Who Creates Jobs?
Estimating Job Creation Rates at the Firm Level*

Peter Huber*, Harald Oberhofer**, Michael Pfaffermayr***

January 16, 2013

Abstract

This paper shows that applying simple employment-weighted OLS estimation to Davis, Haltiwanger and Schuh (1996) firm level job creation rates taking the values 2 and -2 for entering and exiting firms, respectively, provides biased and inconsistent parameter estimates. Consequently, we argue that entries and exits should be analyzed separately and propose an alternative, consistent estimation procedure assuming that the size of continuing firms follows a lognormal distribution. A small-scale Monte Carlo analysis confirms the analytical results. Using a sample of Austrian firms, we demonstrate that the impact of small firms on net job creation is substantially underestimated when applying employment-weighted OLS estimation.

Keywords: Job creation; DHS growth rate; firm size; firm age; maximum likelihood estimation; three-part model; Monte Carlo simulation.

JEL: C18; C53; D22; E24; L25; L26; M13.

* This paper substantially benefited from discussion with and comments by John Haltiwanger, Ron Jarmin and Javier Miranda. We would also like to thank the participants of the 5th Workshop on Empirical Industrial Organization 2012 at the Austrian Institute of Economic Research in Vienna, of the Workshop on Firm Growth and Innovation 2012 in Tarragona and seminar participants at the University of Innsbruck for their comments and suggestions. Financial support from the 'Oesterreichische Nationalbank' (OeNB, grant numbers 13370 and 14383) is gratefully acknowledged.

*Austrian Institute of Economic Research, Arsenal, Objekt 20, A-1030 Vienna, Austria. Email: huber@wifo.ac.at.

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

***Department of Economics, University of Innsbruck, Universitaetsstr. 15, A-6020 Innsbruck, Austria, Austrian Institute of Economic Research and CESifo, E-mail: Michael.Pfaffermayr@uibk.ac.at.

1 Introduction

During the last three decades the question on which firms are the most important net job creators has triggered heated discussions in both, the academic community and among policy makers. Starting with the early insights provided by Birch (1979) the debate centered around the issue whether small or large firms are more successful in creating jobs. A recent study by Neumark *et al.* (2011) reinforced the crucial role of small firms for net job creation while Haltiwanger, Jarmin and Miranda (2012) (hereafter HJM), applying employment-weighted OLS estimation, highlighted the so far neglected role of firm age. Based on an impressive sample of US firms the latter find that young firms, irrespective of their size, are the most important contributors to job creation.

A careful empirical analysis of job creation across different types of firms (small, large, young and old ones) has to take into account several different sources of net job creation or destruction. By definition, newly founded firms create jobs while exiting firms destroy jobs. Additionally, continuing firms adjust their size and might either increase or decrease their level of employment. Taking these arguments together, when analyzing the determinants of net job creation, one has to simultaneously examine firm entry, firm exit and firm growth to address this issue accurately. In their seminal contributions, Davis and Haltiwanger (1992) (hereafter DH) and Davis, Haltiwanger and Schuh (1996) (hereafter DHS) proposed a new measure of job creation that permits such an integrated treatment of firm exit, firm entry and firm growth (of continuing firms). In particular, the DHS measure (as discussed in more detail below) implies a growth rate of exiting (entering) firms of -2 (2) while continuing firms might exhibit any growth rates in the open interval of -2 and 2. This job creation measure provides a convenient way to (descriptively) calculate the relative importance of firm entry and firm exit for net job creation.

The discontinuities at -2 and 2 of the DHS growth rate distribution might, however, cause econometric problems when applying ordinary least squares (OLS) to firm level data.[1] The literature so far has suggested three different approaches to overcome this problem: First, some studies account for these discontinuities by the inclusion of entry and/or exit dummy variables, respectively, and estimate the resulting model by simple (employment-weighted) OLS (see, e.g., Burgess *et al.* 2000; Faberman 2003; Haltiwanger and Vodopivec 2002, 2003; Voulgaris *et al.* 2005 as well as HJM). Second, some authors argue that the DHS

[1] Moreover, Foote (2006, p. 161) argues that it is unclear how the infinite percentage changes for entering and exiting firms should correctly be incorporated into a single growth rate that allows to analyze net job creation at the aggregate level.

growth rate induces censoring at the interval $[-2, 2]$ and, thus, apply Tobit estimation (see, e.g., Ibsen and Westergaard-Nielsen 2005; Ilmakunnas and Maliranta 2005; Guertzgen 2009). The Tobit model, however, is not appropriate in this case, since it assumes a distribution of firm level job creation rates with support outside the interval $[-2, 2]$. Finally, Baldwin *et al.* (1998), Stiglbauer *et al.* 2003, Armington and Acs (2004), Fuchs and Weyh (2010), HJM or Moscarini and Postel-Vinay (2012) among others, construct cell averages of the DHS growth rate in order to avoid these problems. HJM also demonstrate that this latter approach is equivalent to applying employment-weighted OLS to firm level data.[2]

This paper shows that employment-weighted OLS estimation applied to DHS job creation rates at the firm level provides misleading estimates of conditional means. The reason for this is that for such data with boundary points at -2 and 2 OLS, in general, leads to biased and inconsistent slope estimates which, in turn, also affects the estimates of the employment-weighted conditional means. Consequently, cell-averaged OLS regressions also deliver biased conditional mean estimates. Moreover, we explore the impact of this bias on studies such as Voulgaris *et al.* (2005) and HJM that analyze the contribution of firm size and age to overall job creation in this way.

The bias of the simple OLS estimator does not come as a surprise as the distribution of the generalized DHS growth rate is discontinuous at the boundary values of -2 and 2. Applying (employment-weighted) OLS to a model with the DHS growth rate on the left hand side comprises two sources for a bias: First, there is an approximation bias when viewing the DHS growth rate as a linear approximation of the log change in firm size. Second and more importantly, the boundary values for entering and exiting firms induce a bias stemming from a lack of variation in the DHS job creation rate of these firms on the one hand, but variation in the explanatory variables on the other hand. In addition, this model pools over the groups of entering, continuing and exiting firms assuming the same marginal effect of age and size, respectively. For that reason, we propose an alternative maximum likelihood estimator that treats continuing firms, entrants and exiting firms separately. In line with the firm growth literature (see, e.g., Sutton 1997; Hart 2000 and Coad 2009 for surveys), our approach is based on the assumption of a lognormal firm size distribution for the continuing firms. This three-part procedure allows to consistently estimate the effects of firm size and age for job creation and to aggregate (average) marginal effects for specific groups of firms.

[2]Actually, the parameters of the weighted OLS regression coincide with the cell means in a saturated model that includes all interaction effects as shown by, e.g. Searle (1987, p. 102).

A small-scale Monte Carlo analysis confirms our analytical results, indicating that the alternative ML estimator delivers consistent estimates while the OLS estimator of the HJM approximation is biased and inconsistent. There we also show that this bias carries-over to the estimation of employment-weighted conditional means when applying weighted OLS. Finally, we apply unweighted and weighted OLS as well as the maximum likelihood estimator to a sample of Austrian firms and show that employment-weighted OLS estimation provides unreliable conditional means. This is again due to the bias of the OLS estimator which in quantitative terms is non-negligible. Our estimates indicate that, in total, the unweighted HJM estimator underestimates the overall impact of heterogeneity in firm size and age on job creation by approximately 18% while the weighting scheme proposed by HJM leads to an overestimated overall impact of 5%. For the unweighted case, the bias mainly originates from a severe underestimation of job creation by small firms. Similar to Neumark *et al* (2011), our empirical estimates support the view that in Austria the small rather than young firms are the most important net job creators.

2 The econometrics of job creation rates

DH, DHS and HJM suggest to measure net job creation from period $t-1$ to t of firm i by

$$g_{it} = 2\frac{y_{it}-y_{i,t-1}}{y_{it}+y_{i,t-1}} = \begin{cases} -2 & \text{if } y_{it}=0 \text{ (exit)} \\ 2\frac{y_{it}/y_{i,t-1}-1}{y_{it}/y_{i,t-1}+1} & \text{if } y_{i,t-1}\neq 0 \text{ and } y_{it}\neq 0 \\ 2 & \text{if } y_{i,t-1}=0 \text{ (entry)}, \end{cases}$$

where y_{it} denotes a firm's number of employees in t. The main advantage of this measure is that (net) job creation rates are defined for *all* observations, i.e., also for entries ($y_{i,t-1}=0$) and exits ($y_{it}=0$) and that it can easily be used to calculate aggregated figures. However, this convenience comes at the cost of discontinuities of the distribution of g_{it} at -2 and 2. In a list of 10 alternative measures surveyed by Tornqvist *et al.* (1985), g_{it} (denoted there as H_3) is shown to be a useful measure for relative changes, but to be non-additive. The log difference $\ln(y_{it}/y_{i,t-1})$ is found to be preferable as it is the only measure of relative change that is symmetric, additive and normed. Of course, the drawback of the log differences as a measure of relative change is that it is not defined for $y_{it}=0$ and $y_{i,t-1}=0$, respectively.

In specifying the underlying data generating process and to analyze econometric models

for the conditional mean of g_{it} as carried out in HJM, we follow the large literature on the determinants of firm growth (see, e.g., Sutton 1997; Hart 2000 and Coad 2009 for surveys). In particular, for continuing firms we consider the log change in firm size $l_{it} = \ln(y_{it}) - \ln(y_{i,t-1})$ and derive the implied growth rate $g_{it} = g(l_{it})$ assuming

$$l_{it} = \ln z_{it} + \ln \eta_{it},$$

where $\ln z_{it}$ denotes the conditional mean.[3] $\ln \eta_{it}$ is an iid random disturbance with expectation 0. For continuing firms the conditional expectation of g_{it} is nonlinear in l_{it} and the DHS growth rate is given by

$$g(l_{it}) = \begin{cases} -2 & \text{if } y_{it} = 0 \text{ (exit)} \\ 2\frac{e^{l_{it}}-1}{e^{l_{it}}+1} & y_{it} \neq 0 \text{ and } y_{i,t-1} \neq 0 \\ 2 & \text{if } y_{i,t-1} = 0 \text{ (entry).} \end{cases}$$

For $y_{it} \neq 0$ and $y_{i,t-1} \neq 0$, the linear approximation of $g(l_{it})$ at 0 is given by

$$g(l_{it}) \;=\; 2\frac{e^{l_{it}}-1}{e^{l_{it}}+1} \approx 0 + 2\,\frac{e^{l_{it}}+1-\left(e^{l_{it}}-1\right)}{\left(e^{l_{it}}+1\right)^2}e^{l_{it}}\bigg|_{l_{it}=0} l_{it} = 2\frac{2}{(2)^2}l_{it} = l_{it},$$

implying that for continuing firms the specification used in HJM may be interpreted as a linear approximation of g_{it} in terms of l_{it}.

In order to parameterize the change in firm size, we follow HJM and assume that the conditional mean implied by the approximation of g_{it} by l_{it} is specified as $\ln z_{it} = \mathbf{x}'_{it}\beta + \alpha_n d_{it,n}$. $d_{it,n}$ takes the value of 1 if the firm enters and 0 otherwise. $\mathbf{x}_{it}$ is a $(K \times 1)$ vector of exogenous variables with the corresponding parameter vector β. Note, in line with HJM we

[3] Alternatively, one could consider the transformation $g_{it} \to l(g_{it})$. This means specifying the true model as $g_{it} = z_{it} + \eta_{it}$ and deriving the implied linear approximation $l(g_{it})$. Note, this model does not guarantee that $-2 \leq g_{it} \leq 2$ and l_{it} might not be defined for continuing firms. In general, the predicted values of g_t will deviate from 2 (-2) in case of entry (exit) when assuming this alternative data generating process. Analytically, for $-2 < g_{it} < 2$ we have

$$l(g_{it}) \;=\; \ln(2+g_{it}) - \ln(2-g_{it}) = \ln(2+z_{it}+\eta_{it}) - \ln(2-z_{it}-\eta_{it}), \quad \text{and}$$

$$l(g_{it}) = \begin{cases} -\infty & \text{if } g_{it} = -2 \text{ (exit)} \\ \ln(2+g_{it}) - \ln(2-g_{it}) & \\ \infty & \text{if } g_{it} = 2 \text{ (entry)} \end{cases}$$

This model implies that the conditional expectation of g_{it} is linear, while that of l_{it} is non-linear. However, at $-2 < g_{it} < 2$ the linear approximation of l_{it} at 0 yields $l_{it} \approx g_{it}$, since

$$\ln(2+g_{it}) - \ln(2-g_{it}) \approx \ln(2) - \ln(2) + \frac{1}{2}g_{it} - \frac{1}{2}(-1)g_{it} = g_{it}.$$

do not include a dummy for exiting firms. The resulting model pools over all three groups of entering, exiting and continuing firms and, hence, uses *all* observations. Formally, it is given by

$$g_{it} \approx \quad l_{it} = \begin{cases} \mathbf{x}'_{it}\beta + \varepsilon_{it} & \text{if } y_{i,t-1} \neq 0 \text{ (survival or exit)} \\ \mathbf{x}'_{it}\beta + \alpha_n + \varepsilon_{it} & \text{and } y_{i,t-1} = 0 \text{ if a firm enters.} \end{cases} \tag{1}$$

In the Appendix we show that applying (weighted or unweighted) OLS to this model with the DHS growth rate as the dependent variable yields biased and inconsistent estimates. In general, the bias results from pooling the observations of the groups of exiting, entering and continuing firms in a single model assuming that the marginal impact of the exogenous variables is the same for all three groups and by the lack of variation of the disturbances g_{it} in case of exiting and entering firms. These firms have either $g_{it} = 2$ or $g_{it} = -2$ and, therefore, non-stochastic error terms given by $\varepsilon_{it} = -2 - \mathbf{x}'_{it}\beta$ and $\varepsilon_{it} = 2 - \mathbf{x}'_{it}\beta - \alpha_n$, respectively. Furthermore, applying the Frisch, Waugh and Lovell theorem (see, e.g., Davidson and Mackinnon 1993) to get rid of the entry dummy shows that the bias stemming from the entrants also depends on the variation in the right hand side variables within this group as collected in $\mathbf{x}_{it}$. However, it seems that the inclusion of dummies for entering and exiting firms substantially reduces the bias.

HJM emphasize that it proves convenient to apply weighted regressions using cell weights $w_{it} = y_{it} + y_{i,t-1} / \left(\sum_{k=1}^{n_t^h} y_{kt} + y_{k,t-1} \right)$ for each cell h. Its worth nothing, that the bias carries-over to weighted OLS regressions as applied by HJM. In the two way model with main size and age effects, but without interaction effects, the weighted OLS estimator of β implies a model prediction that coincides with the cell means referring to the respective size and age groups. While this may be a convenient way to describe the data in a cross tabulation, it is does not yield an estimator of the marginal effects or conditional means as reflected in the model parameters.[4] Specifically, our result implies that the employment-weighted conditional means calculated by HJM are only imperfect estimates for the true marginal effects of firm size and firm age for net job creation, respectively. In addition, an unresolved issue is that the weights as applied by HJM and the related literature are endogenous as these themselves depend on g_{it} via y_{it} (for more details, see the discussion on the calculation of counterfactuals in Section 4 below). To our knowledge, an analytical solution to this endogenous weights problem so far does not exist, so we take up this issue in our Monte Carlo experiments below.

[4] In ANOVA terms the predicted means for age and size groups have to be interpreted as contrasts, i.e., a linear combinations of the estimated parameters. Given that we found that the parameters estimates are biased, the estimated contrasts will be biased as well.

As an alternative that avoids this bias, one can formulate a three-part model that allows to estimate separate equations for the entering, exiting and continuing firms. For simplicity, we assume that η_{it} is distributed as *iid* lognormal so that a consistent ML estimator can easily be derived.[5] The density with respect to the log difference in firm size of the continuing firms (i.e., conditional on $y_{it} \neq 0$ and $y_{i,t-1} \neq 0$) is given by

$$f(l_{it}|y_{it} \neq 0, y_{i,t-1} \neq 0) = \frac{1}{1-p_{it}-q(x_{it})} \frac{1}{\sqrt{2\pi}} \frac{1}{\sigma} e^{-\frac{1}{2}\frac{(l_{it}-\ln(z_{it}))^2}{\sigma^2}}, \tag{2}$$

where we denote the probability of entry by p_{it} and that of exit by $q(x_{it})$, respectively. The parameter vector of the model for the probability of exit $q(x_{it})$ may be estimated by a separate Probit model. By contrast, the probability of entry p_{it} can hardly be estimated using an econometric model at the firm level and, thus, will be treated as constant within industries and years. Then, the contribution to the likelihood function referring to period t can be written as (see the Appendix, for more details)

$$L_t(\gamma, \beta, \sigma, \mathbf{p}, \mathbf{q}) \quad = \quad \Pi_{n_{n,t}} p_{it} \Pi_{n_{x,t}} q(\gamma; x_{it}) \Pi_{n_t - n_{x,t} - n_{n,t}} f(\beta, \sigma; l_{it}, x_{it}).$$

In the Appendix we also demonstrate that under constant entry rates (within industries and years) the ML estimator of p_{it} can be derived as $\frac{n_{n,t}}{n_t}$, where $n_{n,t}$ denotes the number of entering firms, $n_{x,t}$ the number of exiting firms and n_t is the number of all firms in the sample.[6] Lastly, the parameters of the specification for the continuing firms, β and σ^2, can be estimated by maximizing the likelihood based on the density from equation (2) excluding the observations referring to entering and exiting firms. As demonstrated in Footnote 5, the ML estimates of the β parameters are numerically equivalent to the OLS estimates in a regression with the left hand side variable measured as log difference in firm size.

In order to analyze the overall impact of firm size and firm age on net job creation, we have to aggregate individual job creation rates within different firm groups. Based on the DHS approach the job creation rates for various groups of firms (e.g., industry, size and

[5] To obtain the density in terms of g_{it} we use the transformation $\eta_{it} = \frac{1}{z_{it}}\left(\frac{2+g_{it}}{2-g_{it}}\right)$ and $\frac{\partial \eta_{it}}{\partial g_{it}} = \frac{1}{z_{it}} \frac{4}{(2-g_{it})^2}$. Therefore, the distribution of the growth rate of the continuing firms can be derived as

$$f(\eta(g_{it})| -2 < g_{it} < 2) = \frac{1}{1-p_{it}-q(x_{it})} \frac{1}{\sqrt{2\pi}} \frac{4}{(2+g_{it})(2-g_{it})} \frac{1}{\sigma} e^{-\frac{1}{2}\frac{(\ln(2+g_{it})-\ln(2-g_{it})-\ln(z_{it}))^2}{\sigma^2}}$$

Note that $\ln(2+g_{it}) - \ln(2-g_{it}) = l_{it}$ and, thus, the parameter estimates of the resulting model are numerically identical to the OLS estimates of a model with l_{it} as dependent variable.

[6] In a more general specification the entry rates may be explained by industry specific variables instead of taking them as exogenously given.

age classes) with index h that are populated by n_t^h firms are calculated as

$$g_t^h = \sum_{i=1}^{n_t^h} \frac{y_{it} + y_{i,t-1}}{\sum_{k=1}^{n_t^h} y_{kt} + y_{k,t-1}} g_{it} = 2\frac{\sum_{i=1}^{n_t^h} (y_{it} - y_{i,t-1})}{\sum_{k=1}^{n_t^h} y_{kt} + y_{k,t-1}} \tag{3}$$

or, equivalently, as

$$\begin{aligned} g_t^h &= 2\frac{\sum_{i=1}^{n_t^h - n_{x,t}^h - n_{n,t}^h} y_{it} - y_{i,t-1}}{\sum_{k=1}^{n_t^h} y_{kt} + y_{k,t-1}} + 2\frac{\sum_{i=1}^{n_{t,n}^h} y_{it}}{\sum_{k=1}^{n_t^h} y_{kt} + y_{k,t-1}} \\ &+ 2\frac{-\sum_{i=1}^{n_{t,x}^h} y_{i,t-1}}{\sum_{k=1}^{n_t^h} y_{kt} + y_{k,t-1}} \\ &:= g_{t,c}^h + g_{t,n}^h + g_{t,x}^h. \end{aligned} \tag{4}$$

Equations (3) and (4) show that there is no need to estimate a pooled model as in HJM (see column 5 in their Table 2). By contrast, it is possible to recover the net job creation rates by calculating g_t^h or predictions thereof for the continuing firms and adding the corresponding (weighted) rates referring to the entering and exiting firms, respectively.

3 Monte Carlo simulation

In order to analyze the properties of the above discussed estimators in finite samples, we generate DHS job creation rates between time T and 0 in a cross-section of continuing firms according to the following econometric model:

$$\ln(y_{iT}) - \ln(y_{i0}) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \alpha_n d_{i,n} + \ln \eta_i,$$

where x_{i1} (firm size) takes the value of 1, if a generated lognormal random variable is larger than its median value and x_{i2} (firm age) is a dummy which equals 1 if a generated uniform random variable is larger than its respective median. Both of them are fixed in repeated samples. $beta_1$ and β_2 denote the conditional mean parameters to be estimated. Log initial size, $\ln(y_{i0})$, is generated as iid $N(0,1)$ and also kept fixed in repeated samples. It is also independent of η_i. So by construction this setting rules out any bias due to regression to the mean effects. Lastly, $\ln \eta_i$ is generated as iid $N(0, \sigma^2)$. The true data generating process assumes parameter values $\beta_1 = \beta_2 = -0.1$ and $\beta_0 = 0.05$. Put differently, the marginal effects for being large and being old are assumed to be -0.1, respectively. The

contrasts defined by the difference of the employment-weighted conditional means between the large and small and old and young firms, respectively, should likewise yield these values of -0.1.

In order to obtain groups of entering and exiting firms, we generate a Bernoulli random variable $d_{i,x}$ that sets $y_{iT} = 0$ with probability q and a second one, $d_{i,n}$, which sets $y_{i0} = 0$ with probability p. For the entering firms (at $d_{i,n} = 1$) we set x_{i2} to zero, since in empirical data the entering firms are, by definition, in the youngest age group. Lastly, we calculate the DHS growth rate which is given by $g_{iT} = 2\frac{y_{iT}-y_{i0}}{y_{iT}+y_{i0}}$.

In the Monte Carlo analysis, we carry out alternative experiments that consider combinations of $p \in \{0, 0.2\}$, $q \in \{0, 0.2\}$ and $\sigma \in \{0.1, 0.5\}$. Four different estimators are of interest: The HJM employment-weighted OLS estimator as well as its non-weighted counterpart based on all observations, the OLS estimator using only the continuing firms without the dummy for entering firms and, lastly, the ML estimator with the log difference in firm size as the dependent variable. In Table 1 we report the bias, the root mean square error (RMSE) (both multiplied by 100) and the rejection rates under H_0 of t-tests for $\beta_1 = -0.1$ and $\beta_2 = -0.1$. Thereby, the bias is given by the average difference between the estimated β_1 and β_2 parameters from their true values of -0.1, respectively, while the RMSE takes the averages of the squared differences between the estimated and the true parameters. Finally, the rejection rate is given by the share of replications where a t-test on the true parameter estimates of β_1 and β_2 being -0.1, respectively, rejects.

This small-scale Monte Carlo analysis illustrates the two mentioned potential sources of a bias. Since the data generating process is based on a lognormal model, the differences in the performance between the OLS estimator using only continuing firms and the ML estimator result from the approximation bias. In addition, a comparison of the HJM (weighted and unweighted) OLS estimators with the ML estimator illustrates the additional bias originating from both, the approximation and the pooling of the model using the HJM job creation rate as the dependent variable.

The first series of experiments assumes that $\sigma = 0.1$. In this case, extreme values of y_{iT} and y_{i0} exhibit low probability weights, respectively, and the linear approximation of g_{iT} should yield a small approximation error. Without entry and exit, i.e., $p = 0$ and $q = 0$, the bias and the RMSE are of similar size for the four considered estimators and the simulated rejection rates correspond to the nominal size of the t-test of 0.05. Here, the only exception is the weighted HJM estimator which overrejects the firm age conditional mean estimates. In the experiments with a positive exit rate $q = 0.2$ and no

Table 1: Monte Carlo Simulations: 1,000 observations and 10,000 replications

		σ=0.1				σ=0.5			
		$p=0$ $q=0$	$p=0$ $q=0.2$	$p=0.2$ $q=0$	$p=0.2$ $q=0.2$	$p=0$ $q=0$	$p=0.$ $q=0.2$	$p=0.2$ $q=0$	$p=0.2$ $q=0.2$
Bias*100									
HJM weighted-estimator using all observations and an entry dummy									
	β_1	0.021	0.250	1.166	1.650	0.096	0.214	1.195	1.541
	β_2	0.020	0.209	-0.016	0.244	0.105	0.137	0.040	0.142
HJM unweighted-estimator using all observations and an entry dummy									
	β_1	0.045	2.000	2.026	3.998	0.598	2.439	2.449	4.314
	β_2	0.040	2.066	-0.025	2.484	0.571	2.471	0.473	2.856
OLS estimator using only the continuing firms without an entry dummy									
	β_1	0.045	0.044	0.039	0.038	0.598	0.594	0.567	0.566
	β_2	0.040	0.033	0.030	0.030	0.571	0.541	0.525	0.522
Maximum likelihood estimator									
	β_1	0.006	0.005	0.000	-0.001	0.030	0.026	-0.002	-0.005
	β_2	0.001	-0.006	-0.009	-0.010	0.003	-0.030	-0.045	-0.048
Mean squared error*100									
HJM weighted-estimator using all observations and an entry dummy									
	β_1	0.006	0.158	0.020	0.225	0.176	0.377	0.185	0.433
	β_2	0.011	0.288	0.014	0.461	0.307	0.663	0.385	0.992
HJM unweighted-estimator using all observations and an entry dummy									
	β_1	0.004	0.289	0.044	0.399	0.093	0.386	0.132	0.476
	β_2	0.004	0.294	0.005	0.426	0.093	0.382	0.113	0.529
OLS estimator using only the continuing firms without an entry dummy									
	β_1	0.004	0.005	0.005	0.007	0.093	0.116	0.116	0.151
	β_2	0.004	0.005	0.005	0.007	0.093	0.115	0.114	0.151
Maximum likelihood estimator									
	β_1	0.004	0.005	0.005	0.007	0.101	0.126	0.127	0.166
	β_2	0.004	0.005	0.005	0.007	0.101	0.126	0.124	0.166
Rejection rate									
HJM weighted-estimator using all observations and an entry dummy									
	β_1	0.034	0.009	0.388	0.030	0.050	0.032	0.101	0.058
	β_2	0.172	0.115	0.185	0.142	0.189	0.160	0.209	0.191
HJM unweighted-estimator using all observations and an entry dummy									
	β_1	0.060	0.108	0.966	0.221	0.076	0.114	0.233	0.209
	β_2	0.057	0.111	0.061	0.146	0.074	0.116	0.092	0.150
OLS estimator using only the continuing firms without an entry dummy									
	β_1	0.060	0.057	0.058	0.053	0.076	0.072	0.072	0.065
	β_2	0.057	0.058	0.055	0.055	0.074	0.070	0.068	0.068
Maximum likelihood estimator									
	β_1	0.053	0.051	0.052	0.048	0.053	0.051	0.052	0.048
	β_2	0.050	0.050	0.049	0.050	0.050	0.050	0.049	0.050

entry $p = 0$ both the weighted and unweighted HJM estimators perform much worse, as both explanatory variables vary across firms in this group. However, also with positive entry $p = 0.2$, but no exit $q = 0$ we find large biases of the HJM estimators. Lastly the bias is most pronounced when both entry and exit occurs ($p = 0.2$ and $q = 0.2$). However, the simulation results also reveal that the weighted HJM-OLS estimator tends to exhibit smaller biases as compared to its unweighted counterpart. One reason for this result could be that under the imposed data generating process with random entry and exit, but a skewed firm size distribution, the discontinuities at -2 and 2 tend get lower weights as compared to the unweighted case. With positive entry or exit rates the t-tests tend to overreject and amount to as much as 97 percent at $p = 0$ and $q = 0.2$ in the absence of weighting, rendering the t-test useless in this case. In contrast, the OLS estimator that only uses continuing firms as well as the ML-estimator are hardly biased and the t-tests are properly sized.

Setting $\sigma = 0.5$ leads to similar results, although now more extreme values of y_{iT} and y_{i0} tend to occur so that we observe a somewhat higher approximation bias in addition to the bias resulting from the inclusion of entering and exiting firms. The t-tests for the unweighted HJM estimators again overreject with actual sizes between 0.074 and 0.233 at nominal size 0.05. As expected, the increase in σ leads to slightly oversized t-tests for the model that only uses continuing firms, ranging from 0.065 to 0.076. By contrast, the ML estimator is not affected by an increased σ.

4 Who creates jobs in Austria?

For our empirical application we utilize data from the Austrian Social Security Database (ASSD).[7] The ASSD is an administrative data set which includes records for all employees in Austria. More precisely, the data set represents a daily calender of employment relationships between individuals and firms and, thus, allows for each point in time to calculate the (overall) number of employees in a respective firm.[8] Four our purposes, we calculate annual employment figures taking June 7th as our reference day for the time period from 1993 to 2009. In line with HJM we concentrate on firms operating in all non-farm business sectors.

[7]These data have been used extensively for empirical research in labor economics (see, e.g., Card *et al* 2007; del Bono *et al.* 2012) and industrial organization (see, e.g., Huber and Pfaffermayr 2010; Huber *et al* 2012).

[8]Fink *et al.* 2010 provide a comprehensive discussion on how to extract firm level information from the ASSD.

The database comprises approximately 3 million firm-year observations, for which we apply the different estimators discussed above. We construct eight dummy variables for firm size and firm age, respectively. The classification into different size classes is based on the average firm size in the years t and $t-1$. This classification is based on what HJM refer to as *current size* and allows to control for regression to the mean effects. It is also well in line with previous literature on how to tabulate economic variables over consecutive time periods (see, van de Stadt and Wansbeek 1990 for a formal treatment). Table 2 reports the estimation results, where we additionally control for industry and time fixed effects. The largest (firm size $>$ 250 employees) and oldest firms (firm age $>$ 20 years) form our reference group.

Columns (1) and (2) show the results from OLS estimators for the full sample and an entry dummy. In the first column we report employment-weighted conditional means in the spirit of HJM while column (2) depicts the results for unweighted OLS. These estimators provide relatively good model fits as indicated by R^2-measures of 0.38 and 0.54, respectively. With regard to the employment-weighted conditional means for different firm size and firm age groups, we are able to, qualitatively, replicate the results obtained by HJM.[9] In comparison to the largest and oldest firms, smaller firms exhibit significantly lower net job creation rates, while younger firms tend to be crucial for job creation. Interestingly for one year old firms, the employment-weighted conditional means are estimated to be positive while its marginal effect from the non-weighted OLS estimator is negative. This deviating result is clearly driven by differences in the impact of firm entry on the employment-weighted conditional means in comparison to the non-weighted simple OLS estimator. More specifically, in column (2) the marginal effect of firm entry amounts to approximately 2.45 while the corresponding employment-weighted conditional mean is given by 2.14. This difference supports the view that employment-weighting might be able to reduce the bias induced by OLS estimation. Note, however, by definition the DHS growth rate for entering firms exactly amounts to a value of 2 and, thus, this (two-way) model specification of firm size and firm age provides inaccurate employment-weighted conditional means for e.g., all entrants irrespective of their firm size. As already discussed above this deviation stems from the bias in the OLS estimation, which occurs independently of whether OLS is accompanied by employment-weighting or not.

In the next step we focus on the OLS results for only continuing firms and compare them with our alternative ML approach. To start with, the exclusion of entering and exiting firms and the entry dummy variable worsens the fit of these models dramatically.

[9]Here, we compare our results with column (5) of Table 2 in HJM.

Table 2: Estimation Results for Firm-Level Net Employment Growth in Austria

	One-Part Models			Three-Part Model[a]	
	HJM Weighted	HJM Unweighted	OLS Continuing	PROBIT Exit	ML Continuing
Constant	-0.2010^{***}	-0.1563^{**}	-0.1775^{***}	-2.8337^{***}	-0.2026^{***}
	(0.0475)	(0.0656)	(0.0423)	(0.1533)	(0.0473)
Size 1 to 2	-0.2998^{***}	-0.4967^{***}	-0.0417^{***}	2.1560^{***}	-0.0364^{***}
	(0.0029)	(0.0025)	(0.0018)	(0.0483)	(0.0035)
Size 3 to 5	-0.0961^{***}	-0.1171^{***}	-0.0248^{***}	1.1099^{***}	-0.0187^{***}
	(0.0025)	(0.0024)	(0.0018)	(0.0484)	(0.0035)
Size 6 to 10	-0.0559^{***}	-0.0670^{***}	-0.0122^{***}	0.8287^{***}	-0.0062^{*}
	(0.0024)	(0.0024)	(0.0018)	(0.0485)	(0.0035)
Size 11 to 20	-0.0438^{***}	-0.0511^{***}	-0.0091^{***}	0.7100^{***}	-0.0042
	(0.0024)	(0.0024)	(0.0018)	(0.0486)	(0.0035)
Size 21 to 50	-0.0346^{***}	-0.0389^{***}	-0.0077^{***}	0.6191^{***}	-0.0047
	(0.0024)	(0.0024)	(0.0019)	(0.0489)	(0.0036)
Size 51 to 100	-0.0266^{***}	-0.0280^{***}	-0.0079^{***}	0.4876^{***}	-0.0082^{***}
	(0.0027)	(0.0027)	(0.0021)	(0.0507)	(0.0039)
Size 101 to 250	-0.0170^{***}	-0.0196^{***}	-0.0060^{***}	0.3558^{***}	-0.0081^{*}
	(0.0027)	(0.0027)	(0.0021)	(0.0531)	(0.0042)
Age 0 (Entry)	2.1356^{***}	2.4473^{***}	-	-	-
	(0.0020)	(0.0011)	-	-	-
Age 1	0.0596^{***}	-0.0189^{***}	0.1426^{***}	0.2282^{***}	0.1573^{***}
	(0.0064)	(0.0019)	(0.0010)	(0.0039)	(0.0013)
Age 2 to 3	0.0219^{***}	0.0216^{***}	0.0642^{***}	0.0617^{***}	0.0675^{***}
	(0.0041)	(0.0015)	(0.0008)	(0.0037)	(0.0009)
Age 4 to 5	0.0270^{***}	0.0397^{***}	0.0416^{***}	-0.0196^{***}	0.0433^{***}
	(0.0027)	(0.0015)	(0.0008)	(0.0042)	(0.0009)
Age 6 to 7	0.0303^{***}	0.0442^{***}	0.0301^{***}	-0.0604^{***}	0.0314^{***}
	(0.0028)	(0.0016)	(0.0008)	(0.0047)	(0.0009)
Age 8 to 10	0.0204^{***}	0.0404^{***}	0.0233^{***}	-0.0693^{***}	0.0243^{***}
	(0.0027)	(0.0015)	(0.0007)	(0.0046)	(0.0008)
Age 11 to 20	0.0148^{***}	0.0293^{***}	0.0136^{***}	-0.0709^{***}	0.0147^{***}
	(0.0018)	(0.0011)	(0.0005)	(0.0038)	(0.0006)
Goodness of fit[b]	0.3827	0.5397	0.0177	0.1823	0.0162
Observations	3,000,451	3,000,451	2,385,882	2,708,555	2,385,882

Notes: Robust standard errors in parenthesis. The models include 3-digit industry and year fixed effects which are not reported. [a]The first part refers to the estimates of industry-time specific entry rates. The Probit explains firm exit and the ML part estimates the determinants of job creation for the continuing firms. [b]The Pseudo-R^2 is reported for the Probit model.

This, however, is not surprising since, in case of unweighted OLS estimation, the net job creation rates of all entering firms are perfectly explained by the entry dummy. More importantly, however, the unweighted OLS estimates for continuing firms indicate negative and significant marginal firm size effects throughout suggesting that the largest firms are the most important job creators. By contrast, focusing on the ML estimator presented in column (5) we are hardly able to identify significant firm size effects for firms with more than five and less than 51 employees. Put differently, the net job creation rates of these small and medium sized firms are statistically identical with the ones of the largest firms. Focusing on the firm age effects, both the ML approach as well as the unweighted OLS estimator indicate that surviving young firms most positively contribute to job creation. The latter result is well in line with the employment-weighted conditional means for continuing firms reported by HJM (see, e.g., Figures 4 and 7 in HJM). From a methodological point of view, the unweighted OLS estimates for continuing firms are relatively similar to the maximum likelihood results indicating that the approximation bias involved when applying OLS seems to be relatively moderate for the data at hand.[10] This is especially true for the firm age effects while we do observe some systematic differences in the marginal firm size effects across both different estimators.

Finally, column (3) of Table 2 reports the results from a Probit model for firm exit indicating that firm size and firm age are crucial determinants of firm survival. In particular, these estimates clearly imply that smaller and younger firms are more likely to be forced out of the market. The marginal effects of both, firm size and firm age, monotonically decrease suggesting that the largest and oldest firms that are captured by the constant are the most likely to survive. These findings are well in line with the large literature on market exit of firms (see, e.g., Caves 1998 for a survey) and with the discussion provided by HJM.

In order to more explicitly analyze the impact of firm size and age heterogeneity on net job creation, we classify firms as small or large and young or old relative to the median of the firm size and the firm age distributions, respectively.[11] In a baseline scenario, we predict net job creation rates taking the observed heterogeneity in firm size and age distributions into account. From these figures we subtract the predictions from a counterfactual that calculates (overall) job creation rates for a hypothetical situation, where all firms are

[10]Note, that our discussion from Section 2 demonstrates that the differences between both results are not due to different assumptions concerning the data generating process. This becomes obvious from Footnote 2 which shows that our alternative ML estimator is numerically equivalent to applying OLS to the log difference in firm size.

[11]To give one example, a firm which is larger than the median size but younger than the median age would be classified as *large-young*.

of median size and median age. As a result, we obtain a quantification of the overall contribution of the groups of small or large and young or old firms to net job creation. This approach also allows us to unify the results from the three-part model by calculating expected exit rates from the Probit model reported in Table 2. In contrast, for this exercise we simply add the jobs created by entering firms. Consequently, entry is taken to be exogenously given and does not affect the marginal effects of age and size. In this way, one obtains the overall quantitative effects of firm heterogeneity. These are to be interpreted as average marginal impacts of age and size as we hold the remaining (industry and time) dummy variables constant.

For this analysis we have to take into account that the aggregate net job creation rates defined in equations (3) and (4) use weights that themselves depend on g_{it}. Therefore, based on the definition of g_{it} in equation (1), we calculate new weights for the baseline and for the counterfactual scenarios inserting

$$y_{it} + y_{i,t-1} = \frac{4}{2 - g_{it}} y_{i,t-1}$$

into both equations for the continuing firms and exiting firms, respectively. For entrants we use their birth size (y_{it}) as weights. In order to deal with possible differences of weights between the baseline and the counterfactual, we utilize the average weights of the two scenarios for calculating the labor shares and the overall marginal effects. Table 3 reports the results for both the HJM weighted and unweighted OLS estimators as well as for the alternative maximum likelihood procedure. With regard to the latter we also distinguish between a model with constant exit probabilities (i.e., without a Probit for exit) and the full three-part procedure. For each estimator and firm group we separately report the average marginal effects and the labor shares. Multiplying the marginal effects with the labor shares yields the overall contribution of each respective firm group

To start with, the different estimators indicate that entry and exit dynamics are crucial for overall job creation. This can be seen in the upper parts of each of the 4 panels in Table 3. The HJM OLS estimators and the ML estimator with endogenous survival probabilities indicate a negative (positive) impact of small (large) firms on net job creation. By contrast, when applying the ML estimator with constant exit probabilities, these marginal effects are negative for all (small and large) old firms. This result documents that assuming constant exit rates across firms of different size and age might lead to a misjudgment regarding their importance for net job creation.

In order to calculate each firm group's impact on overall job creation we have to take into

Table 3: Average marginal impact of age and size on net job creation (differences in percentage points)

	HJM weighted estimator using all observations			HJM unweighted estimator using all observations		
	Old	Young	All	Old	Young	All
Difference in percentage points						
Large	5.934	7.048	-	6.114	6.976	-
Small	−9.165	−6.689	-	−17.246	−15.470	-
Labor share						
Large	0.626	0.245	0.871	0.626	0.245	0.871
Small	0.047	0.082	0.129	0.047	0.082	0.129
All	0.673	0.327	1.000	0.673	0.327	1.000
Contribution to difference in percentage points						
Large	3.713	1.726	5.440	3.826	1.708	5.534
Small	−0.433	−0.548	−0.982	−0.816	−1.268	−2.084
All	3.280	1.178	4.458	3.010	0.440	3.450
	ML estimator with constant exit probabilities			ML estimator with endogenous exit probabilities		
	Old	Young	All	Old	Young	All
Difference in percentage points						
Large	−0.584	4.007	-	5.600	9.584	-
Small	−2.109	2.583	-	−13.587	−14.003	-
Labor share						
Large	0.646	0.240	0.886	0.642	0.239	0.881
Small	0.046	0.069	0.114	0.047	0.072	0.119
All	0.691	0.309	1.000	0.689	0.311	1.000
Contribution to difference in percentage points						
Large	−0.377	0.961	0.584	3.594	2.288	5.882
Small	−0.096	0.177	0.081	−0.644	−1.009	−1.653
All	−0.473	1.138	0.665	2.950	1.279	4.229

Notes: The counterfactual assumes that all firms are 8-10 years old and have 51-100 employees. This is the mode in the age/size class table upon which the dummy variable design is based on.

account that the labor share employed in small firms only amounts to approximately 10 percent and weight their contribution to the total effect of firm heterogeneity accordingly.

Turning to each firm group's overall contribution to net job creation (see the last three lines in each of the four panels in Table 3), the (weighted and unweighted) HJM estimators indicate that the old-large firms contribute most to net job creation (3.71 and 3.83 %-points), while small-young firms are responsible for the largest bulk of destroyed jobs (-0.55 and -1.27%-points). Job creation rates in Austrian, therefore, seem to differ substantially from those found by HJM for the US economy. By contrast, when applying maximum likelihood estimation with constant exit probabilities, the HJM result re-emerges implying that the young (small and large) firms contribute positively to overall job creation (1.14 %-points). Moreover, this estimator indicates that also large-old firms contribute negatively to overall net job creation (-0.38 %-points).

Finally, the three-part model that combines the estimation of exit probabilities with the ML estimates for continuing firms, again suggests that the group of small firms contributes negatively to overall net job creation (-1.65 %-points). By contrast, larger firms exhibit positive job creation rates with the quantitative effect being most pronounced for the group of large-old firms. Consequently, in qualitative terms the consistent three-part model leads to similar results as both the weighted and unweighted HJM OLS estimators. In quantitative terms, however, the unweighted HJM estimator overestimates the job creation rates of large-old firms and underestimates the impact of the other three groups of firms. Specifically, this implies that the impact of the group of all small firms (old and young) and and young-large firms is underestimated. In total, for our sample of Austrian firms, the HJM estimator underestimates the overall impact of the heterogeneity in firm size and age on net job creation by approximately 18.4% (i.e., $\frac{(4.229-3.450)}{4.229}$). The weighted HJM estimator also overestimates (underestimates) the impact of large-old (large-young) firms, but, in contrast to its unweighted counterpart, also overestimates the marginal effects for both groups of small firms. Overall, the latter weighted estimator overestimates the impact of firm heterogeneity by approximately 5% (i.e., $\frac{(4.458-4.229)}{4.458}$). This last result again supports the view that the weighting scheme proposed by HJM is able to reduce the bias of the OLS-estimator for the DHS job creation rate as dependent variable.

5 Conclusions

The analysis of (net) job creation by firms with different characteristics has a long tradition in economics. For a long time the discussion has been dominated by the question which (small or large) firms are the more important net job creators. Only recently, HJM highlighted the role of firm age where their results indicate that young firms are the most crucial (net) creators of jobs.

From a methodological point of view, the incorporation of firm entry and firm exit in the analysis of job creation rates has been a serious challenge and DH and DHS proposed a generalized measure of firm growth which permits such an integrated treatment. Their measure of job creation is defined for the closed interval of -2 to 2, where -2 (2) corresponds to firm entry (exit). While this measure might provide a convenient way to descriptively analyze net job creation at the aggregate level, it might also cause some problems when applied to firm level data and, thus, the DHS growth rate demands a specific econometric treatment.

This paper analytically shows that simple OLS estimation of the DHS growth rate at the firm level leads to biased and inconsistent slope parameter estimates. The bias results from two sources, namely from an approximation bias that stems from the implicit linear approximation of the model when applying OLS. The second source of bias comes from the discontinuities of the job creation rate distribution at -2 and 2 that are typically controlled for by the inclusion of entry and/or exit dummies. Moreover, this bias also carries-over to the estimation of employment-weighted conditional means when applying weighted OLS in the spirit of HJM. Consequently, the obtained conditional means might be imperfect estimates for the role of firm size and firm age for net job creation, respectively.

As an alternative one can estimate a three-part model for the DHS growth rate that provides consistent slope parameter estimates under the assumption of a lognormal distribution of the firm size of continuing firms. Moreover, the job creation effects of entering and/or exiting firms can be estimated with Probit models. A small-scale Monte Carlo simulation exercise confirms that the ML estimator provides consistent estimates, while the OLS procedure, that uses all firms (as proposed by HJM), causes biased slope parameters. In a similar vein, this Monte Carlo simulation exercise reveals that employment-weighted conditional means are also biased when applying weighted OLS regression.

Finally, we apply these different estimators to a sample of Austrian firms in order to analyze the impact of heterogeneity in firm size and age. Taking differences in the exit probabilities across firm size and age classes into account, our results indicate that the group of small firms contributes negatively to overall net job creation. By contrast, the contribution of large firms is positive. Unlike the evidence provided by HJM for the US, our data do not reveal a clear-cut result for the role of firm age for job creation. Furthermore, the application of the unweighted HJM procedure underestimates the overall impact of the heterogeneity in firm size and age on net job creation by approximately 18% while the weighting scheme proposed by HJM leads to an overestimated overall impact of only 5%.

References

Armington, Catherine and Zoltan Acs (2004), Job Creation and Persistence in Services and Manufacturing, Journal of Evolutionary Economics, 14(3), pp. 309-325.

Baldwin, John, Timothy Dunne and John Haltiwanger (1998), A Comparison of Job Creation and Job Destruction in Canada and the United States, Review of Economics and Statistics, 80(3), pp. 347-356.

Birch, David L. (1979), The Job Generation Process, MIT Program on Neighborhood and Regional Change, Cambridge, MA.

Burgess, Simon, Julia Lane and David Stevens (2000), The Reallocation of Labour and the Lifecycle of Firms, Oxford Bulletin of Economics Statistics 62(s1), pp. 885-907.

Card, David, Raj Chetty and Andrea Weber (2007), Cash-on-Hand and Competing Models of Intertemporal Behavior: New Evidence from the Labor Market, Quarterly Journal of Economics 122(4), pp. 1511-1560.

Caves, Richard E. (1998), Industrial organization and new findings on the turnover and mobility of firms, Journal of Economic Literature, 36(4), pp. 1947-1982.

Coad, Alex (2009), The Growth of Firms: A Survey of Theories and Empirical Evidence, Edward Elgar, Cheltenham.

Davidson, Russell and James G. MacKinnon (1993), Estimation and Inference in Econometrics, Oxford University Press, Oxford, New York.

Davis, Steven J. and John C. Haltiwanger (1992), Job Creation, Gross Job Destruction, and Employment Reallocation, Quarterly Journal of Economics 107(3), pp. 819-863.

Davis, Steven J., John C. Haltiwanger and Scott Schuh (1996), Job Creation and Destruction, MIT Press, Cambridge MA.

del Bono, Emilia, Andrea Weber and Rudolf Winter-Ebmer (2012), Clash of Career and Family: Fertility Decisions after Job Displacement, Journal of the European Economic Association, 10(4), pp. 659-683.

Faberman, R. Jason (2003), Job Flows and Establishment Characteristics: Variations Across U.S. Metropolitan Areas, William Davidson Institute at the University of Michigan Stephen M. Ross Business School, William Davidson Institute Working Papers Series, No 610.

Fink, Martina, Esther Kalkbrenner, Andrea Weber and Christine Zulehner (2010), Extracting Firm Information from Administrative Records: The ASSD Firm Panel, Working Paper 1004, NRN: The Austrian Center for Labor Economics and the Analysis of the Welfare State.

Foote, Christopher (2006), Comment on Volatility and Dispersion in Business Growth Rates: Publicly Traded versus Privately Held Firms by Davis, Haltiwanger, Jarmin and Miranda, NBER Macroeconomics Annual, 21(2006), pp. 157-166.

Fuchs, Michaela and Antje Weyh (2010), The Determinants of Job Creation and Destruction: Plant-Level Evidence for Eastern and Western Germany, Empirica, 37(4), pp. 425-444.

Guertzgen, Nicole (2009), Firm Heterogeneity and Wages under Different Bargaining Regimes: Does a Centralised Union Care for Low-Productivity Firms?, Jahrbuecher fuer Nationaloekonomie und Statistik, 229(2-3), pp.239-253.

Haltiwanger, John and Milan Vodopivec (2002), Gross Worker and Job Flows in a Transition Economy: An Analysis of Estonia, Labour Economics 9(5), pp. 601-630.

Haltiwanger, John and Milan Vodopivec (2003), Worker Flows, Job Flows and Firm Wage Policies: An Analysis of Slovenia, Economics of Transition 11(2), pp. 253-290.

Haltiwanger, John C., Ron S. Jarmin and Javier Miranda (2012), Who Creates Jobs? Small vs. Large vs. Young, Review of Economics and Statistics, forthcoming.

Hart, Peter E. (2000), Theories of Firms' Growth and the Generation of Jobs, Review of Industrial Organization 17(3), pp. 229-248.

Huber, Peter and Michael Pfaffermayr (2010), Testing for Conditional Convergence in Variance and Skewness: The Firm Size Distribution Revisited, Oxford Bulletin of Economics and Statistics 72(5), pp. 648-668.

Huber, Peter, Harald Oberhofer and Michael Pfaffermayr (2012), Job Creation and the Intra-distribution Dynamics of the Firm Size Distribution, Working Papers in Economics and Finance 2012-05, University of Salzburg.

Ibsen, Rikke and Niels Westergaard-Nielsen (2005), Job Creation and Destruction over the Business Cycles and the Impact on Individual Job Flows in Denmark 1980-2001, Allgemeines Statistisches Archiv/Journal of the German Statistical Society 89(2), pp. 183-207.

Ilmakunnas, Pekka and Mika Maliranta (2005), Worker Inflow, Outflow, and Churning, Applied Economics 37(10), pp. 1115-1133.

Moscarini, Guiseppe and Fabien Postel-Vinay (2012), The Contribution of Large and Small Employers to Job Creation in Times of High and Low Unemployment, American Economic Review, forthcoming.

Neumark, David, Brandon Wall and Junfu Zhang (2011), Do Small Businesses Create More Jobs? New Evidence for the United States from the National Establishment Time Series , Review of Economics and Statistics 93(1), pp. 16-29.

Searle, Shayle R. (1987), Linear Models for Unbalanced Data, Wiley, New York.

Stiglbauer, Alfred, Florian Stahl, Rudolf Winter-Ebmer and Josef Zweimüller (2003), Job Creation and Job Destruction in a Regulated Labor Market: The Case of Austria, Empirica, 30(2), pp. 127-148.

Sutton, John (1997), Gibrat's Legacy, Journal of Economic Literature, 35(1), pp. 40-59.

Tornqvist, Leo, Pentti Vartia and Yrjo O. Vartia (1985), How Should Relative Changes Be Measured?, The American Statistician 39(1), pp. 43-46.

van de Stadt, Huib and Tom Wansbeek (1990), Miscellanea, Regression Effects in Tabulating From Panel Data, Journal of Official Statistics 6(3), pp. 311-317.

Voulgaris Fotini, Theodore Papadogonas and George Agiomirgianakis (2005), Job Creation and Job Destruction in Greek Manufacturing, Review of Development Economics 9(2), pp. 289-301.

Appendix

A The bias of the OLS estimator for the DHS job creation rate

A.1 A general proof

In order to derive the bias of the OLS estimator under the approximated model when using all observations, we partition the model as follows

$$\mathbf{g} = \begin{bmatrix} -2\mathbf{e}_x \\ \mathbf{g}_c \\ 2\mathbf{e}_n \end{bmatrix} = \begin{bmatrix} \mathbf{X}_x & 0 \\ \mathbf{X}_c & 0 \\ \mathbf{X}_n & \mathbf{e}_n \end{bmatrix} \begin{bmatrix} \beta \\ \alpha \end{bmatrix} + \begin{bmatrix} -2\mathbf{e}_x - \mathbf{X}_x\beta \\ \varepsilon_c \\ 2\mathbf{e}_n - \mathbf{X}_n\beta \end{bmatrix},$$

where $\mathbf{e}_x$ and $\mathbf{e}_n$ are vectors of ones.[12] Index x labels exiting firms (with $y_{it} = 0$ and $y_{i,t-1} \neq 0$) and index c refers to the continuing firms (with $y_{it} \neq 0$ and $y_{i,t-1} \neq 0$). The third set of observations with index n denotes entering firms (with $y_{it} \neq 0$ and $y_{i,t-1} = 0$). Let $\mathbf{M}_n = \mathbf{I}_n - \mathbf{D}(\mathbf{D}'\mathbf{D})^{-1}\mathbf{D}'$ with $\mathbf{D} = [\mathbf{0}', \mathbf{0}', \mathbf{d}']'$ and apply the Frisch, Waugh, Lovell theorem (see, e.g., Davidson and Mackinnon, 1993) to obtain

$$\widehat{\beta} = (\mathbf{X}'\mathbf{M}_n\mathbf{X})^{-1}\mathbf{X}'\mathbf{M}_n\mathbf{g}$$

with

$$\begin{aligned} \mathbf{M}_n &= \mathbf{I}_{NT} - \begin{bmatrix} \mathbf{0} \\ \mathbf{e}_n \end{bmatrix} (\mathbf{e}_n'\mathbf{e}_n)^{-1} \begin{bmatrix} \mathbf{0} & \mathbf{e}_n' \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{I}_{(n_x+n_c)\times(n_x+n_c)} & \mathbf{0}_{(n_x+n_c)\times n_n} \\ \mathbf{0}_{n_n\times(n_x+n_c)} & \mathbf{I}_{(n_n\times n_n)} - \mathbf{e}_n(\mathbf{e}_n'\mathbf{e}_n)^{-1}\mathbf{e}_{n(n_n\times n_n)}' \end{bmatrix}, \end{aligned}$$

[12] In a weighted regressions one multiplies the left hand side and the right hand side of the model by some non-stochastic weight w. Thus, the analysis below remains unaffected by this model transformation. However, weighting induces heteroskedastic disturbances so that the estimation of robust standard errors is called for.

where n_x, n_c and n_n denotes the number of exiting, continuing and entering firms, respectively. Inserting the true model $\mathbf{g} = \mathbf{X}\beta + \mathbf{D}\alpha + \varepsilon$ and using $\mathbf{M}_D\mathbf{D} = \mathbf{0}$ yields

$$\begin{aligned}\widehat{\beta} &= \left(\mathbf{X}'\mathbf{M}_D\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{M}_D\left(\mathbf{X}\beta + \mathbf{D}\alpha + \varepsilon\right) \\ &= \beta + \left(\mathbf{X}'\mathbf{M}_D\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{M}_D\varepsilon.\end{aligned}$$

The DHS job creation rate is defined as follows: For exiting firms with $y_{it} = 0$ we measure $g_{it} = -2$ and for entering firms with $y_{i,t-1} = 0$ we have $g_{it} = 2$. In these cases the error terms $\varepsilon_{it,x} = -2 - \mathbf{x}'_{it,x}\beta$ and $\varepsilon_{it,n} = 2 - \mathbf{x}'_{it,n}\beta$ are non-stochastic, i.e.,

$$\varepsilon_{it} = \begin{cases} -2 - \mathbf{x}'_{it,x}\beta & \text{if } y_{it} = 0 \text{ (exit)} \\ \varepsilon_{it,c} & \text{if } y_{it} \neq 0 \text{ and } y_{i,t-1} \neq 0 \\ 2 - \mathbf{x}'_{it,n}\beta - \alpha & \text{if } y_{i,t-1} = 0 \text{ (entry).} \end{cases}$$

It follows that

$$\left(\mathbf{I}_{n_n} - \mathbf{d}\left(\mathbf{d}'\mathbf{d}\right)^{-1}\mathbf{d}'\right)\left(\mathbf{g}_n - \mathbf{X}_n\beta + \mathbf{d}\alpha\right) = \left(\mathbf{I}_{n_n} - \mathbf{d}\left(\mathbf{d}'\mathbf{d}\right)^{-1}\mathbf{d}'\right)\mathbf{X}_n\beta,$$

since $\left(\mathbf{I}_{n_n} - \mathbf{d}\left(\mathbf{d}'\mathbf{d}\right)^{-1}\mathbf{d}'\right)2\mathbf{e}_n = 0$ and $\left(\mathbf{I}_{n_n} - \mathbf{d}\left(\mathbf{d}'\mathbf{d}\right)^{-1}\mathbf{d}'\right)\mathbf{d}\alpha = 0$.

To analyze the consistency of this estimator we assume that the following limits exist:

1. $\lim_{n\to\infty} n^{-1}\left(\mathbf{X}'_x\mathbf{M}_D\mathbf{X}_x\right) = \mathbf{Q}$, which is non-singular.
2. $\lim_{n\to\infty} n_x^{-1}\mathbf{X}'_x\mathbf{e}_x = \mathbf{Q}_x$.
3. $\lim_{n\to\infty} n_x^{-1}\mathbf{X}'_x\mathbf{X}_x = \mathbf{Q}_{xx}$.
4. $\lim_{n\to\infty} n_n^{-1}\mathbf{X}'_n\left(\mathbf{I}_{n_n} - \mathbf{d}\left(\mathbf{d}'\mathbf{d}\right)^{-1}\mathbf{d}'\right)\mathbf{X}_n = \mathbf{Q}_n$.
5. $\lim_{n\to\infty}\frac{n_x}{n} = \delta_x$, $\lim_{n\to\infty}\frac{n_n}{n} = \delta_n$, $\lim_{n\to\infty}\frac{n_c}{c} = \delta_c$.

Then we have

$$\begin{aligned}E[\widehat{\beta} - \beta] &= -n\left(\mathbf{X}'_x\mathbf{M}_D\mathbf{X}_x\right)^{-1}\cdot \\ &\quad \left[2\frac{n_x}{n}n_x^{-1}\mathbf{X}'_x\mathbf{e}_x + \frac{n_x}{n}n_x^{-1}\mathbf{X}'_x\mathbf{X}_x\beta) + \frac{n_n}{n}n_n^{-1}\mathbf{X}'_n\left(\mathbf{I}_{n_n} - \mathbf{d}\left(\mathbf{d}'\mathbf{d}\right)^{-1}\mathbf{d}'\right)\mathbf{X}_n\beta\right] \\ \lim_{n\to\infty} E[\widehat{\beta} - \beta] &= -\delta_x\mathbf{Q}^{-1}\left(2\mathbf{Q}_x + \mathbf{Q}_{xx}\beta\right) - \delta_n\mathbf{Q}^{-1}\mathbf{Q}_n\beta.\end{aligned}$$

Using a standard law of large numbers, and under the assumptions stated above, it follows that $plim_{n\to\infty}\frac{n_c}{n}n_c^{-1}\mathbf{X}_c'\varepsilon_c = 0$ so that

$$plim_{n\to\infty}\left(\widehat{\beta}-\beta\right) = -\delta_x\mathbf{Q}^{-1}\left(2\mathbf{Q}_x+\mathbf{Q}_{xx}\beta\right)-\delta_n\mathbf{Q}^{-\mathbf{1}}\mathbf{Q}_n\beta.$$

Observe that

$$\mathbf{X}_n'\left(\mathbf{I}_{n_n}-\mathbf{d}\left(\mathbf{d}'\mathbf{d}\right)^{-1}\mathbf{d}'\right)\mathbf{X}_n = \mathbf{X}_n'\mathbf{X}_n-\left(1/n_n\right)\mathbf{X}_n'\left(\mathbf{e}_n\mathbf{e}_n'\right)\mathbf{X}_n.$$

From the last equation it becomes obvious that the slope parameters β are biased and estimated inconsistently for two reasons: First, the entries in $\mathbf{X}_n$ may not be invariant within the group of entering firms. Second, $\lim_{n\to\infty} n_x^{-1}\mathbf{X}_x'\left(2\mathbf{e}_x+\mathbf{X}_x\beta\right)\neq 0$, which will occur if $-2\mathbf{e}_x\neq\mathbf{X}_x\beta$. This is a consequence of the pooling assumption.

A.2 A simple example for the bias in a two way interaction model

This subsection provides a simple illustration of the bias in a two way interaction model (see, e.g., Searle, 1987, chapter 4) that only contains dummies for size and age groups and interactions thereof (index by $i=1,...,s$ and $j=1,...,a$, respectively). Thereby, the first age group refers to (zero aged) entering firms. For simplicity, we consider a cell i and j in in a cross-section of DHS job creation rates, g_{ijk}, between two periods t and $t-1$ and skip the time index. The econometric model is then given by

$$g_{ijk} = \mu_{ij}+\varepsilon_{ijk}.$$

The corresponding cell weights are denoted by $w_{ijk}=\frac{y_{ij,kt}+y_{ij,kt-1}}{\sum_{k=1}^{n_{ij}}y_{kt}+y_{k,t-1}}$, where $\sum_{k=1}^{n_{ij}}w_{ijk}=1$ and, for now, we treat them as fixed and exogenously given. The OLS estimator minimizes the weighted sum of squared residuals (WS). This, the first order condition,

the corresponding OLS estimator and its expectation, respectively, read as

$$\begin{aligned}
WS &= \sum_{i=1}^{s}\sum_{j=1}^{a}\sum_{k=1}^{n_{ij}} w_{ijk}(g_{ijk}-\mu_{ij})^2 \\
\frac{\partial WS}{\partial \mu_{ij}} &= -2\sum_{k=1}^{n_{ij}} w_{ijk}(g_{ijk}-\mu_{ij}) \Rightarrow \hat{\mu}_{ij} = \sum_{k=1}^{n_{ij}} w_{ijk} g_{ijk} \\
\hat{\mu}_{ij} &= \left[\begin{array}{ccc} -2\mathbf{e}'_{x,ij} & \mathbf{g}'_{c,ij} & 2\mathbf{e}'_{n,ij} \end{array} \right] \mathbf{w}_{ij} = -2\sum_{k=1}^{n_{ij,x}} w^x_{ijk} + \sum_{k=1}^{n_{c,ij}} w^c_{ijk} g^c_{ij,k} + 2\sum_{k=1}^{n_{ij,n}} w^n_{ijk} \\
&= -2\sum_{k=1}^{n_{ij,x}} w^x_{ijk} + \sum_{k=1}^{n_{c,ij}} w^c_{ijk}\left(\mu_{ij}+\varepsilon^c_{ij,k}\right) + 2\sum_{k=1}^{n_{ij,n}} w^n_{ijk} \\
E[\hat{\mu}_{ij}] &= -2\sum_{k=1}^{n_{ij,x}} w^x_{ijk} + \sum_{k=1}^{n_{c,ij}} w^c_{ijk}\mu_{ij} + 2\sum_{k=1}^{n_{ij,n}} w^n_{ijk}.
\end{aligned}$$

Under equal cell weights $w_{ijk} = \frac{1}{n_{ij}}$, this reduces to $\frac{n_{ij,c}}{n_{ij}}\mu_{ij} + \frac{2(n_{n,ij}-n_{x,ij})}{n_{ij}}$, which, in general differs from μ_{ij}. In this case, the bias of $\hat{\mu}_{ij}$ is given by $E[\hat{\mu}_{ij}-\mu_{ij}] = \frac{n_{ij,c}-n_{ij}}{n_{ij}}\mu_{ij} + \frac{2(n_{n,ij}-n_{x,ij})}{n_{ij}}$. It only vanishes if $n_{ij,c} = n_{ij}$ implying that $n_{n,ij} = n_{x,ij} = 0$. To see the connection to the general model remember that the entry dummy is subsumed in the age categories. Then each column in $\mathbf{X}$ refers to to a dummy for one cell and the bias can be derived as

$$\begin{aligned}
\widehat{\beta}-\beta &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\varepsilon \\
&= (\mathbf{X}'\mathbf{X})^{-1}\left[\begin{array}{ccc} \mathbf{X}'_x & \mathbf{X}'_c & \mathbf{X}'_n \end{array} \right] \left[\begin{array}{c} -2\mathbf{e}_x - \mathbf{X}_x\beta \\ \varepsilon_c \\ 2\mathbf{e}_n - \mathbf{X}_n\beta \end{array} \right]
\end{aligned}$$

and

$$E[\widehat{\beta}-\beta] = (\mathbf{X}'\mathbf{X})^{-1}\left(-2\mathbf{X}'_x\mathbf{e}_x - \mathbf{X}'_x\mathbf{X}_x\beta + \mathbf{X}'_c\varepsilon_c + 2\mathbf{X}'_n\mathbf{e}_n - \mathbf{X}'_n\mathbf{X}_n\beta\right).$$

Since each group has n_{ij} elements and $(\mathbf{X}'\mathbf{X})^{-1}$ is a diagonal matrix with elements $1/n_{ij}$, this expression reduces to

$$E[\hat{\mu}_{ij}-\mu_{ij}] = \frac{1}{n_{ij}}\left(-2n_{ij,x} - n_{ij,x}\mu_{ij} + 2n_{ij,n} - n_{ij,n}\mu_{ij}\right).$$

Thereby, we use that a typical element of the row vector $\mathbf{X}'_x\mathbf{e}_x$ is given by $n_{ij,x}$ and similarly for $\mathbf{X}'_n\mathbf{e}_n$. The typical element of $\mathbf{X}'_x\mathbf{X}_x\beta$ and $\mathbf{X}'_n\mathbf{X}_n\beta$, is $n_{ij,x}$ and $n_{ij,n}$, respectively.

B ML estimation for log differences

For simplicity, we assume constant entry probabilities across industries in deriving the log-likelihood of the three-part model. This yields

$$\begin{aligned}\ln L(\gamma,\beta,\sigma,\mathbf{p},\mathbf{q}) &= \sum_{t=1}^{T} n_{t,n}\ln(p_t)+(n_t-n_{t,n})\ln(1-p_t)\\ &\quad+\sum_{t=1}^{T}\sum_{i=1}^{n_{t,x}}\ln(q(\gamma;\mathbf{x}_{it,x}))-\sum_{t=1}^{T}\sum_{i=1}^{n_{t,c}}\ln(1-q(\gamma;\mathbf{x}_{it,x}))\\ &\quad+\sum_{t=1}^{T}\sum_{i=1}^{n_{t,c}}\ln f(\beta,\sigma;l_{it},\mathbf{x}_{it,c}).\end{aligned}$$

where $q(\gamma;\mathbf{x}_{it,x})$ denotes the conditional probability of exit. The score is given by

$$\begin{aligned}\frac{\partial \ln L(\gamma,\beta,\sigma,\mathbf{p},\mathbf{q})}{\partial p_t} &= \frac{n_{t,n}}{p_t}-\frac{n_t-n_{t,n}}{1-p_t}\\ \frac{\partial \ln L(\gamma,\beta,\sigma,\mathbf{p},\mathbf{q})}{\partial \gamma} &= \sum_{t=1}^{T}\sum_{i=1}^{n_{t,x}}\frac{\partial \ln(q(\gamma;\ \mathbf{x}_{it,x}))}{\partial\gamma}-\sum_{t=1}^{T}\sum_{i=1}^{n_{t,c}}\frac{\partial \ln(1-q(\gamma;\ \mathbf{x}_{it,x}))}{\partial\gamma}\\ \frac{\partial \ln L(\gamma,\beta,\sigma,\mathbf{p},\mathbf{q})}{\partial \beta} &= \sum_{t=1}^{T}\sum_{i=1}^{n_{t,c}}\frac{\partial \ln f(\beta,\sigma;\ l_{it},\mathbf{x}_{it,c})}{\partial\beta}\\ \frac{\partial \ln L(\gamma,\beta,\sigma,\mathbf{p},\mathbf{q})}{\partial \sigma} &= \sum_{t=1}^{T}\sum_{i=1}^{n_{t,c}}\frac{\partial \ln f(\beta,\sigma;\ l_{it},\mathbf{x}_{it,c})}{\partial\sigma}.\end{aligned}$$

From the first equation it immediately follows that $\widehat{p}_t=\frac{n_{t,n}}{n_t}$. Assuming a normal distribution, the score referring to (γ, β, σ) is the same as that of a two-part model where the probability of exit is specified with a Probit model. The contribution of the continuing firms to the likelihood is independent of the Probit equation as it only depends on (β, σ). This implies that the corresponding ML-estimator is equivalent to applying OLS to the model with $\ln l_{it}=\ln(y_{it})-\ln(y_{i,t-1})$ as the dependent variable for continuing firms only. See footnote 2 in the text for the formal derivation.