JOURNAL OF ECONOMIC SURVEYS

doi: 10.1111/j.1467-6419.2009.00602.x

ALTERNATIVE ESTIMATING AND TESTING EMPIRICAL STRATEGIES FOR FRACTIONAL REGRESSION MODELS

Esmeralda A. Ramalho and Joaquim J.S. Ramalho

Departamento de Economia and CEFAGE-UE, Universidade de Évora

José M.R. Murteira

Faculdade de Economia, Universidade de Coimbra and CEMAPRE

Abstract. In many economic settings, the variable of interest is often a fraction or a proportion, being defined only on the unit interval. The bounded nature of such variables and, in some cases, the possibility of nontrivial probability mass accumulating at one or both boundaries raise some interesting estimation and inference issues. In this paper we (i) provide a comprehensive survey of the main alternative models and estimation methods suitable to deal with fractional response variables, (ii) propose a full testing methodology to assess the validity of the assumptions required by each alternative estimator and (iii) examine the finite-sample properties of most of the estimators and tests discussed through an extensive Monte Carlo study. An application concerning corporate capital structure choices is also provided.

Keywords. Conditional mean tests; Fractional regression models; Non-nested hypotheses; Two-part models; Zero outcomes

1. Introduction

In many economic settings, the variable of interest (y) is often a proportion, being defined and observed only on the standard unit interval, i.e. $0 \leq y \leq 1$. Examples include pension plan participation rates, firm market share, proportion of debt in the financing mix of firms, fraction of land area allocated to agriculture and proportion of exports in total sales. The bounded nature of such variables and, in some cases, the possibility of nontrivial probability mass accumulating at one or both boundaries raise some interesting estimation and inference issues. In particular, the standard practice of using linear models to examine how a set of explanatory variables influences a given proportional or fractional response variable is not appropriate in general, since it does not guarantee that the predicted values of the dependent variable are restricted to the unit interval. Nevertheless, only in the last decade have

researchers begun to take seriously the functional form issues raised by fractional data, proposing the so-called fractional regression models that take into account the specific characteristics of fractional response variables; see Papke and Wooldridge's (1996) seminal paper.

Frequently, in applied work, researchers' main interest lies in the estimation of the conditional mean of y, given a set of regressors.[1] In this case, practitioners face two main decisions: (i) which functional form to assume for the conditional expectation of y and (ii) which method to employ in the estimation of the resulting model. In addition, in the case of boundary observations, practitioners have also to decide whether one- or two-part models should be used. So far, most authors have chosen to assume a logistic form for the conditional mean of y, without assessing whether alternative functional forms are more appropriate, and to use the robust quasi-maximum likelihood (QML) method suggested by Papke and Wooldridge (1996), without checking whether a more efficient estimation method could be used. However, in both cases, there are a number of alternatives that may be employed and various simple test procedures that may be used to assess their adequacy. Similarly, the option between a single- and a two-part model is usually made *a priori* and, as far as we know, has never been tested.

In this paper we survey the main alternative regression models and estimation methods that are available for dealing with fractional response variables and propose a full testing methodology to assess the validity of the assumptions required by each estimator. We briefly discuss tests for distributional assumptions and examine in detail tests for conditional mean assumptions, which may also be used for choosing between one-part and two-part models. In addition to the tests that are commonly employed in the econometrics literature (RESET tests) or in the statistics literature of binary models (goodness-of-link tests), we suggest a new class of tests that are valid for testing the correct specification of any conditional mean model (including two-part models) and investigate the application of non-nested tests in this framework. We provide an integrated approach for all conditional mean tests, implementing all of them as Lagrange multiplier (LM) tests for omitted variables, which are calculated using simple artificial regressions.

To the best of our knowledge, no simulation study concerning fractional response variables has ever been undertaken. Therefore, in this paper we also carry out an extensive Monte Carlo simulation study that evaluates the finite-sample properties of most of the estimation methods and tests discussed in the paper under many alternative data-generating processes. To illustrate the usefulness in empirical work of the various techniques discussed in the paper, we apply some of them to the analysis of corporate capital structure decisions, where the variable of interest is usually a leverage (debt to capital assets) ratio, which is defined only on the unit interval and is often null for many firms.

The paper is organized as follows. Section 2 describes the notational framework and discusses the main issues raised when the variable of interest is fractional. Section 3 examines the main alternative regression models and estimation methods that are commonly used with fractional response variables. Section 4 discusses some specification tests for those models and methods. The Monte Carlo simulation

study is described in Section 5. Section 6 is dedicated to the empirical application. Finally, Section 7 contains some concluding remarks and suggestions for future research. An appendix summarizing some practical procedures for dealing with fractional responses is also provided.

2. Framework

Consider a random sample of $i = 1, \ldots, N$ individuals and let y be the fractional variable of interest, $0 \leq y \leq 1$, and x a vector of k covariates. Let θ be the vector of parameters to be estimated and $f(y|x, \theta)$ denote the conditional density of y, which may be known or unknown.

For many years, three main approaches have been followed to model fractional response variables. The first of them, still used by many empirical researchers, is simply to ignore the bounded nature of y and assume a linear conditional mean model for y:

$$E(y|x) = x\theta \tag{1}$$

However, given that y is strictly bounded from above and below, it is in general unreasonable to assume that the effect of any explanatory variable is constant throughout its entire range. Moreover, this linear specification cannot guarantee that the predicted values of y lie between 0 and 1 without severe constraints on the range of x or *ad hoc* adjustments to fitted values outside the unit interval.

Aware of these problems, some empirical researchers opted for assuming the logistic relationship

$$E(y|x) = \frac{e^{x\theta}}{1 + e^{x\theta}} \tag{2}$$

which is indeed a natural choice for modelling proportions since it ensures that $0 < E(y|x) < 1$. However, instead of estimating equation (2) directly, which would require some nonlinear technique, most authors prefer to estimate by least squares the log-odds ratio model defined by

$$E\left(\log \frac{y}{1-y} \middle| x\right) = x\theta \tag{3}$$

which basically corresponds to the linearization of the equation that results from solving $y = e^{x\theta}/(1 + e^{x\theta})$ with respect to $x\theta$. This approach has two main drawbacks. On the one hand, from equation (3) it would not be straightforward to recover $E(y|x)$ and, thus, to interpret the estimates of θ, which would still be the main interest of the analysis; see *inter alia* Papke and Wooldridge (1996) for details. On the other hand, the transformed dependent variable in equation (3) is not well defined for the boundary values 0 and 1 of y, requiring *ad hoc* adjustments if such values are observed in the sample (such as adding an arbitrarily chosen small constant to all observations of y).

Finally, when there are many observations at the upper and/or lower limits of the response variable, it is relatively common to use Tobit models for data censored

at one and/or zero. Again, there are some problems with this approach. First, only in the two-limit Tobit model are in fact the predicted values of y restricted to the unit interval. However, that model can only be applied when we have observations in both limits, which is often not the case. Second, conceptually, as some authors argue (e.g. Maddala, 1991), a Tobit model is appropriate to describe *censored* data in the interval [0, 1] but its application to data *defined* only in that interval is not easy to justify: observations at the boundaries of a fractional variable are a natural consequence of individual choices and not of any type of censoring. Third, the Tobit model is very stringent in terms of assumptions, requiring normality and homoskedasticity of the dependent variable, prior to censoring.

Given the limitations of these models, some alternative approaches that account for the bounded nature of the variable of interest have recently been proposed. Some of them can only be used when there are no observations at the boundaries, while others may also be employed when one or both the limits are observed with a positive probability. However, all of them have in common the utilization of functional forms for the conditional mean of y that enforce the conceptual requirement that $E(y|x)$ is in the unit interval. In the next section we discuss the main alternative functional forms and regression models suggested in prior research.

3. Regression Models for Fractional Response Variables

Two main approaches for modelling fractional data without boundary observations have been proposed so far. The first only requires the correct specification of the (nonlinear) conditional expectation of the fractional response variable. The second alternative is a fully parametric approach, where a particular conditional distribution is assumed for the fractional dependent variable. Only the first approach can also be, in general, applicable to cases where there are a finite number of boundary observations, although in such cases it is often a better choice to use a two-part model, where first a discrete choice model is assumed to describe the fact that y is a boundary observation or not, and then, only for those individuals with $y \in (0, 1)$, a conditional mean or a parametric model is employed. Next, we discuss these three alternative approaches and discuss in which cases one-part or two-part models should be used for modelling fractional responses characterized by a large cluster of data at zero.

3.1 *Nonlinear Models for the Conditional Mean*

The simplest solution for dealing with fractional response variables only requires the assumption of a functional form for y that imposes the desired constraints on the conditional mean of the dependent variable:

$$E(y|x) = G(x\theta) \tag{4}$$

where $G(\cdot)$ is a known nonlinear function satisfying $0 \leq G(\cdot) \leq 1$. This approach was first formally proposed by Papke and Wooldridge (1996), who suggested as possible specifications for $G(\cdot)$ any cumulative distribution function. An obvious

Table 1. Alternative Nonlinear Conditional Mean Specifications for Fractional Response Variables.

Model designation	Distribution function	$G(x\theta)$	$g(x\theta)$	$h(\mu)$
Cauchit	Cauchy	$\frac{1}{2}+\frac{1}{\pi}\arctan(x\theta)$	$\frac{1}{\pi}\frac{1}{(x\theta)^2+1}$	$\tan[\pi(\mu-0.5)]$
Logit	Logistic	$\frac{e^{x\theta}}{1+e^{x\theta}}$	$G(x\theta)[1-G(x\theta)]$	$\ln\frac{\mu}{1-\mu}$
Probit	Standard normal	$\Phi(x\theta)$	$\Phi(x\theta)$	$\Phi^{-1}(\mu)$
Loglog	Extreme maximum	$e^{-e^{-x\theta}}$	$e^{-x\theta}G(x\theta)$	$-\ln[-\ln(\mu)]$
Complementary loglog	Extreme minimum	$1-e^{-e^{x\theta}}$	$e^{x\theta}[1-G(x\theta)]$	$\ln[-\ln(1-\mu)]$

choice for $G(\cdot)$ is the logistic function (2) which, however, instead of being first linearized as discussed above, must be directly estimated using nonlinear techniques.

In Table 1 we present some popular choices for $G(\cdot)$ and corresponding derivatives with respect to the index $x\theta$, $g(x\theta)=\partial G(x\theta)/\partial x\theta$, and the so-called link functions, $h(\mu)$, which will be defined later on. As is well known, while the logistic, standard normal and Cauchy specifications for $G(\cdot)$ are symmetric about the point 0.5 and, consequently, approach 0 and 1 at the same rate, the loglog and complementary loglog models are not symmetric: the former increases sharply at small values of $G(\cdot)$ and slowly when $G(\cdot)$ is near 1, while the latter exhibits the opposite behaviour. The Cauchy distribution presents the heaviest tails, which implies that this specification is more robust to outliers than the logistic and standard normal formulations.

The model defined by equation (4) may be consistently estimated by nonlinear least squares (NLS), as in Hermalin and Wallace's (1994) empirical application, or, as suggested by Papke and Wooldridge (1996), by QML. The latter authors proposed a particular QML method based on the Bernoulli log-likelihood function, which is given by

$$LL_i(\theta)=y_i\log[G(x_i\theta)]+(1-y_i)\log[1-G(x_i\theta)] \tag{5}$$

As the Bernoulli distribution is a member of the linear exponential family (LEF), the QML estimator of θ defined by

$$\hat{\theta}=\arg\max_{\theta}\sum_{i=1}^{N}LL_i(\theta) \tag{6}$$

is consistent and asymptotically normal, regardless of the true distribution of y conditional on x, provided that $E(y|x)$ in equation (4) is indeed correctly specified

(see Gouriéroux *et al.* (1984) for details). Moreover (see Papke and Wooldridge, 1996), there are some cases where this QML estimator is efficient in a class of estimators containing all LEF-based QML and weighted NLS estimators. The asymptotic distribution of the QML estimator is given by

$$\sqrt{N}(\hat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}(0, V) \tag{7}$$

where $V = A^{-1}BA^{-1}$, with $A = E[-\nabla_{\theta\theta'} LL(\theta)]$ and $B = E[\nabla_{\theta} LL(\theta)\, \nabla_{\theta'} LL(\theta)]$. Consistent estimators for A and B are given by $\hat{A} = N^{-1} \sum_{i=1}^{N} \hat{g}_i^2 x_i' x_i [\hat{G}_i(1 - \hat{G}_i)]^{-1}$ and $\hat{B} = N^{-1} \sum_{i=1}^{N} \hat{u}_i^2 \hat{g}_i^2 x_i' x_i [\hat{G}_i(1 - \hat{G}_i)]^{-2}$, respectively, where $\hat{G}_i = G(x_i\hat{\theta})$, $\hat{g}_i = g(x_i\hat{\theta})$ and $\hat{u}_i = y_i - \hat{G}_i$.

For some examples of applications where these methods have been employed, see Hausman and Leonard (1997), Wagner (2001) and Czarnitzki and Kraft (2004), who use the QML method based on the logistic specification to estimate regression models for, respectively, television rating on NBA games, the proportion of exports in a firm production and the share of turnover with new and improved products. An earlier application, based on NLS and the cumulative normal function, may be found in Hermalin and Wallace (1994).

3.2 *Parametric Models: The Beta Fractional Regression Model*

Even when interest is confined to the parameters of the conditional mean function (4), in addition to assuming a given functional form for $E(y|x)$, the researcher may also be willing to specify the conditional distribution $f(y|x, \theta)$ in order to obtain more efficient estimators. There are several statistical distributions that are appropriate for data confined to the unit interval and, hence, may be used in this context. However, all the most commonly used distributions suffer from two drawbacks: (i) as they do not belong to the LEF, the resulting estimators may be non-robust to deviations from the assumed distribution; and (ii) they are defined only in the open interval (0, 1) and therefore cannot be used when there are limit observations.

Due to its known flexibility that allows a great variety of asymmetric forms, the most popular choice for $f(y|x, \theta)$ is the beta distribution; see *inter alia* Brehm and Gates (1993), Haab and McConnell (1998) and Paolino (2001) for some applications of the beta fractional regression model.[2] Although the beta distribution has been used extensively in statistics for more than a century, the literature on the beta regression model is scarce and very recent. Indeed, only in the past decade does the beta regression model seem to have been used for the first time; see *inter alia* Brehm and Gates (1993). Their approach was based on the standard beta density function, which is given by

$$f(y; p, q) = \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)} y^{p-1}(1-y)^{q-1} \tag{8}$$

where $\Gamma(\cdot)$ denotes the gamma function, $0 < y < 1$ and $p > 0$ and $q > 0$ are shape parameters, both of which were specified by Brehm and Gates (1993) as exponential functions of the covariates. However, estimating a covariate's relationship to a shape

parameter is rarely of interest. Therefore, the most recent approaches to the beta regression model work with a different parametrization of the beta density, the same that we adopt in this paper.

As found independently by Paolino (2001) and Ferrari and Cribari-Neto (2004), the interpretation of the parameters of the beta regression model is greatly simplified if a mean-dispersion parametrization of the beta density is used. Let $p = \mu\phi$ and $q = (1 - \mu)\phi$. Then, it follows that

$$f(y; \mu, \phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma[(1-\mu)\phi]} y^{\mu\phi-1}(1-y)^{(1-\mu)\phi-1} \tag{9}$$

which implies

$$E(y) = \frac{p}{p+q} = \mu \tag{10}$$

and

$$\text{Var}(y) = \frac{pq}{(p+q)^2(p+q+1)} = \frac{\mu(1-\mu)}{\phi+1} \tag{11}$$

so that μ is the mean of the response variable and ϕ can be interpreted as a precision parameter in the sense that, for fixed μ, the larger the value of ϕ, the smaller the variance of y. A similar approach was followed by Haab and McConnell (1998) and Kieschnick and McCullough (2003), which, however, kept p in the model instead of introducing the precision parameter.

Based on this approach, two different beta regression models have been proposed. In the simplest case, we simply assume $\mu = G(x\theta)$ as in the models discussed in the previous section and treat ϕ (or p) as a nuisance parameter. Alternatively, considering that a researcher may be interested in analysing whether a variable contributes to the variance of y beyond its effect upon the mean, Paolino (2001) and also Smithson and Verkuilen (2006) assume $\mu = G(x\theta)$ and $\phi = \exp(z\alpha)$, where z is a set of independent variables, potentially distinct from those included in x, and α is a vector of parameters. In both cases, consistent and efficient estimators for the parameters of interest are obtained by maximizing the log-likelihood function based on equation (9) with respect to θ and ϕ or α. The asymptotic distribution of the resulting ML estimators is similar to that given in equation (7) but with V defined as the information matrix, which corresponds simply to either A^{-1} or B.

3.3 *Two-part Models*

The parametric model described in the previous section is not defined at the boundary values of fractional response variables. Moreover, although most conditional mean models may be used in applications where some portion of the sample is at the extreme values of 0 and/or 1, this may not be the best option for modelling cases where the number of corner observations is large. For such cases, where the observations at one or both boundaries occur with too large a frequency than seems to be consistent with a simple model, a better approach may be the employment of two-part models, where the discrete component is modelled

as a binary or multinomial model and the continuous component as a fractional regression model.[3]

In this framework, three distinct situations may arise, depending on whether the outcomes are restricted to the intervals [0, 1], (0, 1] or [0, 1). In this paper, for expository purposes, we consider only the last case, but adapting the model discussed below for the other two cases is straightforward. We chose to focus our attention on the [0, 1) case because it is probably the most common one in economics. Indeed, most of the examples cited in the introduction of the paper, namely firm market share, proportion of debt in the financing mix of firms, fraction of land area allocated to agriculture and proportion of exports in total sales, may be modelled using the approach described next.

With two-part models for response variables observed on the interval [0, 1), the first part consists of a standard binary choice model and governs participation, i.e. the probability of observing a positive outcome. Define

$$y^* = \begin{cases} 0 & \text{for } y = 0 \\ 1 & \text{for } y \in (0, 1) \end{cases} \tag{12}$$

Then,

$$\Pr(y^* = 1|x) = E(y^*|x) = F(x\beta_{1P}) \tag{13}$$

where β_{1P} is a vector of variable coefficients and $F(\cdot)$ is usually one of the distribution functions described in Table 1. The resulting model may be estimated by ML using the whole sample.

The second part of the model governs positive choices, i.e. the magnitude of non-zero outcomes. In this case, a function similar to that defined in equation (4) is also a valid specification:

$$E[y|x, y \in (0, 1)] = M(x\beta_{2P}) \tag{14}$$

As before, $M(x\beta_{2P})$ may be estimated by QML or, if a conditional distribution is assumed for y, by ML. In both cases, estimation is based on the subsample that comprises only the individuals with positive outcomes. For simplicity, we assume that the same regressors appear in both parts of the model but this can relaxed and, in fact, should be if there are obvious exclusion restrictions.

Noting that $E(y|x)$ may be decomposed as

$$E(y|x) = E(y|x, y = 0) \cdot \Pr(y = 0|x) + E[y|x, y \in (0, 1)] \cdot \Pr[y \in (0, 1)|x]$$

and that the first term on the right-hand side of this expression is identically zero, the two-part model may be described simply by

$$\begin{aligned} E(y|x) &= E[y|x, y \in (0, 1)] \cdot \Pr[y \in (0, 1)|x] \\ &= M(x\beta_{2P}) \cdot F(x\beta_{1P}) \end{aligned} \tag{15}$$

where its two components are to be estimated separately. Naturally, misspecification of either $M(x\beta_{2P})$ or $F(x\beta_{1P})$ leads to misspecification of the conditional mean (15). Moreover, comparing equations (4) and (15) shows that one-part and two-part

decision mechanisms yield different functional forms for the conditional mean. Hence, using equation (4) overlooking the two-part decision mechanism produces a serious misspecification problem and leads to results that are of little use: the parameters θ appearing in equation (4) are a mixture of the parameters β_{1P} and β_{2P} in equation (15) and have no clear interpretation. A similar misspecification problem arises if the data are described by a one-part model and a two-part model is used.

From equation (15), we can calculate the effect on y of a unitary change in x_j:

$$\frac{\partial E(y|x)}{\partial x_j} = \frac{\partial M(x\beta_{2P})}{\partial x_j} F(x\beta_{1P}) + M(x\beta_{2P}) \frac{\partial F(x\beta_{1P})}{\partial x_j} \tag{16}$$

Thus, the total change in y can be disaggregated in two parts: (i) the change in y of those that have positive outcomes, weighted by the probability of having positive outcomes; and (ii) the change in probability of having positive outcomes, weighted by the expected value of y for those that have positive outcomes. This decomposition is similar to that found by McDonald and Moffitt (1980) for the Tobit model.

As both β_{1P} and β_{2P} are estimated separately, β_{1P} (β_{2P}) will have the typical asymptotic distribution of ML (QML or ML) estimators. See Ramalho and Silva (2009) and Cook *et al.* (2008) for empirical applications of the two-part fractional regression model using, respectively, conditional mean and beta models in the second part.

3.4 *One-part versus Two-part Models*

As discussed above, there are many examples of fractional data characterized by a large number of observations at zero. In such cases, practitioners have to decide whether one- or two-part models should be used. Clearly, this decision depends crucially on the interpretation placed upon the observed zeros. On the one hand, the zeros may be interpreted as the result from a utility maximizing or similar decision, in which case a one-part model is the appropriate model; for an example, see Wagner (2001), who argues that firms choose the profit-maximizing volume of exports, which might be zero or a positive quantity, and therefore uses a one-part model to explain the exports/sales ratio. In other cases, the zeros and the positive values may be best described by different mechanisms, in which case it is more reasonable to model separately the participation and the amount decisions using two-part models. For instance, consider the relationship between smoking and cigarettes price (Madden, 2008): while it is likely that some individuals decide not to smoke no matter how cheap cigarettes are, it is expected that for the subsample of smokers an increase in cigarette prices may lead to a reduction in the consumption of cigarettes.

In contrast to these examples, in many cases we cannot establish *a priori*, using only theoretical economic arguments, whether one- or two-part models should be used. That is, some of the competing theories may imply the use of one-part models, while others may favour the use of two-part models. For an example of such a case,

see the empirical application described in Section 6. Thus, in addition to the role of theoretical economic reasoning in deciding between one- and two-part models, it is essential to have available a set of statistical tests that might help discriminate between those models. However, to the best of our knowledge, the option between a single- and a two-part fractional regression model has never been tested. In the next section we propose various specification tests for fractional regression models, some of which may be used for choosing between one- and two-part models.

4. Specification Testing

The alternative estimators for fractional regression models described in the previous section are based on different assumptions. Next, we analyse several statistics for testing some of those assumptions and, thus, the statistical validity of those models. As all models require the correct specification of the conditional mean of y, we focus primarily on functional form tests, i.e. tests for assessing assumption (4) in one-part models and assumptions (13), (14) and (15) in two-part models. Note that, in spite of the functional form assumed for the conditional mean of y being the basic assumption of any fractional regression model, very rarely has it been tested in applied work. At the end of this section, tests for assessing the distributional assumptions made in the parametric beta regression model are also briefly discussed.

4.1 *Tests for Conditional Mean Assumptions*

In this section we propose four alternative classes of tests for assessing conditional mean assumptions. All the tests are valid for testing the functional form assumed for both one-part models and the two components of two-part models. Therefore, to simplify the exposition, below we focus on tests for H_0: $E(y|x) = G(x\theta)$, but their adaptation for testing H_0: $E(y^*|x) = F(x\beta_{1P})$ or H_0: $E[y|x, y \in (0, 1)] = M(x\beta_{2P})$ is straightforward. In addition, we show that one of the tests suggested may also be adapted for testing the full specification of two-part models, H_0: $E(y|x) = M(x\beta_{2P})\cdot F(x\beta_{1P})$.

The four classes of tests discussed below are the following: (i) RESET-type tests, where polynomials in the fitted $x\theta$ values are included in $G(\cdot)$ to detect general kinds of functional form misspecification; (ii) goodness-of-link tests, which are based on generalized link functions that incorporate one or more of the links associated with the competing $G(\cdot)$ functions as particular cases; (iii) goodness-of-functional-form tests, based on generalized functional forms, which encompass $G(\cdot)$ as a special case; and (iv) generic non-nested tests, where the alternative competing specifications for $G(\cdot)$ are tested one against the others and which may also be used for testing the full specification of two-part models. To the best of our knowledge, only the RESET test has already been applied in the framework of fractional regression models.

Below, we provide an integrated approach for all tests, implementing all of them as LM statistics for omitted variables, which are calculated using simple artificial

regressions. Therefore, before presenting each test in detail, we first discuss the general form of those artificial regressions.

4.1.1 *Artificial Regressions for LM Test Statistics*

All the four classes of conditional mean tests suggested in this paper may be interpreted as tests for the omission of a J-dimensional vector z in the model $E(y|x, z) = G(x\theta + z\gamma)$, where γ is the vector of parameters associated with z and $G(\cdot)$ is the postulated functional form. Under the null hypothesis H_0: $\gamma = 0$, z is not relevant and $G(x\theta)$ is an appropriate specification for $E(y|x)$. As we show below, the only thing that distinguishes each one of the LM tests proposed is the composition of the vector z. To test for H_0: $\gamma = 0$, all the LM tests may be evaluated with NLS, QML or ML estimators and have a χ^2_J distribution. According to the estimator considered, a different artificial regression should be used to compute the tests.

For the case of ML estimation of the binary component of two-part models, Davidson and MacKinnon (1984) show that an LM statistic for the omission of z with good small-sample properties may be simply computed as LM = ESS, where ESS is the explained sum of squares of the auxiliary regression

$$\tilde{u} = \tilde{g}x^*\delta + \text{error} \tag{17}$$

$\tilde{u} = \hat{u}\hat{\omega}$, $\tilde{g} = \hat{g}\hat{\omega}$, $\hat{\omega} = [\hat{G}(1 - \hat{G})]^{-0.5}$, a circumflex indicates evaluation under H_0 at $\hat{\varphi} = (\hat{\theta}, 0)$ and $x^* = (x', z')$. This artificial regression may also be used for testing the functional form of one-part or the second component of two-part fractional regression models by computing LM $= nR^2$, where R^2 is the constant-unadjusted R^2 from regression (17), if assumption (16) of Papke and Wooldridge (1996) is satisfied, i.e. $\text{Var}(y|x, x) = \lambda G(x\theta)[1 - G(x\theta)], \lambda > 0$; see also Wooldridge (1991a). This is the case when these tests are evaluated with ML estimators based on the beta distribution, under which $\lambda = (1 + \phi)^{-1}$; see equation (11).

Evaluating the tests with ML estimators based on the beta model has the drawback of requiring a particular heteroskedasticity assumption for the conditional variance of y. If this assumption fails, the tests may lead to the rejection of H_0 even though $E(y|x)$ is correctly specified. Therefore, in general, it is preferable to evaluate the tests with QML or NLS estimators and compute heteroskedasticity-robust LM statistics. In the former case, the tests may be calculated as LM = ESS $= N -$ SSR, where SSR is the sum of squared residuals from the artificial regression

$$1 = \tilde{u}\tilde{r}_1\delta_1 + \tilde{u}\tilde{r}_2\delta_2 + \cdots + \tilde{u}\tilde{r}_J\delta_J + \text{error} \tag{18}$$

and $\tilde{r}_j$ are the residuals that result from regressing each element $\tilde{g}z_j$, $j = 1, \ldots, J$, on the entire vector $\tilde{g}x$; see Wooldridge (1991a, b) and Papke and Wooldridge (1996) for details. In the NLS case, a similar computation may be used for the LM statistic but based on the artificial regression

$$1 = \hat{u}\hat{r}_1\delta_1 + \hat{u}\hat{r}_2\delta_2 + \cdots + \hat{u}\hat{r}_J\delta_J + \text{error} \tag{19}$$

which differs from the previous one by setting $\hat{\omega} = 1$; see Wooldridge (2002, p. 368).

4.1.2 *RESET-type Tests*

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. Indeed, using standard approximation results for polynomials, it can be shown that any index model of the form $E(y|x) = H(x\theta)$ can be arbitrarily approximated by $G(x\theta + \sum_{j=1}^{J} \gamma_j (x\theta)^{j+1})$ for J large enough. Therefore, testing the hypothesis $E(y|x) = G(x\theta)$ is equivalent to testing for $\gamma = 0$ in the augmented model $E(y|x, z) = G(x\theta + z\gamma)$, where $z = [(x\hat{\theta})^2, \ldots, (x\hat{\theta})^{J+1}]$. The first few terms in the expansion are the most important and, in practice, only the quadratic and cubic terms are usually considered.

4.1.3 *Goodness-of-link Tests*

Testing the functional form $G(\cdot)$ is equivalent to testing the so-called link function. The link function, from now on denoted by $h(\cdot)$, is a widely used concept in the generalized linear models (GLM) literature, and may be simply defined as the function that relates the linear predictor $x\theta$ to the conditional expected value $\mu = E(y|x)$, i.e. $h(\mu) = x\theta$; see McCullagh and Nelder (1989) for details. Thus, to each particular link function $h_A(\mu)$ corresponds a different functional form $G_A(x\theta)$ and vice versa. The link functions for the cauchit, logit, probit, loglog and complementary loglog functional forms are given in Table 1.

In the GLM framework, the most common approach to test the adequacy of a given link function involves the construction of a generalized link function indexed by some vector of parameters α, which includes the hypothesized link function as a special case for some specific values of α. Following Pregibon (1980), let $h(\mu; \alpha)$ be a generalized link function that embeds both the hypothesized link, $h_A(\mu) = h(\mu; \alpha_A)$, and the (unknown) true link, $h_0(\mu) = h(\mu; \alpha_0)$. A first-order Taylor series expansion of $h(\mu; \alpha_0)$ around α_A yields the approximation

$$h(\mu; \alpha_0) \approx h(\mu; \alpha_A) + \nabla_\alpha h(\mu; \alpha_A)(\alpha_0 - \alpha_A) \tag{20}$$

Replacing the correct link function $h(\mu; \alpha_0)$ by $x\theta$ and solving for $h(\mu; \alpha_A)$ gives rise to the following approximation for the postulated link:

$$h_A(\mu) = x\theta + z\gamma \tag{21}$$

where $\gamma = \alpha_A - \alpha_0$ and

$$z = \nabla_\alpha h(\mu; \alpha_A) \tag{22}$$

which are usually known in the GLM literature as carrier functions. If the assumed link function is correct, then $\alpha_A = \alpha_0$ and $\gamma = 0$.

So far, goodness-of-link tests for H_0: $\gamma = 0$ have been directly based on equation (21) and used exclusively with binary models estimated by ML. Here, in order to allow straightforward computation of robust versions based on NLS or QML estimation, we suggest the implementation of those tests using the LM statistics outlined above. That is, instead of working directly with equation (21), we test for the relevance of z in the generalized functional form $E(y|x, z) = G_A(x\theta + z\gamma)$ that corresponds to the approximate link function (21).

The GLM literature provides many alternative generalized link functions, especially for the logit model. In this case, we have available, among others, the generalizations proposed by Prentice (1976), Pregibon (1980), Aranda-Ordaz (1981), Whittemore (1983), Stukel (1988) and Czado (1994). In contrast, only a few generalizations for the other specifications analysed in this paper have been proposed so far, such as Stukel's (1988) model that also encompasses the probit, loglog and complementary loglog links, and Koenker and Yoon's (2009) augmented model that nests the cauchit link. In the simulation study of Section 5, we merely consider tests based on Stukel's (1988) and Koenker and Yoon's (2009) generalized link functions, since the former is the most encompassing one and the latter is the only one that allows the assessment of cauchit models. The carrier functions for these tests for the cauchit, logit, probit, loglog and complementary loglog specifications are, respectively,

$$
\begin{aligned}
z &= \nabla_\alpha cdf(student_\alpha)^{-1}(\mu)|_{\alpha=1} \\
z &= \left[0.5(x\hat{\theta})^2 I_{(x\hat{\theta}\geq 0)}; -0.5(x\hat{\theta})^2 I_{(x\hat{\theta}\leq 0)}\right] \\
z &= \left[\frac{1}{0.165}\left(\ln\frac{\mu}{1-\mu}e^{-0.165x\hat{\theta}} - x\hat{\theta}\right) I_{(x\theta\geq 0)};\right. \\
&\quad \left.\frac{1}{0.165}\left(\ln\frac{\mu}{1-\mu}e^{0.165x\hat{\theta}} + x\hat{\theta}\right) I_{(x\theta\leq 0)}\right] \\
z &= \left\{-\frac{1}{0.037}\left[\ln\frac{\mu}{1-\mu}(1+0.037x\hat{\theta}) - x\hat{\theta}\right] I_{(x\theta\geq 0)};\right. \\
&\quad \left.\frac{1}{0.620}\left(\ln\frac{\mu}{1-\mu}e^{0.620x\hat{\theta}} + x\hat{\theta}\right) I_{(x\theta\leq 0)}\right\} \\
z &= \left\{\frac{1}{0.620}\left(\ln\frac{\mu}{1-\mu}e^{-0.620x\hat{\theta}} - x\hat{\theta}\right) I_{(x\theta\geq 0)};\right. \\
&\quad \left.-\frac{1}{0.037}\left[\ln\frac{\mu}{1-\mu}(1-0.037x\hat{\theta}) + x\hat{\theta}\right] I_{(x\theta\leq 0)}\right\}
\end{aligned}
$$

4.1.4 *Goodness-of-functional-form Tests*

While each goodness-of-link test is valid for testing the functional form of *particular* fractional regression models, the two tests that we propose next may be applied to test the specification of *any* model. As the new tests are based on direct generalizations of $G(x\theta)$, we call them 'goodness-of-functional-form tests',

although they may also be interpreted as goodness-of-link tests, as shown next. We first present the two generalized functional forms proposed and discuss briefly their characteristics, and then derive the corresponding link functions. Following the approach of the previous section, we obtain the resulting carrier functions z in equation (22), which in this case can be substantially simplified.

The first generalized functional form proposed extends for other models a generalization of the type that is usually employed to introduce asymmetry in the logit model, which consists simply in raising the logit functional form to a positive constant α. See *inter alia* Poirier (1980), Smith (1989) and Nagler (1994), who called the resulting model *generalized logistic model*, *Burrit model* and *scobit model*, respectively. In this paper we propose applying this extension of the logit model to any functional form $G(\cdot)$:

$$E(y|x) = G(x\theta)^{\alpha} \tag{23}$$

where $\alpha > 0$ such that $0 < E(y|x) < 1$. As equation (23) describes only some particular forms of asymmetry, we also propose the alternative specification

$$E(y|x) = 1 - [1 - G(x\theta)]^{\alpha} \tag{24}$$

where the form of asymmetry is complementary.

Figure 1 contains representations of both equations (23) and (24) for several values of α for the logit and loglog cases. In equation (23) the curve of the functional form is shifted to the right and to the left for $\alpha > 1$ and $0 < \alpha < 1$, respectively, the impact being more substantial on the left tail. It is clear that the behaviour of the curves described by equation (24) is complementary to that of equation (23). As both equations (23) and (24) reduce to $G(\cdot)$ when $\alpha = 1$, testing whether $G(x\theta)$ is the correct specification of $E(y|x)$ corresponds to testing for H_0: $\alpha = 1$ in both cases.

Models (23) and (24) give rise to the generalized link functions $h(\mu;\alpha) = h(\mu^{1/\alpha})$ and $h(\mu;\alpha) = h[1 - (1 - \mu)^{1/\alpha}]$, respectively, for $\mu = E(y|x)$. Using the procedures described in the previous section, two new goodness-of-link tests for checking the relevancy of carriers z given in equation (22) may be straightforwardly derived. In this case, as we are testing for $\alpha = 1$, those carriers may be greatly simplified, not involving the calculation of link functions or their derivatives. Define μ^* as the argument of $h(\mu;\alpha)$ such that $h(\mu;\alpha) = h(\mu^*)$, where $\mu^* = \mu^{1/\alpha}$ and $\mu^* = 1 - (1 - \mu)^{1/\alpha}$ for models (23) and (24), respectively. As the carriers (22) may be written as $z = \nabla_{\mu^*} h(\mu^*)|_{\alpha=1} \nabla_{\alpha} \mu^*|_{\alpha=1}$ and it is straightforward to show that $\nabla_{\mu^*} h(\mu^*)|_{\alpha=1} = [\nabla_{x\theta} \mu|_{\alpha=1}]^{-1} = \hat{g}^{-1}$ and $\nabla_{\alpha} \mu^*|_{\alpha=1} = \nabla_{\alpha} \mu|_{\alpha=1}$, they can be simplified to

$$z = \nabla_{\alpha} \mu|_{\alpha=1} \hat{g}^{-1} \tag{25}$$

where $\nabla_{\alpha} \mu|_{\alpha=1} = \hat{G} \ln(\hat{G})$ and $\nabla_{\alpha} \mu|_{\alpha=1} = -(1 - \hat{G}) \ln(1 - \hat{G})$ for tests based on, respectively, equations (23) and (24).

Analysing the structure of equation (25), it is clear that among the functional forms considered in this paper and described in Table 1, the tests based on equations (23) and (24) cannot be applied to, respectively, loglog and complementary

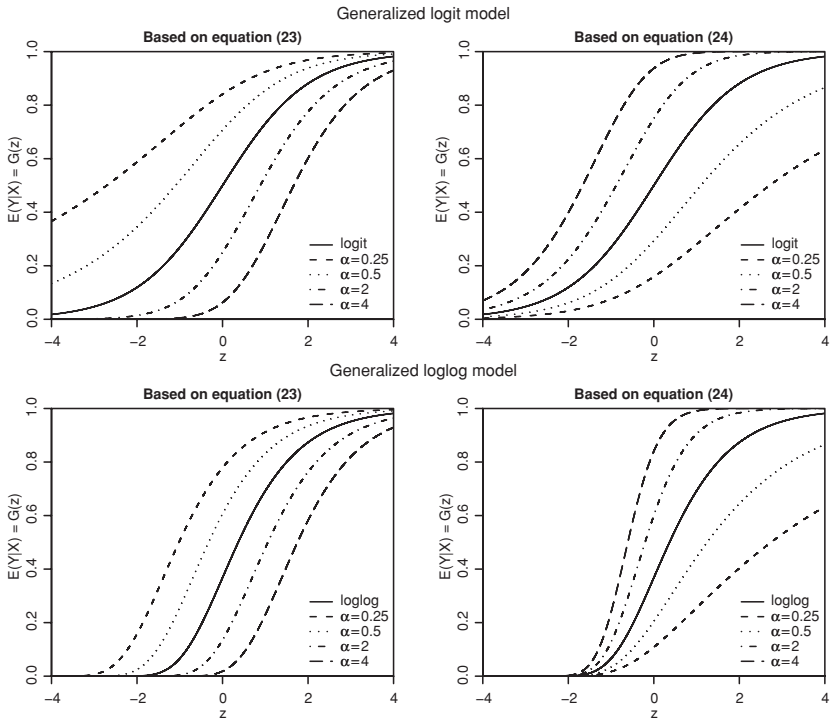

Figure 1. Generalized Functional Form for $E(Y|X)$.

loglog models when the index includes a constant term. Indeed, in such cases $z = -1$, since $\hat{g} = e^{-x\hat{\theta}}e^{-e^{-x\hat{\theta}}}$ and $\hat{g} = e^{x\hat{\theta}}e^{-e^{x\hat{\theta}}}$ equal, respectively, $-\hat{G}\ln\hat{G}$ and $(1-\hat{G})\ln(1-\hat{G})$.

When a logit specification is used for $G(x\theta)$, the carrier functions z used separately by our two tests coincide with the two carrier functions that define Prentice's (1976) goodness-of-link test for logit models, which were derived from a generalized link function indexed by two additional parameters (m_1, m_2). Actually, in the logit case, Prentice's (1976) generalized link function incorporates as special cases both (23), for $m_2 = 1$, and (24), for $m_1 = 1$. Therefore, on the one hand, Prentice's (1976) approach may be seen as a generalization of ours in the logit case and, on the other hand, his approach is more limited since, unlike ours (with the two exceptions already referred to), it cannot be easily applied to other possible specifications for $G(x\theta)$.

4.1.5 *P Test for Non-nested Hypotheses*

As the alternative functional forms available for fractional regression models are non-nested, the various test procedures for non-nested regression models proposed in the econometric literature can be used to test alternative competing

specifications for $E(y|x)$. Here, we focus on the P test statistic proposed by Davidson and MacKinnon (1981), which is probably the simplest way of comparing nonlinear regression models; see *inter alia* Gouriéroux and Monfort (1994) for other alternatives. As far as we know, the P test has never been applied in a context similar to ours before; see, however, the recent paper by Santos Silva *et al.* (2008) for a related approach.

Suppose that $G(x\theta)$ and $T(x\eta)$ are admissible functional forms for $E(y|x)$ and assume homoskedasticity and NLS estimation. In this framework, as shown by Davidson and MacKinnon (1981), testing H_0: $G(x\theta)$ against H_1: $T(x\eta)$, i.e. checking whether $G(x\theta)$ is an appropriate specification for $E(y|x)$ after taking into account the information provided by the alternative model, is equivalent to testing the null hypothesis H_0: $\delta_2 = 0$ in the auxiliary regression

$$(y - \hat{G}) = \hat{g}x\delta_1 + \delta_2(\hat{T} - \hat{G}) + \text{error} \tag{26}$$

where δ_2 is a scalar parameter and the circumflex means evaluation with the NLS estimators $\hat{\theta}$ or $\hat{\eta}$, which are obtained by estimating separately the models defined by $G(\cdot)$ and $T(\cdot)$, respectively. To test H_0: $T(x\eta)$ against H_1: $G(x\theta)$, we need to use another P statistic, which is calculated using a similar auxiliary regression to equation (26) but with the roles of the two models interchanged. Comparing equations (17) and (26), we see that testing for H_0: $\delta_2 = 0$ in the latter equation corresponds to testing for the relevance of $z = (\hat{T} - \hat{G})\hat{g}^{-1}$ in $G(x\theta + z\gamma)$. With fractional regression models, which are typically heteroskedastic and usually are estimated by QML or ML, it is in general preferable to test the relevance of this z variable as explained in Section 4.1.1.

In contrast to the previous classes of tests, which may only be applied to assess the correctness of the functional form assumed in one-part models or in the two separate components of two-part models, the P test may also be applied to test the full specification of two-part models, $E(y|x) = M(x\beta_{2P}) \cdot F(x\beta_{1P})$, against both one-part models, $E(y|x) = G(x\theta)$, and other two-part models, say $E(y|x) = Q(x\rho_{2P}) \cdot S(x\rho_{1P})$, and vice versa. To check whether $E(y|x) = G(x\theta)$ is appropriate after taking into account the information provided by the alternative $E(y|x) = M(x\beta_{2P}) \cdot F(x\beta_{1P})$ and vice versa, the artificial regression (26) must be re-expressed as

$$(y - \hat{G}) = \hat{g}x\delta_1 + \delta_2(\hat{M} \cdot \hat{F} - \hat{G}) + \text{error} \tag{27}$$

and

$$(y - \hat{M} \cdot \hat{F}) = \hat{m}\hat{F}x\delta_{11} + \hat{M}\hat{f}x\delta_{12} + \delta_2(\hat{G} - \hat{M} \cdot \hat{F}) + \text{error} \tag{28}$$

respectively, where $\hat{f}$ and $\hat{m}$ are the partial derivatives of F and M with respect to, respectively, β_{1P} and β_{2P}. Similarly, to check whether $E(y|x) = Q(x\rho_{2P}) \cdot S(x\rho_{1P})$ is appropriate after taking into account the information provided by the alternative $E(y|x) = M(x\beta_{2P}) \cdot F(x\beta_{1P})$ and vice versa, the artificial regressions of interest are

$$(y - \hat{Q} \cdot \hat{S}) = \hat{q}\hat{S}x\delta_{11} + \hat{Q}\hat{s}x\delta_{12} + \delta_2(\hat{M} \cdot \hat{F} - \hat{Q} \cdot \hat{S}) + \text{error} \tag{29}$$

and

$$(y - \hat{M} \cdot \hat{F}) = \hat{m}\hat{F}x\delta_{11} + \hat{M}\hat{f}x\delta_{12} + \delta_2(\hat{Q} \cdot \hat{S} - \hat{M} \cdot \hat{F}) + \text{error} \qquad (30)$$

respectively, where $\hat{q}$ and $\hat{s}$ are the partial derivatives of Q and S with respect to, respectively, ρ_{2P} and ρ_{1P}.

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 is also important to examine whether the beta distribution is appropriate for modelling the fractional response variable in order to obtain efficient ML estimators. The standard test for misspecification of a parametric likelihood function is the information matrix test introduced by White (1982), which, however, can be very burdensome to compute. Moreover, the simplified outer product of gradients (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 information matrix test have been proposed and most authors advocate the use of bootstrap-based critical values. The investigation of the performance of alternative information matrix tests in the framework of the beta fractional regression model would deserve a paper on its own and hence we do not pursue this line of research here. In the empirical application carried out later, we use the bootstrapped OPG information matrix test analysed by Horowitz (1994), which he found to work very well in Tobit and binary probit models. In our case, in each bootstrap replication we generate values for y by random sampling from the beta distribution based on the actual values of x and the ML parameter estimates from the actual sample.

5. Monte Carlo Simulation Study

In this section we investigate the finite-sample performance of most of the estimators and tests discussed throughout this paper in a Monte Carlo simulation study. All experiments consider a single covariate X_1 generated from the normal distribution with mean zero and variance 1 and are based on 10,000 replications, which were performed using the *R* software.

5.1 *Performance of Alternative Estimation Methods*

In our first set of experiments, we compare the performance of three alternative estimation methods (NLS, QML and ML) in terms of bias and precision under the assumptions that both the functional form of the conditional mean and the distribution of y given x were correctly specified by the analyst. We consider four different functional forms (cauchit, logit, probit and loglog) for the conditional mean of the response variable and generate samples of $N = \{100, 200, 500, 1000\}$ according to the beta distribution.

In order to mimic a wide class of data sets that may be available for empirical work, for each functional form assumed for $E(y|x)$ we simulated samples

characterized by different means, variances and levels of asymmetry; see the histograms in Ramalho *et al.* (2009) for samples of 500,000 observations generated using the true value of the parameters of interest $\theta = (\theta_0, 0.5)$ and the shape parameter ϕ. The distribution of the data is approximately symmetric for $\theta_0 = 0$, apart from the loglog case, and is clearly asymmetric for the other values considered for θ_0. Increasing ϕ, the variance of y given x is reduced as well as the weight of observations with small values of y, which makes the distribution of y range from U-shaped to an inverted U-shaped curve for $\theta_0 = 0$. For loglog models, instead of fixing $\theta_0 = -1$ as in the other models, we considered $\theta_0 = -0.5$, so that the distribution of y given x was more similar to that of the other models.

Table 2 reports the mean and standard deviation across replications of the alternative estimators of θ_1 for $N = \{200, 500\}$, while Figure 2 displays the root mean squared error (RMSE) of those estimators for a shape parameter ϕ ranging from 0.5 to 20 for $N = \{100, 500\}$. We find that, in terms of bias, the three estimators displayed a very similar performance, being in general approximately unbiased. In all cases, the ML estimator presents the smallest standard deviation and RMSE, while the QML estimator is clearly more precise than the NLS estimator. However, Figure 2 suggests that those differences vanish as N and ϕ increase.

Next, we generate the response variable according to the simplex distribution.[4] The shape parameter of this distribution was chosen in order to produce a similar range of distributions to those obtained for the beta case. In particular, the means and variances are identical to those simulated before. As the generation of simplex-distributed data is very time-consuming, we considered only the logit case. Table 3, which displays various summary statistics for each estimator, clearly shows that the performance of the QML and NLS estimators hardly changes relative to that documented in Table 2. In contrast, the ML estimator based on the beta distribution is no longer unbiased. In fact, despite the well-known ability of the beta distribution to describe a variety of shapes, as this distribution does not belong to the LEF, the beta ML estimators are not robust to deviations from the assumed distribution. However, since in most cases its standard deviation is again the lowest, the ML estimator still displays the smallest RMSE in some cases for $N = 200$. This advantage of the ML estimator seems to disappear as N increases, since for $N = 500$ its bias remains approximately unchanged, while the dispersion of all estimators gets closer.

5.2 *Effects of the Misspecification of the Conditional Mean*

To analyse the effects of the misspecification of the conditional mean, we focus on QML estimators, since they do not require distributional assumptions and performed better than NLS estimators in the former experiments. Moreover, as the distribution assumed for the data is irrelevant in QML estimation, we merely generate response data from the beta distribution. As the QML estimators for θ are not directly comparable for the four different forms of $E(y|x)$ under analysis, we measure the effects of assuming a misspecified functional form by comparing the partial effects

Table 2. Monte Carlo Parameter Estimates for θ_1 (Beta-distributed Response Variable; $\theta_1 = 0.5$).

		$N = 200$						$N = 500$					
		NLS		QML		Beta-ML		NLS		QML		Beta-ML	
θ_0	ϕ	Mean	St. dev.	Mean	St. dev.	Mean	St. dev.	Mean	St. dev.	Mean	St. dev.	Mean	St. dev.
							Link function: cauchit						
0	1	0.511	0.117	0.510	0.115	0.505	0.096	0.503	0.071	0.503	0.071	0.501	0.060
	2.5	0.508	0.086	0.507	0.085	0.505	0.079	0.503	0.054	0.503	0.053	0.502	0.050
	5	0.504	0.066	0.504	0.065	0.503	0.063	0.502	0.041	0.502	0.041	0.502	0.040
	20	0.502	0.035	0.501	0.035	0.501	0.035	0.501	0.022	0.501	0.022	0.501	0.022
−1	1	0.514	0.147	0.512	0.142	0.509	0.111	0.504	0.091	0.504	0.089	0.503	0.071
	5	0.504	0.083	0.504	0.081	0.504	0.077	0.502	0.052	0.502	0.051	0.502	0.048
	10	0.503	0.061	0.502	0.060	0.503	0.058	0.501	0.038	0.501	0.037	0.501	0.036
	40	0.501	0.031	0.501	0.031	0.501	0.031	0.500	0.020	0.500	0.019	0.500	0.019
							Link function: logit						
0	1	0.506	0.113	0.506	0.112	0.500	0.092	0.502	0.069	0.502	0.068	0.499	0.056
	2.5	0.502	0.084	0.502	0.083	0.500	0.077	0.501	0.053	0.501	0.052	0.500	0.048
	5	0.502	0.064	0.502	0.063	0.500	0.061	0.501	0.040	0.501	0.040	0.500	0.039
	20	0.501	0.034	0.501	0.034	0.501	0.034	0.501	0.022	0.501	0.021	0.500	0.021
−1	1	0.506	0.127	0.505	0.122	0.500	0.088	0.502	0.079	0.501	0.076	0.499	0.054
	5	0.503	0.073	0.502	0.070	0.501	0.063	0.500	0.046	0.500	0.044	0.499	0.040
	10	0.501	0.053	0.501	0.052	0.501	0.049	0.501	0.034	0.501	0.033	0.500	0.031
	40	0.500	0.028	0.500	0.027	0.500	0.026	0.500	0.017	0.500	0.017	0.500	0.017

Journal of Economic Surveys (2011) Vol. 25, No. 1, pp. 19–68

Table 2. *Continued.*

		$N = 200$						$N = 500$					
		NLS		QML		Beta-ML		NLS		QML		Beta-ML	
θ_0	ϕ	Mean	St. dev.	Mean	St. dev.	Mean	St. dev.	Mean	St. dev.	Mean	St. dev.	Mean	St. dev.
							Link function: probit						
0	1	0.505	0.076	0.504	0.073	0.495	0.052	0.502	0.047	0.502	0.045	0.496	0.032
	2.5	0.503	0.057	0.502	0.055	0.499	0.046	0.501	0.035	0.500	0.034	0.499	0.028
	5	0.502	0.043	0.502	0.042	0.500	0.038	0.501	0.027	0.501	0.026	0.500	0.024
	20	0.501	0.023	0.501	0.022	0.500	0.022	0.500	0.014	0.500	0.014	0.500	0.014
−1	1	0.509	0.100	0.505	0.085	0.498	0.042	0.503	0.062	0.502	0.054	0.499	0.026
	5	0.503	0.057	0.502	0.049	0.500	0.036	0.501	0.036	0.500	0.031	0.500	0.022
	10	0.502	0.041	0.501	0.036	0.500	0.030	0.500	0.026	0.500	0.023	0.500	0.019
	40	0.500	0.022	0.500	0.019	0.500	0.018	0.500	0.014	0.500	0.012	0.500	0.011
							Link function: loglog						
0	1	0.506	0.083	0.505	0.076	0.496	0.046	0.502	0.051	0.502	0.047	0.498	0.028
	2.5	0.503	0.061	0.503	0.057	0.497	0.041	0.501	0.039	0.501	0.036	0.499	0.025
	5	0.502	0.047	0.502	0.044	0.499	0.035	0.501	0.029	0.501	0.027	0.499	0.022
	20	0.501	0.025	0.501	0.023	0.500	0.022	0.500	0.016	0.500	0.015	0.500	0.014
−0.5	1	0.509	0.093	0.507	0.077	0.490	0.032	0.504	0.057	0.504	0.048	0.491	0.020
	5	0.502	0.052	0.502	0.044	0.496	0.027	0.501	0.034	0.501	0.028	0.497	0.016
	10	0.501	0.039	0.501	0.032	0.498	0.023	0.501	0.025	0.501	0.021	0.498	0.014
	40	0.500	0.020	0.500	0.017	0.500	0.015	0.500	0.013	0.500	0.011	0.500	0.009

Journal of Economic Surveys (2011) Vol. 25, No. 1, pp. 19–68

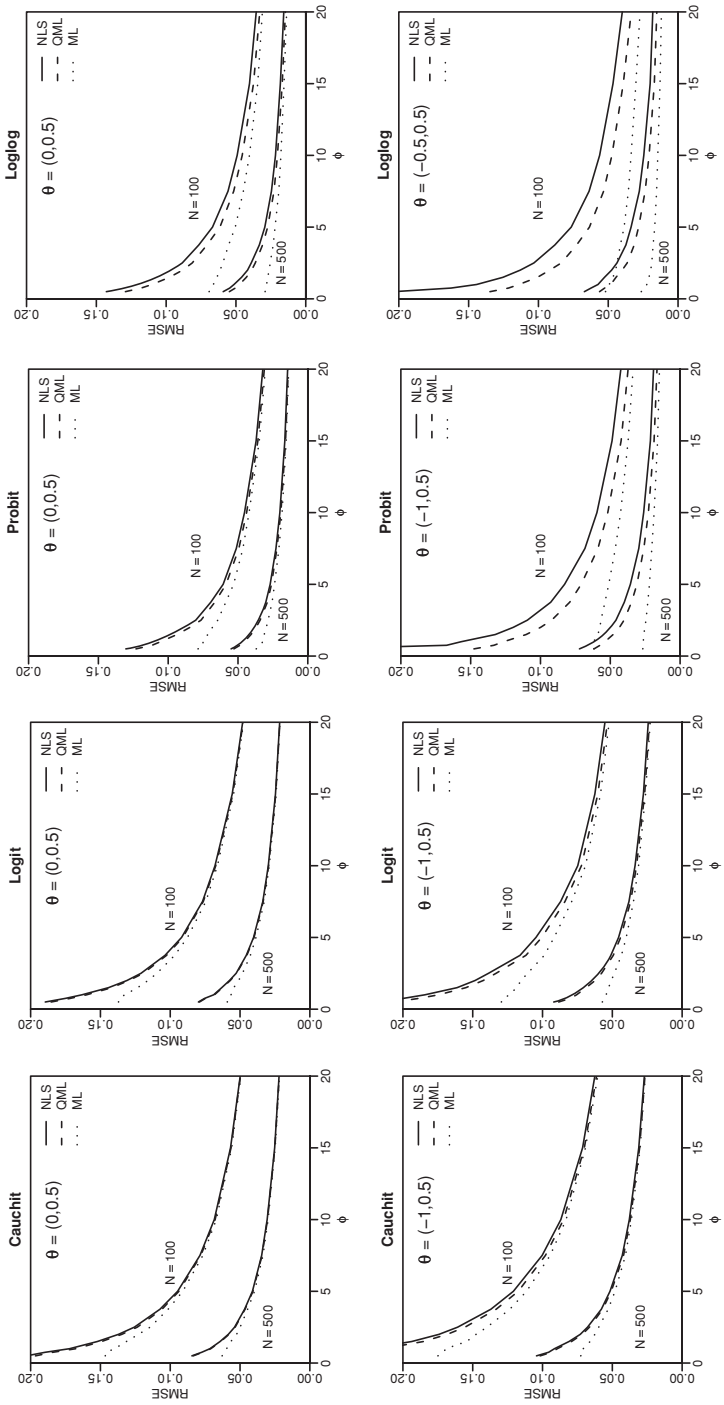

Figure 2. RMSE Comparison of Alternative Estimators for Fractional Regression Models (Beta-distributed Response Variable).

Table 3. Monte Carlo Parameter Estimates for θ_1 (Simplex-distributed Response Variable; Logit Model; $\theta_1 = 0.5$).

		NLS				QML				Beta-ML			
θ_0	ϕ	Mean	Median	St. dev.	RMSE	Mean	Median	St. dev.	RMSE	Mean	Median	St. dev.	RMSE
							$N = 200$						
0	48	0.502	0.503	0.117	0.117	0.501	0.503	0.114	0.114	0.411	0.413	0.079	0.119
	12	0.504	0.503	0.082	0.082	0.503	0.503	0.080	0.080	0.456	0.457	0.066	0.079
	4.8	0.501	0.501	0.066	0.066	0.500	0.502	0.064	0.064	0.476	0.478	0.059	0.063
	1	0.501	0.501	0.032	0.032	0.501	0.501	0.031	0.031	0.495	0.495	0.031	0.031
−1	69	0.503	0.501	0.135	0.135	0.499	0.500	0.126	0.126	0.325	0.326	0.078	0.192
	6.5	0.502	0.502	0.075	0.075	0.501	0.501	0.069	0.069	0.430	0.429	0.060	0.092
	3	0.499	0.501	0.066	0.066	0.498	0.500	0.061	0.061	0.456	0.458	0.057	0.072
	0.6	0.501	0.501	0.028	0.028	0.501	0.500	0.025	0.025	0.490	0.490	0.025	0.027
							$N = 500$						
0	48	0.493	0.499	0.091	0.091	0.493	0.499	0.090	0.090	0.405	0.411	0.066	0.116
	12	0.501	0.501	0.056	0.056	0.500	0.501	0.055	0.055	0.454	0.455	0.046	0.065
	4.8	0.497	0.500	0.056	0.056	0.496	0.500	0.055	0.055	0.473	0.477	0.051	0.058
	1	0.501	0.501	0.021	0.021	0.501	0.500	0.020	0.020	0.495	0.495	0.020	0.021
−1	69	0.489	0.497	0.106	0.106	0.488	0.497	0.102	0.103	0.318	0.324	0.064	0.193
	6.5	0.500	0.501	0.054	0.054	0.499	0.501	0.050	0.050	0.427	0.429	0.043	0.085
	3	0.492	0.500	0.065	0.066	0.492	0.500	0.064	0.064	0.450	0.457	0.059	0.077
	0.6	0.500	0.500	0.022	0.022	0.500	0.500	0.020	0.020	0.489	0.490	0.020	0.023

computed both for the model used for generating the data and for the other three (misspecified) models.

The partial effects of a covariate x_j on the outcome are given by $g(x\hat{\theta})\hat{\theta}_j$ (see Table 1), and their mean across replications is represented in Figure 3 for $N = 500$ and $\phi = 5$ for the $\{0, 0.02, 0.04, \ldots, 0.98, 1\}$ population quantiles of X_1. Clearly, apart from cases where a logit functional form is used in estimation but data are generated according to the probit model and vice versa, misspecification of the functional form may produce very important distortions in the estimation of partial effects. In particular, note that the deviations between the partial effects estimated by cauchit and loglog models may be tremendous. Nevertheless, the direction of the partial effects is always correctly estimated.

In addition to measuring partial effects for individuals with specific characteristics, as we did in Figure 3, in empirical work it is customary to present also the average response of all individuals, $N^{-1}\hat{\theta}_j \sum_{i=1}^{N} g(x_i\hat{\theta})$, and the response of the average individual, $g(\bar{x}\hat{\theta})\hat{\theta}_j$, where $\bar{x}$ denotes the mean of the covariates. In Table 4 we report the results obtained for $N = 500$ and $\phi = 5$. The values underlined denote the partial effect estimated for the true models. In the case of the response of the average individual, we achieve similar conclusions to those of Figure 3; i.e. the bias can be very large in some cases. For example, when $\theta = (-0.5, 0.5)$ and the loglog model is used to generate the data, the biases of the partial effects estimated according to the cauchit, logit and probit model are, respectively, 41.5%, 11.9% and 7.5%. In contrast, the estimation of average sample effects seems to be much more robust to misspecification of the functional form, especially when logit or probit models are employed. Indeed, in these experiments the bias for these two models is always less than 1%, while the maximum bias for the loglog and cauchit models is, respectively, 3.5% and 10.0%.

5.3 *Tests for the Functional Form When There Are No Boundary Observations*

Given the results of the previous section, the selection of the correct functional form for the conditional mean of y is clearly a very relevant issue in modelling fractional data. Therefore, next we investigate the finite-sample properties of the four classes of tests for the conditional mean assumptions discussed before. In particular, we compute (i) two versions of the RESET test, RESET2 and RESET3, which are based on the addition of, respectively, two and three powers of $x\hat{\theta}$, the first being the most widely used in empirical work and the second version automatically calculated by the package Stata in linear models; (ii) a goodness-of-link test (GOL), either that of Koenker and Yoon (2009) for cauchit models or that of Stukel (1988) for logit, probit and loglog models; (iii) the two tests for the goodness-of-functional-form proposed in this paper based on the general functional forms (23) and (24), which are designated, respectively, GOFF1 and GOFF2; and (iv) three non-nested P tests, which differ only in the alternative model considered for testing each null hypothesis. We use the same design as in previous sections and, again, focus on QML estimation and beta-distributed response variables.

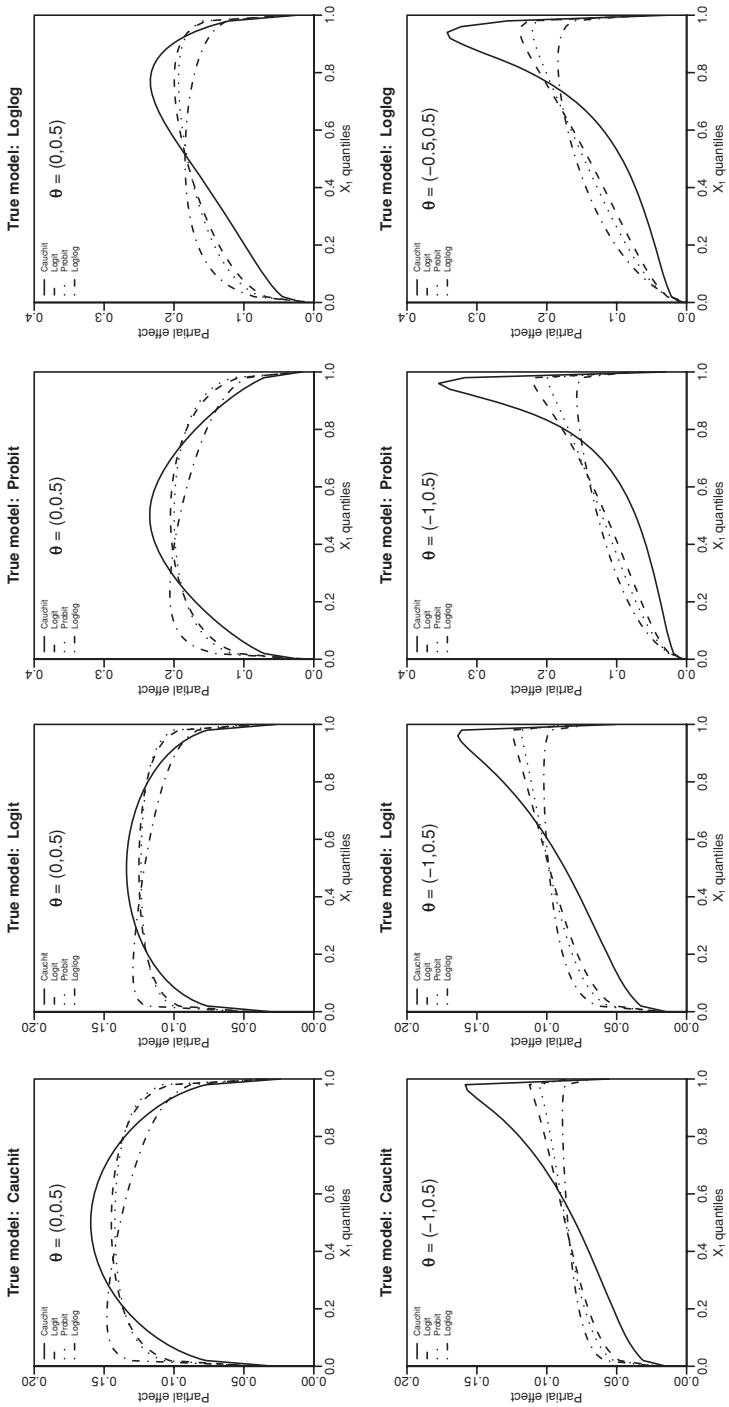

Figure 3. Partial Effects (Beta-distributed Response Variable; $N = 500$; $\phi = 5$).

Table 4. Monte Carlo Partial Effects (Beta-distributed Response Variable; $N = 500$; $\phi = 0.5$).

	Average response of all individuals				Response of the average individual			
True model	Cauchit	Logit	Probit	Loglog	Cauchit	Logit	Probit	Loglog
				$\theta = (0, 0.5)$				
Cauchit	0.134	0.134	0.134	0.132	0.159	0.145	0.142	0.138
	(0.008)	(0.008)	(0.008)	(0.008)	(0.013)	(0.010)	(0.010)	(0.009)
Logit	0.117	0.118	0.118	0.117	0.134	0.125	0.124	0.121
	(0.008)	(0.008)	(0.008)	(0.008)	(0.012)	(0.010)	(0.010)	(0.009)
Probit	0.175	0.179	0.179	0.175	0.234	0.205	0.200	0.192
	(0.007)	(0.008)	(0.008)	(0.007)	(0.016)	(0.011)	(0.010)	(0.010)
Loglog	0.158	0.164	0.165	0.164	0.180	0.182	0.181	0.184
	(0.008)	(0.008)	(0.008)	(0.007)	(0.012)	(0.010)	(0.010)	(0.010)
				$\theta = (-1, 0.5)$ or $\theta = (-0.5, 0.5)$				
Cauchit	0.085	0.085	0.085	0.082	0.080	0.087	0.086	0.085
	(0.008)	(0.008)	(0.008)	(0.008)	(0.007)	(0.009)	(0.009)	(0.009)
Logit	0.092	0.096	0.096	0.094	0.087	0.098	0.098	0.099
	(0.008)	(0.008)	(0.008)	(0.008)	(0.007)	(0.009)	(0.009)	(0.009)
Probit	0.108	0.120	0.120	0.117	0.073	0.115	0.121	0.127
	(0.008)	(0.007)	(0.007)	(0.007)	(0.005)	(0.007)	(0.008)	(0.009)
Loglog	0.127	0.140	0.141	0.140	0.093	0.140	0.147	0.159
	(0.008)	(0.007)	(0.007)	(0.007)	(0.007)	(0.008)	(0.008)	(0.009)

Note: Below the partial effects we report standard errors in parentheses.

Journal of Economic Surveys (2011) Vol. 25, No. 1, pp. 19–68

Table 5 contains the results for the size analysis for $N = \{500, 1000\}$. Clearly, irrespective of the version considered, the RESET test displays the poorest finite-sample properties, its estimated size being different from the nominal size at the 5% level in most of the cases simulated. The performance of the GOL test was also relatively modest in logit and probit models characterized by an asymmetric distribution of the response variable, being undersized in 11 out of the 12 cases analysed. With regard to the *P* test, its behaviour appears to be somewhat sensitive to the alternative hypothesis considered: in some cases (e.g. H_1: cauchit), it revealed some tendency to over-reject the null hypothesis; in others, its performance was very good (e.g. H_1: loglog). Finally, both the GOFF tests exhibited estimated sizes very close to the nominal one in most cases.

In the power analysis we computed the percentage of rejections of the three false null hypotheses considered for each one of the four alternative models simulated. Table 6 reports the results obtained for the case where probit is the true model; for the full set of results, see Ramalho *et al.* (2009). In general, the power of all tests increases when the sample size or the level of asymmetry in the distribution of $y(\theta_0 \neq 0)$ increases or the conditional variance of y decreases (ϕ increases). All tests display very satisfactory power properties in the two sets of cases where the differences between the functional form assumed in the data generation and that used in the estimation are substantial: (i) the true conditional mean is of the loglog form and one of the three symmetric models is assumed and vice versa (last six columns of Table 6); and (ii) both the true and the hypothesized models are symmetric about $x\theta = 0$ but the distribution of y is asymmetric (all columns of Table 6 relative to $\theta_0 = -1$). In these two sets of cases, the only (expected) exceptions to this good behaviour of all tests occur when the data are generated according to a logit model and a probit model is estimated or, in some cases, when the variability of y is very large ($\phi = 1$). Moreover, note that in these cases the *P* test is, in general, the most powerful one and that, in spite of the fact that we are considering uncorrected powers, the GOFF tests (especially the GOFF1 version) display better power properties than RESET tests in many experiments.

A very different scenario arises when we consider the remaining cases, i.e. when both the true and the postulated models are symmetric about $x\theta = 0$ and the distribution of the response variable is also approximately symmetric (all columns of Table 6 relative to $\theta_0 = 0$ except those concerning the loglog model). In this case, the power properties of all tests are much more modest than in the previous experiments. In particular, the GOFF tests have very low power, which is not surprising since both of them are based on generalizations that introduce asymmetry in the cauchit, logit and probit models, which is not present in these experiments. For similar reasons, the power of the *P* test is now much lower when the alternative is the asymmetric loglog model instead of another symmetric specification. In contrast, when two symmetric models are contrasted, the *P* test is again the most powerful of all tests in most cases.

Overall, these experiments show that GOFF and *P* tests are indeed good alternatives to the more popular RESET and GOL tests to assess conditional

Table 5. Monte Carlo Estimated Sizes (%) for a Nominal Size of 5% for Tests for the Functional Form.

	Cauchit						Logit					
	$\theta_0 = 0$			$\theta_0 = -1$			$\theta_0 = 0$			$\theta_0 = -1$		
H_0	$\phi = 1$	$\phi = 5$	$\phi = 20$	$\phi = 1$	$\phi = 10$	$\phi = 40$	$\phi = 1$	$\phi = 5$	$\phi = 20$	$\phi = 1$	$\phi = 10$	$\phi = 40$
						$N = 500$						
RESET2	6.0*	5.1	5.0	5.8*	4.4*	4.4*	6.0*	5.0	4.6	6.2*	4.9	4.3*
RESET3	6.0*	4.6	4.4*	6.4*	4.2*	3.9*	5.6*	4.2*	4.0*	6.6*	4.5*	3.9*
GOL	5.6*	5.1	5.0	5.4	4.8	4.8	5.7*	4.9	5.0	4.1*	2.8*	3.0*
GOFF1	4.9	4.8	5.1	5.3	4.7	4.8	4.9	5.2	4.6	5.3	4.8	4.7
GOFF2	4.9	4.7	5.2	5.3	4.3*	4.5*	4.9	5.1	4.8	5.1	4.7	4.7
P tests												
H_1: cauchit	–	–	–	–	–	–	5.6*	4.8	4.8	5.7*	4.8	4.7
H_1: logit	5.5*	5.5*	5.1	5.3	4.8	4.9	–	–	–	–	–	–
H_1: probit	5.5*	5.5*	5.1	5.3	4.8	4.8	5.6*	4.8	4.8	5.9*	4.9	4.8
H_1: loglog	5.2	4.9	4.9	5.2	4.8	4.8	4.8	5.3	4.6	5.6*	4.9	4.8
						$N = 1000$						
RESET2	5.5*	5.5*	5.0	5.3	4.8	4.6	5.3	4.6	4.7	5.7*	5.4	4.6
RESET3	5.8*	5.1	4.5*	6.0*	4.8	4.4*	5.4	4.5*	4.4*	6.4*	5.0	4.3*
GOL	4.9	5.2	5.1	4.7	5.1	5.0	5.0	4.8	4.8	3.6*	3.8*	3.3*
GOFF1	4.5*	5.2	4.6	4.7	5.0	5.1	4.8	4.9	4.8	4.8	5.2	4.7
GOFF2	4.6	5.3	4.8	4.7	4.7	5.0	4.9	5.0	4.9	4.9	5.1	4.9
P tests												
H_1: cauchit	–	–	–	–	–	–	4.7	5.2	4.9	4.8	5.3	4.8
H_1: logit	5.1	5.5*	5.1	5.0	5.2	4.9	–	–	–	–	–	–
H_1: probit	5.3	5.4	5.0	5.0	5.3	5.0	5.0	5.3	4.7	5.0	5.2	4.9
H_1: loglog	5.2	5.4	5.1	5.1	5.3	4.9	4.7	4.9	4.7	4.8	5.1	4.6

Journal of Economic Surveys (2011) Vol. 25, No. 1, pp. 19–68

Table 5. *Continued.*

	Probit						Loglog					
	$\theta_0 = 0$			$\theta_0 = -1$			$\theta_0 = 0$			$\theta_0 = -0.5$		
H_0	$\phi = 1$	$\phi = 5$	$\phi = 20$	$\phi = 1$	$\phi = 10$	$\phi = 40$	$\phi = 1$	$\phi = 5$	$\phi = 20$	$\phi = 1$	$\phi = 10$	$\phi = 40$
						$N = 500$						
RESET2	7.0*	6.0*	4.8	6.8*	5.8*	5.0	6.5*	5.4	5.6*	6.3*	5.6*	5.5*
RESET3	6.8*	5.5*	4.5*	7.5*	6.9*	5.2	5.8*	5.6*	5.3	4.8	6.7*	5.9*
GOL	5.3	5.4	4.7	4.1*	3.7*	3.3*	5.0	4.9	5.3	5.2	4.6	4.8
GOFF1	4.7	5.0	4.7	4.7	5.2	5.2	–	–	–	–	–	–
GOFF2	4.8	4.7	4.7	5.1	5.2	5.1	5.2	4.8	5.2	5.2	5.1	5.3
P tests												
H_1: cauchit	5.9*	5.3	5.0	5.6*	5.6*	5.1	5.9*	5.3	5.7*	5.8*	6.5*	6.5*
H_1: logit	6.4*	5.7*	5.0	5.8*	5.5*	5.3	5.4	5.0	5.3	5.5*	5.8*	5.6*
H_1: probit	–	–		–	–	–	5.3	4.8	5.2	5.2	5.4	5.4
H_1: loglog	4.8	4.8	4.6	4.8	5.1	5.2	–	–	–	–	–	–
						$N = 1000$						
RESET2	6.8*	5.5*	5.4	7.4*	5.7*	5.2	6.9*	5.9*	5.3	6.6*	5.7*	5.5*
RESET3	7.9*	5.6*	5.0	8.8*	6.8*	5.4	7.4*	6.5*	5.7*	6.9*	7.0*	6.7*
GOL	5.4	5.0	5.1	5.0	4.1*	4.0*	5.0	4.9	4.8	5.0	4.7	4.4*
GOFF1	5.0	5.0	5.2	5.4	5.1	5.1	–	–	–	–	–	–
GOFF2	4.8	4.9	5.1	5.7*	4.9	5.2	5.1	5.1	4.9	4.9	5.0	4.9
P tests												
H_1: cauchit	5.8*	4.9	4.9	6.2*	5.1	5.3	6.2*	5.7*	5.3	5.8*	6.7*	6.2*
H_1: logit	5.9*	5.0	4.9	6.0*	4.9	5.3	5.6*	5.4	5.0	5.3	5.4	5.2
H_1: probit	–	–	–	–	–	–	5.0	5.1	4.8	5.0	4.9	4.8
H_1: loglog	4.8	4.9	5.0	5.4	4.8	5.2	–	–	–	–	–	–

Note: The values with an asterisk are significantly different from the nominal size at the 5% level (95% confidence interval limits: 4.58 and 5.43).

Journal of Economic Surveys (2011) Vol. 25, No. 1, pp. 19–68

Table 6. Monte Carlo Estimated Powers (%) for a Nominal Size of 5% for Tests for the Functional Form – True Model: Probit.

	Cauchit						Logit						Loglog					
	$\theta_0 = 0$			$\theta_0 = -1$			$\theta_0 = 0$			$\theta_0 = -1$			$\theta_0 = 0$			$\theta_0 = -1$		
H_0	$\phi = 1$	$\phi = 5$	$\phi = 20$	$\phi = 1$	$\phi = 10$	$\phi = 40$	$\phi = 1$	$\phi = 5$	$\phi = 20$	$\phi = 1$	$\phi = 10$	$\phi = 40$	$\phi = 1$	$\phi = 5$	$\phi = 20$	$\phi = 1$	$\phi = 10$	$\phi = 40$
									$N = 500$									
RESET2	27.8	52.5	93.3	86.7	100.0	100.0	9.4	9.4	10.6	13.5	27.7	64.0	18.0	44.7	93.9	7.9	35.7	93.5
RESET3	26.6	46.9	89.8	85.3	100.0	100.0	9.0	8.4	8.8	15.6	27.6	59.1	16.2	38.0	90.2	5.9	26.5	88.7
GOL	24.0	53.0	95.5	76.7	100.0	100.0	7.9	8.4	10.0	8.3	21.5	59.3	12.0	34.8	88.3	7.2	20.1	62.3
GOFF1	5.5	6.3	7.2	76.4	100.0	100.0	5.0	5.2	5.1	9.2	25.7	66.6	–	–	–	–	–	–
GOFF2	5.4	6.4	6.9	53.8	99.9	100.0	4.9	5.3	4.9	7.3	20.9	58.7	18.5	53.7	96.9	9.6	49.9	97.4
P tests																		
H_1: cauchit	–	–	–	–	–	–	7.6	8.6	12.2	11.4	29.7	70.6	8.0	24.8	73.3	8.5	42.8	93.1
H_1: logit	29.9	60.1	97.0	90.2	100.0	100.0	–	–	–	–	–	–	16.4	50.5	95.3	9.2	47.5	96.5
H_1: probit	30.6	60.5	97.0	90.5	100.0	100.0	9.0	9.6	13.1	12.3	30.9	72.2	19.9	55.0	96.6	10.1	49.7	97.3
H_1: loglog	21.2	38.9	80.8	90.0	100.0	100.0	5.2	5.3	5.3	10.7	28.9	70.5	–	–	–	–	–	–
									$N = 1000$									
RESET2	42.4	80.3	99.9	98.9	100.0	100.0	10.0	10.9	16.9	17.7	43.7	91.2	32.1	76.2	100.0	13.8	71.4	99.9
RESET3	41.9	76.5	99.7	98.5	100.0	100.0	11.0	10.2	14.2	20.4	42.2	88.8	29.4	70.2	99.8	11.1	62.0	99.9
GOL	40.4	81.6	99.9	96.7	100.0	100.0	8.0	9.2	14.8	12.2	36.8	89.9	23.2	65.4	99.7	10.8	39.9	92.3
GOFF1	5.6	6.5	9.5	96.8	100.0	100.0	4.8	5.1	5.1	12.9	42.4	92.6	–	–	–	–	–	–
GOFF2	5.8	6.8	9.4	85.2	100.0	100.0	5.0	5.1	5.2	10.2	35.5	87.4	36.7	83.5	100.0	18.7	82.8	100.0
P tests																		
H_1: cauchit	–	–	–	–	–	–	8.0	10.2	19.0	15.7	46.9	94.3	14.9	49.6	95.0	14.8	74.7	98.4
H_1: logit	47.3	86.5	100.0	99.4	100.0	100.0	–	–	–	–	–	–	33.5	80.6	99.4	17.3	80.4	99.6
H_1: probit	47.8	86.7	100.0	99.5	100.0	100.0	8.9	11.3	20.3	16.8	49.0	95.0	38.0	84.0	99.8	18.9	82.3	99.9
H_1: loglog	30.3	62.4	98.0	99.4	100.0	100.0	5.0	5.3	5.6	14.8	46.9	94.8	–	–	–	–	–	–

Journal of Economic Surveys (2011) Vol. 25, No. 1, pp. 19–68

mean assumptions in fractional regression models: the GOFF tests are the best in terms of size and often they are among the most powerful tests, while the *P* tests, despite over-rejecting the true null hypothesis in some cases, display clearly the best power properties in most cases. However, when the response variable is symmetrically distributed, the GOFF tests have the important drawback of failing too often to reject symmetric but ill-specified models for the conditional mean of y.

5.4 *Tests for the Functional Form When There Are Boundary Observations*

Finally, we investigate the ability of the functional form tests to detect specification failures in the conditional mean due to the estimation of one-part models when the data-generating process is governed by two-part models, and vice versa. In this context, a large number of combinations of functional forms for $G(x\theta)$, $F(x\beta_{1P})$ and $M(x\beta_{2P})$ could have been considered but, in terms both of the data-generating process and of the null hypothesis to be tested, we restricted our attention to the case where a logit functional form is adopted for $G(x\theta)$ in one-part models and for both $F(x\beta_{1P})$ and $M(x\beta_{2P})$ in two-part models. Although this set-up corresponds to a very simple case, it is probably the most usual approach in applied work; see for example Cook *et al.* (2008) and Ramalho and Silva (2009).

In order to generate samples of fractional data with a given proportion of zero outcomes, we used two distinct data-generating processes. For one-part logit models, as the ratio of a bounded integer variable and its upper limit L is a fractional variable, we first generated a binomial-distributed variable y^* with parameters $L = 16$ and mean $G(x\theta)$ and then obtained a fractional variable $y \in [0, 1]$ by calculating $y = y^*/L$. For two-part logit models, we first generated a binary variable y^* according to the functional form specified for $F(x\beta_{1P})$ and then, only for the sampling units for which $y^* = 1$, we used a beta distribution based on $M(x\beta_{2P})$ and a shape parameter $\phi = 15$ for generating the positive, fractional outcomes. In both cases, we considered $N = \{500, 1000\}$.

Let $\beta_{1P} = (\beta_{1P_0}, \beta_{1P_1})$ and $\beta_{2P} = (\beta_{2P_0}, \beta_{2P_1})$. We set $\theta_1 = \beta_{1P_1} = \beta_{2P_1} = 1$ and chose θ_0 and β_{1P_0} in such a way that the proportion of zero outcomes in each model was 10%, 30% or 50%. The value of the remaining parameter, β_{2P_0}, was chosen in order to obtain identical values for the conditional mean and variance of y in both one-part and two-part logit models. We computed two distinct sets of tests. On the one hand, we computed the same tests considered in the previous section, which were applied separately to one-part models and the two components of two-part models. On the other hand, we used the *P* test to compare both one-part models and the full specification of two-part models (and vice versa) and alternative full specifications of two-part models. For computing this test, we considered nine alternative full specifications for two-part models, each of which corresponded to a different combination of the cauchit, logit and loglog functional forms.

Table 7 reports the results obtained for the size analysis. As in the previous section, the empirical size of the GOFF tests is not significantly different from the nominal one in most cases, the GOL test is undersized most of the time, and the RESET statistics are clearly oversized, especially the RESET3 version. With regard

Table 7. Monte Carlo Estimated Sizes (%) for a Nominal Size of 5% for Tests for the Functional Form in Presence of Boundary Observations.

	$N = 500$			$N = 1000$		
	10%[a]	30%[a]	50%[a]	10%[a]	30%[a]	50%[a]
			H_0: one-part logit model			
RESET2	5.2	5.4	5.9*	5.2	5.5*	4.8
RESET3	5.4	7.3*	9.4*	5.6*	6.5*	7.1*
GOL	4.5*	3.1*	5.3	4.4*	2.8*	4.5*
GOFF1	5.2	5.0	5.2	5.4	5.1	4.7
GOFF2	5.2	4.8	4.7	5.4	4.9	4.1*
P tests (H_1: one-part model)						
H_1: cauchit	5.5*	5.3	6.0*	5.3	5.4	5.1
H_1: probit	5.5*	5.1	5.4	5.4	5.2	5.0
H_1: loglog	5.4	5.0	5.3	5.3	5.2	5.0
P tests (H_1: two-part model)						
H_1: cauchit + cauchit	5.9*	5.5*	5.8*	5.8*	6.0*	5.0
H_1: cauchit + logit	5.7*	4.9	5.3	5.1	5.0	4.8
H_1: cauchit + loglog	5.4	5.1	5.4	5.5*	5.0	4.7
H_1: logit + cauchit	5.8*	6.0*	6.1*	5.5*	6.0*	5.5*
H_1: logit + logit	5.5*	4.9	5.3	5.5*	4.9	4.9
H_1: logit + loglog	5.4	5.1	5.2	5.5*	5.0	4.9
H_1: loglog + cauchit	5.7*	5.9*	5.8*	5.5*	5.7*	5.5*
H_1: loglog + logit	5.4	4.8	5.3	5.3	5.1	4.7
H_1: loglog + loglog	5.5*	5.0	5.2	5.5*	5.1	4.8
			H_0: two-part logit model			
First part						
RESET2	4.8	5.1	5.0	4.8	4.9	5.2
RESET3	4.8	6.4*	6.0*	4.8	5.9*	5.8*
GOL	4.4*	5.1	4.9	3.9*	4.9	5.3
GOFF1	5.2	5.1	4.8	4.9	5.1	5.0
GOFF2	5.2	5.0	4.9	4.8	5.1	5.1
P tests						
H_1: cauchit	6.3*	6.5*	6.4*	6.2*	6.5*	6.1*
H_1: probit	5.7*	6.7*	7.7*	6.0*	6.4*	6.5*
H_1: loglog	5.0	5.3	4.9	5.1	5.0	5.1
Second part						
RESET2	5.2	5.3	5.3	6.0*	5.3	5.8*
RESET3	5.1	6.5*	6.1*	5.9*	6.3*	7.8*
GOL	4.5*	3.1*	4.1*	5.6*	2.8*	3.6*
GOFF1	4.6	5.2	4.8	5.3	4.7	5.4
GOFF2	4.8	4.9	4.5*	5.2	4.7	4.9
P tests						
H_1: cauchit	4.8	5.3	5.4	5.5	5.1	5.3
H_1: probit	4.7	5.2	5.0	5.6*	5.0	5.6*
H_1: loglog	4.6	5.0	4.7	5.4	4.9	5.4

Table 7. *Continued.*

	$N = 500$			$N = 1000$		
	10%[a]	30%[a]	50%[a]	10%[a]	30%[a]	50%[a]
Full specification						
P tests (H_1: one-part model)						
H_1: cauchit	4.6	5.0	4.7	5.0	5.4	5.1
H_1: logit	4.3*	5.2	4.9	5.2	5.1	4.9
H_1: probit	4.5*	5.0	4.6	5.0	4.8	4.7
H_1: loglog	4.3*	4.7	4.7	4.7	4.1*	4.5*
P tests (H_1: two-part model)						
H_1: cauchit + cauchit	4.6	4.8	4.7	5.1	5.1	5.0
H_1: cauchit + logit	4.7	4.7	5.1	4.7	4.6	4.8
H_1: cauchit + loglog	4.3*	4.9	4.5*	4.9	4.7	4.7
H_1: logit + cauchit	4.7	4.9	4.8	5.0	4.9	4.7
H_1: logit + loglog	4.2*	4.7	4.5*	4.5*	4.3*	4.4*
H_1: loglog + cauchit	4.6	5.1	4.9	5.0	5.0	5.0
H_1: loglog + logit	4.5*	5.0	4.9	5.2	4.9	4.6
H_1: loglog + loglog	4.3*	4.6	4.7	4.6	4.4*	4.6

Note: The values with an asterisk are significantly different from the nominal size at the 5% level (95% confidence interval limits: 4.58 and 5.43).
[a]Percentage of the sampled observations with $Y = 0$.

to the *P* test, on the one hand, it continues to display some tendency to over-reject the null hypothesis in some cases and, on the other hand, it seems to be much less reliable when applied to binary models and to be slightly undersized when used to test alternative full specifications of two-part models.

The finite-sample power properties of the tests are documented in Table 8. Note that we have restricted this analysis to two simple cases: estimation of a one-part logit model when the true model is a two-part logit model (first panel of Table 8) and (ii) the opposite case (second panel). Again, most of the highest percentage of rejections of the false null hypothesis are obtained by some versions of the *P* test. However, the power of this statistic is very low when we test the full specification of the two-part logit model against either alternative one-part models or other two-part models. This implies that when using the *P* statistic for testing two-part models it will be better, in general, to focus on the separate analysis of the two components of those models. With regard to the other tests, all of them display very satisfactory power properties. Note that with boundary observations the distribution of y will be, in general, asymmetric, and hence the GOFF tests are particularly useful in this framework.

6. Empirical Application: The Determinants of Corporate Capital Structure

In this section we apply the techniques described so far to the regression analysis of the capital structure decisions of Portuguese small and medium enterprises (SMEs),

Table 8. Monte Carlo Estimated Powers (%) for a Nominal Size of 5% for Tests for the Functional Form in Presence of Boundary Observations.

	$N = 500$			$N = 1000$		
	10%[a]	30%[a]	50%[a]	10%[a]	30%[a]	50%[a]
			H_0: one-part logit model			
RESET2	16.5	33.5	33.4	26.6	56.3	54.7
RESET3	17.8	36.2	38.4	26.4	56.0	55.9
GOL	12.5	24.5	28.8	21.8	47.8	46.8
GOFF1	14.5	29.1	28.3	24.6	53.1	51.0
GOFF2	11.1	22.6	21.7	19.2	42.8	41.2
P tests (H_1: one-part model)						
H_1: cauchit	18.8	37.3	39.4	30.8	62.1	62.2
H_1: probit	19.4	36.7	35.6	31.4	61.2	59.9
H_1: loglog	16.4	34.5	34.5	27.7	59.1	59.0
P tests (H_1: two-part model)						
H_1: cauchit + cauchit	11.5	18.4	20.1	19.6	36.4	41.5
H_1: cauchit + logit	24.2	38.5	32.0	35.6	62.2	55.7
H_1: cauchit + loglog	17.1	34.9	32.3	28.4	59.1	56.8
H_1: logit + cauchit	14.6	24.2	19.7	25.0	47.3	40.3
H_1: logit + logit	20.0	37.0	34.7	32.3	61.7	59.0
H_1: logit + loglog	16.3	33.5	33.1	27.4	58.3	57.4
H_1: loglog + cauchit	13.9	21.5	15.5	24.1	41.6	32.5
H_1: loglog + logit	21.2	39.6	39.1	33.6	63.9	62.6
H_1: loglog + loglog	16.4	34.5	34.8	27.5	59.1	59.4
			H_0: two-part logit model			
First part						
RESET2	20.4	27.6	19.2	44.9	58.2	43.4
RESET3	18.5	24.7	16.6	40.1	51.0	36.2
GOL	24.8	31.4	20.6	51.4	62.2	44.8
GOFF1	32.0	38.9	28.4	55.9	65.8	52.1
GOFF2	34.9	42.4	30.2	63.6	73.4	57.8
P tests						
H_1: cauchit	50.9	50.2	10.4	84.2	75.1	11.0
H_1: probit	50.3	51.3	11.7	77.9	74.5	11.8
H_1: loglog	15.6	28.9	21.8	43.1	61.7	48.0
Second part						
RESET2	92.9	99.8	98.4	99.9	100.0	100.0
RESET3	91.1	99.6	96.6	100.0	100.0	100.0
GOL	85.4	99.3	97.7	99.3	100.0	100.0
GOFF1	84.5	98.9	96.7	99.0	100.0	100.0
GOFF2	72.6	97.1	94.2	94.8	99.9	99.8
P tests						
H_1: cauchit	92.1	98.9	91.3	99.8	100.0	99.7
H_1: probit	95.5	99.8	98.6	100.0	100.0	100.0
H_1: loglog	91.9	99.8	98.7	99.7	100.0	100.0

Table 8. *Continued.*

	$N = 500$			$N = 1000$		
	10%[a]	30%[a]	50%[a]	10%[a]	30%[a]	50%[a]
Full specification						
P tests (H_1: one-part model)						
H_1: cauchit	6.3	4.9	4.5	4.8	3.4	2.8
H_1: logit	6.0	4.3	3.7	4.3	3.3	3.3
H_1: probit	4.7	3.8	3.7	3.5	4.6	6.0
H_1: loglog	3.8	3.0	4.5	6.0	4.4	3.0
P tests (H_1: two-part model)						
H_1: cauchit + cauchit	3.6	4.1	3.4	3.3	5.0	3.8
H_1: cauchit + logit	3.9	5.0	4.0	4.5	5.8	5.4
H_1: cauchit + loglog	4.7	5.3	4.6	5.8	7.5	7.4
H_1: logit + cauchit	4.2	5.3	2.9	6.3	10.4	3.9
H_1: logit + loglog	3.9	3.9	3.1	6.6	7.2	4.6
H_1: loglog + cauchit	5.2	4.5	2.5	5.1	4.8	2.6
H_1: loglog + logit	3.4	2.6	2.8	2.5	5.4	5.5
H_1: loglog + loglog	3.7	3.8	2.9	6.8	9.2	6.3

[a]Percentage of the sampled observations with $Y = 0$.

i.e. their option between debt and equity. First, we discuss the main characteristics of our data and variables, then we discuss briefly some alternative capital structure theories, and finally we present the econometric results of our analysis.

6.1 *Data and Variables*

We consider as a measure of financial leverage the ratio of long-term debt (LTD, defined as the total company's debt due for repayment beyond 1 year) to long-term capital assets (defined as the sum of LTD and equity); see Rajan and Zingales (1995) for an extensive discussion on this and other alternative measures of leverage and for a survey of capital structure theories. We use the definition of SMEs adopted by the European Commission (recommendation 2003/361/EC), including in this category enterprises that employ fewer than 250 persons and have either an annual turnover not exceeding 50 million euros or an annual balance sheet total not exceeding 43 million euros.

We use a subset of the data considered by Ramalho and Silva (2009), which also included information on large firms. Our data set is relative to the year of 1999 and comprises 4421 SMEs, among which 74.8% present a null leverage ratio. Other studies have also documented that a substantial proportion of firms in most countries follow a zero-debt policy; see *inter alia* Petersen and Rajan (1994), Brounen *et al.* (2005) and Strebulaev and Yang (2007). The high percentage of firms that do not use debt at all makes the standard practice of using linear regression

models to explain capital structure decisions (which is still used in most empirical studies) clearly inappropriate. Therefore, a few authors (e.g. Rajan and Zingales, 1995; Cassar, 2004) have opted for using a Tobit approach for data censored at zero. However, as we argue in Section 2, the stringent assumptions associated with the Tobit model and the impossibility of using a two-limit Tobit model (there are no 'zero-equity' firms) make it clear that the use of fractional regression models is a better option for modelling leverage ratios.

In all the alternative regression models considered below, we used similar explanatory variables to those employed by Ramalho and Silva (2009), although in some cases we opted for different proxies: non-debt tax shields (NDTS), measured by the ratio between depreciation and earnings before interest, taxes and depreciation; tangibility (TANGIB), the proportion of tangible assets; size (SIZE), the natural logarithm of total assets; profitability (PROFITAB), the ratio between earnings before interest, taxes and depreciation and total assets; growth (GROWTH), the yearly percentage change in total assets; age (AGE), the number of years since the foundation of the firm; liquidity (LIQUIDITY), the sum of cash and marketable securities, divided by current assets; and four industry dummies.

6.2 *Alternative One- and Two-part Capital Structure Theories*

To date, most capital structure empirical studies have focused on the use of one-part models to explain leverage ratios, which follows directly from the fact that most capital structure theories provide a single explanation for all possible values of leverage ratios. This is the case, for example, of the two most popular explanations of capital structure decisions, the trade-off and the pecking-order theories. According to the former, firms choose the proportion of debt in their capital structure that maximizes their value, which may imply leverage ratios of any value in the unit interval, including zero. Regarding the latter, the pecking-order theory argues that firms do not possess an optimal capital structure. Instead, the firm leverage at each moment merely reflects its external financing requirements, which may be null or any positive amount. For more details, see the recent survey by Frank and Goyal (2008).

In contrast to these traditional approaches, Strebulaev and Yang (2007), in a recent paper suggestively entitled 'The mystery of zero-leverage firms', argue that zero-leverage behaviour is a persistent phenomenon and that standard capital structure theories are unable to provide a reasonable explanation for it. Another interesting recent finding about capital structure decisions is that while larger firms are more likely to have some debt, conditional on having some debt, larger firms are less levered. In particular, Faulkender and Petersen (2006) found that excluding zero-debt firms from leverage regressions changes the sign of the coefficient associated with the variable size from positive to negative, while Kurshev and Strebulaev (2007) argue that 'the positive relationship (between firm size and leverage) is an artifact of the presence of small unlevered firms in the economy. When we control for unlevered firms, the relationship between firm size and leverage becomes slightly but statistically significant negative'. Clearly, firm size

seems to affect in an inverse way the decisions on (i) to issue or not to issue debt and (ii) (for those firms that do decide to use debt) how much debt to issue.

Kurshev and Strebulaev (2007) put forward a theoretical explanation for these opposite effects of firm size on leverage. They conjecture that it is the presence of fixed costs of external financing, and the consequent infrequent refinancing of firms, that causes these differences between small and large firms, since the former are much more affected in relative terms. According to these authors (i) small firms choose higher leverage at the moment of refinancing to compensate for less frequent rebalancing, which explains why, conditional on having debt, they are more levered than large firms; (ii) as they wait longer times between refinancings, small firms, on average, have lower levels of leverage; and (iii) in each moment, there is a mass of firms opting for no leverage, since small firms may find it optimal to postpone their debt issuances until their fortunes improve substantially relative to the costs of issuance. Clearly, a two-part fractional regression model may be the best option for modelling leverage ratios: first, a binary choice model is used to explain the probability of a firm raising debt; then, a fractional regression model is employed to explain the relative amount of debt issued by firms that do use debt. Indeed, with this type of model the variable size (and others) is allowed to influence each decision in a different fashion.

Based on these conjectures, Ramalho and Silva (2009) decided to use a two-part fractional regression model to explain capital structure decisions. Cook *et al.* (2008) have also used a similar model, but did not provide any theoretical justification for their option. In both papers a logistic specification was adopted for the two levels of the model. Ramalho and Silva (2009) considered uniquely QML estimation and used only the RESET test to assess the specification of their model, while Cook *et al.* (2008) estimated a one-part model by QML and a two-part model by ML (based on the beta distribution) and did not perform any test, using the Spearman rank correlation between predicted and actual leverage ratios to choose their final model.

Since both one- and two-part models provide plausible theoretical explanations for capital structure decisions, next all the alternative formulations for one-part and two-part models and specification tests discussed before are applied to the analysis of the capital structure decisions of Portuguese SMEs.

6.3 *Econometric Analysis*

We consider five alternative specifications for the $G(x\theta)$, $F(x\beta_{1P})$ and $M(x\beta_{2P})$ functional forms: cauchit, logit, probit, loglog and complementary loglog. Given the existence of zero outcomes, only conditional mean models may be used for $G(x\theta)$. We consider only QML estimation, since our simulation study revealed that in no case is its performance inferior to that of NLS estimators. In two-part models, $F(x\beta_{1P})$ is estimated in all cases by ML based on the Bernoulli distribution and $M(x\beta_{2P})$ is estimated by both Bernoulli-based QML and beta-based ML. The specification test strategy proposed in the paper is then employed to select the best model(s).

Table 9. Regression Results for One-part Models.

		QML				
	OLS	Cauchit	Logit	Probit	Loglog	Cloglog
NDTS	−0.001**	−0.101	−0.045	−0.021	−0.015	−0.042
	(0.000)	(0.073)	(0.029)	(0.014)	(0.010)	(0.027)
TANGIB	0.071***	2.274***	1.219***	0.627***	0.470***	1.122***
	(0.014)	(0.391)	(0.181)	(0.095)	(0.074)	(0.167)
SIZE	0.027***	0.696***	0.369***	0.193***	0.149***	0.340***
	(0.002)	(0.060)	(0.024)	(0.012)	(0.009)	(0.022)
PROFITAB	−0.132***	−7.453***	−3.369***	−1.688***	−1.227***	−3.141***
	(0.022)	(1.298)	(0.457)	(0.230)	(0.173)	(0.428)
GROWTH	0.000	0.003	0.001	0.000	0.000	0.001
	(0.000)	(0.002)	(0.001)	(0.001)	(0.000)	(0.001)
AGE	0.000	−0.005	−0.005***	−0.003***	−0.002***	−0.004***
	(0.000)	(0.004)	(0.002)	(0.001)	(0.001)	(0.002)
LIQUIDITY	−0.051***	−4.848***	−1.331***	−0.620***	−0.422***	−1.286***
	(0.011)	(0.947)	(0.255)	(0.122)	(0.087)	(0.243)
CONSTANT	−0.259***	−12.413***	−7.141***	−3.857***	−2.806***	−6.814***
	(0.025)	(1.000)	(0.380)	(0.192)	(0.146)	(0.352)
RESET2 test	0.000***	0.000***	0.004***	0.186	0.593	0.000***
RESET3 test	0.000***	0.000***	0.001***	0.087*	0.434	0.000***
GOL test	–	0.000***	0.001***	0.048**	0.684	0.187
GOFF1 test	–	0.000***	0.001***	0.092*	–	0.000***
GOFF2 test	–	0.000***	0.001***	0.110	0.864	–
P test						
H_1: OLS	–	0.000***	0.000***	0.023**	0.528	0.000***
H_1: cauchit	0.000***	–	0.002***	0.167	0.747	0.000***
H_1: logit	0.000***	0.000***	–	0.141	0.770	0.000***
H_1: probit	0.000***	0.000***	0.001***	–	0.827	0.000***
H_1: loglog	0.000***	0.000***	0.001***	0.102	–	0.000***
H_1: cloglog	0.000***	0.000***	0.001***	0.111	0.806	–
R^2	0.100	0.097	0.116	0.117	0.118	0.115

Notes: OLS, ordinary least squares. Below the coefficients we report standard errors in parentheses; for the test statistics we report *P* values; ***, ** and * denote coefficients or test statistics that are significant at 1%, 5% or 10%, respectively; all regressions include industry dummies.

Tables 9 and 10 report the results obtained for one-part and two-part models, respectively. For comparison purposes, we report also the results obtained for a one-part linear regression model. The tests that appear in Table 10 are relative to the specification of the individual components of two-part models. In addition, for beta regression models we report the results of the bootstrapped OPG information matrix statistic described in Section 4.2, which was based on 999 bootstrap samples.

Table 10. Regression Results for Two-part Models.

	First part					Second part									
	ML					QML					ML				
	Cauchit	Logit	Probit	Loglog	Cloglog	Cauchit	Logit	Probit	Loglog	Cloglog	Cauchit	Logit	Probit	Loglog	Cloglog
NDTS	−0.078	−0.053	−0.028	−0.024	−0.044	−0.000	0.001	0.001	0.002	0.001	0.001	0.003	0.002	0.003	0.002
	(0.053)	(0.033)	(0.018)	(0.015)	(0.027)	(0.022)	(0.027)	(0.017)	(0.019)	(0.021)	(0.022)	(0.025)	(0.016)	(0.017)	(0.020)
TANGIB	2.103***	1.708***	0.948***	0.822***	1.341***	0.056	0.094	0.061	0.078	0.056	−0.045	−0.016	−0.006	0.010	−0.031
	(0.247)	(0.202)	(0.116)	(0.107)	(0.158)	(0.146)	(0.159)	(0.097)	(0.099)	(0.127)	(0.133)	(0.146)	(0.090)	(0.091)	(0.117)
SIZE	0.688***	0.598***	0.345***	0.300***	0.469***	−0.116***	−0.128***	−0.078***	−0.081***	−0.098***	−0.097***	−0.107***	−0.066***	−0.068***	−0.082***
	(0.043)	(0.030)	(0.017)	(0.016)	(0.023)	(0.021)	(0.023)	(0.014)	(0.014)	(0.019)	(0.020)	(0.022)	(0.013)	(0.014)	(0.017)
PROFITAB	−3.105***	−2.502***	−1.383***	−1.163***	−1.982***	−2.510***	−2.747***	−1.677***	−1.694***	−2.170***	−2.175***	−2.365***	−1.441***	−1.443***	−1.883***
	(0.664)	(0.515)	(0.287)	(0.248)	(0.420)	(0.404)	(0.408)	(0.245)	(0.238)	(0.335)	(0.401)	(0.412)	(0.249)	(0.243)	(0.336)
GROWTH	−0.000	−0.001	−0.000	−0.001	−0.000	0.003***	0.004***	0.002***	0.002***	0.003***	0.003***	0.003***	0.002***	0.002***	0.003***
	(0.002)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
AGE	0.002	−0.001	−0.000	0.000	−0.001	−0.005***	−0.005***	−0.003***	−0.003***	−0.004***	−0.005***	−0.005***	−0.003***	−0.003***	−0.004***
	(0.003)	(0.003)	(0.001)	(0.002)	(0.002)	(0.002)	(0.002)	(0.001)	(0.001)	(0.002)	(0.001)	(0.002)	(0.001)	(0.001)	(0.001)
LIQUIDITY	−2.513***	−1.602***	−0.828***	−0.589***	−1.439***	−0.192	−0.175	−0.103	−0.081	−0.169	−0.170	−0.158	−0.094	−0.078	−0.147
	(0.358)	(0.247)	(0.135)	(0.113)	(0.209)	(0.195)	(0.206)	(0.125)	(0.124)	(0.168)	(0.171)	(0.186)	(0.114)	(0.115)	(0.148)
CONSTANT	−10.620***	−9.326***	−5.386***	−4.379***	−7.675***	1.578***	1.651***	0.998***	1.360***	−0.925***	1.330***	1.374***	0.830***	1.184***	0.721***
	(0.670)	(0.464)	(0.258)	(0.233)	(0.363)	(0.338)	(0.370)	(0.226)	(0.228)	(0.296)	(0.310)	(0.341)	(0.209)	(0.212)	(0.271)
RESET2 test	0.000***	0.835	0.109	0.000***	0.003***	0.523	0.672	0.680	0.770	0.418	0.641	0.633	0.627	0.605	0.591
RESET3 test	0.000***	0.406	0.036**	0.000***	0.001***	0.371	0.259	0.235	0.188	0.246	0.702	0.639	0.627	0.609	0.615
GOL test	0.000***	0.464	0.061*	0.000***	0.018**	0.557	0.270	0.174	0.951	0.153	0.394	0.500	0.445	0.604	0.355
GOFF1 test	0.000***	0.689	0.101	–	0.003***	0.313	0.651	0.659	–	0.268	0.370	0.371	0.366	–	0.353
GOFF2 test	0.000***	0.903	0.058*	0.000***	–	0.276	0.579	0.713	0.819	–	0.360	0.370	0.366	0.328	–
P test															
H_1: cauchit	–	0.623	0.794	0.011**	0.606	–	0.749	0.637	0.355	0.359	–	0.350	0.335	0.275	0.326
H_1: logit	0.000***	–	0.043**	0.013**	0.001***	0.334	–	0.906	0.652	0.202	0.376	–	0.356	0.312	0.336
H_1: probit	0.000***	0.497	–	0.242	0.004***	0.301	0.916	–	0.739	0.221	0.372	0.368	–	0.320	0.344
H_1: loglog	0.000***	0.797	0.086**	–	0.003***	0.164	0.539	0.598	–	0.176	0.342	0.368	0.365	–	0.333
H_1: cloglog	0.000***	0.996	0.129	0.016**	–	0.914	0.615	0.790	0.644	–	0.394	0.370	0.365	0.309	–
IM test	–	–	–	–	–	–	–	–	–	–	0.516	0.230	0.188	0.104	0.380
R^2	0.208	0.211	0.210	0.205	0.210	0.176	0.176	0.176	0.176	0.176	0.175	0.175	0.175	0.174	0.174

Notes: Below the coefficients we report standard errors in parentheses; for the test statistics we report *P* values; ***, ** and * denote coefficients or test statistics that are significant at 1%, 5% or 10%, respectively; all regressions include industry dummies.

Journal of Economic Surveys (2011) Vol. 25, No. 1, pp. 19–68

The first striking point to emerge from the analysis of these results is that all the five QML/ML estimators considered for $E(y|x)$, $E(y^* = 1|x)$ and $E(y|x, y > 0)$ produce the same conclusions in terms of the sign and significance of the regression coefficients in each model, with only one exception (in the one-part cauchit model the variable AGE is not statistically significant). This result was somewhat expected since in Figure 3 we had already found that misspecification of the functional form, although creating serious distortions in the magnitude of partial effects, does not affect the correct estimation of their direction. Similarly, in the fractional component of two-part models, using QML or ML is indifferent in terms of the sign and significance of the regression coefficients but not in terms of their magnitude: in almost all cases the absolute values of the coefficient estimates yielded by the beta model are less than those obtained by the corresponding conditional mean model estimated by QML. Moreover, ML estimators display the least standard errors in almost all cases. Finally, note that the linear model is the only model that indicates that the variable NDTS is statistically significant and that (apart from the cauchit model) the variable AGE is not.

While the choice of a specific functional form for each one of the three conditional means of y in analysis seems to be important only for calculating the magnitude of partial effects, the choice between a one-part and a two-part model is clearly a very important issue. Indeed, in the two-part model some variables are important only for one of the two sequential leverage decisions made by firms (TANGIB, GROWTH, AGE and LIQUIDITY), while the variable SIZE displays opposite effects on the two levels of the model. If our specification tests reveal that a two-part model is preferable over a single model, then the empirical evidence provided in this paper will clearly favour the recent theoretical arguments put forward by both Strebulaev and Yang (2007) and Kurshev and Strebulaev (2007) over traditional capital structure approaches.

The analysis of the results of the specification tests indicates clearly that only a few specifications are correct. For one-part models, the hypothesis of correct specification of the linear regression model is clearly rejected by all tests. Actually, only the loglog specification for $G(x\theta)$ is never rejected. Given that leverage ratios are clearly asymmetrically distributed and that the number of zero outcomes is very large, a loglog functional form would indeed be our first choice for a one-part model. With regard to two-part models, in the first level, again, only one specification seems to be appropriate to describe the probability of a firm using debt: the logistic functional form. In contrast, for explaining $E(y|x, y > 0)$ all functional form tests fail to reject any of the five models estimated for both QML and ML estimators. Similarly, the information matrix test provides no evidence of the unsuitability of the beta distribution to describe the conditional distribution of LTD. Therefore, given their superior efficiency properties, we consider only the ML estimators for two-part models from now on.

Tables 9 and 10 also contain an R^2-type measure for each model, which was computed as the square of the correlation between the predicted and actual values of LTD and, thus, is comparable across any model and estimation method. The values found for R^2 are very similar in most cases but nevertheless they give further

evidence that the selected models provide a better fit than or a similar fit to the competitor models. Indeed, the highest R^2 in one-part and the first component of two-part models is displayed by the selected loglog and logit models, respectively. On the other hand, the R^2 of the alternative specifications considered for the second stage of two-part models is virtually identical. Note also that the R^2 of the linear regression model is about 18% smaller than that of the one-part loglog model, in spite of OLS choosing $\hat{\theta}$ to maximize the R^2 over all linear functions of the covariates, while the QML/ML methods do not maximize it given the functional form assumed in each case. Moreover, the linear regression model yields predicted outcomes below zero for 7.6% of the firms in our sample, which is a clear indicator of its unsuitability for modelling leverage ratios.

Given that the results of the functional form tests that assess separately $G(x\theta)$, $F(x\beta_{1P})$ and $M(x\beta_{2P})$ suggest that one one-part model and five alternative two-part models may be suitable to describe our data, in the next stage of our specification analysis we applied the versions of the *P* test that allow for the testing of one-part models against the full specification of two-part models, and vice versa, and of alternative full specifications for two-part models, one against the others. We tested only the specifications previously selected by the other tests. In Table 11 we report the *P* values of the *P* test for the one-part loglog model against 25 alternative two-part models and for each one the five two-part models previously selected against five alternative one-part models and 24 alternative two-part models.

The first panel of Table 11 shows clearly that one-part models are not at all appropriate for modelling leverage ratios. Indeed, the correct specification of the one-part loglog model was rejected against most of the alternative two-part models considered.[5] On the other hand, the new set of tests provided no evidence against some of the five alternative two-part models selected before, which allows us to conclude that two-part models are, undoubtedly, the best choice for modelling capital structure decisions. Noting that the two-part models that use a probit or loglog specification in their second level are never (the latter) or almost never (the former) rejected, we opted for them as the best two-part models for explaining the capital structure decisions of Portuguese SMEs.

In Table 12 we present estimates of the partial effects for the two models selected. We computed partial effects for the first and second part of the model, given by $\partial \Pr(y^* = 1|x)/\partial x_j = \beta_{1P_j} f(x\beta_{1P})$ and $\partial E[y|x, y \in (0, 1)]/\partial x_j = \beta_{2P_j} m(x\beta_{2P})$, respectively, and total partial effects, given by equation (16). These three types of partial effects describe the effect of a unitary change in the covariate x_j on the conditional probability of using LTD, on the proportion of LTD used by the firms that already use it, and on the proportion of LTD used by all firms, respectively. In each case, we calculated average sample effects and population partial effects evaluated at the mean of the covariates ($\bar{x}$),[6] which were calculated as, respectively, $\widehat{\text{ASE}} = \hat{\beta}_{1P_j}(1/N)\sum_{i=1}^{N} f(x_i\hat{\beta}_{1P})$ and $\widehat{\text{PPE}} = \hat{\beta}_{1P_j} f(\bar{x}\hat{\beta}_{1P})$ for $\partial \Pr(y^* = 1|x)/\partial x_j$ and similarly for the other partial effects. As seen in Table 12, the two alternative models yield very similar total partial effects. Note that the total partial effect of

Table 11. *P* Tests Involving the Full Specification of Two-part Models (*P* Values).

	H_0: loglog one-part model				
H_1: first part/second part	Cauchit	Logit	Probit	Loglog	Cloglog
Cauchit	0.362	0.278	0.265	0.205	0.352
Logit	0.000***	0.000***	0.000***	0.000***	0.000***
Probit	0.000***	0.000***	0.000***	0.000***	0.000***
Loglog	0.000***	0.000***	0.000***	0.000***	0.000***
Cloglog	0.006***	0.002***	0.001***	0.000***	0.005***
	H_0: logit + cauchit two-part model				
H_1: one-part	Cauchit	Logit	Probit	Loglog	Cloglog
	0.135	0.613	0.017**	0.075*	0.728
H_1: first part/second part	Cauchit	Logit	Probit	Loglog	Cloglog
Cauchit	0.007***	0.009***	0.010***	0.011**	0.010***
Logit	–	0.064*	0.051*	0.033**	0.128
Probit	0.021**	0.010***	0.009***	0.007***	0.001***
Loglog	0.037**	0.024**	0.022**	0.017**	0.028**
Cloglog	0.117	0.790	0.962	0.438	0.537
	H_0: logit + logit two-part model				
H_1: one-part	Cauchit	Logit	Probit	Loglog	Cloglog
	0.092*	0.329	0.092*	0.063*	0.265
H_1: first part/second part	Cauchit	Logit	Probit	Loglog	Cloglog
Cauchit	0.013**	0.017**	0.019**	0.025**	0.016**
Logit	0.158	–	0.044**	0.040**	0.092*
Probit	0.397	0.047**	0.031**	0.012**	0.194
Loglog	0.185	0.073*	0.061*	0.031**	0.143
Cloglog	0.041**	0.022**	0.057*	0.731	0.033**
	H_0: logit + probit two-part model				
H_1: one-part	Cauchit	Logit	Probit	Loglog	Cloglog
	0.301	0.312	0.972	0.398	0.282
H_1: first part/second part	Cauchit	Logit	Probit	Loglog	Cloglog
Cauchit	0.184	0.206	0.216	0.249	0.192
Logit	0.401	0.196	–	0.148	0.177
Probit	0.812	0.402	0.320	0.131	0.746
Loglog	0.582	0.400	0.362	0.231	0.573
Cloglog	0.199	0.095*	0.154	0.665	0.101
	H_0: logit + loglog two-part model				
H_1: one-part	Cauchit	Logit	Probit	Loglog	Cloglog
	0.995	0.608	0.603	0.913	0.621
H_1: first part/second part	Cauchit	Logit	Probit	Loglog	Cloglog
Cauchit	0.645	0.547	0.535	0.517	0.574
Logit	0.612	0.919	0.996	–	0.989
Probit	0.449	0.577	0.595	0.560	0.650
Loglog	0.480	0.565	0.577	0.548	0.620
Cloglog	0.798	0.881	0.818	0.796	0.864

Table 11. *Continued.*

	H_0: logit + cloglog two-part model				
H_1: one-part	Cauchit	Logit	Probit	Loglog	Cloglog
	0.044**	0.351	0.041**	0.047**	0.222
H_1: first part/second part	Cauchit	Logit	Probit	Loglog	Cloglog
Cauchit	0.005***	0.010***	0.012**	0.019**	0.009***
Logit	0.220	0.019**	0.011**	0.008***	–
Probit	0.136	0.007***	0.005***	0.003***	0.023**
Loglog	0.079*	0.019**	0.014**	0.007***	0.041**
Cloglog	0.007***	0.754	0.548	0.047**	0.025**

Note: ***, ** and * denote coefficients or test statistics that are significant at 1, 5 or 10%, respectively.

the variable SIZE is positive, which is in accordance with the positive relationship between firm size and leverage that is found systematically by empirical studies based on one-part models.

Finally, in Table 13, for comparison purposes, we report estimates of partial effects computed from linear and fractional one-part models. Naturally, only total partial effects can be computed in this case. The linear model clearly underestimates all partial effects, in particular those of TANGIB, PROFITAB and LIQUIDITY, where the bias in the estimations of the ASEs is about 26%, 51% and 47%, respectively. On the other hand, while the ASEs estimated by some one-part models (logit, probit, cloglog) are not very different from those produced by the selected two-part models, the differences in the estimation of the PPEs are much more important, with all one-part models underestimating most partial effects (e.g. for LIQUIDITY, TANGIB and PROFITAB the bias is above 16.8%, 8.5% and 7.9% respectively in all cases).

7. Concluding Remarks

This paper focused on models, estimators and specification tests for fractional response variables. Particular attention was dedicated to issues overlooked so far, such as the relevance of choosing the most suitable specification for the conditional mean of the response variable instead of choosing *a priori* the logit or other specific model, the failure in the specification of that conditional mean when one-part decision mechanisms are misspecified as two-part models and vice versa, and the use of GOL and non-nested tests in this framework. New goodness-of-functional-form tests were also proposed and simple procedures for computing LM versions of all tests were discussed.

The extensive Monte Carlo simulation study carried out provided very useful information on the finite-sample performance of the alternative estimators and tests analysed in the paper. First, we confirmed that QML is more attractive than NLS estimation in this framework and that beta-based ML estimators are not robust to

Table 12. Partial Effects for the Models Selected.

	Average sample effects					Population partial effects				
	First part	Second part		Total		First part	Second part		Total	
	Logit	Probit	Loglog	Logit + probit	Logit + loglog	Logit	Probit	Loglog	Logit + probit	Logit + loglog
NDTS	−0.008	0.001	0.001	−0.003	−0.003	−0.008	0.001	0.001	−0.003	−0.003
TANGIB	0.255	−0.002	0.003	0.096	0.097	0.267	−0.002	0.004	0.094	0.096
SIZE	0.089	−0.024	−0.024	0.028	0.028	0.094	−0.024	−0.025	0.028	0.028
PROFITAB	−0.373	−0.531	−0.511	−0.273	−0.271	−0.391	−0.535	−0.531	−0.242	−0.242
GROWTH	0.000	0.001	0.001	0.000	0.000	0.000	0.001	0.001	0.000	0.000
AGE	0.000	−0.001	−0.001	0.000	0.000	0.000	−0.001	−0.001	0.000	0.000
LIQUIDITY	−0.239	−0.034	−0.028	−0.099	−0.097	−0.250	−0.035	−0.029	−0.095	−0.095

Table 13. Partial Effects for One-part Models.

	Average sample effects						Population partial effects					
	OLS	Cauchit	Logit	Probit	Loglog	Cloglog	OLS	Cauchit	Logit	Probit	Loglog	Cloglog
NDTS	−0.001	−0.003	−0.003	−0.003	−0.003	−0.004	−0.001	−0.001	−0.003	−0.003	−0.003	−0.003
TANGIB	0.071	0.074	0.095	0.094	0.090	0.094	0.071	0.033	0.072	0.079	0.086	0.068
SIZE	0.027	0.023	0.029	0.029	0.029	0.028	0.027	0.010	0.022	0.024	0.027	0.021
PROFITAB	−0.132	−0.244	−0.262	−0.252	−0.236	−0.263	−0.132	−0.108	−0.199	−0.214	−0.223	−0.192
GROWTH	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
AGE	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
LIQUIDITY	−0.051	−0.158	−0.104	−0.093	−0.081	−0.108	−0.051	−0.070	−0.079	−0.078	−0.077	−0.078

deviations from the assumed distribution. When the beta assumption is valid, we find that ML outperforms in a sizeable way QML estimation only when the sample size is small and/or the variance of y given x is very large. Second, we showed that for estimating the magnitude of partial effects it is in general very important to choose the correct specification for the conditional mean of y. Finally, we found that both the RESET and GOL tests, which are the most popular tests for assessing the conditional mean assumption made in the related binary regression models in the econometrics and statistics literature, respectively, are not the best option for dealing with fractional regression models. Indeed, the GOFF tests are clearly the best in terms of size and often they are among the most powerful tests, while the P tests, despite over-rejecting the true null hypothesis in some cases, display the best power properties in most cases. However, in cases where the response variable is symmetrically distributed, the GOFF tests exhibit very low power when applied to other symmetric but ill-specified models for the conditional mean of y.

All the techniques discussed in the paper were applied to the regression analysis of the capital structure decisions of Portuguese SMEs. We confirmed recent conjectures by Strebulaev and Yang (2007) that traditional capital structure theories, which consider a single model to explain all financial leverage decisions made by firms, are unable to provide a reasonable explanation for the high percentage of firms that do not use debt at all. Indeed, the specification tests used in our empirical application revealed clearly that the capital structure decisions of Portuguese SMEs, 74.8% of which do not use debt, are best represented by two-part fractional regression models, which is in accordance with the recent papers by Kurshev and Strebulaev (2007) and Ramalho and Silva (2009), who argue that the mechanisms that determine whether a firm uses debt at all are different from the mechanisms that determine the proportion of debt used by firms that do use debt. In particular, we found that firm size may have opposite effects on the two levels of the model, while other variables are important only for one of the two sequential financial leverage decisions made by firms.

Finally, it is important to stress that this paper merely considered estimation and inference of cross-sectional fractional regression models when the outcome is univariate. Therefore, issues such as the internalization of heterogeneity through the use of panel data models (see the recent papers by Wagner (2003) for the logit case and Papke and Wooldridge (2008) and Wagner (2008) for the probit case) or the use of models suitable to deal with multivariate fractional outcomes (e.g. the proportion of income spent in different classes of goods) were not investigated and are important avenues for future research.

Acknowledgements

The authors thank João Santos Silva for helpful comments. Financial support from Fundação para a Ciência e a Technologia is also gratefully acknowledged (grant PTDC/ECO/64693/2006).

Notes

1. Other examples include the estimation of quantiles for fractional data, recently discussed in Machado and Santos Silva (2008).
2. The only alternative to the beta regression model considered so far is based on the simplex distribution developed by Barndorff-Nielsen and Jorgensen (1991). See Song and Tan (2000) and Kieschnick and McCullough (2003) for applications of this distribution in a regression framework.
3. These two-part, hurdle or discrete–continuous mixture models are relatively common in the econometric literature of count data; see Mullahy (1986) for a seminal paper.
4. The simplex density function is

$$f(y;\mu,\phi)=\frac{\exp[-0.5(y-\mu)^2/y(1-y)\mu^2(1-\mu)^2]}{\sqrt{2\pi\phi[y(1-y)]^3}} \quad 0<\mu<1,\ \phi>0$$

   See note 2 for some references on this distribution. Although not reported below, we also computed an ML estimator based on the simplex distribution. The results obtained, which lead to similar conclusions to those described in this paper for the ML estimator based on the beta distribution, are available from the authors upon request.
5. Although not reported, application of similar versions of the P test to other one-part models confirmed categorically their unsuitability for describing the Portuguese SMEs capital structure choices.
6. Except for the industry dummies, we set the dummy relative to the industry comprising the highest percentage of firms at one and the others at zero.

References

Aranda-Ordaz, F.J. (1981) On two families of transformations to additivity for binary response data. *Biometrika* 68(2): 357–363.

Barndorff-Nielsen, O.E. and Jorgensen, B. (1991) Some parametric models on the simplex. *Journal of Multivariate Analysis* 39: 106–116.

Brehm, J. and Gates, S. (1993) Donut shops and speed traps: evaluating models of supervision on police behavior. *American Journal of Political Science* 37(2): 555–581.

Brounen, D., Jong, A. and Koedik, K. (2005) Capital structure policies in Europe: survey evidence. *Journal of Banking and Finance* 30(5): 1409–1442.

Cassar, G. (2004) The financing of business start-ups. *Journal of Business Venturing* 19: 261–283.

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

Cook, D., Kieschnick, R. and McCullough, B.D. (2008) Regression analysis of proportions in finance with self selection. *Journal of Empirical Finance* 15: 860–867.

Czado, C. (1994) Parametric link modification of both tails in binary regression. *Statistical Papers* 35: 189–201.

Czarnitzki, D. and Kraft, K. (2004) Firm leadership and innovative performance: evidence from seven EU countries. *Small Business Economics* 22: 325–332.

Davidson, R. and MacKinnon, J.G. (1981) Several tests for model specification in the presence of alternative hypotheses. *Econometrica* 49(3): 781–793.

Davidson, R. and MacKinnon, J.G. (1984) Convenient specification tests for logit and probit models. *Journal of Econometrics* 25: 241–262.

Faulkender, M. and Petersen, M.A. (2006) Does the source of capital affect capital structure? *Review of Financial Studies* 19(1): 45–79.
Ferrari, S.L.P. and Cribari-Neto, F. (2004) Beta regression for modelling rates and proportions. *Journal of Applied Statistics* 31(7): 799–815.
Frank, M.Z. and Goyal, V.K. (2008) Tradeoff and pecking order theories of debt. In B.E. Eckbo (ed.), *Handbook of Corporate Finance – Empirical Corporate Finance* (Vol. 2, pp. 135–202). Amsterdam: Elsevier.
Gouriéroux, C. and Monfort, A. (1994) Testing non-nested hypotheses. In R.F. Engle and D.L. McFadden (eds), *Handbook of Econometrics* (Vol. IV, pp. 2585–2637). Amsterdam: Elsevier Science.
Gouriéroux, C., Monfort, A. and Trognon, A. (1984) Pseudo maximum likelihood methods: applications to Poisson models. *Econometrica* 52(3): 701–720.
Haab, T.C. and McConnell, K.E. (1998) Referendum models and economic values: theoretical, intuitive, and practical bounds on willingness to pay. *Land Economics* 74(2): 216–229.
Hausman, J.A. and Leonard, G.K. (1997) Superstars in the national basketball association: economic value and policy. *Journal of Labor Economics* 15(4): 586–624.
Hermalin, B.E. and Wallace, N.E. (1994) The determinants of efficiency and solvency in savings and loans. *Rand Journal of Economics* 25(3): 361–381.
Horowitz, J.L. (1994) Bootstrap-based critical values for the information matrix test. *Journal of Econometrics* 61: 395–411.
Kieschnick, R. and McCullough, B.D. (2003) Regression analysis of variates observed on (0, 1): percentages, proportions and fractions. *Statistical Modelling* 3: 193–213.
Koenker, R. and Yoon, J. (2009) Parametric links for binary choice models: a Fisherian-Bayesian colloquy. *Journal of Econometrics*, forthcoming.
Kurshev, A. and Strebulaev, I.A. (2007) Firm size and capital structure. Mimeo.
Lancaster, T. (1984) The covariance matrix of the information matrix test. *Econometrica* 52(4): 1051–1053.
Machado, J.A.F. and Santos Silva, J.M.C. (2008) Quantiles for fractions and other mixed data. Discussion Paper No. 656, Department of Economics, University of Essex.
Maddala, G.S. (1991) A perspective on the use of limited-dependent and qualitative variables models in accounting research. *Accounting Review* 66(4): 788–807.
Madden, D. (2008) Sample selection versus two-part models revisited: the case of female smoking and drinking. *Journal of Health Economics* 27(2): 300–307.
McCullagh, P. and Nelder, J.A. (1989) *Generalized Linear Models* (2nd edn). London: Chapman and Hall.
McDonald, J.F. and Moffitt, R.A. (1980) The uses of tobit analysis. *Review of Economics and Statistics* 62(2): 318–321.
Mullahy, J. (1986) Specification and testing of some modified count data models. *Journal of Econometrics* 33: 341–365.
Nagler, J. (1994) Scobit: an alternative estimator to logit and probit. *American Journal of Political Science* 38(1): 230–255.
Pagan, A. and Vella, F. (1989) Diagnostic tests for models based on individual data: a survey. *Journal of Applied Econometrics* 4: S29–S59.
Paolino, P. (2001) Maximum likelihood estimation of models with beta-distributed dependent variables. *Political Analysis* 9(4): 325–346.
Papke, L.E. and Wooldridge, J.M. (1996) Econometric methods for fractional response variables with an application to 401(k) plan participation rates. *Journal of Applied Econometrics* 11(6): 619–632.
Papke, L.E. and Wooldridge, J.M. (2008) Panel data methods for fractional response variables with an application to test pass rates. *Journal of Econometrics* 145(1/2): 121–133.
Petersen, M.A. and Rajan, R.G. (1994) The benefits of lending relationships: evidence from small business data. *Journal of Finance* 49(1): 3–37.

Poirier, D.J. (1980) A Lagrange multiplier test for skewness in binary models. *Economics Letters* 5: 141–143.
Pregibon, D. (1980) Goodness of link tests for generalized linear models. *Applied Statistics* 29(1): 15–24.
Prentice, R.L. (1976) A generalization of the probit and logit methods for dose response curves. *Biometrics* 32: 761–768.
Rajan, R.J. and Zingales, L. (1995) What do we know about capital structure? Some evidence from international data. *Journal of Finance* 50(5): 1421–1460.
Ramalho, J.J.S. and Silva, J.V. (2009) A two-part fractional regression model for the financial leverage decisions of micro, small, medium and large firms. *Quantitative Finance* 9(5): 621–636.
Ramalho, E.A., Ramalho, J.J.S. and Murteira, J.M.R. (2009) Alternative estimating and testing empirical strategies for fractional regression models. CEFAGE-UE Working Paper 2009/08, University of Evora.
Ramsey, J.B. (1969) Tests for specification errors in classical linear least-squares regression analysis. *Journal of the Royal Statistical Society, Series B* 31(2): 350–371.
Santos Silva, J.M.C., Tenreyro, S. and Windmeijer, F. (2008) Is it different for zeros? Mimeo, University of Essex.
Smith, R.J. (1989) On the use of distributional misspecification checks in limited dependents variable models. *Economic Journal* 99: 178–192.
Smithson, M. and Verkuilen, J. (2006) A better lemon squeezer? Maximum-likelihood regression with beta-distributed dependent variables. *Psychological Methods* 11(1): 54–71.
Song, P.X.-K. and Tan, M. (2000) Marginal models for longitudinal continuous proportional data. *Biometrics* 56: 496–502.
Strebulaev, I.A. and Yang, B. (2007) The mystery of zero-leverage firms. Mimeo, Graduate School of Business, Stanford University.
Stukel, T.A. (1988) Generalized logistic models. *Journal of the American Statistical Association* 83(402): 426–431.
Wagner, J. (2001) A note on the firm size–export relationship. *Small Business Economics* 17: 229–237.
Wagner, J. (2003) Unobserved firm heterogeneity and the size–exports nexus: evidence from German panel data. *Review of World Economics* 139(1): 161–172.
Wagner, J. (2008) Exports and firm characteristics – first evidence from fractional probit panel estimates. Working Paper Series in Economics 97, University of Luneburg.
White, H. (1982) Maximum likelihood estimation of misspecified models. *Econometrica* 50(1): 1–25.
Whittemore, A.S. (1983) Transformations to linearity in binary regression. *SIAM Journal of Applied Mathematics* 43(4): 703–710.
Wooldridge, J.M. (1991a) On the application of robust, regression-based diagnostics to models of conditional means and conditional variances. *Journal of Econometrics* 47: 5–46.
Wooldridge, J.M. (1991b) 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*, Cambridge, MA: MIT Press.

Appendix: Practical Procedures

The aim of this appendix is to provide practitioners with a simple guide for dealing with fractional responses.

A1: *Model Estimation*

The results reported in this paper were obtained using the statistical software *R*, which requires some programming experience. However, Stata possesses already canned commands that allow most of the models discussed in the paper to be computed in a single command line, as described next.

Stata command line for estimating conditional mean models by QML:

$$\text{glm } y \; X1 \; \ldots \; Xk, \text{link}(ff) \text{ family(binomial) robust}$$

where $Xj, j = 1, \ldots, k$, denote the explanatory variables and *ff* is the designation of the functional form chosen for $G(\cdot)$ (probit, logit, loglog or cloglog – without programming, it is not possible to consider a cauchit specification in Stata). When used in the second stage of two-part models

$$\text{glm } y \; X1 \; \ldots \; Xk \text{ if } y > 0, \; \text{link}(ff) \text{ family(binomial) robust}$$

Stata command line for estimating the beta regression model:

$$\text{betafit } y, \; \text{muvar}(X1 \; \ldots \; Xk)$$

which requires the previous installation of the package betafit.ado. It is also possible to estimate the variant of the beta regression model considered by Paolino (2001) and Smithson and Verkuilen (2006) using

$$\text{betafit } y, \; \text{muvar}(X1 \; \ldots \; Xk) \; \text{phivar}(Z1 \; \ldots \; Zm)$$

where $Zj, j = 1, \ldots, m$, are the variables that enter in the specification of the shape parameter.

A2: *LM Statistics*

All the tests for conditional mean assumptions may be implemented as LM statistics, which require only the computation of linear regressions and, hence, may be performed in a straightforward way using Stata or any other statistical software. Next, we summarize the computation of these statistics.

Binary and beta regression models:

1. Obtain the predicted outcomes $\hat{G}$, the derivatives $\hat{g}$ and the residuals $\hat{u}$ from the null model.
2. Construct the weights $\hat{\omega} = [\hat{G}(1 - \hat{G})]^{-0.5}$, the variables $\tilde{u} = \hat{u}\hat{\omega}$ and $\tilde{g} = \hat{g}\hat{\omega}$ and the vectors $\tilde{g}x$ and $\tilde{g}z$, where x denotes the covariates from the null model and z the omitted variables that characterize the tests discussed in Section 4.1.
3. Regress $\tilde{u}$ on $\tilde{g}x$ and $\tilde{g}z$.
4. Compute LM = ESS (binary model) or LM $= nR^2$ (beta model).

Fractional regression models estimated by QML:

1. Obtain the predicted outcomes $\hat{G}$, the derivatives $\hat{g}$ and the residuals $\hat{u}$ from the null model.

2. Construct the weights $\hat{\omega} = [\hat{G}(1 - \hat{G})]^{-0.5}$, the variables $\tilde{u} = \hat{u}\hat{\omega}$ and $\tilde{g} = \hat{g}\hat{\omega}$ and the vectors $\tilde{g}x$ and $\tilde{g}z$.
3. Regress separately each element of the J-dimensional vector $\tilde{g}z$ on the entire vector $\tilde{g}x$ and save the residuals from each regression (denote them by $\tilde{r}_j,\ j = 1, \ldots, J$).
4. Find the products between $\tilde{u}$ and $\tilde{r}_j$ (for all observations) and form the J-dimensional vector $\tilde{u}\tilde{r}$.
5. Run the regression of 1 on $\tilde{u}\tilde{r}$ without an intercept.
6. Compute LM = ESS = n − SSR.