The Persistence of Informality: Evidence from Panel Data

Alpaslan Akay
Melanie Khamis

DISCUSSION PAPER SERIES

I Z A

Forschungsinstitut
zur Zukunft der Arbeit
Institute for the Study
of Labor

The Persistence of Informality:
Evidence from Panel Data

Alpaslan Akay
IZA

Melanie Khamis
*Wesleyan University
and IZA*

ABSTRACT

The Persistence of Informality: Evidence from Panel Data[*]

Informality is a growing phenomenon in the developing and transition country labor market context. In particular, it is noticeable that working in an informal employment relationship is often not temporary. The degree of persistence of informality in the labor market might be due to different sources: structural state dependence due to past informality experiences and spurious state dependence due to time-invariant unobserved individual effects, which can alter the propensity of being in the informal sector independently from actual informality experiences. The purpose of our paper is to study the dynamics of informality using a genuine panel data set in the Ukrainian labor market. By estimating a dynamic panel data probit model with endogenous initial conditions, we find a highly significant degree of persistence due to previous informality experiences. This result implies that policies attempting to reduce current levels of informality may have a long-lasting effect on the labor market.

JEL Classification: D60, I31

Keywords: informality, unobserved heterogeneity, state dependence, transition countries

Corresponding author:

Melanie Khamis
Wesleyan University
Department of Economics
238 Church Street
Middletown, CT 06459-0007
USA
E-mail: mkhamis@wesleyan.edu

1. Introduction

Informality in the labor market is a prevalent phenomenon in the developing, middle-income and transition country context. A large proportion of the labor market participants is considered to be working informally at one point in time or permanently in the sector. Subjective data obtained from informality surveys conducted by the World Bank (2007) highlight that the reasons and motivations for working and remaining in an informal employment relationship can be manifold and summarized as follows: institutional barriers, type of human capital, preferences, tastes, ability, wage differences and previous informality experiences among others.[1] Some of these market or individual specific factors may lead the informality status of workers to be persistent over time. For instance, institutional barriers to enter into the formal sector such as registration costs or employers not wanting to register workers could be important for the persistence of informality status (De Soto, 1989). Lack or limited enforcement of registration or cumbersome rules and regulation may also provide incentives for employers to offer informal jobs leading workers to participate and remain in the informal sector for extended spells (De Soto, 1989; Kanbur, 2009).

Individual human capital considerations may also be an important factor to enter and stay in the informal sector. For some individuals the informal sector serves as a training ground to gain skills in order to get a job in the formal sector in later years (Maloney, 2004). Contrary to this, it could also be that working in the informal

---

[1] The degree to which these reasons and motivations are voluntary or involuntary for the informal sector participants is part of a large debate in the informality literature. For a summary of this debate see World Bank (2007).

sector might be perceived as a low-skill task and a signal for low productivity, which renders the individual to be unsuccessful in future job searches in the formal sector.

One of the other reasons of persistence in the informality status may be due to different valuation of the benefits of social security system by workers in the informal sector. Workers may be myopic towards the future and not value the benefits of social security. Individuals might not find it beneficial to contribute to the social security system due to a history of non-contribution. This can occur when the cost to pay social security at present is higher, or at least perceived to be higher, than the benefits received in the future (Levy, 2008).

These factors may characterize the labor market and lead to persistence in the informality status once the individuals enter into the informal sector. The aim of this paper is to examine whether there is persistence in the informality status of workers in the Ukrainian labor market. Particularly, we aim to disentangle *structural state dependence* which is caused by past informality experiences of workers from *spurious state dependence* which is caused by other sources of persistence such as time-invariant unobserved individual characteristics. Individuals might have different unobserved motivations and preferences for flexible work hours and different valuations of the social security benefits (Maloney, 2004; Dohmen et al., 2011). Individuals may also prefer specific tasks in a job which might only be possible in an informal arrangement (Maloney, 2004; De Mel et al., 2010). Hence, these unobserved characteristics might also alter the propensity of being in the

informal sector independently from actual informality experiences (Heckman, 1978, 1981a, b).

In order to assess the degree of persistence we estimate a dynamic random-effects probit model controlling for past informality experience, observed and unobserved individual characteristics. It is also very important to consider the *initial values problem* in the estimation of such a model (Heckman, 1981a, b; Wooldridge, 2005). This problem occurs when the process generating the informality states is not observed from the beginning and the initial values could be endogenously determined by observed and unobserved time-invariant individual characteristics. It is crucial to deal with the endogenous initial values especially with a short panel dataset (as we have in this paper) to identify the structural state dependence. We deal with this problem using two existing methods, Heckman's (1981a, b) reduced-form approximation and Wooldridge's (2005) method.

Our results can be summarized as follows: the parameter of past informality experience, i.e. structural state dependence, is highly significant on the current informality status in the context of the Ukrainian labor market. Our analysis suggests that a failure to control for the endogenous initial values leads to a wrong inference: the models seriously overestimate the structural state dependence and also the variance of the unobserved individual effect is not identified. Controlling for the endogenous initial values generates lower estimates of structural state dependence and a sizeable variance of the unobserved individual effect. The models estimated

here lead to nonlinear conditional expectations and thus we calculate the average partial effect of past informality experience for various important groups of interest. Our analysis suggests that a past informality experience increases the current probability of being in the informal sector by 7-9 percentage points, *ceteris paribus*, compared to the workers who do not have previous informal sector experience. A detailed analysis suggests that the past informal sector experience is more pronounced on the current informality status for young single males with low education.

The remaining part of the paper is organized as follows: in Section 2 the data source and descriptive statistics are presented. Section 3 introduces the dynamic random-effects probit model and the two solution methods for the initial values problem, the Heckman's (1981a, b) reduced-form approximation and the method of Wooldridge (2005). Section 4 presents the results. Section 5 discusses the robustness checks. Section 6 concludes.

2. Data

2.1. The survey

To understand the nature of informality over time we employ the Ukrainian Longitudinal Monitoring Survey (ULMS) for the years 2003, 2004 and 2007. The survey is nationally representative of the Ukrainian work force and contains household and individual questionnaires. The household questionnaire elicits responses regarding income and expenditure patterns and living arrangements and conditions. The individual questionnaire contains detailed information on various individual characteristics and a large section on current and past labor market experiences.

From this individual section we create our dependent variable, informality in the labor market: we focus on informal employment relationships and do not consider other measures of informality in our paper.[2] Informality in our data is defined according to non-registration of work contract while formal workers are registered. We generate our dependent variable using answers to the following question: *Tell me, please, are you officially registered at this job, that is on a work roster, work agreement or contract?* The answer to the question is either *registered* (formal) or *not registered* (informal). Therefore, our dependent variable is an indicator variable which takes the value 1 if the individual is an informal worker and 0 otherwise. The

---

[2] This follows Kanbur (2009) who recommends the classification of informality according to a specific regulation.

dataset includes both employees and self-employed individuals working either formally or informally. However, we have restricted our analysis to an employee-only sample as this provides the most clear-cut way to distinguish formal and informal employment structure in the labor market.

In addition to data on registration, the ULMS allows us to exploit a rich set of explanatory variables in our analysis. Age (from age 15 onwards), gender (female/male dummy), marital status (married, single, divorced/separated/widowed), education levels (10 levels), total household income, sector of occupation (8 different sectors: agriculture, industry, sales, transportation, public administration, education, services and other sectors which include occupations not included in the previous categories) and regions (Kiev, Center, East, West, South) are the variables that we use in our empirical analysis. These variables are standard in the literature on informality and have also been employed in previous studies with the ULMS (Lehmann and Pignatti, 2007; Dohmen et al., 2011). One of the important limitations of the dataset is that the duration between waves is different. We include several history variables, which account for the time gap in the panel dataset between waves 2004 and 2007. These variables indicate whether individuals changed jobs, marital status or residence during that time period not covered by the main survey. We later use these variables to test the sensitivity of our results.

2.2. Descriptive statistics

In Table 1 the descriptive statistics of the selected variables from the ULMS dataset are presented. The second column presents the statistics for the whole sample. The third and fourth columns relate to the characteristics of individuals split by informal and formal workers. Here, informal workers are counted as the ones that experience at least one spell of informality during the three waves. Formal workers are the number of observations relating to formal work experience, with no informality experience during the three waves of the ULMS. The last three columns of Table 1 show the characteristics of individuals at a point in time of each wave, 2003, 2004, and 2007.

About 4 percent of our sample is considered working in the informal sector as an employee at some point during the three waves while the remaining observations are entirely formal over this time period. Our sample is restricted only to formal and informal employees and also the data set is organized to be a balanced panel for our empirical analysis.[3] Thus, the mean informality rate is lower than the mean informality levels reported in the other studies using the same dataset (Lehmann and Pignatti, 2007). One important concern is that the attrition in the sample at use might be non-random relating to informality. We find that the informality dummy shows a

---

[3] The mean informality rate in the non-restricted dataset is 0.13 (0.09 for 2003, 0.13 for 2004 and 0.15 for 2007). We restrict the dataset only to employees by deleting the self-employed individuals. The mean informality rate then becomes 0.086 (0.06 for 2003, 0.10 for 2004 and 0.10 for 2007). We also delete individuals who are observed less than three waves and also the missing values in various variables. This would explain why our reported informality rates are lower compared to the literature (e.g., Lehmann and Pignatti, 2007)

very stable pattern over time which is in line with informality patterns in the non-restricted sample with self-employed individuals. The mean informality rates in our sample selection are 0.038, 0.044, and 0.040 for 2003, 2004 and 2007 respectively.

The descriptive statistics for informal and formal workers are similar to the findings from the literature in the Ukrainian context using the same data set (Dohmen et al., 2011). The characteristics of informal workers differ on several dimensions from workers who were in the formal sector over the three waves. Average age of the overall estimation sample is about 42 years while the informal workers are younger at about 35 years of age. Informal labor market participants are less likely to be married. More male workers are working in the informal sector. Education levels of informal sector workers are lower than education levels of formal sector participants. Looking at additional characteristics (history variables) for the period between 2004 and 2007 we find that about 2 to 3 percent of individuals report to have changed residence or to have gotten married. About 20 percent report a job change during that period and this is larger for the ones who worked at some point in the informal sector during this time period.

Table 1 about here

In order to give an impression of the dynamics of informality states in the Ukrainian labor market, we present in Table 2 the run patterns of informality status using only the estimation sample. Here, (0, 0, 0) means that an individual is formally employed

9

across the three waves. On the other hand, the triple (1, 1, 1) implies the individual is informally employed during the three waves and this indicates a high degree of persistence. The other possibilities indicate lower levels of persistence to and from informality to formality or vice versa.

Table 2 about here

3. Econometric framework

3.1. The model

In order to distinguish structural state dependence from other sources of persistence we specify a dynamic random-effects probit model by controlling for previous period informality status, observed and unobserved individual characteristics, and endogenous initial values. The dynamic panel random-effects model is specified as:

$$d_{it} = \mathbf{1}\left(x_{it}\beta + \lambda d_{i,t-1} + u_{it} > 0\right), \tag{1}$$

$$u_{it} = \eta_i + \varepsilon_{it}, \tag{2}$$

$$d_{i1} = \mathbf{1}\left(x_{i1}\beta_1 + u_{i1} > 0\right) \tag{3}$$

where d_{it} is a binary dependent variable indicating whether an individual i is informally employed during the current period t (where $i = 1,...,I$ and $t = 1,2,3$); x_{it} is a vector of current socio-demographic and economic characteristics (such as

10

education and marital status); β is the corresponding vector of parameters to be estimated; $d_{i,t-1}$ is an *observed* binary variable indicating whether an individual i was in the informal sector during the previous period $(t-1)$; and the parameter λ represents the *structural state dependence* following Heckman (1981a, b).

The error term in the model (1) and (3) has two components as displayed in (2). The first term (η_i) captures the time-invariant unobserved individual effects (such as motivation and ability). To control for these characteristics is crucial in order to be able to identify structural state dependence (λ). The second term (ε_{it}) is the usual error term, which is assumed to have a normal distribution with zero mean and unit variance due to the identification of a discrete choice model. The actual disturbance process is assumed to be serially uncorrelated. However, in this model controlling for unobserved individual effects automatically induces a serial correlation. The correlation between two sequential error terms is, $Corr(u_{it}, u_{is}) = \sigma_\eta^2 / \sigma_\eta^2 + 1$, $(t, s = 1, ..., T_i; t \neq s)$, where σ_η^2 is the variance of unobserved individual effects to be estimated. We follow a fully parameterized random-effects approach with a maximum likelihood estimator. The log likelihood function that is used in the estimation process is as follows:

$$\log L = \sum_{i=1}^{I} \ln \left[\int_{-\infty}^{\infty} \left\{ f_1(d_{i1} \mid \{\mathbf{x}_{it}\}_{t=1}^{T}, \eta_i) \prod_{t=2}^{T} f_{it}(d_{it} \mid d_{i,t-1}, \mathbf{x}_{it}, \eta_i; \beta) \right\} f(\eta_i) d\eta_i \right], \quad (4)$$

$$f_{it}(d_{it} \mid d_{i,t-1}, \mathbf{x}_{it}, \eta_i; \beta) = \Phi \left[(2d_{it} - 1)(\mathbf{x}'_{it}\beta + \lambda d_{i,t-1} + \sigma_\eta \eta_i) \right], \quad (5)$$

where Φ is the distribution function of standard normal random variable and $f_1(d_{i1} \mid \{\mathbf{x}_{it}\}_{t=1}^{T}, \eta_i)$ is the conditional distribution of initial values. The specification of this distribution is necessary in order to be able to identify structural state dependence.

3.2. Initial values problem

The likelihood function in (4) can be easily maximized using a Gaussian-Hermite quadrature when the conditional distribution of initial values $f_1(d_{i1} \mid \{\mathbf{x}_{it}\}_{t=1}^{T}, \eta_i)$ is known. However, the distribution is unknown since the system given in (1)-(3) has started many periods before our sample panel dataset was observed. In this case the initial values would be endogenously determined with the evolution of parameters, observed and unobserved characteristics of individuals. In order to identify the structural state dependence and to disentangle it from other sources of persistence the initial values should be considered as endogenous with a probability distribution conditioned on observed and unobserved individual characteristics.

There are two main methods for doing this: Heckman's (1981a, b) reduced-form approximation and the Wooldridge's (2005) method which is a simple alternative. Heckman's method is based on available pre-sample information with which the conditional distribution of initial values is approximated via a reduced-form. This approximation allows a flexible specification of the relationships between initial

values, observed and unobserved individual characteristics. Wooldridge (2005) introduces a simple alternative to Heckman's reduced-form approximation. The method suggests that the unobserved individual effects should be considered conditional on initial values and the time-varying exogenous variables in a similar way to the correlated random-effects model of Chamberlain (1984).

Recent studies suggest that there may be differences in the magnitude of state dependence and the estimated variance of unobserved individual effects between these two methods, especially for very short panels (Arulampalam and Stewart, 2009; Akay, 2011). Our sample size is small with $T=3$ and we use both methods to check for the sensitivity of the results. We define the reduced-form which is employed in the Heckman's method as follows,

$$d_{i1} = \mathbf{1}\big(x_{i1}\theta + u_{i1} > 0\big), \tag{6}$$

$$u_{i1} = \psi\eta_i + \varepsilon_{i1}, \tag{7}$$

and the conditional distribution of the initial values is approximated as

$$f_1(d_{i1} \,|\, \{\mathbf{x}_{it}\}_{t=1}^{T}, \eta_i) = \Phi\Big[(2d_{i1} - 1)(x_{i1}\theta + \psi\sigma_\eta \eta_i) \Big]\,. \tag{8}$$

We estimate the parameters β, θ, ψ and σ_η simultaneously by inserting (8) into the likelihood function (4) without imposing any restrictions (Heckman, 1981a, b; Hsiao, 2003). The main assumption in this particular formulation of Heckman's reduced-

form approximation method is that the first period error term u_{i1} is correlated with

the unobserved individual effects η_i but it is uncorrelated with u_{it} for $t > 1$.

For the Wooldridge method we first define an auxiliary distribution of unobserved

individual effects as follows:

$$\eta_i = \pi_0 + \pi_1 \overline{x}_i + \pi_2 d_{i1} + \alpha_i, \tag{9}$$

where π_0 (a constant), π_1 and π_2 are parameters to be estimated; $\overline{x}_i = (1/T)\sum_{i=1}^{T} x_{it}$ is

the *within-means* of time-variant exogenous variables and α_i is the new unobserved

individual effect which is assumed to be normally distributed with

$\alpha_i \mid \overline{x}_i, d_{i1} \sim N\left[\pi_0 + \pi_1 \overline{x}_i + \pi_1 d_{i1}, \sigma_\alpha^2\right]$. Inserting (9) into (4) generates a conditional

likelihood which can be estimated as a standard random-effects probit model. This

method is also very similar to the Chamberlain's (1984) correlated-effects (quasi-

fixed-effects) model since the auxiliary distribution of unobserved individual effects

includes some of time-variant exogenous variables. We only have limited

information on the time-varying characteristics of the individuals. Moreover, we

only have three waves and some time-varying variables such as education do not

show enough variation in the sample across time. We mainly use age and income in

the auxiliary distribution of unobserved individual effects given in (9).

4. Results

Our main aim is to test whether there is significant structural state dependence on the probability of being employed in the informal sector. We first start with a standard static model in order to compare our results to the previous literature. We then extend our analysis to dynamic models in which we check the importance of endogenous initial values under the two different estimation methods. Finally, we test the robustness of our results.

4.1. Main results

We present the full set of parameter estimates from various model specifications in Table 3. In each of these specifications the dependent variable is the current informality status and the models include previous period informality status together with various observed and unobserved individual characteristics. We first present a benchmark specification based on a static random-effects model, which is similar to models estimated in the previous literature (Lehmann and Pignatti, 2007; Dohmen et al., 2011; Arias and Khamis, 2008; World Bank, 2007). This allows us to validate the results from our dataset and to see whether our selection generates similar results compared to the literature. We also examine whether time-invariant unobserved individual characteristics explain an important proportion of the variation in the probability of being employed in the informal sector. This specification is presented

in the second column (Model I).[4] Most of the results are in line with the literature, which estimates the propensity of being in the informal sector in the Ukraine for the cross-section sample in 2007 or in 2003 and 2004 (Dohmen et al. 2011; Lehmann and Pignatti 2007). In particular, the results relating to household income, marital status, education and age exhibit similar patterns in our estimations compared to the results found in the literature (Dohmen et al., 2011; Lehmann and Pignatti 2007). The household income is negatively correlated with the probability of being in informal sector but it is not significant here. We find that the relationship between age and informality follows a U-shape, with a high propensity for informality during younger and older ages and a lower propensity for informality for the age cohorts in between. Marital status is significant and negatively correlated with the informal sector participation. Education tends to be a highly significant determinant of informality in the Ukraine. A higher number of years of education (higher education levels in our case) is negatively correlated with informality. Gender is not a significant determinant of informal sector participation in our initial model (Model I). The results also suggest that unobserved individual characteristics explain an important part of the variation in the probability of being employed in the informal sector ($\rho = 0.52$).

Our main aim is to identify the structural state dependence and to disentangle it from other sources of persistence. Hence, we are mainly interested in the magnitude and

---

[4] Note that all models estimated here are based on the assumption that the observed and unobserved characteristics are not orthogonal to each other. The correlation is controlled for using Chamberlain's (1984) correlated-effects model (quasi-fixed-effects model).

statistical significance of two key parameters: the parameter of the lagged

informality status (λ) and the variance of unobserved individual effects (σ_α^2). In the

third column of Table 3, we present the estimation results obtained by using the

exogenous initial values assumption (Model II). This specification implies that the

first period state of being in the informal sector is exogenous (not a function of

individual observed and unobserved characteristics) which is not a very plausible

assumption itself. The parameter of lagged informality status, i.e., structural state

dependence, is estimated by the exogenous initial values assumption to be around

1.40 (Table 3, Model II) and it is highly significant. However, it is important to note

that the variance of unobserved individual effect is not identified and it is estimated

as zero, which is not expected given the large variance found in the case of the static

random-effects model (Model I).

In the fourth and fifth columns of Table 3 we present results which control for the

endogenous initial values using the Wooldridge's and Heckman's methods (Model

III and IV). The Wooldridge's method is a simple way to deal with the initial values

problem in the dynamic probit models with random effects. However, it may

generate bias for panels of very short durations (Arulampalam and Stewart, 2009;

Akay, 2011). We also do not have many time-variant characteristics to specify a

flexible conditional distribution for the unobserved individual effects. The estimated

parameter of lagged dependent variable is reduced to 0.84 which is almost half of the

size of the parameter generated with the assumption of exogenous initial values.

However, as before the variance of unobserved individual effect is not identified and estimated to be very close to zero.

The preferred method in this paper to deal with the initial values problem is the Heckman's reduced-form approximation given the available sample characteristics and the duration of the panel dataset. Since we do not have additional pre-sample information of individuals, we follow the suggestions of Heckman (1981a, b) to use the first wave in the reduced-form equation (6). The characteristics we include are age, education, marital status, regions and income. The estimated parameter of lagged informality status is highly significant and estimated as 0.80, which is even smaller compared to the Wooldridge method. An important result is that the variance of the unobserved individual effect is identified and estimated as 0.49. This reflects a large variance for the heterogeneity distribution among informal sector workers in Ukraine.

Table 3 about here

4.2. Average partial effects

The results reported in Table 3 suggest that there is a highly significant structural state dependence of informality. However, the models estimated in this paper have nonlinear conditional expectations and the *ceteris paribus* interpretation is only possible using the partial (marginal) effect of a variable. There are various ways of

calculating partial effects. The nonlinearity inherent in the dynamic random-effects probit specification allows us to calculate the partial effects for each individual. The individual partial effects of lagged dependent variable $d_{i,t-1}$ can then be calculated using the conditional expectation of the probit model:[5]

$$\hat{m}_{it}^{d,-1}\left(\mathbf{x}_{it}, d_{i,t-1}\right) = \Phi\left(x_{it}\hat{\beta} + \hat{\lambda}(d_{i,t-1}=1)\right) - \Phi\left(x_{it}\hat{\beta} + \hat{\lambda}(d_{i,t-1}=0)\right), \qquad (10)$$

where $\hat{m}_{it}$ is the partial effect function; Φ is the cumulative distribution function of standard normal random variable; and $\hat{\beta}$ and $\hat{\lambda}$ are the estimated parameters. A consistent estimator of the population average partial effect (*ape*) can be calculated by simply averaging each individual-time partial effects in the observed sample:[6]

$$ape(\mathbf{x}_{it}, d_{i,t-1}) = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=2}^{T}\hat{m}_{it}^{d,-1}\left(\mathbf{x}_{it}, d_{i,t-1}\right). \qquad (11)$$

We present the estimated *ape* and their population averaged standard errors in Table 4. The *ape* generated by the model which assumes exogenous initial values is 0.20 which implies that previous period informality status increases the probability of being in the informal sector in the current period by 20 percentage points. It is large and highly significant as expected. The estimated *ape* of structural state dependence

---

[5] We calculate the standard errors of the partial effects using the delta method.
[6] One of the other alternatives is to integrate the partial effect function over the unobserved individual effects using Monte Carlo integration methods. In this paper we assume that the effect of unobserved individual effect can be ignored after averaging the partial effects for the whole population (see Wooldridge, 2005).

is substantially reduced once we control for the endogenous initial values. The average partial effect of lagged informality status on the current informality status is around 7-9 percentage point, *ceteris paribus*.

Table 4 about here

In the second part of Table 4, we sort individual partial effects by using socio-demographic characteristics such as gender and marital status and estimate the *ape* of lagged informality in each case. The estimated *ape* is significant in most cases and varies in size. It is clearly larger for males. We generate two dummy variables to indicate low educated (less than education level 7) and high educated individuals (education level 7 or higher). The estimated *ape* of lagged informality status is much larger among low educated individuals. We also sort individuals by marital status and calculate the *ape* of lagged informality status. It is larger for single people compared to married or divorced/separated/widowed individuals. One of the important finding is that the *ape* of lagged informality status is larger for young people and it gradually is reduced by age. We also find that the *ape* of lagged informality status differs by sectors. As we would expect, it is not significant for the public administration and education sectors. However, it is substantially larger for the sale and service sectors.

5. Robustness Checks

Alternative specifications - The dynamic probit model with random effects relies on strict parametric assumptions for the auxiliary distribution of unobserved individual effects and initial values. To check the sensitivity of previous results with respect to these assumptions, we estimate a dynamic linear probability model using a GMM (Generalized Method of Moments) estimator with first-differences (Arellano and Bond, 1991). This specification provides a semi-parametric identification of the model parameters by eliminating the unobserved individual effects and initial values problem (Stewart, 2007; Alessie et al., 2004). The full estimation results of the first-difference GMM estimator are presented in the second column of Table 5 (Model I).[7] In this specification we use the same set of independent variables as before. However, due to first-differences the time-invariant variables are swept away and the number of observations in the final estimation sample is also reduced. We experiment with various alternative sets of instruments including the initial informality status d_{i1} as one of the instrument (Stewart, 2007). Table 5 reports the p-value of the overidentifying restrictions based on the Sargan test. The result suggests that the overidentifying restrictions are not rejected at any conventional significance level (p-value=0.979). This specification generates linear conditional expectations and the estimated parameters directly correspond to the average partial effects. The *ape* of state dependence is around 0.36 which is larger than the results reported above. However, it is positive and highly significant, which is in line with the

---

[7] We have estimated also a system-GMM specification and found similar results (Blundell and Bond, 1998). The results can be provided from authors upon request.

previous results. This result also confirms that there is a substantial structural state dependence on the informality status of workers in Ukraine.

Unequal durations between periods - We only have three years of panel data and these were collected in the years 2003, 2004 and 2007. The time duration for 2003-2004 and 2004-2007 differs. This may bias the estimated parameter of structural state dependence. In order to deal with this issue we use a *history module* covered by the 2007 questionnaire about events that take place between 2004 and 2007. To test the sensitivity of previous results we estimate models using three indicator variables for any job, marital status and residential changes that occurred between 2004 and 2007. We then compare the magnitude of the structural state dependence and the variance of unobserved individual effects. The results are presented in the third, fourth and fifth columns of Table 5 (Models II-IV). We find that the estimated parameters are almost the same as before. However, the estimated structural state dependence is slightly reduced (except Heckman's reduced-form approximation specification). In order to check the sensitivity of the results further we calculate the *ape* as previously reported. These results are presented in Table 6 for the whole sample and for some groups of interest. The estimated *ape* is slightly lower compared to Table 4 but the main results remain stable.

Table 5 about here

6. Conclusions

The objective of this paper is to examine the dynamics of informality in the Ukrainian labor market. We attempt to identify persistence or structural state dependence due to past informality experience and to disentangle it from other sources such as the persistence due to unobserved time-invariant individual characteristics. To address this issue we estimate a dynamic panel data random-effects model with endogenous initial values and quasi-fixed-effects specification based on three waves of genuine panel data on informality.

The model which assumes initial informality status of individuals in the labor market to be exogenous substantially overestimates the persistence due to past participation in the informal sector. This assumption leads to zero variance for the unobserved influences on informality status. Once, we account for endogenous initial values using Heckman's reduced-form approximation, we find a significant degree of heterogeneity among the individuals in our sample and a lower degree of persistence due to past informality experience. We calculated individual partial effects for the whole population and for various subgroups and find that the effect of past informality experience on the current informality status is highly significant leading to 7-9 percentage points larger probability to be employed in the informal sector, *ceteris paribus*. The effect of the past informality experience is more pronounced for young single and less educated males.

One of the limitations of this study is the short time dimension of the panel dataset. This dataset, according to our knowledge, is one of the few genuine panel data, which make it possible to study the dynamics of informality in the developing and transition country contexts. However, we cannot assess the precise nature of structural state dependence because the size of the sample reduces substantially when we consider the information on the reasons for being in the informal sector (e.g. whether non-registration is voluntary or involuntary, following Dohmen et al., 2011). We also have not included transitions to and from unemployment or inactivity. With these limitations in mind our results still provide significant insights into the nature of informality over time and the persistence of informality in the labor market in a transition country. Hence, current informality experience of an individual seems to have future effects on the individual's labor market participation in the informal sector. Our results suggest that policies attempting to reduce current levels of informality may have a long-lasting effect on the labor market.

References

Akay, A. (2011). Finite Sample Comparison of Alternative Methods for Estimating Dynamic Panel Data Models, *forthcoming in Journal of Applied Econometrics*.

Alessie, R., S., Hochguertel, and A. van Soest (2004). Ownership of Stocks and Mutual Funds: A Panel Data Analysis. *The Review of Economics and Statistics*, 86: 783-796.

Arellano, M., and S. Bond (1991). Some Tests of Specification for Panel Data: Monte Carlo Evidence and an Application to Employment Equations, *The Review of Economic Studies*, 58: 277-297.

Arias, O., and M. Khamis (2008). Comparative Advantage, Segmentation and Informal Earnings: A Marginal Treatment Effects Approach. IZA Discussion Paper No. 3916.

Arulampalam, W., and M. Stewart (2009). Simplified Implementation of then Heckman Estimator of the Dynamic Probit Model and a Comparison with Alternative Estimators, *Oxford Bulletin of Economics and Statistics* 71: 659-681.

Blundell, R., and S. Bond (1998). Initial Conditions and Moment Restrictions in Dynamic Panel Data Models, *Journal of Econometrics*, 87: 115-143.

Chamberlain, G. (1984). Panel Data, in Z. Griliches, and M. D. Intriligator (eds) *Handbook of Econometrics* 2, Amsterdam: North Holland: 1247-1320.

De Mel S., D. McKenzie and C. Woodruff (2010). Who Are the Microenterprise Owners? Evidence from Sri Lanka and Tokman v. de Soto. In *International Differences in Entrepreneurship*, Lerner, J, Schoar, A (eds). NBER.

De Soto H. (1989). The Other Path—The Invisible Revolution in the Third World. I.B. Tauris & Co Ltd Publishers: London.

Dohmen, T., M. Khamis and H. Lehmann (2011). Risk Attitudes and the Incidence of Informality among Workers. Manuscript.

Heckman, J.J. (1978). Simple statistical models for discrete panel data developed and applied to test the hypothesis of true state dependence against the hypothesis of spurious state dependence., Annales de l.INSEE 30-31: 227-269.

Heckman J.J. (1981a). Heterogeneity and state dependence. In *Studies in Labor Markets*, Rosen S (ed.). University of Chicago Press: Chicago, IL.

Heckman, J.J. (1981b). The Incidental Parameters Problem and the Problem of Initial Conditions in Estimating a Discrete Time-Discrete Data Stochastic Process, in C. Manski, and D. McFadden (eds) *Structural Analysis of Discrete Panel Data with Econometric Applications*, Cambridge: MIT Press: 179-196.

Hsiao, C. (2003). *Analysis of Panel Data*, 2nd edition, New York: Cambridge University Press.

Kanbur, R. (2009). "Conceptualizing Informality: Regulation and Enforcement." *Indian Journal of Labour Economics*, 52: 33-42.

Lehmann, H., and N. Pignatti (2007). Informal Employment and Labor Market Segmentation in Transition Economies: Evidence from Ukraine. IZA Discussion Paper No.3269.

Levy S. (2008). Good Intentions, Bad Outcomes: Social Policy, Informality and Economic Growth in Mexico. Brookings Institution Press: Washington, D.C.

Maloney, W. F. (2004). Informality Revisited. *World Development*, 32: 1159-1178.

Stewart M. (2007). The Interrelated Dynamics of Unemployment and Low-Wage Unemployment", *Journal of Applied Econometrics*, 22: 511-531.

Wooldridge, J.M. (2005). Simple Solutions to the Initial Conditions Problem in Dynamic, Nonlinear Panel-Data Models with Unobserved Heterogeneity, *Journal of Applied Econometrics* 20: 39-54.

World Bank. (2007). Informality: exit and exclusion. Washington D.C.

Table 1. Descriptive Statistics

	Whole sample	Informal Workers	Formal Workers	2003	2004	2007
Informal dummy	0.041	…	…	0.038	0.044	0.040
	(0.198)	…	…	(0.192)	(0.206)	(0.196)
Log(household income)	6.797	6.795	6.797	6.237	6.627	7.528
	(0.989)	(0.863)	(1.000)	(0.793)	(0.619)	(1.023)
Married	0.704	0.559	0.717	0.701	0.687	0.725
	(0.456)	(0.497)	(0.450)	(0.458)	(0.464)	(0.447)
Single	0.130	0.265	0.118	0.148	0.136	0.106
	(0.336)	(0.442)	(0.323)	(0.356)	(0.342)	(0.308)
Divorced/Separated/Widowed	0.166	0.176	0.165	0.151	0.178	0.169
	(0.372)	(0.381)	(0.371)	(0.358)	(0.382)	(0.375)
Female	0.527	0.426	0.536	0.527	0.526	0.526
	(0.499)	(0.495)	(0.499)	(0.499)	(0.499)	(0.499)
Male	0.473	0.574	0.464	0.473	0.474	0.474
	(0.499)	(0.495)	(0.499)	(0.499)	(0.499)	(0.499)
Education level	6.554	5.481	6.649	6.320	6.609	6.733
	(1.829)	(1.860)	(1.796)	(1.969)	(1.771)	(1.714)
Age	42.429	35.972	43.001	40.755	41.774	44.758
	(10.743)	(10.028)	(10.617)	(10.606)	(10.619)	(10.605)
Regions						
Kiev	0.052	0.046	0.052	0.052	0.052	0.052
	(0.222)	(0.210)	(0.223)	(0.222)	(0.222)	(0.222)
Center	0.242	0.167	0.248	0.242	0.242	0.242
	(0.428)	(0.373)	(0.432)	(0.428)	(0.428)	(0.428)
West	0.152	0.093	0.157	0.152	0.152	0.152
	(0.359)	(0.290)	(0.364)	(0.359)	(0.359)	(0.359)
East	0.308	0.343	0.305	0.308	0.308	0.308
	(0.462)	(0.475)	(0.460)	(0.462)	(0.462)	(0.462)
South	0.246	0.352	0.237	0.246	0.246	0.246
	(0.431)	(0.478)	(0.425)	(0.431)	(0.431)	(0.431)

Sectors						
Agriculture	0.080	0.102	0.078	0.084	0.089	0.066
	(0.271)	(0.303)	(0.268)	(0.278)	(0.285)	(0.248)
Industry	0.257	0.204	0.262	0.259	0.262	0.251
	(0.437)	(0.403)	(0.440)	(0.438)	(0.440)	(0.434)
Sales	0.078	0.321	0.057	0.081	0.073	0.081
	(0.268)	(0.468)	(0.231)	(0.272)	(0.260)	(0.272)
Transportation	0.102	0.068	0.105	0.098	0.106	0.102
	(0.303)	(0.252)	(0.306)	(0.297)	(0.308)	(0.302)
Public Administration	0.047	0.015	0.050	0.042	0.048	0.051
	(0.212)	(0.123)	(0.218)	(0.201)	(0.214)	(0.220)
Education	0.265	0.037	0.285	0.269	0.262	0.264
	(0.441)	(0.189)	(0.451)	(0.444)	(0.440)	(0.441)
Services	0.064	0.114	0.059	0.067	0.061	0.063
	(0.244)	(0.319)	(0.236)	(0.250)	(0.239)	(0.244)
Other	0.107	0.139	0.105	0.100	0.099	0.123
	(0.310)	(0.346)	(0.306)	(0.300)	(0.298)	(0.329)
History 2004-2007						
Job leavers	0.187	0.565	0.153	…	…	…
	(0.390)	(0.497)	(0.360)	…	…	…
Marital status change	0.025	0.028	0.025	…	…	…
	(0.156)	(0.165)	(0.155)	…	…	…
Residential change	0.030	0.046	0.029	…	…	…
	(0.171)	(0.210)	(0.167)	…	…	…
#Observations	3984	324	3660	1328	1328	1328

Note: Standard deviations are in parenthesis. Source: Ukrainian Longitudinal Monitoring Survey (2003, 2004 and 2007).

Table 2. Run patterns of informality in Ukraine

Run patterns			#Observations	Percent%
2003	2004	2007		
0	0	0	3660	91.87
0	0	1	78	1.96
0	1	0	66	1.66
0	1	1	27	0.68
1	0	0	51	1.28
1	0	1	18	0.45
1	1	0	48	1.20
1	1	1	36	0.90
			3984	

Note: 1 is informally employed and 0 indicates formal employment based on the above definitions.

Table 3. Main results

Variables	I The Static probit model (Correlated random-effects model)	II The dynamic probit model (correlated random-effects and exogenous initial values)	III The dynamic probit model (Wooldridge method)	IV The dynamic probit model (Heckman's method with correlated random-effects)
Lagged informality status =1	…	1.398***	0.842***	0.802***
		(0.146)	(0.210)	(0.194)
Log(household income)	-0.030	0.010	0.007	-0.002
	(0.103)	(0.092)	(0.092)	(0.072)
Married=1	-0.276*	-0.051	-0.004	-0.052
	(0.201)	(0.160)	(0.165)	(0.078)
Single=1	-0.066	0.030	0.078	0.058
	(0.251)	(0.199)	(0.203)	(0.090)
Female=1	-0.054	-0.136	-0.174*	-0.168*
	(0.172)	(0.126)	(0.129)	(0.104)
Education	-0.207***	-0.145***	-0.145***	-0.175***
	(0.042)	(0.033)	(0.034)	(0.034)
Age	0.103	0.058	0.057	0.065
	(0.133)	(0.152)	(0.153)	(0.066)
Age-squared	-0.001	-0.001	-0.001	-0.001
	(0.002)	(0.002)	(0.002)	(0.001)
Center=1	0.156	0.294	0.390	0.354*
	(0.409)	(0.332)	(0.350)	(0.240)
West=1	0.002	0.263	0.399	0.309
	(0.434)	(0.347)	(0.365)	(0.259)
East=1	0.429	0.469*	0.567**	0.558***
	(0.387)	(0.318)	(0.337)	(0.238)
South=1	0.648**	0.625**	0.750**	0.767***
	(0.390)	(0.318)	(0.338)	(0.232)
Nuisance parameters				
Wooldridge's method	…	…	0.819***	…
Initial period informality			(0.217)	
Mean(household income)	0.085	-0.048	-0.072	-0.049

	(0.182)	(0.139)	(0.140)	(0.104)
Mean(age)	-0.040	0.012	0.019	0.005
	(0.144)	(0.153)	(0.155)	(0.077)
Mean(age-squared)	0.000	0.000	0.000	0.000
	(0.002)	(0.002)	(0.002)	(0.001)
Heckman's method Reduced-form equation				
Age	…	…	…	-0.057***
				(0.013)
Log-household income	…	…	…	0.122***
				(0.049)
Education	…	…	…	-0.201***
				(0.062)
Married	…	…	…	-0.760***
				(0.235)
Single	…	…	…	-0.583**
				(0.277)
Center	…	…	…	-0.547**
				(0.242)
West	…	…	…	-1.067***
				(0.259)
East	…	…	…	-0.352**
				(0.196)
South	…	…	…	-0.132
				(0.154)
Constant	…	…	…	0.057
				(0.069)
ψ	…	…	…	3.447***
				(0.488)
Variance of the unobserved effects	1.136***	<0.001	0.002	0.488***
	(0.131)	(0.039)	(0.622)	(0.102)
#Observations	3984	2656	2656	2656
Wald chi-squared test (p-value)	0.000	0.000	0.000	0.000
Log-likelihood	-471.87	-307.77	-300.78	-494.31

Sector dummies (7 dummies)	yes	yes	yes	yes

Note: We use 30 quadrature points in Gauss-Hermite quadrature for the integrals in likelihood functions. [***], [**], [*] indicate significance levels at 0.01, 0.05 and 0.10, respectively. Standard errors are in parenthesis.

Table 4. Average partial effects of lagged informality status

Average partial effect of lagged informality status for…	Exogenous Initial Values		Wooldridge's Method		Heckman's method		#obs
Whole sample	0.204	***	0.088	**	0.067	*	3984
	(0.064)		(0.044)		(0.036)		
Sorting by socio-demographics							
Females	0.169	***	0.069	*	0.050	*	2098
	(0.059)		(0.036)		(0.029)		
Males	0.242	***	0.109	**	0.085	*	1886
	(0.070)		(0.052)		(0.044)		
Low educated	0.273	***	0.125	**	0.100	**	1910
	(0.074)		(0.057)		(0.050)		
High educated	0.140	**	0.054	*	0.036		2074
	(0.055)		(0.031)		(0.024)		
Married	0.193	***	0.082	*	0.060	*	2806
	(0.063)		(0.042)		(0.035)		
Single	0.288	***	0.136	**	0.115	**	518
	(0.076)		(0.061)		(0.050)		
Divorced/Separated/Widowed	0.182	***	0.077	**	0.058	*	660
	(0.060)		(0.038)		(0.032)		
Age<35	0.285	***	0.132	**	0.112	**	1002
	(0.074)		(0.059)		(0.050)		
Age>35 and Age<50	0.211	***	0.090	**	0.066	*	1905
	(0.066)		(0.045)		(0.038)		
Age>50	0.115	**	0.043	*	0.027		1077
	(0.050)		(0.027)		(0.021)		
Sorting by sectors							
Industry	0.186	***	0.073	*	0.052	*	1025
	(0.062)		(0.039)		(0.032)		
Sales	0.420	***	0.224	***	0.182	**	311
	(0.071)		(0.079)		(0.074)		
Transportation	0.220	***	0.090	*	0.067	*	406
	(0.076)		(0.050)		(0.041)		
Public Administration	0.132		0.049		0.033		188
	(0.084)		(0.042)		(0.030)		
Education	0.067	*	0.019		0.011		1055
	(0.040)		(0.017)		(0.010)		
Services	0.355	***	0.169	**	0.139	**	254
	(0.082)		(0.073)		(0.065)		

Note: [***], [**], [*] indicate significance levels at 0.01, 0.05 and 0.10, respectively. Standard errors are in parenthesis.

Table 5: Alternative estimators and the effect of unequal durations between periods in the panel dataset

Variables	Dynamic Linear Probability Model	Dynamic Probit Random-Effects Model with History Variables		
	I	II	III	IV
	First-difference GMM estimation Arellano-Bond	Exogenous initial values	Wooldridge's method	Heckman's method
Lagged informality status =1	0.358***	1.375***	0.789***	0.863***
	(0.077)	(0.148)	(0.208)	(0.210)
Wooldridge's method Initial period informality	…	…	0.897*** (0.218)	…
Job leavers between 2004 and 2007	…	0.711***	0.726***	0.765***
		(0.118)	(0.120)	(0.132)
Marital status change between 2004 and 2007	…	-1.279***	-1.504***	-1.428***
		(0.520)	(0.554)	(0.509)
Residential change between 2004 and 2007	…	0.142	0.234	0.216
		(0.284)	(0.284)	(0.228)
Log(household income)	-0.002	-0.003	-0.007	-0.013
	(0.007)	(0.098)	(0.098)	(0.084)
Married=1	0.002	-0.034	0.023	-0.028
	(0.026)	(0.167)	(0.173)	(0.136)
Single=1	0.013	0.059	0.117	0.092
	(0.032)	(0.206)	(0.211)	(0.148)
Female=1	…	-0.077	-0.118	-0.098
		(0.134)	(0.137)	(0.158)
Education	0.012	-0.161***	-0.163***	-0.187***
	(0.011)	(0.035)	(0.035)	(0.039)
Age	-0.025	0.053	0.054	0.058
	(0.011)	(0.157)	(0.159)	(0.100)
Age-squared	<0.0001	-0.001	-0.001	-0.001
	(0.0001)	(0.002)	(0.002)	(0.001)
Center	…	0.434	0.539*	0.499*
		(0.358)	(0.379)	(0.354)
West	…	0.450	0.606*	0.511*
		(0.375)	(0.396)	(0.368)

East	…	0.715**	0.836**	0.817***
		(0.345)	(0.367)	(0.338)
South	…	0.792**	0.941***	0.937***
		(0.344)	(0.367)	(0.335)
Sargan test (p-value)	0.979	…	…	…
Variance of the unobserved effects	…	0.001	0.003	0.429***
		(0.033)	(0.088)	(0.125)
#Observations	1328	2656	2656	2656
Wald chi-squared test (p-value)	0.000	0.000	0.000	0.000
Log-likelihood	…	401.84	703.66	700.71
Sector dummies (7 dummies)	yes	yes	yes	Yes

Note: We use 30 quadrature points in Gauss-Hermite quadrature for the integrals in likelihood functions. [***], [**], [*] indicate significance levels at 0.01, 0.05 and 0.10, respectively. The last three models (II, III, and IV) are specified using Chamberlain approach for the unobserved individual effects. Nuisance parameters are not presented here but they can be reported upon request. Standard errors are in parenthesis.

Table 6. Sensitivity analysis: average partial effects

Average partial effect of lagged informality status for…	Exogenous Initial Values		Wooldridge's Method		Heckman's method		#obs
Whole sample	0.176	***	0.071	*	0.070	*	3984
	(0.060)		(0.037)		(0.036)		
Sorting by socio-demographics							
Females	0.145	***	0.056	*	0.054	*	2098
	(0.054)		(0.030)		(0.029)		
Males	0.210	***	0.088	**	0.088	**	1886
	(0.067)		(0.044)		(0.044)		
Low educated	0.240	***	0.103	**	0.104	**	1910
	(0.071)		(0.050)		(0.049)		
High educated	0.117	**	0.041	*	0.039	*	2074
	(0.049)		(0.025)		(0.024)		
Married	0.167	***	0.066	*	0.064	*	2806
	(0.058)		(0.035)		(0.034)		
Single	0.246	***	0.111	**	0.116	**	518
	(0.072)		(0.051)		(0.053)		
Divorced/Separated/Widowed	0.159	***	0.063	*	0.062	**	660
	(0.055)		(0.032)		(0.031)		
Age<35	0.250	***	0.111	**	0.117	**	1002
	(0.072)		(0.051)		(0.054)		
Age>35 and Age<50	0.180	***	0.071	*	0.068	*	1905
	(0.061)		(0.037)		(0.037)		
Age>50	0.100	**	0.035		0.031	*	1077
	(0.046)		(0.023)		(0.018)		

Note: [***], [**], [*] indicate significance levels at 0.01, 0.05 and 0.10, respectively. Standard errors are in parenthesis.