Schwiebert, Jörg; Wagner, Joachim

Conference Paper

A Generalized Two-Part Model for Fractional Response Variables with Excess Zeros

Beiträge zur Jahrestagung des Vereins für Socialpolitik 2015: Ökonomische Entwicklung - Theorie und Politik - Session: Microeconometrics, No. B04-V2

Provided in Cooperation with:
Verein für Socialpolitik / German Economic Association

Suggested Citation: Schwiebert, Jörg; Wagner, Joachim (2015) : A Generalized Two-Part Model for Fractional Response Variables with Excess Zeros, Beiträge zur Jahrestagung des Vereins für Socialpolitik 2015: Ökonomische Entwicklung - Theorie und Politik - Session: Microeconometrics, No. B04-V2, ZBW - Deutsche Zentralbibliothek für Wirtschaftswissenschaften, Leibniz-Informationszentrum Wirtschaft

This Version is available at:
https://hdl.handle.net/10419/113059

A Generalized Two-Part Model for Fractional Response Variables with Excess Zeros

Jörg Schwiebert* and Joachim Wagner**

Leuphana University Lüneburg

February 2015

Abstract

The fractional probit (or fractional logit) model is used when the outcome variable is a fractional response variable, i.e. a variable taking a value between zero and one. In case of excess zeros, the fractional probit model might not be the optimal modeling device since this model does not predict zeros. As a solution, the two-part model has been proposed, which assumes different processes for having a (non-)zero outcome and, conditionally on having a non-zero outcome, the actual outcome. However, the two-part model assumes independence of these processes. This paper proposes a generalization of the two-part model which allows for dependence of these processes and which also nests the two-part model as a special case. A simulation study indicates that the proposed estimator performs well in finite samples. Two empirical examples illustrate that the model proposed in this paper improves upon the fractional probit and two-part model in terms of model fit and also leads to different marginal effects.

Keywords: Fractional Logit, Fractional Probit, Fractional response variable, Two-part model
JEL codes: C25, C35, C51

*Leuphana University Lüneburg, Institute of Economics, Scharnhorststr. 1, 21335 Lüneburg, Germany, phone: +49.4131.677-2312, e-mail: schwiebert@leuphana.de.

**Leuphana University Lüneburg, Institute of Economics, Scharnhorststr. 1, 21335 Lüneburg, Germany, phone: +49.4131.677-2330, e-mail: wagner@leuphana.de.

1 Introduction

There are many examples in applied econometrics where the outcome variable of interest is a fractional response variable, i.e. a variable which only takes values between zero and one. A specific example for a fractional response variable is the share of exports in total sales (Wagner, 2001).

Econometric analysis of fractional response variables by ordinary least squares has the drawback that the model predictions may fall outside of the [0,1]-interval, so that the predictions are not consistent with the nature of the fractional outcome variable. To overcome this drawback, Papke and Wooldridge (1996) introduced the fractional probit model (or fractional logit model), which ensures that the predictions lie in the [0,1]-interval.

The fractional probit model assumes that the conditional mean of the fractional response variable y given a set of explanatory variables x is specified as

$$E[y|x] = \Phi(x'\beta), \tag{1}$$

where $\Phi(\cdot)$ is the standard normal cumulative distribution function and β is a vector of unknown parameters. This specification of the conditional mean ensures that the model's predicted values lie between zero and one, which is in line with the fractional nature of the outcome variable y.

However, in many empirical applications there is a large portion of observations having an outcome of zero. The fractional probit model may not be suitable in this case because it does not predict zeros. More importantly, the occurrence of excess zeros might indicate that the zero outcomes and the non-zero outcomes have been generated by two distinct processes. As a solution, the two-part model has been proposed (Ramalho et al., 2011), which assumes different processes for having a (non-)zero outcome and, conditionally on having a non-zero outcome, the actual outcome. As will be demonstrated below, however, the two-part model assumes independence of these processes.

This paper proposes a generalization of the (ordinary) two-part model which allows for

dependence of these processes and which also nests the two-part model as a special case. The model cannot only be applied when there are excess zeros, but also when outcomes are non-randomly missing in the spirit of the sample selection bias problem (Heckman, 1979). Estimation of this generalized two-part model can be carried out via quasi maximum likelihood. This paper also contains a small-scale simulation study illustrating that the proposed estimator performs well in finite samples and that the simpler two-part and fractional probit models yield biased estimates when there are dependencies between the two processes mentioned above. Furthermore, this paper includes two empirical examples which demonstrate that accounting for the dependence between the processes may be important in empirical work – both in terms of model fit and in terms of marginal effects.

It might be argued that a Tobit model could also be used when dealing with fractional response variables, especially in case of excess zeros. This approach has drawbacks, however. A Tobit model with a lower bound at zero would ensure that predictions are greater than zero; but it might also generate predictions being greater than one, which would be inconsistent with the nature of the fractional outcome variable. On the other hand, a Tobit model with a lower bound at zero and an upper bound at one would indeed yield predictions between zero and one; but this model would only be applicable if there was also an excess portion of ones. Moreover, it is not clear whether assuming censoring for a variable *defined* on the [0,1]-interval actually makes sense (see Ramalho et al., 2011, p. 22). Thus, the Tobit model is not considered further in this paper due to these conceptual reasons.

The remainder of this paper is organized as follows. Section 2 introduces the econometric model and discusses issues of identification, estimation and inference. Section 3 contains the simulation study. In Section 4 the proposed estimator and existing estimators are applied to two empirical examples. Finally, Section 5 concludes the paper.

2 Econometric Approach

This paper proposes a model which assumes distinct processes for having a (non-)zero outcome and, conditionally on having a non-zero outcome, the actual outcome. The first

process determines whether an observation i, $i = 1, \ldots, n$, has a zero outcome or not:

$$z_i = 1(w_i'\gamma + u_i > 0), \tag{2}$$

where z is an indicator variable equal to one if the observation has a non-zero outcome and zero otherwise; w is a vector of explanatory variables, γ is a vector of unknown parameters and u is the error term. Hence, an observation i has a non-zero outcome if and only if $w_i'\gamma + u_i > 0$. For convenience, it is assumed that the error term u has a standard normal distribution, implying that

$$Pr(z_i = 1|w_i) = \Phi(w_i'\gamma). \tag{3}$$

The second process determines the actual outcome, conditionally on having a non-zero outcome. In the two-part model the second part is described by the following conditional mean assumptions:

$$E[y_i|x_i, w_i, z_i = 0] = 0 \tag{4}$$

$$E[y_i|x_i, w_i, z_i = 1] = \Phi(x_i'\beta), \tag{5}$$

where the first equality directly follows from the definition of the variable z. The second equation implies that the conditional mean of the non-zero outcomes is specified as in the fractional probit model.

A drawback of the two-part model is that it assumes that $E[y_i|x_i, w_i, z_i = 1]$ does not depend on the first process, i.e. on z and its determinants. However, if the second process does depend on the first process, estimates from the two-part model will generally be biased. We thus propose the following generalization of the (ordinary) two-part model:

$$E[y_i|x_i, w_i, z_i = 1] = \frac{\Phi_2(x_i'\beta, w_i'\gamma; \rho)}{\Phi(w_i'\gamma)}, \tag{6}$$

where $\Phi_2(\cdot; \rho)$ denotes the bivariate standard normal distribution function with correlation coefficient ρ. Note that the first process is accounted for through the presence of $w_i'\gamma$.

Our formulation of the conditional mean has two advantages. The first advantage is that the conditional mean is always between zero and one, implying that also the predictions are between zero and one. The second advantage is that the two-part model is nested as a special case. If the correlation parameter ρ is equal to zero, the generalized model reduces to the simpler two-part model.

It is important to note that the vectors of explanatory variables x and w both affect the outcome of the second process, but in a slightly different way. The impact of x is *direct*, while the impact of w is only indirect via z.

The generalized two-part model can be estimated by quasi maximum likelihood (QML). The log-likelihood function is given by

$$\begin{aligned}\log L(\theta) = \sum_{i=1}^{n} l_i(\theta) \equiv \sum_{i=1}^{n} \{&(1-z_i)\log(1-\Phi(w_i'\gamma)) + z_i \log \Phi(w_i'\gamma) \\ &+ z_i \left[(1-y_i)\log\left(1 - \frac{\Phi_2(x_i'\beta, w_i'\gamma; \rho)}{\Phi(w_i'\gamma)}\right) + y_i \log \frac{\Phi_2(x_i'\beta, w_i'\gamma; \rho)}{\Phi(w_i'\gamma)} \right] \Big\}, \end{aligned} \tag{7}$$

where $\theta = (\beta', \gamma', \rho)'$ denotes the parameter vector to be estimated.

As described in Papke and Wooldridge (1996) or, in more detail, in Gourieroux et al. (1984), the QML approach ensures that the model parameters will be consistently estimated provided that the conditional means have been correctly specified. The QML approach thus imposes fewer assumptions than would be imposed if the whole (conditional) distribution of the outcome variable y was specified.

The QML estimator is given by

$$\hat{\theta} = \arg\max_{\theta} \log L(\theta). \tag{8}$$

The QML estimator has an asymptotic normal distribution and its estimated asymptotic variance matrix is of the sandwich-type (White, 1982), i.e.

$$Est.Asy.Var.(\hat{\theta}) = \left(-\sum_{i=1}^{n} \frac{\partial^2 l_i(\hat{\theta})}{\partial\theta\partial\theta'} \right)^{-1} \left(\sum_{i=1}^{n} \frac{\partial l_i(\hat{\theta})}{\partial\theta} \frac{\partial l_i(\hat{\theta})}{\partial\theta'} \right) \left(-\sum_{i=1}^{n} \frac{\partial^2 l_i(\hat{\theta})}{\partial\theta\partial\theta'} \right)^{-1}. \tag{9}$$

As usual, the standard errors of $\hat{\theta}$ can be obtained as the square root of the main diagonal

elements of this matrix.

Our generalized two-part model is conceptually similar to the Heckman sample selection model (Heckman, 1979). It can also be applied to situations where the outcome variable y has non-randomly missing observations. In that case, the first process does not generate zeros or non-zeros, but missing values and non-missing values. Hence, our model can also be interpreted as a sample selection model for fractional response variables.

Due to the similarity of our model with the Heckman sample selection model, it shares a common property which is often considered a drawback in applied empirical work. As in the Heckman sample selection model, an exclusion restriction is required for identification, i.e. the existence of a variable directly affecting the first process but not the second process. That means, there must exist at least one variable in w which does not appear in x. Actually, our model *is* identified by our functional form assumptions on the conditional means; but applied researchers typically do not wish to identify model parameters by functional form assumptions alone. Applied researchers thus often favor ordinary two-part models, as these do not require the existence of exclusion restrictions. However, if the second process and the first process are interrelated in the manner described above, the ordinary two-part model will yield biased estimates, so that researchers *should* use a generalized model which accounts for the dependencies between the first and second process.

Since the ordinary two-part model is nested within the generalized two-part model, it is straightforward to test if the ordinary two-part model is a valid description of the data generation process. Given the log-likelihood function specified above, a test of the null hypothesis $H_0 : \rho = 0$ has to be carried out. The null hypothesis implies that the generalized model reduces to the ordinary two-part model. In a QML setting, the Wald testing procedure probably provides the easiest way to test the null hypothesis. The test is a simple test of significance of the parameter ρ. If the null hypothesis is being rejected, this might indicate that the dependencies between first and second process are important and, thus, the generalized model should be preferred over the ordinary two-part model.

As in the fractional probit model and ordinary two-part model, a direct interpretation

of the model parameters is difficult. Researchers thus usually prefer interpretation of marginal effects rather than interpretation of parameters. In the generalized model, the marginal effect of a variable x_k is the change in

$$E[y_i|x_i, w_i] = Pr(z_i = 0|w_i)E[y_i|x_i, w_i, z_i = 0] + Pr(z_i = 1|w_i)E[y_i|x_i, w_i, z_i = 1] \quad (10)$$

$$= 0 + \Phi_2(x_i'\beta, w_i'\gamma; \rho) \quad (11)$$

due to a small change in x_k. Thus, the marginal effect is given by

$$\frac{\partial E[y_i|x_i, w_i]}{\partial x_k} = \frac{\partial \Phi_2(x_i'\beta, w_i'\gamma; \rho)}{\partial x_k} \quad (12)$$

for a given individual i, where x_k may be included in both x and w. The average marginal effect is computed as

$$\frac{1}{n}\sum_{i=1}^{n} \frac{\partial \Phi_2(x_i'\beta, w_i'\gamma; \rho)}{\partial x_k}, \quad (13)$$

which is simply the marginal effect averaged over all individuals.

3 Simulation Evidence

This section contains a small-scale simulation study in which data are generated according to the generalized two-part model developed in the last section. The purpose is to examine the performance of the QML estimator of the generalized two-part model, as well as to examine the bias which occurs if an ordinary two-part model, a fractional probit model or a linear model are used.

The data are generated as follows. The first process is characterized by

$$z_i = 1(\gamma_0 + \gamma_1 w_{i1} + \gamma_2 w_{i2} + u_i > 0), \quad (14)$$

$i = 1, \ldots, n$, where the u_i's are i.i.d. draws from a standard normal distribution. The

covariates w_{i1} and w_{i2} are generated as

$$w_{i1} = v_i + \eta_{1i} \tag{15}$$

$$w_{i2} = v_i + \eta_{2i}, \tag{16}$$

where the v_i's, η_{1i}'s and η_{2i}'s are also i.i.d. draws from a standard normal distribution. Hence, the covariates are assumed to exhibit some correlation, which is quite realistic in applications.

The second process is characterized by the conditional mean assumptions

$$E[y_i|x_i, w_i, z_i = 0] = 0 \tag{17}$$

$$E[y_i|x_i, w_i, z_i = 1] = \frac{\Phi_2(\beta_0 + \beta_1 x_i, \gamma_0 + \gamma_1 w_{i1} + \gamma_2 w_{i2}; \rho)}{\Phi(\gamma_0 + \gamma_1 w_{i1} + \gamma_2 w_{i2})}, \tag{18}$$

where $x_i = w_{i1}$ for all $i = 1, \ldots, n$. Note that the variable w_2 is not assumed to directly affect the second process; this is our exclusion restriction required for the identification of the model.

In order to generate data for the fractional response variable y which satisfy these conditional mean assumptions and which also satisfy the restriction that the response variable must lie in the [0,1]-interval, the beta distribution is used. The beta distribution is particularly convenient because it is defined for variables whose range is the [0,1]-interval and because it can be parameterized in terms of its mean. The probability density function of the beta distribution parameterized in this way is given by

$$f(y; \mu, \psi) = \frac{\Gamma(\psi)}{\Gamma(\mu\psi)\Gamma((1-\mu)\psi)} y^{\mu\psi - 1}(1-y)^{(1-\mu)\psi - 1}, \tag{19}$$

where μ denotes the mean, ψ is a shape parameter and $\Gamma(\cdot)$ is the gamma function (see Ramalho et al., 2011, p. 25). The fractional response variable y is thus generated

according to the rule

$$y_i \begin{cases} = 0 & \text{if } z_i = 0 \\ \sim f(y_i; \frac{\Phi_2(\beta_0+\beta_1 x_i, \gamma_0+\gamma_1 w_{i1}+\gamma_2 w_{i2}; \rho)}{\Phi(\gamma_0+\gamma_1 w_{i1}+\gamma_2 w_{i2})}, \psi) & \text{if } z_i = 1 \end{cases} . \tag{20}$$

The true values of the parameters are assumed to be: $\beta_0 = -1$, $\beta_1 = 0.5$, $\gamma_0 = 0$, $\gamma_1 = \gamma_2 = 1$, $\psi = 10$. The dependence parameter ρ is set to the values 0, 0.5 and 0.9 in order to analyze the estimator performance for different degrees of dependence.

The sample size n is set to 500, 1,000 and 2,000. Each simulation encompasses 1,000 repetitions. Over these repetitions, the mean of the parameter estimates as well as the root mean squared error are calculated. Also, the mean and standard deviation of the (average) marginal effect of variable x are calculated. Finally, the mean and standard deviation of model evaluation criteria are calculated. These model evaluation criteria are the Akaike information criterion (AIC), the Bayesian information criterion (BIC) and the R^2. These are defined by the following formulas (see, e.g., Greene, 2012):

$$AIC = \log(RSS/n) + 2K/n \tag{21}$$

$$BIC = \log(RSS/n) + (K \log(n))/n \tag{22}$$

$$R^2 = 1 - RSS/TSS, \tag{23}$$

where RSS denotes the residual sum of squares, TSS the total sum of squares and K is the number of estimated parameters. The residuals are defined as the deviation of y from the conditional expectation $E[y|x, w]$; TSS is defined in the usual way. Among a set of competing models, the model with the lowest values of AIC and BIC and the highest value of R^2 performs best. It would also have been possible to use the adjusted R^2 instead of the ordinary R^2. However, given the quite large sample sizes used in the simulation design the differences between R^2 and the adjusted R^2 can be reasonably expected to be small.

Simulation results based on four models are presented. The first model is the generalized two-part model developed in this paper. Since this model is the "true" data

generation model, the estimates should be unbiased. The second model is the ordinary two-part model. This model is expected to yield unbiased estimates only in the case $\rho = 0$, since in that case the generalized two-part model reduces to the ordinary two-part model. The third model is the fractional probit model. This model assumes that

$$E[y_i|x_i] = \Phi(\beta_0 + \beta_1 x_i), \tag{24}$$

and thus no distinction between zero outcomes and non-zero outcomes is made. The fourth and last model is a linear model which can be estimated by OLS. It assumes that

$$E[y_i|x_i] = \beta_0 + \beta_1 x_i, \tag{25}$$

hence no distinction between zero and non-zero outcomes is made and the fractional nature of the dependent variable y is ignored. Estimates based on the fractional probit model and linear model are expected to be biased regardless of the value of ρ.

The simulation results are presented in Tables 1-3. Table 1 considers the case $\rho = 0$, Table 2 the case $\rho = 0.5$ and Table 3 the case $\rho = 0.9$. As expected, the parameter estimates from the generalized two-part model are virtually unbiased for all values of ρ and all sample sizes. Also as expected, the parameter estimates from the ordinary two-part model are unbiased only when $\rho = 0$; in the other cases, the bias increases with the degree of correlation ρ. The parameter estimates from the fractional probit and linear models are clearly biased, irrespective of the value of ρ and/or the sample size. It is interesting to note, however, that the fractional probit model leads to less biased estimates than the ordinary two-part model when $\rho = 0.9$.

Since the parameter estimates are not comparable across models due to the different assumptions identifying their pseudo-true values, it may be more reasonable to compare marginal effects. As expected, the estimated marginal effects for variable x are virtually identical for the generalized two-part model and ordinary two-part model in case of $\rho = 0$. As ρ increases, the marginal effects begin to differ. Interestingly, the marginal effects of the fractional probit model and linear model are virtually identical for all sample sizes

and seem not to strongly depend on the value of ρ. Nevertheless, compared with the generalized two-part model which is the "true" model, the results indicate an upward bias of the estimated marginal effects from the fractional probit and linear models.

The model evaluation criteria in Tables 1-3 suggest for all sample sizes and values of ρ that the generalized two-part model performs best, which would have been expected since this model represents the "true" model. Only in the case $\rho = 0$ generalized and ordinary two-part model perform virtually identically (as expected). Again, the fit of the ordinary two-part model relative to the generalized model deteriorates as ρ increases. Interestingly, the fractional probit model outperforms the ordinary two-part model when $\rho = 0.9$. In the other cases the ordinary two-part model performs better than the fractional probit model. Finally, as expected, the linear model is generally the worst model in terms of model fit.

In summary, the simulation results show that parameter estimates and estimated marginal effects from the ordinary two-part model are biased when there is dependence between the first and second process, i.e. when $\rho \neq 0$. Moreover, the estimates from the fractional and linear models are biased as well. The results also indicate that AIC, BIC and R^2 are useful measures to find the best model in terms of model fit, at least when the underlying true model is the generalized two-part model.

4 Empirical Examples

4.1 Empirical application 1: Export behavior of firms

The first application deals with the export behavior of German enterprises from manufacturing industries. Germany is one of the "big players" on the world market for goods, ranking number three after the USA and China in 2013 (Word Trade Organization, 2014, p. 34). Exports play a decisive role in shaping the dynamics of the German economy as a whole, its regions and industries, and its firms. Therefore, reliable empirical evidence on the determinants of export participation of firms and of the share of exports in total sales of the firms is important.

The data set used in this empirical application is described in the appendix. Descriptive statistics are also provided there.

The fractional outcome variable considered here is the share of exports in total sales. In our data set, a large share of firms (about 20%) does not export at all, i.e., there is a large portion of firms having an outcome of zero. Thus, a two-part model seems to be an appropriate modeling device. Since the two-part model involves two distinct processes generating the data, explanatory variables characterizing these processes have to be specified.

The first process determines whether a firm is an exporter or not. Explanatory variables affecting the first process are assumed to be firm size ($fsize$), human capital intensity (hc), R&D intensity (rd), capital intensity (kl), firm age (old), a dummy indicating whether the firm was foreign-owned (fof), a set of industry dummies and a dummy indicating whether a firm was located in West Germany ($west$). In the appendix it is briefly discussed why these variables are assumed to affect export behavior.

The second process determines the size of a firm's share of exports in total sales, given that a firm is an exporter. As discussed in Section 2, an exclusion restriction is required for identification, i.e. the existence of a variable directly affecting the first process but not the second process. In our specification of the exclusion restriction we follow Arndt et al. (2012), who used a Heckman selection model to explain export behavior of German firms. Their outcome variable of interest was not the share of exports in total sales but the volume of exports. In order to identify their selection model, Arndt et al. (2012) excluded a dummy variable indicating whether a firm was located in East Germany from the variables directly affecting the volume of exports. Since such a variable is also available in our data set ($west$), we proceed in the same way. Thus, the location of a firm in West Germany (yes/no) is assumed to directly affect the first process whether a firm is an exporter or not, but not the second process about the size of the share of exports in total sales, given that a firm is an exporter. Since the remaining variables characterizing the first process are also assumed to be determinants of the second process, our set of variables characterizing the second process includes the same variables as the first process

apart from *west*.

We estimated five different models. The first model is the generalized two-part model proposed in this paper. The second model is an ordinary two-part model with the same exclusion restriction as in our generalized two-part model. However, since the ordinary two-part model does not require an exclusion restriction, we also estimated this model without exclusion restriction, i.e., the variable *west* is also included into the set of variables directly affecting the second process. Since we believe that our exclusion restriction is valid, the inclusion of this third model should be interpreted in terms of a robustness check. The empirical results given below indicate that there are no large differences between the two-part models with and without exclusion restriction. The fourth and fifth models under consideration are the fractional probit model and the linear model (OLS), respectively. Both models do not explicitly account for the excess zeros, and the linear model does not even account for the fractional nature of the outcome variable. These latter two models are used to analyze whether results based on these simple models differ from those obtained from the seemingly more appropriate two-part models.

Since estimates of the model parameters are not comparable across models, as discussed in the last section, we computed (average) marginal effects which are comparable across models. Also as in the last section, we computed the values of AIC, BIC and R^2 in order to assess the performance of the models. The results of these computations are given in Table 4. The parameter estimates are not reported due to brevity, but are available from the authors upon request. All estimations and computations have been done in Stata.

Table 4 shows that the marginal effects of firm size ($fsize$) and R&D intensity (rd) are larger in the generalized two-part model than in the remaining models. Especially the fractional probit model and OLS seem to understate these effects. The marginal effects of the remaining variables are relatively similar. The dependence parameter ρ of the generalized two-part model is estimated as -0.484 with a standard error of 0.053. A Wald test indicates that this parameter is significantly different from zero at the 1% significance level, hence the ordinary two-part model (with exclusion restriction) is rejected. Since

the marginal effects of the two-part models with and without exclusion restriction are rather similar in Table 4, the Wald test may also indicate that the ordinary two-part model without exclusion restriction is rejected. The model evaluation criteria are quite close. AIC and R^2 favor the generalized two-part model, while BIC favors OLS. Since BIC imposes a larger penalty on the number of parameters than the other criteria, this result indicates that BIC favors OLS due to the parsimony of the linear model.

The generalized and ordinary two-part models also allow to compute marginal effects at the intensive and extensive margins of export. The marginal effect at the intensive margin is the marginal effect for those firms who do export. Formally, the marginal effect of a variable x_k at the intensive margin is given by

$$\frac{\partial E[y_i|x_i, w_i, z_i = 1]}{\partial x_k} \tag{26}$$

for a given firm i, where x_k may be included in both x and w. The marginal effect at the extensive margin is the marginal effect on the export decision (yes/no). Formally, the marginal effect of a variable w_k at the extensive margin is given by

$$\frac{\partial Pr(z_i = 1|w_i)}{\partial w_k}. \tag{27}$$

Since both generalized and ordinary two-part models rely on identical first processes, the marginal effects at the extensive margin should be very close across the models. However, since the specification of $E[y_i|x_i, w_i, z_i = 1]$ is different across the two-part models, differences in the marginal effects at the intensive margin are expected.

The (average) marginal effects at the intensive and extensive margins are given in Table 5. As expected, the marginal effects at the extensive margins are very close. However, at the intensive margins differences occur, most notably in the variables $fsize$ and rd. The ordinary two-part models with and without exclusion restriction seem to understate the respective effects.

In summary, the empirical results indicate that the generalized two-part model leads to different marginal effects than the ordinary two-part models with/without exclusion

restriction and the fractional probit and linear models, at least in the two variables $fsize$ and rd. Moreover, a Wald test indicates that the generalized two-part model should be preferred over the ordinary two-part models with/without exclusion restrictions. This result is partly confirmed by the model evaluation criteria, as both AIC and R^2 suggest that the generalized two-part model performs best.

4.2 Empirical application 2: Product diversification of firms

The second empirical application deals with product diversification of firms. In Germany, nearly 40 percent of all manufacturing enterprises with at least 20 employees are single-product firms according to a detailed classification of products, and they do not diversify in product-space. Multi-product enterprises producing a large number of goods are a rare species (Wagner, 2009). Given that the links between product diversification of a firm and various dimensions of its performance like stability of employment and profitability (see Braakmann and Wagner, 2011a, 2011b) are important for an understanding of the dynamics of firms and markets, reliable empirical evidence on the determinants of the degree of product diversification of firms is important.

The data set used in this empirical application is described in the appendix. Descriptive statistics are also provided there.

The fractional outcome variable considered here is the share of the most important product of a firm in its total sales, which measures the degree of product diversification. Since there is a large portion of firms producing a single good only (about 35% in our data set), the outcome variable is equal to one for a large share of firms. Thus, a two-part model seems to be an appropriate modeling device. In order to fit into our framework, we transform the outcome variable by considering (1– share of the most important product of a firm in its total sales). This transformation ensures that the excess ones are transformed into excess zeros, which fit well in the framework developed in this paper. As in the first empirical example, explanatory variables characterizing the first and second process of the two-part model have to be specified.

The first process determines whether a firm is a multi-product firm or not. Explana-

tory variables affecting the first process are assumed to be firm size ($fsize$), human capital intensity (hc), R&D intensity (rd), capital intensity (kl), firm age (old), a dummy indicating whether the firm was foreign-owned (fof) and a set of industry dummies. In the appendix it is briefly discussed why these variables are assumed to affect the product diversification of firms.

The second process determines the size of (1– a firm's share of the most important product in total sales), i.e. the magnitude of product diversification, given that a firm is a multi-product firm. Again, an exclusion restriction is required for identification, i.e. the existence of a variable directly affecting the first process but not the second process. In our case this variable is firm age. Usually a firm is founded to pursue an idea for a new or much improved product. This means that a newly founded firm will often produce a single product only. If this firm survives, usually it will add other products to its portfolio. Therefore, firm age is expected to affect whether a firm produces exactly one good or more than one good. On the other hand, given that a firm produces more than one good, the degree of product diversification of a firm is expected to depend on the amount and quality of resources available to the firm (and not further on its age). Thus, the firm age is assumed to directly affect the first process whether a firm is a multi-product firm or not, but not the second process about the magnitude of product diversification, given that a firm is a multi-product firm. Since the remaining variables characterizing the first process are also assumed to be determinants of the second process, our set of variables characterizing the second process includes the same variables as the first process apart from old.

We estimated the same models as in the first application. The results are given in Tables 6 and 7. Table 6 contains the (average) marginal effects and model evaluation criteria, while Table 7 includes the (average) marginal effects at the intensive and extensive margins. Table 6 reveals differences in marginal effects across models, most notably for the variables $fsize$ and hc. While the effect of $fsize$ is larger in the generalized two-part model than in the remaining models, it is smaller in case of hc. The remaining variables also exhibit differences in marginal effects across models, albeit to a smaller

amount. In contrast to the first application, all three model evaluation criteria indicate that the generalized two-part model performs best. This result is complemented by the significance of the dependence parameter ρ at the 1% significance level; the parameter ρ is estimated as -0.339 with a standard error of 0.045. Hence, the ordinary two-part model with exclusion restriction is rejected. Since the ordinary two-part model without exclusion restriction again performs similarly to the model with exclusion restriction, one may conclude that the two-part model without exclusion restriction is rejected as well.

Furthermore, the marginal effects at the intensive margin also differ across models, most notably for the variables $fsize$, hc and rd; see Table 7. The marginal effects at the extensive margin are again very close, as expected.

Compared with the first application, the second example reveals stronger differences in marginal effects across models. Moreover, the model evaluation criteria suggest more clearly that the generalized two-part model should be preferred, as all criteria came to the same result.

Taken together, both examples suggest that the generalized two-part model outperforms the ordinary two-part model and the simpler fractional probit and linear models. Since the marginal effects also differ, the results clearly indicate that it is important in applications to use the generalized two-part model when there exist dependencies between the two processes generating the data.

5 Conclusions

This paper proposed a generalization of the two-part model for fractional response variables. The ordinary two-part model ignores the dependence between the two processes generating the data, which generally leads to biased estimates when dependence exists. A simulation study illustrated these biases.

The most challenging problem for the application of the proposed model in practice seems to be to find a compelling exclusion restriction. Two empirical examples were presented where such an exclusion restriction could be identified. However, the lack of an available exclusion restriction should not lead practitioners to use the ordinary two-part

model which does not require such an exclusion restriction. As we showed in this paper, using the "wrong" model leads to biased estimates when the two processes generating the data are indeed dependent.

References

Arndt, C., Buch, C.M. and Mattes, A. (2012). Disentangling barriers to internationalization. *Canadian Journal of Economics* 45, 41-63.

Braakmann, N. and Wagner, W. (2011a). Product diversification and stability of employment and sales: first evidence from German manufacturing firms. *Applied Economics* 43, 3977-3985.

Braakmann, N. and Wagner, J. (2011b). Product diversification and profitability in German manufacturing firms. *Jahrbücher für Nationalökonomie und Statistik* 231, 326-335.

Fritsch, M. Görzig, B., Hennchen, O. and Stephan, A. (2004). Cost structure surveys for Germany. *Schmollers Jahrbuch / Journal of Applied Social Science Studies* 124, 557-566.

Gourieroux, C., Monfort, A. and Trognon, A. (1984). Pseudo maximum likelihood methods: applications to Poisson models. *Econometrica* 52, 701-720.

Greene, W.H. (2012). *Econometric Analysis*. 7th, international edition. Pearson, Boston, Mass.

Heckman, J.J. (1979). Sample selection bias as a specification error. *Econometrica* 47, 153-161.

Malchin, A. and Voshage, R. (2009). Official firm data for Germany. *Schmollers Jahrbuch / Journal of Applied Social Science Studies* 129, 501-513.

Montgomery, C.A. (1994). Corporate diversification. *Journal of Economic Perspectives* 8, 163-178.

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.

Raff, H. and Wagner, J. (2014). Foreign ownership and the extensive margins of exports: evidence for manufacturing enterprises in Germany. *The World Economy* 37, 579-591.

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, 19-68.

Wagner, J. (2001). A note on the firm size – export relationship. *Small Business Economics* 17, 229-237.

Wagner, J. (2009). Produktdifferenzierung in deutschen Industrieunternehmen 1995 - 2004: Ausmaß und Bestimmungsgründe. *Jahrbücher für Nationalökonomie und Statistik* 229, 615-642.

Wagner, J. (2011a). Exports and firm characteristics in Germany: A Survey of Empirical Studies (1991 to 2011). *Applied Economics Quarterly* 57, 145-160.

Wagner, J. (2011b). Exports and firm characteristics in German manufacturing industries: new evidence from representative panel data. *Applied Economics Quarterly* 57, 107-143.

Wagner, J. (2012). Average wage, qualification of the workforce and export performance in German enterprises: evidence from KombiFiD data. *Journal of Labour Market Research* 45, 161-170.

Wagner, J. (2014a). A note on firm age and the margins of exports: first evidence from Germany. University of Lüneburg Working Paper Series in Economic No. 303 (forthcoming, *International Trade Journal*).

Wagner, J. (2014b). Still different after all these years. Extensive and intensive margins of exports in East and West German manufacturing enterprises. University of Lüneburg Working Paper Series in Economic No. 313.

White, H.L. (1982). Maximum likelihood estimation of misspecified models. *Econometrica* 50, 1-25.

World Trade Organization (2014). World Trade Report 2014. Trade and development: recent trends and the role of the WTO. Geneva: WTO Publishing.

Appendix

Empirical application 1: data and definition of variables

The empirical investigation of the participation in exports and the share of exports in total sales of firms from German manufacturing industries uses a tailor-made data set that combines high quality firm-level data from three official sources. The first source of firm level information is the regular survey of establishments from manufacturing industries by the Statistical Offices of the German federal states. The survey (known as the *Monatsbericht*, or monthly report) covers all establishments from manufacturing industries that employ at least twenty persons in the local production unit or in the company that owns the unit. Participation of firms in the survey is mandated in official statistics (see Malchin and Voshage, 2009, for details). For this study the monthly establishment data were aggregated to annual data and at the enterprise level to match the unit of observation in the other data sources (described below).

The second source of data is the cost structure survey for enterprises in the manufacturing sector. This survey is carried out annually as a representative random sample survey in about 15,000 firms. The sample is stratified according to the number of employees and the industries; all firms with 500 and more employees are covered by the cost structure survey (see Fritsch et al., 2004).

These data were matched with the enterprise register system (*Unternehmensregister-*

System) which is the third source of data.

With these linked data sets it is possible to investigate the margins of exports in firms from manufacturing industries in Germany. The definition of the variables used in the empirical models is discussed in detail below.

Exporter status (exp): The extensive margin measures the participation of a firm in exports (or not).

Share of exports in total sales (expshare): The intensive margin of export is the percentage share of all sales due to exports.

Information on the exporter status of a firm and on the share of exports in total sales of a firm is based on information on export sales and total turnover taken from the first data source (the monthly report). This information is available for all firms from manufacturing industries with at least twenty employees.

Firm size (fsize): A positive link between firm size and margins of exports qualifies as a stylized fact. This positive link is due to fixed costs of exporting and efficiency advantages of larger firms due to scale economies, advantages of specialization in management and better conditions on the markets for inputs. Large firms can be expected to have cost advantages on credit markets while small firms often face higher restrictions on the capital market leading to a higher risk of insolvency and illiquidity. Furthermore, there might be disadvantages of small firms in the competition for highly qualified employees. There are limits to the advantage of size, because coordination costs mount as the scale of operations increases, and at some point any further expansion might cease to be profitable. Therefore, a positive relationship between firm size and exports, at least up to a point, is expected. For Germany empirical evidence in line with this is reported in a number of studies (see Wagner, 2011a, for a survey). Firm size is measured here by the number of employees in a firm (also included in squares to take care of non-linearity). The source is the first data set (the monthly report).

Human capital intensity (hc): Given that Germany is relatively rich in human capital, firms that use human capital intensively can be expected to have a comparative advantage on international markets. Empirical studies find that the qualification of the workforce

is an important factor for the international competitiveness of German firms (Wagner, 2011b). Human capital intensity is measured here by the average wage per employee. Direct information on the qualification of the employees in a firm is not available in the data used in this study, but Wagner (2012) demonstrates that the average wage is indeed a good proxy variable for the qualification of the workforce in German manufacturing firms. The source is for information on the amount of wages paid and the number of employees is the first data set (the monthly report).

R&D intensity (rd): Activities in research and development that are closely related to product and process innovations are known to be positively linked to success in exports in German firms (see Wagner, 2011a, 2011b). R&D intensity is measured here by the share of employees that are active in R&D in all employees in a firm. This intensity measure is based on information on R&D employees and total employees taken from the second data source (the cost structure survey).

Capital intensity (kl): The amount of capital used per employee is traditionally expected to be positively linked to exports in a relatively capital-abundant country like Germany. In the data used in this study, however, there is no direct information on the capital stock of the firms. Therefore, the amount of depreciation per employee is used as a proxy variable that can be expected to be (more or less) proportional to the amount of capital per head. Information on the amount of depreciation and the number of employees is taken from the second data source (the cost structure survey).

Firm age (old): Although some newly founded firms are "born globals" that export from the start, typically it takes years before firms eventually export to one foreign market, and then enter further markets progressively. Firms gain expertise in entering new foreign markets from experience, and this lowers the fixed costs of entry to any further new market. A similar argument can be made with regard to the number of products exported. At any point in time, therefore, firm age and the margins of exports can be expected to be closely linked. Germany is a case in point. Wagner (2014a) reports that older firms are more often exporters, export more and more different goods to more different destination countries. Information on firm age is not available from the data used in this study.

However, we know whether a firm was already active in 1995 (the first year data from the monthly report are available for). Firms that were active in 1995, and that were founded before 1996 accordingly, are classified as old firms (based on this information from the first data source, the monthly report).

Foreign owned firm (fof): Firms that are subsidies of a multinational enterprise that has its headquarter in a foreign country are termed foreign owned firms. Foreign ownership is known to have a positive impact on the margins of exports, because these firms can use the international networks and trade contacts of their parent companies and are involved in international supply chains (see Raff and Wagner, 2014, for a discussion of the literature, a theoretical model, and empirical evidence for Germany). A firm is considered to be foreign owned if more than 50 percent of the voting rights of the owners or more than 50 percent of the shares are controlled (directly or indirectly) by a firm or a person/institution located outside Germany. Information on foreign ownership status of an enterprise is taken from the fourth source of data, the enterprise register system.

Industry: Dummy variables for 2digit-industries are included in the empirical models to control for industry specific effects like competitive pressure, policy measures, demand shocks etc. The source is the first data set (the monthly report).

West (west): A dummy variable indicating whether (the headquarter of) a firm is located in West Germany or not (i.e., in the former communist East Germany). It is well known that the propensity to export and the share of exports in total sales is considerably higher in West German firms compared to East German firms even more than 20 years after the re-unification of both parts of Germany back in 1990 (see Wagner, 2014b).

Descriptive statistics of these variables (apart from the industry dummies) are summarized in Table 8.

Empirical application 2: data and definition of variables

Why do some firms diversify, i.e. why do they produce more than one good and spread activities across markets, while others focus their economic activity on one product only?

According to the resource view (Montgomery, 1994, p. 167f.) firms that have an excess

capacity in productive factors – for example, special knowledge the firm has accumulated through time, and that can be used in other markets without reducing the use in the market the firm is already active in - can reap economies of scope by expanding into different product markets. Alternatively, the firm may sell this specific asset to another firm active in this market. However, it is reasonable to expect that market failure does exist when it comes to trade in intangible assets like knowledge, and this is an incentive to internalize the use of the assets. Furthermore, productive factors of this type are often closely linked to persons who cannot simultaneously work for several firms producing different products.

These theoretical considerations can guide the specification of an empirical model for the determinants of product diversification. The definition of the variables included in this model, its source, and its links to theory are discussed in detail below.

Information on the number of products of a firm and on the amount of sales due to each product come from the regular *survey of products produced* (*Produktionsstatistik*) performed by the German statistical offices. This information is used to construct the two dependent variables of the empirical models:

Multi-product firm (multi): A dummy variable taking on the value of one if a firm produces more than one product.

Degree of diversification (degree): Computed as (1–share of sales due to most important product in total sales) of a firm.

Information on the independent variables included in the empirical models come from three different sources that are discussed in some detail in the first part of this appendix, namely the monthly report (firm size, wage per employee, firm age, and industry affiliation), the cost structure survey (R&D intensity, capital intensity) and the enterprise register system (foreign ownership).

Firm size (fsize): Firm size is measured by the number of employees in a firm (also included in squares to take care of non-linearity).

Human capital intensity (hc): Human capital intensity is measured by the average wage per employee.

R&D intensity (rd): Research and development intensity is measured by the share of employees active in R&D in all employees in a firm.

A positive relationship between product diversification and firm size, human capital intensity, and R&D intensity can be expected because an increase in each of these firm characteristics tends to go hand in hand with an increase in the resources that are available for diversification.

Capital intensity (kl): The amount of depreciation per employee is used as a proxy variable due to the lack of more direct information on the capital stock of the firm. This variable is included as a control variable only.

Firm age (old): Due to missing information with regard to the founding year of the firm in the data, firms that were active in 1995 already are classified as old firms in a dummy variable defined accordingly. Given that manufacturing firms are often founded to realize the idea for one new product it is expected that the link between firm age and product diversification is positive, because older firms will more often have accumulated the resources needed to diversify in product space.

Foreign owned firm (fof): A firm is considered to be foreign owned if it is controlled to more than 50 percent by a firm or a person located outside Germany. This variable is included as a control variable in the empirical model.

Industry: Dummy variables for 2digit-industries are included to control for differences in the level of diversification between industries.

Descriptive statistics of these variables (apart from the industry dummies) are summarized in Table 9.

Tables

Table 1: Simulation results for $\rho = 0$

	Generalized Two-part		Two-part		Fractional Probit		OLS	
	Mean	RMSE/SD	Mean	RMSE/SD	Mean	RMSE/SD	Mean	RMSE/SD
				n=500				
Parameters								
β_0	-1.000	0.056	-1.001	0.040	-1.417	0.419	0.156	1.156
β_1	0.500	0.033	0.500	0.027	0.693	0.195	0.118	0.382
γ_0	-0.007	0.089	-0.007	0.089				
γ_1	1.014	0.111	1.014	0.111				
γ_2	1.010	0.109	1.010	0.109				
ρ	0.000	0.085						
Marginal effect of x	0.105	0.005	0.105	0.005	0.118	0.005	0.118	0.006
Model evaluation criteria								
AIC	-4.409	0.102	-4.412	0.102	-4.250	0.094	-3.885	0.064
BIC	-4.358	0.102	-4.369	0.102	-4.233	0.094	-3.868	0.064
R^2	0.751	0.027	0.751	0.027	0.704	0.031	0.575	0.028
				n=1,000				
Parameters								
β_0	-0.999	0.041	-0.999	0.028	-1.414	0.415	0.157	1.157
β_1	0.500	0.024	0.500	0.019	0.692	0.193	0.118	0.382
γ_0	0.000	0.064	0.000	0.064				
γ_1	1.007	0.078	1.007	0.078				
γ_2	1.011	0.077	1.011	0.077				
ρ	0.000	0.063						
Marginal effect of x	0.105	0.004	0.105	0.004	0.118	0.004	0.118	0.004
Model evaluation criteria								
AIC	-4.417	0.070	-4.419	0.070	-4.250	0.065	-3.885	0.047
BIC	-4.388	0.070	-4.394	0.070	-4.240	0.065	-3.875	0.047
R^2	0.753	0.018	0.753	0.018	0.705	0.021	0.576	0.019
				n=2,000				
Parameters								
β_0	-1.000	0.029	-1.000	0.020	-1.414	0.414	0.157	1.157
β_1	0.500	0.017	0.500	0.014	0.692	0.192	0.118	0.382
γ_0	0.000	0.045	0.000	0.045				
γ_1	1.006	0.053	1.006	0.053				
γ_2	1.006	0.052	1.006	0.052				
ρ	0.001	0.044						
Marginal effect of x	0.104	0.003	0.104	0.003	0.118	0.003	0.118	0.003
Model evaluation criteria								
AIC	-4.418	0.050	-4.418	0.050	-4.249	0.047	-3.885	0.033
BIC	-4.401	0.050	-4.404	0.050	-4.244	0.047	-3.879	0.033
R^2	0.751	0.013	0.751	0.013	0.705	0.016	0.575	0.014

Note: The root mean squared errors (RMSE) refer to the parameters, while the standard deviations (SD) refer to the marginal effect of x and the model evaluation criteria. The true values of the parameters are $\beta_0 = -1$, $\beta_1 = 0.5$, $\gamma_0 = 0$, $\gamma_1 = 1$ and $\gamma_2 = 1$. The simulation results are based on 1,000 repetitions.

Table 2: Simulation results for $\rho = 0.5$

	Generalized Two-part		Two-part		Fractional Probit		OLS	
	Mean	RMSE/SD	Mean	RMSE/SD	Mean	RMSE/SD	Mean	RMSE/SD
				n=500				
Parameters								
β_0	-0.997	0.049	-0.780	0.224	-1.253	0.256	0.175	1.175
β_1	0.500	0.029	0.403	0.101	0.629	0.131	0.121	0.380
γ_0	-0.001	0.090	-0.001	0.090				
γ_1	1.024	0.109	1.025	0.111				
γ_2	1.022	0.111	1.022	0.112				
ρ	0.493	0.089						
Marginal effect of x	0.109	0.006	0.103	0.005	0.121	0.005	0.121	0.006
Model evaluation criteria								
AIC	-4.039	0.099	-4.009	0.103	-3.963	0.085	-3.729	0.062
BIC	-3.989	0.099	-3.966	0.103	-3.946	0.085	-3.712	0.062
R^2	0.673	0.034	0.661	0.036	0.642	0.034	0.548	0.030
				n=1,000				
Parameters								
β_0	-0.998	0.035	-0.777	0.225	-1.251	0.253	0.175	1.175
β_1	0.500	0.021	0.401	0.101	0.627	0.129	0.120	0.380
γ_0	-0.003	0.061	-0.003	0.062				
γ_1	1.007	0.078	1.008	0.080				
γ_2	1.009	0.075	1.009	0.075				
ρ	0.497	0.061						
Marginal effect of x	0.109	0.004	0.103	0.004	0.121	0.004	0.120	0.004
Model evaluation criteria								
AIC	-4.042	0.069	-4.010	0.073	-3.960	0.060	-3.730	0.046
BIC	-4.012	0.069	-3.985	0.073	-3.950	0.060	-3.720	0.046
R^2	0.670	0.023	0.658	0.024	0.639	0.023	0.546	0.020
				n=2,000				
Parameters								
β_0	-1.000	0.025	-0.778	0.223	-1.251	0.252	0.175	1.175
β_1	0.500	0.015	0.400	0.101	0.626	0.127	0.120	0.380
γ_0	0.000	0.044	0.000	0.044				
γ_1	1.002	0.053	1.002	0.054				
γ_2	1.003	0.052	1.003	0.052				
ρ	0.499	0.046						
Marginal effect of x	0.109	0.003	0.102	0.003	0.120	0.002	0.120	0.003
Model evaluation criteria								
AIC	-4.046	0.051	-4.013	0.053	-3.963	0.044	-3.730	0.032
BIC	-4.030	0.051	-3.999	0.053	-3.957	0.044	-3.724	0.032
R^2	0.670	0.017	0.659	0.018	0.640	0.017	0.546	0.015

Note: The root mean squared errors (RMSE) refer to the parameters, while the standard deviations (SD) refer to the marginal effect of x and the model evaluation criteria. The true values of the parameters are $\beta_0 = -1$, $\beta_1 = 0.5$, $\gamma_0 = 0$, $\gamma_1 = 1$ and $\gamma_2 = 1$. The simulation results are based on 1,000 repetitions.

Table 3: Simulation results for $\rho = 0.9$

	Generalized Two-part		Two-part		Fractional Probit		OLS	
	Mean	RMSE/SD	Mean	RMSE/SD	Mean	RMSE/SD	Mean	RMSE/SD
				n=500				
Parameters								
β_0	-0.995	0.043	-0.620	0.383	-1.143	0.152	0.189	1.189
β_1	0.499	0.029	0.327	0.177	0.582	0.088	0.121	0.379
γ_0	-0.006	0.092	-0.005	0.094				
γ_1	1.024	0.107	1.025	0.116				
γ_2	1.027	0.112	1.026	0.113				
ρ	0.882	0.064						
Marginal effect of x	0.112	0.006	0.100	0.006	0.121	0.005	0.121	0.006
Model evaluation criteria								
AIC	-3.642	0.120	-3.556	0.131	-3.607	0.108	-3.466	0.084
BIC	-3.591	0.120	-3.514	0.131	-3.590	0.108	-3.449	0.084
R^2	0.573	0.043	0.533	0.051	0.551	0.042	0.485	0.035
				n=1,000				
Parameters								
β_0	-0.996	0.030	-0.619	0.383	-1.142	0.146	0.189	1.189
β_1	0.498	0.021	0.325	0.177	0.580	0.083	0.121	0.380
γ_0	-0.003	0.060	-0.002	0.062				
γ_1	1.013	0.072	1.013	0.079				
γ_2	1.024	0.079	1.024	0.080				
ρ	0.887	0.043						
Marginal effect of x	0.112	0.004	0.099	0.004	0.121	0.004	0.121	0.004
Model evaluation criteria								
AIC	-3.644	0.085	-3.556	0.092	-3.604	0.076	-3.465	0.062
BIC	-3.615	0.085	-3.532	0.092	-3.594	0.076	-3.455	0.062
R^2	0.570	0.030	0.530	0.035	0.549	0.028	0.482	0.024
				n=2,000				
Parameters								
β_0	-0.998	0.021	-0.619	0.382	-1.142	0.144	0.189	1.189
β_1	0.499	0.014	0.325	0.176	0.580	0.081	0.121	0.380
γ_0	-0.004	0.042	-0.004	0.043				
γ_1	1.008	0.048	1.008	0.052				
γ_2	1.017	0.053	1.017	0.054				
ρ	0.891	0.030						
Marginal effect of x	0.112	0.003	0.099	0.003	0.121	0.003	0.121	0.003
Model evaluation criteria								
AIC	-3.648	0.060	-3.559	0.065	-3.605	0.053	-3.466	0.042
BIC	-3.631	0.060	-3.545	0.065	-3.599	0.053	-3.460	0.042
R^2	0.570	0.021	0.529	0.025	0.549	0.021	0.482	0.017

Note: The root mean squared errors (RMSE) refer to the parameters, while the standard deviations (SD) refer to the marginal effect of x and the model evaluation criteria. The true values of the parameters are $\beta_0 = -1$, $\beta_1 = 0.5$, $\gamma_0 = 0$, $\gamma_1 = 1$ and $\gamma_2 = 1$. The simulation results are based on 1,000 repetitions.

Table 4: Empirical application 1: estimated average marginal effects

	Generalized two-part		Two-part with exclusion		Two-part without exclusion		Fractional probit		OLS	
Variable	Marg. eff.	(SE)	Marg. eff.	(SE)	Marg. eff.	(SE)	Marg. eff.	(SE)	Marg. eff.	(SE)
fsize	0.564	(0.156)	0.363	(0.085)	0.362	(0.085)	0.171	(0.033)	0.211	(0.038)
hc	0.570	(0.024)	0.585	(0.023)	0.570	(0.025)	0.577	(0.026)	0.605	(0.026)
rd	0.713	(0.079)	0.564	(0.054)	0.571	(0.054)	0.433	(0.040)	0.550	(0.048)
kl	0.866	(0.287)	0.894	(0.287)	0.978	(0.295)	0.805	(0.308)	0.637	(0.339)
old	0.019	(0.004)	0.022	(0.004)	0.021	(0.004)	0.022	(0.004)	0.018	(0.004)
fof	0.109	(0.007)	0.105	(0.006)	0.106	(0.006)	0.107	(0.006)	0.119	(0.007)
west	0.013	(0.004)	0.007	(0.002)	0.018	(0.006)	0.021	(0.006)	0.019	(0.006)
Model evaluation criteria										
AIC	-3.024		-3.017		-3.017		-3.011		-3.013	
BIC	-2.992		-2.986		-2.985		-2.995		-2.997	
R^2	0.325		0.320		0.320		0.313		0.314	

Note: The sample size is $n = 14,382$. Standard errors have been calculated using the delta method. The model evaluation criteria have been calculated as described in Section 4.

Table 5: Empirical application 1: estimated average marginal effects at intensive and extensive margins

Variable	Generalized two-part Marg. eff.	(SE)	Two-part with exclusion Marg. eff.	(SE)	Two-part without exclusion Marg. eff.	(SE)
Intensive margin						
fsize	0.418	(0.105)	0.168	(0.032)	0.168	(0.032)
hc	0.537	(0.024)	0.535	(0.025)	0.517	(0.028)
rd	0.551	(0.062)	0.346	(0.042)	0.355	(0.042)
kl	1.149	(0.321)	1.203	(0.324)	1.304	(0.335)
old	0.008	(0.004)	0.009	(0.004)	0.008	(0.004)
fof	0.107	(0.007)	0.102	(0.006)	0.103	(0.006)
west	0.008	(0.002)	0	(-)	0.014	(0.007)
Extensive margin						
fsize	0.985	(0.343)	0.908	(0.328)	0.908	(0.328)
hc	0.565	(0.041)	0.574	(0.041)	0.574	(0.041)
rd	1.167	(0.176)	1.127	(0.170)	1.127	(0.170)
kl	-0.398	(0.451)	-0.421	(0.441)	-0.421	(0.441)
old	0.059	(0.006)	0.060	(0.006)	0.060	(0.006)
fof	0.069	(0.010)	0.068	(0.010)	0.068	(0.010)
west	0.029	(0.008)	0.029	(0.009)	0.029	(0.009)

Note: The sample size is $n = 14,382$. Standard errors have been calculated using the delta method.

Table 6: Empirical application 2: estimated average marginal effects

Variable	Generalized two-part Marg. eff.	(SE)	Two-part with exclusion Marg. eff.	(SE)	Two-part without exclusion Marg. eff.	(SE)	Fractional probit Marg. eff.	(SE)	OLS Marg. eff.	(SE)
fsize	1.734	(0.124)	1.356	(0.097)	1.357	(0.097)	0.388	(0.056)	0.427	(0.059)
hc	-0.052	(0.020)	-0.019	(0.020)	-0.017	(0.020)	0.036	(0.020)	0.033	(0.020)
rd	0.033	(0.037)	0.051	(0.037)	0.049	(0.037)	0.094	(0.036)	0.089	(0.037)
kl	-0.832	(0.284)	-0.798	(0.277)	-0.807	(0.277)	-0.652	(0.257)	-0.617	(0.225)
fof	-0.015	(0.006)	-0.011	(0.006)	-0.012	(0.006)	-0.004	(0.006)	-0.004	(0.006)
old	0.015	(0.003)	0.011	(0.002)	0.008	(0.004)	0.009	(0.004)	0.009	(0.004)
Model evaluation criteria										
AIC	-3.034		-3.030		-3.030		-3.011		-3.009	
BIC	-3.004		-3.001		-3.000		-2.996		-2.994	
R^2	0.123		0.119		0.119		0.099		0.098	

Note: The sample size is $n = 14,294$. Standard errors have been calculated using the delta method. The model evaluation criteria have been calculated as described in Section 4.

Table 7: Empirical application 2: estimated average marginal effects at intensive and extensive margins

Variable	Generalized two-part Marg. eff.	(SE)	Two-part with exclusion Marg. eff.	(SE)	Two-part without exclusion Marg. eff.	(SE)
Intensive margin						
fsize	0.883	(0.100)	0.248	(0.035)	0.249	(0.035)
hc	-0.082	(0.023)	-0.033	(0.022)	-0.029	(0.023)
rd	0.018	(0.043)	0.045	(0.043)	0.043	(0.043)
kl	-0.849	(0.321)	-0.782	(0.307)	-0.796	(0.307)
fof	-0.009	(0.007)	-0.004	(0.007)	-0.004	(0.007)
old	0.007	(0.002)	0	(-)	-0.006	(0.004)
Extensive margin						
fsize	3.764	(0.293)	3.756	(0.295)	3.756	(0.295)
hc	0.006	(0.042)	0.008	(0.042)	0.008	(0.042)
rd	0.068	(0.080)	0.067	(0.080)	0.067	(0.080)
kl	-0.899	(0.473)	-0.898	(0.475)	-0.898	(0.475)
fof	-0.029	(0.012)	-0.029	(0.012)	-0.029	(0.012)
old	0.034	(0.008)	0.036	(0.008)	0.036	(0.008)

Note: The sample size is $n = 14,294$. Standard errors have been calculated using the delta method.

Table 8: Empirical application 1: descriptive statistics

	Mean	SD
exp	0.802	0.399
expshare	0.265	0.267
fsize	0.026	0.171
hc	0.332	0.118
rd	0.025	0.058
kl	0.006	0.009
old	0.530	0.499
fof	0.139	0.346
west	0.824	0.381

Note: The number of observations is $n = 14,382$. $fsize$ is measured in 10,000's, hc in 100,000's and kl in millions.

Table 9: Empirical application 2: descriptive statistics

	Mean	SD
multi	0.650	0.477
degree	0.215	0.233
fsize	0.026	0.172
hc	0.332	0.117
rd	0.025	0.058
kl	0.006	0.009
old	0.531	0.499
fof	0.139	0.346

Note: The number of observations is $n = 14,294$. $fsize$ is measured in 10,000's, hc in 100,000's and kl in millions.