Network for Studies on Pensions, Aging and Retirement

NETSPAR ACADEMIC SERIES

The Dynamics of Self-employment in the Netherlands

Elisabeth Beusch

MSc Thesis 2015-036

The Dynamics of Self-employment in the Netherlands*

Elisabeth Beusch†
ANR: 641894
e.beusch@uvt.nl

Version date: 29th October 2015

Abstract

This paper presents a dynamic multinomial logit model to explain the transitions in and out of self-employment using a particular set of Dutch micro-paneldata, the LISS panel. By taking account of unobserved heterogeneity in the model we are able to differentiate between what Heckman (1981b) calls true and spurious state dependence. We find that past and initial labour states combined have the largest influence on choice probabilities. We also find a relative low covariance between unobserved preferences for self-employment and non participation in the labour force, whereas the preference to be self-employed has a positive covariance with the one for unemployment. The simulation results from the model support what has been found in descriptive studies: men have a higher probability to choose self-employment than women, and men with a small family are also more likely to choose self-employment than those without family.

*This paper uses data of the LISS (Longitudinal Internet Studies for the Social sciences) panel administered by CentERdata (Tilburg University, The Netherlands).
†I would like to thank Bas van Heiningen, Carlos Eduardo Sandoval Moreno and Yeqiu Zheng for helpful comments.

1. Introduction

Interest in the self-employed has increased in the Netherlands as the share of self-employed in the working population has been growing continuously over the past years. This increase in the self-employed is being driven by one particular group of self-employed, the so called "zzp'ers"—zelfstandigen zonder personel, i.e. (self-employed) entrepreneurs without any employees. (CBS, 2014) Because this is a new phenomena policy makers are interested in the effects that the growth of their number may have on the labour market, social security systems, or government finances. To this goal the Dutch government set up a working group in May 2014 for inter-departmental policy research (Interdepartementaal Beleidsonderzoek (IBO)).[1] The IBO aims to explain among others why the number of "zzp'ers" has increased, how their increased number may have affected the economy, and to which extent fiscal, social and labour policies should be adjusted to counteract possible negative effects and increase positive effects. (CBS, 2014)

As a response to the questions asked by the IBO the CBS has written a report (CBS, 2014) that summarises the background information they have on the self-employed. The information in this report is however mostly descriptive but also covers entries and exits from self-employment. The goal of my paper is thus to estimate a dynamic model of self-employment which should function as a first draft[2] to develop a better model in the future that is able to explain the patterns in the data described by the CBS.

Within the academic literature there are two strands that have been looking at self-employment. First, there is a whole literature in business economics that looks at the choice to be self-employed, mostly from a perspective where individuals are entrepreneurs, i.e. where they are considered as small businesses. (Blanchflower, 2000) offers a good overview of this literature. Two points should be noted concerning this literature: As far as I know the literature it only considers binomial regression models, and restricts itself to the choice between self-employment and wage-employment. Second, it features mostly static models—I cannot remember having seen a dynamic framework. (See Beusch (2015) for a dynamic logit model.)

The second strand of literature is in labour economics, and has been using dynamic multinomial models already for quite some time. In this literature self-employment is often not the main interest of the study but enters as one of the choices considered in the model. See e.g. Gong et al. (2004) who model the choice between informal and formal sector work in Mexico, and Buddelmeyer and Wooden (2011) who model dynamics between casual and other types of employment in Australia. Oguzoglu (2010) follows Gong et al. (2004) to model the influence of disability on the decision of employment state, and Zucchelli et al. (2012) consider self-employment as an alternative to part-time employment for the elderly under possible ill-health. Another case in point is Prowse (2012) who includes self-employment

[1] See e.g. nu.nl/zzp/3783595/werkgroep-brengt-positie-zzpers-in-kaart.html, retrieved on 3 May 2015.

[2] This model has thus been mostly an exercise to develop my coding skills next to better understanding self-employment

when modelling the labour participation of women. And last but not least Been and Knoef (2013) in particular, who model self-employment decisions for the elderly in the Netherlands.

As I already have stated, the goal of this paper is humble. I am therefore not adding much to the existing literature but rather filling a small gap. The contribution is in two points: First, I am estimating a model for the population between 23 and 60 years, whereas Been and Knoef (2013) focus on the elderly only. Second, while I use with the LISS panel a relatively rich data set I am nevertheless trying to estimate a relatively simple model that only includes basic covariates which should be available in almost any other dataset too and hence would allow for the model to be estimated using other datasets as well. This is however a point I will not focus on in the remainder of the paper. A drawback of this setting is that I am not able to make a statement concerning the reasons why an individual enters self-employment. Thus, I can not discuss hypotheses, as e.g. the necessity hypothesis (Been and Knoef, 2013) which would argue that individuals enter self-employment because they cannot find wage-employment following unemployment for example.

In order to explain self-employment dynamics, I follow the literature in economics and model self-employment as a choice of an individual in the labour market using a dynamic multinomial choice framework with unobserved heterogeneity. Using a multinomial logit model will allow me to avoid possible sample selection bias, which may occur when one estimates a binomial model only. Including unobserved heterogeneity, by allowing for correlated random effects following Train (2009), on the other hand will allow me to differentiate between what Heckman (1981b) calls spurious and true state dependence. Note that this distinction is important to be made if we want to uncover the dynamics in the data. I solve the problem of initial conditions that arises in dynamic models following Wooldridge (2005) and Albarrán et al. (2015).

The main finding is that past and initial labour states combined have the largest impact on labour state choice probabilities. Nevertheless, adding individual and household characteristics to the covariates improves the model. I also find that the covariance between individual preferences for self-employment and non-participation in the labour force is relatively low. The covariance between unemployment, and self-employment or non-participation in the labour force is however positive and quite large when compared to the variances. The simulation results also show some support that the model matches observations made in the literature before. I find that for the specific individuals modelled men have a higher probability to choose self-employment than women, and all else equal, individuals with a small family are also more likely to choose self-employment than single individuals.

The remainder of the paper is as follows. The next section gives a definition of self-employment and discusses it in the context of the LISS panel. Section 3 introduces the LISS panel and includes some descriptive statistics of the data used. Section 4 first gives a short databased motivation for the chosen model. In a second part it describes the model and methods and I use to estimate the dynamics of self-employment. Section 5 presents and section 6 discusses the estimation results, while section 7 concludes.

2. Definition of self-employment

The International Labour Organisation (ILO) classifies employment status based on six types of work contracts that are based on the type of economic risk and authority which are involved when practising a job. These six groups are further defined on a broader level by distinguishing between paid-employment jobs on one side and self-employment jobs on the other side. The definition of the latter is given as follows:

> Self-employment jobs are those jobs where the remuneration is directly dependent upon the profits (or the potential for profits) derived from the goods and services produced (where own consumption is considered to be part of profits). The incumbents make the operational decisions affecting the enterprise, or delegate such decisions while retaining responsibility for the welfare of the enterprise. (In this context "enterprise" includes one-person operations.)

From a theoretical point of view most of the literature on self-employment uses the definition by the ILO, or a very similar one. There are however some issues if one defines self-employment in practise. Both Parker (2004) and CBS (2014) offer a discussion on potential issues; the former from a general perspective and the latter with a focus on the Netherlands.

One particular issue with Dutch data pointed out by the CBS are so called majority shareholder directors ("dga's"). While the tax authority considers "dga's" as employees, apparently 90% of them considered themselves as self-employed when asked in a survey. Since the LISS panel, which is based on surveys, is not combined with fiscal data in this paper this problem is not directly of relevance. It should however be kept in mind whenever Dutch data is e.g. compared to international data[3], and when we define self-employment in this study we also have to choose in some cases how we want to treat the "dga's", as will be discussed in section 3.1.

3. Data

As stated earlier in the text, this paper makes use of the LISS panel. CentERdata provides the following description of the LISS panel on their website[4]:

> The LISS panel is a representative sample of Dutch individuals who participate in monthly Internet surveys. The panel is based on a true probability sample of households drawn from the population register. Households that could not otherwise participate are provided with a computer and Internet connection. A longitudinal survey is fielded in the panel every year, covering a large variety of

[3] Tthe ILO does not specify how "dga's" should be treated. The CBS counts "dga's" as self-employed persons in the publication I cite but it is for example not clear if they are included in the number of self-employed in the Statline data used in section 3.2.

[4] See reference guidelines at lissdata.nl/assets/uploaded/References_LISS.pdf retrieved on May 13 2015

domains including work, education, income, housing, time use, political views, values and personality.

Based on the discussion in section 3.1 I only use the *Work and Schooling* core study as well as some information on the individuals and the households from the background variables. The core studies are on an annual basis and currently there are seven waves available, covering the years 2008-2014. The background variables are collected on a monthly basis and merged to the core study based on the month when the core study was answered. The data is thus not in terms of calendar years but spanning from April/May of one year to the same months in the year after. The survey of the core study generally has a response rate ranging from approx. 75-80% and thus we have answers from around 5500-6500 persons of which a few are incomplete. Furthermore, the individuals are not always the same as some leave the panel and others join in later waves when refreshment samples are recruited.

I further restrict the sample and only include individuals from 23 up to and including 60 years of age. The reason for this is that there some sectors in the Netherlands where the minimum wage is a function of the worker's age until 23. As a result of this it seems that young workers in these sectors may have a higher risk of becoming unemployed close to their birthdays. In addition to this, students who are finishing their education become also harder to classify, as they may currently hold a (side) job, while studying and, given that April/May is relatively close to the end of the school year, also may see themselves as first time job seekers. Depending on what they choose as the "best" description we may thus classify students that are in the same situation differently. The age limit at 60 years on the other hand stems from the idea that older individuals face a different consideration set than younger workers as an exit from the labour force includes (early) retirement for them which is not a general option for e.g. a 40 year old worker. In addition to this, I also only include household heads and their wedded or unwedded partners in my dataset, excluding other family members, or individuals living in shared houses.

Last, I am also restricted to those individuals for whom I have at least two consecutive observations, because my aim is to estimate a dynamic model. Because I assume the labour state choice to be a dynamic process, and because the estimation of dynamic correlated random effects models has not yet been fully discussed for unbalanced panels, I also decided to exclude any observations of an individual that follow after a period of non-observation, so that the unbalancedness in my panel is of a closer description to the case considered by Albarrán et al. (2015). After all these restrictions, I am left with complete information in the variables I am interested in for 4598 individuals and a total of 15248 observations for them.

Before I proceed to give some descriptive statistics followed by a motivation for the model described in section 4, it should be noted that the LISS panels is actually a rich data set with many variables that may offer many interesting avenues for research. I however restrict myself to a basic set of personal and household characterstics, that are mostly given in the background variables, instead of exploring other variables in the core studies. I do this first because most of the work related questions are unfortunately only asked to the employed, or the unemployed respectively, and don't allow me to have the same variables explaining all labour states. Second, I believe that it is also important to have a relatively simple benchmark model first before e.g. developing a structural model that would take such different variables into account.

3.1. Self-employment identification in the LISS panel

There are two instances within the LISS panel's different core studies where self-employment can be identified. In addition to these two core studies, we can also identify self-employed individuals in the background variables. However, the variables of self-employment that we can construct, while comparable across the different waves within each definition, are not always comparable across all of the three sources.

The background variables are a collection of variables which the household representative has to answer for every household member, even those that do not participate in the LISS panel, and which are updated monthly with each wave. They contain one question where participants are asked about their primary occupation. The answering individual is presented several options of which she is asked to choose the one describing herself or the other person best. Of the 14 options the one of interest to this study is *"Option 3: autonomous professional, freelancer, or self-employed"* which stands in contrast to the employment states *"Option 1: paid employment"* and *"Option 2: works or assists in family business"*. The other options include various states of unemployment or non-participation in the labour force. Thus, there are two issues that follow immediately from this question which one has to keep in mind if one wants to work with the background variables' information: First, based on how the options are presented, it is not clear how "dga's" will answer the question, and neither is it clear where another household member would place a "dga's" from their household. Based on the CBS's survey, it seems likely that a majority of the "dga's" themselves will answer that they are self-employed while a minority will answer with option 1, i.e. that they follow a paid employment. If the "dga" is not the representative that fills out the background survey we may assume that the other person asks them for their opinion but there is no certain way to tell how they would fill out the form. Hence, it is not possible to construct a definition where all "dga's" are treated uniformly if the background variables are used. Second, it is neither clear how individuals will decide between option 2, working in a family business, and self-employment, if they are e.g. the owner of the family business. As CBS (2014) counts both to the self-employed this is however a minor issue for the definition of self-employment in the LISS panel itself.

Second, in the *Work and Schooling* core study there is one question[5] that asks participants about the form of employment which they follow in their principal job. The answer options not only differentiate between permanent and temporary employment contracts, as well as on-call

[5] Question cw121

employees or temp-staffers, but also give four different options of self-employment: "Option 5: self-employed/freelancer", "Option 6: independent professional", "Option 7: director of a limited liability or private limited company", and lastly "Option 8: Majority shareholder director". Furthermore, unlike in the background variables, the instructions given should pre-empt that "dga's" give different answer, as they state that *"[a] majority shareholder director, also, generally receives an income as an employee. Nevertheless, if this [applies] to you, we request that you indicate that you [are] a (majority shareholder) director"*. Within the *Work and Schooling* core study I am thus able to identify all "dga's" uniformly and separately from other forms of self-employment. For directors of a limited liability or private limited company (Dutch: NV or BV, respectively) on the other hand the instructions read: *"A director of a limited liability or private limited company [...] is generally on the payroll of that company. In that case, please enter that you [are an] employee in permanent or temporary employment"*. Hence, it is not clear how a director of a small to medium sized NV or BV, who founded the company himself, and whom we may see as a self-employed individual, would answer this question. One last complication in the *Work and Schooling* core study is that all these questions are asked to all individuals that are employed or were ever employed. To identify the current employees and self-employed we thus have to condition on an individual's current labour state given by a subset of questions[6] that are similar to the 14 options in the background variables except for not including self-employment among them.

Lastly, one can also identify the self-employed based on the core study *Economic Situation: Income*.[7] This core study asks participants about their work situation in the year prior to the one when they fill out the survey. With this question one can identify freelancers, "zzp'ers", company owners, participants in a partnership (both "maatschap" and "VOF"), as well as owners of a private limited liability company or limited partnership, and those making profits/losses through some (not further specified) form of enterprise. Once more "dga's" however pose a problem. In a different question, before the self-employment identifying questions, participants are asked if they received income from paid work in employment in the last year. It further suggests that "dga's" and directors of an NV or BV answer this question in the affirmative by including the statement that *"A managing director of a private or public limited company is (usually) employed by the company as an employee. A majority shareholder director generally receives income as an employee as well"*. As this is only framed as a suggestion and not an instruction it is again not clear how a "dga" who chose that form of business e.g. for fiscal reasons would answer this question, and one cannot be sure that all "dga's" will be treated uniformly in a self-employment definition based on this core study. In addition to the problem concerning "dga's" and directors of NVs or BVs, the core study on income also only allows one to identify the unemployed as those individuals that received

[6] Questions cw088-cw102 — if an individual gives several responses I use cw104 to identify the best description among those. See appendix C for the variable coding used

[7] In the order of discussion the following questions are concerned: ci037-ci045, ci008, ci081-ci101.

(various types of) unemployment benefits, and consequently those not part of the labour force, as those who answered neither of the three in the affirmative. Note also, that individuals may answer in each of the three sets of questions in the affirmative if they are working in several kind of jobs, and were temporarily unemployed in that year. This thus makes this particular set of data harder to use in my model, where I will not allow for part-time employment. Lastly, it should also be kept in mind that the *Economic Situation: Income* study asks the participants for information from a year prior which cannot be matched directly to the personal characteristics and household information that is given for the current year. Thus, if one wants to use this dataset one is left with fewer observations than with the others as not all other control variables used can be computed backwards.

In summary, out of the three instances in the LISS panel where one can identify the self-employed, two—the background variables and the core study on income—don't guarantee a uniform treatment of "dga's" or directors of NVs or BVs. In the background variables we are more likely to include "dga's" in the self-employed whereas in the core study on income it seems more likely that we will exclude them from the self-employed. Directors of NVs or BVs are treated with the same uncertainty in both cases. As CBS (2014) points out that "dga's" have become more numerous in recent years, I believe that it is important to include them among the self-employed. And as the CBS also includes "dga's" in their measure of the self-employed, I therefore choose the *Work and Schooling* core study to construct self-employment indicators, where all "dga's" will be included. Using the *Work and Schooling* study furthermore also allows me to do some additional checks in section 3 as it allows me to exclude them. In the following I will include "dga's" in all the analyses, unless otherwise stated. Lastly, the *Work and Schooling* study further has the advantage over the income study that I will be able to have more observations in my dataset. Unfortunately there is no solution for a clear treatment of directors of NVs or BVs, and neither of the three definitions allows for an unambiguous treatment of family members that work in a small family owned business which can be seen either as being self-employed or an employee.

3.2. Descriptive statistics

The explanatory variables that I will use for the estimations in section 5 are both individual characteristics like age, gender, or education, which have been shown to have some correlation with the choice to be self-employed (CBS, 2014), and household specific variables (whether an individual lives with a partner and has children, as well as the size of the household), which have been found to have explanatory power in regressions (Blanchflower, 2000). In addition to these, I also include an indicator variable that measures whether an individual lives in a self-owned dwelling. The idea behind this variable is that it should be seen as an indirect measure of wealth. Controlling for wealth is of interest as Blanchflower (2000) points out some studies which found that capital endowments are explanatory variables of self-employment. This variable is thus included to try to approximate for this effect. However

Table 1: Means of regression covariates

	employees	self-employed	unemployed	not working	all
Age	44.90	47.20	46.80	48.41	45.76
Female	0.52	0.43	0.55	0.79	0.57
Living with partner	0.78	0.81	0.63	0.78	0.78
Has children	0.55	0.58	0.42	0.48	0.54
Female with partner	0.40	0.36	0.38	0.65	0.44
Female with children	0.29	0.26	0.24	0.42	0.31
Medium education	0.60	0.54	0.63	0.73	0.62
Higher education	0.38	0.44	0.31	0.20	0.35
Self-owned dwelling	0.80	0.83	0.62	0.66	0.77
Number of hh members	2.87	3.08	2.42	2.78	2.85

Source: LISS panel, own calcuations.

its inclusion requires the assumption that it is strictly exogenous, and that house ownership is not endogenous to the choice of labour state. This argument cannot be made for income which is indirectly linked to house ownership as e.g. banks place conditions on income when awarding mortgages. However, as unemployment should be seen as a temporary state choice, and an individual should not be able to stay out of the labour force for a long time, except in case of invalidity, unless they have a substantial wealth to live from, I assume that there is no violation of the exogeneity assumption.

Unfortunately it is not possible to include sector-controls since we do not observe the sectors for (all) the unemployed or those out of the labour force. Such controls would be of interest especially when modelling the choice to be self-employed as there are some sectors (e.g. agriculture or construction) where the share of the self-employed relative to all the employed is particularly high (Beusch, 2015).

Table 1 reports the means of these variables over all time periods.[8] One possible issue with the data that can be seen in the table is that women are slightly over-represented in the sample as they make up 57% of all observations. The average age seems more or less similar across labour states but is slightly lower among employees. In its report the CBS writes that self-employed individuals are more frequent among the old but since the standard deviation (not reported) is similar to the standard deviation of age for the other labour states this observation cannot be made for the self-employed individuals in my data.[9] The rest of the variables seems more or less in line with what the CBS or Blanchflower (2000) reports: The share of individuals living with a partner differs little across the different labour states, although the lower share among the unemployed is a bit unexpected. More interestingly

[8]The characteristics do not vary a lot over time except for the unemployed which consist of a relatively small sample in each period. Hence a larger variation in their values is not very surprising.

[9]Since age enters my function for the indirect utility only in a linear function, this is reassuring. Otherwise, one should enter age as a polynomial function.

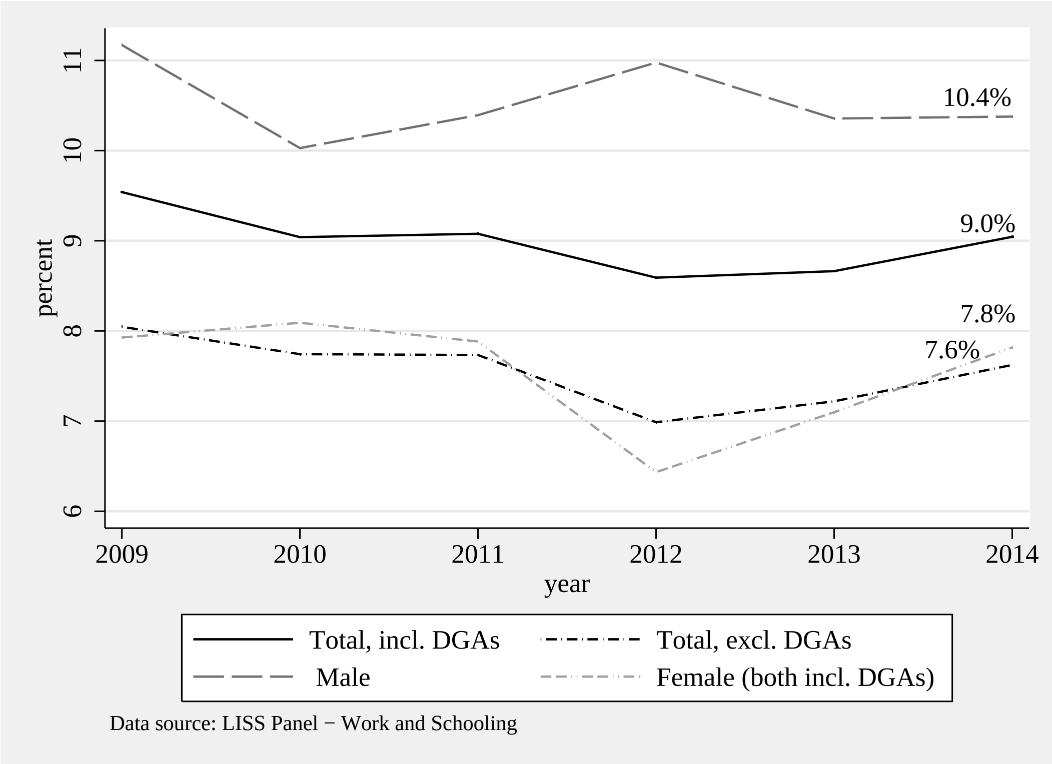

Figure 1: Share of self-employed relative to all employed

though we can see that when interacted with the dummy variable for women, we find among the individuals that choose to not be in the labour force, a higher percentage of women that live with a partner than in the other labour states. Similar results hold for the indicators whether the individual has children, and also once it is interacted with the gender dummy. We can also see that less individuals with a higher education are choosing not to be in the labour force. Furthermore one should note that the dummy variables for medium and higher education cover almost 100% of each labour state's population - since the two indicators covered less of the sample in the LISS panel before my restrictions, this hints at a slight under-representation of individuals with a low level of education in my sample in particular. Last, one should take notice that the indicator for living in a self-owned dwelling shows that large shares of employees and self-employed own their dwelling whereas less of the unemployed or not working individuals do. Since these groups are also separated by the relative size of their incomes, it might thus be that this variable suffers from endogeneity problems. However, these possible problems will be ignored in the remainder of this paper.

In correspondence with CBS (2014) I am interested to see how the share of the self-employed develops over time in the data that I use. Figure 1 shows the development of the share of the self-employed relative to all employed individuals in my data. First, we can see that whether we include "dga's" or not does not change the general trend in the share of the self-employed in the LISS panel, and only shifts the line upwards by a more or less fixed amount of percentage points.[10] Hence we do not directly observe a downward trend for "dga's" since 2009 as reported by the CBS. Second we can also not see any general upward trend in the share of self-employed—if anything we find a slightly downward sloping, or U-shaped development—which also stands in contrast to the CBS' findings. The magnitude of the self-employment shares we find also hint at a very strong under-representation of self-employed individuals in the LISS panel. Using CBS data Beusch (2015) reports shares of 16.6% for the total population and 17.2% for males and 16.1% for females respectively in 2013. For women the actual share of the self-employed is thus almost twice as high as in my data. In terms of the development between genders one can see that the share of self-employed is changing differently for men and women. It is decreasing substantially for women in 2012 while it is increasing for men in the same year. Furthermore the differences in terms of percentage points between the two genders are also larger in the LISS panel than in the CBS data in Beusch (2015).

These are unfortunately not the only problems that I find when we look at the descriptive statistics of my data. One additional problem in the data concerns the unemployment rate calculated for my sample. It is at least one percentage point too low for the years 2009-2012, and even two percentage points too low for the last two years.[11] Therefore either the LISS panel under-represents the unemployed, or my definition[12] of labour states based on the answers given in the *Work and Schooling* core study is underestimating them.

3.3. Motivation

A strong motivation behind estimating a multinomial logit model is to avoid possible sample selection bias, which may occur if only binary estimates are considered. But also when looking at the data, one can find some motivation for a multinomial logit model, as well as for the dynamic model and why unobserved heterogeneity should be included too.

When estimating a multinomial logit model, the implicit assumption is that individuals are able to choose from all states. In our data and set-up this is allowed, as unemployment is not defined based on receiving unemployment benefits—in which case individuals that were not working in the previous period could only enter the labour force through employment or self-employment—but based on answers to questions that also ask whether the participant is currently looking for work. Table 2 and 3 show the transition probabilities observed in the dataset. The first thing to note from table 3 is that the transition probabilities are different for men and women. This has also been pointed out for a sample of 50 to 63 year old in Been and Knoef (2013). Nevertheless this paper estimates the model jointly for men and women but acknowledges that there may be room for improvement with regards to this. Second

[10] It also does this if we calculate the shares by gender.

[11] The shares of the different labour states corresponding to the years in figure 1 are shown in table 11 in appendix B. I compare them to EUROSTAT unemployment rates for the 25 to 74 year olds, extracted on July 3, 2015.

[12] I tried other definitions that take more factors into account but could not improve the share of the unemployed using these - hence I decided to use the most parsimonous.

Table 2: Observed transition probabilities (in %)

Chosen labour state past \ current	all 0	1	2	3
0: employee	94.22	0.68	1.56	3.54
1: self-employed	4.38	88.49	1.05	6.09
2: unemployed	29.95	5.08	35.56	29.41
3: not in labour force	12.16	2.22	3.28	82.34
overall	71.67	7.12	2.67	18.55

Based on 15248 observations pairs.

Source: LISS Panel, own calculations.

Table 3: Observed transition probabilities by gender (in %)

Chosen labour state	male				female			
past \ current	0	1	2	3	0	1	2	3
0: employee	95.75	0.81	1.70	1.74	92.83	0.55	1.43	5.19
1: self-employed	3.37	91.74	1.18	3.71	5.68	84.28	0.87	9.17
2: unemployed	32.93	5.39	35.93	25.75	27.54	4.83	35.27	32.37
3: not in labour force	10.56	3.46	4.92	81.06	12.56	1.91	2.87	82.67
overall	78.78	9.32	2.79	9.12	66.24	5.44	2.58	25.74

Based on 6601 (male) and 8647 (female) observation pairs.

Source: LISS Panel, own calculations.

we can see that except for the unemployed the probability that an individual is observed in the same labour state in the next period is at least 80% and higher for employees and the self-employed than those choosing to not participate in the labour force. Finally one can see that the probability is higher for those in unemployment and those not in the labour force to choose self-employment in the next period than for employees.

The transitions matrices however do not allow us to distinguish between Heckman (1981b)'s true and spurious state dependence. If we want to distinguish between the two it is necessary to not just model observed heterogeneity but one also has to take the unobserved heterogeneity into account. That unobserved heterogeneity is likely to exist is also supported by the results taken from the International Social Survey Programme's Work Orientation module $I - III$ (ISSP Research Group, 1991, 1999, 2013). In this survey one question asks the participants whether they would choose employment or self-employment if they could choose freely between different kinds of jobs. Results based on this question are presented in table 4. There are two main points that should be taken away from this table. First, not all individuals answer the same. While this may seem obvious it is nevertheless an indication that different people have different tastes. If we do not consider this unobserved heterogeneity,

Table 4: Preference for self-employment by groups

	1989		1997		2005	
Groups	%	N	%	N	%	N
All	34.0	1690	31.9	2267	28.6	925
Employed	34.2	684	33.2	1021	28.9	530
Unemployed	35.9	39	31.3	67	36.4	33
Self-employed	80.8	125	n.a.	n.a.	85.7	77

Note: N denotes the total number of persons per category.
Source: ISSP Work Orientation module, own calculations.

we may thus risk biased estimates. Second we can also see that in particular in 2005 the employed and unemployed have different preferences for self-employment. This in turn may hint at correlations between preferences for different labour states. I.e. individuals who have chosen employment instead of not participating in the labour force may also be more likely to prefer self-employment. Or an unemployed person may rather try to set-up an own business than go out of the labour force.

4. Model

This section first presents the model framework chosen and in a second part the estimation procedure.

4.1. Dynamic multinomial model of labour states

This subsection discusses the dynamic multinomial logit model with correlated random effects, which I use to explain each individual's labour state choice, in detail. This type of model is also more generally known as a mixed logit model. Mixed logit models have the advantage that they are highly flexible and able to approximate any random utility model. And, unlike the standard (multinomial) logit model they allow for random taste variation, unrestricted substitution patterns, and correlation in unobserved factors over time. Furthermore, the lagged dependent variables can be added, unlike in probit models, without changing the estimation procedure.(Train, 2009, chapter 6)

The labour states are assumed to be exclusive, i.e. the model abstracts from part-time employment, assuming that all individuals work full time whenever they are choosing a working state. This may seem like a strong assumption considering that almost half of the population in the Netherlands was working part-time in 2014[13] but when we look at the self-reported hours worked for individuals in our panel, we find a median value of 34, respectively

[13]See eurostat, statistics explained: employment statistics (retrieved October 15, 2015).

35 hours[14] for employees and the self-employed. Thus, this simplifying assumption should nevertheless still fit our data relatively well.

Throughout this paper it is further assumed that any individual can choose to be in any of the four states at time t: employment ($j = 0$), self-employment ($j = 1$), unemployment ($j = 2$), or to not participate in the labour force ($j = 3$). Admittedly this may be seen as a strong assumption since unemployment is in most cases not a state that individuals choose. However, since we can observe that some individuals enter/exit self-employment from/into unemployment, this is an attempt to include these observations in the estimation, and a structural model of labour state choice is left for future research.

The model is hence similar to the first-order Markov model proposed by Heckman (1981b). Like in Heckmans's model, the model used in this paper also includes past labour states as explanatory variables to control for dynamic effects. Habit persistence on the other hand is estimated differently from Heckman's approach. Here individual random effects are included as controls for unobserved characteristics. As such my model also allows to distinguish between what Heckman calls "true" structural state dependence and "spurious" state dependence, i.e. unobserved heterogeneity. I furthermore allow for correlation between the random effects, in which I closely follow the simulated maximum likelihood procedures described in Train (2009). The estimation approach I take is neither new to the literature: e.g. both Gong et al. (2004) and Been and Knoef (2013) use similar models.

I model the observed choice of a labour state as the outcome of a maximisation process. Each individual re-evaluates her state every period, and chooses whichever labour state j that maximises her utility for that period. In terms of the econometric specification I thus consider the discrete choice model where an individual derives indirect utility y^*_{ijt} from state j at time t. That is

$$y_{ijt} = \begin{cases} 1 & \text{if } y^*_{ijt} > y^*_{ikt} \qquad \text{for } j,k = 0,1,2,3; j \neq k; i = 1,\ldots,N; t = 2,\ldots,T \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

and $y_{it} = (y_{i0t} \ldots y_{i3t})$ is a column vector[15] summarising an individual's choice.

Indirect utility y^*_{ijt} from choosing state j is assumed to be latent and in its reduced form determined by

$$y^*_{ijt} = y_{it-1}\gamma_j + X_{it}\beta_j + \alpha_{ij} + \epsilon_{ijt} \tag{2}$$

[14]These values are calculated using only those individuals that indicate up to 55 hours of work. This leaves us with 10742 (instead of 10927) observations for the employees and 927 (instead of 1084) for the self-employed. The corresponding means are slightly lower at approximately 30 and 31.5 hours, with standard deviations of 14 and 15 respectively.

[15]The notation in this model will make use of column vectors, where one might be more used to row vectors. The choice is not too important but keeping the notation here in column vectors allows to read the Matlab code for the estimation more easily. When not otherwise mentioned, vectors are assumed to be columns vectors.

where y_{it-1} is a vector that describes the labour state choice from the last period, X_{it} is a vector of k observed exogenous variables, and the coefficient vectors γ_j and β_j are to be estimated. The variables included in X_{it} are individual as well as household characteristics that are assumed to have an influence on indirect utility y^*_{ijt}. These variables have been discussed in section 3.2. Next to these covariates X_{it} also include controls for time effects.

It should also be noted that the model in this paper abstracts from any higher order learning on the job. That is apart from the first lag of past labour state choices, no further lags enter the model. Because of a lack of observed data, I neither include a measure of job tenure. The model thus does not take into account that an individual that has been self-employed for several years may be more likely to remain so when compared to an individual that has been self-employed for one year only.

The parameter α_{ij} is a random effect that reflects time-invariant[16] unobserved heterogeneity across individuals and ϵ_{ijt} is and identically and independently distributed error term. Furthermore, ϵ_{ijt} is assumed to be independent of X_{it} and α_{ij} and drawn from a Type 1 extreme value distribution, from which the logit model follows.

4.1.1. The initial conditions problem

Because equation (2) includes the lagged dependent variables, an initial conditions problem as described by Heckman (1981a) arises. First, we only have incomplete measures of individuals' first labour state choice in the LISS panel and are therefore not able to use them.[17] Our measure of initial conditions is therefore given by the first observation that we have of an individual. Second, and more importantly, it is likely wrong to assume that the first choice, which we observe, y_{i0} is independent of the unobserved random effects $(\alpha_{i0}, \ldots, \alpha_{i3})$.[18] Here, I solve for the initial conditions following the approach suggested by Wooldridge (2005). I thus specify a parametric model for the density of α_i conditional on the initial observation of the dependent variable y_{i0} and α_{ij} can be written as

$$\alpha_{ij} = y_{i0}\delta_j + \mu_{ij} \tag{3}$$

which amounts to including y_{i0} as an additional row vector of regressors when estimating the model.

This approach to handle the initial conditions problem was however constructed for bal-

[16] Hence, also the approach to model individual preferences abstracts from any higher order process of habit formation.

[17] There is one question (cw123) in the "Work and Schooling" core study that ask individuals in what type of organisation they worked in their very first job. One answer option is "self-employed" but unfortunately the structure of the question does not allow us to identify individuals who were unemployed or chose to not work after the obligatory years of schooling.

[18] Note that even if we were to observe an individuals very first choice of labour state after the end of obligatory schooling, this choice would likely neither be independent of α_i.

anced panels whereas my estimation is done on an unbalanced panel.[19] Wooldridge (2005, p.44) writes that if the sequence of observations of the dependent variable and the sample selection mechanism are independent conditional on the initial conditions and the exogenous variables, the maximum likelihood estimation using the balanced sub-panel will be consistent. The LISS panel however is not only unbalanced because of sample attrition but also because some individuals only join in later waves (to make up for those that have left). If I were to restrict myself to the biggest balanced sub-panel, assuming that the necessary conditions hold, I would thus lose more than half of my observations, and be left with a bit more than 25% of all individuals. Restricting the estimation to the balanced subpanel may thus imply important loss in efficiency.

Wooldridge (2009) proposes strategies for correlated random effects models with unbalanced panels but restricts himself to non-dynamic models, which can neither be applied directly to a dynamic modelling approach. The only study that considers the problem of estimating dynamic non-linear random effects models with unbalanced panels is by Albarrán et al. (2015). The authors show that even if the selection mechanism is completely at random unbalancedness can lead to inconsistent estimation in dynamic models and more conditions have to be satisfied for consistent estimation. In particular it is necessary that the process generating y_{i0_i} is in the steady state, and that the sequence in which periods an individual is observed, is independent from the shocks to the initial conditions. As a solution the authors suggest that one estimates parameters that are specific to each sub-panel. This approach is however not feasible in a multinomial logit model with choice-invariant variables as the number of parameters would become too large. Thus I only include one small set of additional covariates in the initial conditions equation to control for the year in which an individual was first observed, and no interaction terms. δ thus becomes a $8 \times J$ coefficient matrix to be estimated.[20]

4.1.2. Correlation in the random effects

In a last step I allow for correlation in the random effects. Like Gong et al. (2004) and Been and Knoef (2013) I assume that μ_i is drawn from a J-dimensional multivariate normal distribution with mean zero and covariance W.

Following Train (2009, chapter 9) I use a Choleski transformation for the multivariate normals. As Train (p.238) writes, the advantage of using this approach is that "for any pattern of covariance, there is some set of loadings from independent components that reproduces that covariance". I thus only have to make an assumption concerning the distribution of the unobserved heterogeneity but not of the covariance. Hence,

$$\mu_i = \xi_i' L' \tag{4}$$

[19] See table 12 in appendix B for su-panel patterns in the data.

[20] We have $J-1 = 3$ states in the initial conditions, as well as six periods in which we can observe an individual for the first time, of which we drop one in order to identify all the parameters.

where ξ_i is a $J \times 1$ vector of independent standard normal distributed variables, and L is the lower triangular Cholesky factorization of μ_i's covariance matrix W such that $LL' = W$.

Note that by allowing μ to be multivariate the independence of irrelevant alternatives assumption is no longer imposed. The estimates of the covariances will give an indication whether individuals who prefer one labour state are also more likely to prefer a particular other labour state. E.g. if the covariance for state 1 and 2 (being self-employed or unemployed) is positive, we should expect an individual, ceteris paribus, to have a higher probability of choosing self-employed when it has a high individual parameter for unemployment (ξ_2).

Substituting (4) and (3) in (2), and writing the utilities in vectorised form we have

$$y_{it}^* = y_{it-1}\gamma + y_{i0}\delta + X_{it}\beta + \xi_i' L' + \epsilon_{it} \qquad i = 1, \ldots, N \ ; \ t = 1, \ldots, T \tag{5}$$

where y_{it}^* is a $1 \times J$ vector of indirect utilities for individual i at time t, γ and δ are $J \times J$ matrix of parameters, β is a $k \times J$ matrix of parameters, and L contains the parameters of the covariance structure. All elements of β, γ δ, and L have to be estimated.

4.1.3. Likelihood function

The probability to observe a particular individual choosing labour state j at time t conditional on ξ_i in the multinomial logit model is then given by

$$Prob(y_{it} = j | X_{it}, y_{it-1}, y_{i0}, \xi_i) = \frac{exp(y_{it-1}\gamma_j + y_{i0}\delta_j + X_{it}\beta_j + \xi_i' L_j')}{\sum_{k=0}^{J} exp(y_{it-1}\gamma_k + y_{i0}\delta_k + X_{it}\beta_k + \xi_i' L_k')} \tag{6}$$

where L_j is the j^{th} row of L.

It follows that the conditional probability of observing a sequence of choices for individual i is

$$Prob(y_i | X_{it}, y_{it-1}, y_{i0}, \xi_i) = \prod_t \prod_j Prob(y_{it} = j | X_{it}, y_{it-1}, y_{i0}, , \xi_i)^{\mathbb{D}_{ijt}} \tag{7}$$

where $\mathbb{D}_{ijt}$ is an indicator function denoting whether state j is chosen by the individual. The unconditional probability, or likelihood function, is then given by

$$Prob(y_i | X_{it}, y_{it-1}, y_{i0}) = \int_{\xi_i} Prob(y_i | X_{it}, y_{it-1}, y_{i0}, \xi_i) f(\xi_i) d\xi_i \tag{8}$$

where $f(\xi_i)$ denotes the multivariate distribution of ξ_i. The log likelihood function to be estimated is thus

$$log\mathcal{L} = \sum_{i=1}^{N} logProb(y_i | X_{it}, y_{it-1}, y_{i0}) \tag{9}$$

4.1.4. Identification

In its current form the multinomial logit model described by equations (6) – (9) is not identified as there are too many parameters. For identification purposes I take $j = 0$ as the base category. β_0, γ_0, δ_0 and also the first column in L are normalised to zero, and alternatives $j = 1, 2, 3$ are estimated relative to the base category.[21] E.g. the covariance matrix is thus given by

$$W = \begin{pmatrix} l_{11}^2 & & \\ l_{11}l_{21} & l_{21}^2 + l_{22}^2 & \\ l_{11}l_{31} & l_{21}l_{31} + l_{22}l_{32} & l_{31}^2 + l_{32}^2 + l_{33}^2 \end{pmatrix}$$

and I estimate $J - 1 = 3$ sets of coefficients for each variable.

4.2. Maximum Simulated Likelihood

In order to estimate the model the probabilities given by equation (8), i.e. $\int_{\xi_i}$ in particular, have to be simulated. As Train (2009, chapter 10) writes, $log\hat{P}$ is not an unbiased estimator for $logP$ because of the non-linear log operation, even if $\hat{P}$ is an unbiased estimator of P. Thus the bias in the simulator for $logProb(y_i|X_{it}, y_{it-1}, y_{i0})$ translates into bias in the maximum simulated likelihood (MSL) estimator. This bias however diminishes as more draws are used in the simulation. Using a large number of draws for ξ_i on the other hand increases the computational burden of the estimation. One way to lighten this burden is, to instead of independent random draws, use an alternative method that provides better coverage of the domain and therefore greater accuracy for a given number of draws. One method that fulfils these requirements are Halton draws. Both Bhat (2001) and Train (2009, chapter 9.3.3) showed that e.g. 100 Halton draws can provide more precise results than 1000 random draws.

In order to simulate $\int_{\xi_i}$ I therefore take 150 draws for each individual from a $J - 1$-dimensional Halton sequence in which I follow closely the method described in Train (2009, chapter 9.3.3). Based on the discussion in the paragraph above, 150 draws should be sufficient, also because the panel is based on more than four thousand individuals, i.e. large in itself. In addition I also randomise the Halton draws following the procedure described by Bhat (2003). As a base of the Halton sequences I use the vector of primes given by $[3, 13, 7]$. This may be a mistake because as (Haan and Uhlendorff, 2006, p.233) mention"the number of draws should not be an integer multiple of any of the primes used" but clearly 150 is a integer multiple of 3. All the results and the findings in subsection 5.2 should therefore possibly seen with precaution.[22]

[21] Note that dimensions are now reduced by 1, and all vectors and matrices are now of dimensions $(\cdot \times J - 1)$.

[22] This mistake was found too late to be corrected for this version of the paper. Interestingly enough Train (2009, chapter 9.3.3) made the same mistake by using primes $[2, 3, 5, 7, 11]$ on 125 draws. As his results are relatively stable across permutations this mistake may possibly not be all too grave.

Returning to the model, the probability of observing an individual's observed sequence of labour state choices in the simulation is given by

$$Prob(y_i|X_{it}, y_{it-1}, y_{i0}) = \frac{1}{R}\sum_{r=1}^{R}\prod_{t=2}^{T}\left(\prod_{j=1}^{J} Prob(y_i|X_{it}, y_{it-1}, y_{i0}, \xi_i)^{\mathbb{D}_{ijt}}\right)^{\mathbb{S}_{it}}$$

The simulated loglikelihood function is thus given by

$$log\mathcal{SL} = \sum_{i=1}^{N} log\left[\frac{1}{R}\sum_{r=1}^{R}\prod_{t=2}^{T}\left(\prod_{j=1}^{J} Prob(y_i|X_{it}, y_{it-1}, y_{i0}, \xi_i)^{\mathbb{D}_{ijt}}\right)^{\mathbb{S}_{it}}\right] \tag{10}$$

where R is the number of Halton draws taken to simulate $\int_{\xi_i}$, and $\mathbb{S}_{ijt}$ is an indicator function to control for the unbalanced panel, denoting whether an individual's observation enters in the estimation.

All model specifications for the correlated random effects models reported in section 5 are estimated using an own code written in Matlab 2014a. I solve them as an unconstrained minimisation[23] problem using KNITRO as a solver and supplying the gradient as defined in appendix A. As the best guess for the starting values I first estimate a pooled multinomial regression on the same covariates, including the lags, for each specification. The only difference is the non-inclusion of the unobserved individual heterogeneity, and I make use of the function "mnrfit" in Matlab's statistics toolbox instead of my own code. If ignoring the unobserved heterogeneity leads to a large biase in the estimates these starting values may be far off. This in turn could pose a problem if the empirical function is very flat or "bumpy", in which case the solver may not find the minimum. The use of different starting values is discussed in section 5.2.

5. Results

5.1. Estimation results

As a benchmark I first estimate static multinomial panel regressions with correlated random effects which are reported in table 8. Both regressions also include time effects for the years 2010-2014, which are only significant in the equation for the unemployed. The only difference in the regressions shown in table 8 is that model A uses a smaller set of covariates, excluding the information on whether an individual is living with a partner, and whether he or she has children. In both of the models most coefficients are statistically significant at the 5%, or even 1% confidence level. Interestingly the coefficient for the gender dummy variable is no longer statistically significant once the additional covariates enter the regression. As most coefficients change once the additional characteristics are included in model C it seems

[23]Matlab only solves minimisation problems and therefore I estimate the negative log-likelihood function.

important to include these interaction terms. In line with this are also the results from the likelihood ratio test comparing the two models. Based on the calculated test-statistic we reject the null hypothesis that the parameters are the same. I.e. we reject the restricted model B in favour of the unrestricted model C. An interpretation of the coefficients is not straight-forward in the multinomial framework. Hence the focus will be on odds-ratios for the moment. In model C the interpretation of those ratios actually fits some theories as will be discussed later. Based on the odds ratios we find that a woman who lives alone without a partner and children is, ceteris paribus, less likely to choose self-employment compared to any other labour state. She is approximately 0.89 times as likely to be self-employed than employed, 0.66 times as likely to be self-employed than unemployed, and approximately four times more likely to remain out of the labour force than to be self-employed. The picture is different however for a woman who lives with a partner and one child in a three persons household: she is almost twice as likely to be self-employed than employed or unemployed, and a bit more than 8 times less likely to be self-employed than not participating in the labour force. For men on the other hand the picture is almost the opposite: A single man living alone is more likely to choose self-employment relative to any other labour state with a factor ranging from 1.6 (being out of the labour force) to approximately 3 (wage employment). But if he is living with a partner and a child, he is only 0.75 times as likely to choose self-employment over employment, whereas he is almost 5, respectively 17 times more likely to choose self-employment over unemployment or being out of the labour force. This would thus give support to a hypothesis where women choose self-employment as a form of more flexible employment that still allows looking after children by e.g. working from home, whereas men may choose self-employment more out of necessity to avoid unemployment. This is pure speculation though and more work would have to be done to fully understand choices behind self-employment.

In a next step dynamic factors, i.e. lagged dependent variables and the initial conditions, are added to the regression model. Model *D.1* in table 9 reports the results for the extension of model C.[24] First, by comparing the values of the log-likelihood we can directly see without having to calculate a likelihood ratio test that including the dynamic factors improves the model a lot, and we should choose model *D.1* over model C. Second, we can also see that the lagged dependent variables enter highly statistically significant in the model—only the lag for the past unemployment indicator in the unemployment equation is significant at the 10% significance-level whereas all others are significant at the 1% level. Moreover, we can see that the own lag in each labour state increases the probability to be in that state the most when compared to the other states. Similar results hold for the initial states observed in the data,

[24] The results for the extension of model B are reported in table 14 in appendix B. Since the coefficients are similar between the "small" and "large" model, and moreover because the likelihood ratio tests suggest that the smaller models are always rejected in favour of those with a larger set of covariates in X, the results in table 14 are not discussed. Note that also a model with only dynamic factors and time effects was estimated (see table 13 in appendix B) which is rejected in favour of model D in the likelihood ratio test, but is preferred to model C based on Bayes' information criterion.

except for unemployment. This is not surprising, considering that we should not have too many long-term unemployed individuals in the data set, and initial unemployment is thus less likely an explanatory factor of unemployment two or more periods later. Furthermore, the initial conditions time effects, which were added in order to control for possible differences in the observation starting periods in the unbalanced panel, are all statistically insignificant. A likelihood ratio test also fails to reject the null hypothesis and hence we cannot reject the null that the restricted model in *D.2* without such initial conditions time effects is better. The discussion will thus focus on model *D.2* in the remainder of this section.

What is also interesting to note is the difference in the estimation of the Cholesky decomposition of the covariance-variance matrix across model *C* and *D.2*. We can see that in the static model a large variance and covariance in the unobserved heterogeneity is needed to fit the data, in particular for the self-employed. However, once the lagged dependent variable and initial conditions are included in the regression the Cholesky loadings are estimated to be much smaller. Furthermore the implied covariance-variance matrix in the static model showed a high covariance between the preference to be self-employed and not in the labour force whereas the dynamic model suggests a very low covariance for the same. In the dynamic model it is only the covariance between unemployment preference and self-employment or not being in the labour force that are relatively high. The covariances are of a similar magnitude as the variance needed in the unobserved unemployment preferences to fit the model to the data. Based on these estimates we should expect to see an individual with a high preference to be unemployed to have a higher probability to choose self-employment or going out of the labour force. Overall, the estimates of the Cholesky factorisation suggest that true state dependence is of importance in the model, but also that unobserved heterogeneity should be taken into account to differentiate true from spurious state dependence. This is further illustrated if we take a look at the starting values (not reported) in which the lagged dependent variable's coefficient is larger than in our final estimate. I.e. without taking unobserved heterogeneity into account we would overestimate state dependence and the coefficient would capture both true and spurious state dependence.

Furthermore one can see in table 9 that few of the coefficients are still statistically significant in the dynamic model. We find that if coefficients have any statistically significant explanatory power then in explaining the choice to be out of the labour force, or to be unemployed. Except for the indicator variable for living in a self-owned dwelling none of the individual or household characteristics enter self-employment significantly. If we calculate the same odds-ratios as earlier, with the additional assumption that the respective individuals are currently employees and also were employees when we initially observed them, we also find different results. A single woman living on her own without a child is twice as likely to be an employee than self employed or not in the labour force, i.e. has an odds-ratio of approximately 0.5 in both cases, and 0.65 times more likely to be self-employed than unemployed. If she however is living in a household with a partner and child, the odds-ratio for her to be self-employed is a bit higher (0.7) and she is a bit less likely to choose self-employment over not being in the labour force compared to a single woman (odds-ratio 0.43). But she is on the other hand about 1.6 times more likely to choose self-employment over unemployment. For men we find that a single man is slightly less likely to choose to be self-employed versus any other labour state, the odds being the lowest at 0.87 in comparison to being out of the labour force. And a man with a small family is always more likely to choose self-employment but the odds-ratios are higher in the cases where self-employment is compared to unemployment (3.3) or not working (4.2) than compared to being an employee (1.7). Note that in this model we can no longer use the same interpretation as earlier in the static model. Not because the results changed but because the odds-ratios are conditional on past and initial labour state choices. For an individual of either gender that was self-employed in the last period we would e.g. find odds-ratios that are more in favour of self-employment. And we can already see from the size of the coefficients for the individual and household characteristics relative to the lagged and initial values that the latter two, and especially the initial values, will have more influence on the choice probabilities than most of the variables contained in X_i.

5.2. Robustness checks

5.2.1. Starting values

To check whether the estimates are robust with respect to the chosen starting values, model *B.1*[25] is estimated for different vectors of starting values. The starting values for this robustness test are generated by taking the original starting values and subtracting or adding up to 0.5 in increments of 0.1 from each value, as well as adding 1 to 4 in increments of 1. The changes in the starting values were selected without any a priori reflection on their meaning for the estimates, and it turns out that the log-likelihood function is no longer defined if either 3 or 4 is added to all original starting values. For all other changes, except for the case where 2 is added, the solver finds the same solution on average[26]. When 2 is added a different solution is however found. Table 15 in appendix B reports the differences in the estimates. What should first be noted is that this second solution has a slightly higher function value which hence suggests that original solution is not the global maximum of the function. It further suggests that the log-likelihood function might be quite flat, and the solver might have gotten stuck in a local maximum. Furthermore we can see in table 15 that the coefficients differ mostly for the lagged variables as well as the initial labour states, and the elements of the Cholesky decomposition. More work is thus required on the estimates and the results discussed above should be taken with precaution.

[25]This choice was made because of the computational time needed.

[26]The standard deviation across these estimates is very small, differing from zero at most on the fourth decimal.

5.2.2. Halton draws

As Train (2009, chapter 9.3.3) points out, it is not yet completely understood how well Halton draws work precisely in simulation-based estimation and caution should be applied. He also discusses some anomalies that arise in his MSL estimation using Halton draws. In particular he finds that results may change—although they don't do so substantially in his estimation—when different permutations of the primes are used. Additionally the results may also be influenced when one uses a different set of primes for the dimensions, or also if different numbers of Halton draws are used. I therefore perform three robustness checks related to Halton draws.

First, I therefore estimate the model *D.1* using different numbers of draws starting from 50, and increasing by increments of 50 until 300 draws. If the chosen amount of 150 draws is sufficient one should only find variation in the estimates using less draws and the results based on a higher number of draws should not differ from those in our estimation with 150 draws. Unsurprisingly, the estimates using only 50 or 100 Halton draws differ from those with 150. However, the estimated coefficients when using 200 Halton draws differ from ours, suggesting that 150 draws may not be enough. Furthermore, the results from this robustness check show that the estimated coefficients do not seem to converge to a stable solution even when the Halton draws are increased to 300.[27] Table 16 in appendix B reports the means and standard deviations across the coefficients based on 200-300 Halton draws. It should be noted though that while the differences in the estimates are too big to be considered to be numerical errors, most of the original estimates still lie within one standard deviation of the means reported in table 16, and while they should be considered with precaution, they may not be completely off, although we cannot tell in which direction they are biased.

In order to see if my estimates suffer from the same problems as Train's, I next first change the primes used to generate the Halton draws by permuting their order. Table 17 reports the means and standard deviations across these results for model *B.1*[28]. In a last step I also choose different primes and re-estimate the same model for all their permutations. Table 18 reports these results similarly to table 17. It is especially the outcome of these two last robustness checks that is worrisome because as one can see from both tables the results vary a lot for different primes and/or their respective order. The results not just indicate very different means when compared to our original results in table 14 but also a relatively large variation across the estimated coefficients.

Summarising these robustness checks, I find that the results are in particular unstable with regards to the primes that are chosen as the basis of the Halton draws. It should be noted, as was mentioned in section 4.2, that a mistake was made concerning the primes chosen initially.

[27] Considering that it took 17 hours to solve the model with 300 Halton draws compared to 1 hour for 250 draws, using an even larger amount of draws may not be computationally feasible.

[28] A smaller model was chosen here to estimate all models to save computational time. Note that if the Halton draws were to exhibit their normal properties, it should not matter whether we are estimating the correct, or a misspecified model as they should not influence the estimation results at all.

Table 5: Simulated transition probabilities (in %)

Chosen labour state	all			
past \ current	0	1	2	3
0: employee	94.75	0.46	1.36	3.43
1: self-employed	2.60	91.01	0.81	5.58
2: unemployed	27.85	3.80	46.98	21.36
3: not in labour force	10.49	1.84	2.73	84.94
overall	71.32	6.72	2.59	19.37

1000 simulations for all 4598 individuals over unbalanced horizons.
Source: LISS Panel, own calculations.

Hence it is not clear if the results are not robust because of this initial mistake or if more work is needed to fully understand how Halton draws can be used within the context of my estimation.

5.3. Simulation

To simulate the transition probabilities in the data, I first take one draw of μ_i for each individual from the multivariate normal distribution with mean zero and variance W, the latter implied by the estimated Cholesky loadings in L. In a next step I forward simulate each individuals labour state choice starting from the first period the individual enters the panel for all the remaining, up to five periods. For this I use the observed covariates in X, as well as the observed initial conditions, and I draw error terms from a type 1 extreme value distribution in addition. All of this is done for a thousand repetitions over the draws for each ϵ_{ijt} and always for the same draw for μ_i. The transition probabilities are then calculated over all simulated periods, and then averaged across repetitions. The results from the simulation are shown in table 5.

When we compare the simulation results over the whole sample in table 5 with the observed transition probabilities in table 2 we see that the probabilities are similar for employees and those not in the labour force, though slightly biased towards state-dependence for the latter, but rather off for the self-employed and unemployed. For the latter two the state-dependence is quite overestimated. One can see that in contrast to this the transitions from self-employment to wage-employment are underestimated, and also the exits from self-employment into unemployment or out of the labour force are slightly underestimated. Similarly exit into self-employment is underestimated for unemployment transitions, and the same holds for exit from the labour force following unemployment. Especially the latter is strongly underestimated.

In order to better understand the coefficients discussed earlier in this section table 6 and 7

Table 6: Simulated transition probabilities: single, living alone

Chosen labour state	male				female			
past \ current	0	1	2	3	0	1	2	3
0: employee	95.68	0.27	1.76	2.30	96.06	0.15	1.49	2.30
1: self-employed	8.09	83.52	1.96	6.42	9.81	79.95	2.06	8.17
2: unemployed	38.47	3.68	46.96	10.90	40.67	2.55	45.28	11.50
3: not in LF	28.12	1.84	5.34	64.70	29.17	1.29	4.71	64.84
overall	93.25	2.54	1.62	2.59	89.89	0.98	1.22	7.91

1000 simulations over 1000 Halton draws for one representative individual.
44 years of age, owns dwelling, lives with partner and child, was employed in 2008.

show simulated transition probabilities[29] for the already discussed cases. The other covariates (education, age, whether the dwelling is self-owned, etc.) are chosen to fit the averages described in table 1. Table 6 shows the simulated transition probabilities for a single individual in a one person household, who owns the dwelling they live in and is 44 years of age at the beginning of the observations, in wage-employment and has a mid-level education. In table 7 the results are presented for a similar individual that in contrast lives in a three-persons household with their partner and child.

The first thing to notice is that these transition probabilities differ from the sample average with respect to the transitions for an individual that is not participating in the labour force. We see that for all of the constructed example the individuals are less likely to remain in the state and have a much higher probability to be (wage) employed in the next period. With respect to self-employment we can see from the two tables is that these men have overall a higher probability to choose self-employment across all types of past labour states when compared to women. And we also find that for the chosen case, if the individual has a small household they also have a higher probability to choose self-employment in the next period given any of the four states. Lastly, we can see that men with a small family in our simulation setting have a particular high probability to transition into self-employment when they are unemployed.

6. Discussion

The results presented in section 5 still leave many points of discussion open.

First, before anything else the use of Halton draws should be examined more closely to find out if they have been used in a wrong way, and whether the mistake in the choice of primes is the reason behind the results found in the robustness checks.

Second, as it was noted in section 3.2 the transition probabilities in the sample differ between women and men. It is also an unlikely assumption that women and men behave

[29] The simulation is similar to the one for the population but in addition we take 1000 different Halton draws for the constructed individual over which we do 1000 simulations for the error terms.

Table 7: Simulated transition probabilities: living with partner and child

Chosen labour state	male				female			
past \ current	0	1	2	3	0	1	2	3
0: employee	97.18	0.60	1.06	1.16	95.81	0.23	0.76	3.20
1: self-employed	8.72	86.96	0.90	3.42	9.22	80.27	0.91	9.60
2: unemployed	47.07	7.11	38.89	6.93	43.01	4.03	36.23	16.73
3: not in LF	33.37	4.39	3.85	58.39	27.74	1.56	2.32	68.38
overall	89.99	1.29	3.07	5.65	91.15	0.69	2.57	5.59

1000 simulations over 1000 Halton draws for one representative individual.

44 years of age, owns dwelling, lives with partner and child, was employed in 2008.

in the same manner on the labour market. One should thus consider estimating different parameters for each gender, e.g. by splitting the sample by gender. This may also help in case that women are indeed over-represented in the sample, instead of introducing weights.

Third, it has not been discussed how age may influence the choice to be self-employed. Considering that CBS (2014) writes that older individuals are more likely to be self-employed, the model should try also accommodate this. As the analysis in this paper is based on sample that is very heterogeneous in terms of age, it would be interesting to see if different age groups have other transition probabilities. I.e. if possible the lagged dependent variables should be interacted with age-group indicators.

Fourth, CBS (2014) also writes that in more recent years many of the new “zzp’ers” have entered self-employment from unemployment. This gives rise to the hypothesis that self-employment may be chosen out of necessity as an alternative to unemployment. Hence one should think of ways to identify such transitions similar to e.g. Been and Knoef (2013) for the elderly. A first step could also be to at least include age-gender-education specific unemployment rates as control variables.

Fifth, the data in the LISS panel should be further explored. Since the background data should be updated monthly by the household representative, one may find more information and higher frequency changes in there. This would however mostly help with the unemployment definition but less with other labour states as one year periods seem reasonable to look at labour state decisions in general. The LISS panel however includes many variables and more information on most labour states. If one wants to continue with the LISS data on this topic it should be a priority to find a model in which the covariates that enter the decision may differ across labour states.

A final point of discussion is the absence of any income variables in the estimation. While those would introduce endogeneity it would nevertheless be reasonable to include them. So far, the model assumes that the covariates influence some indirect utility an individual receives from being in a labour state. In a next step it could be interesting to see how expected wages may enter an individual’s decision as monetary incentives might be the main reason for choosing an employment, especially if one considers the necessity hypothesis. E.g. Taylor (1996) found in a structural model that differences in predicted log-earnings are a determinant of self-employment. Similar to this line of thought one may also assume that the state-dependence observed in this model may be due to investments or costs that come with changing ones labour state. Setting up one's own business certainly comes with some costs and hence an individual may prefer to remain in wage employment even though its utility may be higher in self-employment (when costs don't enter the equation). Lastly, it should also be noted that the inclusion of expected income would also allow to consider how changes in pension benefits tied to employment states affect individuals' choices.

7. Conclusion

This paper examines the dynamics of self-employment in the Netherlands. To this purpose I define four labour states: employment, self-employment, unemployment, and not participating in the labour force. Using a multinomial logit framework with unobserved heterogeneity I then model the labour state choices observed in the data for a sample of individuals aged 23 to 60 based on the LISS panel.

The empirical specification allows me to differentiate between spurious and true state dependence. I find that individual heterogeneity enters the model significantly and hence should not be ignored when we try to model individuals' labour state choices. I furthermore find that the lagged dependent variable, i.e. past labour state choices, enter significantly, and that the observed initial conditions also are important in explaining the choices.

As far as I can judge based on my simulation results, my model is also able to attribute higher choice probabilities for self-employment to men, when we consider the average age in the data. I also find that for the given example the probabilities to choose self-employment are higher for an individual with a small family consisting of a partner and one child, than for a similar individual that lives alone. Overall, despite of the bad results found in the robustness checks, the model still seems to perform rather well in matching the transition probabilities observed in the data. One should thus be able to use it as a benchmark for further work related to the dynamics of self-employment in the Netherlands.

Table 8: Static multinomial panel regressions with correlated random effects

	Model A			Model C		
	self-empl.	*unempl.*	*outofLF*	*self-empl.*	*unempl.*	*outofLF*
Constant	-28.2269	-8.7263	-12.5568	-24.2317	-7.4596	-10.0098
	(1.8215)	(0.9279)	(0.9205)	(1.6947)	(0.9082)	(0.8706)
Age	0.1828	0.0898	0.1717	0.1364	0.0819	0.1648
	(0.0222)	(0.0128)	(0.0131)	(0.0217)	(0.0124)	(0.0128)
Female	0.9724	1.5234	4.4433	-1.2051	-0.0007	0.7335
	(0.4213)	(0.2469)	(0.3131)	(0.8354)	(0.3760)	(0.4536)
Has partner				-1.5467	-1.5421	-2.0502
				(0.7224)	(0.4170)	(0.4660)
Has kids				-2.0368	-1.2648	-2.9464
				(0.6368)	(0.4566)	(0.4242)
i: f x partner				1.2113	1.5761	3.3641
				(0.9113)	(0.4564)	(0.5119)
i: f x kids				0.9883	0.3256	1.8464
				(0.6969)	(0.4374)	(0.4467)
midd.educ	0.2555	-1.0198	-1.5619	0.8307	-0.9926	-1.7204
	(0.7585)	(0.4770)	(0.5294)	(0.8038)	(0.4868)	(0.5134)
high educ	1.1412	-1.6499	-3.4637	1.7765	-1.6005	-3.5125
	(0.8009)	(0.4884)	(0.5643)	(0.8786)	(0.5000)	(0.5471)
own dwelling	1.5403	-1.1332	-1.8627	0.8772	-1.1398	-2.0376
	(0.4129)	(0.2185)	(0.2225)	(0.4433)	(0.2389)	(0.2361)
# of hh members	0.9288	-0.0705	0.1647	1.0926	0.3073	0.6246
	(0.1546)	(0.0766)	(0.0811)	(0.2630)	(0.1596)	(0.1294)
L	12.5538			12.9983		
	(0.6016)			(0.6495)		
	4.0742	2.4500		3.9874	2.3636	
	(0.3222)	(0.2154)		(0.3190)	(0.2053)	
	5.3495	3.5369	3.3459	5.2870	3.1878	3.2204
	(0.2779)	(0.2129)	(0.1813)	(0.2742)	(0.1894)	(0.1815)
implied W	157.5991	51.1469	67.1571	168.9569	51.8292	68.7218
	51.1469	22.6018	30.4607	51.8292	21.4857	28.6158
	67.1571	30.4607	52.3226	68.7218	28.6158	48.4852
LogLikelihood	*-7265.56*			*-7219.02*		
N	*4598*			*4598*		
total observations	*15248*			*15248*		

Standard errors in parentheses.
All regressions include time effects (not reported).

Table 9: Dynamic multinomial panel regressions with correlated random effects

	Model D.1			**Model D.2**		
	self-empl.	*unempl.*	*not in LF*	*self-empl.*	*unempl.*	*not in LF*
Constant	-9.6334	-5.8593	-6.7221	-9.1291	-5.6807	-6.5897
	(1.2018)	(0.6991)	(0.5804)	(1.0745)	(0.6892)	(0.5723)
Age	0.0110	0.0339	0.0634	0.0100	0.0272	0.0610
	(0.0141)	(0.0089)	(0.0074)	(0.0124)	(0.0089)	(0.0071)
Female	-0.8046	-0.2378	0.0041	-0.6358	-0.2315	-0.0528
	(0.6509)	(0.2962)	(0.3080)	(0.5779)	(0.3003)	(0.3081)
Has partner	-0.0660	-0.7018	-0.4644	-0.1270	-0.6511	-0.4240
	(0.4842)	(0.3180)	(0.3057)	(0.4764)	(0.3254)	(0.3094)
Has kids	0.6910	-0.0582	-0.8682	0.7826	0.0397	-0.7559
	(0.5312)	(0.3447)	(0.2962)	(0.5132)	(0.3565)	(0.3033)
i: f x partner	0.6253	0.6582	1.1519	0.7181	0.5407	1.2024
	(0.7019)	(0.3649)	(0.3492)	(0.6461)	(0.3708)	(0.3559)
i: f x kids	-0.8798	-0.4353	0.2886	-0.9767	-0.4622	0.2177
	(0.5089)	(0.3290)	(0.3032)	(0.4850)	(0.3360)	(0.3037)
midd.educ	0.4758	-0.6323	-0.8551	0.1579	-0.6927	-0.8401
	(0.7342)	(0.4482)	(0.3564)	(0.6916)	(0.4451)	(0.3599)
high educ	0.6883	-0.9885	-1.8102	0.3073	-1.0149	-1.7199
	(0.7512)	(0.4563)	(0.3778)	(0.7029)	(0.4552)	(0.3785)
own dwelling	0.6340	-0.3593	-0.5760	0.4960	-0.3015	-0.5841
	(0.3146)	(0.1962)	(0.1702)	(0.2888)	(0.2037)	(0.1690)
# of hh members	-0.0313	0.0089	0.1260	-0.0372	-0.0134	0.0984
	(0.1802)	(0.1309)	(0.0901)	(0.1736)	(0.1369)	(0.0941)
l.self-empl.	4.7493	1.1405	2.7086	4.5803	1.1152	2.6128
	(0.4451)	(0.5976)	(0.3762)	(0.4383)	(0.6183)	(0.3727)
l.unempl.	2.2378	2.1730	1.5889	2.0569	2.0778	1.8109
	(0.4948)	(0.2526)	(0.2533)	(0.4746)	(0.2549)	(0.2503)
l.outofLF	2.9516	1.2752	2.1616	2.3565	1.5684	2.1843
	(0.4463)	(0.2392)	(0.1320)	(0.3943)	(0.2353)	(0.1304)
initial: self	8.3007	3.5180	2.4636	8.4916	3.7819	3.0737
	(1.2451)	(0.7565)	(0.5541)	(1.0771)	(0.7940)	(0.5631)
initial: unem	2.5415	3.3523	3.5718	2.9856	3.4002	3.3910
	(0.8492)	(0.4894)	(0.4293)	(0.7675)	(0.4823)	(0.4231)
initial: outLF	1.6819	3.0333	5.6603	2.8266	2.5038	5.5868
	(0.6394)	(0.3967)	(0.3609)	(0.5720)	(0.3621)	(0.3528)
init.obs: time cntrl		Yes			No	

Standard errors in parentheses.

All regressions include time effects (not reported).

Table 10: Continuation of table 9

	Model D.1			Model D.2		
	self-empl.	*unempl.*	*not in LF*	*self-empl.*	*unempl.*	*not in LF*
L	2.7900			2.8295		
	(0.3753)			(0.3159)		
	1.0489	1.5764		1.2997	1.4338	
	(0.3050)	(0.2179)		(0.2897)	(0.2488)	
	0.2912	2.0092	1.1803	0.9848	0.8624	1.9019
	(0.2645)	(0.1952)	(0.2347)	(0.2047)	(0.2036)	(0.1468)
implied W	7.7840	2.9264	0.8125	8.0062	3.6775	2.7865
	2.9264	3.5850	3.4727	3.6775	3.7450	2.5165
	0.8125	3.4727	5.5148	2.7865	2.5165	5.3308
LogLikelihood	*-5131.11*			*-5137.50*		
N	*4598*			*4598*		
total observations	*15248*			*15248*		

Standard errors in parentheses.

References

Albarrán, Pedro, Raquel Carrasco, and Jesús M Carro, "Estimation of Dynamic Nonlinear Random Effects Models with Unbalanced Panels," 2015.

Been, Jim and Marike Knoef, "The necessity of self-employment towards retirement: evidence from labor market dynamics and search requirements in unemployment insurance," in "28th Annual Congress of the European Economic Association" 2013.

Beusch, Elisabeth, "A Description of Self-Employment in the Netherlands and Potential Explanatory Variables," 2015. Major paper in econometrics.

Bhat, Chandra R, "Quasi-Random Maximum Simulated Likelihood Estimation of the Mixed Multinomial Logit Model," 2001.

_ , "Simulation Estimation of Mixed Discrete Choice Models Using Randomized and Scrambled Halton Sequences," 2003.

Blanchflower, David G, "Self-employment in OECD countries," *Labour economics*, 2000, *7* (5), 471–505.

Buddelmeyer, Hielke and Mark Wooden, "Transitions out of casual employment: the Australian experience," *Industrial Relations: A Journal of Economy and Society*, 2011, *50* (1), 109–130.

CBS, "Achtergrondkenmerken en ontwikkelingen van zzp'ers in Nederland," cbs.nl December 2014.

Gong, Xiaodong, Arthur Van Soest, and Elizabeth Villagomez, "Mobility in the urban labor market: a panel data analysis for Mexico," *Economic Development and Cultural Change*, 2004, *53* (1), 1–36.

Haan, Peter and Arne Uhlendorff, "Estimation of multinomial logit models with unobserved heterogeneity using maximum simulated likelihood," *The Stata Journal*, 2006, *6* (2), 229–245.

Heckman, James J, "The Incidental Parameters Problem and the Problem of Initial Conditions in Estimating a Discrete Time-Discrete Data Stochastic Process," in Charles F. Maski and Daniel L. McFadden, eds., *Structural Analysis of Discrete Data and Econometric Applications*, MIT Press, Cambridge, MA, 1981, pp. 179–195.

_ , "Statistical Models for Discrete Panel Data," in Charles F. Maski and Daniel L. McFadden, eds., *Structural Analysis of Discrete Data and Econometric Applications*, MIT Press, Cambridge, MA, 1981, pp. 114–178.

ISSP Research Group, *International Social Survey Programme: Work Orientation III - ISSP 1989*, za1840 data file version 1.0.0 ed. 1991.

_ , *International Social Survey Programme: Work Orientation III - ISSP 1997*, za3090 data file version 1.0.0 ed. 1999.

_ , *International Social Survey Programme: Work Orientation III - ISSP 2005*, za4350 data file version 2.0.0 ed. 2013.

Oguzoglu, Umut, "Disability and multi-state labour force choices with state dependence," 2010.

Parker, Simon C, *The Economics of Self-employment and Entrepreneurship*, Cambridge University Press, 2004.

Prowse, Victoria, "Modeling employment dynamics with state dependence and unobserved heterogeneity," *Journal of Business & Economic Statistics*, 2012, *30* (3), 411–431.

Taylor, Mark P, "Earnings, independence or unemployment: why become self-employed?," *Oxford Bulletin of Economics and Statistics*, 1996, *58* (2), 253–266.

Train, Kenneth E, *Discrete choice methods with simulation*, Cambridge University Press, 2009.

Wooldridge, Jeffrey M, "Simple solutions to the initial conditions problem in dynamic, nonlinear panel data models with unobserved heterogeneity," *Journal of applied econometrics*, 2005, *20* (1), 39–54.

_ , "Correlated random effects models with unbalanced panels," 2009.

Zucchelli, Eugenio, Mark N Harris, and Xueyan Zhao, "Ill-health and transitions to part-time work and self-employment among older workers," *Available at SSRN 2269449*, 2012.

A. First order derivatives

Let for notational ease $\mathsf{P}(y_{ijtr}) \equiv Prob(y_{ijtr}|X_{it}, y_{it-1}, y_{i0}, \xi_i)$ and similarly for all other probabilities. The simulated loglikelihood function from equation (10) is then given by

$$logS\mathcal{L} = \sum_{i=1}^{N} log \left[\frac{1}{R} \sum_{r=1}^{R} \prod_{t=2}^{T} \left(\prod_{j=1}^{J} \mathsf{P}(y_{ijtr})^{\mathbb{D}_{ijt}} \right)^{\mathbb{S}_{it}} \right]$$

Furthermore let vector θ be a vector with κ elements containing the stacked columns of all the coefficient matrices β, γ, δ, as well as a vector of all the elements of the Cholesky factorisation $[l_{11}, l_{21}, l_{22}, l_{31}, l_{32}, l_{33}]'$. And lastly, let V_it be a vector of the covariates, where X_it, y_{it-1}, y_{i0} are stacked accordingly to the elements in θ. Note that e.g. for a covariate in X_{it} its index in V_it κ is jointly defined by j and k. The last rows in V_it are given by the vector $[\xi_1, \xi_1, \xi_2, \xi_1, \xi_2, \xi_3]'$.

Then using section 4.1.3 the first order derivative can be written as

$$\begin{aligned}
\frac{\partial logS\mathcal{L}}{\partial \theta_\kappa} &= \sum_{i=1}^{N} \left[\frac{1}{\mathsf{P}(y_i)} \left(\frac{1}{R} \sum_{r=1}^{R} \mathsf{P}(y_{ir}) \left(\sum_{t=1}^{T} \mathbb{S}_{it} \frac{\mathsf{P}(y_{itr})'}{\mathsf{P}(y_{itr})} \right) \right) \right] \\
&= \sum_{i=1}^{N} \left[\frac{1}{\mathsf{P}(y_i)} \left(\frac{1}{R} \sum_{r=1}^{R} \mathsf{P}(y_{ir}) \left(\sum_{t=1}^{T} \mathbb{S}_{it} \left[\mathbb{D}_{i\kappa t} - \mathsf{P}(y_{i\kappa tr}) \right] V_{i\kappa t} \right) \right) \right]
\end{aligned}$$

B. Additional tables

Table 11: Labour state shares in the observed population

	2009	2010	2011	2012	2013	2014
Shares in % of sample						
employees	72.1	71.3	71.6	71.7	72.1	71.2
self-employed	7.6	7.1	7.2	6.7	6.8	7.1
unemployed	1.6	2.6	2.4	3.1	3.3	3.6
non working	18.7	19.1	18.8	18.5	17.8	18.2
Shares in % of labour force						
employees	88.7	88.0	88.2	88.0	87.6	87.0
self-employed	9.4	8.8	8.8	8.3	8.3	8.7
unemployed	2.0	3.2	3.0	3.8	4.1	4.4

Source: LISS panel, own calcuations.

Table 12: Sub-panel patterns in the data and their frequency

Nr. of individuals	%	Pattern
1203	26.16	111111
249	5.42	111110
261	5.68	111100
317	6.89	111000
551	11.98	110000
695	15.12	100000
57	1.24	011111
16	0.35	011110
20	0.43	011100
18	0.39	011000
61	1.33	010000
220	4.78	001111
70	1.52	001110
75	1.63	001100
138	3.00	001000
33	0.72	000111
9	0.20	000110
14	0.30	000100
387	8.42	000011
126	2.74	000010
78	1.70	000001
4598	100.00	

Table 13: Dynamic multinomial regression without individual characteristics

	Model E		
	self-empl.	*unempl.*	*outofLF*
Constant	-8.2521	-6.2493	-5.5614
	(0.5655)	(0.3225)	(0.2227)
l.self-empl.	4.5308	0.9320	2.4065
	(0.4170)	(0.5965)	(0.3706)
l.unempl.	1.9113	2.0104	1.7333
	(0.4787)	(0.2482)	(0.2514)
l.outofLF	2.2265	1.4617	2.0945
	(0.4007)	(0.2334)	(0.1315)
initial: self	8.7630	4.3005	3.6942
	(1.0618)	(0.7707)	(0.5822)
initial: unem	2.8809	3.9369	4.1586
	(0.7580)	(0.5024)	(0.4552)
initial: outLF	2.7796	2.9481	6.5884
	(0.5999)	(0.3746)	(0.3859)
init.obs: 2009	0.7162	0.7327	-0.1212
	(0.6760)	(0.3625)	(0.3501)
init.obs: 2010	0.0816	0.2537	-0.4346
	(0.3967)	(0.2517)	(0.2353)
init.obs: 2011	1.0518	0.1648	-0.2785
	(1.1032)	(0.8695)	(0.9108)
init.obs: 2012	0.1903	-0.2855	-0.4506
	(0.4229)	(0.3205)	(0.2846)
init.obs: 2013	1.1097	0.0276	-0.8775
	(0.9301)	(0.7889)	(0.5892)
L	2.9552		
	(0.3148)		
	1.4593	1.5242	
	(0.2889)	(0.2580)	
	1.1790	1.0737	2.0543
	(0.2232)	(0.2314)	(0.1544)
W	8.7332	4.3126	3.4842
	4.3126	4.4530	3.3571
	3.4842	3.3571	6.7628
LogLikelihood	*-5275.04*		
N	*4598*		
total observations	*15248*		

Standard errors in parentheses.

Includes time effects (not reported).

Table 14: Dynamic multinomial panel regressions with correlated random effects

	Model B.1			**Model B.2**		
	self-empl.	*unempl.*	*not in LF*	*self-empl.*	*unempl.*	*not in LF*
Constant	-9.7969	-5.9125	-7.0234	-9.4060	-5.7821	-7.0307
	(1.1460)	(0.6735)	(0.5586)	(1.0947)	(0.6604)	(0.5463)
Age	0.0119	0.0321	0.0609	0.0055	0.0305	0.0610
	(0.0141)	(0.0088)	(0.0071)	(0.0132)	(0.0085)	(0.0069)
Female	-0.7642	0.0410	0.9780	-0.7399	0.0549	0.9761
	(0.2680)	(0.1684)	(0.1597)	(0.2604)	(0.1658)	(0.1581)
midd.educ	0.4040	-0.6085	-0.8186	0.3592	-0.6301	-0.8271
	(0.7371)	(0.4402)	(0.3494)	(0.7421)	(0.4329)	(0.3493)
high educ	0.6131	-0.9595	-1.8254	0.5671	-0.9711	-1.8352
	(0.7465)	(0.4470)	(0.3701)	(0.7523)	(0.4398)	(0.3692)
own dwelling	0.7030	-0.3788	-0.4322	0.6605	-0.3906	-0.4346
	(0.3118)	(0.1810)	(0.1544)	(0.2972)	(0.1779)	(0.1532)
# of hh members	0.0842	-0.1500	-0.0432	0.0688	-0.1593	-0.0399
	(0.0922)	(0.0622)	(0.0515)	(0.0914)	(0.0609)	(0.0511)
l.self-empl.	4.6951	1.2583	2.7278	4.6328	1.2587	2.7227
	(0.4480)	(0.5997)	(0.3707)	(0.4462)	(0.5918)	(0.3660)
l.unempl.	2.4487	2.1742	1.6477	2.3786	2.1887	1.6472
	(0.5335)	(0.2461)	(0.2518)	(0.5290)	(0.2455)	(0.2504)
l.outofLF	2.8721	1.3299	2.1804	2.7991	1.3192	2.1833
	(0.4542)	(0.2345)	(0.1309)	(0.4333)	(0.2330)	(0.1305)
initial: self	8.4561	3.2326	2.5459	8.5968	3.2732	2.5926
	(1.2385)	(0.7490)	(0.5390)	(1.2157)	(0.7528)	(0.5363)
initial: unem	2.3529	3.3423	3.6687	2.5217	3.3482	3.6627
	(0.8442)	(0.4849)	(0.4350)	(0.8388)	(0.4644)	(0.4298)
initial: outLF	1.7703	2.9176	5.7283	1.9386	2.9405	5.7190
	(0.6493)	(0.3756)	(0.3597)	(0.6355)	(0.3734)	(0.3553)
init.obs: time cntrl		Yes			No	
L	2.8624			2.9004		
	(0.3705)			(0.3591)		
	0.8849	1.6446		0.9266	1.6175	
	(0.3055)	(0.2192)		(0.3064)	(0.2183)	
	0.3324	1.8119	1.4835	0.3800	1.8035	1.4798
	(0.2546)	(0.1982)	(0.1959)	(0.2441)	(0.1990)	(0.1960)
implied W	8.1934	2.5329	0.9515	8.4125	2.6876	1.1023
	2.5329	3.4878	3.2741	2.6876	3.4751	3.2693
	0.9515	3.2741	5.5943	1.1023	3.2693	5.5867
LogLikelihood	*-5152.80*			*-5159.30*		
N	*4598*			*4598*		
total observations	*15248*			*15248*		

Standard errors in parentheses.

All regressions include time effects (not reported).

Table 15: Differences in estimated coefficients

	self-empl.	*unempl.*	*not in LF*
Constant	0.1580	-0.0156	0.0404
Age	0.0062	-0.0015	-0.0007
Female	0.1694	-0.0534	-0.0334
midd. educ	-0.3119	-0.0463	-0.0386
high educ	-0.3898	-0.0244	0.0100
own dwelling	-0.0857	0.0306	-0.0094
# of hh members	0.0062	0.0044	0.0009
t.2010	0.0128	0.0054	0.0032
t.2011	0.0069	0.0097	0.0123
t.2012	-0.0263	0.0141	0.0152
t.2013	-0.0216	0.0213	0.0102
t.2014	0.0037	0.0229	0.0162
l.self-empl.	-0.1248	-0.1297	-0.1579
l.unempl.	-0.3460	-0.0638	0.0959
l.outofLF	-0.5519	0.1188	-0.0326
initial: self	0.0877	0.5475	0.7702
initial: unem	0.5437	0.1074	0.0519
initial: outLF	1.1279	-0.1592	0.1207
init.obs: 2009	-0.0707	0.0388	-0.0285
init.obs: 2010	-0.0537	-0.0077	-0.0715
init.obs: 2011	-0.0084	-0.0009	-0.0032
init.obs: 2012	0.0113	-0.0014	-0.0019
init.obs: 2013	-0.0673	-0.0166	-0.0101
Cholesky	-0.0059		
	-0.3940	0.2016	
	-0.7435	0.6838	-0.3398
Loglikelihood			-0.2347

Calculated as “original estimates” - “estimates new solution”.

Table 16: Summary of results using 200, 250, and 300 Halton draws

	Means			Standard Deviations		
	self-empl.	*unempl.*	*not in LF*	*self-empl.*	*unempl.*	*not in LF*
Constant	-9.5112	-5.9240	-6.8033	0.0239	0.0811	0.1066
Age	0.0114	0.0330	0.0635	0.0021	0.0024	0.0011
Female	-0.8263	-0.2486	-0.0222	0.1066	0.0281	0.0150
Has partner	-0.0663	-0.7144	-0.4509	0.0152	0.0169	0.0213
Has kids	0.7510	-0.0523	-0.9020	0.0326	0.0490	0.0148
i: f x partner	0.7951	0.6836	1.1762	0.0698	0.0686	0.0305
i: f x kids	-0.9200	-0.4319	0.3081	0.0392	0.0199	0.0367
midd.educ	0.3216	-0.6398	-0.8088	0.0455	0.0420	0.0085
high educ	0.4854	-1.0011	-1.7300	0.0632	0.0531	0.0401
own dwelling	0.6575	-0.3287	-0.5989	0.0243	0.0649	0.0063
# of hh members	-0.0515	0.0041	0.1250	0.0019	0.0072	0.0026
t.2010	0.1954	0.5761	0.2004	0.0209	0.0062	0.0042
t.2011	0.1747	0.2390	0.1006	0.0250	0.0068	0.0018
t.2012	-0.0152	0.5829	0.0930	0.0314	0.0099	0.0032
t.2013	0.2851	0.6916	0.0613	0.0413	0.0041	0.0104
t.2014	0.2053	0.8394	0.2624	0.0497	0.0052	0.0104
l.self-empl.	4.7131	1.3254	2.6077	0.0795	0.0533	0.0778
l.unempl.	2.6151	2.0370	1.6615	0.1125	0.1077	0.0998
l.outofLF	2.6587	1.3380	2.1169	0.0942	0.1462	0.0400
initial: self	8.3256	3.0369	2.8844	0.3079	0.1541	0.1917
initial: unem	2.1720	3.4592	3.5944	0.1623	0.0463	0.1707
initial: outLF	2.2239	2.8922	5.7750	0.2968	0.3359	0.1550
init.obs: 2009	0.6923	0.6860	-0.0043	0.0562	0.0358	0.0339
init.obs: 2010	0.2619	0.2108	-0.1668	0.0348	0.0089	0.0229
init.obs: 2011	1.4217	0.3183	0.2750	0.0955	0.0186	0.0287
init.obs: 2012	0.3125	-0.0550	0.0057	0.0807	0.0127	0.0589
init.obs: 2013	1.5735	0.3353	-0.0274	0.0935	0.0493	0.0025
Cholesky	2.7969			0.0904		
	0.7491	1.7804		0.1199	0.0557	
	0.6793	1.5777	0.3922	0.1695	0.3989	1.9157
LogLikelihood	-5125.0756			6.1532		

Table 17: Summary of results using all permutations of primes (3,7,13)

	Means			Standard Deviations		
	self-empl.	*unempl.*	*not in LF*	*self-empl.*	*unempl.*	*not in LF*
Constant	-8.9408	-5.9291	-6.9687	0.7503	0.1682	0.0741
Age	0.0105	0.0305	0.0602	0.0042	0.0026	0.0010
Female	-0.6584	0.0500	0.9807	0.0525	0.0329	0.0139
midd.educ	0.3150	-0.6488	-0.8327	0.1348	0.0294	0.0501
high educ	0.4903	-0.9613	-1.8012	0.1817	0.0338	0.0558
own dwelling	0.7375	-0.3786	-0.4291	0.0302	0.0388	0.0301
# of hh members	0.0627	-0.1622	-0.0512	0.0132	0.0114	0.0070
t.2010	0.1873	0.5633	0.1976	0.0102	0.0090	0.0049
t.2011	0.1662	0.2301	0.0985	0.0170	0.0111	0.0035
t.2012	-0.0212	0.5711	0.0894	0.0206	0.0177	0.0071
t.2013	0.2606	0.6765	0.0533	0.0301	0.0195	0.0153
t.2014	0.1805	0.8155	0.2513	0.0358	0.0240	0.0171
l.self-empl.	5.0315	1.4036	2.7527	0.4195	0.2143	0.1940
l.unempl.	2.7539	2.0285	1.7069	0.5442	0.0885	0.0691
l.outofLF	2.8499	1.4128	2.1635	0.4072	0.0768	0.0322
initial: self	7.0875	2.7314	2.4447	1.3751	0.7924	0.7496
initial: unem	1.6754	3.4880	3.6942	0.9781	0.1483	0.1320
initial: outLF	1.6046	2.7310	5.7414	0.9551	0.2120	0.1422
init.obs: 2009	0.5846	0.7228	0.0111	0.1866	0.0440	0.0334
init.obs: 2010	0.2165	0.2263	-0.2353	0.0355	0.0456	0.0573
init.obs: 2011	1.3011	0.3411	0.3302	0.0622	0.0531	0.0611
init.obs: 2012	0.3338	-0.0860	-0.0886	0.1091	0.0380	0.0141
init.obs: 2013	1.5450	0.3736	0.0462	0.1180	0.0457	0.0718
Cholesky	2.4728			0.3903		
	0.5635	1.7885		0.6122	0.1263	
	0.3661	1.4792	1.7005	0.6035	0.2180	0.1286
LogLikelihood	-5148.9501			3.2695		

Table 18: Summary of results using all permutations of primes (3,7,11)

	Means			Standard Deviations		
	self-empl.	*unempl.*	*not in LF*	*self-empl.*	*unempl.*	*not in LF*
Constant	-7.9377	-5.9368	-6.9967	1.5042	0.1224	0.1040
Age	0.0117	0.0297	0.0606	0.0041	0.0024	0.0011
Female	-0.5516	-0.0116	0.9685	0.0824	0.0577	0.0261
midd.educ	0.1334	-0.5990	-0.8137	0.0284	0.0681	0.0291
high educ	0.1823	-0.8745	-1.7620	0.0584	0.0917	0.0267
own dwelling	0.5985	-0.3292	-0.4492	0.0788	0.0515	0.0064
# of hh members	0.0379	-0.1512	-0.0514	0.0228	0.0150	0.0090
t.2010	0.2413	0.5864	0.1880	0.0327	0.0367	0.0122
t.2011	0.2181	0.2486	0.0898	0.0187	0.0400	0.0186
t.2012	0.0380	0.5998	0.0788	0.0516	0.0584	0.0226
t.2013	0.3167	0.7152	0.0381	0.0184	0.0673	0.0254
t.2014	0.2595	0.8514	0.2335	0.0221	0.0715	0.0346
l.self-empl.	5.4009	1.2373	2.8203	0.9434	0.5216	0.3290
l.unempl.	1.9699	2.1739	1.7763	0.7624	0.2129	0.1323
l.outofLF	2.5603	1.5457	2.1608	0.3894	0.1558	0.0311
initial: self	5.7845	3.4804	2.4516	3.0664	1.6735	1.1262
initial: unem	2.3662	3.3227	3.6276	1.5457	0.1881	0.1336
initial: outLF	1.9793	2.5220	5.7465	1.1682	0.2987	0.1197
init.obs: 2009	0.6507	0.6908	-0.0073	0.2809	0.0528	0.0207
init.obs: 2010	0.1768	0.2316	-0.2380	0.0411	0.0269	0.0348
init.obs: 2011	1.0749	0.3743	0.3560	0.1213	0.0266	0.0438
init.obs: 2012	0.2606	-0.0928	-0.0348	0.0474	0.0492	0.0254
init.obs: 2013	1.2164	0.3950	0.0326	0.2346	0.0302	0.0391
Cholesky	1.7755			1.4246		
	1.1604	1.0020		0.9620	0.8384	
	0.5949	1.2499	1.2754	0.6501	0.9908	1.0023
LogLikelihood	-5159.7847			3.7336		

C. Variable coding

General note: Unless otherwise mentioned numbers of top-levels indicate new codings and on lower levels the original codings from LISS panel variables following corresponding variables.

C.1. Education

Variable: *cw[yyx]005*

Recoded to CBS categories for levels of education. See Wikipedia: Education in the Netherlands (June 10 2015) and CBS: Onderwijsniveau SOI 2003 (June 10 2015). The CBS categories 2-4 are then further coded as having received a medium level of education, and categories 5 and 6 are seen as having received higher education.

8 Not (yet) completed any education
- 1 did not complete any education
- 2 did not complete primary school

1 Primary school
- 3 primary school

2 VMBO
- 4 lower and continued special education
- 5 VGLO (continued lower education)
- 6 LBO (lower professional education)
- 7 lower technical school, household school
- 8 MULO, ULO, MAVO (lower/intermediate secondary education; US: junior high school)
- 9 VMBO vocational training program (preparatory intermediate vocational school)
- 10 VMBO theoretical or combined program (preparatory intermediate vocational school)

3 HAVO/VWO
- 11 MMS (intermediate girls' school)
- 12 HBS (former pre-university education, US: senior high school)
- 13 HAVO (higher general secondary education; US: junior high school)
- 14 VWO (pre-university education, US: senior high school)
- 15 gymnasium, atheneum, lyceum (types of pre-university education programs)
- 16 KMBO (short intermediate professional education), VHBO (preparatory higher professional education)

4 MBO

17 MBO professional training program (intermediate professional education)

18 MBO-plus to access HBO, short HBO education (less than two years) (higher professional education

5 HBO

19 HBO (higher professional education), institutes of higher education, new style

20 teacher training school

21 conservatory and art academy

6 WO

22 academic education (including technical and economic colleges, former style) bachelor's degree (kandidaats)

23 academic education (including technical and economic colleges, former style) master's degree (doctoraal)

24 academic education, bachelor

25 academic education, master

26 doctor's degree (Ph.D, including doctoral research program to obtain Ph.D)

7 Other[30]

27 other

C.2. Employment: Labour force

Variable: *cw[yyx]088-102* (3-digit numbers in list below correspond to variable name.)
Definition of labour force taken from CBS Definitions (online):

People:

- who have paid work (employed labour force), or
- who don't have paid work, recently looked for work and are directly available for it (unemployed labour force).

1 **Labour force**

– *Employed*

088 I perform paid work (even if is it just for one or several hours per week or for a brief period)

102 I perform paid work, but am looking for more or other work

[30]Further re-coded, if possible, using *cw[yyx]006*.

- *Unemployed*
  - 091 I am looking for work following the loss of my previous job
  - 092 I have performed paid work, but am released from the obligation to find a new job following the loss of my previous job
  - 093 I am a first-time job seeker
  - 094 I am seeking work following a lengthy interruption

0 **Not in labour force**

- *not looking for work*
  - 089 I am not working now, but have performed paid work in the past
  - 090 I perform unpaid work while retaining my benefit or allowance
  - 095 I am a pupil / student / trainee with an expenses claim only
  - 096 I take care of the household
  - 097 I live off private means
  - 101 I perform voluntary work
- *retired*
  - 098 I have taken early retirement
  - 099 I am a pensioner
- *disabled*
  - 100 I am partly or wholly disabled for work