Regression Analysis of Multivariate Fractional Data*

José M.R. Murteira

Faculdade de Economia, Universidade de Coimbra, and CEMAPRE

Joaquim J.S. Ramalho

Departamento de Economia and CEFAGE-UE, Universidade de Évora

This version: March, 2012

Abstract

The present article discusses alternative regression models and estimation methods for dealing with multivariate fractional response variables. Both conditional mean models, estimable by nonlinear least squares and quasi-maximum likelihood, and fully parametric models (Dirichlet and Dirichlet-multinomial), estimable by maximum likelihood, are considered. In contrast to previous papers but similarly to the univariate case, a new parameterization is proposed here for the parametric models, which allows the same specification of the conditional mean of interest to be used in all models, irrespective of the specific functional form adopted for it. The text also discusses at some length the specification analysis of fractional regression models, proposing several tests that can be performed through artificial regressions. Finally, an extensive Monte Carlo study evaluates the finite sample properties of most of the estimators and tests considered.

JEL classification code: C35.

Key Words: Multivariate fractional data; Quasi-maximum likelihood estimator; Dirichlet regression; Regression-based specification tests; RESET test.

*Financial support from Fundacao para a Ciencia e a Tecnologia (grant PTDC/EGE-ECO/119148/2010) is gratefully acknowledged. Address for correspondence: José Murteira, Faculdade de Economia, Universidade de Coimbra, Av. Dias da Silva, 165, 3004-512 Coimbra, Portugal. E-mail: jmurt@fe.uc.pt.

1 Introduction

In several economic settings, the dependent variable of interest is often a proportion or, more generally, a vector of proportions, $\boldsymbol{y} \equiv (y_1, y_2, \ldots, y_M)'$, corresponding to a set of shares for a given number (M) of exhaustive, mutually exclusive categories. Examples include pension plan participation rates, fraction of land allocated to agriculture, percentage of weekly time devoted to each of a given set of human activities, market shares of firms, fractions of income spent on various classes of goods, asset portfolio shares, and proportions of different types of debt within the financing mix of firms. While in the first two cases there are only two categories ($M = 2$, usually a characteristic and its opposite, or absence) and a single proportion is modelled, the remaining examples illustrate the more general situation (encompassing the former), where $M > 2$ and the joint behaviour of a multivariate fractional variable is of interest.

The regression analysis of fractional data, inherently bounded within the unit simplex, raises a number of interesting research issues that challenge conventional approaches of estimation and inference. For the univariate case, the main issues are discussed in the seminal paper by Papke and Wooldridge (1996), who propose the robust quasi-maximum likelihood method (QML) of Gouriéroux, Monfort and Trognon (1984) for the estimation of the so-called fractional regression models, on the basis of a Bernoulli quasi-likelihood and a logit conditional mean function. Although less common, maximum likelihood (ML) estimation on the basis of the beta distribution has also been proposed in the literature (e.g. Paolino, 2001; Ferrari and Cribari-Neto, 2004). In a recent paper, Ramalho, Ramalho and Murteira (2011) survey the main alternative regression models and estimation methods that are available for dealing with univariate fractional response variables and propose a unified testing methodology to assess the validity of the assumptions required by each model and estimator.

In a multivariate setting, as in the univariate case, researchers' main interest frequently lies in the estimation of the conditional means of the elements of $\boldsymbol{y}$, given a set of explanatory variables, $E(\boldsymbol{y}|\boldsymbol{X})$. One seminal methodological contribution to this goal is provided by Woodland (1979), who presents maximum likelihood estimation of systems of share equations on the basis of the Dirichlet distribution, a well known multivariate generalization of the beta distribution. Like the latter, the Dirichlet is not applicable when the response variables assume either value in $\{0, 1\}$ with nontrivial probability, a constraint that can be violated in several situations.[1] More recently, QML estimation based on the multivariate Bernoulli (MB) probability function (p.f.) has also

[1] For instance, in demand analysis the phenomenon of 'zero expenditures' becomes increasingly important when individual data are analyzed and shorter time periods are observed (*e.g.*, the tobacco share of a family budget may be zero in a certain period).

become relatively popular; see *inter alia* Sivakumar and Bhat (2002), Ye and Pendyala (2005), and Mullahy and Robert (2010), who model, respectively, commodity flows, transportation time and household time allocation. When interest is confined to the conditional mean parameters, QML can prove a satisfactory approach, often dealing well with boundary observations. However, unlike in the univariate case, the specifications used for $E(\boldsymbol{y}|\boldsymbol{X})$ in conditional mean models, estimable by QML (or nonlinear least squares - NLS), and fully parametric models, estimable by ML, usually differ and are often not compatible. Moreover, the specification analysis of multivariate fractional models has not merited much attention in the literature.

The present paper considers both conditional mean models and fully parametric models for multivariate fractional responses. For fully parametric models, a new parameterization is proposed that enables the use of any valid specification of $E(\boldsymbol{y}|\boldsymbol{X})$ and facilitates a ready evaluation of the covariates' relationships to $E(\boldsymbol{y}|\boldsymbol{X})$. The multinomial logit model stands out as the most analytically tractable and widely used conditional mean specification, so, although not confined to it, the text devotes special attention to this model and some of its extensions. Meanwhile, alternatively to Dirichlet regression, the paper also discusses multinomial-based specifications, potentially advantageous when the data are obtained as ratios of observable integers, possibly exhibiting boundary values with nontrivial probability.[2]

The specification analysis of multivariate fractional models is also discussed at some length in the present text. In particular, the paper proposes several tests of the first moment assumptions, which can be obtained by making use of the robust testing procedure introduced by Wooldridge (1991), adequately performed upon QML or NLS estimation. In addition, some specification tests for other assumptions implied by fully parametric models are also briefly discussed. All the proposed tests can be implemented through artificial least squares regressions.

The remainder of the paper is organized as follows. Section 2 describes the notation and critically reviews previous modelling approaches for share regressions. Section 3 discusses alternative regression models and estimation methods that are available for use with multivariate fractional response variables. Section 4 proposes specification tests for the various models and methods considered in the paper. Section 5 is dedicated to a Monte Carlo study, illustrating the behaviour of several estimators and tests. Finally, section 6 concludes the paper and suggests future research.

[2] One word about terminology seems advisable here: in microeconometrics the adjective "multinomial" usually refers to models based on a p.f. that is termed "multivariate Bernoulli" in the statistics literature. In the latter context, as is well known, the term "multinomial" refers to a different p.f. (encompassing the MB). In this paper, use of both p.f.'s is discussed, so, to avoid ambiguity, the statistical terminology is preferred.

2 Framework

Let $\boldsymbol{y} \equiv (y_1, \ldots, y_M)'$ denote the M-vector of fractional dependent variables, or shares, confined, by definition, to the unit $(M-1)$-simplex,[3]

$$\mathcal{S}^{M-1} \equiv \left\{ \boldsymbol{y} \in \mathcal{R}^M : \sum\nolimits_{m=1}^{M} y_m = 1, y_m \geq 0, \forall m \right\}.$$

For quite some time, the most popular econometric specifications of systems of share equations did not take into account the intrinsic characteristics of fractional responses. Typically, each share y_m was decomposed into a deterministic function of covariates, $D_m(\boldsymbol{X};\boldsymbol{\beta})$, and a stochastic disturbance term, u_m,

$$y_m = D_m(\boldsymbol{X};\boldsymbol{\beta}) + u_m, \quad m = 1, ..., M, \tag{1}$$

with $\boldsymbol{\beta}$ a parameter vector. Then, usually: (i) a multivariate normal distribution was assumed for u_m; (ii) to deal with the singularity of the share equation system, one equation (the M-th, say) was deleted from the system and the corresponding predicted share was calculated as

$$D_M\left(\boldsymbol{X};\hat{\boldsymbol{\beta}}\right) = 1 - \sum_{m=1}^{M-1} D_m\left(\boldsymbol{X};\hat{\boldsymbol{\beta}}\right);$$

(iii) the restrictions observed on y_m were not fully taken into account in the specification of $D_m(\boldsymbol{X};\boldsymbol{\beta})$. Clearly, this setup fails to guarantee that, similarly to actual shares, predicted shares fall into the unit simplex, due to a nonzero probability of greater than unity or negative predictions.

In view of this problem, various alternative approaches have been suggested. Hermalin and Wallace (1994), Wang, *et al.* (2006), Pu, *et al.* (2008) and Yin, *et al.* (2010) use a probit or logit fractional specification for each of the deterministic components of (1). However, each equation is estimated individually, so predicted shares do not necessarily fall within the unit simplex, irrespective of deleting one equation from the system (the predicted share for equation M may be negative) or not (the predicted shares do not sum up to unity).

Aitchison (1982) and Fry, Fry and McLaren (1996) propose a one-to-one transformation from the unit simplex $\mathcal{S}^{M-1}$ to the real set $\mathcal{R}^{M-1}$, namely the additive log-ratio transformation defined by $r_m = \log(y_m/y_M)$, $m = 1, ..., M-1$. This yields the model

$$r_m = \log\left[D_m(\boldsymbol{X};\boldsymbol{\beta})/D_M(\boldsymbol{X};\boldsymbol{\beta})\right] + v_m, \quad m = 1, ..., M-1, \tag{2}$$

with v_m assumed to follow a multivariate normal distribution.[4] The inverse transformation from $\mathcal{R}^{M-1}$ to $\mathcal{S}^{M-1}$ is the additive logistic transformation, which implies a multinomial logit

[3] This type of data are known in the statistical literature as 'compositional data' (Aitchison, 1982).

[4] This method is widely used in fields like geology, pedology, geochemistry and biology (see the survey by Aitchison and Egozcue, 2005), and political science (see *e.g.* Katz and King, 1999).

specification for the y_m's. While effectively restricting predicted shares to the unit simplex, this method presents some disadvantages such as not being well defined for the boundary value 0, thus requiring *ad hoc* adjustments if that value is observed in the sample (*e.g.* replacing the resultant infinite values of r_m by an arbitrarily chosen large number).

One approach for dealing with boundary values in the fractional context has been the use of multivariate tobit models, namely for data censored at zero (*e.g.* Heien and Wessells, 1990). However, with such models there is again a nonzero probability of some shares, or their summation, being greater than unity. One alternative, adopted by, *e.g.*, Poterba and Samwick (2002) and Klawitter (2008), is to assume that shares follow a multivariate normal distribution truncated at the boundaries of the $(M-1)-$unit simplex. Use of this approach, however, may be discouraged by the fact that, as for tobit-type models, estimation is often fraught with computational complexity, which may lead researchers to adopt questionable assumptions. For instance, Poterba and Samwick (2002) assume non-correlated disturbances across latent variables equations underlying shares of financial assets, in order to avoid a log-likelihood with eight-dimensional normal integrals.

Given the limitations of the foregoing approaches, this paper considers various alternative regression models that fully account for the bounded, unit-sum nature of fractional variables without requiring transformations of the response variables. As described in the next sections, these models differ on a number of respects, such as the adoption, or not, of full distributional assumptions for shares, and the possibility, or not, of dealing with boundary observations. In any case, they all have in common the use of functional forms for $E(\boldsymbol{y}|\boldsymbol{X})$ which enforce the conceptual requirement that, as for $\boldsymbol{y}$, its elements belong to the unit simplex.[5]

In the ensuing text a random sample of $i = 1, ..., N$ observations on $\boldsymbol{y}$ and $\boldsymbol{X}$ is supposed to be available for estimation of the parameters of interest, usually those of the conditional mean function, $E(\boldsymbol{y}|\boldsymbol{X})$. Let $E(\boldsymbol{y}|\boldsymbol{X}) = \boldsymbol{G}(\boldsymbol{X};\boldsymbol{\beta}_0) \equiv [G_1(\boldsymbol{X};\boldsymbol{\beta}_0), ..., G_M(\boldsymbol{X};\boldsymbol{\beta}_0)]'$, the column M-vector of the conditional mean functions of $\boldsymbol{y}$, with $\boldsymbol{\beta}_0$ denoting the true value of $\boldsymbol{\beta}$. To simplify the notation, $E(\boldsymbol{y}|\boldsymbol{X})$ and its components, $E(y_m|\boldsymbol{X})$, $m = 1, ..., M$, will often be referred to without explicit mention to its arguments: $\boldsymbol{G} \equiv \boldsymbol{G}(\boldsymbol{X};\boldsymbol{\beta})$, $G_m \equiv G_m(\boldsymbol{X};\boldsymbol{\beta})$. When intended, the corresponding individual entities may be denoted as $\boldsymbol{G}_i \equiv \boldsymbol{G}(\boldsymbol{X}_i;\boldsymbol{\beta})$ and $G_{im} \equiv G_m(\boldsymbol{X}_i;\boldsymbol{\beta})$. Given the definition of the elements of $\boldsymbol{y}$, their conditional means are also subject to the constraints $G_m \geq 0$, $\forall m$, and $\sum_{m=1}^{M} G_m = 1$. Usually, G_m is specified as a function

[5]For brevity sake, multivariate two-part and similar models are not included in the text. Indeed, the consideration of this subject would be elaborate enough so as to deserve a separate paper on its own. Some keynote references in this regard are Wales and Woodland (1983) and Lee and Pitt (1986), who address the estimation of demand systems with nonnegativity constraints.

of M indices of covariates, that is, $G_m = G_m(\boldsymbol{X}\boldsymbol{\beta})$ and $\boldsymbol{X}\boldsymbol{\beta} = [\boldsymbol{x}_1'\boldsymbol{\beta}, ..., \boldsymbol{x}_M'\boldsymbol{\beta}]'$, where $\boldsymbol{X} = [\boldsymbol{x}_1, ..., \boldsymbol{x}_M]'$, with column $K-$vectors $\boldsymbol{x}_m$ conformable to $\boldsymbol{\beta}$. With an appropriate redefinition of covariates and parameters vectors (as described below for the special case of the multinomial logit), alternative invariant covariates and alternative specific parameters can also be considered.

3 Regression Models and Estimation Methods

Two main approaches for dealing with multivariate fractional data are considered here. The first only requires correct specification of the conditional mean of $\boldsymbol{y}$, given covariates, whereas the second is a fully parametric approach based on the assumption of a particular distribution for $\boldsymbol{y}$, whose mean may be specified as in the first approach. Most situations with a finite number of boundary observations preclude application of the second approach, except when the fractional response variables can be interpreted as ratios of integers and these integers are observable.

3.1 Alternative Specifications for $E(\boldsymbol{y}|\boldsymbol{X})$

The specifications used for modelling binary response variables in the univariate case are also employed to describe the conditional mean of fractional responses; see Papke and Wooldridge (1996). Analogously, the specifications that are commonly used to model the probability of an individual choosing between M mutually exclusive alternatives may also be employed to describe $E(\boldsymbol{y}|\boldsymbol{X})$ in the multivariate context, since they satisfy the bounded, unit-sum nature of the conditional means of fractional variables. Next, three of the most popular of those specifications are briefly reviewed.

In the multivariate context, special attention has been devoted to the multinomial logit specification, which can be expressed as

$$G_m = \frac{\exp(\boldsymbol{x}_m'\boldsymbol{\beta})}{\sum_{l=1}^{M}\exp(\boldsymbol{x}_l'\boldsymbol{\beta})},\ m = 1, ..., M. \tag{3}$$

This formulation is general enough to allow for alternative invariant covariates and alternative specific parameters, if interactions of alternative specific indicators with alternative invariant explanatory variables are included as covariates. For instance, if $G_m = \exp(\boldsymbol{z}_m'\boldsymbol{\alpha} + \alpha_m z) / \left[\sum_{l=1}^{M}\exp(\boldsymbol{z}_l'\boldsymbol{\alpha} + \alpha_l z)\right]$, then, in expression (3), $\boldsymbol{x}_m' \equiv [\boldsymbol{z}_m', d_1 z, ..., d_M z]$ and $\boldsymbol{\beta} \equiv [\boldsymbol{\alpha}', \alpha_1, ..., \alpha_M]'$, where d_l, $l = 1, ..., M$, denotes an indicator variable equal to one if $l = m$.[6] To avoid ambiguity, the special case with only alternative invariant covariates and alternative specific parameters will hereafter be designated "MNL".

[6] As is well known, the unit-sum identity of the conditional means imply normalization of coefficients associated with alternative-invariant covariates.

A limitation of the multinomial logit model, usually known as independence of irrelevant alternatives (IIA), is that discrimination among alternatives reduces to a series of pairwise comparisons which are unaffected by the characteristics of alternatives other than the pair under consideration. One alternative approach that does not have this weakness is the nested logit, which is the most common member of the "generalized extreme-value" class of models, widely used in discrete choice analysis (see, *e.g.*, Train, 2009, Ch. 4). In the present context, this model is suitable when proportions are attributed to alternatives in a way that involves a sequence of allocation decisions over some level hierarchy. Examples of its use with fractional responses are provided by Ye and Pendyala (2005) and Dubin (2007), who model, respectively, time allocation among different activities and market shares. The nested logit can be expressed as follows: suppose that $M > 2$ and the alternatives can be distributed into L nonoverlapping subsets of categories, $S_1, ..., S_L$, $L < M$. Then, considering only two decision levels, the conditional mean of y_m, where alternative m belongs to the subset S_l, can be expressed as

$$G_m = \frac{\exp\left[\boldsymbol{x}_m'\boldsymbol{\beta}/\left(1+\eta_l\right)\right]\left\{\sum_{j\in S_l}\exp\left[\boldsymbol{x}_j'\boldsymbol{\beta}/\left(1+\eta_l\right)\right]\right\}^{\eta_l}}{\sum_{k=1}^{L}\left\{\sum_{j\in S_k}\exp\left[\boldsymbol{x}_j'\boldsymbol{\beta}/\left(1+\eta_k\right)\right]\right\}^{1+\eta_k}}. \tag{4}$$

This formulation nests the multinomial logit for $\eta_l = 0$, $l = 1, ..., L$.

Another possible generalization of the logit model is the "random parameters logit", which takes $\boldsymbol{\beta}$ as random. This model can result from several concerns, *e.g.*, the possibility of individual heterogeneity of regression parameters, the occurrence of measurement errors or the omission of covariates. When this type of concern is allowed for and no repeated observations on individuals are available, econometric analysis must be based on conditional means marginal with respect to parameter variation. Formally,

$$G_m = E_{\boldsymbol{\beta}}\left(G_m^{\beta}\right) = \int G_m^{\beta} dF_{\boldsymbol{\beta}}\left(\boldsymbol{\beta}\right), \; m = 1, ..., M, \tag{5}$$

where G_m^{β} is typically defined as in (3) and $F_{\boldsymbol{\beta}}$ denotes the joint distribution of $\boldsymbol{\beta}$. Among other possibilities, this distribution can be specified as multivariate normal (the most frequent choice) or, *e.g.*, lognormal (if the elements of $\boldsymbol{\beta}$ are known to be positive).

Even though the main focus of this paper is the empirical analysis of fractional regression models, irrespective of the economic theory that may have generated the system of share equations to be estimated, it should be noted that these models also conform with the constrained economic optimization framework that underlies some applications of multivariate fractional regression models. For example, Considine and Mount (1984) demonstrated that a multinomial logit specification can represent a "well-behaved" set of demand functions and Dubin (2007) produced a similar proof for the nested logit model.

3.2 Conditional Mean Models

As in the univariate case, the simplest solution for dealing with multivariate fractional response variables is the use of conditional mean models, i.e. models that only involve the specification of $E(\boldsymbol{y}|\boldsymbol{X})$. Apart from some complex specifications that may be adopted for G_m (*e.g.*, the random parameters logit specification), the parameters of the model for $E(\boldsymbol{y}|\boldsymbol{X})$ may be estimated, in general, by, among other methods, systems NLS or QML. In the former case, as in (1), one can write this model as a system of nonlinear regression equations of the form

$$y_m = G_m + u_m, \quad m = 1, ..., M, \quad \sum_{m=1}^{M} y_m = 1.$$

Under random sampling and standard assumptions (namely correct and twice continuously differentiable specification of $\boldsymbol{G}$), the NLS estimator, $\hat{\boldsymbol{\beta}}_{NLS}$, minimizer of the sum of squared residuals, $\sum_{i=1}^{N} \hat{\boldsymbol{u}}_i'\hat{\boldsymbol{u}}_i$, $\hat{\boldsymbol{u}}_i \equiv \boldsymbol{y}_i - \boldsymbol{G}\left(\boldsymbol{X}_i;\hat{\boldsymbol{\beta}}_{NLS}\right) \equiv \boldsymbol{y}_i - \hat{\boldsymbol{G}}_i$, is consistent and asymptotically normal, that is,

$$\sqrt{N}\left(\hat{\boldsymbol{\beta}}_{NLS} - \boldsymbol{\beta}_0\right) \xrightarrow{d} \mathcal{N}\left(\boldsymbol{0}, \boldsymbol{A}_0^{-1}\boldsymbol{B}_0\boldsymbol{A}_0^{-1}\right),$$

where

$$\boldsymbol{A}_0 \equiv E\left[\nabla_{\boldsymbol{\beta}}\boldsymbol{G}_i(\boldsymbol{\beta}_0)'\nabla_{\boldsymbol{\beta}}\boldsymbol{G}_i(\boldsymbol{\beta}_0)\right],$$

$$\boldsymbol{B}_0 \equiv E\left[\nabla_{\boldsymbol{\beta}}\boldsymbol{G}_i(\boldsymbol{\beta}_0)'\boldsymbol{u}_i\boldsymbol{u}_i'\nabla_{\boldsymbol{\beta}}\boldsymbol{G}_i(\boldsymbol{\beta}_0)\right],$$

$\boldsymbol{u}_i \equiv \boldsymbol{y}_i - \boldsymbol{G}_i(\boldsymbol{\beta}_0)$ and $\mathcal{N}$ denotes the multivariate normal distribution (see, *e.g.*, Wooldridge, 2002, Theorems 12.2 and 12.3). These matrices can be consistently estimated in the usual way upon NLS estimation, by substituting sample averages for population expected values and evaluating $\boldsymbol{\beta}$ at $\hat{\boldsymbol{\beta}}_{NLS}$.

Potentially more efficient estimators of $\boldsymbol{\beta}$ may be obtained by assuming some reasonable model for the conditional second moments and estimating the model by systems weighted NLS. An alternative approach is provided by QML, which is based on the maximization of a linear exponential family (LEF) likelihood. In the present context, a natural choice for this likelihood, generalizing the approach of Papke and Wooldridge (1996) in the univariate case, is provided by the MB p.f. (see Johnson, *et al.*, 1997, Ch. 36). This p.f. is appropriate when there are M alternatives and each individual chooses only one alternative. Let the m-th component of $\boldsymbol{b} \equiv (b_1, ..., b_M)'$ be a binary variable equal to one if alternative m is taken, and zero otherwise. Considering $\pi_m \equiv \Pr(b_m = 1) = E(b_m)$, the MB p.f. can be written as

$$f_{\boldsymbol{b}}(\boldsymbol{b}) = \prod_{m=1}^{M} \pi_m^{b_m}, \quad \sum_{m=1}^{M} \pi_m = 1.$$

In a regression context, the parameters π_m can be replaced by conditional expectations given covariates.

With multivariate fractional variables, substituting $E(y_m|\boldsymbol{X})$ for π_m, the i-th term of the likelihood can be expressed as $L_i^{MB}(\boldsymbol{\beta}) = \prod_{m=1}^{M} G_{im}^{y_{im}}$. This yields the individual contribution to the log-likelihood,

$$\log L_i^{MB}(\boldsymbol{\beta}) = \sum_{m=1}^{M} y_{im} \log G_{im} = \sum_{m=1}^{M-1} y_{im} \log \frac{G_{im}}{G_{iM}} + \log G_{iM}, \tag{6}$$

where $G_{iM} = 1 - \sum_{m=1}^{M-1} G_{im}$ and the last expression evinces the LEF form of the likelihood. The QML estimator $\hat{\boldsymbol{\beta}}_{QML}$ maximizing $LL^{MB}(\boldsymbol{\beta}) \equiv \sum_{i=1}^{N} \log L_i^{MB}(\boldsymbol{\beta})$ is consistent and asymptotically normal regardless of the true conditional distribution of $\boldsymbol{y}$, provided that $\boldsymbol{G}$ is correctly specified (Gouriéroux, *et al.*, 1984). Formally,

$$\sqrt{N}\left(\hat{\boldsymbol{\beta}}_{QML} - \boldsymbol{\beta}_0\right) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \boldsymbol{A}_0^{-1}\boldsymbol{B}_0\boldsymbol{A}_0^{-1}\right),$$

where

$$\begin{aligned} \boldsymbol{A}_0 &\equiv E\left[-\nabla_{\boldsymbol{\beta}\boldsymbol{\beta}'} LL_i^{MB}(\boldsymbol{\beta})\right]_{\boldsymbol{\beta}=\boldsymbol{\beta}_0} \\ \boldsymbol{B}_0 &\equiv E\left[\nabla_{\boldsymbol{\beta}} LL_i^{MB}(\boldsymbol{\beta}) \nabla_{\boldsymbol{\beta}'} LL_i^{MB}(\boldsymbol{\beta})\right]_{\boldsymbol{\beta}=\boldsymbol{\beta}_0}. \end{aligned} \tag{7}$$

Consistent estimators for $\boldsymbol{A}_0$ and $\boldsymbol{B}_0$ are obtained in the usual manner, replacing population expectations by sample averages, with $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}_{QML}$. QML estimation of fractional MNL models has been considered by Sivakumar and Bhat (2002), Ye and Pendyala (2005), Mullahy (2010) and Mullahy and Robert (2010).

As an alternative to NLS or QML, one may resort to ML estimation, which requires full knowledge on the joint conditional density of the response variables.

3.3 The Dirichlet Regression Model

Let $\hat{\boldsymbol{\beta}}_{ML}$ denote the ML estimator of $\boldsymbol{\beta}$. As is well known, under correct specification of the joint conditional density, $f(\boldsymbol{y}|\boldsymbol{X})$,

$$\sqrt{N}\left(\hat{\boldsymbol{\beta}}_{ML} - \boldsymbol{\beta}_0\right) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \boldsymbol{A}_0^{-1}\right),$$

with $\boldsymbol{A}_0$ defined in (7). Several statistical distributions are suited to model data confined to the unit simplex. The most popular choice is the Dirichlet distribution, a multivariate generalization of the beta distribution (see Kotz, *et al.*, 2000, Ch. 49).[7] Its joint density function can be expressed as

$$\begin{aligned} f_{\boldsymbol{y}}^{D}(\boldsymbol{y};\boldsymbol{\gamma}) &= \frac{\Gamma(\gamma_0)}{\prod_{m=1}^{M}\Gamma(\gamma_m)} \prod_{m=1}^{M} y_m^{\gamma_m - 1} \equiv Dirichlet(\boldsymbol{\gamma}), \\ y_m &: \quad y_m > 0, \quad \sum_{m=1}^{M} y_m = 1, \quad m = 1, ..., M, \end{aligned}$$

[7] For alternative distributions for fractional data, see Kotz, *et al.* (2000), Ch. 49 (Sec. 8) and Ch. 50.

where $\boldsymbol{\gamma} \equiv (\gamma_1, ..., \gamma_M)'$ denotes a vector of positive parameters and $\gamma_0 \equiv \sum_{m=1}^{M} \gamma_m$. Under this parameterization,

$$E(y_m) = \frac{\gamma_m}{\gamma_0} \tag{8}$$

and the elements of the covariance matrix of $\boldsymbol{y}$ can be expressed as

$$COV(y_l, y_m) = \frac{\gamma_l(\delta_{lm}\gamma_0 - \gamma_m)}{\gamma_0^2(\gamma_0 + 1)}, \quad l, m = 1, ..., M, \tag{9}$$

where δ_{lm} denotes the Kronecker delta equal to one if $l = m$ and zero otherwise. The Dirichlet distribution is defined only for $y_m \in (0, 1)$ and, therefore, cannot be used when the probability of limit observations is nontrivial.

With an appropriate choice of parameters, the Dirichlet distribution allows for great flexibility. It also constitutes a simple probability structure endowed with some attractive mathematical features. For instance, any subvector of $\boldsymbol{y}$ is absolutely continuous with density having the same form as above. Also, a desirable property for applications is that permutation of $\boldsymbol{y}$ components simply leads to a Dirichlet by permuting the corresponding parameters. Moreover, aggregation of some elements of $\boldsymbol{y}$ also leads to a Dirichlet distribution with the same type of aggregation in the vector of parameters. Furthermore, each component y_m is distributed as $Beta(\gamma_m, \gamma_0 - \gamma_m)$. Finally, if all γ_m parameters are proportionately large, then the Dirichlet can be approximated by a multivariate normal density. Note, however, that the Dirichlet distribution is not a LEF member, so any regression model based on it is not robust to distributional misspecification.

In order to allow for relationships between Dirichlet random vectors and a set of explanatory variables, a regression structure can be considered by introducing covariates in γ_m, $m = 1, ..., M$. However, estimating the covariates' relationships to γ_m may not be of much interest, so this paper proposes the reparameterization $\gamma_m \equiv \phi G_m$, $m = 1, ..., M$, with $\phi > 0$, from which one obtains $\gamma_0 \equiv \phi \sum_{m=1}^{M} G_m = \phi$, and the expression for the Dirichlet density becomes

$$f_{\boldsymbol{y}|\boldsymbol{X}}^{D}(\boldsymbol{y}; \phi, \boldsymbol{\beta}|\boldsymbol{X}) = \frac{\Gamma(\phi)}{\prod_{m=1}^{M}\Gamma(\phi G_m)} \prod_{m=1}^{M} y_m^{\phi G_m - 1}. \tag{10}$$

Consequently, (8) and (9) yield $E(y_m|\boldsymbol{X}) = G_m$ and

$$COV(y_l, y_m|\boldsymbol{X}) = \frac{G_l(\delta_{lm} - G_m)}{\phi + 1}, \quad l, m = 1, ..., M. \tag{11}$$

With this new formulation, $\boldsymbol{\beta}$ has the same interpretation as in conditional mean models and the parameter ϕ can be interpreted as a precision measure in the sense that, for fixed $\boldsymbol{G}$, the larger the value of ϕ, the smaller the elements of the covariance matrix $COV(y_l, y_m)$ – note that $\boldsymbol{y}$ degenerates at $\boldsymbol{G}$ if $\phi \to \infty$.[8]

[8] Instead of treating ϕ as a nuisance parameter, one may also specify it as a function of covariates (possibly distinct from $\boldsymbol{X}$), namely if interest lies in analyzing whether a variable contributes to the variances and covariances of $\boldsymbol{y}$ beyond its effect on the means.

Previous applications of the Dirichlet regression model (*e.g.*, Woodland, 1979; Chotikapanich and Griffiths, 2002) used different parameterizations, which, in contrast to the one proposed in this paper, are not generalizable to any possible specification for $E(y_m|\boldsymbol{X})$. For example, Woodland's (1979) proposal requires that

$$E(y_m|\boldsymbol{X}) = \frac{\mu_m(\boldsymbol{X};\boldsymbol{\beta})}{\sum_{l=1}^{M}\mu_l(\boldsymbol{X};\boldsymbol{\beta})}, \quad m = 1, ..., M,$$

since he sets $\gamma_m \equiv \phi\mu_m(\boldsymbol{X};\boldsymbol{\beta})$, where $\mu_m(\cdot)$ are index functions of the covariates.[9]

3.4 Regression Models for Proportions Obtained as Ratios of Observable Integers

In some applications, the response variables can be interpreted as ratios of integers, a situation that occurs when, *e.g.*, the elements of $\boldsymbol{y}$ are the proportions of individuals in a given group who select each of M mutually exclusive alternatives. When the number of individuals in each group (n) and the number of individuals in a given group who choose alternative m (n_m) are known, one can resort to models that make explicit use of this extra information. The alternative models now described may or may not produce more efficient estimators than the approaches previously discussed (which may be still valid), a fact that depends on the actual covariance structure of the data generating process. Unlike the Dirichlet regression model, the parametric models discussed next are defined for both boundary and interior values of the unit interval.

3.4.1 The Multinomial Regression Model

Consider, as a statistical unit, a group of $n > 0$ individuals (so, N now denotes the number of different groups in the available sample) and let $y_m = n_m/n$, with $n_m \geq 0$ observable integers such that $n = \sum_{m=1}^{M} n_m$. Thus, y_m can be viewed as the proportion of individuals belonging to the same group who select alternative m. Let π_m denote the probability that an individual selects alternative m. Then, $(n_1, ..., n_M) = n \times \boldsymbol{y}$ follows a multinomial p.f. with parameters n and $\boldsymbol{\pi} \equiv (\pi_1, ..., \pi_M)$. Formally,

$$f_{\boldsymbol{y}}^{M}(\boldsymbol{y}; n, \boldsymbol{\pi}) = \frac{n!}{\prod_{m=1}^{M}(ny_m)!}\prod_{m=1}^{M}\pi_m^{ny_m}, \tag{12}$$

where $\pi_M = 1 - \sum_{m=1}^{M-1}\pi_m$. Under this parametrization, $E(y_m) = \pi_m$ and $COV(y_l, y_m) = \pi_l(\delta_{lm} - \pi_m)/n$.

[9] Note that Woodland (1979) did not impose the constraint $\mu_m(\cdot) > 0$ that is necessary to enforce the required positivity of the Dirichlet parameters. In fact, he used a linear specification for $\mu_m(\cdot)$ in his empirical analysis.

A regression model can be accommodated by considering covariates in π_m. As before, let $\pi_m = G_m$, which leads to a conditional p.f. $f^M_{\boldsymbol{y}|\boldsymbol{X}}(\boldsymbol{y}; n, \boldsymbol{\beta}|\boldsymbol{X})$ and a conditional covariance matrix of $\boldsymbol{y}|\boldsymbol{X}$ with typical element

$$COV^M(y_l, y_m|\boldsymbol{X}) = \frac{G_l(\delta_{lm} - G_m)}{n}, \quad l, m = 1, ..., M. \tag{13}$$

The individual contribution to the log-likelihood can then be written as

$$\log L^M_i(\boldsymbol{\beta}) = \log\left[\frac{n_i!}{\prod_{m=1}^M (n_i y_{im})!}\right] + n_i \log L^{MB}_i(\boldsymbol{\beta}),$$

where $\log L^{MB}_i(\boldsymbol{\beta})$ is the individual contribution to the MB log-likelihood defined in (6).

The multinomial p.f. is a member of the LEF, so it can be used for QML estimation of the G_m parameters. If the data are actually generated by a multinomial law, then fully efficient ML estimation is achieved.

3.4.2 The Dirichlet-Multinomial Regression Model

Extra-multinomial dispersion can be allowed for by considering a joint distribution for $\boldsymbol{\pi}$. Mosimann (1962) shows that, if $\boldsymbol{\pi}$ follows a Dirichlet distribution, then $n \times \boldsymbol{y}$ follows a Dirichlet-multinomial (DM) mixture p.f.. In a regression context, with the proposed mean-dispersion parameterization for the Dirichlet conditional distribution,

$$f^D_{\boldsymbol{\pi}|\boldsymbol{X}}(\boldsymbol{\pi}; \phi, \boldsymbol{\beta}|\boldsymbol{X}) = \frac{\Gamma(\phi)}{\prod_{m=1}^M \Gamma(\phi G_m)} \prod_{m=1}^M \pi_m^{\phi G_m - 1},$$

one can formally write the DM conditional p.f. as

$$f^{DM}_{\boldsymbol{y}|\boldsymbol{X}}(\boldsymbol{y}; n, \phi, \boldsymbol{\beta}|\boldsymbol{X}) = \frac{n!\Gamma(\phi)}{\Gamma(\phi + n)} \prod_{m=1}^M \frac{\Gamma(\phi G_m + n y_m)}{\Gamma(\phi G_m)(n y_m)!}. \tag{14}$$

Several remarks about this expression seem appropriate. First, for $M = 2$, (14) reduces to the beta-binomial p.f. (see Johnson, *et al.*, 2005, Ch. 6); see *inter alia* Heckman and Willis (1977) and Santos Silva and Murteira (2009) for examples of the use of the beta-binomial model in a regression context. Second, the DM mixture has beta-binomial univariate marginals with parameters such that $E(y_m|\boldsymbol{X}) = E(\pi_m|\boldsymbol{X}) = G_m$ and

$$COV^{DM}(y_l, y_m|\boldsymbol{X}) = \frac{G_l(\delta_{lm} - G_m)}{n} \frac{\phi + n}{\phi + 1} \tag{15}$$

(see Johnson, *et al.*, 1997, Ch. 36). Thus, the DM mixture preserves the conditional means of the dependent variables, with reference to the multinomial p.f.. It is also obvious that the DM approach accommodates extra-multinomial dispersion, since $COV^{DM}(y_l, y_m|\boldsymbol{X}) = a \cdot COV^M(y_l, y_m|\boldsymbol{X})$, with $a = (\phi + n)/(\phi + 1) > 1$. Also, expression (11) (with π_m replacing

y_l and y_m) implies that $\lim_{\phi\to\infty} V(\pi_m|\boldsymbol{X}) = 0$, so the conditional distribution of π_m becomes degenerate at G_m and the DM model collapses to the multinomial as the parameter ϕ grows infinitely large. It is noted, however, that the DM p.f. is not a LEF member; therefore, ML estimation of its parameters is not robust to distributional misspecification.

One example of the use of the DM model (with a different parameterization) is provided by Mullahy (2010), in the context of financial asset portfolio shares.

4 Specification Analysis

All the alternative estimators for fractional regression models described above require the correct specification of the conditional mean of $\boldsymbol{y}$. Therefore, the primary focus of this section is on tests for assessing the appropriateness of $\boldsymbol{G}(\boldsymbol{X};\boldsymbol{\beta})$ as a model for $E(\boldsymbol{y}|\boldsymbol{X})$. The second part of the section considers tests for assessing distributional assumptions other than the conditional mean, implied by the Dirichlet, multinomial and Dirichlet-multinomial models.

4.1 Tests for the Conditional Mean

All the tests proposed in this section are tests for the exclusion of an $L-$dimensional vector of parameters $\boldsymbol{\eta}$ in the generalized model

$$E(\boldsymbol{y}|\boldsymbol{X}) = \boldsymbol{H}(\boldsymbol{X};\boldsymbol{\beta},\boldsymbol{\eta}) = [H_1(\boldsymbol{X};\boldsymbol{\beta},\boldsymbol{\eta}),...,H_M(\boldsymbol{X};\boldsymbol{\beta},\boldsymbol{\eta})]'.$$

Under the null hypothesis H_0: $\boldsymbol{\eta} = \boldsymbol{0}$, $\boldsymbol{H}(\boldsymbol{X};\boldsymbol{\beta},\boldsymbol{0}) = \boldsymbol{G}(\boldsymbol{X}\boldsymbol{\beta})$ is an appropriate specification for $E(\boldsymbol{y}|\boldsymbol{X})$. Such tests can be carried out in the usual manner, through a Lagrange multiplier (LM), Wald or likelihood ratio test. Given that the model under the alternative may be difficult to estimate, LM tests are proposed, which can be carried out by making use of Wooldridge's (1991) robust regression-based procedure, implemented upon QML or NLS estimation.

To this effect consider $\boldsymbol{y}^- \equiv (y_1,\ldots,y_{M-1})'$, $\boldsymbol{G}^- \equiv (G_1,\ldots,G_{M-1})'$ and $\boldsymbol{H}^- \equiv (H_1,\ldots,H_{M-1})'$, vectors of nonredundant fractional responses and corresponding conditional means under the null and alternative hypotheses, respectively. Also, let the alternative functional form of $E(y_m|\boldsymbol{X})$ be denoted by $H_m \equiv H_m(\boldsymbol{X};\boldsymbol{\beta},\boldsymbol{\eta})$. LM tests for the null conditional mean specification can be computed according to the following procedure, where $\widehat{(\cdot)}$ represents evaluation at the restricted NLS or QML estimators $\left(\hat{\boldsymbol{\beta}}',\boldsymbol{0}'\right)'$:

i. For the i-th observation, $i = 1,...,N$, obtain the $(M-1)\times L$ matrix $\tilde{\boldsymbol{W}}_i \equiv \hat{\boldsymbol{C}}_i^{-1/2}\cdot\nabla_{\boldsymbol{\eta}'}\hat{\boldsymbol{H}}_i^-$ and the $(M-1)\times K$ matrix $\tilde{\boldsymbol{X}}_i \equiv \hat{\boldsymbol{C}}_i^{-1/2}\cdot\nabla_{\boldsymbol{\beta}'}\hat{\boldsymbol{H}}_i^- = \hat{\boldsymbol{C}}_i^{-1/2}\cdot\nabla_{\boldsymbol{\beta}'}\hat{\boldsymbol{G}}_i^-$, with $\boldsymbol{C}_i$ an $(M-1)-$square matrix whose expression is detailed below and depends upon the estimator that is used.

ii. Compute the matrix OLS regression of $\tilde{\boldsymbol{W}}_i$ on $\tilde{\boldsymbol{X}}_i$, $i = 1, ..., N$, and obtain the $(M-1) \times L-$matrix of residuals, $\tilde{\boldsymbol{R}}_i$.

iii. For the i-th observation, $i = 1, ..., N$, obtain the $(1 \times L)-$vector $\tilde{\boldsymbol{e}}_i' \tilde{\boldsymbol{R}}_i$, where $\tilde{\boldsymbol{e}}_i \equiv \hat{\boldsymbol{C}}_i^{-1/2} \hat{\boldsymbol{e}}_i$ and $\hat{\boldsymbol{e}}_i \equiv \boldsymbol{y}_i^- - \hat{\boldsymbol{G}}_i^-$.

iv. Compute the OLS regression of the constant 1 on the $(1 \times L)-$vector $\tilde{\boldsymbol{e}}_i' \tilde{\boldsymbol{R}}_i$ and obtain the corresponding sum of squared residuals SSR.

v. Compute the LM statistic as $N - SSR$, which, under H_0, is asymptotically distributed as a chi-squared random variable with L degrees of freedom.

As mentioned, the formal expression of the $(M-1)-$square matrix $\boldsymbol{C}_i$ varies according to the estimation method that is used. Following Wooldridge (1991), under an LEF log-likelihood of the form

$$LL_i = a\left(\boldsymbol{G}_i^-, \boldsymbol{\nu}_i\right) + b\left(\boldsymbol{y}_i^-, \boldsymbol{\nu}_i\right) + \boldsymbol{y}_i^{-\prime} \boldsymbol{c}\left(\boldsymbol{G}_i^-, \boldsymbol{\nu}_i\right) \equiv a_i + b_i + \boldsymbol{y}_i^{-\prime} \boldsymbol{c}_i,$$

with $\boldsymbol{\nu}_i$ a vector of covariates and nuisance parameters, $a(\cdot)$ and $b(\cdot)$ scalars, and $\boldsymbol{c}(\cdot)$ an $(M-1)-$column vector of functions, $\boldsymbol{C}_i$ is defined as the inverse of the $(M-1)-$square matrix of derivatives $\nabla_{\boldsymbol{G}_i^{-\prime}} \boldsymbol{c}_i$. In the MB-QML case,

$$\boldsymbol{c}_i = \left(\log \frac{G_{i1}}{G_{iM}}, \ldots, \log \frac{G_{iM-1}}{G_{iM}}\right)', \quad G_{iM} = 1 - \sum_{m=1}^{M-1} G_{im},$$

from which,

$$\boldsymbol{C}_i = \begin{bmatrix} G_{i1}^{-1} + G_{iM}^{-1} & G_{iM}^{-1} & \cdots & G_{iM}^{-1} \\ G_{iM}^{-1} & G_{i2}^{-1} + G_{iM}^{-1} & \cdots & G_{iM}^{-1} \\ \vdots & \vdots & & \vdots \\ G_{iM}^{-1} & G_{iM}^{-1} & \cdots & G_{i,M-1}^{-1} + G_{iM}^{-1} \end{bmatrix}^{-1}.$$

Under multinomial-based QML,

$$\boldsymbol{c}_i = n_i \left(\log \frac{G_{i1}}{G_{iM}}, \ldots, \log \frac{G_{iM-1}}{G_{iM}}\right)',$$

so the previous matrix $\boldsymbol{C}_i$ should be scaled by the factor n_i^{-1}. With NLS estimation (which can be interpreted as QML based on a Gaussian likelihood with uncorrelated, unit-variance, errors), after replacing G_{iM} by $1 - \sum_{m=1}^{M-1} G_{im}$, one obtains

$$\boldsymbol{c}_i = \begin{bmatrix} 2G_{i1} + G_{i2} + \cdots + G_{iM-1} \\ G_{i1} + 2G_{i2} + \cdots + G_{iM-1} \\ \cdots \\ G_{i1} + G_{i2} + \cdots + 2G_{iM-1} \end{bmatrix} \Rightarrow \boldsymbol{C}_i = \left(\boldsymbol{I}_{M-1} + \boldsymbol{\iota}_{M-1} \boldsymbol{\iota}_{M-1}'\right)^{-1},$$

with $\boldsymbol{I}_{M-1}$ the identity matrix of order $(M-1)$ and $\boldsymbol{\iota}_{M-1}$ an $(M-1)-$column vector of ones. This inverse matrix can be checked to equal $\boldsymbol{I}_{M-1} - M^{-1}\boldsymbol{\iota}_{M-1}\boldsymbol{\iota}'_{M-1}$.[10]

The remainder of this section describes one general procedure valid to assess any conditional mean specification (RESET-type tests), as well as two special cases designed to assess the validity of the multinomial logit against, respectively, the nested logit and the random parameters logit. To the best of our knowledge, none of the three tests have been previously used to assess multivariate fractional models.

4.1.1 RESET-type Test

The RESET test was proposed originally by Ramsey (1969) as a general test for functional form misspecification for the (single equation) linear regression model but since then it has been applied to many other models. In fact, Papke and Wooldridge (1996) use the RESET test as a general functional form diagnostic for models of univariate fractional responses. Moreover, in the multivariate framework, various generalizations of the original RESET test have been presented for linear models (*e.g.*, Giles and Keil, 1997; Shukur and Edgerton, 2002; Alkhamisia, *et al.*, 2008). However, to the best of the authors' knowledge, the RESET test has never been applied to multivariate nonlinear regressions. The following extends the RESET test to the multivariate fractional case.

Assume that the components of the mean vector function under the null hypothesis, $G_m(\boldsymbol{X}\boldsymbol{\beta})$, $m = 1, ..., M$, are continuously differentiable and injective (the case for any specification based on continuous, strictly monotonous functions). Then, $\boldsymbol{G}(\boldsymbol{X}\boldsymbol{\beta})$ is invertible, in which case one may write the alternative vector model as

$$\boldsymbol{H}(\boldsymbol{X}\boldsymbol{\beta}) = \boldsymbol{G}\left\{\boldsymbol{G}^{-1}\left[\boldsymbol{H}(\boldsymbol{X}\boldsymbol{\beta})\right]\right\} \Leftrightarrow$$
$$H_m(\boldsymbol{X}\boldsymbol{\beta}) = G_m\left\{\boldsymbol{G}^{-1}\left[\boldsymbol{H}(\boldsymbol{X}\boldsymbol{\beta})\right]\right\}, \quad m = 1, ..., M,$$

where $\boldsymbol{G}^{-1}$ denotes the inverse vector function of $\boldsymbol{G}$.[11]

[10] For $M = 2$, the above definitions of $\boldsymbol{C}_i$ yield, respectively, $\hat{\boldsymbol{C}}_i^{-1/2} = \left[\hat{G}_{i2}\left(1 - \hat{G}_{i2}\right)\right]^{-1/2}$ (MB-QML estimation) and $\hat{\boldsymbol{C}}_i^{-1/2} = 1$ (NLS estimation). As expected, these constitute the weights that intervene in the artificial regressions used for testing the specification of the fractional conditional mean in the univariate case (see Ramalho, *et al.*, 2011, Section 4.1.1).

[11] A $K-$vector function of K variables

$$\boldsymbol{G} \equiv [G_1(\boldsymbol{Z}), ..., G_K(\boldsymbol{Z})]', \qquad \boldsymbol{Z} \equiv (Z_1, ..., Z_K),$$

where the G functions are continuously differentiable with domain and counterdomain open subsets of $\mathcal{R}^K$, is invertible in the neighborhood of the point $\boldsymbol{z}$, if and only if the Jacobian determinant of $\boldsymbol{G}$ with respect to $\boldsymbol{Z}$ is non-zero at $\boldsymbol{z}$. That is, an inverse vector function, $\boldsymbol{Z} = \boldsymbol{G}^{-1}$, exists in some neighborhood of $\boldsymbol{G}(\boldsymbol{z})$. The assumption that the G functions are injective prevents the Jacobian determinant from ever being zero, so $\boldsymbol{G}$ is

In general, the components of the vector $\boldsymbol{G}^{-1}[\boldsymbol{H}(\boldsymbol{X}\boldsymbol{\beta})]$ are nonlinear functions of its arguments (the elements of $\boldsymbol{X}\boldsymbol{\beta}$), which can be arbitrarily well approximated by Taylor polynomials of given order. For J large enough, each element of $\boldsymbol{G}^{-1}$ {name it $g_m \equiv g_m[\boldsymbol{H}(\boldsymbol{X}\boldsymbol{\beta})]$, $m = 1, ..., M$} can thus be approximated by $\sum_{j=1}^{J} P_{mj}$, where each $P_{mj} \equiv P_{mj}(\boldsymbol{X}\boldsymbol{\beta})$ represents a homogeneous polynomial of degree j in the elements of $\boldsymbol{X}\boldsymbol{\beta}$. That is, P_{mj} denotes a linear combination of the powers and cross-products (with degree j) of all the elements of $\boldsymbol{X}\boldsymbol{\beta}$. For the m-th component of $\boldsymbol{G}^{-1}$, one can write $P_{m1} = (\boldsymbol{X}\boldsymbol{\beta})'\boldsymbol{\delta}_m$ with $\boldsymbol{\delta}_m$ a vector of constants, $P_{m2} = (\boldsymbol{X}\boldsymbol{\beta})'\boldsymbol{\Delta}_m(\boldsymbol{X}\boldsymbol{\beta})$ with $\boldsymbol{\Delta}_m$ a matrix of constants, and so forth. If, for each component g_m, the appropriate constant in $\boldsymbol{\delta}_m$ is equal to one and the remaining constants (in δ_m, Δ_m, ...) are all zero, then g_m equals $\boldsymbol{x}_m'\boldsymbol{\beta}$ (the m-th element of $\boldsymbol{X}\boldsymbol{\beta}$), in which case $\boldsymbol{G}^{-1}[\boldsymbol{H}(\boldsymbol{X}\boldsymbol{\beta})]$ collapses to $\boldsymbol{X}\boldsymbol{\beta}$. Thus, the null model is nested within this approximation to $\boldsymbol{H}(\boldsymbol{X}\boldsymbol{\beta})$, a result that suggests an easy way to test the statistical validity of the null model.

The approximation to the alternative model can be generally expressed as

$$G_m\left[\boldsymbol{x}_1'\boldsymbol{\beta} + \boldsymbol{w}_{m1}'\boldsymbol{\eta}_{m1}, ..., \boldsymbol{x}_M'\boldsymbol{\beta} + \boldsymbol{w}_{mM}'\boldsymbol{\eta}_{mM}\right], \quad m = 1, ..., M,$$

where $\boldsymbol{w}_{ml}$ and $\boldsymbol{\eta}_{ml}$, $l = 1, ..., M$, denote, respectively, the vectors of added powers and cross-products, and associated coefficients. In all generality, this approximation augments the null model $E(y_m|\boldsymbol{X}) = G_m(\boldsymbol{X}\boldsymbol{\beta})$ by adding powers and cross-products of all the elements of $\boldsymbol{X}\boldsymbol{\beta}$ to each of its arguments. Therefore, testing the null hypothesis is equivalent to testing the significance of these (estimated) added terms in the augmented model. Clearly, the number of additional terms can be quite large, even with small J and M, so a fully fledged version of the RESET-type test can use a large number of degrees of freedom and be cumbersome to implement. Consequently, besides choosing a small enough value for J (2 or 3, say), one may carry out a simplified form of the test, by adopting exclusion restrictions that limit the number of terms to be added as arguments of $\boldsymbol{G}$ under the alternative hypothesis. For instance, a null MNL specification with $M = 3$ and $\boldsymbol{X}\boldsymbol{\beta} = [\boldsymbol{x}'\boldsymbol{\beta}_1, \boldsymbol{x}'\boldsymbol{\beta}_2]'$ can be assessed by testing the joint significance of η_1 and η_2 in

$$\frac{\exp\left[\boldsymbol{x}'\boldsymbol{\beta}_m + \eta_m\left(\boldsymbol{x}'\hat{\boldsymbol{\beta}}_m\right)^2\right]}{\sum_{l=1}^{3}\exp\left[\boldsymbol{x}'\boldsymbol{\beta}_l + \eta_l\left(\boldsymbol{x}'\hat{\boldsymbol{\beta}}_l\right)^2\right]}, \quad m = 1, 2, 3,$$

where $\boldsymbol{\beta}_3 = \boldsymbol{0}$ and $\eta_3 = 0$.

In general, let $\boldsymbol{\eta}$ denote the full vector of coefficients associated with the added terms; then, the null hypothesis H_0: $\boldsymbol{\eta} = \boldsymbol{0}$ is easily tested with a LM procedure. The LM test

invertible in all its domain. The existence of an inverse function of $\boldsymbol{G}$ is equivalent to saying that the system of equations $G_k = G_k(Z_1, ..., Z_K)$, $k = 1, ..., K$, can be solved for $Z_1, ..., Z_K$, as functions of $G_1, ..., G_K$.

is performed by considering, for the $m-$th row of the matrix $\nabla_{\boldsymbol{\eta}'}\hat{\boldsymbol{H}}_i^-$, a vector with typical element $\nabla_{\left(\boldsymbol{x}_{il}'\boldsymbol{\beta}\right)}\hat{G}_m\left(\boldsymbol{X}_i\boldsymbol{\beta}\right)\otimes\boldsymbol{w}_{iml}'$, where $\boldsymbol{w}_{iml}$ denotes the i-th observation of $\boldsymbol{w}_{ml}$, $l=1,...,M$. For instance, under the previous null MNL model, with the $m-$th element of $\boldsymbol{X}\boldsymbol{\beta}$ equal to $\boldsymbol{x}'\boldsymbol{\beta}_m$, $m=1,2$, $w_{11}=\left(\boldsymbol{x}'\hat{\boldsymbol{\beta}}_1\right)^2$, $w_{22}=\left(\boldsymbol{x}'\hat{\boldsymbol{\beta}}_2\right)^2$, $w_{12}=w_{21}=0$, $\boldsymbol{\eta}=[\eta_1,\eta_2]'$ and

$$\nabla_{\boldsymbol{\eta}'}\hat{\boldsymbol{H}}_i^- = \begin{bmatrix} \hat{G}_{i1}\left(1-\hat{G}_{i1}\right)w_{11} & -\hat{G}_{i1}\hat{G}_{i2}w_{22} \\ -\hat{G}_{i1}\hat{G}_{i2}w_{11} & \hat{G}_{i2}\left(1-\hat{G}_{i2}\right)w_{22} \end{bmatrix}.$$

4.1.2 Tests of the Multinomial Logit

The RESET test can be used to assess the validity of any null model for the conditional mean of $E\left(\boldsymbol{y}|\boldsymbol{X}\right)$. The following procedures are less general, specifically intended to assess the multinomial logit specification.

4.1.2.1 Test of the Multinomial Logit Against the Nested Logit

As previously mentioned, the standard logit model exhibits the IIA property, which implies zero correlation between fractions associated with any two alternative categories. When this stringent property does not hold, a more general specification is needed. The nested logit model is an analytically tractable generalization of the multinomial logit, suitable when proportions are distributed among alternatives as a result of a hierarchical decision sequence.

The expression of the nested logit model is given in (4). This expression nests the multinomial logit model, which corresponds to the null hypothesis H_0: $\eta_l=0$, $l=1,...,L$. The LM test of this hypothesis can be implemented by using, for each conditional mean function, the $(1\times L)$ vector of partial derivatives $\nabla_{\boldsymbol{\eta}'}\hat{H}_{im}$ with typical element

$$\nabla_{\eta_l}\hat{H}_{im}=\hat{G}_{im}\left(1\left(m,l\right)\left\{\log\left[\sum_{j\in S_l}\exp\left(\boldsymbol{x}_{ij}'\hat{\boldsymbol{\beta}}\right)\right]-\boldsymbol{x}_{im}'\hat{\boldsymbol{\beta}}\right\}-\right.$$
$$\left.\sum_{j\in S_l}\hat{G}_{ij}\left\{\log\left[\sum_{j\in S_l}\exp\left(\boldsymbol{x}_{ij}'\hat{\boldsymbol{\beta}}\right)\right]-\boldsymbol{x}_{ij}'\hat{\boldsymbol{\beta}}\right\}\right),\quad l=1,\ldots,L,$$

with $1\left(m,l\right)$ denoting an indicator function equal to one if alternative m belongs to subset S_l, and zero otherwise.

4.1.2.2 Test of the Multinomial Logit against the Random Parameters Logit

The random parameters logit expressed in (5) provides another generalization of the basic multinomial logit model. Frequently, this model involves non-analytical expressions, requiring simulation or numerical approximation to be computed. As a consequence, approximate models that facilitate estimation and inference in this context are useful. One such approximation,

that allows for a full range of correlation structures across alternatives and does not depend on the form of the distribution of $\boldsymbol{\beta}$, is provided by the "heterogeneity adjusted logit" (HAL) model, proposed by Chesher and Santos Silva (2002). This approximation also nests the basic multinomial logit, so the approach enables easy assessment of the latter model.

The HAL approximation to the random parameters logit can be expressed as

$$H_m = \frac{\exp\left(\boldsymbol{x}_m'\boldsymbol{\beta} + \sum_{j=1}^{M}\sum_{\substack{k=1 \\ j,k\neq m^*}}^{M}\eta_{jk}w_m^{jk}\right)}{\sum_{l=1}^{M}\exp\left(\boldsymbol{x}_l'\boldsymbol{\beta} + \sum_{j=1}^{M}\sum_{\substack{k=1 \\ j,k\neq m^*}}^{M}\eta_{jk}w_l^{jk}\right)}, \quad m = 1, ..., M,$$

where m^* is arbitrarily chosen from the set $\{1, ..., M\}$, $w_m^{jk} = 0$ for $m = m^*$ and, for $m \neq m^*$,

$$w_m^{jk} = \begin{cases} \frac{1}{2} - G_m, & j = k = m \\ 0, & j = k \neq m \\ -G_j, & j \neq k = m \\ -G_k, & k \neq j = m \\ 0, & j \neq m, k \neq m \end{cases},$$

G_m denotes the multinomial logit conditional mean $E(y_m|\boldsymbol{X})$, and η_{jk} are parameters. The choice of m^* is equivalent to measuring parameter heterogeneity relative to alternative m^*. The additional variables, w_m^{jk}, and the interpretation of the η_{jk} parameters vary with the choice of m^* but the approximate means, H_m, are invariant to this choice (see Chesher and Santos Silva, 2002, Sec. 2, for details).

In this setting, the multinomial logit corresponds to the null hypothesis H_0: $\boldsymbol{\eta} = \mathbf{0}$, with $\boldsymbol{\eta}$ the vector of coefficients of added terms, η_{jk}. The LM test for the omission of these terms can then be implemented by using

$$\nabla_{\boldsymbol{\eta}'}\hat{H}_{im} = \hat{G}_{im}\left(1 - \hat{G}_{im}\right)\boldsymbol{w}_m',$$

where $\boldsymbol{w}_m$ denotes the vector of added terms, w_m^{jk}, in H_m.

4.2 Tests for Other Distributional Assumptions

Testing the correct specification of $E(\boldsymbol{y}|\boldsymbol{X})$ is clearly the most important issue in fractional regression models. Nevertheless, once the functional form is selected, one may also examine whether a given distribution is appropriate for modelling, so as to obtain efficient ML estimators. This, in turn, prompts the convenience of assessing the statistical validity of the selected parametric model. The standard test for misspecification of a parametric likelihood function is the information matrix (IM) test introduced by White (1982), which, however, can be very

burdensome to compute. A less ambitious approach, corresponding to a restricted version of the IM test and not too difficult to implement, is to resort to a conditional moment (CM) test (Newey, 1985, Tauchen, 1985) of the moment assumptions imposed by the specification that is adopted for the conditional distribution of the response variables. In case of rejection of the null hypothesis, the model imposing the moment assumptions under test should obviously be discarded and a different model should be entertained.

If the CM test is conducted under the assumption of correct specification of $E(\boldsymbol{y}|\boldsymbol{X})$, then typically one will be interested in assessing the validity of the second moment conditions,

$$E\left[\left(y_{il}-G_{il}\right)\left(y_{im}-G_{im}\right)-\alpha_i G_{il}\left(\delta_{lm}-G_{im}\right)|\,\boldsymbol{X}_i\right]=0,\quad 1\leq l\leq m\leq M-1,$$

where α_i is given by $1/(\phi+1)$ (Dirichlet), $1/n_i$ (multinomial) or $(\phi+n_i)/[(\phi+1)n_i]$ (Dirichlet-multinomial). The OPG version of the test statistic can be computed as N times the uncentered R^2 from the auxiliary OLS regression

$$1=\hat{\boldsymbol{m}}_i'\boldsymbol{\lambda}_1+\hat{\boldsymbol{s}}_i'\boldsymbol{\lambda}_2+error,$$

where $\hat{\boldsymbol{m}}_i\equiv\boldsymbol{m}\left(\boldsymbol{y}_i,\boldsymbol{x}_i,n_i,\hat{\boldsymbol{\beta}},\hat{\phi}\right)$ denotes the i-th observation of the vector of moment conditions imposed by the model under consideration, $\hat{\boldsymbol{s}}_i$ refers to the i-th element of the corresponding score vector, and $\widehat{(\cdot)}$ now denotes evaluation at ML estimates. The expressions of the individual contribution to the score vector are given by, respectively,

Dirichlet

$$\boldsymbol{s}_i=\left[\begin{array}{c}\nabla_\phi\log L_i^D\\ \nabla_{\boldsymbol{\beta}}\log L_i^D\end{array}\right]=\left[\begin{array}{c}\Psi(\phi)+\sum_{m=1}^M\left\{G_{im}\left[\log y_{im}-\Psi(\phi G_{im})\right]\right\}\\ \sum_{m=1}^M\left[\phi\log y_{im}-\Psi(\phi G_{im})\right]\nabla_{\boldsymbol{\beta}}G_{im}\end{array}\right],$$

Multinomial

$$\boldsymbol{s}_i=\nabla_{\boldsymbol{\beta}}\log L_i^M=n_i\sum_{m=1}^M\frac{y_{im}}{G_{im}}\nabla_{\boldsymbol{\beta}}G_{im},$$

Dirichlet-multinomial

$$\boldsymbol{s}_i=\left[\begin{array}{c}\nabla_\phi\log L_i^{DM}\\ \nabla_{\boldsymbol{\beta}}\log L_i^{DM}\end{array}\right]=\left[\begin{array}{c}\Psi(\phi)-\Psi(\phi+n_i)+\sum_{m=1}^M G_{im}\left[\Psi(\phi G_{im}+n_i y_{im})-\Psi(\phi G_{im})\right]\\ \phi\sum_{m=1}^M\left[\Psi(\phi G_{im}+n_i y_{im})-\Psi(\phi G_{im})\right]\nabla_{\boldsymbol{\beta}}G_{im}\end{array}\right],$$

where $\Psi(\cdot)$ denotes the digamma function (first derivative of the log-gamma function, $\log\Gamma(\cdot)$) and

$$\nabla_{\boldsymbol{\beta}}G_{im}=\sum_{l=1}^M g_{im}^{(l)}\boldsymbol{x}_{il},\ \ m=1,...,M,$$

with $g_{im}^{(l)}\equiv\nabla_{\left(\boldsymbol{x}_{il}'\boldsymbol{\beta}\right)}G_{im}$, the partial derivative of G_{im} with respect to the l-th component of $\boldsymbol{X}_i\boldsymbol{\beta}$. Under the null hypothesis of correct moment specification, the test statistic is asymptotically

distributed as a chi-squared random variate with number of degrees of freedom equal to the dimension of the $\hat{\boldsymbol{m}}_i$ vector.

The multinomial model can also be tested against the DM, as the latter nests the former model under H_0: $\delta \equiv 1/\phi \to 0^+$. Given that the Wald and likelihood ratio tests' null asymptotic distributions are affected by the location of the null hypothesis on the boundary of the parameter space, also in this case it is easier to implement an LM test, which retains its usual asymptotic chi-squared null law.

5 Monte Carlo Study

This section uses Monte Carlo methods to illustrate the finite-sample performance of most of the estimators and tests discussed throughout this paper. In the first subsection, all experiments involve estimation (by NLS, QML or ML) of a conditional mean function which is correctly specified as MNL. The second subsection illustrates the small sample size and power of the conditional mean tests discussed in the paper when the null hypothesis is a MNL. All experiments are based on 5,000 replications of samples of size $N = 100$, 250, 500 or 1000, with computations performed using the R software.

5.1 Performance of Alternative Estimators

The experiments in this subsection assume correct specification of $E(\boldsymbol{y}|\boldsymbol{X})$ as MNL with $M = 5$ shares, thus involving only alternative-invariant covariates. Formally, the ('true' and specified) model for the conditional mean of the dependent variables can be expressed as

$$G_m = \frac{\exp(\boldsymbol{x}'\boldsymbol{\beta}_m)}{\sum_{l=1}^{3} \exp(\boldsymbol{x}'\boldsymbol{\beta}_l)}, \tag{16}$$

where $\boldsymbol{x} \equiv (1, x_2)'$, with conformable parameter vectors $\boldsymbol{\beta}_m \equiv (\beta_{1m}, \beta_{2m})$, $m = 1, 2, 3, 4$, and $\boldsymbol{\beta}_5 = \mathbf{0}$. The variable x_2 is newly drawn in each replica, obtained as i.i.d. draws from a displaced $Exponential(1)$ distribution with mean zero.

Different parameter values are considered in each of two different designs (named A and B) for $\boldsymbol{\beta}_m$, $m = 1, 2, 3, 4$. While $\beta_{2m} = 1$, $m = 1, 2, 3, 4$, in both designs, the values assigned to β_{1m} are chosen in order to yield two very distinct distributions of shares for the different alternatives. As can be seen from Table 1, the mean shares for the five alternatives are identical in Design A and quite unbalanced in Design B, where $G_m \simeq 2G_{m-1}$, $m = 2, 3, 4, 5$.

The observations on $\boldsymbol{y}$ are obtained as i.i.d. draws from three different distributions: the Dirichlet p.f. presented in (10), with $\phi \in \{10, 20, 30, 40, 50\}$; the multinomial p.f. presented in (12), with G_m instead of π_m and n obtained as an i.i.d. draw from a discrete uniform

Table 1

Experimental Designs

Design	β_{11}	β_{12}	β_{13}	β_{14}	Mean shares (%)*				
					y_1	y_2	y_3	y_4	y_5
A	0.23	0.23	0.23	0.23	20.0	20.0	20.0	20.0	20.0
B	−2.72	−2.02	−1.32	−0.63	3.2	6.4	13.0	25.8	51.6

* Mean shares obtained from a simulated sample of size 100,000.

p.f. $\mathcal{U}(1, n_{\max})$, where $n_{\max} \in \{11, 21, 31, 41, 51\}$; and the Dirichlet-multinomial mixture p.f. presented in (14), with n and ϕ defined as in the previous experiments. In each case, the values of ϕ and/or n imply different degrees of variability of the response variables and, thus, are bound to influence the precision of the various estimators.

For the case of a Dirichlet-distributed response variable and $N = 250$, Figure 1 displays the root mean squared errors (RMSE) of three alternative estimators of β_{2m}, $m = 1, 2, 3, 4$: NLS, MB-QML and Dirichlet-ML (D-ML), all consistent under the two designs considered. As expected, in all cases D-ML exhibits an efficiency advantage over the other two estimators, which may be substantial for smaller values of ϕ (i.e. when y displays more variability) but is largely attenuated for higher values of ϕ. NLS performs invariably worst, namely for small values of ϕ, proving to be the method whose estimates' precision is most sensitive to the variability of the dependent variables. The precision of the estimates is also very sensitive to the relative importance of each alternative: in Design B, irrespective of the estimator considered, the RMSE of the estimates of the parameters associated with alternatives exhibiting lower mean shares are often much larger.

[Figure 1 about here]

In the second experiment, where the response variables are obtained as integers ratios from a conditional multinomial distribution, the conditional mean parameters are estimated by NLS, MB-QML, D-ML, multinomial-based ML (MULT) and Dirichlet-multinomial ML (DM-ML). For the D-ML estimator to be computed, the samples were modified by replacing the zero and unit values of the responses with, respectively, 10^{-6} and $1 - 10^{-6}$. Note that the D-ML estimator is expected to be inconsistent in this case, even if no boundary values were observed. Figure 2 plots the RMSE of the parameters' estimates, for the five different values considered for $n_{\max}$.

[Figure 2 about here]

With regard to the relative performance of NLS and MB-QML methods, once again NLS is the least efficient. The MULT (ML) estimator appears more efficient than MB-QML, which is due to the fact that the former estimator makes use of potentially useful information (on n) that is ignored by the latter. Incidentally, the closeness of the MULT and DM-ML estimators' performance is also expected because the latter method nests the former. However, unless there is reason to suspect that the data suffer from extra-multinomial dispersion, the MULT estimator should be preferred to DM-ML. Otherwise, with no extra-multinomial dispersion (the case here), convergence of the DM-ML method is often difficult to achieve and (understandably) always for quite large ϕ estimates (see the remarks on eq. (14)). As a practical consequence, difficulty in obtaining DM-ML estimates may be taken as indication that there is simply no unobserved heterogeneity to account for, so the MULT or MB-QML approaches may well suffice. In what concerns the performance of D-ML, its estimates are biased, as expected. In fact, the difference between its RMSE and that of the other estimators is entirely due to its bias.

Figure 3 sums up the RMSE results of the experiment where the response variables have a conditional Dirichlet-multinomial distribution. The conditional mean parameters are estimated with the same five methods used in the previous experiment (zeros and ones are again modified in the case of D-ML estimation). The first two rows of Figure 3 refer to the case where ϕ is set to 10 and $n_{\max} \in \{11, 21, 31, 41, 51\}$, while in the last two rows $n_{\max}$ is set to 11 and $\phi \in \{10, 20, 30, 40, 50\}$. Note that for the same values of G_m and n, the variances of the dependent variables are now considerably higher than in the previous experiment (in some cases, they can be more than five times higher – compare expressions (15) and (13))

[Figure 3 about here]

Again, the inconsistent D-ML estimator fares much worse than the consistent estimators and the NLS estimator displays the highest RMSE of the remaining estimators. The MULT estimator outperforms MB-QML in all cases, so use of the available information on n seems advantageous in what concerns QML methods. However, the best performer is DM-ML, which shows again the importance of using ML estimators whenever reliable information on the data distribution is available. Nevertheless, the gains in precision relatively to the MULT estimator are relatively unimportant in most cases.

Finally, in Figure 4 the performance of alternative estimators under different sample sizes, $N \in \{100, 250, 500, 1000\}$, is investigated for some selected cases of Design B. In most cases, the RMSE's of all estimators decrease substantially as N grows, with the efficiency advantage of ML over QML and NLS estimation being much less relevant for large sample sizes. Actually, for N large, using ML instead of QML produces sizeable gains in precision only for the regression

coefficients of the alternatives displaying very low mean shares. On the other hand, note how the estimates produced by NLS are often much less precise than those of their competitors, even for $N = 500$. This is mainly a consequence of the extreme values that NLS occasionally yields. For the cases where the responses variables are not Dirichlet-distributed, the D-ML estimator displays a stable RMSE across different sample sizes, as a result os its inconsistency.

[Figure 4 about here]

5.2 Performance of Alternative Conditional Mean Tests

Next, the performance of alternative tests for conditional mean assumptions is investigated in the particular case where the null hypothesis is the MNL model (16) and the responses have a Dirichlet distribution with nuisance parameter $\phi \in \{10, 50\}$. The three specification tests proposed in Section 4.1 are included in this study: the RESET test, which is a general test for model misspecification; and the tests designed to be sensitive to departures from the multinomial logit in the direction of the nested logit or the random parameters logit, which are denoted by NESTED and CSS, respectively.

For all tests, two different versions are computed; one more general, indexed by the subscript 'g', and a simplified version that results from the adoption of some exclusion restrictions, indexed by the subscript 's'. For the RESET test, only the square of the fitted power $x'\hat{\beta}_m$ is added to equation m and the associated parameters η_m are allowed to differ across alternatives in one case ($RESET_g$) or are constrained to be identical across equations in another case ($RESET_s$). For the NESTED statistic, it is assumed that the practitioner thinks that the alternatives may be grouped in two nests, one grouping two alternatives and the other the remaining three categories. With this information, two versions of NESTED are implemented; one that considers all the ten possible combinations of such two nests ($NESTED_g$), and another that is based on the following two nests about which the empirical researcher is particularly suspicious: S_1, containing alternatives $m = 1, 2$; and S_2, containing alternatives $m = 3, 4, 5$ ($NESTED_s$). Finally, the full version of the CSS test is implemented (CSS_g) as well as a simplified version that only assumes randomness of the parameters β_{21} and β_{22}, independently distributed from each other (CSS_s).

All test versions are implemented as LM statistics based on MB-QML estimators and have asymptotic chi-square distributions. However, the number of degrees of freedom of their distributions is very different: it is 1 for $RESET_s$, 2 for $NESTED_s$ and CSS_s, 4 for $RESET_g$, 10 for CSS_g and 20 for $NESTED_g$. Especially in small samples, the high number of degrees of freedom displayed by the general versions of the tests may affect substantially their finite sample

properties. However, being more general, the full versions of all tests are sensitive to a wider range of model misspecifications.

5.2.1 Empirical size

To investigate the size properties of the tests in finite samples, the data are again generated as in the experiments of the previous section. Figure 5 displays the percentage of rejections of the (correct) null hypothesis for a nominal level of 5% for both Designs A and B and $N \in \{100, 250, 500, 1000\}$. The horizontal lines represent the limits of a 95% confidence interval for the nominal size.

[Figure 5 about here]

Figure 5 reveals that the general versions of the tests are much more conservative than the corresponding simplified versions. In fact, CSS_g and $NESTED_g$ are undersized for all the sample sizes simulated and $RESET_g$ is also undersized in most cases. Unlike the general versions, the performances of the simplified variants of each test are very heterogeneous. The $NESTED_s$ test is undersized in all cases. The CSS_s test performs relatively well in Design A but is undersized in Design B in most cases. The $RESET_s$ test is clearly the best performer, displaying an empirical size that is not significantly different from the nominal level of 5% (most cases) or only slightly higher.

5.2.2 Empirical power

The power properties of the tests are examined considering four distinct types of misspecification sources. In the first two cases, the functional form adopted for the structural model is the correct one but a relevant covariate is omitted or mismeasured. In the remaining two experiments, the correct structural model is either a nested logit or a random parameters logit. All the test variants are applied in the four cases, even the NESTED and CSS statistics that were constructed to be particularly sensitive to the last two types of model misspecification, respectively. The results are summarized in Figure 6 for $N = 250$.

[Figure 6 about here]

The first row of Figure 6 considers the case of the omission of a quadratic term of an included regressor. In particular, the conditional mean of the responses variables is still generated from the MNL model (16), but now $\boldsymbol{x} \equiv \left(1, x_2, x_2^2\right)'$, $\boldsymbol{\beta}_m \equiv (\beta_{1m}, \beta_{2m}, \beta_{3m})$, $\beta_{3m} = \theta$, where $m = 1, 2, 3, 4$ and $\theta \in \{0, 0.05, 0.1, 0.15, 0.2, 0.25\}$, and $\beta_{35} = 0$. In general terms, the power of the tests

increase as θ (i.e. the relative importance of the omitted covariate) and ϕ (i.e. the precision of the parameter estimates) increase, as could be anticipated. Unsurprisingly, the RESET tests are clearly the best performers in these experiments, with $RESET_s$ displaying the highest power, which is a consequence of β_{3m} being constant across alternatives, as assumed by this RESET version. In contrast, the specific versions of the NESTED and CSS statistics exhibit very low power, while their general variants are able to detect that some form of misspecification is present in the estimated model but display in general much less power than the RESET tests.

The case of covariate measurement error is analyzed in the second row of Figure 6. The conditional mean of the responses variables is generated from the MNL model (16) as in the size experiments, but estimation is based on $\boldsymbol{x}^* \equiv (1, x_2^*)'$, where $x_2^* = x_2 + u$. The measurement error u is generated from a Student-t distribution with five degrees of freedom, scaled to have variance $\theta \in \{0, 0.1, 0.2, 0.3, 0.4, 0.5\}$. Again, as expected, the most powerful tests are the two RESET versions. Because the measurement error affects all share equations in a similar way, $RESET_s$ performs better than $RESET_g$. Regarding the other tests, the conclusions are relatively similar to the previous experiments, with the main difference being that both CSS versions have very low power in Design B.

A very different picture appears in the third row of Figure 6, where the data are generated according to the nested logit specification (4). The same nests S_1 and S_2 defined above for constructing $NESTED_s$ are employed to generate the data, with the parameter η_l, $l = 1, 2$, that appears in (4) being set to $\theta \in \{-0.75, -0.6, -0.45, -0.3, -0.15, 0\}$. Now, the best performers are the simplified versions of the NESTED and CSS tests (recall that the latter assumes randomness of only β_{2m}, $m = 1, 2$, and that nest S_1 contains alternatives $m = 1, 2$), which, unlike in the previous experiments, are more powerful than their corresponding generalized versions. Also in contrast to the previous two experiments, $RESET_g$ is now more powerful than $RESET_s$, which reflects the fact that the simulated misspecification does not affect each share equation in a similar way. More importantly, note that $RESET_g$ does not lag far behind the $NESTED_s$ and CSS_s tests and performs similarly to, or better than, $NESTED_g$ and CSS_g.

Finally, in the last row of Figure 6 the data are generated assuming the random parameters logit specification (5) for $E(\boldsymbol{y}|\boldsymbol{X})$. As in the construction of CSS_s, only β_{2m}, $m = 1, 2$, are random, having independent distributions. In particular, we set $\beta_{2m} = 1 \pm \theta$, $m = 1, 2$, with $\theta \in \{0, 0.2, 0.4, 0.6, 0.8, 1\}$. Similar conclusions to the nested logit case are achieved, with the RESET tests being again more powerful than $NESTED_g$ and CSS_g. As information on the precise form of the nests or parameter randomness is often not available, this promising behaviour of the RESET tests illustrates the usefulness of general misspecification tests also in this context.

6 Concluding Remarks

This paper presents alternative estimating and testing empirical strategies for cross-section multivariate fractional regression models. These include models of the conditional mean, estimable through NLS or QML methods, and fully parametric regression models, estimable by ML. Among QML methods, the multivariate Bernoulli stands out as a tool of choice, due to its user friendliness and appropriate statistical properties, requiring only correct specification of the conditional mean of the response variables. In any case, when the data under study consist of ratios of observable integers, the multinomial and multinomial-based mixture models are viable alternatives which may provide more efficient estimators. The multinomial and the Dirichlet-multinomial mixture can also prove useful when the data contain boundary observations, which are incompatible with the Dirichlet-ML approach.

The simulation study included in the paper gives evidence of the relative advantage of QML (multivariate Bernoulli and multinomial) approaches, which, besides being easy to use, compete well with the ML estimators (Dirichlet and multinomial-Dirichlet), even when the latter are implemented under fully correct distributional assumptions, especially when the sample size is large, the responses variables are not too dispersed and some fractions are not too small in relative terms. The same cannot be said of the NLS estimator, which is found to behave very poorly in several situations. Thus, namely given the availability of the multivariate Bernoulli and multinomial QML estimators, use of the NLS method seems inadvisable.

The article also discusses the specification analysis of multivariate fractional regression models, with an emphasis on tests of the conditional mean specification. Along with tests that are applicable to any conditional mean functional form (RESET-type tests), specific tests of the multinomial logit model are also proposed. All conditional mean specification tests are proposed as LM tests, implemented upon QML estimation and using artificial OLS regressions. The Monte Carlo study reveals that RESET-type tests are particularly useful in this context, given its simplicity and good performance across a range of possible model misspecifications.

The present text has suggested several hints for future related work. Among others, the extension of some of the proposed techniques to multivariate fractional panel data stands out as an important avenue for future research.

References

Aitchison, J. (1982), "The Statistical Analysis of Compositional Data (with discussion)", *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, 44(2), 139-177.

Aitchison, J. and J. Egozcue (2005), "Compositional Data Analysis: Where Are We and Where Should We Be Heading?", *Mathematical Geology*, 37(7), 829-850.

Alkhamisia, M., G. Khalaf and G. Shukur (2008), "The Effect of Fat-tailed Error Terms on the Properties of System-wise RESET Test", *Journal of Applied Statistics*, 35(1), 101-113.

Chesher, A. and J. Santos Silva (2002), "Taste Variation in Discrete Choice Models", *The Review of Economic Studies*, 69, 1, 147-168.

Chotikapanich, D. and W. E. Griffiths (2002), "Estimating Lorenz Curves Using a Dirichlet Distribution", *Journal of Business & Economic Statistics,* 20(2), 290-295.

Considine, T.J. and T.D. Mount (1984), "The Use of Linear Logit Models for Dynamic Input Demand Systems", *Review of Economics and Statistics*, 66, 434-443.

Dubin, J. (2007), "Valuing Intangible Assets with a Nested Logit Market Share Model", *Journal of Econometrics*, 139, 285-302.

Ferrari, S. and F. Cribari-Neto (2004), "Beta Regression for Modelling Rates and Proportions", *Journal of Applied Statistics*, 31(7), 799-815.

Fry, J.M., T.R.L. Fry and K.M. McLaren (1996), "The Stochastic Specification of Demand Share Equations: Restricting Budget Shares to the Unit Simplex", *Journal of Econometrics*, 73, 377-385.

Giles, D. and A. Keil (1997), "Applying the RESET Test in Allocation Models: a Cautionary Note", *Applied Economics Letters*, 4, 359-363.

Gouriéroux, C., A. Monfort and A. Trognon (1984), "Pseudo Maximum Likelihood Methods: Theory.", *Econometrica*, 52, 681-700.

Heckman, J. J and R. J. Willis, (1977), "A Beta-logistic Model for the Analysis of Sequential Labor Force Participation by Married Women", *Journal of Political Economy*, 85, 27-58.

Heien, D. and C. R. Wessells (1990), "Demand Systems Estimation with Microdata: A Censored Regression Approach", *Journal of Business & Economic Statistics*, 8(3), 365-371.

Hermalin, B.E. and N.E. Wallace (1994), "The determinants of efficiency and solvency in savings and loans", *Rand Journal of Economics*, 25(3), 361-381.

Johnson, N., A. Kemp, and S. Kotz (2005), *Univariate Discrete Distributions, 3rd. ed.*, Wiley.

Johnson, N., S. Kotz and N. Balakrishnan (1997), *Discrete Multivariate Distributions*, Wiley.

Katz, J. N. and G. King (1999), "A Statistical Model for Multiparty Electoral Data", *Political Science*, 93(1), 15-32.

Klawitter, M. (2008), "The Effects of Sexual Orientation and Marital Status on How Couples Hold Their Money", *Review of Economics of the Household*, 6(4), 423-446.

Kotz, S., N. Balakrishnan and N. Johnson (2000), *Continuous Multivariate Distributions*, *Vol. 1*, Wiley.

Lee, L.-F. and M. Pitt (1986), "Microeconometric Demand Systems With Binding Nonnegativity Constraints: The Dual Approach", *Econometrica*, 54(5), 1237-1242.

Mosimann, J. (1962), "On the Compound Multinomial Distribution, the Multivariate Beta-distribution, and Correlation Among Proportions", *Biometrika*, 49, 65–82.

Mullahy, J. (2010), "Multivariate Fractional Regression Estimation of Econometric Share Models", *2nd. International Health Econometrics Workshop*, Rome, 2010.

Mullahy, J. and S. Robert (2010), "No Time to Lose: Time Constraints and Physical Activity in the Production of Health", *Revue of the Economics of the Household*, 8(4), 409-432.

Newey, W. (1985), "Maximum Likelihood Specification Testing and Conditional Moment Tests", *Econometrica*, 53, 1047-1070.

Paolino, P. (2001), "Maximum Likelihood Estimation of Models with Beta-distributed Dependent Variables", *Political Analysis*, 9(4), 325-346.

Papke, L. E. and J. M. Wooldridge (1996), "Econometric Methods for Fractional Response Variables with an Application to 401(k) Plan Participation Rates", *Journal of Applied Econometrics*, 11(6), 619-632.

Poterba, J. and A. Samwick (2002), "Taxation and Household Portfolio Composition: US Evidence from the 1980s and 1990s", *Journal of Public Economics*, 87, 5-38.

Pu, C., V. Lan, Y. Chou and C. Lan (2008), "The Crowding-out Effects of Tobacco and Alcohol Where Expenditure Shares are Low: Analyzing Expenditure Data for Taiwan, *Social Science & Medicine*, 66(9), 1979-1989.

Ramalho, E., J. Ramalho and J. Murteira (2011), "Alternative Estimating and Testing Empirical Strategies for Fractional Regression Models", *Journal of Economic Surveys*, 25(1), 19-68.

Ramsey, J.B. (1969), "Tests for Specification Errors in Classical Linear Least-Squares Regression Analysis", *Journal of the Royal Statistical Society B*, 31, 350-371.

Santos Silva, J. M. C. and J. Murteira (2009), "Estimation of Default Probabilities Using Incomplete Contracts Data", *Journal of Empirical Finance*, 16(3), 457-465.

Shukur, G. and D. Edgerton (2002), "The Small Sample Properties of the Reset Test as Applied to Systems of Equations", *Journal of Statistical Computation and Simulation*, 72(12), 909-924.

Sivakumar, A. and C. Bhat (2002), "Fractional Split-Distribution Model for Statewide Commodity-Flow Analysis", *Transportation Research Record*, 1790, 80-88.

Tauchen, G. (1985), "Diagnostic Testing and Evaluation of Maximum Likelihood Models", *Journal of Econometrics*, 30, 415-443.

Train, K. E. (2009), *Discrete Choice Methods with Simulation*, *2nd. ed.*, Cambridge University Press.

Wales, T. J. and A. D. Woodland (1983), "Estimation of Consumer Demand Systems with Binding Non-negativity Constraints", *Journal of Econometrics*, 21, 263-285.

Wang, H., L. Zhang, and W. Hsiao (2006). "Ill health and its potential influence on household consumptions in rural China", *Health Policy*, 78(2-3), 167-177.

White, H. (1982), "Maximum Likelihood Estimation of Misspecified Models", *Econometrica*, 50(1), 1-25.

Woodland, A. D. (1979), "Stochastic Specification and the Estimation of Share Equations", *Journal of Econometrics*, 10, 361-383.

Wooldridge, J. M. (1991), "Specification Testing and Quasi-maximum Likelihood Estimation", *Journal of Econometrics*, 48, 29-55.

Wooldridge, J. M. (2002), *Econometric Analysis of Cross Section and Panel Data*, MIT Press.

Ye, X. and R. M. Pendyala (2005), "A Model of Daily Time Use Allocation Using Fractional Logit Methodology", in: H.S. Mahmassani, Ed., *Transportation and Traffic Theory: Flow, Dynamics, and Human Interaction*, Elsevier Science Ltd, pp. 507.524.

Yin, R.S., Q. Xiang, J. T. Xu and X.Z. Deng (2010), "Modeling the driving forces of the land use and land cover changes along the upper Yangtze river of China", *Environmental Management*, 45(3), 454-65.

Figure 1: RMSE comparison of alternative estimators for multivariate fractional regression models
(Dirichlet−distributed response variable; N = 250)

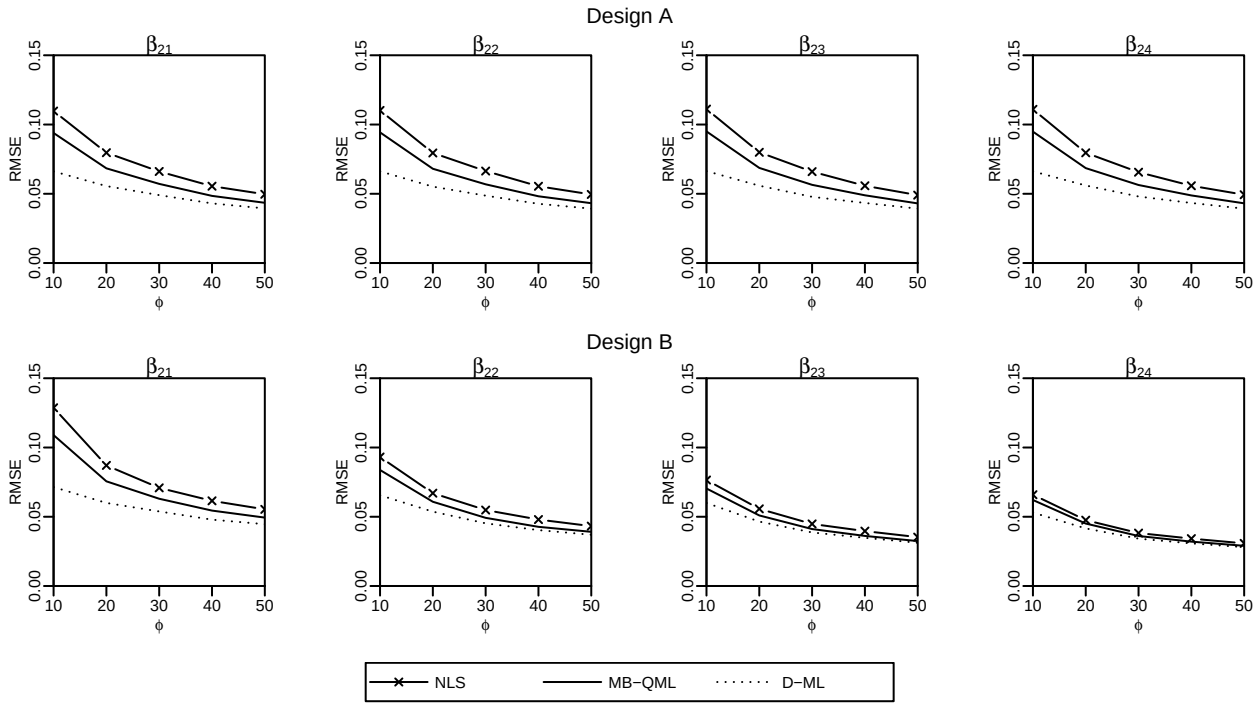

Figure 2: RMSE comparison of alternative estimators for multivariate fractional regression models
(Multinomial−distributed response variable; N = 250)

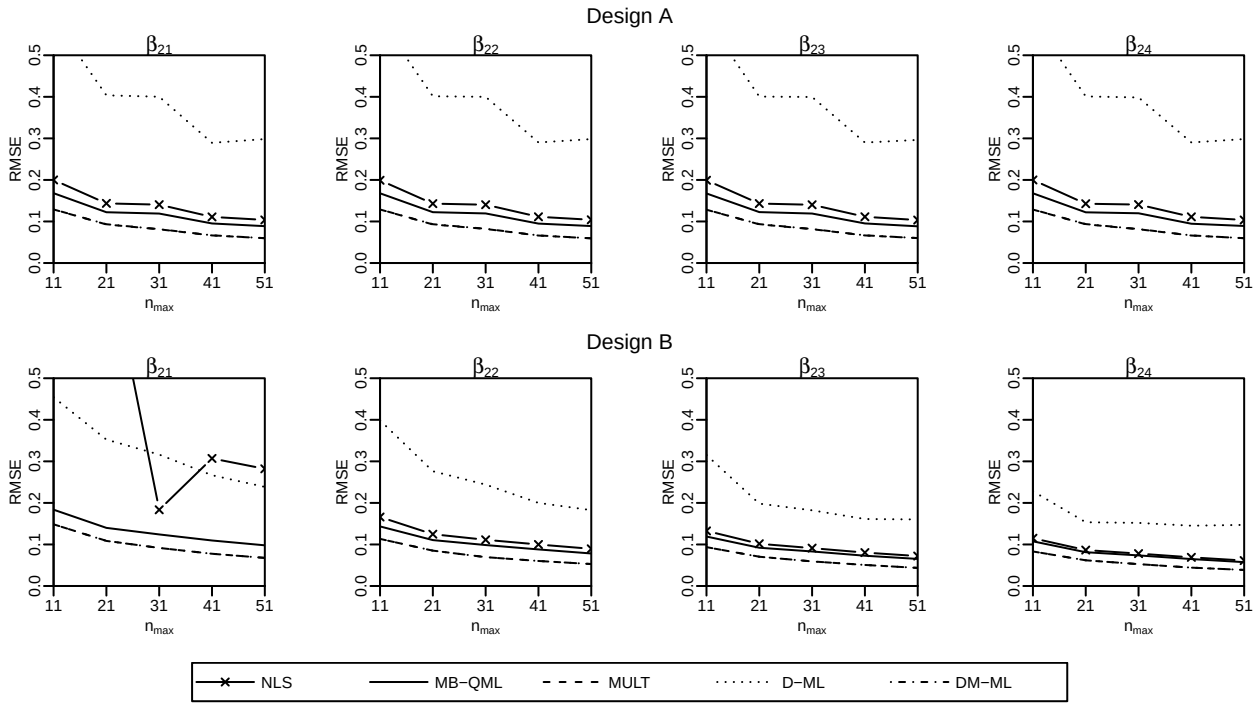

Figure 3: RMSE comparison of alternative estimators for multivariate fractional regression models
(Dirichlet−Multinomial−distributed response variable; N = 250)

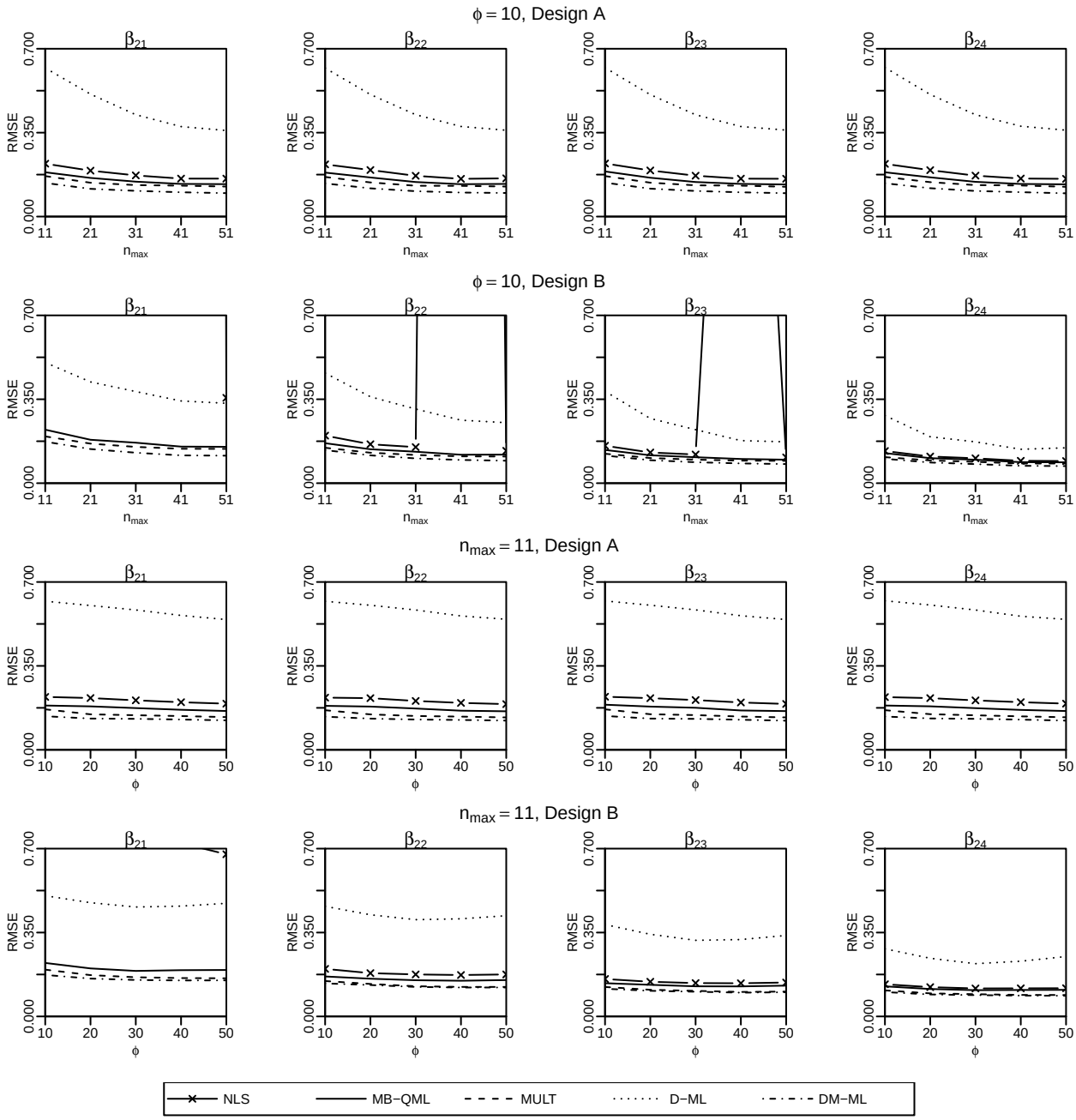

Figure 4: RMSE comparison of alternative estimators for multivariate fractional regression models
(Different sample sizes; Design B)

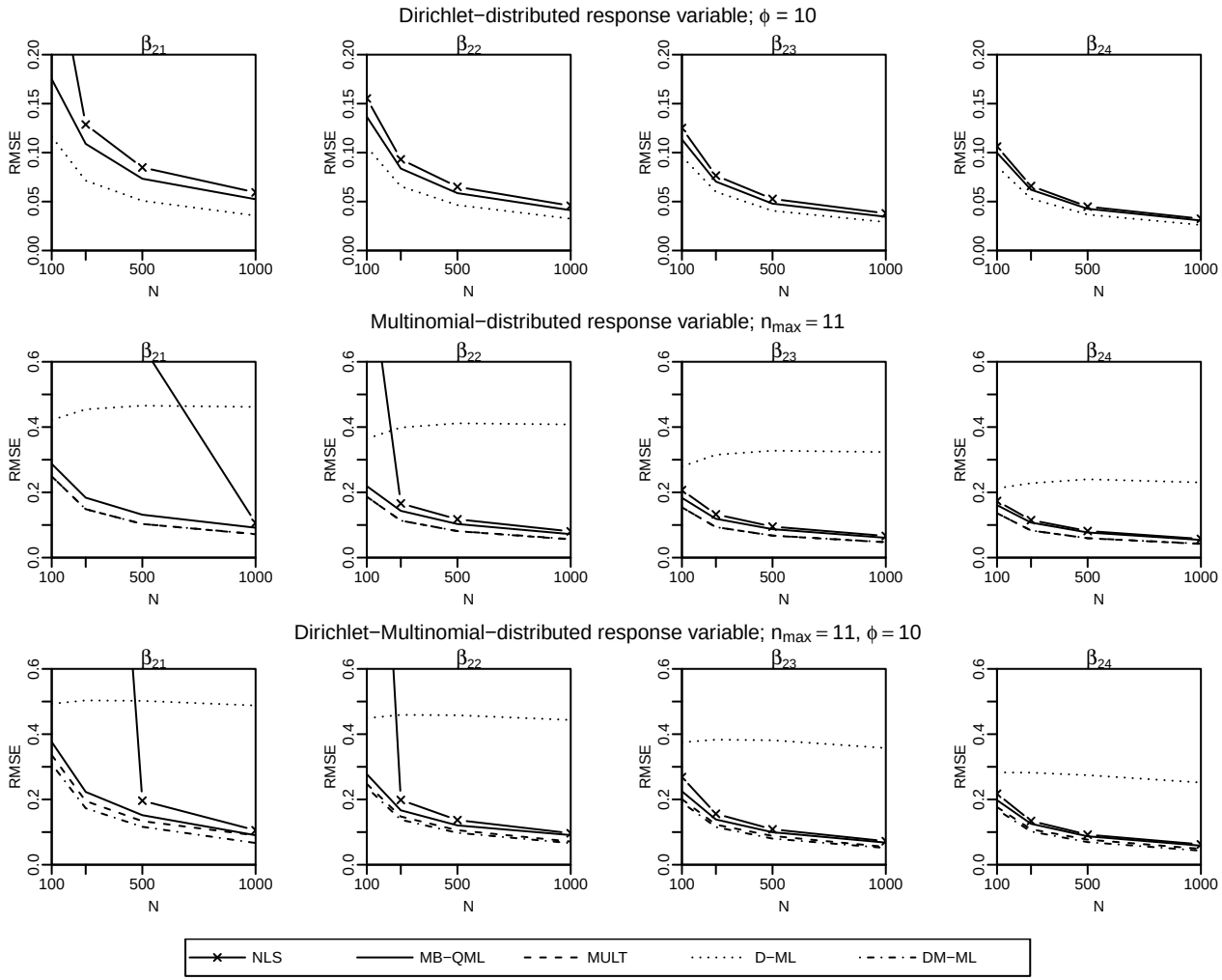

Figure 5: Empirical size (Dirichlet−distributed response variable)

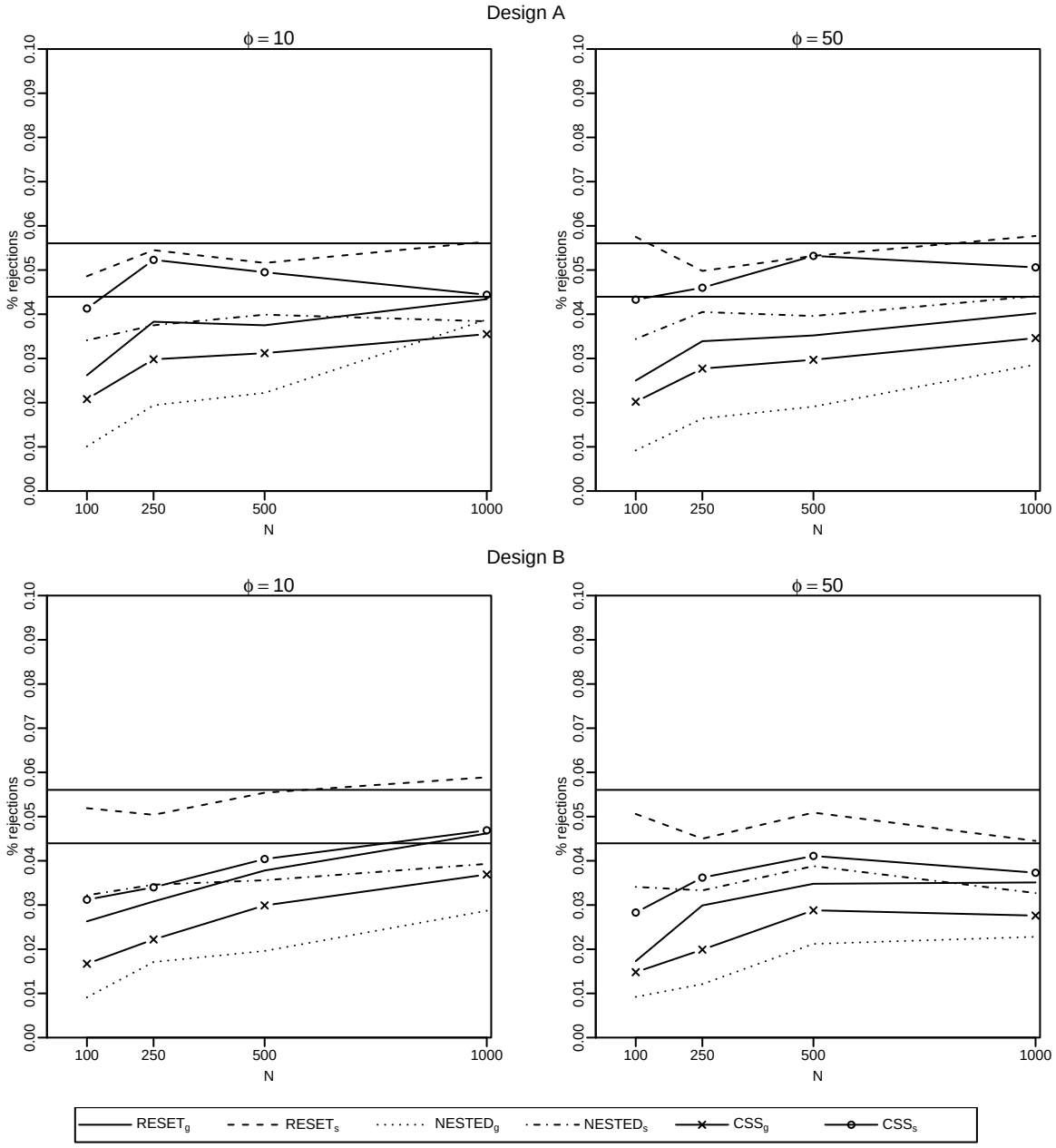

Figure 6: Empirical power (Dirichlet−distributed response variable; N = 250)