0 \\ 0 & 0 & ... & 0 & 0 \\ ... & ... & ... & ... & ... \end{pmatrix} + ...$$

$$+ \sum_{A_K=1} \begin{pmatrix} 0 & ... & ... & ... & ... \\ ... & ... & 0 & ... & 0 \\ ... & 0 & B_{hi}\left(1-\frac{2n_{hK}}{n_{*K}}\right)+\left(\frac{n_{hK}}{n_{*K}}\right)^2 & 0 & \frac{n_{hK}n_{mK}}{n_{*K}^2}-B_{hi}\frac{n_{mK}}{n_{*K}}-B_{mi}\frac{n_{hK}}{n_{*K}} \\ ... & ... & 0 & ... & 0 \\ ... & 0 & ... & ... & ... \\ \frac{n_{hK}n_{mK}}{n_{*K}^2}-B_{hi}\frac{n_{mK}}{n_{*K}}-B_{mi}\frac{n_{hK}}{n_{*K}} & 0 & ... & ... & B_{mi}\left(1-\frac{2n_{mK}}{n_{*K}}\right)+\left(\frac{n_{mK}}{n_{*K}}\right)^2 \end{pmatrix},$$

which can be simplified to

$$\hat{\Psi} = \begin{pmatrix} n_{11}\left(1-\frac{n_{11}}{n_{*1}}\right) & 0 & ... & ... & ... & -\frac{n_{H-2,1}n_{H-1,1}}{n_{*1}} & 0 & ... \\ 0 & ... & 0 & ... & ... & 0 & ... & 0 \\ ... & 0 & n_{1K}\left(1-\frac{n_{1K}}{n_{*K}}\right) & 0 & ... & ... & 0 & -\frac{n_{H-2,K}n_{H-1,K}}{n_{*K}} \\ 0 & ... & ... & ... & ... & ... & ... & ... \\ -\frac{n_{11}n_{H-1,1}}{n_{*1}} & 0 & ... & ... & ... & n_{H-1,1}\left(1-\frac{n_{H-1,1}}{n_{*1}}\right) & 0 & ... \\ 0 & ... & 0 & ... & ... & 0 & ... & 0 \\ ... & 0 & -\frac{n_{1K}n_{H-1,K}}{n_{*K}} & ... & ... & ... & 0 & n_{H-1,K}\left(1-\frac{n_{H-1,K}}{n_{*K}}\right) \end{pmatrix}$$

because $\sum_{A_{ki}=1} B_{hk} = n_{hk}$ and $\sum_{A_{ki}=1} 1 = n_{*k}$ for $k = 1, ..., K$.

Therefore,

$$Q = \begin{pmatrix} \frac{1}{n_{*1}}\frac{n_{11}}{n_{*1}}\left(1-\frac{n_{11}}{n_{*1}}\right) & 0 & ... & ... & ... \\ 0 & ... & 0 & ... & ... \\ ... & 0 & \frac{1}{n_{*K}}\frac{n_{1K}}{n_{*K}}\left(1-\frac{n_{1K}}{n_{*K}}\right) & 0 & ... \\ 0 & ... & ... & ... & ... \\ -\frac{1}{n_{*1}}\frac{n_{11}n_{H-1,1}}{n_{*1}^2} & 0 & ... & ... & ... \\ 0 & ... & 0 & ... & ... \\ ... & 0 & -\frac{1}{n_{*K}}\frac{n_{1K}n_{H-1,K}}{n_{*K}^2} & ... & ... \\ -\frac{1}{n_{*1}}\frac{n_{11}n_{H-1,1}}{n_{*1}^2} & 0 & ... \\ 0 & ... & 0 \\ ... & 0 & -\frac{1}{n_{*K}}\frac{n_{1K}n_{H-1,K}}{n_{*K}^2} \\ ... & ... & ... \\ \frac{1}{n_{*1}}\frac{n_{H-1,1}}{n_{*1}}\left(1-\frac{n_{H-1,1}}{n_{*1}}\right) & 0 & ... \\ 0 & ... & 0 \\ ... & 0 & \frac{1}{n_{*K}}\frac{n_{H-1,K}}{n_{*K}}\left(1-\frac{n_{H-1,K}}{n_{*K}}\right) \end{pmatrix} \quad \text{(A3)}$$

having $(H-1)\times(H-1)$ blocks of size K, each of which will be diagonal.

A.2.2 *Multinomial model*

Recall that the inverse of the estimated information matrix is:

$$\mathcal{I}(\theta) = \begin{pmatrix} \mathcal{I}_1(\theta) & 0 & 0 \\ 0 & ... & 0 \\ 0 & 0 & \mathcal{I}_K(\theta) \end{pmatrix},$$

with

$$\mathcal{I}_k(\theta)^{-1} = \frac{1}{nE(A_{ki})} \begin{pmatrix} P_{1k}(1-P_{1k}) & -P_{1k}P_{2k} & ... & -P_{1k}P_{H-1,k} \\ -P_{2k}P_{1k} & ... & ... & ... \\ ... & ... & ... & ... \\ -P_{H-1,k}P_{1k} & ... & ... & P_{H-1,k}(1-P_{H-1,k}) \end{pmatrix}.$$

In practice, we use the estimate of P_{hk} , for $k = 1, ..., K$ and $h = 1, ..., H-1$ to estimate the inverse information matrix. Given that $\hat{P}_{hk} = \frac{n_{hk}}{n_{*k}}$ and $E(A_{ki}) = \frac{n_{*k}}{n}$ then

$$\mathcal{I}_k(\theta)^{-1} = \frac{1}{nE(A_{ki})} \begin{pmatrix} \frac{n_{1k}}{n_{*k}}(1-\frac{n_{1k}}{n_{*k}}) & -\frac{n_{1k}}{n_{*k}}\frac{n_{2k}}{n_{*k}} & ... & -\frac{n_{1k}}{n_{*k}}\frac{n_{H-1,k}}{n_{*k}} \\ -\frac{n_{1k}}{n_{*k}}\frac{n_{2k}}{n_{*k}} & ... & ... & ... \\ ... & ... & ... & ... \\ -\frac{n_{1k}}{n_{*k}}\frac{n_{H-1,k}}{n_{*k}} & ... & ... & \frac{n_{H-1,k}}{n_{*k}}(1-\frac{n_{H-1,k}}{n_{*k}}) \end{pmatrix},$$

which is exactly the same as the heteroskedasticity-robust variance in the multivariate regression (A3), except for re-ordering. Therefore, given that the point estimators are the same, the Wald tests will also be the same.

A.3 *Relationship between test statistics when* $H = 2$ **and** $H = 3$

Following Stewart (1995), the W, LR and LM tests can be written as functions of the eigenvalues $(\lambda_1, ..., \lambda_{H-1})$ of the matrix GE^{-1}, where $G = \tilde{\Sigma}_R'\tilde{\Sigma}_R - \hat{\Sigma}_U'\hat{\Sigma}_U$ and $E = \hat{\Sigma}_U'\hat{\Sigma}_U$ with $\hat{\Sigma}_U$ and $\tilde{\Sigma}_R$ being the unrestricted and restricted MLE of the residual matrix in the multivariate regression model, respectively.

Specifically, the three tests can be written as:

$$\left.\begin{array}{c} W = n\sum_i \lambda_i \\ LM = n\sum_i \frac{\lambda_i}{1+\lambda_i} \\ \exp(LR) = n\prod_i (1+\lambda_i) \end{array}\right\}.$$

For $H = 3$, there are only two eigenvalues (λ_1 and λ_2), so

$$\left.\begin{array}{c} W = \lambda_1 + \lambda_2 \\ LM = \frac{\lambda_1}{1+\lambda_1} + \frac{\lambda_2}{1+\lambda_2} = \frac{\lambda_1+\lambda_2+2\lambda_1\lambda_2}{\lambda_1+\lambda_2+\lambda_1\lambda_2+1} \\ \exp(LR) = (1+\lambda_1)(1+\lambda_2) = \lambda_1 + \lambda_2 + \lambda_1\lambda_2 + 1 \end{array}\right\}.$$

Therefore, the set of values of W, LM and LR compatible with the previous expressions is a two-dimensional manifold in the three dimensional $(W, LM, \exp(LR))$ space.

This is in contrast to the case of tests on the coefficients of a multiple regression involving a single ($H = 2$) regressand or tests on the coefficients of a multivariate regression that involve a single regressor ($K = 1$), in which case $W = \lambda_1$, $LM = \lambda_1/(1 + \lambda_1)$ and $\exp(LR) = (1 + \lambda_1)$, so that all three tests lie on a line (a unidimensional manifold) in the three dimensional $(W, LM, \exp(LR))$ space.

A.4 *F approximations*

The F approximations of the Pillai trace (V), Wilks' lambda (Λ) and Lawley-Hotelling (LH) tests that Stata uses are:

$$V_F = \frac{(2n + s + 1)V}{(2m + s + 1)(s - V)}$$

$$\Lambda_F = \frac{(1 - \Lambda^{\frac{1}{t}})df_2}{\left(\Lambda^{\frac{1}{t}}\right) df_1}$$

$$LH_F = \frac{2(sn + 1)LH}{s^2(2m + s + 1)}$$

where p is the number of columns of y variables, v_h is the hypothesis degrees of freedom, v_e is the error degrees of freedom, $s = \min(p, v_h), m = (|v_h - p| - 1)/2$, $n = (v_e - p - 1)/2$, $df_1 = pv_h$, $df_2 = wt + 1 - pv_h/2$, $w = v_e + v_h - (p + v_h + 1)/2$ and $t = \left(\frac{p^2 v_h^2 - 4}{p^2 + v_h^2 - 5}\right)^{1/2}$ (see Stata (2012), Manova, entry for more details).

A.5 *Bootstrap*

Although Monte Carlo simulations helps us in assessing how good the asymptotic approximation of a test statistic is, in practice, they are not useful for inferences in a given sample because we do not know the true values of the parameters.

For that reason, bootstrap provide an alternative to asymptotic approximations for obtaining p-values by resampling methods. In this section, I will explain how to compute the bootstrap for the models in section 3.

For a given sample, I calculate n_{SL} and n_S to estimate the marginal probabilities $\hat{\pi}_{*L}$ and $\hat{\pi}_{S\diamond}$. Then I use those estimated values to independently draw the actions of the kicker, as well as whether or not the goal is scored.

However, a Monte Carlo exercise in which the p-values of the tests are computed using many random bootstraps simulations is computationally very costly. Fortunately, I can also use the exact procedure explained in section 4.1 to obtain all possible contingency tables corresponding to $\hat{\pi}_{*L}$ and $\hat{\pi}_{S\diamond}$, and from there, the exact distribution of bootstrap test statistics. For example, for 5 penalty kicks and two actions per player, there are 286 different contingency tables, while for $n = 20$ there are 1771 contingency tables.

However, I have found that many of those distributions are repeated for different values of $\hat{\pi}_{*L}$ and $\hat{\pi}_{S\diamond}$. More precisely, for an even number of observations there are

$$\frac{\frac{n+2}{2} \times \frac{n}{2}}{2} + 1$$

different distributions, while for an odd number of observations there are

$$\frac{\frac{n-1}{2} \times \frac{n+1}{2}}{2} + 1.$$

This is due to the symmetry relationships that arise because the variables of the model are mutually exclusive dummy variables. For example, $n_L = 2$ and $n_S = 3$ will give the same distribution for the test statistics as $n_L = 3$ and $n_S = 2$.

Therefore, in the 2×2 case, the only difference between the exact test and the bootstrap test distribution is that the former uses the unknown probabilities while the latter uses the estimated probabilities.

However, in the case of three actions, the number of possible contingency tables is very large, and finding their exact bootstrap distribution is very tedious. For that reason, I compute the bootstrap p-value using Monte Carlo simulations rather than the exact test, using once again the estimated values of the marginal probabilities.

A.6 *Kickers and goalkeepers*

Players are divided between kickers and goalkeepers. In brackets is the identification number used in Table 2, and in parentheses it appears the teams they play for.

A.6.1 *Kickers*

[1] Messi *(Barcelona), [2] Cristiano Ronaldo (Real Madrid/Manchester United), [3] Falcao (Atlético de Madrid/Monaco), [4] Gerrard (Liverpool), [5] Guiseppe Rossi (Villareal/Fiorentina), [6] Hulk* (Oporto/Zenit), [7] Ibrahimovic (Inter Milan/Milan/PSG), [8] Kanoute (Sevilla), [9] Negredo* (Almería/Sevilla), [10] Soldado (Getafe/Tottenham), [11] Villa (Valencia/Atlético de Madrid), [12] Xabi Prieto (Real Sociedad).

*kickers are left-footed. All others are right footed.

A.6.2 *Goalkeepers*

[1] Aouate (Deportivo La Coruña/Mallorca), [2] Diego Alves (Almería/Valencia), [3] Diego López (Villareal/Real Madrid), [4] Iraizoz (Athletic Club Bilbao), [5] Moya (Mallorca/Getafe/), [6] Palop (Sevilla), [7] Ricardo (Osasuna), [8] Roberto (Granada), [9] Ruben (Rayo Vallecano), [10]Tono (Racing Santander/Granada/Rayo Vallecano).

Tables

Table 5[7]

Test for Equality of Scoring Probabilities with 2 Actions

		Frequency		Scoring Rates								
Player	#Obs.	L	R	L	R	F-test	A. Pva	LR	A. Pva	LM	A. Pva	B. Pva
Kicker 1	29	0.69	0.31	0.9	0.78	0.746	0.395	0.790	0.373	0.779	0.377	0.442
Kicker 2	44	0.36	0.64	1	0.96	0.565	0.456	0.588	0.442	0.584	0.444	0.399
Kicker 3	16	0.50	0.50	0.88	0.88	0	1	0	1	0	1	1
Kicker 4	32	0.59	0.41	0.95	0.85	0.898	0.350	0.944	0.331	0.930	0.334	0.387
Kicker 5	21	0.52	0.48	0.91	0.8	0.472	0.50	0.515	0.472	0.509	0.475	0.544
Kicker 6	22	0.36	0.64	0.79	0.75	0.033	0.856	0.037	0.847	0.036	0.847	0.849
Kicker 7	41	0.34	0.66	0.93	0.96	0.224	0.638	0.235	0.627	0.235	0.627	0.688
Kicker 8	9	0.67	0.33	0.67	1	1.166	0.315	1.387	0.238	1.285	0.256	0.284
Kicker 9	25	0.76	0.24	0.68	1	2.547	0.124	2.626	0.105	2.493	0.114	0.115
Kicker 10	20	0.25	0.75	1	0.80	1.125	0.302	1.212	0.270	1.176	0.278	0.287
Kicker 11	20	0.7	0.3	0.93	1	0.415	0.527	0.456	0.499	0.451	0.501	0.403
Kicker 12	16	0.69	0.31	1	0.80	2.406	0.143	2.537	0.111	2.346	0.125	0.081
Goalkeeper 1	13	0.77	0.23	0.80	1	0.634	0.442	0.729	0.393	0.709	0.399	0.381
Goalkeeper 2	16	0.50	0.50	0.50	0.63	0.225	0.641	0.256	0.612	0.253	0.614	0.697
Goalkeeper 3	10	0.60	0.40	0.83	0.75	0.084	0.779	0.104	0.746	0.104	0.746	0.759
Goalkeeper 4	15	0.73	0.27	1	1	0	1	0	1	0	1	1
Goalkeeper 5	10	0.10	0.90	1	0.88	0.40	0.544	0.487	0.484	0.476	0.490	0.445
Goalkeeper 6	10	0.70	0.30	0.71	0.67	0.018	0.896	0.022	0.880	0.022	0.880	0.894
Goalkeeper 7	10	0.44	0.56	1	0.80	2.333	0.170	2.589	0.107	2.25	0.133	0.101
Goalkeeper 8	13	0.46	0.54	1	0.86	1.087	0.346	1.226	0.277	1.17	0.291	0.245
Goalkeeper 9	11	0.67	0.33	0.88	1	1.60	0.241	1.823	0.176	1.666	0.196	0.168
Goalkeeper 10	12	0.25	0.75	0.33	1	2.045	0.186	2.252	0.133	2.037	0.153	0.150

Note: *Indicates we reject the null at the 10% significance level, **5% level, ***1% level. Additionally, A and B denote de asymptotic and bootstrap pvalues.

Table 6 [8]

Test for Equality of Scoring Probabilities with 3 Actions

		Frequency			Scoring Rates											
Player	#Obs.	L	C	R	L	C	R	F-test	A. Pva	B. Pva	LR	A. Pva	B. Pva	LM	A. Pva	B. Pva
Kicker 1	36	0.58	0.11	0.31	0.90	1	0.82	0.525	0.596	0.566	1.127	0.568	0.563	1.110	0.573	0.563
Kicker 2***	50	0.34	0.10	0.56	1	0.60	0.96	7.119	0.001	0.012	13.232	0.001	0.012	11.626	0.002	0.012
Kicker 3	21	0.43	0.19	0.38	0.89	1	0.88	0.232	0.794	0.774	0.536	0.764	0.773	0.529	0.767	0.773
Kicker 6	26	0.54	0.02	0.31	0.79	1	0.75	0.540	0.589	0.604	1.194	0.550	0.601	1.167	0.557	0.601
Kicker 8	20	0.55	0.20	0.25	0.80	1	1	1.434	0.265	0.214	3.118	0.210	0.212	2.887	0.236	0.212
Kicker 9	28	0.68	0.11	0.21	0.68	1	1	1.854	0.177	0.152	3.873	0.144	0.143	3.617	0.163	0.143
Kicker 12	20	0.55	0.20	0.25	1	1	0.80	1.593	0.232	0.135	3.437	0.179	0.134	3.157	0.206	0.134
Goalkeeper 2	18	0.50	0.11	0.39	0.56	0.50	0.57	0.013	0.986	0.972	0.032	0.984	0.972	0.032	0.984	0.972
Goalkeeper 3	12	0.50	0.08	0.42	0.67	1	0.57	1.125	0.366	0.278	2.677	0.262	0.225	2.40	0.301	0.225
Goalkeeper 4**	18	0.67	0.11	0.22	1	0.50	1	6.666	0.008	0.036	11.447	0.003	0.035	8.470	0.014	0.035
Goalkeeper 5	15	0.13	0.27	0.60	1	0.85	0.89	0.345	0.714	0.633	0.840	0.656	0.602	0.817	0.664	0.602

Note: *Indicates we reject the null at the 10% significance level, **5% level, ***1% level. Additionally, A and B denote de asymptotic and bootstrap pvalues.

Table 7[9]

Test for Serial Independence with 2 Actions

		Transition Matrix										
Player	#Obs.	L|L	R|L	L|R	R|R	F-test	A. Pva	LR	A. Pva	LM	A. Pva	B. Pva
Kicker 1	28	0.73	0.27	0.62	0.38	1.614	0.215	1.687	0.193	1.637	0.20	0.218
Kicker 2	43	0.40	0.60	0.32	0.68	0.073	0.787	0.076	0.781	0.076	0.781	0.793
Kicker 3	15	0.50	0.50	0.57	0.43	0.066	0.80	0.076	0.781	0.076	0.782	0.861
Kicker 4	31	0.50	0.50	0.69	0.31	1.113	0.299	1.168	0.279	1.146	0.284	0.298
Kicker 5	20	0.50	0.50	0.60	0.40	0.183	0.673	0.203	0.652	0.202	0.653	0.717
Kicker 6	21	0.54	0.46	0.88	0.12	2.595	0.123	2.689	0.101	2.524	0.112	0.122
Kicker 7	40	0.21	0.79	0.42	0.58	1.732	0.196	1.782	0.181	1.743	0.186	0.201
Kicker 8	8	0.67	0.33	0.50	0.50	0.136	0.724	0.179	0.671	0.177	0.673	0.704
Kicker 9*	24	0.67	0.33	1	0	2.75	0.111	2.826	0.092	2.666	0.102	0.092
Kicker 10	19	0	1	0.29	0.71	1.789	0.198	1.901	0.167	1.809	0.178	0.166
Kicker 11	19	0.77	0.23	0.67	0.33	0.201	0.659	0.224	0.635	0.222	0.636	0.685
Kicker 12	15	0.70	0.30	0.80	0.20	0.149	0.705	0.171	0.678	0.170	0.679	0.721
Goalkeeper 1**	12	0.71	0.29	0.60	0.40	4.464	0.060	4.429	0.035	3.703	0.054	0.028
Goalkeeper 2*	15	0.50	0.50	0.57	0.43	3.572	0.081	3.641	0.056	3.233	0.072	0.097
Goalkeeper 3	9	0.25	0.75	0.80	0.20	0.070	0.797	0.090	0.763	0.090	0.764	0.897
Goalkeeper 4	13	0.78	0.22	0.50	0.50	1.186	0.297	1.320	0.250	1.260	0.261	0.231
Goalkeeper 5	9	0	1	0.13	0.88	0.259	0.626	0.327	0.567	0.321	0.570	0.604
Goalkeeper 6	9	0.33	0.67	0.33	0.67	1.166	0.315	1.387	0.238	1.285	0.256	0.214
Goalkeeper 7	9	0.33	0.67	0.40	0.60	0.026	0.875	0.035	0.850	0.035	0.850	0.902
Goalkeeper 8	9	0.25	0.75	0.60	0.40	0.070	0.797	0.090	0.763	0.090	0.764	0.908
Goalkeeper 9	11	0.50	0.50	1	0	0.070	0.797	0.090	0.763	0.090	0.764	0.893
Goalkeeper 10	10	0	1	0.25	0.75	1.60	0.241	1.823	0.176	1.666	0.196	0.285

Note: *Indicates we reject the null at the 10% significance level, **5% level, ***1% level. Additionally, A and B denote de asymptotic and bootstrap pvalues.

Table 8 [10]

Test for Serial Independence with 3 Actions

		Transition Matrix																
Player	#Obs.	L|L	C|L	R|L	L|C	C|C	R|C	L|R	C|R	R|R	Wilks' F	A. Pva	B. Pva	LH F	A. Pva	B. Pva	Pillai F	A. Pva	B. Pva
Kicker 1	35	0.60	0.15	0.25	0.75	0	0.25	0.45	0.10	0.45	1.596	0.186	0.140	1.623	0.180	0.135	1.564	0.194	0.145
Kicker 2	49	0.44	0	0.56	0.60	0	0.40	0.21	0.18	0.61	1.518	0.203	0.162	1.524	0.201	0.161	1.511	0.205	0.163
Kicker 3	20	0.44	0.12	0.44	0.67	0	0.33	0.38	0.38	0.24	0.679	0.611	0.562	0.654	0.628	0.561	0.702	0.595	0.557
Kicker 6	25	0.54	0.08	0.38	0.50	0	0.50	0.63	0.37	0	1.890	0.129	0.087	1.962	0.118	0.081	1.805	0.144	0.097
Kicker 8	19	0.45	0.18	0.37	0.75	0.25	0	0.75	0.25	0	0.860	0.498	0.394	0.849	0.506	0.386	0.865	0.495	0.408
Kicker 9	27	0.58	0.11	0.31	1	0	0	0.83	0.17	0	0.871	0.488	0.331	0.853	0.499	0.334	0.888	0.477	0.331
Kicker 12	19	0.46	0.27	0.27	1	0	0	0.60	0.20	0.20	0.644	0.634	0.545	0.627	0.646	0.537	0.658	0.625	0.556
Goalkeeper 2	17	0.50	0	0.50	0.50	0	0.50	0.42	0.29	0.29	0.568	0.687	0.480	0.546	0.703	0.479	0.586	0.674	0.498
Goalkeeper 3	11	0.40	-	0.60	0	-	1	0.80	-	0.20	0.697	0.606	0.261	0.605	0.666	0.265	0.787	0.550	0.239
Goalkeeper 4	17	0.50	0.50	0	0.73	0.10	0.17	0.50	0	0.50	0.731	0.578	0.316	0.682	0.610	0.327	0.779	0.547	0.306
Goalkeeper 5**	14	0.50	0	0.50	0	0.50	0.50	0.12	0.12	0.76	2.873	0.049	0.022	3.283	0.034	0.020	2.384	0.082	0.029

Note: *Indicates we reject the null at the 10% significance level, **5% level, ***1% level. Additionally, A and B denote de asymptotic and bootstrap pvalues.

Figure 1: Tests for Independence with 2 Actions: Palacios-Huerta (2017) Design, n=5

Figure 1.a: LPM LM

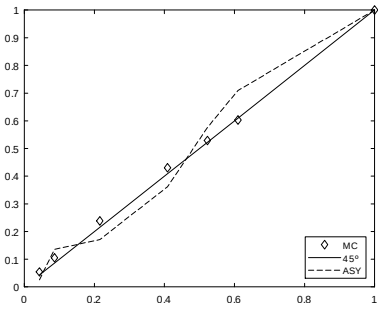


Figure 1.b: LPM LR

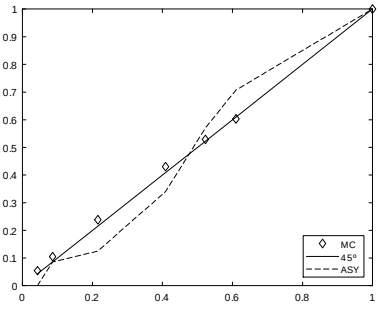


Figure 1.c: LPM F-test

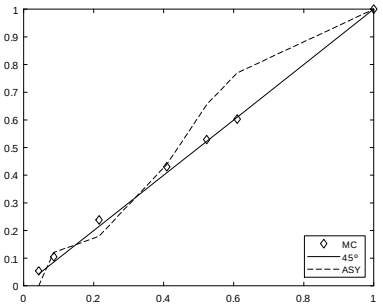


Figure 1.d: LPM Wald

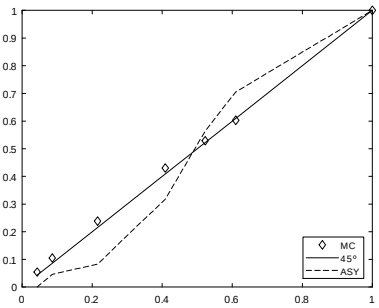


Figure 1.e: LPM Het Rob. Wald

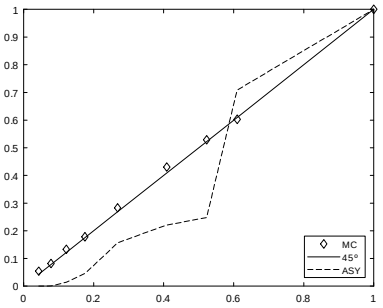


Figure 1.f: Logit LR

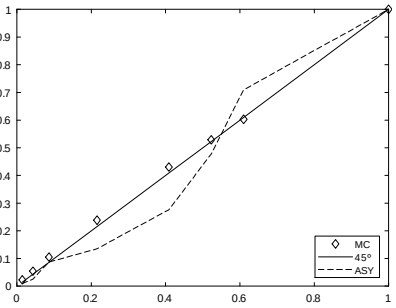
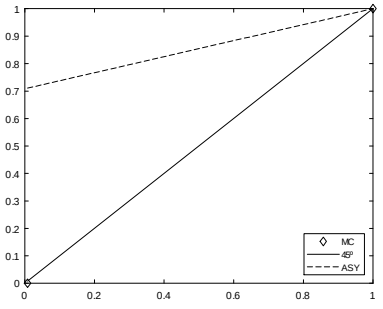


Figure 1.g: Logit Wald



Notes: The graphs show cross plots of exact p-values in the horizontal axis and Monte Carlo and asymptotic p-values in the vertical axis. The diamonds refer to the Monte Carlo p-values and the dashes to the asymptotic ones.

Figure 2: Tests for Independence with 2 Actions: Palacios-Huerta (2017) Design, n=20

Figure 2.a: LPM LM

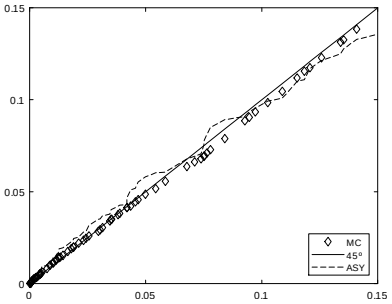


Figure 2.b: LPM LR

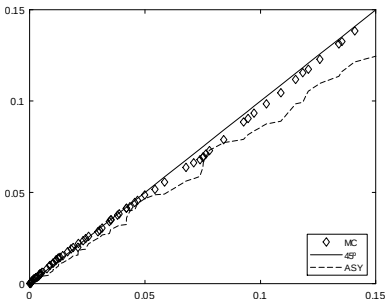


Figure 2.c: LPM F-test

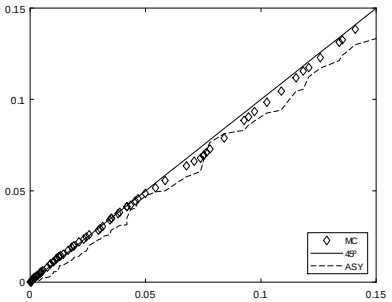


Figure 2.d: LPM Wald

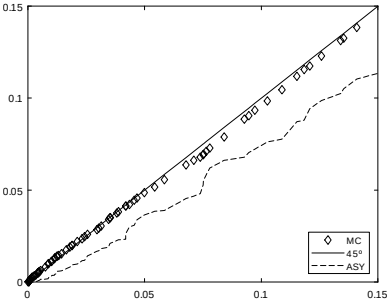


Figure 2.e: LPM Het. Rob. Wald

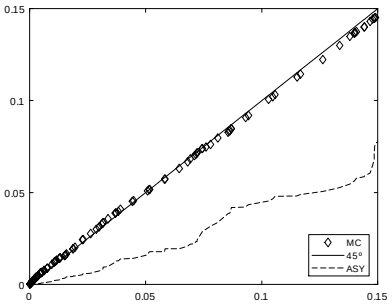


Figure 2.f: Logit LR

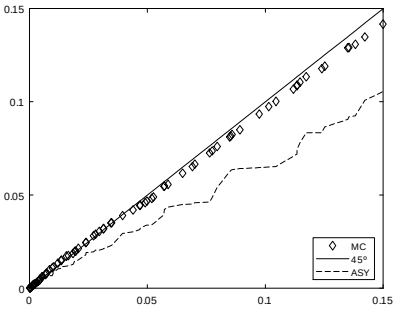
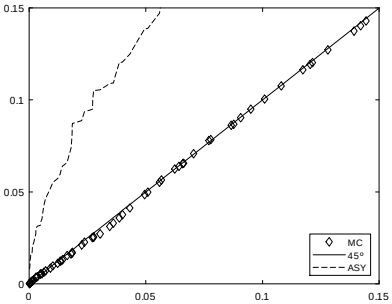


Figure 2.g: Logit Wald



Notes: The graphs show cross plots of exact p-values in the horizontal axis and Monte Carlo and asymptotic p-values in the vertical axis. The diamonds refer to the Monte Carlo p-values and the dashes to the asymptotic ones.

Figure 3: P-value Plots for Test of Equal Scoring Probabilities with 3 Actions, n=20

Figure 3.a: LPM LM

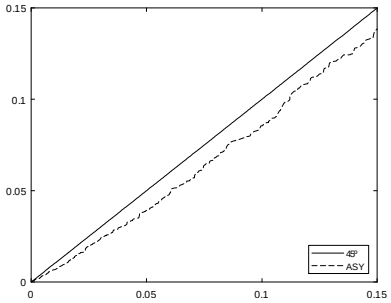


Figure 3.b: LPM LR

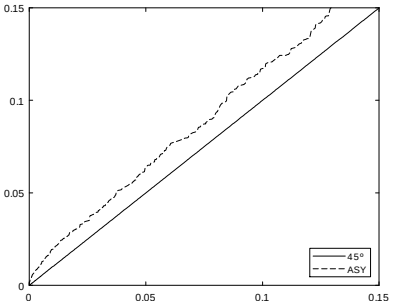


Figure 3.c: LPM F-test

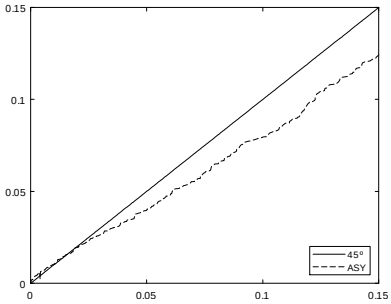


Figure 3.d: LPM Wald

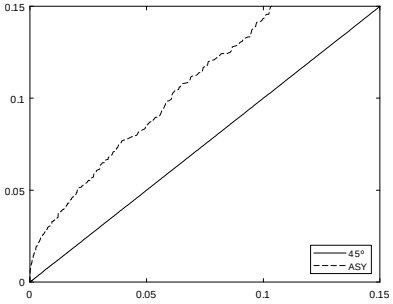


Figure 3.e: LPM Het. Rob. Wald

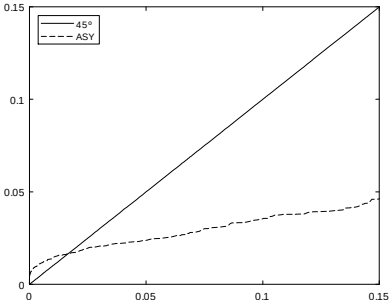


Figure 3.f: Logit LR

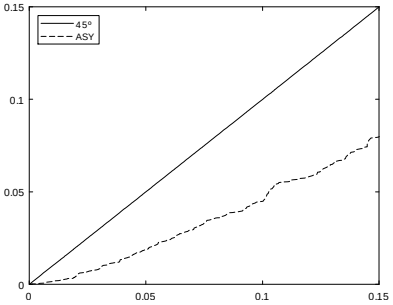
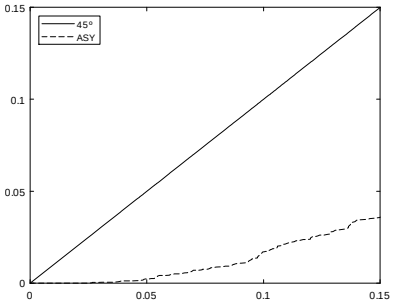


Figure 3.g: Logit Wald



Notes: The graphs display the empirical distribution functions of the asymptotic p-values in the Monte Carlo simulations (see Davidson and MacKinnon (1998)).

Figure 4: P-value Plots for Serial Correlation Tests with 3 Actions, n=20

Figure 4.a: LPM LM

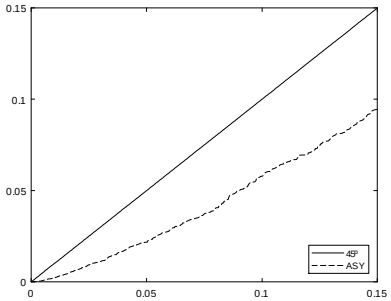


Figure 4.b: LPM Pillai

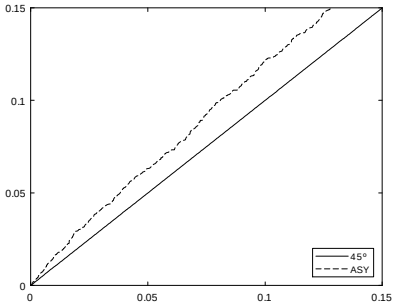


Figure 4.c: LPM LR

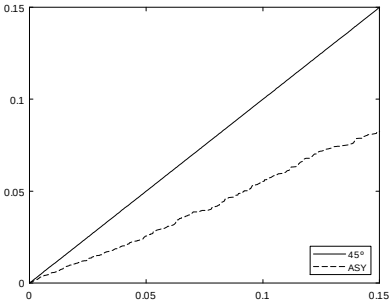


Figure 4.d: LPM Wilks

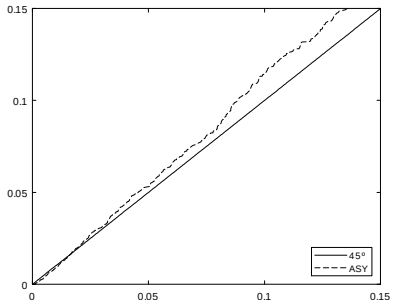


Figure 4.e: LPM Wald

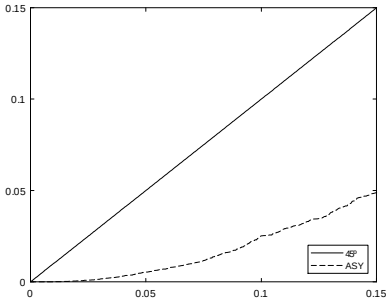


Figure 4.f: LPM LH

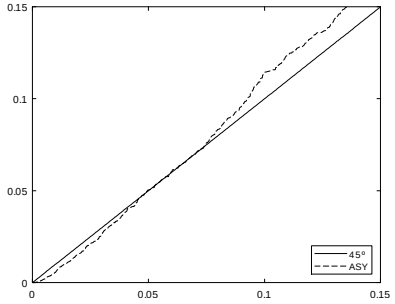


Figure 4.g: Logit LR

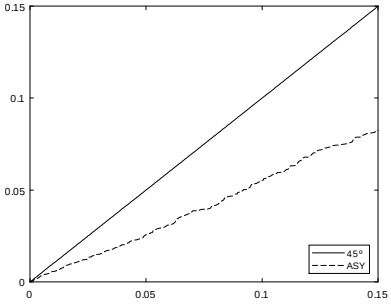
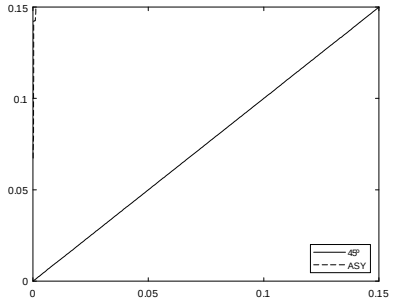


Figure 4.g: Logit Wald



Notes: The graphs display the empirical distribution functions of the asymptotic p-values in the Monte Carlo simulations (see Davidson and MacKinnon (1998)).