Schwiebert, Jörg

Conference Paper

A Sample Selection Model for Fractional Response Variables

Beiträge zur Jahrestagung des Vereins für Socialpolitik 2016: Demographischer Wandel - Session: Microeconometrics, No. G01-V1

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

Suggested Citation: Schwiebert, Jörg (2016) : A Sample Selection Model for Fractional Response Variables, Beiträge zur Jahrestagung des Vereins für Socialpolitik 2016: Demographischer Wandel - Session: Microeconometrics, No. G01-V1, ZBW - Deutsche Zentralbibliothek für Wirtschaftswissenschaften, Leibniz-Informationszentrum Wirtschaft, Kiel und Hamburg

This Version is available at:
https://hdl.handle.net/10419/145527

A Sample Selection Model for Fractional Response Variables

Abstract

This paper develops a sample selection model for fractional response variables, i.e., variables taking values between zero and one. It is shown that the proposed model is consistent with the nature of the fractional response variable, i.e., it generates predictions between zero and one. A simulation study shows that the model performs well in finite samples and that competing models, the Heckman selection model and the fractional probit model (without selectivity), generate biased estimates. An empirical application to the impact of education on women's perceived probability of job loss illustrates that the choice of an appropriate model is important in practice. In particular, the Heckman selection model and the fractional probit model are found to underestimate (in absolute terms) the impact of education on the perceived probability of job loss.

Keywords: Fractional probit model, Fractional response variable, Sample selection bias, Sample selection model

JEL codes: C24, C25

1 Introduction

Since Heckman's (1979) seminal paper on the sample selection bias problem, the issue of non-random sample selectivity is well-known to economists. In particular, it is known that estimates are generally biased when the estimation sample is a non-random sample from the overall population. In this paper I develop a sample selection model for fractional response (dependent) variables. This study extends a recent paper by Schwiebert and Wagner (2015), who proposed a generalized two-part model for fractional response variable with excess zeros. The model developed in Schwiebert and Wagner (2015) is conceptually similar to the model considered in this paper, but this paper focuses on the issue of sample selectivity in the spirit of Heckman (1979) while Schwiebert and Wager (2015) focus on two-part modeling issues.

Fractional response variables are variables which take values between zero and one. Such variables include true fractions, e.g. the share of exports in total sales (Wagner, 2001), but also other variables bounded between zero and one, e.g. the perceived probability that a certain event (like job loss) occurs.

In case of continuous dependent variables, the Heckman sample selection model can be used to correct for the potential non-random sample selectivity. Heckman (1979) showed that augmenting the linear regression equation with an inverse Mills ratio term and estimating this augmented regression equation by ordinary least squares (OLS) yields unbiased estimates of the parameters of interest. One might raise the question why the Heckman selection model cannot also be used in case of fractional response (dependent) variables. The reason is that the Heckman selection model assumes an underlying linear relationship between the dependent variable and the explanatory variables. The linearity assumption is however not appropriate when the dependent variable is a fractional response variable, since a linear model generates predictions which might fall outside the range of the fractional response variable, i.e., the $[0, 1]$-interval.

To illustrate this issue, assume for the moment that there is no sample selectivity. Suppose that y denotes the fractional response variable, x is a vector of explanatory variables and β a vector of corresponding parameters. In a linear model, it is assumed

that the mean of y conditional on x is given by

$$E[y|x] = x'\beta. \tag{1}$$

After estimating this model, say by OLS, the predictions generated by the model are characterized by $\hat{y} = x'\hat{\beta}$, where $\hat{y}$ is the prediction of the dependent variable y and $\hat{\beta}$ is the OLS estimate of β. When y is a fractional response variable, however, it cannot be guaranteed that these predictions are bounded between zero and one, which should be the case for a fractional response variable. In other words, a linear model is not consistent with the fractional nature of the dependent variable. The same critique applies to the Heckman selection model, which assumes a linear underlying population model.

Papke and Wooldridge (1996) developed a model which is suitable for fractional response variables. They however did not consider sample selectivity issues. Their idea is to specify the conditional mean of y given x as follows:

$$E[y|x] = G(x'\beta), \tag{2}$$

where $G(\cdot)$ is a cumulative distribution function (cdf). Since a cdf is bounded between zero and one, this model generates predictions which are consistent with the nature of the fractional response variable. When the cdf is the logistic cdf, the model is known as the fractional logit model; when the cdf is the standard normal cdf, it is known as the fractional probit model. However, as in case of the linear model both fractional logit and fractional probit models will give biased estimates when non-random sample selectivity is an issue.

The sample selection model for fractional response variables developed in this paper combines elements from the Heckman selection model and the fractional probit model. In particular, the model incorporates potential non-random sample selectivity in a way which is consistent with the fractional nature of the dependent variable. It is shown that the assumptions made on the impact of selection on the fractional response variable imply that the underlying population model is a fractional probit model, which is consistent with

the nature of the fractional response variable. Moreover, the assumptions imply that when there is no non-random sample selectivity, the model also reduces to the fractional probit model. The point is that no matter whether there is a sample selection bias problem or not, the model will always be consistent with the fractional nature of the dependent variable.

A simulation study is provided which illustrates that the proposed model yields estimates being different from the Heckman selection model, indicating that the assumption of a linear model characterizing the underlying overall population is not appropriate in case of fractional response variables. This paper also contains an empirical application to study the impact of education on women's perceived probability of job loss. As indicated above, the perceived probability that a certain event – like job loss – occurs, can also be interpreted as a fractional response variable, although the term "fractional" might be misleading. Since the perceived probability of job loss is only observed for women who are working, the observed sample can be considered a non-random sample from the overall population of women. Thus, a sample selection model for fractional response variables appears to be an appropriate modeling device.

The fractional probit or logit model has been introduced by Papke and Wooldridge (1996) and has been extended to panel data by Papke and Wooldridge (2008). Wooldridge (2010) describes how to estimate a fractional response model in the presence of endogenous explanatory variables. A survey on fractional response models is provided by Ramalho et al. (2011). Ramalho et al. (2011) also consider two-part models for cases when there is a large portion of observations located at the bounds of the fractional response variable. The fractional response model has often been used in empirical applications; examples include Papke and Wooldridge (1996; 2008), Wagner (2001), Ramalho et al. (2011) and Gallani, Krishnan and Wooldridge (2015).

The paper is organized as follows. Section 2 proposes the sample selection model for fractional response variables and discusses issues of specification, estimation and inference. Section 3 provides simulation evidence on the finite sample properties of the estimator of the model and shows how this model performs in comparison with competing models.

Section 4 contains the empirical application of the model to real data. Finally, Section 5 concludes the paper.

2 Econometric Model

For explanatory purposes, I will first show how the issue of sample selectivity is dealt with in the Heckman (1979) selection model. As in the introduction, let y denote the dependent variable and x the vector of covariates with corresponding parameter vector β. Heckman (1979) assumes that the population model is characterized by the following linear relationship:

$$y^* = x'\beta + \varepsilon. \tag{3}$$

Here, y^* denotes a latent dependent variable and ε is the error term. The observed dependent variable y is related to y^* as follows:

$$y = \begin{cases} y^* \text{ if } z = 1 \\ \text{“missing” otherwise} \end{cases} . \tag{4}$$

The variable z is a selection indicator; if $z = 1$, y is observed and otherwise y is not observed (“missing”). The selection process determining z is characterized by the following selection equation:

$$z = 1[w'\gamma + u > 0], \tag{5}$$

where w is a vector of covariates and γ an associated vector of parameters; u is the error term. Non-random sample selection occurs when ε and u are dependent. The crucial assumption of the Heckman selection model is

$$\begin{pmatrix} \varepsilon \\ u \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2 & \rho\sigma \\ \rho\sigma & 1 \end{pmatrix} \right), \tag{6}$$

i.e., that ε and u are jointly normally distributed with correlation ρ. If the correlation coefficient ρ is different from zero, an ordinary least squares estimation of the equation

$$y = x'\beta + \varepsilon \tag{7}$$

will give inconsistent estimates of β. However, Heckman (1979) showed that – based on the bivariate normality assumption on the error terms – OLS estimation of the augmented regression equation

$$y = x'\beta + \beta_\lambda \lambda(w'\gamma) + v, \tag{8}$$

yields consistent estimates of β. Here, $\beta_\lambda \equiv \rho\sigma$ is a parameter, $\lambda(w'\gamma) \equiv \phi(w'\gamma)/\Phi(w'\gamma)$ is the inverse Mills ratio term and v denotes the error term of the augmented regression equation; moreover, $\phi(\cdot)$ is the standard normal probability density function (pdf) and $\Phi(\cdot)$ is the standard normal cdf.

Unfortunately, in case of fractional response variables it is not possible to derive a sample selection model from an underlying population model, at least not in the way Heckman (1979) proceeded. Since the dependent variable y is only observed if the selection indicator z is equal to one, estimation of a sample selection model requires a specification of the conditional mean $E[y|x, w, z = 1]$. In the Heckman selection model,

$$E[y|x, w, z = 1] = x'\beta + \beta_\lambda \lambda(w'\gamma) \tag{9}$$

follows from the linear specification of the underlying population model and the bivariate normality assumption on the error terms.

To develop a sample selection model for fractional response variables, I proceed in the opposite way. Instead of specifying an underlying population model and imposing an assumption on the joint distribution of error terms, I directly specify the conditional mean $E[y|x, w, z = 1]$. The selection process is assumed to be described by the same selection equation as in the Heckman selection model. In particular, I make the following

assumption:

Assumption 1:

(a) $E[y|x, w, z = 1] = \frac{\Phi_2(x'\beta, w'\gamma, \rho)}{\Phi(w'\gamma)}$, *where* $\Phi_2(\cdot, \cdot, \rho)$ *denotes the bivariate standard normal distribution with correlation* ρ*;*

(b) $Pr(z = 1|w) = \Phi(w'\gamma)$.

Here, ρ is different from the correlation coefficient defined in the Heckman selection model, but it is also an indicator of non-random sample selectivity. Note that Assumption 1 (b) implies that the selection equation is of the probit type, as in the Heckman selection model.

While Assumption 1 is sufficient to estimate the model, it is not clear what is actually being estimated – i.e., what is the interpretation of β? The interpretation of γ is quite straightforward, it contains the coefficients from the selection equation. A sensible interpretation of β however requires an additional assumption on the conditional mean of y given $z = 0$:

Assumption 2: $E[y|x, w, z = 0] = \frac{\Phi_2(x'\beta, -w'\gamma, -\rho)}{\Phi(-w'\gamma)}$.

With this additional assumption it follows that:

Theorem 1: *Suppose that Assumption 1 and 2 hold. Then, the underlying population model is characterized by* $E[y|x, w] = \Phi(x'\beta)$.

Proof:

$$\begin{aligned} E[y|x, w] &= Pr(z = 0|w) \cdot E[y|x, w, z = 0] + Pr(z = 1|w) \cdot E[y|x, w, z = 1] \\ &= \Phi_2(x'\beta, -w'\gamma, -\rho) + \Phi_2(x'\beta, w'\gamma, \rho) \\ &= \Phi(x'\beta). \end{aligned}$$

□

Hence, the population model is a fractional probit model and β are the corresponding coefficients associated with this model. Since the fractional probit model is consistent with the nature of the fractional response dependent variable, the parameter vector β has a clear and economically appealing interpretation.

The specification of the conditional means in Assumptions 1 (a) and 2 may still seem to be arbitrary. The specification has several advantages, however, which makes it economically and statistically appealing. First, both conditional means are very similar to conditional probabilities; indeed, it can be shown that both conditional means are bounded between zero and one, so that they are consistent with the nature of the fractional response dependent variable. Second, when the correlation parameter ρ is equal to zero, both conditional means reduce to $\Phi(x'\beta)$. Hence, in case of no non-random sample selectivity, both conditional means reduce to the fractional probit model of the underlying population, as it should be. This is also true for the Heckman selection model, since when $\rho = 0$ it follows that $\beta_\lambda = 0$ and so the inverse Mills ratio term cancels out. Third, as shown above, the specification identifies a well-defined underlying population model – the fractional probit model – which is consistent with the nature of the fractional response dependent variable.

The specific form of the conditional mean $E[y|x, w, z = 1]$ can also be motivated as follows. If y was a binary variable and related to $x'\beta$ and ε as $y = 1[x'\beta + \varepsilon > 0]$, and if ε and u were jointly normally distributed with unit variances and correlation ρ, the conditional probability that $y = 1$ given $z = 1$ would exactly equal $E[y|x, w, z = 1]$ as defined in Assumption 1. This parallels the motivation of the fractional probit model, since in that model the conditional mean $E[y|x]$ could also be interpreted as the conditional probability that $y = 1$ if y was a binary variable.

After developing the sample selection model for fractional response variables, I now discuss issues regarding specification, estimation and inference. The first issue concerns the variables in x and w. As in the Heckman selection model, an exclusion restriction is required for proper identification, i.e., w should contain at least one variable which is not included in x. Put differently, the variable excluded from x should affect the dependent variable y only indirectly via the selection process, but should not have a direct impact on y. Actually the model proposed above *is* identified by the functional form assumptions on the conditional means. However, imposing an exclusion restriction is highly recommended since in empirical practice it is often not convincing to identify parameters from functional

form assumptions alone.

The proposed model can be estimated by quasi maximum likelihood (QML); see Gourieroux, Monfort and Trognon (1984) for a general treatment and Papke and Wooldridge (1996) for a description in the context of fractional response models. The log-likelihood function of the sample selection model for fractional response variables is given by

$$\log L(\theta) = \sum_{i=1}^{n} l_i(\theta) \equiv \sum_{i=1}^{n} \{(1-z_i)\log(1-\Phi(w_i'\gamma)) + z_i \log \Phi(w_i'\gamma) + z_i \left[(1-y_i)\log\left(1 - \frac{\Phi_2(x_i'\beta, w_i'\gamma; \rho)}{\Phi(w_i'\gamma)}\right) + y_i \log \frac{\Phi_2(x_i'\beta, w_i'\gamma; \rho)}{\Phi(w_i'\gamma)} \right] \}, \quad (10)$$

where $\theta = (\beta', \gamma', \rho)'$ denotes the parameter vector to be estimated, i indexes individuals and n is the sample size. An advantage of QML is that only the conditional means given above – and not the full distribution – have to be correctly specified in order to obtain consistent estimates of the model parameters. The QML estimator $\hat{\theta}$ has an asymptotic normal distribution and its estimated asymptotic variance matrix is of the sandwich-type (White, 1982), i.e.

$$Est.Asy.Var.(\hat{\theta}) = \left(-\sum_{i=1}^{n} \frac{\partial^2 l_i(\hat{\theta})}{\partial\theta\partial\theta'} \right)^{-1} \left(\sum_{i=1}^{n} \frac{\partial l_i(\hat{\theta})}{\partial\theta} \frac{\partial l_i(\hat{\theta})}{\partial\theta'} \right) \left(-\sum_{i=1}^{n} \frac{\partial^2 l_i(\hat{\theta})}{\partial\theta\partial\theta'} \right)^{-1}. \quad (11)$$

Standard errors of the estimated parameters can be derived from this matrix in the usual way.

As described above, a correlation parameter of $\rho = 0$ means that there is no non-random sample selection. In that case, the sample selection model for fractional response variables reduces to the fractional probit model. In case of QML estimation, a Wald test is the easiest way to test for the absence of non-random sample selectivity. The null hypothesis is $\rho = 0$, and testing amounts to a simple significance test of ρ.

3 Simulation Evidence

This section contains simulation evidence on the finite sample properties of the QML estimator of the sample selection model for fractional response variables. It also provides

evidence on the bias of estimates from the Heckman selection model and the fractional probit model when the sample selection model for fractional response variables is the true data generation model.

The simulated data are generated as follows. The selection equation is assumed to be

$$z_i = 1(\gamma_0 + \gamma_1 w_{i1} + \gamma_2 w_{i2} + u_i > 0), \tag{12}$$

$i = 1, \ldots, n$, where the u_i's are i.i.d. draws from a standard normal distribution. The explanatory variables w_{i1} and w_{i2} are generated according to

$$w_{i1} = v_i + \eta_{1i} \tag{13}$$

$$w_{i2} = v_i + \eta_{2i}, \tag{14}$$

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

The next step is to generate the fractional response variable y. Conceptually, I proceed as described in the last section, by specifying the conditional mean $E[y|x, w, z = 1]$ as follows:

$$E[y_i|x_i, w_i, z_i = 1] = \frac{\Phi_2(\beta_0 + \beta_1 x_i, \gamma_0 + \gamma_1 w_{i1} + \gamma_2 w_{i2}; \rho)}{\Phi(\gamma_0 + \gamma_1 w_{i1} + \gamma_2 w_{i2})}, \tag{15}$$

where $x_i = w_{i1}$ for all $i = 1, \ldots, n$. Note that w_{i2} affects the conditional mean only through the selection process. This is the exclusion restriction needed for proper identification of the parameters.

Based on the specification of the conditional mean, the fractional response variable y can be generated. It is convenient to use the beta distribution for the generation of y, since draws from a beta distribution are bounded between zero and one, which is required for a fractional response variable. A further advantage of the beta distribution is that it can be parameterized in terms of its mean, so that the assumption on the conditional

mean given above can be implemented quite easily. The beta distribution parameterized in terms of its mean is given by

$$f(y;\mu,\psi) = \frac{\Gamma(\psi)}{\Gamma(\mu\psi)\Gamma((1-\mu)\psi)} y^{\mu\psi-1}(1-y)^{(1-\mu)\psi-1}, \tag{16}$$

where μ denotes the mean, ψ is a shape parameter and $\Gamma(\cdot)$ is the gamma function (see Ramalho et al., 2011, p. 25). Thus, y can be generated according to the rule

$$y_i \begin{cases} \text{“is missing”} & \text{if } z_i = 0 \\ \sim f(y_i; \frac{\Phi_2(\beta_0+\beta_1 x_i, \gamma_0+\gamma_1 w_{i1}+\gamma_2 w_{i2};\rho)}{\Phi(\gamma_0+\gamma_1 w_{i1}+\gamma_2 w_{i2})}, \psi) & \text{if } z_i = 1 \end{cases} . \tag{17}$$

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, 0.3 and 0.7 in order to analyze the estimator performance for different degrees of dependence.

Sample sizes of 500, 1,000 and 2,000 are considered. Each simulation comprises 1,000 repetitions. Over these repetitions, the mean of the parameter estimates and the associated root mean squared error (RMSE) are calculated.

As mentioned above, also the Heckman selection model and the fractional probit model are used to generate estimates. The purpose is to show evidence on the bias generated by models which do not properly account for the fractional nature of the dependent variable (the Heckman selection model) or the potential non-random sample selectivity (the fractional probit model). The Heckman selection model assumes that the underlying population model is given by

$$y_i^* = \beta_0 + \beta_1 x_i + \varepsilon_i, \tag{18}$$

while the frational probit model assumes that

$$E[y_i|x_i] = E[y_i|x_i, z_i = 1] = \Phi(\beta_0 + \beta_1 x_i). \tag{19}$$

Note that both assumptions are wrong with respect to the data generation process.

Since the estimated parameters in β are not comparable across models, the simulation results also include evidence on the marginal effects of an increase in x on the dependent variable y in the underlying population model. In case of the Heckman selection model, the underlying population model is linear, hence the marginal effect is simply the coefficient of x, β_1; in case of the sample selection model for fractional response variables and the fractional probit model, the underlying population model is characterized by

$$E[y_i|x_i, w_{i1}, w_{i2}] = \Phi(\beta_0 + \beta_1 x_i). \tag{20}$$

Thus, the marginal effect of an increase in x is

$$ME_i = \frac{\partial E[y_i|x_i, w_{i1}, w_{i2}]}{\partial x} = \frac{\partial \Phi(\beta_0 + \beta_1 x_i)}{\partial x} \tag{21}$$

for a given observation i. The average marginal effect is then given by

$$ME = \frac{1}{n} \sum_{i=1}^{n} ME_i. \tag{22}$$

This average marginal effect is also calculated in each repetition and for each model. Over the repetitions, the mean of the average marginal effect is calculated as well as the corresponding standard deviation.

The estimates for the Heckman selection model are generated using the two-step estimation approach. That is, in a first step the selection equation is estimated by probit and these estimates are used to calculate the inverse Mills ratio term. In a second step, the regression equation is augmented by the inverse Mills ratio term and estimated by OLS. Alternatively, the full model could have been estimated in one step by applying the maximum likelihood method. However, the two-step estimator requires fewer assumptions than the maximum likelihood estimator. In particular, the error terms ε and u need not be bivariate normally distributed, but it suffices that $E[\varepsilon|u] = \delta u$ for some fixed parameter δ and $u \sim \mathcal{N}(0, 1)$ (see Wooldridge, 2010, p. 803). Due to these less restrictive assumptions, the two-step approach seems to be more appropriate when the dependent

variable is a fractional response variable rather than a continuous variable.

The simulation results are given in Tables 1-3. Table 1 contains the results for $\rho = 0$, Table 2 for $\rho = 0.3$ and Table 3 for $\rho = 0.7$. I begin with the finite sample properties of the QML estimator of the sample selection model for fractional response variables. As can be seen from Tables 1-3, the model parameters are estimated well for all sample sizes, irrespective of the degree of dependence. Moreover, as expected, the RMSE's decline with increasing sample size. Also, the marginal effects are virtually identical for different degrees of dependence, which would have been expected since the underlying population model does not depend on the degree of dependence.

Now I investigate what happens if the Heckman selection model or the fractional probit model, respectively, are used for estimation. Tables 1-3 show that the estimates of the selection equation parameters are estimated well when the Heckman selection model is used. This is no coincidence, since in the Heckman selection model the selection equation is assumed to be of the probit type, which is indeed true. The remaining parameters are biased, but they are not really comparable with those from the sample selection model for fractional response variables. However, the marginal effects are comparable. Tables 1-3 show that the marginal effects obtained from the Heckman model substantially differ from the marginal effects derived from the true model, the sample selection model for fractional response variables. Moreover, the difference becomes larger when the degree of dependence increases. These simulation results thus indicate that the results from the Heckman selection model are biased when the true underlying model is the sample selection model for fractional response variables. This suggests that it is important to properly account for the fractional nature of the dependent variable.

Considering the results from the fractional probit model, Table 1 shows that the parameter estimates and marginal effects are identical to those from the sample selection model for fractional response variables. Note that no estimates are given for the selection equation parameters, since in case of the fractional probit model it is assumed that there is no non-random sample selectivity. Since there is indeed no non-random sample selectivity when $\rho = 0$ (Table 1), it is not surprising that the fractional probit model yields the same

estimates as the sample selection model for fractional response variables. However, as dependence increases, the parameter estimates become different and also the marginal effects begin to differ. This indicates that that the results from the fractional probit model are biased when non-random sample selectivity is an issue.

In summary, the simulation results show that the QML estimator of the sample selection model for fractional response variables performs well in finite samples. Furthermore, the results also show that an application of the Heckman selection model, which does not account for the fractional nature of the dependent variable, or the fractional probit model, which does not account for the potential non-random sample selectivity, leads to biased estimation results, particularly to biased marginal effects. This suggests that it is important in practice to use a sample selection model for fractional response variables when (i) the dependent variable is a fractional response variable and (ii) non-random sample selectivity is an issue.

4 Empirical Application

This section contains an empirical application of the proposed sample selection model for fractional response variables to real data. Specifically, I consider the impact of education on the perceived (subjective) probability of job loss. As described by Manski and Straub (2000), job loss is "commonly assumed to be unanticipated by the worker and unaffected by worker behavior on the job; the result of plant closings, elimination of positions, and the like" (Manski and Straub, 2000, p. 467), and can therefore be interpreted as exogenous job destruction (Manski and Straub, 2000, p. 467). I use data from the 2001 wave of the German Socioeconomic Panel (SOEP). Respondents were asked how likely it was that they lost their job within the next two years. Answers could be made in decimal steps, i.e., 0%, 10%, 20%,..., 100%. Since the answers are bounded between 0% (=0) and 100% (=1), the perceived probability of job loss is a fractional response variable.

Job loss leads to substantial pecuniary and non-pecuniary costs; see, e.g., Winkelmann and Winkelmann (1998) and the references cited therein. Winkelmann and Winkelmann (1998) also used SOEP data and found a large negative effect of unemployment on indi-

vidual well-being. It can be expected that also a high perceived probability of job loss has a similar (negative) effect on individual well-being.

Education typically raises the individual amount of human capital and thus increases the employee's value to the employer. Therefore, I expect that education reduces not only the actual but also the perceived probability of job loss, since employees know their value to some extent. If education decreases the perceived probability of job loss, education may be interpreted as some kind of insurance against the non-pecuniary costs associated with job insecurity. Since the non-pecuniary costs of unemployment are quite substantial (Winkelmann and Winkelmann, 1998), it is highly interesting from an economic point of view to investigate if education leads to a lower perceived probability of job loss and thus reduces these costs.

In this application I analyze the impact of education on the perceived probability of job loss for women only. In my sample about one third of women are not working. Since the perceived probability of job loss is reported only by women who are working, a regression of the perceived probability of job loss on education (and further covariates) for those women may lead to a sample selection bias. Hence, a sample selection model should be used. Due to the fractional nature of the dependent variable, the sample selection model for fractional response variables developed in this paper seems to be an appropriate modeling device. I compare the estimates from this model with the estimates from the Heckman selection model and the fractional probit model to investigate to what extent the models lead to different estimates.

The underlying population model has the perceived probability of job loss as the dependent variable. Explanatory variables are (years of) education, age, age squared, dummies for the state of residence, a dummy for foreign nationality, dummies for marital status and the number of children. Age and age squared capture age-specific differences in job loss probabilities, while the state dummies reflect state-specific labor market conditions. People with foreign nationality may have different perceptions of job security than German people and/or may face different labor market opportunities than German people. Marital status and the number of children may affect the employer's decision to

lay people off in light of socially minded reasons, and the employee might know this.

Since non-random sample selectivity might be an issue, the next step is to set up a selection equation which governs the probability that a woman is working. Explanatory variables assumed to affect the selection process are the same covariates that appear in the main equation, and an additional variable which is needed because of the exclusion restriction. As described above, this variable should affect the perceived probability of job loss only indirectly via the selection process, but should otherwise not have a direct impact on the perceived probability of job loss. A variable which can be argued to satisfy these requirements is the total household income minus the wage income of the woman. I call this variable "additional income". Additional income can be expected to have an impact on the selection process: the higher the additional income, the lower the material incentive to work. Furthermore, additional income is rather private information and typically not available to the employer; thus, it should not affect the employer's layoff decisions. Hence, additional income should not affect the actual and perceived probability of job loss directly, but only indirectly via the selection process.

My sample includes women in their prime working age, i.e., between 25 and 54 years of age, who are not self-employed. Self-employed workers were excluded because it is difficult to distinguish between voluntary quits and job losses in case of self-employed workers (see Manski and Straub, 2000, p. 467). Summary statistics of the variables are given in Table 4.

As mentioned above, estimates from three different models will be analyzed: the sample selection model for fractional response variables developed in this paper, the Heckman selection model (two-step estimation) and the fractional probit model. While the Heckman selection model does not account for the fractional nature of the dependent variable, the fractional probit model ignores the potential non-random sample selectivity. Since the model parameters are not comparable and the focus of this application is on the impact of education, I also computed the estimated marginal effect of education on the perceived probability of job loss for all three models. As in the simulation study, this marginal effect is the marginal effect of education in the underlying population model.

The estimation results are given in Table 5. Table 5 includes the estimated parameters of each model as well as the estimate of the correlation parameter ρ in case of the selection models. Moreover, the marginal effect of education in the underlying population model is reported. As described, this marginal effect is comparable across models. No estimates for the state dummies are reported due to brevity.

The standard error of estimated ρ from the Heckman selection model has been obtained by bootstrapping. The reason is that not ρ itself is estimated but the coefficient β_λ of the inverse Mills ratio term. After estimation it is possible to derive an estimate of σ, which can be used to calculate estimated ρ because $\beta_\lambda = \rho\sigma$. Since estimated σ is obtained *after* estimation, the standard error of estimated ρ cannot be obtained simply from an application of the delta method. Therefore, I chose bootstrapping to obtain the standard error. The value reported in Table 5 is based on 1,000 bootstrap iterations.

Table 5 shows estimates for the parameters of the main and selection equation. In case of the fractional probit model, there is no selection equation, hence no results are reported. Since the selection equation for both the sample selection model for fractional response variables and the Heckman selection model are of the probit type, it is no coincidence that the estimates of the selection equation parameters are very close. The "additional income" variable, which has been excluded from the underlying population model, has the expected negative impact on the selection process. Also note that the selection model for fractional response variables and the Heckman selection model both yield a quite similar correlation coefficient of about 0.60. Both estimates are significantly different from zero, which indicates that non-random sample selectivity is indeed an issue.

The estimated marginal effect of education varies over the models, but is generally negative, as expected. The largest value (in absolute terms) is obtained from the sample selection model for fractional response variables, and the lowest value from the fractional probit model. The marginal effect from the Heckman selection model is in between. The differences illustrate that the choice of an appropriate model is important in practice. In particular, the results suggest that models which do not account for the fractional nature of the dependent variable (the Heckman model) or do not account for the non-random

sample selectivity (the fractional probit model) underestimate (in absolute terms) the impact of education on the perceived probability of job loss, at least in this data example.

5 Conclusions

This paper developed a sample selection model for fractional response variables. The model was shown to be consistent with the nature of the fractional response variable, i.e., the model generates predictions between zero and one. Simulation evidence demonstrated that estimation of this model works well in finite samples, and that competing estimators which do either not account for the fractional nature of the dependent variable (the Heckman selection model) or do not account for potential non-random selectivity (the fractional probit model) lead to biased estimates. An empirical application to the impact of education on women's perceived probability of job loss illustrated that it is important in practice to choose an appropriate model. In particular, the Heckman selection model and the fractional probit model seemed to underestimate (in absolute terms) the marginal effect of an increase in education on women's perceived probability of job loss.

The challenge associated with an application of the model in applied research is to find an appropriate exclusion restriction. However, given that such an exclusion restriction is available, the model developed in this paper provides a useful device to correct for potential sample selection bias when the dependent variable is a fractional response variable. Since non-random sample selectivity is an issue frequently encountered in empirical research, there appear to be many potential applications of this model in practice.

References

Gallani, S., Krishnan, R. and Wooldridge, J.M. (2015). Applications of Fractional Response Model to the Study of Bounded Dependent Variables in Accounting Research. Harvard Business School Accounting & Management Unit Working Paper No. 16-016. Available at SSRN: http://ssrn.com/abstract=2642854
or http://dx.doi.org/10.2139/ssrn.2642854.

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

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

Manski, C.F. and Straub, J.D. (2000). Worker perceptions of job insecurity in the mid-1990s: evidence from the survey of economic expectations. *The Journal of Human Resources* 35, 447-479.

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. (2011). Alternative estimating and testing empirical strategies for fractional regression models. *Journal of Economic Surveys* 25, 19-68.

Schwiebert, J. and Wagner, J. (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.

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

Wooldridge, J.M. (2010). *Econometric analysis of cross section and panel data*. 2. ed. MIT Press, Cambridge, Mass.

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

Winkelmann, L. and Winkelmann, R. (1998). Why are the unemployed so unhappy? Evidence from panel data. *Economica* 65, 1-15.

Tables

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

	Sel. model f. frac. resp. var.		Heckman sel. model		Frac. probit model	
	Mean	RMSE/SD	Mean	RMSE/SD	Mean	RMSE/SD
			n=500			
Parameters						
β_0	-1.001	0.057	0.158	1.158	-1.003	0.039
β_1	0.501	0.033	0.165	0.335	0.502	0.027
γ_0	0.002	0.088	0.002	0.088		
γ_1	1.019	0.109	1.020	0.109		
γ_2	1.014	0.113	1.014	0.113		
ρ	-0.003	0.084	0.167	0.234		
Marginal effect of x	0.118	0.005	0.165	0.010	0.118	0.005
			n=1,000			
Parameters						
β_0	-0.998	0.040	0.159	1.159	-0.999	0.027
β_1	0.499	0.023	0.164	0.336	0.499	0.019
γ_0	0.000	0.063	0.000	0.063		
γ_1	1.013	0.077	1.013	0.077		
γ_2	1.006	0.076	1.006	0.076		
ρ	-0.002	0.060	0.168	0.203		
Marginal effect of x	0.118	0.003	0.164	0.007	0.118	0.003
			n=2,000			
Parameters						
β_0	-0.999	0.028	0.158	1.159	-1.000	0.020
β_1	0.500	0.016	0.164	0.336	0.500	0.013
γ_0	0.001	0.044	0.001	0.044		
γ_1	1.004	0.054	1.004	0.054		
γ_2	1.007	0.053	1.007	0.053		
ρ	-0.002	0.043	0.167	0.187		
Marginal effect of x	0.118	0.002	0.164	0.005	0.118	0.002

Note: The root mean squared errors (RMSE) refer to the parameters, while the standard deviations (SD) refer to the marginal effect of x. 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 $\rho = 0.3$

	Sel. model f. frac. resp. var.		Heckman sel. model		Frac. probit model	
	Mean	RMSE/SD	Mean	RMSE/SD	Mean	RMSE/SD
			n=500			
Parameters						
β_0	-1.000	0.051	0.152	1.152	-0.865	0.141
β_1	0.501	0.031	0.170	0.330	0.440	0.065
γ_0	0.001	0.088	0.001	0.088		
γ_1	1.024	0.115	1.024	0.116		
γ_2	1.016	0.113	1.016	0.113		
ρ	0.300	0.086	0.681	0.409		
Marginal effect of x	0.117	0.005	0.170	0.010	0.114	0.005
			n=1,000			
Parameters						
β_0	-1.001	0.03738778	0.152	1.152	-0.864	0.139
β_1	0.501	0.02172999	0.170	0.330	0.440	0.063
γ_0	0.000	0.06235902	0.000	0.062		
γ_1	1.007	0.07473817	1.007	0.075		
γ_2	1.007	0.07608353	1.007	0.076		
ρ	0.302	0.06023886	0.683	0.398		
Marginal effect of x	0.117	0.003	0.170	0.007	0.114	0.003
			n=2,000			
Parameters						
β_0	-0.999	0.02553724	0.153	1.153	-0.863	0.138
β_1	0.499	0.01549113	0.170	0.330	0.439	0.062
γ_0	0.002	0.0426022	0.002	0.043		
γ_1	1.005	0.05509056	1.006	0.055		
γ_2	1.005	0.05406413	1.005	0.054		
ρ	0.298	0.04441957	0.678	0.387		
Marginal effect of x	0.117	0.002	0.170	0.005	0.114	0.002

Note: The root mean squared errors (RMSE) refer to the parameters, while the standard deviations (SD) refer to the marginal effect of x. 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 $\rho = 0.7$

	Sel. model f. frac. resp. var.		Heckman sel. model		Frac. probit model	
	Mean	RMSE/SD	Mean	RMSE/SD	Mean	RMSE/SD
			n=500			
Parameters						
β_0	-0.997	0.045	0.127	1.127	-0.696	0.307
β_1	0.499	0.027	0.184	0.316	0.363	0.140
γ_0	-0.003	0.089	-0.002	0.089		
γ_1	1.014	0.106	1.015	0.111		
γ_2	1.019	0.109	1.019	0.110		
ρ	0.693	0.081	1.000	0.300		
Marginal effect of x	0.118	0.005	0.184	0.011	0.107	0.006
			n=1,000			
Parameters						
β_0	-0.998	0.033	0.126	1.126	-0.695	0.307
β_1	0.499	0.020	0.184	0.316	0.361	0.141
γ_0	0.003	0.060	0.004	0.061		
γ_1	1.009	0.074	1.010	0.077		
γ_2	1.008	0.076	1.008	0.076		
ρ	0.699	0.057	1.000	0.300		
Marginal effect of x	0.118	0.003	0.184	0.008	0.107	0.004
			n=2,000			
Parameters						
β_0	-0.999	0.023	0.126	1.126	-0.696	0.305
β_1	0.500	0.014	0.184	0.316	0.362	0.139
γ_0	-0.002	0.045	-0.002	0.046		
γ_1	1.003	0.051	1.003	0.054		
γ_2	1.005	0.054	1.005	0.054		
ρ	0.697	0.041	1.000	0.300		
Marginal effect of x	0.118	0.002	0.184	0.006	0.107	0.003

Note: The root mean squared errors (RMSE) refer to the parameters, while the standard deviations (SD) refer to the marginal effect of x. 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: Summary statistics

Variable	Description	Obs	Mean	Std. Dev.
pjobloss	Perceived prob. of job loss	3,733	0.194	0.256
educ	Years of education	5,612	11.957	2.440
age	Age	5,612	39.453	8.096
state	State of residence			
..Schleswig-Holstein	Schleswig-Holstein (0/1; base)	5,612	0.026	0.159
..Hamburg	Hamburg (0/1)	5,612	0.013	0.113
..Lower Saxony	Lower Saxony (0/1)	5,612	0.085	0.279
..Bremen	Bremen (0/1)	5,612	0.007	0.081
..North-Rhine-Westfalia	North-Rhine-Westfalia (0/1)	5,612	0.218	0.413
..Hessen	Hessen (0/1)	5,612	0.069	0.254
..Rheinland-Pfalz	Rheinland-Pfalz (0/1)	5,612	0.050	0.218
..Baden-Wuerttemberg	Baden-Wuerttemberg (0/1)	5,612	0.121	0.327
..Bavaria	Bavaria (0/1)	5,612	0.141	0.348
..Saarland	Saarland (0/1)	5,612	0.016	0.124
..Berlin	Berlin (0/1)	5,612	0.035	0.184
..Brandenburg	Brandenburg (0/1)	5,612	0.039	0.195
..Mecklenburg-Vorpommern	Mecklenburg-Vorpommern (0/1)	5,612	0.023	0.149
..Saxony	Saxony (0/1)	5,612	0.073	0.260
..Saxony-Anhalt	Saxony-Anhalt (0/1)	5,612	0.043	0.202
Thuringia	Thuringia (0/1)	5,612	0.041	0.199
foreign	Foreign nationality (0/1)	5,612	0.108	0.311
marital status	Marital status			
..married (liv. tog.)	Married and living together (0/1; base)	5,612	0.722	0.448
..married (sep.)	Married and separated (0/1)	5,612	0.024	0.152
..single	Single (0/1)	5,612	0.156	0.363
..divorced	Divorced (0/1)	5,612	0.083	0.276
..widowed	Widowed (0/1)	5,612	0.016	0.125
no. children	Number of children	5,612	0.963	1.041
add. inc.	Add. monthly income (divided by 1,000)	5,612	1.789	1.190

Note: The data have been taken from the 2001 wave of the German Socioeconomic Panel (SOEP).

Table 5: Estimation results

Variable (Dep.var.: pjobloss)	Sel. model f. frac. resp. Var. Coef.	(Std. Err.)	Heckman sel. model Coef.	(Std. Err.)	Frac. Probit model Coef.	(Std. Err.)
			Population model			
educ	-0.0292	(0.0073)	-0.0075	(0.0021)	-0.0122	(0.0071)
age	-0.0904	(0.0280)	-0.0207	(0.0082)	0.0120	(0.0216)
age squared	0.0011	(0.0004)	0.0002	(0.0001)	-0.0002	(0.0003)
foreign	0.1073	(0.0637)	0.0230	(0.0175)	-0.0051	(0.0642)
marital status						
..married (sep.)	0.1883	(0.1100)	0.0586	(0.0288)	0.2141	(0.1167)
..single	-0.0103	(0.0506)	0.0015	(0.0142)	0.0589	(0.0497)
..divorced	-0.0294	(0.0559)	-0.0024	(0.0157)	0.0755	(0.0532)
..widowed	-0.0003	(0.1132)	0.0018	(0.0335)	0.0407	(0.1134)
no. children	0.1415	(0.0313)	0.0334	(0.0095)	0.0028	(0.0196)
constant	1.4314	(0.6065)	0.7209	(0.1783)	-0.9895	(0.4340)
			Selection equation			
educ	0.0679	(0.0086)	0.0680	(0.0083)		
age	0.2950	(0.0250)	0.2967	(0.0253)		
age squared	-0.0037	(0.0003)	-0.0037	(0.0003)		
foreign	-0.3085	(0.0608)	-0.3084	(0.0610)		
marital status						
..married (sep.)	-0.1850	(0.1218)	-0.1896	(0.1200)		
..single	0.0855	(0.0655)	0.0848	(0.0648)		
..divorced	0.1135	(0.0743)	0.1164	(0.0725)		
..widowed	-0.0332	(0.1504)	-0.0317	(0.1497)		
no. children	-0.3574	(0.0215)	-0.3592	(0.0219)		
add. inc.	-0.1852	(0.0238)	-0.1854	(0.0166)		
constant	-5.2592	(0.4974)	-5.2790	(0.5032)		
ρ	-0.6102	(0.0911)	-0.6019	(0.1167)		
Mar. eff. of educ	-0.0101	(0.0027)	-0.0075	(0.0021)	-0.0032	(0.0019)
State dummies incl.	Yes		Yes		Yes	
No. obs.	5,612		5,612		3,733	

Note: In case of the Heckman selection model, the standard error of estimated ρ has been obtained by bootstrapping. The coefficients associated with the state dummies are not displayed due to brevity. The marginal effect of educ refers to the marginal effect of education in the underlying population model.