Oberhofer, Harald; Pfaffermayr, Michael

Working Paper

Two-Part Models for Fractional Responses Defined as Ratios of Integers

WIFO Working Papers, No. 472

Provided in Cooperation with:
Austrian Institute of Economic Research (WIFO), Vienna

Suggested Citation: Oberhofer, Harald; Pfaffermayr, Michael (2014) : Two-Part Models for Fractional Responses Defined as Ratios of Integers, WIFO Working Papers, No. 472, Austrian Institute of Economic Research (WIFO), Vienna

This Version is available at:
https://hdl.handle.net/10419/129022

WIFO
ÖSTERREICHISCHES INSTITUT
FÜR WIRTSCHAFTSFORSCHUNG

WORKING PAPERS

Two-Part Models for Fractional Responses Defined as Ratios of Integers

Harald Oberhofer, Michael Pfaffermayr

472/2014

WIFO

ÖSTERREICHISCHES INSTITUT FÜR WIRTSCHAFTSFORSCHUNG
AUSTRIAN INSTITUTE OF ECONOMIC RESEARCH

Two-Part Models for Fractional Responses Defined as Ratios of Integers

Harald Oberhofer, Michael Pfaffermayr

WIFO Working Papers, No. 472

June 2014

Abstract

This paper discusses two alternative two-part models for fractional response variables that are defined as ratios of integers. The first two-part model assumes a Binomial distribution and known group size. It nests the one-part fractional response model proposed by Papke and Wooldridge (1996) and thus, allows to apply Wald, LM and/or LR tests in order to discriminate between the two models. The second model extends the first one by allowing for overdispersion. Monte Carlo studies reveal that, for both models, the proposed tests are equipped with sufficient power and are properly sized. Finally, we demonstrate the usefulness of the proposed two-part models for data on the 401(k) pension plan participation rates used in Papke and Wooldridge (1996).

E-mail address: Harald.Oberhofer@sbg.ac.at, Michael.Pfaffermayr@wifo.ac.at
2014/156/W/0

Medieninhaber (Verleger), Hersteller: Österreichisches Institut für Wirtschaftsforschung • 1030 Wien, Arsenal, Objekt 20 • Tel. (43 1) 798 26 01-0 • Fax (43 1) 798 93 86 • http://www.wifo.ac.at/ • Verlags- und Herstellungsort: Wien

Two-Part Models for Fractional Responses defined as Ratios of Integers*

Harald Oberhofer* Michael Pfaffermayr‡

June 12, 2014

Abstract

This paper discusses two alternative two-part models for fractional response variables that are defined as ratios of integers. The first two-part model assumes a Binomial distribution and known group size. It nests the one-part fractional response model proposed by Papke and Wooldridge (1996) and thus, allows to apply Wald, LM and/or LR tests in order to discriminate between the two models. The second model extends the first one by allowing for overdispersion. Monte Carlo studies reveal that, for both models, the proposed tests are equipped with sufficient power and are properly sized. Finally, we demonstrate the usefulness of the proposed two-part models for data on the 401(k) pension plan participation rates used in Papke and Wooldridge (1996).

JEL Codes: C12, C15, C25, C52.

Keywords: Fractional response models for ratios of integers, one-part versus two-part models, Wald test, LM test, LR test.

*We would like to thank Joaquim Ramalho for his comments on a previous version of this paper.

*Department of Economics and Social Sciences and Salzburg Centre of European Union Studies (SCEUS), University of Salzburg and The Austrian Center for Labor Economics and the Analysis of the Welfare State. Address: Residenzplatz 9, 5010 Salzburg, Austria. E-mail: Harald.Oberhofer@sbg.ac.at.

‡Department of Economics, University of Innsbruck and Austrian Institute of Economic Research (WIFO). Address: Universitaetsstrasse 15, 6020 Innsbruck, Austria. E-mail: Michael.Pfaffermayr@uibk.ac.at.

1 Introduction

Many empirical studies deal with fractional response data that are bounded in the [0,1] interval and, in addition, contain a significant amount of observations at the boundary values of 0 or 1. In their seminal paper Papke and Wooldridge (1996) propose a one-part fractional response model that extends the generalized linear model (GLM) literature from statistics.[1] In particular, they introduce a quasi-maximum likelihood (QMLE) approach that only requires the correct specification of the conditional mean to consistently estimate one-part fractional response models. In this framework, there is no need for an *ad hoc* transformation of the boundary values of 0 or 1.[2]

If the data at hand contain a large share of these boundary values, the econometric literature alternatively offers two-part models that assume a different data generating process (DGP) for the zeros or ones, respectively.[3] In order to empirically discriminate between the competing one-part and two-part fractional response models, the literature so far typically applies a P test for non-nested hypotheses as described in Davidson and MacKinnon (1981) and Ramalho, Ramalho and Murteira (2011).

In many empirical applications the fractional response variable is defined as ratio of integers such as e.g., the share of employees participating in a voluntary pension plan (Papke and Wooldridge 1996), where the group size is known. This additional group size information can explicitly be used for the empirical analysis.[4] Starting from this observation, we propose two-part models that exploit information on the group size and, additionally, nest the one-part alternatives following the approach suggested by Lin and Schmidt (1984) and Mullahy (1986). Lin and Schmidt (1984) propose an LM test of the Tobit model against an alternative two-part model that nests the Tobit model as a special case, while Mullahy (1986) applies this approach to count data hurdle models.

In this paper, the first two-part model is based on the binomial likelihood framework, while the second ones additionally allows for overdispersion in the data by assuming a

[1] In a more recent paper, Papke and Wooldridge (2008) discuss fractional response models for panel data. Ramalho *et al.* (2011) provide a comprehensive up-to-date overview on the econometrics of fractional response models.

[2] As an alternative to this QMLE approach, scholars have also proposed to assume a beta distribution and estimate the resulting model via maximum likelihood (see, e.g., Paolino 2001, Kieschnick and McCullough 2003, Ferrari and Cribari-Neto 2004). Beta regression models, however, are not able to deal with the boundary values of 0 and 1 without an *ad hoc* transformation.

[3] See, e.g., Mullahy (1986), Lambert (1992), Cameron and Trivedi (2005) Wooldridge (2002, Problem 19.8), Ramalho and Silva (2009, 2013), Ramalho *et al.* (2011) and Oberhofer and Pfaffermayr (2012). In the beta regression framework, two-part models are proposed by e.g., Cook, Kieschnick and McCullough (2008) and Ospina and Ferrari (2012). Hall (2000) proposes a zero-inflated binomial model.

[4] In the context of multivariate fractional response variables, Murteira and Ramalho (2014) discuss the usefulness of the group size information for formulating econometric models that are based on the multinomial and Dirchilet-multinomial distributions, respectively.

beta-binomial likelihood function. This latter model, for example, is able to account for correlated individual zero and one decisions that could be triggered by group specific random effects.

Applying a maximum likelihood framework, this paper derives explicit formulas for the Wald and the LM tests, respectively, which could be used as alternatives to the available non-nested P test. Two Monte Carlo simulation exercises reveal that both proposed tests are properly sized and equipped with sufficient power to discriminate between the two-part model and its (nested) one-part alternative. Finally, we apply the different estimators to firm-level data on 401(k) pension plan participation rates as used in Papke and Wooldridge (1996) and document that participation decisions are highly correlated within firms.

The remainder of the paper is organized as follows: Section 2 presents the nested two-part models for fractional response variables that are defined as ratio of integers. Section 3 reports the main findings from two small-scale Monte Carlo exercises. Section 4 offers an empirical application for 401(k) plan participation rates and in Section 5 we provide some concluding remarks.

2 Generalized Two-Part Fractional Response Models

2.1 The Generalized Binomial Two-Part Model

The typical fractional response model is based on the Bernoulli or binomial distribution. Assume there are $i = 1, \ldots, N$ groups (e.g., firms) in which $j = 1, \ldots, n_i$ units (workers) are confronted with a 0/1 decision (e.g., to participate in a voluntary pension plan). We focus on situations where the number of units, n_i, is observed as in Papke and Wooldridge (1993, 1996) and assume that n_i is exogenously given so that it is appropriate to condition on it. The probability that unit j in group i opts for 1 (e.g., to participate in a voluntary pension plan) is denoted by θ_i which is assumed to be group- but not unit-specific. The number of units within a group choosing 1 is denoted by k_i and the corresponding observed share (at the group level) is given by $y_i = \frac{k_i}{n_i}$ with $0 \leq y_i \leq 1$ or $k_i = n_i y_i$, respectively.[5] Following Papke and Wooldridge (1996), for such a set-up the conditional expectation of

[5]Note, in comparison to the fractional response model analyzed in Papke and Wooldridge (1996), the individual contributions to the likelihood, the estimated score and the estimated information matrix are all multiplied by n_i in this model (see also Papke and Wooldridge, 1993, pp. 10-11).

the fractional response variable y_i is group-specific and can be specified as

$$E(y_i|x_i, n_i) = G(x_i\beta), \qquad i = 1, \ldots, N, \tag{1}$$

where (the $1 \times k$ vector) x_i refers to a set of i-specific explanatory variables with the corresponding parameter vector β. Typically, $G(.)$ is a cumulative distribution function (cdf) such as the logistic function $G(z) = \exp(z)/(1 + \exp(z))$ which maps z to the $(0, 1)$ interval.[6] In this case, the group specific contributions to the log likelihood can be written as

$$\ln(f(\beta; y_i, x_i)) = n_i(y_i \ln(G(x_i\beta)) + (1 - y_i) \ln(1 - G(x_i\beta))) + const. \tag{2}$$

Following Wooldridge (2002, Problem 19.8), Cameron and Trivedi (2005, p. 680), Ramalho and Silva (2009, 2013), Ramalho *et al.* (2011) and Oberhofer and Pfaffermayr (2012), one may consider a two-part model to explicitly account for an excessive number of boundary values. Here, we concentrate on the case of boundary values at one, but similar arguments apply to the case of an excessive number of zeros. In contrast to the one-part model, the two-part alternative assumes a different data generating process for the boundary values. For notational simplicity, the explanatory variables in the first and second part of the model are assumed to be the same, but in general they could differ. Formally, this (generalized) two-part model can be defined as in Cameron and Trivedi (2005, pp. 545, 680) and is given by

$$f(y_i|x_i, n_i) = \begin{cases} P_1(n_i, x_i) & \text{if} \quad y_i = 1 \quad \text{or} \quad k_i = n_i \\ (1 - P_1(n_i, x_i))\frac{P_2(k_i, x_i)}{1 - P_2(n_i, x_i)} & \text{if} \quad y_i < 1 \quad \text{or} \quad k_i < n_i \end{cases}, \tag{3}$$

where $P_1(n_i, x_i) = P_1(K_i = n_i, x_i)$ refers to the first part of the model and $P_2(k_i, x_i) = P_2(K_i = k_i, x_i)$, $k_i = 1, ..., n_i$ to its second part. Under independent unit decisions, K_i is assumed to be distributed as Binomial with conditional probabilities

$$\begin{aligned} P_1(n_i, x_i) &= \theta_{i1}^{n_i} \\ P_2(n_i y_i, x_i) &= \binom{n_i}{n_i y_i} \theta_{i2}^{n_i y_i} (1 - \theta_{i2})^{n_i - n_i y_i}. \end{aligned} \tag{4}$$

where for $0 < \theta_{i1} < 1$ the probability of y_i amounting exactly to 1 is given by $\theta_{i1}^{n_i}$ under $P_1(n_i, x_i)$ and $\theta_{i2}^{n_i}$ under $P_2(n_i, x_i)$.[7]

[6] See Ramalho *et al.* (2011, 2013) for a comprehensive discussion on alternative functional forms for one-part and two-part fractional response models.

[7] In the one-part model it holds that $P_2(n_i, x_i) = P_1(n_i, x_i)$ or $\theta_{i1} = \theta_{i2}$ and $f(y_i|x_i, n_i)$ reduces to $P_2(k_i, x_i)$.

Under this two-part model, we specify the probability of observing a share of 1 by $P_1(n_i, x_i) = \theta_{i1}^{n_i}$ assuming $\theta_{i1} = G(x_i\gamma)$. The second part of the model for values $y_i < 1$ is based on the conditional distribution:

$$f(y_i|y_i < 1, x_i, n_i) = (1 - P_1(n_i, x_i))\frac{P_2(k_i, x_i)}{1 - P_2(n_i, x_i)}$$

implying that the probability distribution $f(y_i|x_i, n_i)$ is divided by $1 - G(x_i\beta)^{n_i}$ to ensure that the conditional probabilities sum up to 1. The conditional mean of the two-part model, thus, is given by[8]

$$\begin{aligned} E(y_i|x_i, n_i) &= (1 - P_1(n_i, x_i))\, E(y_i|y_i < 1, x_i, n_i) + P_1(n_i, x_i) \\ &= \tfrac{1-G(x_i\gamma)^{n_i}}{1-G(x_i\beta)^{n_i}}(G(x_i\beta) - G(x_i\beta)^{n_i}) + G(x_i\gamma)^{n_i}. \end{aligned} \tag{5}$$

Equation (5) shows that in case of $\gamma = \beta$ the conditional mean of this two-part model reverts to the simple one-part formulation. The standard two-part literature typically uses a simplified version of the conditional mean which ignores the fact that the second part of the model also assigns a non-zero probability to boundary values. For example, Ramalho and Silva (2009, p. 630) specify the conditional mean $E(y_i|y_i > 0, x_i, n_i)$ as $G(x_i\beta)$.

Defining $z_i = 1$ if $y_i = 1$ and 0 otherwise, the likelihood of the two part-model contains the individual contributions

$$\begin{aligned} \ln(f(\gamma, \beta; y_i, x_i)) &= (1 - z_i)[n_i(y_i \ln(G(x_i\beta)) + (1 - y_i) \ln(1 - G(x_i\beta))) - \ln(1 - G(x_i\beta)^{n_i})] \\ &\quad +(1 - z_i) \ln(1 - G(x_i\gamma)^{n_i}) + z_i n_i \ln(G(x_i\gamma)) + constant. \end{aligned} \tag{6}$$

Under this specification maximum likelihood estimation is straight forward, since it separates into the estimation of the model explaining $P(y_i = 1|x_i, n_i)$ using all observations and the estimation of the fractional response model for the observations with $y_i < 1$ only. In the following, we assume that the distributions, upon which the one-part and two-part models are based, are correctly specified and concentrate on maximum likelihood estimation.

The main advantage of the proposed two-part model is that it nests the one-part fractional response model since, as demonstrated in equations (5) and (6), under $\theta_{i1}^{n_i} = \theta_{i2}^{n_i}$

[8] In case of zero boundary values the conditional mean of this two-part fractional response model modifies to

$$E(y_i|x_i, n_i) = P(y_i > 0|x_i, n_i)E(y_i|y_i > 0, x_i, n_i) = \frac{1 - (1 - G(x_i\gamma))^{n_i}}{1 - (1 - G(x_i\beta))^{n_i}}G(x_i\beta).$$

(or equivalently, $\gamma = \beta$) the two-part-model reverts to the one-part fractional response model.[9] In case of x_i being the same for the one-part and the two-part model and their parameters being equal under ($\gamma = \beta$) the two models coincide and have the same likelihood functions. This hypothesis can easily be tested by an LM or a Wald test of $H_0 : \gamma = \beta$. If one or both parts of the two-part model contain additional explanatory variables denoted by w_{1i} and w_{2i} with parameter vectors ϕ_1 and ϕ_2, respectively, the underlying H_0 to test is $\gamma = \beta$, $\phi_1 = 0$, $\phi_2 = 0$.[10]

In Appendix A1, we derive an LM test which is based on the estimated parameters of the one-part fractional response model that are indexed by OP. Similar, to Mullahy (1986) the LM test uses the parametrization $\gamma = \beta + \delta$ and tests H_0: $\delta = 0$ vs. H_0: $\delta \neq 0$. It is easy to calculate and is given by

$$LM = \widehat{s}'_{\delta,OP} \left(A^{-1}(\widehat{\beta}_{OP}) + \left(B(\widehat{\beta}_{OP}) - A(\widehat{\beta}_{OP}) \right)^{-1} \right) \widehat{s}_{\delta,OP}. \tag{7}$$

Thereby, $\widehat{s}_{\delta,OP} = \sum_{i=1}^{N} \widehat{C}_{i\beta,OP}(z_i - G(x_i\widehat{\beta}_{OP})^{n_i})x_i'$ and $\widehat{C}_{i\beta,OP} = n_i \frac{1-G(x_i\widehat{\beta}_{OP})}{1-G(x_i\widehat{\beta}_{OP})^{n_i}}$. $A(\widehat{\beta}_{OP}) = \sum_{i=1}^{N} \widehat{C}^2_{i\beta,OP}(1 - G(x_i\widehat{\beta}_{OP})^{n_i})G(x_i\widehat{\beta}_{OP})^{n_i} x_i' x_i$ and $B(\widehat{\beta}_{OP}) = \sum_{i=1}^{N} n_i((1 - G(x_i\widehat{\beta}_{OP})) * G(x_i\widehat{\beta}_{OP}))x_i' x_i$. Note x_i is defined as $1 \times k$ vector. Under standard assumptions this LM test is asymptotically distributed as $\chi^2(k)$.

Appendix A2 derives a Wald test statistic that uses the parameter estimates of the two-part model with index TP and is given by[11]

$$\widehat{W} = (\hat{\gamma}_{TP} - \hat{\beta}_{TP})' \left(A(\widehat{\gamma}_{TP})^{-1} + \left(B(\widehat{\beta}_{TP}) - A(\widehat{\beta}_{TP}) \right)^{-1} \right)^{-1} (\hat{\gamma}_{TP} - \hat{\beta}_{TP}), \tag{8}$$

which is likewise asymptotically distributed as $\chi^2(k)$.

[9] In a related setting, Lin and Schmidt (1984) derive an LM test for testing a Tobit model against the more general Cragg's two-part model under normality using a similar nesting hypothesis. Mullahy (1986) proposes LM and Hausman test statistics in order to discriminate between one- and two-part (hurdle) count data models.

[10] In the quasi maximum likelihood framework, the literature commonly applies non-nested P tests to discriminate between the one-part and two-part fractional response models. Following Davidson and MacKinnon (1981) and Ramalho *et al.* (2011) the P test for the null hypothesis that the one-part model is the true one and the two-part model is the alternative is based on an artificial regression. However, in their propositions 4.1.2 and 4.3.2 Gourieroux, Monfort and Trognon (1984) prove that under the nested parametrization this test is not applicable.

[11] When applying the Wald test it is crucial to use the weighted form of the likelihood given in (3) or to assume $n_i = n$. This is necessary because the likelihood is not defined for a group size of $n_i = 1$.

2.2 The Beta-Binomial Two-part Model

In many cases the assumption of independent zero-one decisions of the individuals is not plausible as there may exist pronounced overdispersion. To give an example, the presence of group specific random effects generates equi-correlation within groups (see, e.g., Heckman and Willis 1977 and McCulloch and Searle 2001) and violates the independence assumption of the binomial distribution made above. Following Heckman and Willis (1977), Prentice (1986), McCulloch and Searle (2001) and Santos Silva and Murteira (2009) for these situations a beta-binomial model forms a plausible and tractable alternative.[12] While for such situations, one still obtains consistent estimates using the Bernoulli-QMLE for the one-part model under H_0 (see Papke and Wooldridge 1996), the two-part model based on the binomial distribution (6) cannot be estimated consistently when the beta-binomial is the true data generating process. As a result, the above proposed Wald and LM tests are also invalid in this more general setting.

Following Prentice (1986), we assume that the random variable K_i is distributed as beta-binomial taking values $k_i = n_i y_i$ with probabilities

$$\begin{aligned} P_2(K_i = k_i, x_i) &= \binom{n_i}{k_i} \int_0^1 \pi_i^{k_i+a_{i2}-1}(1-\pi_i)^{n_i-k_i+b_{i2}-1} \tfrac{\Gamma(a_{i2}+b_{i2})}{\Gamma(a_{i2})\Gamma(b_{i2})} d\pi_i \\ &= \binom{n_i}{k_i} \tfrac{\Gamma(k_i+a_{i2})\Gamma(n_i-k_i+b_{i2})}{\Gamma(n_i+a_{i2}+b_{i2})} \tfrac{\Gamma(a_{i2}+b_{i2})}{\Gamma(a_{i2})\Gamma(b_{i2})} = \binom{n_i}{k_i} \frac{\prod_{j=0}^{k_i-1}(a_{i2}+j) \prod_{j=0}^{n_i-k_i-1}(b_{i2}+j)}{\prod_{j=0}^{n_i-1}(a_{i2}+b_{i2}+j)}. \end{aligned}$$

This distribution results from a data generating process that is based on $K_i|\pi_{i2} \sim Bin(n_i, \pi_{i2})$ and $\pi_{i2} \sim Beta(a_{i2}, b_{i2})$. In order to parametrize the model, we specify

$$\theta_{i2} := E[\pi_i|x_i] = \tfrac{a_{i2}}{a_{i2}+b_{i2}},\ c = a_{i2} + b_{i2},$$

or

$$a_{i2} = c\theta_{i2},\ b_{i2} = c(1-\theta_{i2}),\ \theta_{i2} = \tfrac{e^{x_i\beta}}{1+e^{x_i\beta}}$$

This model introduces intra-group correlation und, therefore, overdispersion (see Prentice 1986 and McCulloch and Searle 2001). A model with group-specific random effects that follow a $Beta(a_{i2}, b_{i2})$ distribution yields the same model structure (see McCulloch and Searle 2001).

[12] The Dirchilet-multinomial distribution allows to generalize the beta-binomial model for multivariate fractional response data with more than two alternative choices (see, e.g., Johnson, Kemp and Kotz 2005, Mullahy 2010, Murteira and Ramalho 2014).

Letting, $K_i = \sum_{j=1}^{n_i} K_{ij}$, where K_{ij} are correlated Bernoulli-random variables that take the value 1 with probability π_i and $\pi_i \sim Beta(a_{i2}, b_{i2})$, the variance and covariance are given by

$$\begin{aligned} Var[K_i] &= n_i \tfrac{a_i b_i}{c^2} + n_i (n_i - 1) \tfrac{a_i b_i}{c^2(1+c)} \\ &= n_i \theta_{i2}(1-\theta_{i2})(1 + (n_i - 1)\rho), \end{aligned} \tag{9}$$

and

$$Cov[K_{ij}, K_{il}] = \tfrac{a_i b_i}{c^2(1+c)} \text{ and } \rho = corr[K_{ij}, K_{il}] = \tfrac{1}{1+c}, \tag{10}$$

respectively.[13] For the boundary value $K_i = n_i$ we specify $P_1(n_i, x_i)$ analogously and assume that

$$P_1(n_i, x_i) = \frac{\Gamma(n_i + a_{i1})}{\Gamma(n_i + a_{i1} + b_{i1})} \frac{\Gamma(a_{i1} + b_{i1})}{\Gamma(a_{i1})}, \tag{11}$$

where $a_0 = c\theta_{i1}$, $b_{i1} = c(1 - \theta_{i1})$. For $P_1(n_i)$ the nested specification is given by $\theta_{i1} = \frac{e^{x_i\gamma}}{1+e^{x_i\gamma}} = \frac{e^{x_i(\beta+\delta)}}{1+e^{x_i(\beta+\delta)}}$.

We impose the same parameter c for $P_1(n_i)$ and $P_2(k_i)$ as a different value for $P_1(n_i)$ remains unidentified for zero-one dependent variables. For ease of exposition c is assumed to be independent of i. Principally, the parameter c could be made dependent on explanatory variables x_i.[14] In this case, a convenient parametrization would be $c_i = 2(e^{z_i\vartheta} - 1)^{-1}$ so that $\rho = 2\frac{e^{z_i\vartheta}}{1+e^{z_i\vartheta}} - 1$, where z_i denotes the vector or explanatory variables for c_i and ϑ the corresponding parameter vector (see Prentice 1986). This parametrization guarantees that the coefficient of intra-group correlation is restricted to the [0,1] interval.

Given the maximum likelihood estimates (see Appendix B for details) of the unconstrained model, it is straight forward to use Wald or likelihood ratio (LR) tests for model discrimination between the two-part and one-part models. The resulting H_0 to test would again be $H_0: \ \gamma = \beta$ or $\delta = 0$ vs. $H_1: \ \gamma \neq \beta$ or $\delta \neq 0$. An LM test is infeasible in this context, as a simple closed form of the expected Hessian of that model cannot be derived analytically under H_0.

[13] Note, this model only allows for positive intra-group correlation as one has to assume that $a_i > 0$ and $b_i > 0$. Prentice (1986) proposes a transformation of the Beta-binomal distribution that is also able to handle negative intra-group correlation.

[14] In the empirical application in Section 4 we introduce one specification where c depends on the matching rate, firm size and the age of the pension plan. For further details see columns (7) and (8) in Table 3 and the corresponding discussion in the text.

3 Monte Carlo Exercises

3.1 The Binomial Two-Part Model

To investigate the performance of the proposed tests in finite samples we set up two small-scale Monte Carlo simulation exercises. In this section we concentrate on independent individual decisions and, thus, apply the generalized binomial two-part model. In Section 3.2 we introduce a group-specific random effect and examine the performance of our tests for the beta-binomial case.

For the binomial model, we generate Bernoulli random variables using the logistic cdf $G(x_i + 0.5)$, where x_i is distributed uniformly over $[0, 5]$ and held fixed in repeated samples. To obtain a share variable we divide the resulting Bernoulli random number by n_i and take $n_i \sim iid\ N(100, 5)$. In the lower panel of Table 1 we increase $\bar{n}$ to 200. The probabilities for the boundary values of 1 are based on $\theta_{i1} = G(\alpha(x_i + 0.5))^{n_i}$, where α varies from 0.95 to 1.05 so that at $\alpha = 1$ the one-part model is the true one. The dummy variable for boundary values takes the value 1 if a generated uniformly distributed random variable is lower than θ_{i1} and 0 otherwise. We run each Monte Carlo experiment 10,000 times for sample sizes of 500 and 1,000 observations, respectively, and calculate the size of the tests as the shares of rejections at $\alpha = 1$ and a nominal size of 5 percent. Hence, the corresponding 95-percent confidence interval of the simulated size is given by [0.046,0.054]. The power of each test is defined as the share of rejections at $\alpha \neq 1$.

The results of this simulation exercise are summarized in Table 1. The second and the fifth columns report the share of ones for sample sizes of $N = 500$ and $N = 1,000$, respectively. Concentrating on $N = 500$, under H_0, 14 percent of all observations take on the boundary value of 1. Increasing $\bar{n}$ to 200 reduces the share of ones to 6 percent under H_0. Columns 2 and 5 also reveal that an increase in α leads to an increase in the share of boundary values.

The Monte Carlo simulation results indicate that the simulated size of the Wald and the LM tests is within the 95-percent confidence interval in all experiments, ranging from 0.046 to 0.051. In a similar vein, both tests have power in both directions $\alpha < 1$ and $\alpha > 1$, and as expected their power increases with sample size. To give an example, for $N = 1,000$, $\bar{n} = 200$ and $\alpha = 1.05$ the Wald and LM tests amount to 0.791 and 0.823, respectively. To sum up, both the power and size of these tests are suitable and they might be considered as valuable alternatives to the already available non-nested P test if independent 0/1 decisions within groups can be assumed and n_i is observed.

Table 1: Monte Carlo Simulation for the Binomial Model, 10,000 Replications

	α	Share of ones	Wald test	LM test	Share of ones	Wald test	LM test
			$N = 500$			$N = 1,000$	
$\bar{n} = 100$							
	0.95	0.11	0.428	0.385	0.11	0.748	0.725
	0.96	0.11	0.287	0.253	0.12	0.531	0.503
	0.97	0.12	0.175	0.151	0.12	0.326	0.299
	0.98	0.12	0.103	0.087	0.13	0.167	0.149
	0.99	0.13	0.059	0.053	0.13	0.080	0.069
	1.00	**0.14**	**0.046**	**0.049**	**0.14**	**0.049**	**0.051**
	1.01	0.14	0.061	0.071	0.15	0.072	0.082
	1.02	0.15	0.101	0.123	0.15	0.162	0.178
	1.03	0.15	0.178	0.205	0.16	0.323	0.348
	1.04	0.16	0.269	0.302	0.16	0.517	0.539
	1.05	0.16	0.412	0.449	0.17	0.719	0.737
$\bar{n} = 200$							
	0.95	0.04	0.419	0.343	0.04	0.754	0.710
	0.96	0.04	0.272	0.212	0.05	0.544	0.491
	0.97	0.05	0.166	0.126	0.05	0.340	0.291
	0.98	0.05	0.095	0.069	0.05	0.163	0.132
	0.99	0.06	0.058	0.049	0.06	0.074	0.062
	1.00	**0.06**	**0.046**	**0.048**	**0.06**	**0.049**	**0.050**
	1.01	0.06	0.062	0.081	0.07	0.084	0.101
	1.02	0.07	0.113	0.145	0.07	0.179	0.214
	1.03	0.07	0.192	0.242	0.08	0.357	0.402
	1.04	0.08	0.319	0.380	0.08	0.586	0.630
	1.05	0.08	0.460	0.526	0.08	0.791	0.823

3.2 The Beta-Binomial Two-Part Model

This subsection briefly discusses the main findings from a second Monte Carlo exercises that investigates the power and size of Wald and LR tests for the beta-binomial model for fractional response data that are defined as fractions of integers.[15]

We generate beta-binomial random variables assuming that $c_i \in \{1, 3\}$ which induces inter-group correlation ρ of 0.5 and 0.25, respectively. With ρ at hand, we draw random numbers from the Beta distribution in order to obtain group-specific probabilities. These probabilities in turn are used when drawing final (beta-)binomial random variables. The probabilities for the boundary values of 1 are based on $\frac{a_{i1}}{c} = \theta_{i1} = G(\alpha(0.75x_i + 0.5))$, while the second part is based on $G(0.75x_i + 0.5)$. The size of the Wald and LR tests is again measured as share of rejections at $\alpha = 1.0$ and the power can be inferred from experiments with $\alpha \neq 1.0$.

[15] As already pointed out, for this model no closed form solution of the expected Hessian can be derived and thus LM tests are not feasible in this setting.

Table 2: Monte Carlo Simulation for the Beta-Binomial Model, 10,000 Replications

	ρ	α	Share of ones	Wald test	LR test	Share of ones	Wald test	LR test
				$N = 500$			$N = 1,000$	
$\bar{n} = 100$								
	0.5	0.6	0.14	0.343	0.494	0.14	0.734	0.805
	0.5	0.8	0.18	0.068	0.144	0.18	0.167	0.243
	0.5	**1.0**	**0.23**	**0.047**	**0.049**	**0.23**	**0.051**	**0.055**
	0.5	1.2	0.28	0.168	0.113	0.29	0.242	0.180
	0.5	1.4	0.34	0.398	0.277	0.34	0.625	0.511
	0.25	0.6	0.14	0.734	0.805	0.02	0.987	0.991
	0.25	0.8	0.18	0.167	0.243	0.03	0.538	0.582
	0.25	**1.0**	**0.23**	**0.051**	**0.055**	**0.050**	**0.048**	**0.052**
	0.25	1.2	0.29	0.242	0.180	0.08	0.628	0.597
	0.25	1.4	0.34	0.625	0.511	0.12	0.997	0.997

Table 2 reports the results from the Monte Carlo exercise for $n_i \sim iid\ N(100, 5)$, sample sizes of 500 and 1,000 observations, respectively again using 10,000 replications. The within-group correlation is 0.5 (0.25) in the upper (lower) part of Table 2. To start with, this Monte Carlos exercise indicates that the Wald and LR tests that are obtained from the estimation of the Beta-binomial model are all properly sized. More precisely, for $\alpha = 1.0$ the share of rejections ranges from 0.047 to 0.055. Both tests also exhibit sufficient power in both directions. Moreover, the power increases with sample size and tends to be larger for lower values of ρ. To sum up, this small Monte Carlo exercise indicates that the Beta-binomial two-part model is an attractive alternative for fractional response data that are characterized by inter-group correlation which might be induced by group-specific random effects. Moreover, the suggested Wald and LR tests seem to be useful to discriminate between the one-part and two-part models in this more general setting.

4 An Empirical Application: The 401(k) Pension Plan Participation Rates

This section offers an application of the nested two-part fractional response models for fractional responses defined as ratios of integers using 401(k) pension plan participation rates data that have also been used by Papke and Wooldridge (1996). In order to compare our estimation results with the non-weighted one-part model proposed by Papke and

Wooldridge (1996) we also replicate their results.[16] Moreover, the beta-binomial model also allows to highlight that the 401(k) plan participation rates are characterized by non-negligible intra-firm correlation. Finally, we document the usefulness of the proposed Wald, LM and LR tests for discriminating between one-part and two-part fractional response models.

In their empirical application, Papke and Wooldridge (1996) model the participation in 401(k) pension plans using a sample of 4,734 US manufacturing firms. The dependent variable (PRATE) is measured as the fraction of active 401(k) pension plan accounts relative to the overall number of eligible employees which amounts to one in 42.73 percent of all observations (i.e., 2,023 firms). The vector of covariates contains a firm's matching rate (MRATE), the firms overall number of employees (log (EMP)), the pension plan's age (AGE) as well as an indicator variable (SOLE) that takes on the value of 1 if the 401(k) pension plan is the only one offered by the firm. In their most general specification, Papke and Wooldridge (1996) include squared terms of the former three covariates in order to control for additional non-linearities. Further details on different specifications and sub-sample results can be found in Papke and Wooldridge (1996).

Table 3 reports the parameter estimates for various different fractional response models. To start with, column (1) replicates the results from column (4) of Table III in Papke and Wooldridge (1996). These results are based on the non-employment weighted QMLE estimator for the one-part fractional response model. In column (2) we apply the same QMLE estimator but additionally weight the observations by each firm's number of employees. Columns (3) and (4) report the results from the generalized binomial two-part fractional response model. Thereby, column (4) reports the restricted model results where the parameters are the same in both parts of the model. Note, column (4) contains the same parameter estimates as column (2), but the standard errors are much smaller. The reason is that the latter are MLE-estimates under the assumption of known group sizes and independent unit decisions within groups (i.e., absence of overdispersion). Papke and Wooldridge (1996) calculate robust standard errors to account for potential overdispersion in the data. Finally, columns (5) to (8) report estimation results from the beta-binomial fractional response model that explicitly allows for overdispersion in a flexible MLE-setting.

The parameter estimates reported in Table 3 indicate that in qualitative terms the results are similar across all different models as well as across both parts of the two-part models. Thereby, the first part corresponds to the probability to observe full-participation while the second part estimates the share of participants for firms with

[16]Oberhofer and Pfaffermayr 2012 provide a comprehensive replication exercise of Papke and Wooldridge (1996) and alternatively estimate a standard two-part fractional response model.

Table 3: Estimation results: 401(k) plan participation rates

Variable	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
First Part								
$MRATE$			1.313***	1.372***	1.349***	1.779***	1.203***	1.220***
			(0.083)	(0.002)	(0.085)	(0.085)	(0.097)	(0.068)
$MRATE^2$			−0.221***	−0.290***	−0.228***	−0.312***	−0.246***	−0.288***
			(0.022)	(0.001)	(0.022)	(0.022)	(0.024)	(0.018)
$\log(EMP)$			0.403***	−0.602***	−0.203*	−0.597***	−0.310**	−0.674***
			(0.111)	(0.004)	(0.115)	(0.111)	(0.131)	(0.075)
$\log(EMP)^2$			0.029***	0.029***	0.015**	0.032***	0.002	0.035***
			(0.007)	(0.000)	(0.007)	(0.007)	(0.011)	(0.005)
AGE			−0.004	0.058***	−0.004	0.006	0.011	0.044***
			(0.009)	(0.000)	(0.009)	(0.009)	(0.010)	(0.006)
AGE^2			0.000*	−0.001***	0.000*	0.000	0.000**	−0.000***
			(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
$SOLE$			0.389***	0.053***	0.400***	0.416***	0.416***	0.118***
			(0.047)	(0.002)	(0.048)	(0.048)	(0.050)	(0.035)
$CONSTANT$			1.944***	3.466***	3.347***	3.532***	3.748***	3.307***
			(0.425)	(0.018)	(0.437)	(0.428)	(0.501)	(0.298)
Second Part/ One Part								
$MRATE$	1.665***	1.372***	0.514***	1.372***	0.936***	1.779***	0.942***	1.220***
	(0.089)	(0.169)	(0.003)	(0.002)	(0.079)	(0.085)	(0.082)	(0.068)
$MRATE^2$	−0.332***	−0.290***	−0.178***	−0.290***	−0.226***	−0.312***	−0.238***	−0.288***
	(0.021)	(0.041)	(0.001)	(0.001)	(0.019)	(0.022)	(0.022)	(0.018)
$\log(EMP)$	−1.031***	−0.602**	−0.629***	−0.602***	−0.594***	−0.597***	−0.607***	−0.674***
	(0.112)	(0.256)	(0.004)	(0.004)	(0.088)	(0.111)	(0.088)	(0.075)
$\log(EMP)^2$	0.054***	0.029**	0.032***	0.029***	0.032***	0.032***	0.032***	0.035***
	(0.007)	(0.013)	(0.000)	(0.000)	(0.006)	(0.007)	(0.006)	(0.005)
AGE	0.055***	0.058***	0.061***	0.058***	0.048***	0.006	0.050***	0.044***
	(0.008)	(0.008)	(0.000)	(0.000)	(0.006)	(0.009)	(0.006)	(0.006)
AGE^2	−0.001***	−0.001***	−0.001***	−0.001***	−0.001***	0.000	−0.001***	−0.000***
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
$SOLE$	0.064	0.053	−0.170***	0.053***	−0.120***	0.416***	−0.107***	0.118***
	(0.047)	(0.127)	(0.002)	(0.002)	(0.038)	(0.048)	(0.037)	(0.035)
$CONSTANT$	5.105***	3.466***	3.376***	3.466***	3.040***	3.532***	3.119***	3.307***
	(0.431)	(1.204)	(0.019)	(0.018)	(0.336)	(0.428)	(0.333)	(0.298)
Intra-group correlation								
$MRATE$							−0.202***	−0.191***
							(0.037)	(0.011)
$\log(EMP)$							−0.236***	−0.088***
							(0.028)	(0.009)
AGE							0.014***	0.008***
							(0.003)	(0.002)
$CONSTANT$					2.648***	1.517***	3.315***	1.440***
					(0.026)	(0.023)	(0.124)	(0.070)
Tests								
LM/LR Test ($\chi^2(8)$)			204.46***		2702.33***		489.28***	
Wald Test ($\chi^2(8)$)			82322.39***		5020.01***		619.41***	
ρ					0.352	0.397	0.406	0.410
Observations	4,734	4,734	4,734	4,734	4,734	4,734	4,734	4,734

Notes: Parameter estimates are reported. The results in column (1) are identical to column (4) in Table III in Papke and Wooldridge (1996). In the logit model the dependent variable is one if all employees participate in the 401(k) pension plan and zero otherwise. The QMLE of the two-part model is estimated only for $PRATE < 1$. The LR test refers to columns (5) to (8).

less than full participation. The variable SOLE that informs whether the 401(k) pension plan is the only one offered by a firm forms a notable exception. The probability to observe full-participation is significantly higher in firms that only offer the 401(k) pension plan throughout but the non-restricted two-part models indicate that for all other firms a sole offer of the 401(k) pension plan reduces the share of participants.

Columns (5) to (8) in Table 3 highlight the existence of non-negligible intra-group correlation in the data on 401(k) pension plan participation. The estimated ρ varies from 0.35 to 0.41 and (as indicated in columns 5 and 8) is also affected by a firm's characteristics. An increase in a firm's matching rate decreases c_i and, thus, increases the intra-group correlation ρ as can be seen from equation (10). This finding is well in line with our expectations indicating that a larger matching rate increases the intra-group correlation making 100 percent participation more likely. In a similar vein, a larger firm size also increases ρ. By contrast, in firms that offer older pension plans the intra-group correlation is reduced.

At the bottom of Table 3 we report the results from the alternative LM, LR and Wald tests. All of them indicate that the one-part model should be rejected in favour of the two-part fractional response model. Given the discussion on the parameter estimates for the SOLE dummy variable, this result is not very surprising. This, together with the discussion on the intra-group correlation from above suggests that the generalized beta-binomial model might be most accurate for estimating the 401(k) plan participation rates data at hand.

Finally, for a quantitative comparison of the impacts of all covariates across the different econometric models, one has to focus on marginal effects. Here, we again follow Papke and Wooldridge (1996) and plot the predicted participation rates for different matching rates. Thereby, we set firm size and the pension plans age at their median values of 628 employees and 8 years, respectively, and assume that no other pension plan is offered (i.e., SOLE=0). Finally, we vary the matching rate from 0 (i.e., no matching offered at all) to 1, (i.e., 100 percent matching). Figure 1 displays the predicted participation rates for the QMLE proposed by Papke and Wooldridge (1996) and both alternative two-part models. The left panel compares the predictions from the QMLE (column 1 of Table 3) with the ones from the two-part binomial alternative (column 3). In the right panel the alternative predictions are based on the parameters from the two-part beta-binomial model that parametrizes the intra-group correlation with MRATE, log(EMP) and AGE (column 7). The 95 percent confidence intervals are constructed using the delta method.

Figure 1 reveals some interesting results: First of all, in the absence of any matching (i.e., MRATE=0) the predicted conditional mean of PRATE is lowest (highest) for the QMLE (two-part binomial model). This difference is statistically significant as indicated

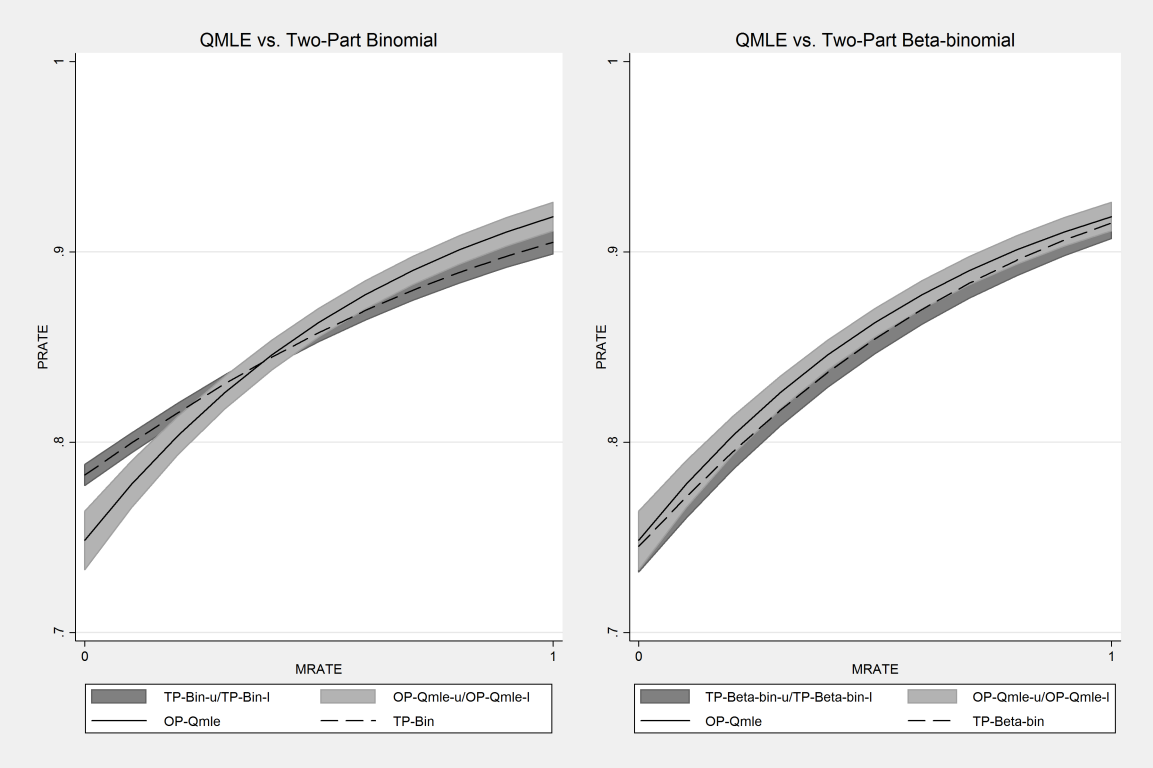

Figure 1: Participation rate versus matching rate: Model predictions

by the respective confidence intervals. Second, as can be seen from the slopes of the corresponding curves, the predicted marginal effect of MRATE, by contrast, is lowest for the generalized binomial model. Most interestingly, however, the QMLE and the (ML based) two-part beta-binomial model deliver relatively similar predictions. This is indicated by the overlap of the respective 95 percent confidence intervals over the whole range of matching rates considers. This finding highlights the usefulness of the QMLE estimator, but also points to relevance of the (two-part) beta-binomial model as a valuable alternative. The QMLE is very easy to implement, but the two-part beta-binomial estimator allows to explicitly specify intra-group correlation, thus, providing additional (parametric) insights in the data generating process.

5 Conclusions

In many applications of fractional responses models the number of units per group is observed and, consequently, the fractional response variable is based on a fraction of integers. In such a situation one can use this additional information and specify two-part models that nest the one-part fractional response model proposed by Papke and

Wooldridge (1996) and account for intra-group correlation and overdispersion induced by group-specific random effects.

These nested two-part models also have the advantage that they allow to apply simple LM, LR and Wald tests to discriminate between one-part and two-part fractional response models. Based on the proposed two-part models, this paper also derives explicit formulas for the Wald and the LM tests. Two Monte Carlo simulation exercises reveal that these tests are properly sized and equipped with sufficient power in the generalized binomial and the beta-binomial fractional response framework. Finally, we apply our alternative estimators to a sample of 401(k) pension plan participation rates and are able to show that these data are characterized by non-negligible intra-group correlation.

6 References

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

Cook D. O., Kieschnick R. and McCullough B. D. (2008). Regression Analysis of Proportions in Finance with Self Selection. *Journal of Empirical Finance* **15**(5), 860–867.

Davidson R. and MacKinnon J. G. (1981). Several Tests for Model Specification in the Presence of Alternative Hypotheses. *Econometrica* **49**(3), 781–793.

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

Gourieroux C., Monfort A. and Trognon A. (1984). Pseudo-maximum Likelihood Methods: Theory. *Econometrica* **52**(3), 681–700.

Hall, D. B. (2000). Zero-inflated Poisson and Binomial Regression with Random Effects: A Case Study. *Biometrics*, **56**(4), 1030–1039.

Heckman J. J. and Willis R. J. (1977). A Beta-logistic Model for the Analysis of Sequential Labor Force Participation by Married Women. *Journal of Political Economy* **85**(1), 27–58.

Johnson N. L., Kemp A. W. and Kotz S. (2005). *Univariate Discrete Distributions*, 3rd edition. John Wiley & Sons: Hoboken, New Jersey.

Kieschnick R. and McCullough B. D. (2003). Regression Analysis of Variates Observed on (0, 1): Percentages, Proportions and Fractions. *Statistical Modelling* **3**(3) 193–213.

Lambert D. (1992). Zero-inflated Poisson Regression, with an Application to Defects in Manufacturing. *Technometrics* **34**(1), 1–14.

Lin T. F. and Schmidt P. (1984). A Test of the Tobit Specification Against an Alternative Suggested by Cragg. *Review of Economics and Statistics* **66**(1), 174–77.

McCulloch C. E. and Searle A. F. M. (2001). *Generalized, Linear and Mixed Models*. John Wiley & Sons: Hoboken, New Jersey.

Mullahy J. (1986). Specification and Testing of Some Modified Count Data Models. *Journal of Econometrics* **33**(3), 341–365.

Mullahy J. (2010). Multivariate Fractional Regression Estimation of Econometric Share Models. NBER Working Papers 16354, National Bureau of Economic Research.

Murteira J. M. R. and Ramalho J. J. S. (2014). Regression Analysis of Multivariate Fractional Data. *Econometric Reviews*, forthcoming.

Oberhofer H. and Pfaffermayr M. (2012). Fractional Response Models - A Replication Exercise of Papke and Wooldridge (1996). *Contemporary Economics* **6**(3), 56–64.

Ospina R. and Ferrari S. L. P. (2012). A General Class of Zero-or-One Inflated Beta Regression Models. *Computational Statistics and Data Analysis* **56**(6), 1609–1623.

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. (1993). Econometric Methods for Fractional Response Variables with an Application to 401(k) Plan Participation Rates. National Bureau of Economic Research Technical Working Paper No. 147.

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.

Prentice R. L. (1986). Binary Regression Using an Extended Beta-Binomial Distribution, With Discussion of Correlation Induced by Covariate Measurement Errors *Journal of the American Statistical Association* **81**(394), 321–327.

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

Ramalho E. A., Ramalho J. J. S. and Murteira J. M. R. (2013). A Generalized Goodness-of-Functional Form Test for Binary and Fractional Regression Models. *Manchester School*, forthcoming.

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 J. J. S. and Silva J. V. (2013). Functional Form Issues in the Regression Analysis of Corporate Capital Structure. *Empirical Economics* **44**(2), 799–831.

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

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

A LM and Wald Tests for the Binomial Two-Part Model

A.1 Derivation of the LM Test

To derive the LM test we reparametrize the model and set $\gamma = \beta + \delta$. Then, the likelihood function is given by

$$\begin{aligned}
&\textstyle\sum_{i=1}^{N} \ln(f(\beta; y_i, x_i)) \\
&= (1 - z_i)\big[\ln(1 - G(x_i\,(\beta + \delta))^{n_i}) + n_i(y_i \ln G(x_i\beta) + (1 - y_i)\ln(1 - G(x_i\beta))) \\
&- \ln(1 - G(x_i\beta)^{n_i}) + const\big] + z_i.\big[n_i \ln(G(x_i\,(\beta + \delta)))\big].
\end{aligned}$$

To derive the score define $z_i = 0$ if $y_i < 0$ and $z_i = 1$ if $y_i = 1$ and denote $G_i' = g_i$. To simplify the notation, we first assume that the model only contains a constant ($x_i = 1$) and introduce the vector of explanatory variables below. The score is given by

$$\begin{aligned}
\frac{\partial \ln(l(\delta, \beta))}{\partial \delta} &= (1 - z_i)\frac{-n_i G(\beta + \delta)^{n_i - 1} g(\beta + \delta)}{1 - G(\beta + \delta)^{n_i}} + z_i \frac{n_i g(\beta + \delta)}{G(\beta + \delta)} \\
\frac{\partial \ln(l(\delta, \beta))}{\partial \beta} &= (1 - z_i)\frac{-n_i G(\beta + \delta)^{n_i - 1} g(\beta + \delta)}{1 - G(\beta + \delta)^{n_i}} \\
&\quad + (1 - z_i) n_i \left(\frac{g(\beta) y_i}{G(\beta)} - \frac{g(\beta)(1 - y_i)}{1 - G(\beta)} \right) \\
&\quad + (1 - z_i)\frac{n_i G(\beta)^{n_i - 1} g(\beta)}{1 - G(\beta)^{n_i}} \\
&\quad + z_i \frac{n_i g(\beta + \delta)}{G(\beta + \delta).}
\end{aligned}$$

Now define $C_{i\delta} = n_i \frac{1 - G(\delta)}{1 - G(\delta)^{n_i}}$, $C_{i\beta+\delta} = n_i \frac{1 - G(\beta+\delta)}{1 - G(\beta+\delta)^{n_i}}$ and observe that $G_i = \frac{e^\beta}{1+e^\beta}$ and $g_i = \frac{e^\beta\left(1+e^\beta\right) - e^\beta e^\beta}{\left(1+e^\beta\right)^2} = \frac{e^\beta}{1+e^\beta}\frac{1}{1+e^\beta} = G_i(1 - G_i)$. Inserting and simplifying yields

$$\begin{aligned}
s_{i\delta} &= \frac{\partial \ln(l(\delta, \beta))}{\partial \delta} = C_{i\beta+\delta}\,(z_i - G(\beta + \delta)^{n_i}) \\
s_{i\beta} &= \frac{\partial \ln(l(\delta, \beta))}{\partial \beta} = n_i\,(y_i - G(\beta)) + (1 - z_i)\,[G(\beta + \delta)^{n_i} C_{i\beta+\delta} - G(\beta)^{n_i} C_{i\beta}]\,,
\end{aligned}$$

where we use $y_i = 1$ if $z_i = 1$. The Hessian thus can be derived as

$$\begin{aligned}
\frac{\partial^2 \ln(l(\delta,\beta))}{\partial \delta^2} &= \frac{\partial C_{i\beta+\delta}}{\partial \delta}\left(z_i - G(\beta+\delta)^{n_i}\right) - C_{i\beta+\delta}^2\left(1 - G(\beta+\delta)^{n_i}\right)G(\beta+\delta)^{n_i} \\
\frac{\partial \ln(l(\delta,\beta))}{\partial \delta \partial \beta} &= \frac{\partial C_{i\beta+\delta}}{\partial \beta}\left(z_i - G(\beta+\delta)^{n_i}\right) - C_{i\beta+\delta}^2\left(1 - G(\beta+\delta)^{n_i}\right)G(\beta+\delta)^{n_i} \\
\frac{\partial^2 \ln(l(\delta,\beta))}{\partial \beta^2} &= -n_i\left(\left(1 - G(\beta)\right)G(\beta)\right) + \frac{\partial(1-z_i)\left[G(\beta+\delta)^{n_i}C_{i\beta+\delta} - G(\beta)^{n_i}C_{i\beta}\right]}{\partial \beta},
\end{aligned}$$

so that under $H_0: \ \delta = 0$ on obtains

$$\begin{aligned}
E[H_i]|_{\delta=0} &= -\begin{bmatrix} C_{i\beta}^2\left(1 - G(\beta)^{n_i}\right)G(\beta)^{n_i} & C_{i\beta}^2\left(1 - G(\beta)^{n_i}\right)G(\beta)^{n_i} \\ C_{i\beta}^2\left(1 - G(\beta)^{n_i}\right)G(\beta)^{n_i} & n_i\left(\left(1 - G(\beta)\right)G(\beta)\right) \end{bmatrix} \\
&:= -\begin{bmatrix} A_i & A_i \\ A_i & B_i \end{bmatrix}.
\end{aligned}$$

Now, introducing a $(1 \times k)$ vector of explanatory variables x_i for unit i under $H_0: \ \delta = 0$ yields

$$\begin{aligned}
s_{i\delta} &= C_{i\beta}\left(z_i - G(x_i\beta)^{n_i}\right)x_i' \\
s_{i\beta} &= n_i\left(y_i - G(x_i\beta)\right)x_i' = 0.
\end{aligned}$$

We define

$$\begin{aligned}
s_\delta &= \sum_{i=1}^{N} s_{i\delta} \\
s_\beta &= \sum_{i=1}^{N} s_{i\beta} \\
A(\beta) &= \sum_{i=1}^{N} C_{i\beta}^2 G(x_i\beta)^{n_i}\left(1 - G(x_i\beta)^{n_i}\right)x_i'x_i \\
B(\beta) &= \sum_{i=1}^{N}\left[n_i\left(G(x_i\beta)(1 - G(x_i\beta))\right)x_i'x_i\right],
\end{aligned}$$

$$I(\beta) = \begin{bmatrix} A & A \\ A & B \end{bmatrix} \text{ and } I(\beta)^{-1} = \begin{bmatrix} A^{-1}\left(I + A\left(B - A\right)^{-1}\right)AA^{-1} & -\left(B - A\right)^{-1} \\ -\left(B - A\right)^{-1} & \left(B - A\right)^{-1} \end{bmatrix}.$$

Note $A^{-1}\left(A + A\left(B - A\right)^{-1}A\right)A^{-1} = A^{-1}\left(I + A\left(B - A\right)^{-1}\right) = A^{-1} + \left(B - A\right)^{-1}$. The estimated LM statistic uses the estimated parameters from the one-part model with index

OP and is given by

$$\begin{aligned}\widehat{LM} &= \left[\begin{array}{cc} \widehat{s}'_{\delta_{OP}} & 0 \end{array}\right] \left[\begin{array}{cc} A(\widehat{\beta}_{OP})^{-1} + \Big(B(\widehat{\beta}_{OP}) - A(\widehat{\beta}_{OP})\Big)^{-1} & -\Big(B(\widehat{\beta}_{OP}) - A(\widehat{\beta}_{OP})\Big)^{-1} \\ -\Big(B(\widehat{\beta}_{OP}) - A(\widehat{\beta}_{OP})\Big)^{-1} & \Big(B(\widehat{\beta}_{OP}) - A(\widehat{\beta}_{OP})\Big)^{-1} \end{array}\right] \left[\begin{array}{c} \widehat{s}_{\delta_{OP}} \\ 0 \end{array}\right] \\ &= \widehat{s}'_{\delta_{OP}} \left(A(\widehat{\beta}_{OP})^{-1} + \Big(B(\widehat{\beta}_{OP})^{-1} - A(\widehat{\beta}_{OP})\Big)^{-1}\right) \widehat{s}_{\delta_{OP}},\end{aligned}$$

which is asymptotically distributed as $\chi^2(k)$.

A.2 Derivation of the Wald Test

For the Wald test consider the model in original parametrization so that the following restriction is tested

$$\left[\begin{array}{cc} I_k & -I_k \end{array}\right] \left[\begin{array}{c} \gamma \\ \beta \end{array}\right] = \gamma - \beta = 0.$$

The expected Hessian of the unrestricted model can be derived similarly as above, and for observation i reads as

$$E[H_i] = -\left[\begin{array}{cc} C_{i\gamma}^2 \left(1 - G(x_i\gamma)^{n_i}\right) G(x_i\gamma)^{n_i} & 0 \\ 0 & n_i \left(\left(1 - G(x_i\beta)\right) G(x_i\beta)\right) - C_{i\beta}^2 \left(1 - G(x_i\beta)^{n_i}\right) G(x_i\beta)^{n_i} \end{array}\right].$$

The Wald test uses the estimated parameters of the two-part model that are indexed by TP. Denoting the estimated variance covariance matrix of $\hat{\gamma_{TP}}$ by $\hat{V}_{\gamma_{TP}}$ and that of $\hat{\beta_{TP}}$ by $\hat{V}_{\beta_{TP}}$ it can easily be shown that

$$\hat{V}_{\gamma_{TP}} + \hat{V}_{\beta_{TP}} = A(\widehat{\gamma}_{TP})^{-1} + \Big(B(\widehat{\beta}_{TP}) - A(\widehat{\beta}_{TP})\Big)^{-1},$$

and

$$\widehat{W} = (\hat{\gamma}_{TP} - \hat{\beta}_{TP})' \left(A(\widehat{\gamma}_{TP})^{-1} + \Big(B(\widehat{\beta}_{TP}) - A(\widehat{\beta}_{TP})\Big)^{-1}\right)^{-1} (\hat{\gamma}_{TP} - \hat{\beta}_{TP}).$$

Asymptotically, this Wald test is also $\chi^2(k)$ distributed.

B The Beta-Binomial Two-Part Fractional Response Model

In the two-part model the probability for $k_i = y_i n_i \leq n_i$ successes in n_i trials can be rewritten as

$$g(y_i|x_i, n_i) = \begin{cases} P_1(n_i, x_i) & \text{if} \quad y_i = 1 \\ (1 - P_1(n_i, x_i)) \frac{P_2(k_i, x_i)}{1 - P_2(n_i, x_i)} & \text{if} \quad y_i < 1. \end{cases}$$

Defining $\Delta \ln \Gamma(y, a) = \ln \Gamma(y + a) - \ln \Gamma(a)$, one can write

$$\begin{aligned} \ln P_2(k_i) &= \ln \binom{n_i}{k_i} + \Delta \ln \Gamma(k_i, a_i) + \Delta \ln \Gamma(n_i - k_i, b_i) - \Delta \ln \Gamma(n_i, a_i + b_i) \\ & \ln \binom{n_i}{k_i} + \Delta \ln \Gamma(k_i, c\theta_{i2}) + \Delta \ln \Gamma(n_i - k_i, c(1 - \theta_{i2})) - \Delta \ln \Gamma(n_i, c). \end{aligned}$$

$P_1(n, x_i)$ is defined analogously:

$$\begin{aligned} P_1(n_i) &= \frac{\Gamma(n_i + a_{i1})}{\Gamma(n_i + a_{i1} + b_{i1})} \frac{\Gamma(a_{i1} + b_{i1})}{\Gamma(a_{i1})} = \frac{\prod_{j=0}^{n_i - 1} (a_{i1} + j)}{\prod_{j=0}^{n_i - 1} (a_{i1} + b_{i1} + j)} = \frac{\prod_{j=0}^{n_i - 1} (c\theta_{i1} + j)}{\prod_{j=0}^{n_i - 1} (c + j)} \\ \ln P_1(n_i) &= \sum_{j=0}^{n_i - 1} \ln(c\theta_{i1} + j) - \sum_{j=0}^{n_i - 1} \ln(c + j). \end{aligned}$$

Similarly,

$$\ln P_2(n_i) = \sum_{j=0}^{n_i - 1} \ln(c\theta_{i2} + j) - \sum_{j=0}^{n_i - 1} \ln(c + j).$$

In order to drive the likelihood of the model, we define $z_i = 1[k_i = n_i] = 1[y_i = 1]$. Then the contribution of each group i to the likelihood is

$$\begin{aligned} &\ln L_i = (1 - z_i)\big[\ln(1 - P_1(\theta_{i1}, c, n_i)) + \ln P_2(\theta_{i2}, c, k_i) \\ &- \ln(1 - P_2(\theta_{i2}, c, n_i)) + const\big] + z_i\big[\ln(P_1(\theta_{i1}, c, n_i))\big] \\ &:= L_{i1} + L_{i2} \\ &\text{with} \\ &L_{i1} = (1 - z_i) \ln(1 - P_1(\theta_{i1}, c, n_i)) + z_i \ln(P_1(\theta_{i1}, c, n_i)) \\ &L_{i2} = -(1 - z_i) \ln(1 - P_2(\theta_{i2}, c, n_i)) + (1 - z_i) \ln P_2(\theta_{i2}, c, k_i)). \end{aligned}$$

Note under H_0 : $P_1(\theta_{i1}, c, n_i) = P_2(\theta_{i2}, c, n_i)$ the contribution of each group i to the likelihood reduces to

$$\ln L_i^1 = (1 - z_i) \ln(1 - P_2(\theta_{i2}, c)) + z_i \ln(P_2(\theta_{i2}, c)).$$

We assume $\theta_{i2} = G(x_i\beta)$, while $\theta_{i1} = G(x_i\left(\beta + \delta\right)$. In order to derive the score one needs the first derivatives of $\ln P_1(\theta_{i1}, c, n_i)$, $\ln P_2(\theta_{i2}, c, n_i)$ and $\ln P_2(\theta_{i2}, c, k_i)$. Using

$$\begin{aligned}
\frac{\partial}{\partial \ln p} \ln(1 - e^{\ln p}) &= -\frac{e^{\ln p}}{1-e^{\ln p}} = -\frac{p}{1-p} \\
\frac{\partial}{\partial \ln p} \frac{z-e^{\ln p}}{1-e^{\ln p}} &= \frac{-e^{\ln p}\left(1-e^{\ln p}\right)+(z-e^{\ln p})e^{\ln p}}{\left(1-e^{\ln p}\right)^2} = e^{\ln p}\frac{-\left(1-e^{\ln p}\right)+(z-e^{\ln p})}{\left(1-e^{\ln p}\right)^2} \\
&= e^{\ln p}\frac{-1+e^{\ln p}+z-e^{\ln p}}{\left(1-e^{\ln p}\right)^2} = \frac{z-1}{\left(1-e^{\ln p}\right)^2}e^{\ln p} = \frac{(z-1)p}{(1-p)^2},
\end{aligned}$$

one can write

$$\begin{aligned}
\frac{\partial \ln L_i}{\partial \delta} &= \left[(1 - z_i)\frac{-P_1(\theta_{i1}(\beta,\delta),c,n_i)}{1-P_1(\theta_{i1}(\beta,\delta),c,n_i)} + z_i\right] \frac{\partial \ln P_1(\theta_{i1}(\beta,\delta),c,n_i)}{\partial \theta_{i1}} \frac{\partial \theta_{i1}(\beta,\delta)}{\partial \delta} \\
&= \left[\frac{-(1-z_i)P_1(\theta_{i1}(\delta),c,n_i)+z_i(1-P_1(\theta_{i1}(\delta),c,n_i))}{1-P_1(\theta_{i1}(\delta),c,n_i)}\right] \frac{\partial \ln P_1(\theta_{i1}(\beta,\delta),c,n_i)}{\partial \theta_{i1}} \frac{\partial \theta_{i1}(\beta,\delta)}{\partial \delta} \\
&= \left[\frac{z_i-P_1(\theta_{i1}(\beta,\delta),c,n_i))}{1-P_1(\theta_{i1}(\beta,\delta),c,n_i)}\right] \frac{\partial \ln P_1(\theta_{i1}(\beta,\delta),c,n_i)}{\partial \theta_{i1}} \frac{\partial \theta_{i1}(\beta,\delta)}{\partial \delta} \\
\frac{\partial \ln L_i}{\partial \beta} &= \left[\frac{z_i-P_1(\theta_{i1}(\beta,\delta),c,n_i))}{1-P_1(\theta_{i1}(\beta,\delta),c,n_i)}\right] \frac{\partial \ln P_1(\theta_{i1}(\beta,\delta),c,n_i)}{\partial \theta_{i1}} \frac{\partial \theta_{i1}(\beta,\delta)}{\partial \beta} \\
&\quad - \left[\frac{z_i-P_2(\theta_{i2}(\beta),c,n_i))}{1-P_2(\theta_{i2}(\beta),c,n_i)}\right] \frac{\partial \ln P_2(\theta_{i2}(\beta),c,n_i)}{\partial \theta_{i2}} \frac{\partial \theta_{i2}(\beta)}{\partial \beta} \\
&\quad + \left((1 - z_i)\frac{\partial \ln P_2(\theta_{i2}(\beta),c,k_i)}{\partial \theta_{i2}} + z_i\frac{\partial \ln P_2(\theta_{i2}(\beta),c,n_i)}{\partial \theta_{i2}}\right) \frac{\partial \theta_{i2}(\beta)}{\partial \beta} \\
\frac{\partial \ln L_i}{\partial c} &= \left[\frac{z_i-P_1(\theta_{i1}(\beta,\delta),c,n_i))}{1-P_1(\theta_{i1}(\beta,\delta),c,n_i)}\right] \frac{\partial \ln P_1(\theta_{i1}(\beta,\delta),c,n_i)}{\partial c} \\
&\quad + \left[\frac{z_i-P_2(\theta_{i2}(\beta),c,n_i))}{1-P_2(\theta_{i2}(\beta),c,n_i)}\right] \frac{\partial \ln P_2(\theta_{i2}(\beta),c,n_i)}{\partial c} \\
&\quad +(1 - z_i)\frac{\partial \ln P_2(\theta_{i2}(\beta),c,k_i)}{\partial c} + z_i\frac{\partial \ln P_2(\theta_{i2}(\beta),c,n_i)}{\partial c}
\end{aligned}$$

and

$$\begin{aligned}
P_1(n_i) &= \Delta \ln \Gamma\left(n_i, cq_i\right) - \Delta \ln \Gamma\left(n_i, c\right) \\
P_2(n_i) &= \Delta \ln \Gamma\left(n_i, cp_i\right) - \Delta \ln \Gamma\left(n_i, c\right),
\end{aligned}$$

where $\Delta \ln \Gamma\left(y, a\right) = \ln \Gamma\left(y + a\right) - \ln \Gamma\left(a\right)$. $\frac{\partial \Delta \ln \Gamma(y,a)}{\partial a} = \psi\left(y + a\right) - \psi\left(a\right) = \Delta\psi(y, a)$, $\psi(x) = \frac{\partial \ln \Gamma(x)}{\partial x}$ denotes di-gamma function.

$$
\begin{aligned}
\frac{\partial \ln P_1(\theta_{i1}(\beta,\delta),c,n_i)}{\partial \theta_{i1}} &= \frac{\partial(\Delta \ln \Gamma(n_i,cq_i(\beta,\delta))-\Delta \ln \Gamma(n_i,c))}{\partial \theta_{i1}} = \Delta\psi(n_i, c\theta_{i1}(\beta,\delta))c \\
\frac{\partial \ln P_2(\theta_{i2}(\beta),c,n_i)}{\partial \theta_{i2}} &= \Delta\psi(n_i, c\theta_{i2}(\beta))c \\
\frac{\partial \ln P_2(\theta_{i2}(\beta),c,k_i)}{\partial \theta_{i2}} &= \Delta\psi(k_i, c\theta_{i2}(\beta))c \\
\frac{\partial \ln P_1(\theta_{i1}(\beta,\delta),c,n_i)}{\partial c} &= \Delta\psi(n_i, c\theta_{i1}(\beta,\delta))\theta_{i1}(\beta,\delta) - \Delta\psi(n_i, c) \\
\frac{\partial \ln P_2(\theta_{i2}(\beta),c,n_i)}{\partial c} &= \Delta\psi(n_i, c\theta_{i2}(\beta))\theta_{i2}(\beta) - \Delta\psi(n_i, c) \\
\frac{\partial \ln P_2(\theta_{i2}(\beta),c,k_i)}{\partial c} &= \Delta\psi(k_i, c\theta_{i2}(\beta))\theta_{i2}(\beta) - \Delta\psi(k_i, c).
\end{aligned}
$$

Defining

$$
\begin{aligned}
u(\theta_{i1}(\beta,\delta),c,n_i) &= \frac{z_i - P_1(\theta_{i1}(\beta,\delta),c,n_i)}{1-P_1(\theta_{i1}(\beta,\delta),c,n_i)} \\
\frac{\partial}{\partial \ln P_1} u(\theta_{i1}(\beta,\delta),c,n_i) &= v(\theta_{i1}(\beta,\delta),c,n_i) = \frac{(z_i-1)P_1(\theta_{i1}(\beta,\delta),c,n_i)}{(1-P_1(\theta_{i1}(\beta,\delta),c,n_i))^2},
\end{aligned}
$$

the derivation of the score can be based on

$$
\begin{aligned}
\frac{\partial \ln L_i^1}{\partial \delta_k} &= u(\theta_{i1}(\beta,\delta),c,n_i)\Delta\psi(n_i,cq_i)c\frac{\partial \theta_{i1}}{\partial \delta_k} = 0 \\
\frac{\partial \ln L_i^2}{\partial \beta_k} &= u(\theta_{i1}(\beta,\delta),c,n_i)\Delta\psi(n_i,cq_i)c\frac{\partial \theta_{i1}}{\partial \beta_k} = 0 \\
\frac{\partial \ln L_i^2}{\partial \beta_k} &= u(\theta_{i1}(\beta,\delta),c_i,n_i)\Delta\psi(n_i,cq_i)c\frac{\partial \theta_{i1}}{\partial \beta_k} \\
&\quad -u(\theta_{i1}(\beta,\delta),c_i,n_i)\Delta\psi(n_i,cp_i)c\frac{\partial \theta_{i2}}{\partial \beta_k} \\
&\quad +(1-z_i)\Delta\psi(k_i,cp_i)c\frac{\partial \theta_{i2}}{\partial \beta_k} + z_i\Delta\psi(n_i,cp_i)c\frac{\partial \theta_{i2}}{\partial \beta_k} \\
\frac{\partial \ln L_i}{\partial c} &= u(\theta_{i1}(\beta,\delta),c_i,n_i)\left(\Delta\psi(n_i,cq_i)\theta_{i1} - \Delta\psi(n_i,c)\right) \\
&\quad -u(\theta_{i2},c_i,n_i)\left(\Delta\psi(n_i,cp_i)\theta_{i2} - \Delta\psi(n_i,c)\right) \\
&\quad +\left(\Delta\psi(k_i,cp_i)\theta_{i2} - \Delta\psi(k_i,c)\right).
\end{aligned}
$$

To implement the likelihood estimator, we provide the score but rely on the numerical derivation of the Hessian. Then, Wald and LR tests are readily available from standard ML estimation routines.