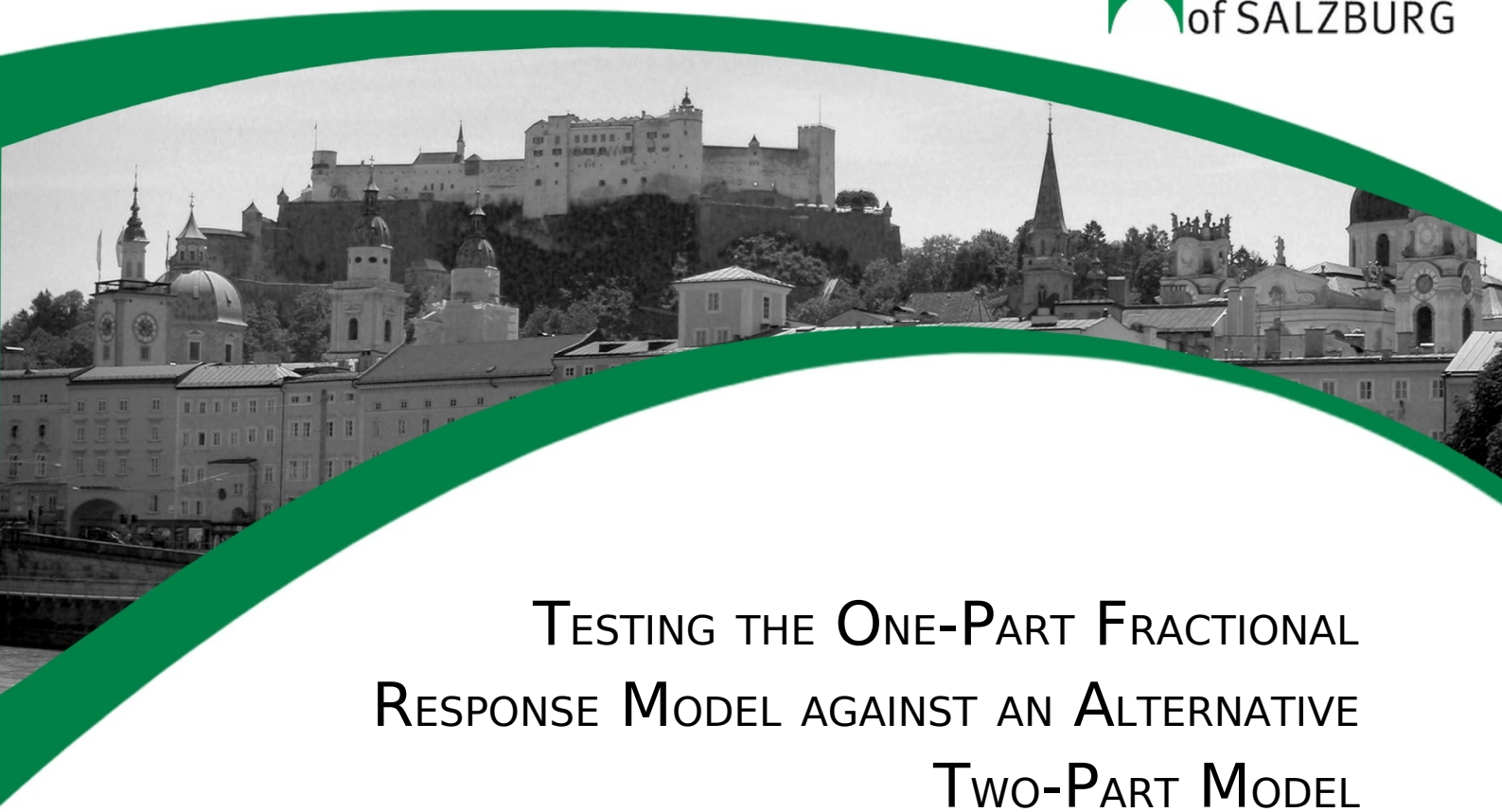
UNIVERSITY of SALZBURG



Testing the One-Part Fractional Response Model against an Alternative Two-Part Model

Harald Oberhofer and Michael Pfaffermayr

Working paper No. 2011-01

WORKING PAPERS IN ECONOMICS AND FINANCE

Testing the One-Part Fractional Response Model against an Alternative Two-Part Model

Harald Oberhofer[a] and Michael Pfaffermayr[b,c]

December 18, 2010

Abstract

This note proposes a generalized two-part model for fractional response variables that nests the one-part model proposed by Papke and Wooldridge (1996). Consequently, a Wald test allows to discriminate between these two competing models. A small scale Monte Carlo simulation demonstrates that the proposed Wald test is properly sized and equipped with higher power than an alternative non-nested P-test.

JEL Codes: C12, C15, C21, C25
Keywords: Fractional response models, two-part model, Wald test, P-test.

[a]Department of Economics and Social Sciences, University of Salzburg, Kapitelgasse 5-7, 5010 Salzburg, Austria. E-mail address: Harald.Oberhofer@sbg.ac.at.
[b]Department of Economics, University of Innsbruck, Universitaetsstrasse 15, 6020 Innsbruck, Austria. Tel.:(+43)512 507 7359. E-mail: michael.pfaffermayr@uibk.ac.at.
[c]Austrian Institute of Economic Research, PO-Box 91, A-1103-Vienna, Austria.

1 Introduction

Many empirical studies deal with share data confined to the $(0, 1)$ interval and, in addition, with a significant amount of observations at the boundaries, 0 or 1. In their seminal paper Leslie Papke and Jeffrey Wooldridge (1996) propose a one-part fractional response model that extends the generalized linear model (GLM) literature from statistics for such data.[1] They introduce a quasi-maximum likelihood estimator (QLME) to obtain a robust method to estimate one-part fractional response models without *ad hoc* transformation of boundary values. In case of a significant share of boundary values, one may alternatively consider a two-part model that accounts for an excessive number of boundary values of ones or zeros, assuming a different econometric model for the boundary values.[2]

The literature so far uses a P-test for non-nested hypotheses to discriminate between one-part and two-part models as described in Davidson and MacKinnon (1981) and Ramalho et al. (2010). This note shows that the one-part fractional response model can be nested in a two-part hurdle model. Hence, one can simply use a Wald test as an alternative to the P-test. A small scale Monte Carlo exercise reveals that the proposed Wald test is properly sized and equipped with higher power than the P-test.

2 The one-part and two-part fractional response models

The fractional response model is based on the Bernoulli distribution. Assume there are $i = 1, \ldots, N$ groups (firms) in which $j = 1, \ldots, n_i$ units (workers) are confronted with a zero or one decision (e.g., participation in a voluntary pension plan). Following Papke and Wooldridge (1993, 1996) we assume that n_i is fixed and exogenously given so that it is appropriate to condition on n_i. The probability that a particular unit j in group i decides for 1 is denoted by θ_i and it is assumed to be group-specific but not unit-specific. The number of units within a group deciding for 1 is denoted by Y_i and the corresponding share is $y_i = \frac{Y_i}{n_i}$ with $0 \leq y_i \leq 1$.

In many empirical applications, the individual decisions of the units remain unobserved and only the share of ones and zeros within a group is known, i.e., a sample of a fractional response variable is available. The conditional distribution is then defined in terms of the number of successes $n_i y_i$ in n_i trials. Additionally, in comparison to the

[1] In a recent paper, Papke and Wooldridge (2008) discuss fractional response models for panel data. Ramalho et. al (2010) provide a comprehensive up-to-date overview on the econometrics of fractional response models.

[2] See Lambert (1992), Wooldridge (2002, Problem 19.8), Ramalho and Vidigal da Silva (2009) and Ramalho, Ramalho and Muteira (2010).

unweighted 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 (see Papke and Wooldridge, 1993, pp. 10-11). Under independent unit decisions, Y_i is Bernoulli distributed with conditional density

$$f(y_i|\mathbf{x}_i, n_i) = \binom{n_i}{n_i y_i} \theta_i^{n_i y_i} (1-\theta_i)^{n_i - n_i y_i}, \tag{1}$$

where (the $1 \times k$ vector) $\mathbf{x}_i$ refers to a set of i-specific explanatory variables with the corresponding parameter vector β. In particular, the probability of the share y_i amounting exactly to zero or one is given by $(1-\theta_i^{n_i})$ and $\theta_i^{n_i}$, respectively. The conditional expectation of the fractional response variable y_i is group-specific and specified as

$$E(y_i|\mathbf{x}_i, n_i) = \theta_i = G(\mathbf{x}_i\beta), \qquad i = 1, \ldots, N. \tag{2}$$

Typically, $G(.)$ is a cumulative distribution function (cdf) like the logistic function $G(z) = \exp(z)/(1+\exp(z))$ which maps z to the $(0,1)$ interval.[3] Finally, the Bernoulli log likelihood is maximized to obtain estimates of the coefficient vector β:

$$\sum_{i=1}^{N} \ln(f_i(\beta)) = \sum_{i=1}^{N} n_i(y_i \ln(G(\mathbf{x}_i\beta)) + (1-y_i)\ln(1-G(\mathbf{x}_i\beta))) + cons. \tag{3}$$

Following Wooldridge (2002, Problem 19.8), Cameron and Trivedi (2005, p. 680), Ramalho and Vidigal da Silva (2009) and Ramalho et al. (2010), we alternatively consider a two-part model to account for an excessive number of boundary values. We concentrate on the case of boundary values of ones, but similar arguments apply to the case of an excessive number of zeros. In the two-part model the boundary values are described by a different data generating process which we specify as $P(y_i = 1) = q_i^{n_i}$ with $0 < q_i < 1$ and $q_i = G(\mathbf{x}_i\gamma)$. For notational simplicity, the explanatory variables in the first and second part of the model are assumed to be the same, but in more general models they could be different. In comparison to the standard two-part model, our specification additionally takes into account the exponent n_i. Formally, the two-part model can be defined as (see Cameron and Trivedi, 2005, pp. 545, 680):

$$g(y_i|\mathbf{x}_i, n_i) = \begin{cases} q_i^{n_i} & \text{if} \quad y_i = 1 \\ (1-q_i^{n_i}) f(y_i|y_i < 1, \mathbf{x}_i, n_i) & \text{if} \quad y_i < 1. \end{cases} \tag{4}$$

[3] See Ramalho et al. (2010) for a comprehensive discussion on different functional forms in one-part and two-part models.

The second part of the model is based on the conditional distribution $f(y_i|y_i < 1, |\mathbf{x}_i, n_i)$ implying that the probability distribution $f(y_i|\mathbf{x}_i, n_i)$ is divided by $1 - \theta_i^{n_i}$ to ensure that the conditional probabilities sum up to 1. The conditional mean of the two-part model, thus, is given by[4]

$$\begin{aligned} E(y_i|\mathbf{x}_i, n_i) &= P(y_i < 1|\mathbf{x}_i, n_i)E(y_i|y_i < 1, \mathbf{x}_i, n_i) + P(y_i = 1|\mathbf{x}_i, n_i) \\ &= \frac{1 - G(\mathbf{x}_i\gamma)^{n_i}}{1 - G(\mathbf{x}_i\beta)^{n_i}}(G(\mathbf{x}_i\beta) - G(\mathbf{x}_i\beta)^{n_i}) + G(\mathbf{x}_i\gamma)^{n_i}. \end{aligned} \tag{5}$$

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 Vidigal da Silva (2009, p. 8) specify the conditional mean $E(y_i|y_i > 0, \mathbf{x}_i, n_i)$ as $G(\mathbf{x}_i\beta)$. The likelihood of the two part-model consists of individual contributions reading as

$$\ln(g(\gamma, \beta)) = \begin{cases} \ln(1 - G(\mathbf{x}_i\gamma)^{n_i}) + n_i(y_i \ln(G(\mathbf{x}_i\beta)) + (1 - y_i)\ln(1 - G(\mathbf{x}_i\beta))) \\ -\ln(1 - G(\mathbf{x}_i\beta)^{n_i}) + constant \qquad \text{if} \quad y_i < 1 \\ n_i \ln(G(\mathbf{x}_i\gamma)) \qquad\qquad\qquad\qquad\qquad\quad \text{if} \quad y_i = 1. \end{cases} \tag{6}$$

Under this specification maximum likelihood estimation is straight forward, since it separates into the estimation of the model explaining $P(y_i = 1|\mathbf{x}_i, n_i)$ using all observations and the estimation of the fractional response model based only on the observations with $y_i < 1$. In the following, we assume that the distributions upon which the one-part and the two-part models are based are correctly specified and concentrate on maximum likelihood estimation.[5]

The main advantage of this two-part-model is that it nests the one-part fractional response model since under $q_i^{n_i} = \theta_i^{n_i}$ the two-part-model reverts to the one-part fractional response model. In case of $\mathbf{x}_i$ being the same for the one-part and two-part model and their parameters being equal, i.e., γ=β, the two models coincide and have the same likelihood functions. This can easily be tested by a Wald test of the hypothesis γ=β. If the first part of the two-part model includes additional explanatory variables denoted by

[4]In case of zero boundary values the conditional mean of this model is given by

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

[5]Actually, one can apply quasi-maximum likelihood estimation (QMLE) to the one-part model assuming that the conditional expectation of y_i is correctly specified (see Gourieroux et al., 1984). The consistency of the estimated parameters of the two-part model require the assumption of the correct specification of $E(y_i|y_i < 1, \mathbf{x}_i, n_i)$ and $P(y_i = 1|\mathbf{x}_i, n_i)$. Since we are interested in the estimation of the two-part model and in testing hypotheses about its estimated parameters, we resort to maximum likelihood estimation.

$\mathbf{z}_i$ with parameter vector ϕ, the underlying hypothesis to test is γ=β, $\phi = 0$. In applying this Wald test it is crucial to use the weighted form of the likelihood given in (3) or to assume $n_i = n$, since it is not defined for $n_i = 1$. Denoting the estimated variance covariance matrix of $\hat{\gamma}$ by $\hat{V}_\gamma$ and that of $\hat{\beta}$ by $\hat{V}_\beta$ it can easily be shown that the Wald test statistic is given by

$$W = (\hat{\gamma} - \hat{\beta})'(\hat{V}_\gamma + \hat{V}_\beta)^{-1}(\hat{\gamma} - \hat{\beta}). \tag{7}$$

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

The literature commonly applies a (non-nested) P-test to discriminate between the one-part and two-part model. Following Davidson and MacKinnon (1981) and Ramalho et al. (2010) the P-test for the null hypothesis that the one-part model is the true one and two-part model is the alternative is based on the artificial regression

$$y_i - \hat{G}_i = \hat{g}_i \mathbf{x}_i \delta_1 + \delta_2((1 - \hat{F}_i)\hat{M}_i + \hat{F}_i - \hat{G}_i) + \varepsilon_i, \tag{8}$$

where, $\hat{g}_i = \frac{\partial G(\mathbf{x}_i\hat{\beta})}{\partial \mathbf{x}_i\hat{\beta}}$. Note, the alternative uses the specification $\hat{F}_i = G(\mathbf{x}_i\hat{\gamma})^{n_i}$ and $\hat{M}_i = G(\mathbf{x}_i\hat{\beta})$ similar to Ramalho et al. (2010). Hence, the two models are non-nested. Under H_0 the two-part model should not provide additional information to estimate the conditional mean, which is equivalent to testing $\delta_2 = 0$ vs. $\delta_2 \neq 0$.

3 Monte Carlo simulations

To investigate the performance of the two tests in finite samples we set up a small Monte Carlo simulation exercise. We generate Bernoulli random variables using logistic function $G(x_i + 0.5)$, 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. In a first set of experiments we assume $n_i = 100$, while the second one takes $n_i \sim iid \quad N(100, 5)$. The probabilities for the boundary values of 1 are based on $q_i = G(\alpha(x_i + 0.5))^{n_i}$, where α varies between 0.95 and 1.05 so that at $\alpha = 1$ the one-part model is the true one. To obtain the dummy variable for boundary values that takes the value 1 if boundary values of 1 are observed, we generate a uniformly distributed random variable and set the value of the dummy variable to 1 if this random variable is lower than q_i and 0 otherwise.

We run each Monte Carlo experiment 10,000 times for sample sizes of 500 and 1,000, respectively, and calculate the size of the tests as share of rejections at $\alpha = 1$ and nominal size 0.05. Hence, with a nominal size of 5 percent a 95-percent confidence interval of the

calculated size is given by [4.48, 5.43]. The power of the tests is defined as share of rejections at $\alpha \neq 1$.

The results of this simulation exercise are summarized in Table 1. Under H_0 and at $n = 100$, 14% of the observations fall on the boundary of 1. Increasing $n = 200$ yields a share of 6% of ones under H_0. In case of varying n_i these shares are of comparable size. Increasing α leads to an increase in the share of boundary values.

The Monte Carlo simulation results indicate that the simulated size of the Wald test is within a 95-percent confidence interval in 7 out 8 experiments. Contrary, the P-test tends to be marginally oversized especially at lower group sizes. The Wald test and the P-test have power in both directions $a < 1$ and $a > 1$, and as expected it increases with sample size. Moreover, the power of the two tests does not differ between the experiments that assume fixed and variable group size. However, in all experiments the power of the P-test is considerably smaller than that of the Wald test.

4 Conclusions

In many applications of the fractional response model the number of units upon which the outcome variable is based can be observed. In such a situation one can specify a two-part model that nests the fractional response model and as an alternative to the available P-tests a Wald test can be applied to discriminate between the two models. A small Monte Carlo simulation exercise shows that the Wald test exhibits higher power than the P-test.

Table 1: Monte Carlo simulation: 10,000 replications

	a	Share of ones	Wald test	P-test	Share of ones	Wald test	P-test
			$N = 500$			$N = 1,000$	
$\bar{n} = 100$							
Fixed	0.95	0.11	0.491	0.121	0.11	0.796	0.210
	0.96	0.12	0.337	0.094	0.12	0.599	0.142
	0.97	0.12	0.210	0.070	0.12	0.380	0.093
	0.98	0.13	0.122	0.050	0.13	0.209	0.065
	0.99	0.13	0.068	0.048	0.13	0.096	0.047
	1.00	0.14	**0.049**	**0.058**	0.14	**0.051**	**0.056**
	1.01	0.15	0.056	0.063	0.14	0.062	0.082
	1.02	0.15	0.090	0.089	0.15	0.133	0.119
	1.03	0.16	0.156	0.114	0.16	0.277	0.167
	1.04	0.16	0.254	0.154	0.16	0.467	0.240
	1.05	0.17	0.394	0.198	0.17	0.685	0.323
Variable	0.95	0.11	0.484	0.128	0.12	0.820	0.238
	0.96	0.12	0.330	0.091	0.13	0.632	0.157
	0.97	0.12	0.212	0.067	0.13	0.391	0.100
	0.98	0.13	0.125	0.052	0.14	0.214	0.063
	0.99	0.13	0.073	0.053	0.15	0.091	0.050
	1.00	0.14	**0.051**	**0.054**	0.15	**0.050**	**0.054**
	1.01	0.15	0.057	0.066	0.16	0.069	0.078
	1.02	0.15	0.086	0.087	0.16	0.140	0.122
	1.03	0.16	0.148	0.112	0.17	0.294	0.182
	1.04	0.16	0.251	0.156	0.18	0.498	0.264
	1.05	0.17	0.381	0.197	0.18	0.693	0.357
$\bar{n} = 200$							
Fixed	0.95	0.04	0.431	0.142	0.04	0.760	0.270
	0.96	0.05	0.285	0.104	0.05	0.552	0.186
	0.97	0.05	0.172	0.080	0.05	0.337	0.120
	0.98	0.05	0.103	0.061	0.05	0.171	0.078
	0.99	0.06	0.055	0.052	0.06	0.071	0.056
	1.00	0.06	**0.049**	**0.053**	0.06	**0.052**	**0.051**
	1.01	0.07	0.057	0.062	0.07	0.079	0.066
	1.02	0.07	0.107	0.079	0.07	0.176	0.106
	1.03	0.08	0.188	0.110	0.08	0.347	0.156
	1.04	0.08	0.324	0.152	0.08	0.580	0.226
	1.05	0.09	0.469	0.200	0.08	0.778	0.321
Variable	0.95	0.040	0.424	0.146	0.04	0.795	0.260
	0.96	0.050	0.285	0.107	0.05	0.609	0.187
	0.97	0.050	0.176	0.075	0.05	0.359	0.123
	0.98	0.050	0.096	0.056	0.06	0.183	0.076
	0.99	0.060	0.053	0.051	0.06	0.077	0.053
	1.00	0.060	**0.044**	**0.053**	0.07	**0.051**	**0.051**
	1.01	0.070	0.055	0.063	0.07	0.084	0.067
	1.02	0.070	0.106	0.081	0.08	0.185	0.104
	1.03	0.080	0.189	0.114	0.08	0.381	0.158
	1.04	0.080	0.319	0.152	0.09	0.614	0.234
	1.05	0.090	0.474	0.211	0.09	0.825	0.345

References

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

Davidson R. and J.G. MacKinnon. 1981. Several Tests for model specification in the presence of alternative hypotheses, Econometrica **49**: 781-793.

Gourieroux C., Monfort A. and Trognon A. 1984. Pseudo-maximum likelihood methods: theory. *Econometrica* **52** : 681-700.

Lambert D. 1992. Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics* **34** : 1-14.

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** : 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** : 121-133.

Ramalho, E.A., Ramalho, J.J.S. and Murteira J.M.R. 2010. Alternative estimating and testing empirical strategies for fractional regression models. *Journal of Economic Surveys* **forthcoming**.

Ramalho, J.J.S. and Vidigal da Silva J. 2009. A two-part fractional regression model for the financial leverage decisions of micro, small, medium and large firms. *Quantitative Finance* **9** : 621-636.

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