FRIEDRICH-ALEXANDER-UNIVERSITÄT ERLANGEN-NÜRNBERG

Lehrstuhl für VWL, insbes. Arbeitsmarkt- und Regionalpolitik
Professor Dr. Claus Schnabel

Diskussionspapiere
Discussion Papers

NO. 101

Estimating labor supply in self-employment: pitfalls and resolutions

DANIEL S. J. LECHMANN

SEPTEMBER 2017

ISSN 1615-5831

Editor: Prof. Dr. Claus Schnabel, Friedrich-Alexander-Universität Erlangen-Nürnberg

Estimating labor supply in self-employment: pitfalls and resolutions[a]

Daniel S. J. Lechmann[b]

Abstract: The small extant literature on the working hours of self-employed workers is deficient, because it often lacks a clear theoretical underpinning and suffers from three common mistakes: including the hourly wage as an explanatory variable, controlling for input factors of production, and not considering endogenous selection of self-employed workers. I introduce a structural causal model that makes clear that neither the wage nor input factors such as the number of employees or the amount of capital invested are determinants of working hours in self-employment. It also shows why selection bias arises when using a sample of self-employed individuals. I present an empirical discrete choice labor supply model that resolves these issues. Estimating this model with German data, I find that both non-labor income and education negatively affect labor supply in self-employment.

Zusammenfassung: Die bisherige, überschaubare Literatur zum Arbeitsangebot von Selbständigen ist mangelhaft: Oft fehlt es an einer theoretischen Basis und empirische Schätzungen leiden an drei verbreiteten Fehlern. Sie enthalten den Studenlohn als erklärende Variable, kontrollieren für Inputfaktoren und berücksichtigen nicht ausreichend die Selektion in Selbständigkeit. Dieser Artikel zeigt anhand eines Strukturellen Kausalen Modells, dass weder der Stundenlohn noch die Menge an Inputfaktoren, wie bspw. die Zahl der Beschäftigten oder die Menge eingesetzten Kapitals, Determinanten der Arbeitszeit in Selbständigkeit sind. Er zeigt ebenfalls, warum es zu Selektionsverzerrung kommt, wenn zur Schätzung eine Stichprobe von Selbständigen verwendet wird. Er stellt ein empirisches Discrete-Choice-Arbeitsangebotsmodell vor, dass diesen Umständen Rechnung trägt. Eine Schätzung des Modells mit deutschen Daten ergibt, dass Bildung und Nichtarbeitseinkommen die Arbeitszeit in Selbständigkeit negativ beeinflussen.

Keywords: Germany, labor supply, self-employment, SOEP, structural causal model, working hours

JEL classification: J22, J23

[a]For helpful comments and suggestions, I want to thank Claus Schnabel and participants in research seminars at the University of Erlangen–Nürnberg.

[b]Friedrich-Alexander University Erlangen–Nürnberg and Labor and Socio-Economic Research Center (LASER). Lange Gasse 20, 90403 Nuremberg, Germany. Email: dsj.lechmann@fau.de.

1 Motivation

Self-employed workers represent a non-negligible share of the labor force. In developed countries, they make up between 6.5% (USA) and 35.4% (Greece) of the total workforce (Germany: 11.0%, cf. OECD 2016). At the same time, they contribute disproportionately to aggregate wealth accumulation and play an important role in employment creation and productivity growth (see, e.g., Parker 2009: Ch. 9.1 and van Praag & Versloot 2007).

Nonetheless, there are only about a dozen studies on the labor supply, i.e., the working hours, of self-employed workers – compared to hundreds of studies on that same subject for dependent employees. Furthermore, studies for self-employed workers are almost exclusively based on US data and none uses German data so far.

Much worse than its limited scope is that the extant empirical literature on the labor supply of self-employed workers is consistently flawed. First, in a false analogy to labor supply models of workers in dependent employment, most studies estimate models that include the wage as an explanatory variable (see, e.g., Pencavel 2015, Ajayi-Obe & Parker 2005, Parker et al. 2005, Camerer et al. 1997). But in contrast to dependent employees, for self-employed workers the hourly wage does not correspond to the marginal rate of transformation and is actually no determinant of their labor supply decision (cf. Pencavel 2016, Boulier 1979). Worse, as the hourly wage of a self-employed worker generally depends on the working hours of that worker, conditioning on the wage results in endogenous selection bias and yields uninformative estimates of the effects of both the wage and the other variables that have been included.

Second, some studies use input factors such as start-up capital or the number of employees as explanatory variables (see, e.g., Carree & Verheul 2009, Ajayi-Obe & Parker 2005, Bitler et al. 2005). This approach overlooks that self-employed workers simultaneously choose the levels of own labor supplied and other input factors demanded. And if one does not perfectly control for all common causes of labor supply and input demand, this simultaneity results in spurious correlation between labor supply and input demand.

Third, the existing literature does not properly take into account that individuals do not randomly select into self-employment (see, e.g., Pencavel 2016, Pencavel 2015, Carree & Verheul 2009, Ajayi-Obe & Parker 2005, Bitler et al. 2005, Parker et al. 2005, Camerer et al. 1997, Boulier 1979). Selection into self-employment depends on the potential utility in that labor market state which in turn depends on working hours in that state. Therefore, estimating a labor supply model using a sample of self-employed workers is not informative about the effects of the explanatory variables on labor supply in self-employment – unless one corrects for the selection bias.

These deficiencies in the literature are possibly related to a lack of a clear theoretical underpinning of empirical models of self-employed workers' labor supply. According to Parker (2009: 345), it is common for specifications including the wage to be estimated "without any motivating discussion in terms of utility maximisation", which "is perhaps indicative of an underdeveloped body of theory relating to entrepreneurs' labor supply".

This paper addresses these concerns with the current state of the theoretical and empirical literature in the following ways. First, I introduce a structural causal model (SCM, see, e.g., Pearl et al. 2016: Ch. 1.5) of the labor supply of self-employed workers – including a motivating discussion in terms of utility maximization (section 2). Second, using this model, I show that conditioning on the wage, input factors of production, or self-employment in empirical labor supply models results in biased estimates (sections 3.1 to 3.3). Third, I present an empirical labor supply model that resolves these issues, including an original solution to the selection problem (section 3.4). Rather than relying on a more or less plausible exclusion restriction for the selection into self-employment, I make use of a discrete choice model that avoids the selection problem altogether. Fourth, I estimate this model using German data (section 4). I provide first estimates of the effects of non-labor income and education on labor supply in self-employment for Germany and demonstrate the sensitivity of the estimates to the mis-specifications discussed above. I conclude with some suggestions for future research in section 5.

2 A structural causal model of the labor supply of self-employed workers

The SCM presented in this section can be motivated by a static neoclassical theory of labor supply for a self-employed worker who owns and manages his own firm, partly resembling theory in Pencavel (2016) and Boulier (1979). Following Pearl (2009: Def. 7.1.1), the SCM consists of a set of exogenous variables X, a set of endogenous variables E, and a set of functions F. Equation 1 displays the set of variables that are assumed to be exogenous. The individual is a price taker, so X contains the price per unit consumption p, the price per unit output q, and the price per unit input (other than own labor) r. y is non-labor income, and e is the level of education of the worker.

$$X := \{p, q, r, y, e\} \tag{1}$$

Equation 2 defines the set of endogenous variables E. $SE = 1$ indicates self-employment while $SE = 0$ indicates dependent employment or not working at all. U_{SE}, U_{DE}, U_{NE} is utility conditional on being self-employed, in dependent employment, and not working, respectively. C_{SE} and H_{SE} is real consumption and labor supply if self-employed, respectively. G is profit, Q production, and Z input other than own labor (e.g., capital, paid labor). Finally, W_{SE} and W_{DE} is the wage in self-employment and dependent employment, respectively.

$$E := \{SE, U_{SE}, U_{DE}, U_{NE}, C_{SE}, H_{SE}, G, Q, Z, W_{SE}, W_{DE}\} \tag{2}$$

For every endogenous variable, there is a function in F that specifies how that variable is determined, see equation 3. The definitions of these functions follow below.

$$F := \{f_{SE}, f_{U_{SE}}, f_{U_{DE}}, f_{U_{NE}}, f_{C_{SE}}, f_{H_{SE}}, f_G, f_Q, f_Z, f_{W_{SE}}, f_{W_{DE}}\} \tag{3}$$

An individual chooses self-employment if the utility if self-employed is higher than the utility in dependent employment or the utility of not working.

$$f_{SE} : SE = \begin{cases} 1 & \text{if } (U_{SE} > U_{DE}) \wedge (U_{SE} > U_{NE}) \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

Since the model is not supposed to explain labor supply in dependent employment but only in self-employment, utility in dependent employment is simply the indirect utility given the real wage and real non-labor income, where the wage is affected by the level of education. Likewise, utility in non-employment is the indirect utility given real non-labor income.

$$f_{U_{DE}} : U_{DE} = V_{DE}(\frac{W_{DE}}{p}, \frac{y}{p}) \tag{5}$$

$$f_{W_{DE}} : W_{DE} = W_{DE}(e) \tag{6}$$

$$f_{U_{NE}} : U_{NE} = V_{NE}(\frac{y}{p}) \tag{7}$$

Utility in self-employment depends on real consumption and working hours given self-employment.

$$f_{U_{SE}} : U_{SE} = U_{SE}(C_{SE}, H_{SE}) \tag{8}$$

The individual consumes his entire real income, where income in self-employment is profits G plus non-labor income y.

$$f_{C_{SE}} : C_{SE} = \frac{G + y}{p} \tag{9}$$

A self-employed worker produces Q units of output subject to a production function that depends on input of own labor H_{SE}, the level of education e, and other inputs Z. Output is sold at price q, so self-employment profits G are sales qQ less costs rZ.

$$f_Q : Q = Q(H_{SE}, Z, e) \tag{10}$$

$$f_G : G = q \cdot Q - r \cdot Z \tag{11}$$

The individual therefore maximizes utility subject to the following consumption constraint implied by equations 9, 10, and 11.

$$C_{SE} = \frac{q \cdot Q(H_{SE}, Z, e) - r \cdot Z + y}{p}$$

So, the Lagrangian becomes

$$\begin{aligned} L(C_{SE}, H_{SE}, Z, \lambda) = {} & U_{SE}(C_{SE}, H_{SE}) \\ & + \lambda(C_{SE} - \frac{q \cdot Q(H_{SE}, Z, e) - r \cdot Z + y}{p}) \end{aligned}$$

The first-order conditions are then

$$\frac{\partial L}{\partial C_{SE}} = \frac{\partial U_{SE}(C_{SE}, H_{SE})}{\partial C_{SE}} + \lambda = 0 \tag{12}$$

$$\frac{\partial L}{\partial H_{SE}} = \frac{\partial U_{SE}(C_{SE}, H_{SE})}{\partial H_{SE}} - \lambda \frac{q}{p} \frac{\partial Q(H_{SE}, Z, e)}{\partial H_{SE}} = 0 \tag{13}$$

$$\frac{\partial L}{\partial Z} = -\lambda(\frac{q}{p} \frac{\partial Q(H_{SE}, Z, e)}{\partial Z} - \frac{r}{p}) = 0 \tag{14}$$

$$C_{SE} = \frac{q \cdot Q(H_{SE}, Z, e) - r \cdot Z + y}{p} \tag{15}$$

Using equations 12 and 13 to eliminate λ gives

$$-\frac{\frac{\partial U_{SE}(C_{SE},H_{SE})}{\partial H_{SE}}}{\frac{\partial U_{SE}(C_{SE},H_{SE})}{\partial C_{SE}}} = \frac{q}{p} \frac{\partial Q(H_{SE}, Z, e)}{\partial H_{SE}} \tag{16}$$

And equation 14 implies

$$\frac{\partial Q(H_{SE}, Z, e)}{\partial Z} = \frac{r}{p} \tag{17}$$

Equation 16 resembles the well-known condition of optimal labor supply for dependent employees in that it states that the marginal rate of substitution equals the marginal rate of transformation. The crucial difference is that the marginal rate of transformation for a dependent employee is the wage, whereas for a self-employed worker it is the real value of marginal product of own labor. And while the wage in dependent employment is typically assumed to be invariant to individual labor supply, the same assumption would be quite implausible with respect to the marginal product of labor for that would require a production technology with constant marginal productivity.[1]

As the manager of his own firm, the self-employed worker not only has to decide on own labor supplied but also on other input factors demanded. Therefore, he additionally needs to satisfy equation 17, which states that the marginal product of inputs other than own labor equals their real price.

The individual then jointly chooses H_{SE} and Z to simultaneously satisfy conditions 16 and 17 given the values of p, q, r, y, and e. Consequently, we arrive at the input demand function

$$f_Z : Z = Z(p, q, r, y, e) \tag{18}$$

and the labor supply function

$$f_{H_{SE}} : H_{SE} = H_{SE}(p, q, r, y, e) \tag{19}$$

Finally, let the wage in self-employment be profits per hour.

$$f_{W_{SE}} : W_{SE} = \frac{G}{H_{SE}} \tag{20}$$

For every SCM, there is a corresponding directed acyclical graph (DAG). Figure 1 displays the DAG that corresponds to the SCM at hand.

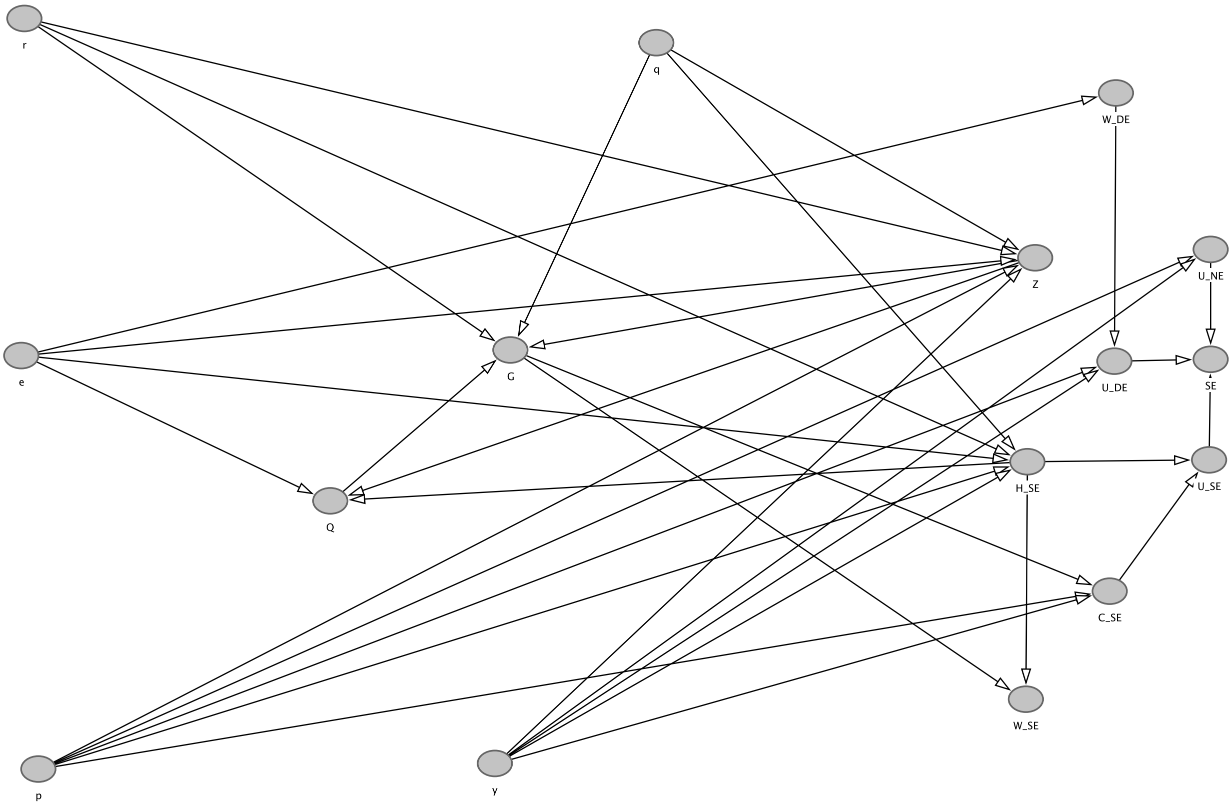
3 Implications for empirical labor supply models

3.1 Conditioning on the wage

Equation 19 shows that labor supply in self-employment H_{SE} does not depend on the wage W_{SE}, neither directly, since W_{SE} does not appear in $f_{H_{SE}}$, nor indirectly, because

[1]It is debatable whether the wage in dependent employment is also a function of working hours. Clearly, in the presence of non-linear taxation, the net wage is a function of working hours. The gross wage is probably constant, though. In contrast, for self-employed workers, under non-constant marginal productivity even the gross wage depends on working hours.

Figure 1: Corresponding DAG for SCM of labor supply of self-employed workers



Source: Own depiction using DAGitty (see Textor et al. 2011).

$f_{H_{SE}}$ contains only exogenous variables. Since W_{SE} does not affect H_{SE}, any attempt to estimate a wage elasticity of labor supply for self-employed workers is futile.

Conditioning on W_{SE} nevertheless when estimating a labor supply model for self-employed workers, results in biased estimates. That is because conditioning on a descendant of a dependent variable – just as conditioning on the dependent variable itself – results in "endogenous selection bias" (cf. Elwert & Winship 2014). W_{SE} is a descendant of H_{SE}, as labor supply affects the wage directly, since it appears in $f_{W_{SE}}$, see equation 20, and also indirectly via the path $H_{SE} \rightarrow Q \rightarrow G \rightarrow W_{SE}$ (cf. Figure 1), because of its influence on production and profits, see equations 10 and 11.

The direction of the bias is indeterminate as the relationship between labor supply H_{SE} and the wage W_{SE} is likely non-monotonic. For small values of H_{SE}, the marginal product of labor may be large and so the positive indirect effect of H_{SE} on W_{SE} likely dominates the negative direct effect, while for large values of H_{SE} and a correspondingly smaller marginal product of labor, the opposite is probably the case.

3.2 Conditioning on input factors of production

As H_{SE} and the amount of other inputs Z are chosen jointly by the same agent, estimating an effect of Z on H_{SE} is pointless. Note that in this case the estimation of a simultaneous equations model makes no sense as well (see, e.g., Wooldridge 2010: 239–241).

If one nevertheless conditions on Z, e.g., by including the number of employees or the amount of capital invested as explanatory variables, the resulting estimates will be biased. First, unless you control perfectly for all common causes of Z and H_{SE}, i.e., for p, q, r, y, and e, you will find an "effect" of Z on H_{SE} that is due to omitted variable bias.[2]

Second, conditioning on an outcome results in spurious correlation between all common causes of that outcome (see, e.g., Elwert & Winship 2014). Conditioning on Z therefore results in spurious correlation between p, q, r, y, and e. And now that these variables are correlated, omitting one of them leads to omitted variable bias for the effect of any of them on H_{SE}. Suppose, for instance, we are interested in the effect of education e on labor supply H_{SE}, and we condition on Z but cannot control for q. Then, correlation between e and H_{SE} is transmitted via the causal path $e \rightarrow H_{SE}$ but, since e and q are correlated conditional on Z, also via the non-causal path $e \rightarrow Z \leftarrow q \rightarrow H_{SE}$ (cf. Figure 1). So in this case, the conditional correlation between e and H_{SE} does not identify the causal effect of e on H_{SE}, whereas the unconditional correlation indeed does.

3.3 Conditioning on self-employment

Individuals choose the employment type that gives the highest utility, where utility depends on working hours and consumption, see equations 4, 5, 7, 8. Therefore, employment type SE is a descendant of the outcome variable H_{SE} because of the path $H_{SE} \rightarrow U_{SE} \rightarrow SE$ (cf. Figure 1). Consequently, using a sample of self-employed workers to estimate the determinants of labor supply in self-employment, i.e., conditioning on

[2] If you did manage to perfectly control for all common causes of Z on H_{SE}, the estimation would suffer from perfect multicollinearity between p, q, r, y, e, and Z.

$SE = 1$, results in endogenous selection bias.[3]

3.4 An empirical discrete choice labor supply model

Taking into account the above, a sensible empirical labor supply model for self-employed workers excludes the wage W_{SE} and the amounts of other input factors Z and includes all observable variables that appear in the labor supply function $f_{H_{SE}}$. If using a sample of self-employed workers to estimate the model, it is essential to correct for selection, e.g., by adjusting for the selection-propensity score (see Angrist 1997). One way to do so would be to apply the well-known method proposed by Heckman (1979), but that would require an exclusion restriction (see, e.g., Puhani 2000), i.e., a variable that is observable for the whole population and (directly or indirectly) affects selection into self-employment without affecting labor supply in self-employment. This author has had a hard time coming up with a variable meeting these requirements, which is why in the SCM above there is no such variable.

As an alternative, I suggest estimating a discrete choice labor supply model for the whole population (which avoids conditioning on $SE = 1$), where the categorical dependent variable combines working hours and employment types. In particular, in the most simple specification the dependent variable Y could be defined as

$$Y = \begin{cases} 1 & \text{if } H > k \wedge SE = 1 \\ 2 & \text{if } H \leq k \wedge SE = 1 \\ 3 & \text{if } DE = 1 \vee NE = 1 \end{cases}$$

where k is some threshold of working hours H, e.g., median working hours in self-employment, and SE, DE, and NE denote self-employment, dependent employment, and non-employment, respectively. Using this specification, it is possible to consistently estimate (approximations to) the conditional probabilities $P(Y = 1|x)$ and $P(Y = 2|x)$, where x can be any of the variables in the labor supply function $f_{H_{SE}}$ that is observable

[3] This issue resembles the well-known problem of sample selection bias in the labor supply literature for dependent employees: Using a sample of workers with strictly positive working hours, thereby conditioning on selection into employment, leads to biased estimates (see, e.g., Cahuc et al. 2014: 46–48).

for the whole population, in particular, y or e.

Specifically, since the functional form of $P(Y|x)$ is unknown, the empirical model consists of the three linear projections[4]

$$L(Y_1|1, y, \boldsymbol{e}, \boldsymbol{c}) = \beta_{1,0} + y\beta_{1,y} + \boldsymbol{e\beta}_{1,e} + \boldsymbol{c\beta}_{1,c} \tag{21}$$

$$L(Y_2|1, y, \boldsymbol{e}, \boldsymbol{c}) = \beta_{2,0} + y\beta_{2,y} + \boldsymbol{e\beta}_{2,e} + \boldsymbol{c\beta}_{2,c} \tag{22}$$

$$L(Y_3|1, y, \boldsymbol{e}, \boldsymbol{c}) = \beta_{3,0} + y\beta_{3,y} + \boldsymbol{e\beta}_{3,e} + \boldsymbol{c\beta}_{3,c} \tag{23}$$

where

$$Y_1 = \begin{cases} 1 & \text{if } Y = 1 \\ 0 & \text{otherwise} \end{cases}$$

and Y_2 and Y_3 analogously. $\boldsymbol{e}$ is a vector of dummies for different levels of education and $\boldsymbol{c}$ is a vector of control variables.

Then, $\beta_{1,x} - \beta_{2,x}$ is the (approximate) effect of x on the probability of working more than k hours in self-employment relative to working less or equal than k hours in self-employment and $\beta_{1,x} + \beta_{2,x} = -\beta_{3,x}$ is the effect of x on the self-employment selection probability. As an example, if $\beta_{1,y} = 0.1$ and $\beta_{2,y} = 0.2$, it follows that $\beta_{1,y} + \beta_{2,y} = 0.3 > 0$ and $\beta_{1,y} - \beta_{2,y} = -0.1 < 0$, which implies that non-labor income positively affects selection into self-employment, while negatively affecting labor supply in self-employment.

4 Estimation results for Germany

4.1 Estimation strategy

OLS consistently estimates the parameters in a linear projection (cf., e.g., Wooldridge 2010: 58). Consequently, I use "systems OLS", i.e., equation-by-equation OLS, to estimate the linear projections 21 to 23. A potentially more efficient alternative would be the systems feasible GLS estimator, but this estimator can be shown to collapse to OLS if

[4]Instead of using linear projections, i.e., best linear approximations to the unknown function $P(Y|x)$, one can just assume a functional form for $P(Y|x)$ and specify a multinomial logit model. I provide results for this more conventional approach in the appendix.

exactly the same regressors appear in each equation (cf. Cameron & Trivedi 2005: 210).

4.2 Data and variables

The data source used is the German Socio-Economic Panel (SOEP 2017). The SOEP is a representative annual panel survey started in 1984 (for details, see, e.g., Wagner et al. 2007). My estimation sample covers the waves 2005 to 2015, since the question on the number of employees Z is slightly different before 2005. Like previous studies, I restrict the sample to male individuals.

Table 1 shows the distribution of weekly working hours of male self-employed workers in Germany. Consistent with previous findings (see, e.g., Parker 2009: 341), self-employed workers report exceptionally long hours, with an average of 47 and a median of 50 hours. I set $k = 50$, so the dependent variables are three dummies indicating working more than median hours in self-employment (Y_1), working less or equal than median hours in self-employment (Y_2), and not being self-employed at all (Y_3).[5]

Table 1: Distribution of weekly working hours of male self-employed workers

Percentiles	Hours
1%	3
20%	39
25%	40
40%	45
50%	50
60%	50
75%	60
80%	60
99%	80
$N = 3,745$	

Source: Own calculations.
Data: SOEP 2005-2015.
Weights: phrf.

I estimate the effects of non-labor income y and education $\boldsymbol{e}$ on labor supply. Non-labor income is the equivalence weighted difference of pre-government household income[6] and

[5] As a robustness check, I use two rather than one hours threshold, namely, $k_1 = 60$ and $k_2 = 40$, i.e., the 75% and 25% percentiles of the hours distribution.

[6] Household income is the sum of household labor earnings, asset flows, private retirement income, and private transfers such as alimony and child support payments. Partly, this variable has been imputed by the data provider (see Grabka 2016).

individual labor earnings last year and is measured in €1,000 per month. $\boldsymbol{e}$ is a vector of five dummies, indicating different levels of education, from tertiary education to incomplete elementary education.

To take account of potential confounders, the vector of control variables $\boldsymbol{c}$ includes age (linear and squared) and the lagged dependent variables (three dummies). Non-labor income likely depends on working hours and employment types in the past. And if there is true state dependence, current labor supply choices also depend on past working hours and employment types (see, e.g., Lechmann & Wunder 2016 and Henley 2004 on state dependence in self-employment). Lagged dependent variables also account for unobserved heterogeneity to some extent (see, e.g., Angrist & Pischke 2009: Chs. 5.3–5.4). Additionally, $\boldsymbol{c}$ includes survey year dummies.

To illustrate the sensitivity of the estimates with respect to the mis-specifications discussed above, I estimate models with and without the wage W_{SE} and the number of employees Z as additional control variables. Here, W_{SE} is the natural logarithm of gross monthly profits divided by weekly working hours times 4.33. The number of employees Z is measured in three categories: $Z \geq 5$, $1 \leq Z \leq 4$, and none (reference category). Table 2 displays the descriptive statistics of the explanatory variables for the estimation sample.

Table 2: Descriptive statistics

	Mean	Std. Dev.
Non-labor income y (€1,000 per month)	0.71	1.32
Tertiary education (dummy)	0.21	
Full maturity certificate (dummy)	0.13	
Intermediate qualification (dummy)	0.30	
Elementary education (dummy)	0.34	
Elementary education incomplete (dummy)	0.02	
Age	42.82	12.28
Wage in self-employment W_{SE} (natural logarithm)	2.72	0.87
No. of employees Z: None (dummy)	0.42	
No. of employees Z: $1 \leq Z \leq 4$ (dummy)	0.36	
No. of employees Z: $Z \geq 5$ (dummy)	0.22	

Source: Own calculations. *Data*: SOEP 2005-2015. *Weights*: phrf.

4.3 Results

Columns 1 and 2 of Table 3 display the OLS estimates of the linear projection 21 for a sample of self-employed workers, i.e., conditional on $SE = 1$.[7] If controlling for the wage W_{SE} and the number of employees Z, the estimated effect of non-labor income on labor supply in self-employment is positive, small, and insignificant (see column 1). In contrast, when dropping W_{SE} and Z, this effect appears to be negative and is statistically significant at the 5% level (see column 2). The effect size is still small, though: An additional €1,000 non-labor income per month decreases the probability of working more than 50 hours by just 0.2 percentage points.

The estimates in column 2 most likely still suffer from selection bias. Columns 3 to 5 of Table 3 therefore display the OLS estimates of the linear projections 21 to 23 using a sample that besides self-employed workers includes dependent employees and non-employed individuals. Column 5 shows that non-labor income positively affects selection into self-employment. An additional €1,000 non-labor income per month decreases the probability of not being self-employed by 0.2 percentage points (statistically significant at the 5% level). However, it does so by increasing only the probability of working not more than 50 hours in self-employment by 0.3 percentage points (see column 4) while the probability of being self-employed and working more than 50 hours remains unchanged (see column 3).[8] So, non-labor income indeed seems to negatively affect labor supply in self-employment as $\hat{\beta}_{1,y} - \hat{\beta}_{2,y} = -0.003 < 0$, but this difference is not statistically significantly different from zero at the 5% level.[9]

The estimates of the education dummies tend to indicate a negative relationship between education and labor supply in both conditional-on-self-employment specifications, which is clearer in the specification without the wage and the number of employees but insignificant in both specifications (see columns 1 and 2 of Table 3). The estimates not conditional on self-employment confirm that the effect of education on

[7]Conditional on $SE = 1$, $\beta_{2,x} = -\beta_{1,x}$ and $\beta_{3,x} = 0$. Therefore, I do not explicitly show the conditional-on-self-employment estimates of the linear projections 22 and 23.

[8]The reported estimates in columns 3 to 5 do not sum to zero because of rounding error.

[9]Generally, the estimates for y may be biased towards zero because of inaccurate measurement of non-labor income.

Table 3: OLS estimates of discrete choice labor supply models

	1	2	3	4	5
Non-labor income y	0.001	–0.002*	0.000	0.003	–0.002*
(€1,000 per month)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Education $\boldsymbol{e}$ (reference: elementary education)					
Tertiary education	–0.043	–0.048	0.000	0.008*	–0.008*
(dummy)	(0.026)	(0.029)	(0.002)	(0.002)	(0.002)
Full maturity certificate	–0.061	–0.042	0.001	0.006*	–0.007*
(dummy)	(0.036)	(0.036)	(0.002)	(0.003)	(0.003)
Intermediate qualification	–0.029	–0.017	0.000	0.003	–0.003*
(dummy)	(0.027)	(0.028)	(0.001)	(0.002)	(0.001)
Elementary education incomplete	0.100	0.119	0.003	–0.003	0.001
(dummy)	(0.080)	(0.076)	(0.006)	(0.004)	(0.003)
Wage in self-employment W_{SE}	–0.066*				
(natural logarithm)	(0.012)				
No. of employees Z (reference: none)					
$1 \leq Z \leq 4$	0.103*				
(dummy)	(0.022)				
$Z \geq 5$	0.170*				
(dummy)	(0.029)				
Lagged outcome (reference: $SE = 0$)					
$H > 50 \wedge SE = 1$	0.500*	0.527*	0.669*	0.190*	–0.859*
(dummy)	(0.038)	(0.039)	(0.023)	(0.017)	(0.016)
$H \leq 50 \wedge SE = 1$	–0.068*	–0.074*	0.124*	0.655*	–0.779*
(dummy)	(0.034)	(0.034)	(0.011)	(0.018)	(0.015)
No. of observations	3,745	3,745	72,664	72,664	72,664
R^2	0.37	0.35	0.49	0.49	0.70

1 and 2: $P(H > 50|SE = 1)$, 3: $P(H > 50 \wedge SE = 1)$, 4: $P(H \leq 50 \wedge SE = 1)$,
5: $P(SE = 0)$. SE: self-employment, H: weekly working hours.
Additional control variables: age (linear and squared), survey year dummies.
Clustered standard errors in brackets. $^*p < 5\%$.
Source: Own calculations. *Data*: SOEP 2005-2015. *Sample*: Men. *Weights*: phrf.

labor supply in self-employment is negative.[10] For instance, having completed tertiary education increases the probability of being self-employed and working not more than 50 hours by 0.8 percentage points compared to elementary education and does not affect the probability of being self-employed and working more than 50 hours (see columns 3 and 4). Consequently, education positively affects selection into self-employment (see column 5).

Overall, these estimates demonstrate that both non-labor income and education

[10]If education increases the marginal product of own labor, the negative effect of education implies that in self-employment the income effect dominates the substitution effect.

negatively affect labor supply in self-employment. Importantly though, this result hinges on using the correct specification. If mistakenly conditioning on the wage, the number of employees, and self-employment, the relationship between non-labor income and education and labor supply remains unclear.

Several robustness checks largely confirm the negative effects of non-labor income and education. First, rather than estimating best linear approximations to the unknown function $P(Y|x)$, I estimated a multinomial logit model for $P(Y|x)$ (see Table 4 in the appendix). Second, I estimated a model with four rather than three choice categories: Being self-employed and working more than 60 hours, more than 40 and not more than 60 hours, less or equal than 40 hours, or not being self-employed at all. The negative effect of non-labor income turns out to be more pronounced in this specification, but the estimated effect of education is less clear (see Table 5 in the appendix). As to the effect of education, in a third robustness check, I excluded non-labor income and the lagged dependent variables. The level of education is determined at the beginning of the working life, whereas current non-labor income and lagged working hours and employment types are determined later during working life and possibly as a function of education. That would make y and the lagged dependent variables "bad controls" for the effect of education on labor supply (see Angrist & Pischke 2009: Ch. 3.2.3). Table 6 in the appendix shows that the estimated negative effect of education is distinctly more pronounced when dropping the potential bad controls.

5 Conclusions

I have introduced a structural causal model of the labor supply of self-employed workers. While this model can be regarded as simplistic, it makes clear that neither the wage nor input factors are determinants of self-employed workers' labor supply and including these variables in empirical models results in biased estimates. The model also clearly illustrates why selection bias arises when using a sample of exclusively self-employed workers for estimation. Pointing out these issues seems important given that the extant literature commonly does use the wage or input factors of production as explanatory variables while not properly taking selection of workers into self-employment into account.

I have proposed an empirical labor supply model that avoids these fallacies. It is easy enough not to include the wage or the number of employees as explanatory variables, but correcting for selection into self-employed is non-trivial. As an alternative, I have suggested using a discrete choice model that can be estimated for the whole population which avoids conditioning on selection into self-employment.

Estimating this model with German data, I have found that both non-labor income and education negatively affect labor supply in self-employment. Importantly, these effects are distinctly visible only if not conditioning on the wage, the number of employees, nor selection into self-employment.

In the structural causal model, non-labor income and education are assumed to be exogenous. Under this assumption, my estimates can be interpreted as effects. Evidently, if non-labor income and education are exogenous, anything that can go wrong is conditioning on too many variables. Nevertheless, I have also tried to take account of potential confounders at least to some extent, mainly by controlling for the lagged dependent variables. Still, up to now, the model omits many variables that are most likely also relevant for self-employed workers' labor supply, some of which might render non-labor income and education endogenous – even conditional on the lagged dependent variables. Therefore, in future research, variables such as (entrepreneurial) ability and (risk) preferences should also be considered and dynamics, imperfect competition, and the household context should be modeled as well.

Besides that, when estimating the effects of variables that are only observable in self-employment, it is not possible to avoid selection bias by means of estimating a discrete choice model similarly to the one I have proposed. So hopefully, researchers will come up with a credible exclusion restriction for a sample selection model in the future.

References

Ajayi-Obe O, Parker SC (2005): The changing nature of work among the self-employed in the 1990s: Evidence from Britain. *Journal of Labor Research* 26: 501–517.

Angrist JD (1997): Conditional independence in sample selection models. *Economics Letters* 54: 103–112.

Angrist JD, Pischke J (2009): *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton: Princeton University Press.

Bitler MP, Moskowitz T, Vissing-Jørgensen A (2005): Testing agency theory with entrepreneur effort and wealth. *Journal of Finance* 60: 539–576.

Boulier BL (1979): Supply decisions of self-employed professionals: The case of dentists. *Southern Economic Journal* 45: 892–902.

Cahuc P, Carcillo S, Zylberberg A (2014): *Labor Economics.* 2nd edition. Cambridge: MIT Press.

Camerer C, Babcock L, Loewenstein G, Thaler R (1997): Labor supply of New York City cabdrivers: One day at a time. *Quarterly Journal of Economics* 11: 407–441.

Cameron AC, Trivedi PK (2005): *Microeconometrics: Methods and Applications.* New York: Cambridge University Press.

Carree M, Verheul I (2009): Time allocation by the self-employed: The determinants of the number of working hours in start-ups. *Applied Economics Letters* 16: 1511–1515.

Elwert F, Winship C (2014): Endogenous selection bias: The problem of conditioning on a collider variable. *Annual Review of Sociology* 40: 31–53.

Grabka MM (2016): SOEP 2015 – Codebook for the $PEQUIV File 1984–2015: CNEF variables with extended income information for the SOEP. *SOEP Survey Papers* 343.

Heckman JJ (1979): Sample selection bias as a specification error. *Econometrica* 47: 153–161.

Henley A (2004): Self-employment status: The role of state dependence and initial circumstances. *Small Business Economics* 22: 67–82.

Lechmann DSJ, Wunder C (2016): The dynamics of solo self-employment: Persistence and transition to employership. *LASER Discussion Papers* 96, forthcoming in *Labour Economics.*

OECD (2016): *OECD Factbook 2015–2016.* http://dx.doi.org/10.1787/factbook-2015-en.

Parker SC (2009): *The Economics of Entrepreneurship.* New York: Cambridge University Press.

Parker SC, Belghitar Y, Barmby T (2005): Wage uncertainty and the labour supply of self-employed workers. *Economic Journal* 115: 190–207.

Pearl J (2009): *Causality: Models, Reasoning, and Inference.* New York: Cambridge University Press.

Pearl J, Glymour M, Jewell NP (2016): *Causal Inference in Statistics: A Primer.* Hoboken: Wiley.

Pencavel J (2015): The labor supply of self-employed workers: The choice of working hours in worker co-ops. *Journal of Comparative Economics* 43: 677–689.

Pencavel J (2016): Whose preferences are revealed in hours of work? *Economic Inquiry* 54: 9–24.

Puhani PA (2000): The Heckman correction for sample selection and its critique. *Journal of Economic Surveys* 14: 53–68.

SOEP (2017): Socio-Economic Panel (SOEP), data for years 1984-2015, version 32.1. doi: 10.5684/soep.v32.1.

Textor J, Hardt J, Knüppel S (2011): DAGitty: A graphical tool for analyzing causal diagrams. *Epidemiology* 5: 745.

van Praag CM, Versloot PH (2007): What is the value of entrepreneurship? A review of recent research. *Small Business Economics* 29: 351–382.

Wagner GG, Frick JR, Schupp J (2007): The German Socio-Economic Panel study (SOEP): Scope, evolution and enhancements. *Schmollers Jahrbuch* 127: 139-169.

Wooldridge JM (2010): *Econometric Analysis of Cross Section and Panel Data.* Cambridge: MIT Press.

Appendix

Table 4: Average partial effects for multinomial logit model

	1	2	3
Non-labor income y	0.000	0.001*	–0.001*
(€1,000 per month)	(0.000)	(0.000)	(0.000)
Education $\boldsymbol{e}$ (reference: elementary education)			
Tertiary education	0.000	0.008*	–0.008*
(dummy)	(0.001)	(0.002)	(0.002)
Full maturity certificate	0.000	0.007*	–0.008*
(dummy)	(0.002)	(0.003)	(0.003)
Intermediate qualification	0.000	0.003*	–0.003*
(dummy)	(0.001)	(0.002)	(0.002)
Elementary education incomplete	0.006*	–0.009*	0.003
(dummy)	(0.003)	(0.003)	(0.003)
Lagged outcome (reference: $SE = 0$)			
$H > 50 \wedge SE = 1$	0.556*	0.184*	–0.740*
(dummy)	(0.023)	(0.016)	(0.023)
$H \leq 50 \wedge SE = 1$	0.105*	0.565*	–0.670*
(dummy)	(0.010)	(0.020)	(0.019)
No. of observations		72,664	

1: $P(H > 50 \wedge SE = 1)$, 2: $P(H \leq 50 \wedge SE = 1)$, 3: $P(SE = 0)$.
SE: self-employment, H: weekly working hours.
Additional control variables: age (linear and squared), survey year dummies.
Delta-method standard errors in brackets. *$p < 5\%$.
Source: Own calculations. *Data*: SOEP 2005-2015. *Sample*: Men. *Weights*: phrf.

Table 5: OLS estimates of discrete choice labor supply model (four choice categories)

	1	2	3	4
Non-labor income y	0.000	–0.001	0.004*	–0.003*
(€1,000 per month)	(0.000)	(0.001)	(0.002)	(0.001)
Education $\boldsymbol{e}$ (reference: elementary education)				
Tertiary education	–0.001	0.006*	0.002	–0.008*
(dummy)	(0.001)	(0.002)	(0.002)	(0.002)
Full maturity certificate	–0.001	0.002	0.006	–0.007*
(dummy)	(0.001)	(0.002)	(0.003)	(0.003)
Intermediate qualification	0.001	0.002	0.000	–0.003*
(dummy)	(0.001)	(0.001)	(0.001)	(0.001)
Elementary education incomplete	0.007	–0.006	–0.000	–0.001
(dummy)	(0.008)	(0.005)	(0.002)	(0.003)
Lagged outcome (reference: $SE = 0$)				
$H > 60 \wedge SE = 1$	0.437*	0.375*	0.032*	–0.844*
(dummy)	(0.043)	(0.036)	(0.012)	(0.029)
$40 < H \leq 60 \wedge SE = 1$	0.070*	0.669*	0.101*	–0.839*
(dummy)	(0.009)	(0.019)	(0.011)	(0.013)
$H \leq 40 \wedge SE = 1$	0.015*	0.191*	0.529*	–0.736*
(dummy)	(0.005)	(0.021)	(0.032)	(0.023)
R^2	0.23	0.52	0.32	0.70
No. of observations	72,664			

1: $P(H > 60 \wedge SE = 1)$, 2: $P(40 < H \leq 60 \wedge SE = 1)$, 3: $P(H \leq 40 \wedge SE = 1)$, 4: $P(SE = 0)$. SE: self-employment, H: weekly working hours.
Additional control variables: age (linear and squared), survey year dummies.
Clustered standard errors in brackets. $*p < 5\%$.
Source: Own calculations. *Data*: SOEP 2005-2015. *Sample*: Men. *Weights*: phrf.

Table 6: OLS estimates of discrete choice labor supply model (excluding non-labor income and lagged dependent variables)

	1	2	3
Education $\boldsymbol{e}$ (reference: elementary education)			
Tertiary education	0.013*	0.034*	–0.048*
(dummy)	(0.004)	(0.005)	(0.007)
Full maturity certificate	0.006*	0.024*	–0.030*
(dummy)	(0.003)	(0.006)	(0.007)
Intermediate qualification	0.002	0.008*	–0.011*
(dummy)	(0.002)	(0.003)	(0.004)
Elementary education incomplete	0.010	–0.007*	–0.004
(dummy)	(0.019)	(0.004)	(0.019)
R^2	0.01	0.02	0.03
No. of observations	72,664		

1: $P(H > 50 \wedge SE = 1)$, 2: $P(H \leq 50 \wedge SE = 1)$, 3: $P(SE = 0)$.
SE: self-employment, H: weekly working hours.
Additional control variables: age (linear and squared), survey year dummies.
Clustered standard errors in brackets. $*p < 5\%$.
Source: Own calculations. *Data*: SOEP 2005-2015. *Sample*: Men. *Weights*: phrf.

In der Diskussionspapierreihe sind kürzlich erschienen:

Recently published Discussion Papers:

101	Lechmann, D.	Estimating labor supply in self-employment: pitfalls and resolutions	09/2017
100	Kuehnle, D. Oberfichtner, M.	Does early child care attendance influence children's cognitive and non-cognitive skill development?	03/2017
99	Hinz, T.	Personnel policy adjustments when apprentice positions are unfilled: Evidence from German establishment data	09/2016
98	Lechmann, D., Wunder, C.	The dynamics of solo self-employment: persistence and transition to employership	05/2016
97	Hirsch, B., Jahn, E. J., Oberfichtner, M.	The urban wage premium in imperfect labour markets	01/2016
96	Hirsch, B., Lechmann, D., Schnabel, C.	Coming to work while sick: An economic theory of presenteeism with an application to German data	04/2015
95	Schnabel, C.	United, yet apart? A note on persistent labour market differences between western and eastern Germany	03/2015
94	Hirsch, B., Oberfichtner M., Schnabel, C.	The levelling effect of product market competition on gender wage discrimination	07/2014
93	Konietzko, T.	Der Einfluss von Arbeitslosigkeit der Ehemänner auf die Zeitallokation von Paaren	07/2014
92	Hirsch, B., Merkl, C., Mueller, S., Schnabel, C.	Centralized vs. Decentralized Wage Formation: The Role of Firms' Production Technology	06/2014
91	Bossler, M., Oberfichtner, M.	The employment effect of deregulating shopping hours: Evidence from German retailing	02/2014

Eine aktualisierte Liste der Diskussionspapiere findet sich auf der Homepage:
http://www.arbeitsmarkt.rw.fau.de/

An updated list of discussion papers can be found at the homepage:
http://www.arbeitsmarkt.rw.fau.de/