Regression Analysis of Multivariate Fractional Data*

José M.R. Murteira
Faculdade de Economia, Universidade de Coimbra, and CEMAPRE

Esmeralda A. Ramalho and Joaquim J.S. Ramalho
Departamento de Economia and CEFAGE-UE, Universidade de Évora

This version: June, 2011

Abstract

The present article discusses alternative regression models and estimation methods for dealing with multivariate fractional response variables. Both conditional mean models, estimable by NLS and quasi-maximum likelihood (QML), and fully parametric models (Dirichlet and Dirichlet-multinomial), estimable by ML, 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. All the proposed types of tests of the conditional mean can be performed through artificial OLS regressions, with test statistics evaluated at QML or NLS estimates. The paper also includes an extensive Monte Carlo study evaluating the finite sample properties of most of the estimators and tests considered.

JEL classification code: C35, C16.

Key Words: Multivariate fractional data; Quasi-maximum likelihood estimator; Dirichlet regression; Dirichlet-multinomial mixture; Specification tests.

* 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, $y \equiv (y_1, y_2, \ldots, y_M)'$, 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, usually, $M > 2$ 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- or binomial-based likelihood and a logit conditional mean function. In a recent paper, Ramalho, Ramalho and Murteira (2011) survey the main alternative regression models and estimation methods that are available for dealing with 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 $y$, given a set of explanatory variables, $E(y|X)$. One seminal methodological contribution to this goal is provided by Woodland (1979), who presents maximum likelihood (ML) estimation of systems of share equations on the basis of the Dirichlet distribution, a well known multivariate generalization of the beta distribution (see Kotz, Balakrishnan and Johnson, 2000, ch. 49). The so-called Dirichlet regression model, which can be regarded as a multivariate generalization of the beta regression model (see, e.g., Ferrari and Cribari-Neto, 2004), is presented by Woodland (1979) as a tractable and theoretically sound specification, taking due account
of the bounded, unit-sum nature of the response variables. In Economics, this model was also used by, *e.g.*, Chotikapanich and Griffiths (2002), who adopt it to approximate the distribution of proportions of income classes and subsequently estimate a Lorenz curve through ML.

Nevertheless, in spite of its worth, ML estimation based on the Dirichlet model is not robust to distributional misspecification. Moreover, like the beta distribution, 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. 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).

In face of these difficulties, one alternative route to the principal goal of estimating the conditional mean of \( y \) is provided by QML on the basis of the multivariate Bernoulli (MB) probability function (p.f.) or, in some cases, the multinomial p.f..\(^1\) As is well known, both p.f.’s belong to the linear exponential family (LEF) of distributions, so, with correct specification of the conditional mean of \( y \), QML yields consistent estimators of the conditional mean parameters. Besides being more robust than the Dirichlet regression model, QML is easier to use and can accommodate boundary observations. The specification of the elements of \( E(y|X) \) must comply with the two basic characteristics of proportions: they are bounded between 0 and 1 and they add up to one. Therefore, clearly, the specifications that are commonly used to model choice probabilities can be used in this context as well. In particular, the multinomial logit model proves a very useful specification for \( E(y|X) \), namely because it is considerably easier to estimate than the multinomial probit and other alternative models (even when \( M \) is not very large). For some examples of applications of the MB-QML method based on the multinomial logit model, see Sivakumar and Bhat (2002), Ye and Pendyala (2005) and Mullahy and Robert (2010), who model, respectively, commodity flows, transportation time and household time allocation.

---

\(^1\)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. Nevertheless, when there is no such danger, well established econometric terms are also used in the ensuing text.
The present article discusses alternative regression models and estimation methods for dealing with multivariate fractional response variables. The paper considers both conditional mean models, estimated by QML, and fully parametric models, estimated by ML. 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 $E(y|X)$ to be used in all models, irrespective of the specific functional form adopted for it. Therefore, although only a multinomial logit specification is considered for $E(y|X)$ throughout the paper, the extension to other conditional mean specifications is immediate. The suggested parameterization has also the advantage of facilitating a ready evaluation of the covariates’ relationships to $E(y|X)$. In addition to the Dirichlet regression model, other fully parametric regression models are discussed, namely multinomial-based specifications, which can be a useful modelling tool when the response variables are obtained as integers ratios and exhibit boundary values with nontrivial probability.

The article also discusses at some length the specification analysis of multivariate fractional models. This is a sensitive issue that, to the best of the present authors’ knowledge, has not merited much attention in the received literature on multivariate fractional regression. Besides a brief overview of specification tests that are available for fully parametric models, the paper also proposes several tests to assess the validity of first moment assumptions. The latter tests are extensions of their counterparts for either the univariate fractional case or the discrete case. All the proposed types of tests of the conditional mean – RESET-type test, goodness-of-functional form (GOFF) test and a test for the multinomial logit against the nested logit – can be obtained by making use of the robust, regression-based testing procedure proposed by Wooldridge (1991), adequately performed upon QML or systems nonlinear least squares (NLS) estimation.

The remainder of the paper is organized as follows. Section 2 describes the notation and critically reviews some common approaches for estimating share equations. Section 3 discusses alternative regression models and estimation methods that are available for use with multivariate fractional response variables, including specific models for the case where the fractional response variables are obtained as ratios of observable integers. 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 Framework

Consider a random sample of \( i = 1, \ldots, N \) individuals, supposed to be available for estimation of the parameters of interest. Let \( \mathbf{y} \equiv (y_1, \ldots, y_M)' \) denote the \( M \)-vector of fractional dependent variables, or shares, confined, by definition, to the unit \((M - 1)\)-simplex,

\[
S^{M-1} \equiv \{ \mathbf{y} \in \mathcal{R}^M : \sum_{m=1}^{M} y_m = 1, y_m \geq 0, \forall m \}.
\]  

(1) This type of data are known in the statistical literature as ‘compositional data’ (Aitchison, 1982).

For many years, the most popular econometric specifications of systems of share equations have not taken into account the intrinsic characteristics of fractional responses. Typically, each observed share \( y_m \) was decomposed into a deterministic component, \( D_m (\mathbf{X}; \mathbf{\beta}) \), and a stochastic disturbance term, \( u_m \),

\[
y_m = D_m (\mathbf{X}; \mathbf{\beta}) + u_m, \quad m = 1, \ldots, M,
\]  

(2) where \( \mathbf{X} \) denotes the matrix of observations on exogenous explanatory variables and \( \mathbf{\beta} \) denotes the column-vector of parameters to be estimated. Then, usually: (i) a multivariate normal distribution was assumed for \( u_m \); (ii) in order to deal with the singularity of the share equation system, one equation \((M)\) was deleted from the system and the corresponding predicted share was calculated as

\[
D_M (\mathbf{X}; \hat{\mathbf{\beta}}) = 1 - \sum_{m=1}^{M-1} D_M (\mathbf{X}; \hat{\mathbf{\beta}});
\]  

(3) and (iii) the restrictions observed on \( y_m \) were not fully taken into account in the specification of \( D_m (\mathbf{X}; \mathbf{\beta}) \). 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 in the literature, most of which, however, are also not entirely satisfactory. Hermalin and Wallace (1994) suggested using a probit specification for the deterministic component of (2), \( D_m (\mathbf{X}; \mathbf{\beta}) = \Phi (\mathbf{X}_m; \mathbf{\beta}_m) \), where \( \Phi (\cdot) \) denotes the normal cumulative distribution function (with covariates and parameters that may be alternative-specific), and estimating the resulting nonlinear seemingly unrelated regression model by systems NLS. Pu, Lan, Chou and Lan (2008) made a similar suggestion but based on a logistic specification and
QML estimation. However, both methods are also insufficient to ensure that the predicted shares fall into 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 will not sum up to unity).

Fry, Fry and McLaren (1996) propose using a method suggested by Aitchison (1982), which is based on the application of a one-to-one transformation from the unit simplex $S^{M-1}$ to the real set $R^{M-1}$, namely the additive logratio transformation defined by $z_m = \log(y_m/y_M), \ m = 1, ..., M - 1$.\footnote{This method is widely used in fields like geology, pedology, geochemistry and biology; see the recent survey by Aitchison and Egozcue (2005). Its application it is also common in political science, where some variants have been proposed; e.g. Katz and King (1999) assume that $v_m$ in (4) has a Student’s $t$ distribution.} This yields the following general functional form for estimation:

$$z_m = \log \left[ D_m (X; \beta) / D_M (X; \beta) \right] + v_m, \ m = 1, ..., M - 1, \quad \text{(4)}$$

where $v_m$ is assumed to follow a multivariate normal distribution. The inverse transformation from $R^{M-1}$ to $S^{M-1}$ is the additive logistic transformation, which implies a multinomial logit specification for the $y_m$’s. While this method effectively restricts the predicted shares to the unit simplex, it presents some disadvantages such as not being well defined for the boundary value of 0 and thus requiring ad hoc adjustments if that value is observed in the sample (e.g. replacing the resultant infinite values of $z_m$ by an arbitrarily chosen large number). In contrast, direct NLS or QML estimation of multinomial logit models, which considers the same specification for the $y_m$’s, can easily handle boundary observations.

A very common approach for dealing with fractional responses has been the use of tobit models 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. A better approach would be to assume that the shares follow a multivariate normal distribution truncated at the boundaries of the $(M - 1)$-unit simplex. However, to the best of the present authors’ knowledge, this method has been adopted only by few researchers (e.g. Poterba and Samwick, 2002; Klawitter, 2008), maybe because it also presents a number of difficulties. First, as some authors argue (e.g. Maddala, 1991), the tobit model is conceptually appropriate to describe censored data in the inter-
val $[0, 1]$ but its application to data \textit{defined} only within that interval is not easy to justify: observations at the boundaries of the support of a fractional variable are a natural consequence of individual choices and not of any type of censoring. Second, the tobit model is very stringent in terms of assumptions, requiring normality and homoskedasticity of the dependent variables prior to censoring. Third, the suggested two-limit tobit can only be applied when the sample comprises observations of the response variables in both limits of the unit interval, often not the case for fractional data. Finally, estimation of the model is frequently fraught with computational complexity that may lead researchers to adopt questionable assumptions. For instance, Poterba and Samwick (2002) impose non-correlated disturbances across latent variables equations underlying the shares of financial assets, so as 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 joint distributional assumptions for the 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 $E(y|X)$ which enforce the conceptual requirement that, as for $y$, its elements belong to the unit simplex.

In what follows, let $E(y|X) = G(X; \beta_0) \equiv [G_1(X; \beta_0), ..., G_M(X; \beta_0)]'$, the column $M$-vector of the conditional mean functions of $y$, with $\beta_0$ denoting the true value of $\beta$. To simplify the notation, $E(y|X)$ and its components, $E(y_m|X)$, $m = 1, ..., M$, will often be referred to without explicit mention to its arguments: $G \equiv G(X; \beta)$ and $G_m \equiv G_m(X; \beta)$. When intended, the individual observation of these quantities may be denoted, respectively, as $G_i \equiv G(X; \beta) = E(y_i|X)$ and $G_{im} \equiv G_m(X; \beta) = E(y_{im}|X_i)$. Given the definition of the elements of $y$, their conditional means are also subject to the constraints $G_m \geq 0, \forall m$, and $\sum_{m=1}^{M} G_m = 1$. Frequently, $G_m$ is specified as a function of indices of covariates, that is, $G_m = G_m(X; \beta)$, with the form of $X$ depending on the specific model that is used. For instance, when all covariates vary with alternatives and the parameters are invariant across alternatives, $X \equiv (x_1, \ldots, x_M)'$, with $x_m'$ conformable to $\beta$; if the covariates are all alternative-invariant and the parameters are alternative-specific, then one can adopt a block-diagonal form, $X \equiv diag(x', m = 1, ..., M)$, and
$$\beta = (\beta_1', \ldots, \beta_M')',$$ where \(x'\) is conformable to \(\beta_m\) so \(X\beta = [x'\beta_1; \ldots; x'\beta_M]'\).³

### 3 Regression Models and Estimation Methods

Two main approaches for modelling multivariate fractional data that take into account the characteristics of the response variables are considered here. The first only requires correct specification of the conditional mean of \(y\), given covariates, whereas the second alternative is a fully parametric approach based on the assumption of a particular conditional distribution for \(y\), whose mean is 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 may be interpreted as ratios of integers.⁴

#### 3.1 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 require the correct specification of \(E(y|X)\). This section presents the formulation adopted throughout the paper for the conditional mean of \(y\) and discusses two alternative estimation methods for the resultant model.

In the univariate case, the specifications used for modelling binary response variables are also typically employed to describe the conditional mean of fractional responses; see Papke and Wooldridge (1996). Analogously, in the multivariate case, the specifications that are commonly used to model the probability of an individual choosing the discrete outcome \(m\) may also be employed to describe \(E(y|X)\) in the fractional context. Indeed, any of those specifications satisfy the bounded, unit-sum nature of both the probabilities of choosing between \(M\) mutually exclusive alternatives (discrete case) and the \(M\) conditional means \(y_m\) (fractional case).

Here, special attention is devoted to the multinomial logit specification, which can be

---

³As is well-known, the unit-sum identity can imply identification conditions on some elements of \(\beta\), namely normalization of coefficients associated with alternative-invariant covariates. In what follows, \(\beta\) denotes the vector of identified parameters.

⁴For brevity sake, multivariate two-part and similar models are not included in the text.
formally expressed, in general, as

$$G_m = \exp \left( \frac{x_m' \beta_m}{\sum_{i=1}^{M} \exp (x_i' \beta_i)} \right), \quad m = 1, \ldots, M. \quad (5)$$

It is noted that, although no other specifications for $G_m$ are explicitly considered in the paper, all the regression models discussed below also accommodate any other $G_m$ specifications that comply with the fractional nature of the response variables. Meanwhile, well known and widely used special cases of the multinomial logit model are the so-called “conditional logit” (CL) model, with alternative-invariant parameters, $\beta_m = \beta, \forall m$, and the conventional “multinomial logit” (MNL), with alternative-invariant covariates,$^5$

$$G_m = \frac{\exp (x_m' \beta_m)}{1 + \sum_{i=2}^{M} \exp (x_i' \beta_i)}, \quad m = 1, \ldots, M, \quad \beta_1 = 0. \quad (6)$$

Although the 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, in addition to its obvious mathematical appropriateness, the multinomial logit model also conforms 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. A similar proof was produced by Dubin (2007) for the nested logit model.

The parameters of the fractional multinomial logit model defined by (5) may be estimated by, among other methods, NLS or QML. Indeed, note that this model can be expressed as a system of nonlinear regression equations of the form

$$y_m = G_m + u_m, \quad m = 1, \ldots, M; \quad \sum_{m=1}^{M} u_m = 1, \quad (7)$$

where $u_m$ denotes an error term. Under random sampling and standard assumptions (namely correct and twice continuously differentiable specification of $G$ – the case for a correct logit model), the systems NLS estimator of $\beta$, minimizing the sum of squared residuals, $\sum_{i=1}^{N} \hat{u}_i' \hat{u}_i, \quad \hat{u}_i \equiv y_i - G \left( X_i; \hat{\beta} \right) = y_i - \hat{G}_i$, is consistent and asymptotically normal. Formally,

$$\sqrt{N} \left( \hat{\beta}_{NLS} - \beta_0 \right) \overset{d}{\rightarrow} \mathcal{N} \left( 0, A_0^{-1} B_0 A_0^{-1} \right), \quad (8)$$

$^5$In line with traditional terminology, the term “multinomial” is used both for the general multivariate logit formulation and the particular model with alternative-invariant covariates. To avoid ambiguity, the latter will hereafter be designated “MNL”.

9
where
\[ A_0 \equiv E \left[ \nabla_{\beta} G_i (\beta_0)' \nabla_{\beta} G_i (\beta_0) \right], \quad (9) \]
\[ B_0 \equiv E \left[ \nabla_{\beta} G_i (\beta_0)' u'_i \nabla_{\beta} G_i (\beta_0) \right], \quad (10) \]
\[ u_i \equiv y_i - G_i (\beta_0) \] 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 \( \beta \) at \( \hat{\beta}_{NLS} \).

Potentially more efficient estimators of \( \beta \) may be obtained by choosing some convenient weighting matrix or assuming some reasonable model for the conditional second moments, \( V (y|X) \), and estimating the model by systems weighted NLS (WNLS). An alternative approach (formally amounting to a particular class of WNLS estimators) is provided by QML, next described.

As is well known, the QML estimator is obtained from maximization of an 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, Kotz and Balakrishnan, 1997, ch. 36). This p.f. is appropriate for the case where there are \( M \) alternatives and each individual chooses only one alternative. Let the \( m \)-th component of \( b \equiv (b_1, ..., b_M)' \) 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
\[ f_b (b) = \prod_{m=1}^{M} \pi_m^{b_m}, \quad \sum_{m=1}^{M} \pi_m = 1. \quad (11) \]
In a regression context, the parameters \( \pi_m \) can be replaced by conditional expectations given covariates.

With multivariate fractional variables, substituting \( E (y_m|X) \) for \( \pi_m \), the individual contribution to the likelihood function can be expressed as
\[ L^{MB}_i (\beta) = \prod_{m=1}^{M} G^{y_{im}}_{im}. \quad (12) \]
The individual contribution to the log-likelihood then results as
\[ LL^{MB}_i (\beta) \equiv \log L^{MB}_i (\beta) = \sum_{m=1}^{M} y_{im} \log G_{im} = \sum_{m=1}^{M-1} y_{im} \log \frac{G_{im}}{G_{iM}} + \log G_{iM}, \quad (13) \]
where \( G_{iM} = 1 - \sum_{m=1}^{M-1} G_{im} \). The \( \beta \) QML estimator based on the MB model, maximizing \( LL (\beta) \equiv \sum_{i=1}^{N} LL^{MB}_i (\beta) \), is consistent and asymptotically normal regardless of the true
conditional distribution of $y$, provided that $G$ is correctly specified (Gouriéroux, Monfort and Trognon, 1984). Formally,

$$
\sqrt{N} \left( \hat{\beta}_{QML} - \beta_0 \right) \xrightarrow{d} \mathcal{N} \left( 0, A_0^{-1} B_0 A_0^{-1} \right),
$$

where

$$
A_0 \equiv E \left[ -\nabla_{\beta^L} LL_i^M (\beta) \right]_{\beta = \beta_0} \quad (15)
$$

and

$$
B_0 \equiv E \left[ \nabla_{\beta} LL_i^M (\beta) \nabla_{\beta^L} LL_i^M (\beta) \right]_{\beta = \beta_0} \quad (16)
$$

Consistent estimators for $A_0$ and $B_0$ are obtained in the usual manner, replacing population expectations by sample averages, with $\beta = \hat{\beta}_{QML}$. QML estimation of fractional MNL models have been considered by Sivakumar and Bhat (2002), Ye and Pendyala (2005), Mullahy (2010) and Mullahy and Robert (2010).

### 3.2 The Dirichlet Regression Model

Even when interest is confined to the parameters of the conditional mean, it may be possible to improve on the efficiency of the QML method by adopting some reasonable conditional second moment assumptions (Gouriéroux, Monfort and Trognon, 1984). As previously mentioned, this amounts to systems WNLS, potentially more efficient than QML (itself a particular case of WNLS). However, as Mullahy (2010) notes, the share nature of the data rarely provides significant guidance about second moments assumptions (beyond the fact that they are bounded). Thus, alternatively, one may resort to ML estimation upon full specification of the conditional density of $y$ given $X$. With correct $f(y|X)$ specification, ML estimators of $\beta$ are obtained as $\hat{\beta}_{ML} \equiv \arg \max LL (\beta)$, with

$$
\sqrt{N} \left( \hat{\beta}_{ML} - \beta_0 \right) \xrightarrow{d} \mathcal{N} \left( 0, A_0^{-1} \right)
$$

and $A_0$ defined in (15).

Several statistical distributions are suited to data confined within the unit simplex. The most popular choice for the joint law of $y$ is the Dirichlet distribution, which can be obtained as a multivariate generalization of the beta distribution (see Kotz, Balakrishnan and Johnson, 2000, ch. 49). The joint density function of the Dirichlet distribution can be expressed as

$$
f_y^D (y; \gamma) = \frac{\Gamma (\gamma_0)}{\prod_{m=1}^{M} \Gamma (\gamma_m)} \prod_{m=1}^{M} y_m^{\gamma_m-1} \equiv Dirichlet (\gamma),
$$

$$
y_m : y_m > 0, \quad \sum_{m=1}^{M} y_m = 1, \quad m = 1, ..., M,
$$

where $\gamma_0$ is the parameter that controls the overall scale of the distribution.
where \( \gamma \equiv (\gamma_1, ..., \gamma_M)^t \) denotes a vector of positive parameters and \( \gamma_0 \equiv \sum_{m=1}^M \gamma_m \). Under this parameterization,

\[
E (y_m) = \frac{\gamma_m}{\gamma_0}
\]  

(19)

and the elements of the covariance matrix of \( y \) can be expressed as

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

(20)

where \( \delta_{lm} \) 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 \( y \) is absolutely continuous with density having the same form as above. Also, a desirable property for applications is that permutation of \( y \) components simply leads to a Dirichlet by permuting the corresponding parameters. Moreover, aggregation of some elements of \( y \) in a smaller vector also leads to a Dirichlet distribution with the same type of aggregation in the vector of parameters. Furthermore, each component \( y_m \) is distributed according to a \( Beta(\gamma_m, \gamma_0 - \gamma_m) \). Finally, if all \( \gamma_m \) parameters are proportionately large, then the Dirichlet can be approximated by a multivariate normal density.

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 \( \gamma_m \), then denoting some specified function of \( X \) and parameters \( \beta \), \( \gamma_m \equiv \gamma_m (X; \beta) \), \( m = 1, ..., M \). However, estimating the covariates’ relationships to the \( \gamma \) parameters may not be of much interest, so a different parameterization may prove advantageous with respect to the interpretation of the parameters of a Dirichlet regression model. This paper proposes the reparameterization \( \gamma_m \equiv \phi G_m \), \( m = 1, ..., M \), with \( \phi > 0 \) and \( G_m \) being given by (5). Thus, one obtains \( \gamma_0 \equiv \phi \sum_{m=1}^M G_m = \phi \), and the expression for the Dirichlet density becomes

\[
f^D_{y|m|X} (y; \phi, \beta|X) = \frac{\Gamma (\phi)}{\prod_{m=1}^M \Gamma (\phi G_m)} \prod_{m=1}^M y_m^{\phi G_m - 1}.
\]  

(21)

Consequently, from (19) and (20), it follows that \( E (y_m|X) = G_m \) and

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

(22)
The parameter $\phi$ can be interpreted as a precision measure in the sense that, for fixed $G$, the larger the value of $\phi$, the smaller the diagonal elements of the covariance matrix $COV(y_i, y_m)$ – note that $y$ degenerates at $G$ if $\phi \to \infty$.

So far, to the best of the present authors’ knowledge, this type of mean-dispersion parameterization has only been introduced in a univariate context (for the beta regression model), proposed independently by Paolino (2001) and Ferrari and Cribari-Neto (2004). Note that both Woodland (1979) and Chotikapanich and Griffiths (2002) have adopted similar mean-dispersion parameterizations for their Dirichlet models, which, however, are specific to the particular models considered by them. For example, Woodland (1979) set

$$\gamma_m = \phi z_m(X; \beta),$$

where $z_m(\cdot)$ denotes specified positive index functions of the covariates, from which it follows that

$$E(y_m|X) = \frac{z_m(X; \beta)}{\sum_{i=1}^{M} z_i(X; \beta)}, \quad m = 1, ..., M. \quad (23)$$

While the parameterization proposed in this paper is valid irrespective of the specification assumed for $G_m$, the one suggested by Woodland (1979) requires the adoption of the so-called universal logit specification for $G_m$, not being applicable in cases where, for example, nested logit or multinomial probit specifications are used for representing $E(y_m|X)$.

The proposed Dirichlet regression model assumes that only the $G_m$’s are functions of covariates and parameters, treating $\phi$ as a nuisance parameter. Alternatively, considering that a researcher may be interested in analyzing whether a variable contributes to the variances and covariances of $y$ beyond its effect on the means, one may also specify $\phi$ as a function of covariates (possibly distinct from $X$). Again, this is analogous to what some authors chose to do in the univariate case for the beta regression model; see Paolino (2001) and Smithson and Verkuilen (2006).

The literature suggests several generalizations of the Dirichlet model, aimed, namely, at overcoming some of its most stringent assumptions. For example, as is evident from (22), all covariances are non-positive. The generalizations proposed include, among others, Barndorff-Nielsen and Jorgensen’s (1991) parametric models, the scaled Dirichlet (Aitchison, 2003, pp. 305-306), Connor and Mosimann’s distribution (Connor and Mosimann, 1969) and the Liouville distribution (Rayens and Srinivasan, 1994). A detailed

\[\text{Note that Woodland (1979) did not impose the constraint } z_m(\cdot) > 0 \text{ that is necessary to enforce the required positivity of the Dirichlet parameters. In fact, he used a linear specification for } z_m(\cdot) \text{ in his empirical analysis.}\]
consideration of these models, however, is beyond the scope of the present text.

### 3.3 Regression Models for Proportions Obtained as Ratios of Observable Integers

In some applications, the response variables may be interpreted as ratios of integers, i.e. the elements of $\mathbf{y}$ 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 that choose alternative $m$ ($n_m$) are known, one can also resort to models that make explicit use of this extra information (not merely of $\mathbf{y}$). The alternative models now described may or may not produce more efficient estimators than the approaches previously discussed (usually still valid), a fact that depends on the actual covariance structure of the data generating process (DGP). In any case, being defined for both boundary and interior values of the unit interval, they can prove advantageous alternatives over the regression models presented so far.

#### 3.3.1 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 $\pi_m$ denote the probability that an individual selects alternative $m$. Then, $(n_1, ..., n_M) = n \times \mathbf{y}$ follows a multinomial p.f. with parameters $n$ and $\pi = (\pi_1, ..., \pi_M)$. Equivalently,

$$f^M_y(\mathbf{y}; n, \pi) = \frac{n!}{\prod_{m=1}^{M} (n y_m)!} \prod_{m=1}^{M} \pi_m^{n y_m},$$

where $\pi_M = 1 - \sum_{m=1}^{M-1} \pi_m$. Under this parametrization, $E(y_m) = \pi_m$,

$$\text{COV}(y_l; y_m) = \frac{\pi_l (\delta_{lm} - \pi_m)}{n}, \quad l, m = 1, ..., M,$$

and the individual contribution to the log-likelihood is given by

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

$$= \log \left[ \frac{n_i!}{\prod_{m=1}^{M} (n_i y_{im})!} \right] + n_i \left( \sum_{m=1}^{M-1} y_{im} \log \frac{\pi_{im}}{\pi_{iM}} + \log \pi_{iM} \right).$$

14
A regression model can be accommodated by considering covariates in $\pi_m$. As before, let $\pi_m = G_m$, which leads to a conditional p.f. $f_{y|X} (y; n, \beta|X)$ and a conditional covariance matrix of $y|X$ with typical element

$$\text{COV}^M (y_i, y_m|X) = \frac{G_l (\delta_{lm} - G_m)}{n}. \tag{27}$$

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

$$\log L^M_i (\beta) = \log \left[ \frac{n_i!}{\prod_{m=1}^M (n_i y_{im})!} \right] + n_i \log L^{MB}_i (\beta), \tag{28}$$

where $\log L^{MB}_i (\beta)$ is the individual contribution to the MB log-likelihood defined in (13).

The multinomial p.f. is a member of the LEF, so correct $G_m$ specification suffices for consistency of the resulting QML estimator of $\beta$. In addition, if the data are indeed multinomial distributed, then ML estimation and inference can be performed.

### 3.3.2 The Dirichlet-Multinomial Regression Model

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

$$f_{\pi|X} (\pi; \phi, \beta|X) = \frac{\Gamma (\phi)}{\prod_{m=1}^M \Gamma (\phi G_m)} \prod_{m=1}^M \pi_m^\phi G_{m-1}, \tag{29}$$

one can formally write the Dirichlet-multinomial conditional p.f. as

$$f_{y|X} (y; n, \phi, \beta|X) = \frac{n! \Gamma (\phi)}{\Gamma (\phi + n)} \prod_{m=1}^M \frac{\Gamma (\phi G_m + n y_m)}{\Gamma (\phi G_m) (n y_m)!} \tag{30}$$

Several remarks about this expression seem appropriate. First, for $M = 2$ the model reduces to the beta-binomial p.f. (see Johnson, Kemp, and Kotz, 2005, ch. 6). Examples of the use of the beta-binomial model in a regression context can be found, among others, in Heckman and Willis (1977) and Santos Silva and Murteira (2009). Second, the DM mixture has beta-binomial univariate marginals with parameters such that

$$E (y_m|X) = E (\pi_m|X) = G_m \tag{31}$$

and

$$\text{COV}^{DM} (y_i, y_m|X) = \frac{G_l (\delta_{lm} - G_m) \phi + n}{n} \frac{\phi}{\phi + 1}. \tag{32}$$
(see Johnson, Kotz and Balakrishnan, 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 \( COV^{DM} (y_t, y_m | X) = a \cdot COV^M (y_t, y_m | X) \), with \( a = (\phi + n) / (\phi + 1) > 1 \). Also, expression (22) (with \( \pi_m \) replacing \( y_t \) and \( y_m \)) implies that \( \lim_{\phi \to \infty} V (\pi_m | X) = 0 \), so the conditional distribution of \( \pi_m \) becomes degenerate at \( G_m \) and the DM model collapses to
the multinomial, as the parameter \( \phi \) grows infinitely large. In any case, it should be noted
that the DM is not a LEF member, so, unlike in the multinomial case, only ML estimation
is possible, which inevitably requires the DM model to be a fully correct specification of
the conditional p.f. of \( n \times y \) given covariates.

One example of the use of the DM model in econometrics (for the discrete case and
with a different parameterization) is provided by Guimarães and Lindrooth (2007), with
a data set on patient choice of hospitals. These authors obtain the DM model from
marginalization of unobserved log-gamma group specific heterogeneity, in the context of
multinomial logit specification of conditional choice probabilities. More recently, Mullahy
(2010) uses the DM specification (also based on a different parameterization) to estimate
models of financial asset portfolio shares. However, in the case considered by Mullahy
(2010), the response variables are not generated by a ratio of integers, so some arbitrary
assumptions on the value of \( n_i \) have to be made. Given that in such a case the DM
model cannot coincide with the DGP and that the DM p.f. is not a LEF member, the
appropriateness of the application of DM models in that context seems questionable.

3.4 Marginal Effects

As with any other nonlinear regression model, the parameters of multivariate fractional
models are not readily interpretable since they intervene jointly to parameterize the mar-
ginal effects of the explanatory variables on \( E(y|X) \). Instead, the calculation of marginal
effects (or elasticities) is typically much more useful in applications. For a continuous
alternative-varying explanatory variable \( x_{mj} \) (CL model), the marginal effect on \( E(y_r|X) \)
due to a unitary change in \( x_{mj} \) is given by the following partial derivative:

\[
\nabla_{x_{mj}} G_r = \frac{1}{\sum_{l=1}^{M} \exp (x_l' \beta)} \left[ \delta_{rm} \beta_j \exp (x_r' \beta) - \frac{\beta_j \exp (x_r' \beta) \exp (x_m' \beta)}{\sum_{l=1}^{M} \exp (x_l' \beta)} \right] = \beta_j G_m (\delta_{rm} - G_r).
\]

(33)
In the case of a continuous alternative-invariant explanatory variable $x_j$ (MNL model), the marginal effect on $E(y_r|x)$ due to a unitary change in $x_j$ is given by:

$$\nabla_{x_j} G_r = G_r \left( \beta_{rj} - \sum_{l=1}^M G_l \beta_{lj} \right).$$ (34)

Finally, in the case of a discrete explanatory variable $x_{mj}$ or $x_j$, the marginal effect may be calculated as the difference between the conditional mean of $y_r$ when the regressor is increased by one unit and the conditional mean value before the increase. In all cases: (i) estimates of marginal effects are obtained by replacing $\beta$ with NLS, QML or ML estimates; (ii) asymptotic standard errors for the estimated derivatives can be obtained through the delta method; and (iii) the sum of all the marginal effects associated with a given explanatory variable over the $M$ shares is zero. For the MNL model, as in the case of multinomial discrete choice analysis, the sign of the marginal effect is not necessarily given by the sign of $\beta_{rj}$—see equation (34).

As in the discrete case, log-odds ratios are formally less complicated than the expressions for the partial derivatives. In the present context, taking conditional means instead of choice probabilities, one can write, respectively,

Model CL: \[ \log \left( \frac{G_m}{G_i} \right) = (x'_m - x'_i) \beta, \]

Model MNL: \[ \log \left( \frac{G_m}{G_i} \right) = x' (\beta_m - \beta_i), \]

from which, with the normalization $\beta_M = 0$, $\log \left( \frac{G_m}{G_M} \right) = x' \beta_m$ in the MNL case. As the definition of the log-odds ratio shows, these ratios do not depend on alternatives other than those involved in the ratio. This is the well-known independence of irrelevant alternatives (IIA) property of the multinomial logit model, which naturally extends to the case of fractional variables. It is commonly recognized that the behavioural implications of the IIA property (hence of the multinomial logit) may be unrealistic in many situations. Therefore, its user friendliness notwithstanding, the multinomial logit should be tested in empirical work. The next section discusses a number of specification tests that are available to assess the correctness of this model.

4 Specification Testing

The alternative estimators for fractional regression models described in Section 3 are based on different assumptions. Next, several test statistics for these assumptions and, thus, for
the statistical validity of these models are discussed. As all models require the correct specification of the conditional mean of $y$, the primary focus of this section is on functional form tests, i.e. tests for assessing the correctness of the assumption $E(y|X) = G(X; \beta)$. Given the widespread use and computational tractability of the multinomial logit model, all procedures are applied to the specification testing of this model. In any case, the first two proposed tests (RESET-type test and GOFF test) can also be used for other models, such as the multinomial probit. At the end of this section, tests for assessing the distributional assumptions of the parametric Dirichlet and DM models are also discussed.

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 $\eta$ in the generalized model $E(y|X) = H(X; \beta, \eta)$. Under the null hypothesis $H_0 : \eta = 0$, $H(X; \beta, 0) = G(X; \beta)$ is an appropriate specification for $E(y|X)$. Such tests can be carried out in the usual manner, through a Lagrange multiplier (LM), Wald or likelihood ratio (LR) test. Given that the model under the alternative may be difficult to estimate, the former strategy is proposed here, by making use of Wooldridge’s (1991) robust regression-based tests, which can be implemented upon QML or NLS estimation. One alternative to the LM tests proposed here is suggested by Mullahy (2010), who bases a conditional moment test of the mean assumptions on the Hosmer-Lemeshow’s approach, commonly used in the assessment of binary logit models (see Hosmer and Lemeshow, 2000, ch. 5).

Let $y^- \equiv (y_2, \ldots, y_M)'$, $G^- \equiv (G_2, \ldots, G_M)'$ and $H^- \equiv (H_2, \ldots, H_M)'$, vectors of nonredundant fractional responses and respective conditional means, under the null (multinomial logit) and alternative hypotheses. Also, let the alternative functional form of $E(y_m|X)$ be denoted by $H_m \equiv H_m(X; \beta, \eta)$. In the MB-QML case, with the log-likelihood defined in (13), tests for the conditional mean specification can be computed as $N – SSR$, where $SSR$ denotes the sum of squared residuals from the OLS regression of the constant 1 on the $(1 \times L)$ -vector $\tilde{e}_i \tilde{R}_i$, where $\tilde{e}_i \equiv \tilde{\tilde{C}}^{-1/2}_i \hat{e}_i$, $\hat{e}_i \equiv y_i^- – \tilde{\tilde{G}}_i^-$, $\tilde{R}_i$ denotes

7If the data come from a Dirichlet or a DM data generating process, then Dirichlet or DM ML estimates, respectively, can also be used. However, in this case the resulting test is not robust to distributional misspecifications, even with correct $E(y|X)$. Given that those densities are not LEF members, $\sqrt{N}$-consistency of the corresponding ML estimator requires fully correct distributional assumptions.
the \((M - 1) \times L\)-matrix of residuals from the matrix regression of \( \tilde{W}_i \equiv \hat{C}_i^{-1/2} \cdot \nabla \eta' \hat{H}_i \) on \( \tilde{X}_i \equiv \hat{C}_i^{-1/2} \cdot \nabla \eta' \hat{H}_i \), and \( (\cdot) \) represents evaluation at the restricted QML estimator \((\hat{\beta}', 0')'\). In the MB-based QML case, \( C_i \) is expressed as

\[
C_i = \begin{bmatrix}
G_{i2}^{-1} + G_{i1}^{-1} & G_{i1}^{-1} & \cdots & G_{i1}^{-1} \\
G_{i1}^{-1} & G_{i3}^{-1} + G_{i1}^{-1} & \cdots & G_{i1}^{-1} \\
\vdots & \vdots & \ddots & \vdots \\
G_{i1}^{-1} & G_{i1}^{-1} & \cdots & G_{i,M}^{-1} + G_{i1}^{-1}
\end{bmatrix}^{-1}; \quad (35)
\]

if the model is estimated by multinomial-based QML, then the previous expression for the \( C_i \) matrix should be multiplied by the factor \( n_i^{-1} \). With NLS estimation, \( C_i = I_{M-1} - M^{-1} \iota_{M-1}' \iota_{M-1} \), with \( I_{M-1} \) the identity matrix of order \((M - 1)\) and \( \iota_{M-1} \) an \((M - 1)-\)column vector of ones.8

In the context of the multivariate logit model the following types of conditional mean tests are considered: (i) RESET-type tests, (ii) goodness-of-functional form tests and (iii) test of the multinomial logit against the nested logit specification. For each test the expression of \( \nabla \eta' \hat{H}_i \) is now obtained.

### 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 linear regression model but, as shown by Pagan and Vella (1989), it can be applied to any type of index models. Papke and Wooldridge (1996) propose the test as a general functional form diagnostic for models of univariate fractional responses. More recently, their suggestion was taken up by Loudermilk (2007), who uses a RESET-type procedure in the context of a dynamic panel data model for fractional dependent variables.

The use and performance of various generalizations of the original RESET test in the context of multivariate linear models has been investigated in several papers (Giles and Keil, 1997, Shukur and Edgerton, 2002, Mantalos and Shukur, 2007, and Alkhamisia, Khalaf and Shukur, 2008). To the best of the present authors’ knowledge, no such enquiry

---

8For \( M = 2 \), the above definitions of \( C_i \) yield, respectively, \( \hat{C}_i^{-1/2} = \left[ \hat{G}_{i2} \left( 1 - \hat{G}_{i2} \right) \right]^{-1/2} \) (QML estimation) and \( \hat{C}_i^{-1/2} = 1 \) (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., 2010, Section 4.1.1).
has yet been conducted for systems of nonlinear regressions: still, it should be noted that,
as for linear systems, use of a RESET-type procedure within the present nonlinear systems
should take into account the fact that $E(y|X)$ is a functional of several linear indices
of covariates (not just one as in the case of a binomial model). Consequently, the test
should assess the joint significance of powers and cross-products of these index functions,
as results from a multivariate Taylor-expansion of the (unspecified) alternative model,
esting the multinomial logit.

To see this, let the unspecified alternative conditional mean vector be generally denoted
as $H(X\beta) \equiv \left[ H_1(X\beta), \ldots, H_M(X\beta) \right]'$. Also, define the vector function $Q(X\beta) \equiv 
\left[ \log H_1(X\beta), \ldots, \log H_M(X\beta) \right]'$; then,

$$G_m \left[ Q(X\beta) \right] = \frac{\exp \left[ \log H_m(X\beta) \right]}{\sum_{i=1}^{M} \exp \left[ \log H_i(X\beta) \right]} = H_m(X\beta) \Leftrightarrow G \left[ Q(X\beta) \right] = H(X\beta),$$

where the second passage involves the equality $\sum_{i=1}^{M} \exp \left[ \log H_i(X\beta) \right] = \sum_{i=1}^{M} H_i(X\beta) = 1$. This means that the alternative model for $E(y|X)$ can be expressed as a multinomial
logit-type model with possibly nonlinear arguments $Q(X\beta)$. Now, a multivariate Taylor-
expansion of $Q_m(X\beta)$ around $X\beta = 0$ yields

$$Q_m(X\beta) \simeq Q_m(0) + (X\beta)\nabla_{(X\beta)}Q_m(0) + \frac{1}{2} (X\beta)^{\prime} \nabla_{(X\beta)(X\beta)^{\prime}}Q_m(0) (X\beta) \equiv (36)$$

$$R_m(X\beta), \quad m = 1, \ldots, M$$

(further terms can of course be included). With $R(X\beta) \equiv \left[ R_1(X\beta), \ldots, R_M(X\beta) \right]'$, one
can finally write $H(X\beta) \simeq G \left[ R(X\beta) \right]$, that is

$$H_m(X\beta) \simeq \frac{\exp \left[ R_m(X\beta) \right]}{\sum_{i=1}^{M} \exp \left[ R_i(X\beta) \right]}, \quad m = 1, \ldots, M.$$ 

Expression (36) indicates that the correctness of $G$ can be assessed by including
squares and cross-products (and possibly higher powers) of the index functions in $X\beta$ as
arguments of the multinomial logit, and subsequently testing for their joint significance.
For instance, under an MNL specification with $M = 3$ and $X\beta = [x^\prime \beta_2, x^\prime \beta_3]^\prime$, a RESET-
type test (with only quadratics) involves testing the significance of the $\eta$ parameters in

$$\frac{\exp (x^\prime \beta_m + w^\prime \eta_m)}{\sum_{i=1}^{3} \exp (x^\prime \beta_i + w^\prime \eta_i)}, \quad m = 1, 2, 3,$$

with $\beta_1 = 0$, $\eta_1 = 0$, and $w$ denotes the vector of nonredundant terms in $\left( X\hat{\beta} \right) \left( X\hat{\beta} \right)'$,
that is, $w = \left[ (x^\prime \hat{\beta}_2)^2, (x^\prime \hat{\beta}_3)^2, (x^\prime \hat{\beta}_2)(x^\prime \hat{\beta}_3) \right]'$. Under a null CL specification, with
$M = 3$, and, accordingly, $X\beta = [x'_1\beta, x'_2\beta, x'_3\beta]'$, $w$ can be expressed as
\[
\begin{align*}
w & \equiv \begin{bmatrix}
(x'_1\beta)^2, (x'_2\beta)^2, (x'_3\beta)^2, \ldots, (x'_2\beta)(x'_3\beta)
\end{bmatrix}', \\
\end{align*}
\]
and each of the $\eta_m$ is a column 6-vector.

Let $\eta \equiv (\eta_1', \ldots, \eta_M')'$, the vector of stacked coefficients' vectors $\eta_m$, of the added terms $w$. An LM test of the null hypothesis $H_0 : \eta = 0$ can be carried out by considering, for each of the rows in $\nabla_{\eta'} \hat{H}_{i-}$,
\[
\nabla_{\eta'} \hat{H}_{im} = \hat{G}_{im} \left( \delta_{im} - \hat{G}_{il} \right) w'_i, \quad l, m = 2, \ldots, M. 
\]

For instance, with $M = 3$,
\[
\nabla_{\eta'} \hat{H}_{i3} = \begin{bmatrix}
\nabla_{\eta_1'} \hat{H}_{i3}, \ \nabla_{\eta_2'} \hat{H}_{i3}
\end{bmatrix} = \begin{bmatrix}
-\hat{G}_{i2} \hat{G}_{i3}, \ \hat{G}_{i3} \left( 1 - \hat{G}_{i2} \right)
\end{bmatrix} \otimes w'_i.
\]

### 4.1.2 Goodness-of-Functional Form Test

The logit likelihood can be nested within a more general specification, in a way that readily enables the assessment of its functional form. The following test is termed ‘goodness-of-functional form’ (GOFF) tests as it generalizes the corresponding GOFF test proposed by Ramalho, Ramalho and Murteira (2011) in the univariate case. The test can also be interpreted as a multivariate version of a "goodness-of-link" test, a procedure used in the generalized linear models literature; see McCullagh and Nelder (1989) for details.

The proposed generalized functional form extends, for the multivariate case, a generalization of the type that is usually employed to introduce asymmetry in the logit model, which consists simply on raising the logit functional form to a positive constant. The alternative model for $E(y_m|X)$ can be expressed as $H_m(X; \beta, \eta) = G_m(X\beta)^{1+\eta_m}$, $\eta_m > -1$, $m = 2, \ldots, M$. From this, the multinomial logit specification results under the null hypothesis $H_0 : \eta = 0$, where $\eta \equiv (\eta_2, \ldots, \eta_M)'$. Then, $\nabla_{\eta'} \hat{H}_{i-} = diag \left( \hat{G}_{im} \log \hat{G}_{im}, \quad m = 2, \ldots, M \right)$. If one single parameter is considered in the alternative model, that is $\eta_2 = \ldots = \eta_M = \eta$, the previous matrix reduces to the vector $\nabla_{\eta'} \hat{H}_{i-} = \left( \hat{G}_{i2} \log \hat{G}_{i2}, \ldots, \hat{G}_{iM} \log \hat{G}_{iM} \right)'$.

### 4.1.3 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. One
such generalization is provided by the nested logit model, the most widely used member of the "generalized extreme-value" class of models (see, e.g., Train, 2009, ch. 4, and the references therein). Ye and Pendyala (2005) raise this concern in the context of time use among several activities, as the fraction of time spent on one occupation may be correlated with that spent on other alternatives, in a way that involves a sequence of allocation decisions over some level hierarchy. One example of the use of a hierarchical approach is provided by Dubin (2007) who uses nested logit market share models to estimate valuations of intangible assets.

Suppose that $M > 2$ and the alternatives can be distributed into $L$ nonoverlapping subsets of similar categories, $S_1, ..., S_L$, $L < M$. Then, the generalized conditional mean function of $y_m$, where alternative $m$ belongs to the subset $S_l$, can be expressed as (see, e.g., Train, 2009, ch. 4.2)

$$H_m (X; \beta, \eta) = \frac{\exp \left[ \mathbf{x}_m \beta_m / (1 + \eta_l) \right] \left\{ \sum_{j \in S_l} \exp \left[ \mathbf{x}_j \beta_j / (1 + \eta_l) \right] \right\}^{\eta_l}}{\sum_{k=1}^{L} \left\{ \sum_{j \in S_k} \exp \left[ \mathbf{x}_j \beta_j / (1 + \eta_k) \right] \right\}^{1 + \eta_k}}$$

(only a two-level hierarchy is considered). It can be seen that $\sum_{l=1}^{L} \sum_{m \in S_l} H_m (X; \beta, \eta) = 1$ and that this formulation nests the multinomial logit for $\eta_l = 0, l = 1, ..., L$. It follows that, for each conditional mean function, the elements of the $(1 \times L)$ vector of partial derivatives $\nabla_{\eta_l} \hat{H}_{lm}$ are given by

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

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

4.2 Tests for Distributional Assumptions

Testing the correct specification of $E (y | X)$ is clearly the most important issue in fractional regression models. However, once the functional form is selected, it may also be important to examine whether a given distribution is appropriate for modeling the fractional response variables in order to obtain efficient ML estimators. 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. Moreover, the simplified OPG version proposed by Chesher (1983) and Lancaster (1984) possesses an asymptotic distribution that is, in general, a very poor approximation to its finite-sample distribution. Therefore, many other forms of the IM test have been proposed and most authors advocate the use of bootstrap-based critical values. The bootstrapped OPG information matrix test, found by Horowitz (1994) to work very well in tobit and binary models, may prove useful in testing for the Dirichlet specification, in line with its implementation in the context of the beta regression model, by Ramalho, Ramalho and Murteira (2011).

A less ambitious approach, corresponding to a restricted version of the IM test and not too difficult to carry out in the context of ML estimation, 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 $E(y|X)$, then one possibility is to assess the validity of the \( \binom{M}{2} \) second moment conditions,

\[
E[(y_{il} - G_{il}) (y_{im} - G_{im}) - \alpha_i G_{il} (\delta_{im} - G_{im})|X_i] = 0, \quad 1 \leq l \leq m \leq M - 1,
\]

where $\alpha_i$ is given by, respectively, $1/(\phi + 1)$ (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

\[
1 = \hat{m}_i' \lambda_1 + \hat{s}_i' \lambda_2 + error,
\]

where $\hat{m}_i \equiv m(y_i, x_i, n_i, \hat{\beta}, \hat{\phi})$ denotes the $i$-th observation of the vector of moment conditions imposed by the model under consideration, $\hat{s}_i$ refers to the $i$-th element of the corresponding score vector, and $\hat{}$ now denotes evaluation at ML estimates.\(^9\) 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 $\hat{m}_i$ vector.

\(^9\)The elements of the score vector are easily obtained in most econometrics packages, which provide specific commands to compute the derivatives of the log-likelihood.
The DM model can also be tested against the multinomial through likelihood-based tests. As mentioned, the DM nests the latter model when $\delta \equiv 1/\phi \to 0^+$. Thus, with a constant $\phi$ parameter in the DM specification (that is, not specified as some function of covariates), validity of the null hypothesis $H_0: \delta = 0$ against the one-sided alternative $H_1: \delta > 0$ can be tested using familiar tests. However, the Wald and LR tests’ null asymptotic distributions are affected by the location of the null hypothesis on the boundary of the parameter space (contrarily to the LM test, which retains its usual asymptotic chi-squared law under the null). With regard to the LR test, Self and Liang (1987) show that the null asymptotic distribution of the corresponding statistic is a 50 : 50 mixture of a chi-squared function with one degree of freedom and a degenerate distribution at zero. Therefore, any upper-tail area calculated with respect to this mixture equals one-half times the corresponding tail area with respect to the chi-squared distribution with one degree of freedom.

5 Monte Carlo Study

This section illustrates the finite-sample performance of most of the estimators and tests discussed throughout this paper through a Monte Carlo simulation. In the first subsection, all experiments involve estimation (by NLS, QML or ML) of a conditional mean function which is correctly specified as multinomial logit. The second subsection illustrates the small sample size and power of several specification tests discussed in the paper.10 All experiments, unless otherwise stated, are based on 10,000 replications, with computations performed using the R software.

5.1 Performance of Alternative Estimators

The five experiments in this subsection assume correct specification of $E(y|X)$ as multinomial logit with $M = 3$ shares, involving alternative-invariant covariates (intercept and $x_2$) and an alternative-specific covariate ($x_3$).11 Formally, the (‘true’ and specified) model for

---

10 At the date of the present version of the paper, simulation results concerning the behaviour of the proposed tests are not yet available.

11 Such a specification is sometimes known as "mixed" logit.
the conditional mean of the dependent variables can be expressed as

$$G_{im} = \frac{\exp (v_i'\beta_m^\nu + \beta_3 x_{i3})}{\sum_{i=1}^3 \exp (v_i'\beta_1^\nu + \beta_3 x_{i3})},$$

where $v_i \equiv (1, x_{i2})'$, with conformable parameter vectors $\beta_m^\nu \equiv (\beta_{1m}, \beta_{2m})$, $m = 2, 3$ and $\beta_1^\nu = 0$. The regressors are newly drawn in each replica, obtained as i.i.d. draws from the following distributions: $x_2 \sim N(0, 1)$ and $x_3$ is obtained as the sum of an exponentially-distributed random variable with mean $10^{-1}$ and the constants $a - 10^{-1}$, with $a = m - 2$, $m = 1, 2, 3$. The covariate $x_3$ is alternative-specific, so it can only assume one of possible three different values in each replica. Therefore, it is drawn from one of the three mentioned distributions, designed to produce considerably stable values around each mean (the variance equals $10^{-2}$ and the means equal $a$).

Different parameter values are considered in each of four different designs (named A, B, C and D), as summarized in Table 1. This combination of covariates and parameters’ values yields various distributions of shares for the different alternatives. As can be seen from Table 1, the mean shares for the three alternatives are quite balanced in Design A and become less and less so through the remaining designs, with Design D producing the greatest dispersion of mean shares.

The reported simulation results in the first four experiments refer to samples with size $N = 100$ – for different sample sizes ($200 \leq N \leq 500$) results follow similar patterns, so they are omitted. The fifth experiment, based on Design D and 1000 replications, compares the performance of alternative estimators for different sample sizes.

### Table 1

<table>
<thead>
<tr>
<th>Experimental Designs</th>
<th>$\beta_{12}$</th>
<th>$\beta_{22}$</th>
<th>$\beta_{13}$</th>
<th>$\beta_{23}$</th>
<th>$\beta_3$</th>
<th>Mean shares$^*$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>$y_1$</td>
</tr>
<tr>
<td>A</td>
<td>-1.0</td>
<td>-1.0</td>
<td>-0.7</td>
<td>1.0</td>
<td>0.3</td>
<td>32.6</td>
</tr>
<tr>
<td>B</td>
<td>-0.5</td>
<td>-1.0</td>
<td>-0.5</td>
<td>1.0</td>
<td>1.0</td>
<td>57.8</td>
</tr>
<tr>
<td>C</td>
<td>-0.5</td>
<td>-1.0</td>
<td>0.5</td>
<td>1.0</td>
<td>1.1</td>
<td>48.1</td>
</tr>
<tr>
<td>D</td>
<td>1.0</td>
<td>-1.0</td>
<td>-1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>83.1</td>
</tr>
</tbody>
</table>

*: Mean shares obtained from a simulated sample of size 100,000.
In the first experiment, the observations on \( y \) are obtained as i.i.d. draws from a conditional Dirichlet distribution, \( f_{y|X}(y; \phi, \beta|X) \), the formal expression of which is presented in (21). Five different values of the parameter \( \phi \) are considered, in order to allow for different degrees of variability of the response variables: \( \phi \in \{10, 20, 30, 40, 50\} \).

With this data, the parameters of the correctly specified conditional mean model are estimated by NLS, MB-QML and Dirichlet-ML (D-ML). All these estimators are consistent under the four designs, so the main objective of the exercise is to investigate the possible efficiency gain of the D-ML estimator, with reference to the former two.

The main results of this experiment are depicted in Figure 1. All the plots graph the values of the root mean squared errors (RMSE) of the conditional mean parameters’ estimates, for the five different values of \( \phi \).

With the exception of covariates’ coefficients under designs C and D and lower \( \phi \) values, the RMSE’s produced by the three estimation methods are not very different. Expectably, D-ML features a slight efficiency advantage over the other two estimators, which, nevertheless, seems attenuated for higher values of \( \phi \). NLS performs invariably worst, namely for small values of \( \phi \), proving to be the method whose estimates’ precision is most sensitive to the variability of the dependent variables.

**Experiment 2**

In the second experiment the response variables are obtained as integers ratios from a conditional multinomial distribution. That is, observations on \( y \) are obtained as i.i.d. draws from a conditional distribution such, that, conditionally on \( X \), \( n^y \) follows a multinomial distribution with the previously described multinomial logit conditional mean (with parameters values belonging to each of the four designs, A through D). The formal expression of this multivariate p.f. is presented in (24), with \( G_m \) instead of \( \pi_m \). For each individual observation, \( n \) is obtained as an i.i.d. draw from a discrete uniform p.f. \( U(1, n_{\text{max}}) \), where \( n_{\text{max}} \in \{11, 21, 31, 41, 51\} \). These values influence the degree of variability of the responses, so they are bound to influence the precision of the various estimators.

With this data, the conditional mean parameters are estimated by NLS, MB-QML, multinomial-based ML (MULT), D-ML 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 \( 1 - 10^{-6} \). The objectives of
this exercise are the following: (i) to evaluate the bias of the D-ML estimator, which is inconsistent in the present case; (ii) to assess of the possible advantage of MULT over NLS and, especially, the MB-QML method, due to the fact that MULT makes use of potentially useful information (on $n$) that is ignored by the latter; (iii) to check the behaviour of the DM-ML estimator, itself a generalization of the MULT approach.

The main results of this experiment are shown in Figure 2. Here, the plots graph the values of the root mean squared errors (RMSE) of the parameters’ estimates, for the five different values of $n_{\text{max}}$.

[Figure 2 about here]

In what concerns the performance of D-ML, its estimates are biased, as expected. This bias is exemplified in Table 2, which contains the averages of the coefficients’ estimates under design A.

<table>
<thead>
<tr>
<th>Table 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Experiment 2, Design A</td>
</tr>
<tr>
<td>Empirical means of Dirichlet-ML Estimates</td>
</tr>
<tr>
<td>[ \begin{array}{cccccc}</td>
</tr>
<tr>
<td>\beta_{12} &amp; \beta_{22} &amp; \beta_{13} &amp; \beta_{23} &amp; \beta_3</td>
</tr>
<tr>
<td>\hline</td>
</tr>
<tr>
<td>DGP &amp; -1.0 &amp; -1.0 &amp; -0.7 &amp; 1.0 &amp; 0.3</td>
</tr>
<tr>
<td>$n_{\text{max}}$ &amp; \hline</td>
</tr>
<tr>
<td>11 &amp; -0.999 &amp; -0.597 &amp; -0.589 &amp; 0.596 &amp; 0.409</td>
</tr>
<tr>
<td>21 &amp; -0.998 &amp; -0.659 &amp; -0.608 &amp; 0.660 &amp; 0.391</td>
</tr>
<tr>
<td>31 &amp; -1.000 &amp; -0.797 &amp; -0.646 &amp; 0.802 &amp; 0.353</td>
</tr>
<tr>
<td>41 &amp; -0.999 &amp; -0.694 &amp; -0.616 &amp; 0.694 &amp; 0.383</td>
</tr>
<tr>
<td>51 &amp; -1.000 &amp; -0.762 &amp; -0.634 &amp; 0.758 &amp; 1.364</td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

With regard to the relative performance of MULT, NLS and MB-QML methods, once again NLS is the less efficient, performing very poorly under all designs except Design A. The MULT (ML) estimator appears slightly more efficient than MB-QML, which is expected. 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
convergence of the DM-ML method is often difficult to achieve and (understandably) always for quite large $\phi$ estimates (see the remarks on eq. (30)). 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.

**Experiment 3**

In this experiment the response variables are obtained as i.i.d. integers ratios from a conditional Dirichlet-multinomial mixture p.f. with one of the above mentioned multinomial logit conditional means. The formal expression of this model is presented in (30). The parameter $n$ is drawn from the same discrete uniform p.f. that is used in Experiment 2 and $\phi$ is set to 10. As a consequence, the variances of the dependent variables are now considerably higher than in Experiment 2 (these variances can be more than five times those in Experiment 2, for the same conditional mean – compare (32) and (27)) Under this DGP, the conditional mean parameters are estimated with the five methods used in Experiment 2 (zeros and ones are again modified in the case of D-ML estimation).

The main objectives of the experiment are the following: (i) to evaluate the potential advantage of MULT over the NLS and MB-QML methods; (ii) to assess the difference of precision of the DM-ML and QML estimators (MB and MULT).

Figure 3 sums up the results of the experiment in terms of RMSE of the conditional mean parameters’ estimates. As in Experiment 2, the various plots graph RMSE results for the four designs and five different values of $n_{\text{max}}$.

Among the methods used, NLS again performs very poorly, namely under designs C and D. The MULT estimator outperforms MB-QML in all the cases considered, so use of the available information on $n$ seems advantageous in what concerns QML methods. The MULT estimator appears slightly less efficient than DM-ML, which is to be expected in the present experiment.

**Experiment 4**

As in Experiment 3, the response variables are obtained as i.i.d. integers ratios from a conditional Dirichlet-multinomial mixture p.f. with one of the above mentioned multinomial logit conditional means. Differently from the previous experiment, $n$ is now taken to be discrete uniform, ranging from 1 to 11, and five values of $\phi$ are considered. The main
objectives of this experiment are basically the same as those of Experiment 3, now for varying \( \phi \) parameter. The inconsistent D-ML estimator is also included but, as expected, it often fares considerably worse than the consistent estimators (with the frequent exception of the NLS estimator). In most cases, MB-QML is less precise than MULT, which exhibits an almost identical performance to DM-ML regardless of the value of \( \phi \). This is somewhat noteworthy, namely for low \( \phi \) values under which the DM and multinomial are less alike.

[Figure 4 about here]

**Experiment 5**

In Experiment 5 the performance of alternative estimators under different sample sizes \( (N) \) is investigated. The first row of figures depicts the behaviour of NLS, MB-QML and D-ML estimators under a Dirichlet population with \( \phi = 10 \) and fixed group size. As before, the NLS estimator performs worst, although, for \( N \geq 500 \) it becomes quite close to MB-QML. As expected, both estimators exhibit a poorer performance than D-ML, namely for low \( N \).

The second row depicts the behaviour of all considered estimators under a conditional MULT population with varying group sizes. Now, the D-ML estimator behaves poorly, as expected, given its inconsistency. In what regards the remaining methods, understandably QML estimators behave worse than ML, with DM- and MULT-ML estimators performing almost identically. Results for the third row (DM population) are almost identical to those in the second row. A slight disagreement is noted with reference to the second row, with a small advantage of DM-ML over MULT-QML, namely for larger sample sizes.

[Figure 5 about here]

### 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 (Dirichlet and Dirichlet-multinomial) regression models, also 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 rather well with the ML estimators (Dirichlet and multinomial-Dirichlet), even when the latter are implemented under fully correct distributional assumptions. 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 unadvisable.

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 and GOFF tests), a test of the multinomial logit against the nested logit is also proposed. All conditional mean specification tests are proposed as LM tests, implemented upon QML estimation and using artificial OLS regressions. Besides these tests, some tests of the fully parametric Dirichlet and Dirichlet-multinomial specifications — or of some of its implied assumptions — are also proposed. The performance of all types of tests is to be studied through an extensive simulation study, presently on going.

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, potentially applicable in several contexts of practical interest.
References


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


7 Appendix

This Appendix contains algebraic derivations of some expressions that are used in the tests for the specification of the conditional mean, $E(y|X)$ (section 4).

1. Expression of $C_i$

Following Wooldridge (1991), under an LEF log-likelihood of the form

$$LL_i = a\left(G_i^\top\nu_i\right) + b\left(y_i^\top\nu_i\right) + y_i^\top c\left(G_i^\top\nu_i\right) \equiv a_i + b_i + y_i^\top c_i,$$

with $\nu_i$ a vector of covariates and nuisance parameters, $a(\cdot)$ and $b(\cdot)$ scalars, and $c(\cdot)$ an $(M - 1)$–column vector of functions, the matrix $C_i$ is defined as the inverse of the matrix of derivatives $\nabla_{G_i^\top}c_i$. In the MB QML case,

$$c_i = \left(\log\frac{G_{i2}}{G_{i1}}, \ldots, \log\frac{G_{iM}}{G_{i1}}\right)',$n
\quad G_{i1} = 1 - \sum_{m=2}^{M} G_{im},$$

from which $C_i$ is given by (35). Under multinomial-based QML,

$$c_i = n_i \left(\log\frac{G_{i2}}{G_{i1}}, \ldots, \log\frac{G_{iM}}{G_{i1}}\right)',$
\quad so the previous matrix $C_i$ should be scaled by the factor $n_i^{-1}$. Under NLS estimation (which can be interpreted as QML based on a Gaussian likelihood with uncorrelated, unit-variance, errors), after replacing $G_{i1}$ by $1 - \sum_{m=2}^{M} G_{im}$, one obtains

$$c_i = \begin{bmatrix}
2G_{i2} + G_{i3} + \cdots + G_{iM} \\
G_{i2} + 2G_{i3} + \cdots + G_{iM} \\
\vdots \\
G_{i2} + G_{i3} + \cdots + 2G_{iM}
\end{bmatrix} \Rightarrow C_i = \left(I_{M-1} + \mathbf{\nu}_{M-1}' \mathbf{\nu}_{M-1}^{-1}\right)^{-1},$$

which can be checked to equal $I_{M-1} - M^{-1} \mathbf{\nu}_{M-1}' \mathbf{\nu}_{M-1}$.
2. Derivatives $\nabla_{\eta_i} E(y_m|X)$ for the test of the multinomial logit against the nested logit

If alternative $m$ belongs to $S_l$, 

$$
\nabla_{\eta_i} H_m = \frac{\exp [x'_m \beta_m / (1 + \eta_l)] \left\{ \sum_{j \in S_l} \exp [x'_j \beta_j / (1 + \eta_l)] \right\}^{\eta_i}}{\sum_{k=1}^{L} \left\{ \sum_{j \in S_k} \exp [x'_j \beta_j / (1 + \eta_k)] \right\}^{1+\eta_k}} + 
\frac{\sum_{k=1}^{L} \left\{ \sum_{j \in S_k} \exp [x'_j \beta_j / (1 + \eta_k)] \right\}^{1+\eta_k} \times 
\left( \sum_{j \in S_l} \exp [x'_j \beta_j / (1 + \eta_l)] \right)^{\eta_i} \log \left( \sum_{j \in S_l} \exp [x'_j \beta_j / (1 + \eta_l)] \right) + 
\eta_l \left\{ \sum_{j \in S_l} \exp [x'_j \beta_j / (1 + \eta_l)] \right\} \sum_{j \in S_l} \exp [x'_j \beta_j / (1 + \eta_l)] \left[ - \frac{x'_j \beta_j}{(1 + \eta_l)^2} \right] - 
\frac{\exp [x'_m \beta_m / (1 + \eta_l)] \left\{ \sum_{j \in S_l} \exp [x'_j \beta_j / (1 + \eta_l)] \right\}^{1+\eta_l} \log \left( \sum_{j \in S_l} \exp [x'_j \beta_j / (1 + \eta_l)] \right)}{\sum_{k=1}^{L} \left\{ \sum_{j \in S_k} \exp [x'_j \beta_j / (1 + \eta_k)] \right\}^{1+\eta_k}} \times 
(1 + \eta_l) \left\{ \sum_{j \in S_l} \exp [x'_j \beta_j / (1 + \eta_l)] \right\} \sum_{j \in S_l} \exp [x'_j \beta_j / (1 + \eta_l)] \left[ - \frac{x'_j \beta_j}{(1 + \eta_l)^2} \right].
$$

Therefore,

$$
\nabla_{\eta_i} \hat{H}_{im} = \frac{\exp (x'_{im} \hat{\beta}_m) (-x'_{im} \hat{\beta}_m)}{\sum_{k=1}^{L} \sum_{j \in S_k} \exp (x'_{ij} \hat{\beta}_j)} + \frac{\exp (x'_{im} \hat{\beta}_m)}{\sum_{k=1}^{L} \sum_{j \in S_k} \exp (x'_{ij} \hat{\beta}_j)} \times 
\log \left( \sum_{j \in S_l} \exp (x'_{ij} \hat{\beta}_j) \right) - \frac{\exp (x'_{im} \hat{\beta}_m)}{\left\{ \sum_{k=1}^{L} \sum_{j \in S_k} \exp (x'_{ij} \hat{\beta}_j) \right\}^{2}} \times 
\left\{ \sum_{j \in S_l} \exp (x'_{ij} \hat{\beta}_j) \log \left( \sum_{j \in S_l} \exp (x'_{ij} \hat{\beta}_j) \right) + \sum_{j \in S_l} \exp (x'_{ij} \hat{\beta}_j) \left( - x'_{ij} \hat{\beta}_j \right) \right\} = 
\exp (x'_{im} \hat{\beta}_m) \times \left( \log \left( \sum_{j \in S_l} \exp (x'_{ij} \hat{\beta}_j) \right) - x'_{im} \hat{\beta}_m \right) - 
\frac{\sum_{k=1}^{L} \sum_{j \in S_k} \exp (x'_{ij} \hat{\beta}_j)}{\sum_{j \in S_l} \exp (x'_{ij} \hat{\beta}_j)} \times \left( \log \sum_{j \in S_l} \exp (x'_{ij} \hat{\beta}_j) \right) - x'_{im} \hat{\beta}_m - 
\frac{\sum_{j \in S_l} \exp (x'_{ij} \hat{\beta}_j)}{\sum_{k=1}^{L} \sum_{j \in S_k} \exp (x'_{ij} \hat{\beta}_j)} \times \left( \log \sum_{j \in S_l} \exp (x'_{ij} \hat{\beta}_j) \right) - x'_{ij} \hat{\beta}_j \right) \right) \right) \right) = 
\hat{G}_{im} \left( \log \sum_{j \in S_l} \exp (x'_{ij} \hat{\beta}_j) \right) - x'_{im} \hat{\beta}_m - \sum_{j \in S_l} \hat{G}_{ij} \left( \log \sum_{j \in S_l} \exp (x'_{ij} \hat{\beta}_j) \right) - x'_{ij} \hat{\beta}_j \right) \right) \right).
If alternative \( m \) belongs to \( S_h \neq S_l \),

\[
\nabla_q H_m = - \frac{\exp \left[ x'_m \hat{\beta}_m / (1 + \eta_h) \right] \left\{ \sum_{j \in S_h} \exp \left[ x'_j \hat{\beta}_j / (1 + \eta_h) \right] \right\}^{\eta_h}}{\left( \sum_{k=1}^L \left\{ \sum_{j \in S_k} \exp \left[ x'_j \hat{\beta}_j / (1 + \eta_k) \right] \right\}^{1+\eta_k} \right)^2} \times \\
\left( \left\{ \sum_{j \in S_l} \exp \left[ x'_j \hat{\beta}_j / (1 + \eta_l) \right] \right\}^{1+\eta_l} \log \left\{ \sum_{j \in S_l} \exp \left[ x'_j \hat{\beta}_j / (1 + \eta_l) \right] \right\} + \\
(1 + \eta_l) \left\{ \sum_{j \in S_l} \exp \left[ x'_j \hat{\beta}_j / (1 + \eta_l) \right] \right\}^{\eta_l} \sum_{j \in S_l} \exp \left[ x'_j \hat{\beta}_j / (1 + \eta_l) \right] \left[ -\frac{x'_j \hat{\beta}_j}{(1 + \eta_l)^2} \right] \right) .
\]

Therefore,

\[
\nabla_q \hat{H}_{im} = - \frac{\exp \left( x'_{im} \hat{\beta}_m \right)}{\left\{ \sum_{k=1}^L \left[ \sum_{j \in S_k} \exp \left( x'_j \hat{\beta}_j \right) \right] \right\}^2} \times \\
\left\{ \sum_{j \in S_l} \exp \left( x'_{ij} \hat{\beta}_j \right) \log \left[ \sum_{j \in S_l} \exp \left( x'_{ij} \hat{\beta}_j \right) \right] + \sum_{j \in S_l} \left[ \exp \left( x'_{ij} \hat{\beta}_j \right) \left( -x'_{ij} \hat{\beta}_j \right) \right] \right\} = \\
\exp \left( x'_{im} \hat{\beta}_m \right) \times \\
\sum_{k=1}^L \left[ \sum_{j \in S_k} \exp \left( x'_j \hat{\beta}_j \right) \right] \sum_{j \in S_l} \left[ \exp \left( x_{ij} \hat{\beta}_j \right) \{ \log \left[ \sum_{j \in S_l} \exp \left( x_{ij} \hat{\beta}_j \right) \right] - x_{ij} \hat{\beta}_j \} \right] = \\
\sum_{k=1}^L \left[ \sum_{j \in S_k} \exp \left( x'_j \hat{\beta}_j \right) \right] \\
- \hat{G}_{im} \sum_{j \in S_l} \hat{G}_{ij} \left\{ \log \left[ \sum_{j \in S_l} \exp \left( x_{ij} \hat{\beta}_j \right) \right] - x_{ij} \hat{\beta}_j \right\} .
\]
Figure 1: RMSE comparison of alternative estimators for multivariate fractional regression models (Dirichlet–distributed response variable; N = 100)
Figure 2: RMSE comparison of alternative estimators for multivariate fractional regression models
(Multinomial–distributed response variable; n random – min(n) = 1; N = 100)

Design A

Design B

Design C

Design D
Figure 3: RMSE comparison of alternative estimators for multivariate fractional regression models
(Dirichlet–Multinomial–distributed response variable; n random − \( \min(n) = 1; \phi = 10, N = 100 \)
Figure 4: RMSE comparison of alternative estimators for multivariate fractional regression models
(Dirichlet–Multinomial–distributed response variable; n random − min(n) = 1; max(n) = 11, N = 100)
Figure 5: RMSE comparison of alternative estimators for multivariate fractional regression models – different sample sizes (Design D)