Predicting gross wages of non-employed persons in Croatia

SLAVKO BEZEREDI, univ. spec. oec.*
IVICA URBAN, PhD*

Article**
JEL: J31, C21, C52
doi: 10.3326/fintp.40.1.1

* This work has been supported in part by the Croatian Science Foundation under project number UIP-2014-09-4096. The authors express gratitude to the reviewers, whose comments have motivated significant additions and upgrades to the basic version of the paper.
** Received: September 22, 2015
Accepted: January 7, 2016

Slavko BEZEREDI
Institute of Public Finance, Smičiklasova 21, 10000 Zagreb, Croatia
e-mail: slavko.bezeredi@ijf.hr

Ivica URBAN
Institute of Public Finance, Smičiklasova 21, 10000 Zagreb, Croatia
e-mail: ivica.urban@ijf.hr

FINANCIAL THEORY AND PRACTICE
40 (1) 1-61 (2016)

Abstract

We present the findings of a study aimed at building a model for predicting wages of non-employed persons in Croatia. The predictions will be used in the calculation of marginal effective tax rate at the extensive margin and in labour supply modelling. The database used is 2012 "EU statistics on income and living conditions". The paper comprehensively explains the data source, variables, subgroups of employed and non-employed, and the results of the linear regression model, the Heckman selection model and the quantile regression model. The quality of predictions obtained by different models is compared and discussed.

Keywords: gross wages, estimation, prediction, unemployed, inactive, Heckman selection model, quantile regressions, Croatia

1 INTRODUCTION

This paper presents the findings of a study aimed at building a model for predicting gross wages of non-employed persons in Croatia, using "EU statistics on income and living conditions" (henceforth SILC) data. These wage predictions will be primarily used as inputs in further research: (a) for the calculation of marginal effective tax rate at the extensive margin (METREM), and (b) for the estimation of discrete choice labour supply models.

METREM measures the net benefit of a household occurring in a hypothetical situation, in which a non-employed person enters employment. The transition from non-employment to employment has a complex effect on household income; social benefits are typically reduced or extinguished, which decreases the gain obtained from employment. Furthermore, part of a gross wage is taxed away in terms of personal income tax and social insurance contributions. The traditional approach computes METREM for several "model family types" (e.g., a single person or a couple with one earner and two children aged 12 years).[1] Such an approach provides a good description of how the tax-benefit system affects household net income, but ignores the heterogeneity of family and personal characteristics in the population. To provide an accurate picture of the distribution of METREM, real datasets and *tax-benefit microsimulation models* should be used in estimation.[2]

EUROMOD is the tax-benefit microsimulation model for the European Union, which provides cross-country comparable measures of direct taxes and social insurance contributions liabilities. The model also provides cash benefit entitlements for the household population of EU member states (Figari et al., 2014). Beginning in 2016, EUROMOD will include the module for simulating the Croatian tax-benefit system. MICROMOD is the tax-benefit microsimulation model

[1] Carone et al. (2004) perform such calculations for OECD countries. Bejaković et al. (2012) calculate METREM for eight hypothetical family types in Croatia. The analysis indicated that some family types have very high METREM (near 100%), such as two-adult families, in which both adults are non-employed, and families with three or more children. For non-employed persons in these families "work does not pay" because the withdrawal of benefits is almost as high as the gain from the net wage.

[2] Such a model is proposed by Immervoll and O'Donoghue (2002), who calculate METREM using EUROMOD.

for Croatia, which will be based on EUROMOD and contain additional elements concerning local government benefits and labour supply estimation.[3]

For most countries, EUROMOD uses SILC data. SILC data for Croatia are based on the survey "Anketa o dohotku stanovništva", compiled since 2010 by the Croatian Bureau of Statistics (CBS) and used by CBS to calculate measures of poverty and living standards (CBS, 2013b).

The effects of the tax-benefit system on household income are evaluated using the microsimulation models of taxes and benefits. However, these models do not *per se* provide one of the key variables needed for the calculation of METREM – *the amount of gross wage that could be earned by a non-employed person who enters employment*. SILC contains information only on wages earned in the income reference period. For persons who have not been working in this period, data on wages are missing.

Wages of non-employed persons can be predicted from the wage equation, which describes the functional relationship between the wage and personal characteristics such as age, marital status, place of living, work experience and occupation. Industry of employment and job characteristics can also be added as independent variables. The wage equation is typically modelled within the linear regression model (LRM) and estimated using the sample of employed persons.[4]

However, such an approach, which uses only data on employed persons, is challenged by Heckman (1976, 1979), who introduces the concept of "sample selection problem". Namely, the wage equation coefficients obtained by the abovementioned model may be biased because the sample is not representative of the whole population. Heckman suggests a model that identifies and corrects the sample selection problem. This model consists of a wage equation and a participation equation, in which the latter estimates the probability of a person to be employed vs. non-employed. The random terms in the wage and participation equation represent unobservable characteristics influencing wage and probability of employment, respectively. If these random terms are correlated, the sample selection problem exists and wage equation parameters must be "corrected".

The Heckman selection model (HSM) has achieved huge popularity among researchers and is widely used in wage estimations. Two areas of application are most frequent: (a) prediction of wages of non-employed persons for labour supply modelling[5], and (b) estimation of the gender wage gap and other wage differen-

[3] MICROMOD will be developed in a project, "Application of Microsimulation Models in the Analysis of Taxes and Social Benefits in Croatia" (Institute of Public Finance). For more details, see: http://www.ijf.hr/eng/research/croatian-science-foundation-projects/1053/ammatsbc/1062/.

[4] The relevant studies for Croatia include Nestić et al. (2015) and Nestić (2005).

[5] A small excerpt of these studies includes van Soest (1995; for the Netherlands in 1987), Labeaga et al. (2008; for Spain in the late 1990s), Pacifico (2009; for Italy in 2002), Berger et al. (2011; for Luxembourg in 2004), Bičáková et al. (2011; for the Czech Republic in 2002), and Mojsoska-Blazevski et al. (2013; for Macedonia in 2011).

tials.[6] HSM has become a standard tool in wage estimation despite criticism and the emergence of alternative approaches for addressing the selection problem (Winship and Mare, 1992; Vella, 1998; Puhani, 2000).

In this study, we use several methods for gross wage estimation and prediction: LRM, HSM and quantile regression model. For HSM estimation, non-employed persons are partitioned into several distinctive groups. Results from different models are compared to reveal the advantages and weaknesses of different methods. The comparison is done by the analysis of residuals and density estimates of predicted wage distributions. Although SILC data for Croatia exist for several years, this study is the first comprehensive work to employ them. Therefore, our descriptions can be useful in promoting wider use of this valuable data source. During the research, we faced several methodological issues, most of them recurring in the literature. In this paper, suggestions are provided on how these problems can be addressed, but further research is needed to address them completely.

The structure of the paper is as follows. Section 2 is devoted to methodological issues. The first part describes LRM and HSM. Then, the formulas for gross wage prediction are derived. A discussion of specification issues in HSM follows. In the last part of this section, the mentioned methodological challenges are discussed. Section 3 first provides a description of the Croatian SILC and the variables constructed for use in regression models. A brief overview follows on the structure of working population based on SILC, which serves as an introduction into the procedure for shaping the subgroups of employed and non-employed persons. Descriptive analysis of the variables by subgroups is then presented. Section 4 analyses participation in employment and non-employment using the probit method. The prediction quality of probit models is assessed using classification tables and several measures of fit. The key results of the paper are found in section 5, which presents the estimates of the wage equation using LRM, quantile regressions model and HSM. Predictions from all of these models are compared for employed and non-employed persons. Section 6 discusses the results and concludes.

2 METHODS FOR WAGE ESTIMATION AND PREDICTION

2.1 LINEAR REGRESSION MODEL AND HECKMAN SELECTION MODEL

The standard approach in econometric modelling of wages assumes that the natural logarithm of wage, w_i, of each person i is linearly dependent on variables that describe C personal characteristics, which are summarised by $X_i = [1, x_{i1}, ..., x_{iC}]$. The relationship between w_i and X_i is called the *wage equation* and is written as follows:

$$w_i = X_i\alpha + \varepsilon_i \tag{1}$$

[6] See Paci and Reilly (2004; for Albania, Bosnia and Herzegovina, Bulgaria, Poland, Serbia, Tajikistan, Uzbekistan in the early 2000s), Pastore and Verashchagina (2008; for Belarus in 1996 and 2001), Khitarishvili (2009; for Georgia in 2000 and 2004), and Avlijaš et al. (2013; for Serbia, Montenegro and Macedonia in the mid-2000s).

where $\alpha = [\alpha_0, \alpha_1, ..., \alpha_C]$ is the set of coefficients common to the whole population. ε_i is the random term, which captures "unobservable characteristics", i.e., the part of wage that is not described by $X_i\alpha$; $\varepsilon_i \sim N(0, \sigma_{\varepsilon\varepsilon})$, i.e., ε_i is normally distributed with variance $\sigma_{\varepsilon\varepsilon}$, and the expected value of ε_i is $E(\varepsilon_i | X_i) = 0$.

By population, we mean all persons in a society, either employed or non-employed.[7] Because $E(\varepsilon_i | X_i) = 0$, the expected wage of a person i randomly drawn from population equals the following:

$$E(w_i | X_i) = X_i\alpha + E(\varepsilon_i | X_i) = X_i\alpha \tag{2}$$

In an attempt to estimate relationship (1), data on actual wages are typically used. Precisely for this reason, data on wages are usually available only for employed people, whereas they are missing for non-employed persons. Thus, the sample of I observations, $i = \{1, ..., I\}$, randomly drawn from the population can be sorted and divided into two parts: K employed persons, $i = \{1, ..., K\}$, and $I - K$ non-employed persons, $i = \{K + 1, ..., I\}$.

Let us assume that we know and correctly measure all of the elements of X_i, and, furthermore, that actual wages reflect, in general, true earning potential. The linear regression model (LRM), using the ordinary least squares method on subsample $i = \{1, ..., K\}$, will provide estimates $\tilde{\alpha} = f\{w_i, X_i; i = 1, ..., K\}$ of true coefficients α. Are estimates $\tilde{\alpha}$ unbiased? According to Heckman (1976, 1979), they may not be because the sample used in estimation, $i = \{1, ..., K\}$, covers only employed persons. Thus, information concerning non-employed persons, $i = \{K + 1, ..., I\}$, is excluded from estimation. As Heckman notes, the expected wage of employed person i, $i = 1, ..., K$, equals the following:

$$E(w_i | X_i, \text{employed}) = X_i\alpha + E(\varepsilon_i | \text{employed}) \tag{3}$$

which is different from $E(w_i | X_i) = X_i\alpha$ in equation (2). In equation (3), Heckman (1979) introduces the concept of "sample selection rule", which implies that the expected wage not only depends on X_i but also on how sample $i = \{1, ..., K\}$ is chosen. To obtain the proper estimates of α based on the available wage data, he proposes the following two-equation model:[8]

$$w_i = X_i\alpha + e_i \tag{4}$$

$$p_i = Z_i\beta + u_i \tag{5}$$

[7] Thus, a certain wage is attributed to everybody, and in this sense, wage w_i is a hypothetical construct, embodying human abilities and corresponding earning potential.

[8] In this presentation of the model, we follow Heckman (1979), with slight adaptation to our wage case. Heckman selection model is extensively used and studied. For textbook presentations, see e.g. Amemiya (1985), Verbeek (2004), Cameron and Trivedi (2005), and Green (2008). For critical reviews, see, e.g., Winship and Mare (1992), Vella (1998), Puhani (2000), Nicaise (2001), Bushway et al. (2007), and Breunig and Mercante (2010).

Equation (4) corresponds to the *wage equation* in (1); $\varepsilon_i \sim N(0, \sigma_{\varepsilon\varepsilon})$ is a random term analogous to ε_i. Equation (5) is the *participation equation*, which describes the relationship between $C + D$ personal characteristics, $Z_i = [1, x_{i1}, ..., x_{iC}, y_{i1}, ..., y_{iD}]$, and the person's employment or non-employment status. β is the set of common coefficients, and $u_i \sim N(0, \sigma_{uu})$ is a random term with variance σ_{uu}, having similar interpretation as e_i and ε_i. If $p_i \geq 0$, a person is employed; the person is non-employed if $p_i < 0$. Denote with $\sigma_e = (\sigma_{ee})^{1/2}$ and $\sigma_u = (\sigma_{uu})^{1/2}$ the standard deviations of e_i and u_i, respectively. The covariance and correlation terms are presented by σ_{eu} and $\rho_{eu} = \sigma_{eu} / (\sigma_e \sigma_u)$, respectively.

Recall the "sample selection rule" from equation (3). The person is employed if $p_i \geq 0$, i.e., if $u_i \geq -Z_i\beta$. Therefore, equation (3) is rewritten as follows:

$$E(w_i \mid X_i, u_i > -Z_i\beta) = X_i\alpha + E(\varepsilon_i \mid u_i > -Z_i\beta) \tag{6}$$

The term $E(\varepsilon_i \mid u_i > -Z_i\beta)$ is not equal to zero if there exists a correlation between unobservable characteristics e_i and u_i. Heckman (1979) obtains the value of $E(\varepsilon_i \mid u_i > -Z_i\beta)$, and equation becomes the following:

$$E(w_i \mid X_i, u_i > -Z_i\beta) = X_i\alpha + \frac{\sigma_{eu}}{\sigma_u}\frac{\phi(Z_i\beta)}{\Phi(Z_i\beta)} \tag{7}$$

where $\phi(\cdot)$ is the standard normal p.d.f. and $\Phi(\cdot)$ is the standard normal c.d.f., i.e., the probability that a person is employed. Commonly, ratio $\lambda_i = \phi(Z_i\beta) / \Phi(Z_i\beta)$ is called the "Heckman's lambda" for person i. λ_i are non-negative, monotonically decreasing and convex in $Z_i\beta$. Because $\sigma_{eu} = \rho_{eu}(\sigma_e \sigma_u)$, we have that $\sigma_{eu} / \sigma_u = \rho_{eu}\sigma_e = \Lambda$. If unobservable characteristics, represented by e_i and u_i are correlated, that will be reflected in $\rho_{eu} \neq 0$ ($\sigma_{eu} \neq 0$) and consequently in $\Lambda \neq 0$.

There are two ways to estimate HSM from equations (4) and (5): maximum likelihood and the "two-step procedure". For differences between these approaches, see, e.g., Verbeek (2004). We employ the maximum likelihood estimation using the Stata program "Heckman selection model (ML)" (command *heckman*), which provides us with estimates $\hat{\alpha}$, $\hat{\beta}$, $\hat{\rho}_{eu}$, $\hat{\sigma}_e$ and $\hat{\Lambda} = \hat{\rho}_{eu}\hat{\sigma}_e$ and with their standard errors (for more details, see section 5.4).

2.2 WAGE PREDICTION FORMULAS

The coefficients $\hat{\alpha}$ should be unbiased and consistent estimators of true coefficients α from equation (1). Following Breunig and Mercante (2010), we define three sets of wage predictions based on HSM:

(1) Unconditional predictions, applicable to the entire sample:

$$\hat{w}_i^{HUC} = X_i\hat{\alpha} \tag{8}$$

(2) Conditional predictions for employed only, defined as follows:

$$\hat{w}_i^{HCE} = X_i\hat{\alpha} + \hat{\Lambda}\frac{\phi(Z_i\hat{\beta})}{\Phi(Z_i\hat{\beta})} \tag{9}$$

(3) Conditional predictions for non-employed only, defined as follows:

$$\hat{w}_i^{HCN} = X_i\hat{\alpha} + \hat{\Lambda}\frac{-\phi(Z_i\hat{\beta})}{1-\Phi(Z_i\hat{\beta})} \tag{10}$$

We also use the wage predictions based on LRM, estimated for the subsample of employed persons, and applicable to the whole sample:

$$\hat{w}_i^{LRM} = X_i\hat{\alpha} \tag{11}$$

The correlation between unobservable characteristics in the participation and wage equations can be either positive ($\hat{\rho}_{eu} > 0 \Rightarrow \hat{\Lambda} > 0$) or negative ($\hat{\rho}_{eu} < 0 \Rightarrow \hat{\Lambda} < 0$). Both cases appear in the empirical literature.[9] If $\hat{\Lambda} < 0$, the predictions $\hat{w}_i^{HUC}$ from equation (8) will be generally greater than the predictions $\hat{w}_i^{LRM}$; furthermore, if $\hat{\Lambda} < 0$, predictions $\hat{w}_i^{HCN}$ from equation (10), obtained for non-employed persons, will be greater than the predictions $\hat{w}_i^{HUC}$ because $-\phi(Z_i\hat{\beta})/(1-\Phi(Z_i\hat{\beta}))$ is always non-positive.

2.3 SPECIFICATION ISSUES IN THE HECKMAN SELECTION MODEL

HSM requires proper specification of both participation and wage equation, i.e., the right choice of the characteristics in X_i and Z_i. Note that Z_i captures all elements of X_i and introduces D additional personal characteristics, y_{i1}, ..., y_{iD}. According to Verbeek (2004), economic arguments require that all elements of X_i are included into Z_i. Conversely, elements y_{i1}, ..., y_{iD} should capture only those characteristics that are not statistically or economically important in the wage equation.

Selecting the model variables for HSM represents a sensible task. If some important variable is omitted from the participation and wage equations, the correlation between error terms, σ_{eu}, will be incorrectly assessed, and the method will suggest misleading values of α. In choosing the variables, we follow the research of others (see references in footnotes 5 and 6) and create a comprehensive set of characteristics, given the availability of data in SILC (see section 3.2).

Section 2.1 speaks generally about the "population" and distinguishes between "employed" and "non-employed". In practice, it is necessary to define precisely what these groups represent. "Non-employed" are a heterogeneous group consist-

[9] Nicaise (2001) explains the phenomenon of negative Λ using the "crowding hypothesis" from labour economics theory. Namely, in periods of high unemployment, due to constraints on the demand side in the labour market, "individuals compete with each other by bidding down wages or by accepting jobs below their level of qualification". Thus, for example, in fear of becoming unemployed, persons with tertiary education may replace those with secondary education on jobs that commonly "belong" to the latter. This effect pushes the expected wage line (to be estimated for employed persons) below its "true" level; HSM should reveal the true line.

ing of individuals who have varying attachments to the labour market and different participation mechanisms. Correct specification of the participation equation requires that non-employed are divided into more homogeneous subgroups, such as unemployed, marginally employed, and the work-able inactive (Breunig and Mercante, 2010). In this study, working age persons are divided into *employed*, *unemployed*, *inactive* and other persons. A special procedure is created to form these subgroups (see section 3.4).

2.4 OTHER METHODOLOGICAL CHALLENGES

HSM is comprehensively used for predicting the wages of non-employed (see references in footnote 5). However, the predictive power and methodological issues concerning the general suitability of HSM for such a purpose have not been thoroughly investigated. One exception is Breunig and Mercante (2010), who conclude that LRM, which uses the subsample of employed persons, has greater predictive power than do HSM and several other selection models.[10]

Based on a literature review and our own investigation, we have identified several methodological issues related to predicting the wages of non-employed.[11] In this paper, we can provide only suggestions for the solutions to these problems; further research is needed to address them completely.

(1) Concerning interpretation of the results, Paci and Reilly (2004) note that the unconditional wage predictions, $\hat{w}_i^{HUC}$, do not represent "actual" wages that could be obtained at the market but rather the "wage offers" of persons randomly drawn from the population that are based on their personal characteristics. Therefore, we ask the following question: are the predictions $\hat{w}_i^{HUC}$ appropriate for use in the calculation of METREM?

Assuming that a non-employed person, who hypothetically enters employment, accepts the ongoing market wage, then the predictions $\hat{w}_i^{LRM}$ have more credibility than do $\hat{w}_i^{HUC}$ (or $\hat{w}_i^{HCN}$) because they reflect better the actual market wages.

(2) Both the HSM and simple wage equation models are concentrated on the "supply side", i.e., the personal characteristics that determine the supply of labour but neglect the "demand side" of the labour market, whose influence can be particularly important in recession periods (e.g., for Croatia in 2011).

The "demand side" can be partly incorporated into current models through the use of *occupation* variables, which may "transmit" the effects of low or high relative demand in particular areas on the wages.

[10] Breunig and Mercante (2010) claim that their paper is "the first to examine the question of the predictive power [of HSM] for the non-selected sample". In a thorough analysis for Australia, they employ HSM and several alternative selection models. They use longitudinal survey data, which enable them to analyse the persons who change their employment status over the period of several years and to compare the predicted wages for the periods of non-employment with actual wages obtained in employment.

[11] Some of these issues were suggested by our reviewers.

(3) Additional problems arise for models that address non-employed persons. Namely, wage predictions for non-employed imply the *ceteris paribus* assumption, according to which the hypothetically newly employed do not affect the overall wage setting mechanism. However, this assumption is obviously unwarranted; a large group of non-employed persons entering employment at a certain moment (given that the market can absorb them) would have a huge effect on all market wages.

In the calculation of METREM, an explicit assumption can be made, i.e., that the model analyses the *hypothetical* transition from non-employment to employment, in which *only one* person enters the market at a time. Such an event would have a negligible effect on the market wage.

(4) Both LRM and HSM consist of a single wage equation; for each variable, a single coefficient is estimated for all sample data. Thus, the partial effect of each variable on the wage is identical across the wage distribution. However, in reality, this assumption may not hold. Using LRM, Nestić (2005) finds for Croatia in 2003 that, controlled for various personal characteristics, the wage premium for employed in the widely defined public sector is 9%. However, the results of quantile regressions show that the premium for employees at the 10th percentile of wage distribution was 15%, for those at the 75th percentile 5%, and for those at the 90th percentile 0%. This evidence demonstrates that a "single" wage equation cannot capture different strengths of influences, particularly at the tails of a wage distribution.

Because we are specifically interested in low-potential wage earners (who are usually more likely to be non-employed), alternative approaches, such as quantile regressions, should be considered for predicting wages of non-employed (see section 5.3).

3 DATA, VARIABLES AND SUBSAMPLES

3.1 DATA SOURCE

The microdata used in this study come from the 2012 edition of Croatian SILC, which is compiled by the Croatian Bureau of Statistics (CBS) using data from the survey "Anketa o dohotku stanovništva" (ADS).[12]

SILC contains a rich set of variables describing demographic and socio-economic characteristics of persons. Because its primary role is the measurement of "income and living conditions", SILC offers a relatively detailed overview of different types of personal and household incomes.[13] However, compared with the Labour Force Survey (LFS), SILC is somewhat less detailed in respect to labour market variables. For example, SILC lacks data on the duration of unemployment or the type of ownership of the firm in which a person is employed.

[12] ADS was introduced in the Croatian statistical system in 2010 and is in line with EU regulations and Eurostat's methodology prescribed for the SILC surveys. For more details, see CBS (2013a, 2013b).
[13] For definitions of SILC variables, see Eurostat (2015).

An important feature of SILC is the "time discrepancy" in reference periods for different variables. Data on demographic characteristics and data on financial, social, and health situations refer to the date of the interview (DIN). Income data refer to the "income reference year" (IRY). Data on economic activity status are collected both for DIN and IRY. In our case, DIN is some date in 2012, and IRY is the entire year 2011.

The sample contains data for 5,838 households and 15,166 persons.[14] SILC contains sampling weights for each person in the sample, which enables the aggregation of sample data to the whole population level. These samples are used in all calculations and estimations in this paper.

3.2 VARIABLES ON PERSONAL CHARACTERISTICS AND INCOME

The description of variables considered in the analysis is shown in table A1 (appendix 2). The variables are divided into several categories: age, marital status, children, education, area of living, health, wage and income, employment, occupation and industry. In this subsection, we describe the main features of the variables, whereas the descriptive analysis of data follows in subsections 3.7, 4.1 and 5.1.

Age. The main variable (*ag_year*) refers to the age of a person in the middle of IRY (i.e., on 30 July 2011). Persons are also divided into four age groups (*ag_1525*, *ag_2540*, *ag_4055* and *ag_5565*).

Marital status. The variables describing marital status conform to formal rules and capture married (*ms_mard*), divorced (*ms_divo*) and widowed (*ms_widw*). A certain number of married persons do not live in households with their spouses, whereas a small number of divorced and widowed live with a partner in a household; these arrangements are not investigated further. However, for persons who claim the "never married" status, separate variables are created for those who live with a partner in a household (*ms_nmhp*) and for those who do not have a partner in a household (*ms_nmnp*).

Children. The children variables capture the numbers of own parents' children in three age groups: 0 to 2 years (*ch_p0002*), 3 to 6 years (*ch_p0306*) and 7 to 15 years (*ch_p0715)*. Under the assumption that the presence of other children in a household – not own parents' children but, for example, grandchildren and nephews – may affect the employment decision, an additional variable is introduced that represents the number of these children aged 0 to 15 years (*ch_o0015*).

Education. There are four basic educational variables relating to unfinished primary school (*ed_nopr*), finished primary school (*ed_prim*), secondary education (*ed_seco*) and tertiary education (*ed_tert*). Because the number of those with un-

[14] The sample used in this study is identical to the sample used in EUROMOD. For EUROMOD purposes, the original SILC 2012 sample is slightly changed; 33 non-respondent households and 18 children born in 2012 were excluded. For more details, see Urban and Bezeredi (2015).

finished primary school is quite small, a new variable (*ed_prnp*) joins them together with persons who have finished primary school.

Area of living. Detailed data on place of living are not available in SILC. However, SILC offers a variable that categorises the municipalities into three groups according to the number of inhabitants per square meter. Using these data, three variables are constructed (*ar_dens*, *ar_intr*, *ar_thin*), which are proxies for urban, semi-urban and rural areas, respectively.

Health. SILC contains several variables describing the self-perceived health status of a person. They are used to create the variables that denote persons with bad or very bad health (*hs_badh*) and persons whose usual activities are limited due to health problems (*hs_lima*).

Occupation. Occupational variables are based on the SILC variable, which refers to the main job of a currently employed person. This variable also registers occupation "held on the last main job" for people who currently do not have a job but have worked in the past. Therefore, for people who have never worked, information on occupation is not available. Among unemployed, and particularly among inactive, those who have never worked account for a high percentage. The use of occupational variables in such cases is not possible. Repercussions for the analysis will be discussed in section 3.5. Due to the relatively small number of persons in occupation "managers", they are joined with "professionals" (*oc_21*). For the same reason, persons in "armed forces occupations" are joined with "technicians and associate professionals" (*oc_30*).

Employment. Work experience (*we_yipw*) is measured as the number of years spent in paid work *before* the beginning of IRY. Another variable (*we_yopw*) represents the "inverse" of work experience, measuring the time out of work since the date when the first work experience was attained. See section 3.8 for a detailed analysis of these variables.

Several variables (*em_locs*, *em_locl*, *em_perj*, *em_mana*) describe the characteristics of the currently held job. "Agricultural household" (*em_agri*) denotes a household in which the primary source of market income comes from self-employment in agriculture.

Income and work. Gross wage captures earnings from employment paid in cash or near cash terms. To obtain the hourly gross wage, yearly gross wage is divided by the yearly number of working hours. Yearly working hours are obtained using information on months spent in work (during IRY) and usual number of work hours per week.

There are several variables capturing income obtained by the observed person's household. These variables cover a large portion of total household income, but the following items are excluded: (a) a person's own income from employment

and self-employment; (b) a person's own income from social insurance (unemployment and sickness benefits); and (c) social assistance benefits received by a person's household. In addition to usual cash incomes (*oi_a* to *oi_f*), one variable (*oi_g*) captures imputed rent from the use of a dwelling and serves as a proxy for the value of housing assets.

Industry. There are 21 industries overall according to NACE Rev. 2, but some industries are aggregated within SILC. Nestić et al. (2015, table D2b) use LFS to calculate the shares of employed persons by industry sector and the type of ownership in 2012. In sectors O, P and Q all persons are employed in the "narrower defined" public sector. Furthermore, in sectors D, E, H and R, the large majority of workers are employed by state-owned enterprises. One of the variables (*in_opq*) can serve as a proxy to employment in the "narrower defined" public sector.

3.3 STRUCTURE OF THE WORKING POPULATION BASED ON SILC

The working age population includes women aged 15 to 60 years and men aged 15 to 65 years. This definition is motivated by the fact that the statutory age for old-age retirement in 2011 is 60.25 (65) years for women (men). The lower limit of 15 years is the age when primary school is finished.

Table 1 presents the structure of the working age population, as defined above. This information is based on SILC questions about self-defined economic status, which is recorded at different time instances: (a) on DIN, and (b) in each month during IRY. The variables on activity status capture the person's own perception and are not comparable with LFS definitions of employment, unemployment, inactivity, and so forth. Henceforth, the quotation marks in the naming of activity statuses are used to signify that they are self-reported, and do not conform to usual economic and statistical definitions.

Section (a) of table 1 presents the structure according to economic status on DIN. For readers acquainted with the Croatian economy, a curious result is that there are 537 thousand "unemployed" persons. According to the Labour Force Survey (LFS), in 2012 there were approximately 300 thousand unemployed (year average), whereas the number of registered unemployed was 324 thousand (CBS, 2015). How many "unemployed" are unemployed when some of the LFS definitions apply? This number can be determined by checking the answers to several questions also available in SILC. Twenty-eight per cent of "unemployed" did not actively seek a job in the four weeks preceding the interview. Based solely on this fact, they would not be treated as unemployed, but rather as inactive. An additional 3% of the "unemployed" should not be treated as unemployed because they either (or both) worked at least 1 hour in the previous week and were not available for work in the subsequent 2 week period. Thus, the number of "unemployed" who comply with LFS definitions would be 370 thousand.

Section (b) of table 1 shows the structure based on the activity statuses during IRY for four groups of interest: "employed", "self-employed", "unemployed" and "persons

fulfilling domestic tasks and care responsibilities" (FDTCR). Figures show the total numbers of persons who report one of the mentioned statuses in at least one month in 2011. For example, 618 thousand persons were "unemployed" for one month or more. Each group is divided into three subgroups according to the number of months spent in the respective status. Thus, 86% of all "employed" were at work for all 12 months, whereas 67% of "unemployed" were out of work during the entire year.

Table 1

Structure of the working age population according to SILC 2012

	All		Women		Men	
	In thous.	%	In thous.	%	In thous.	%
Section (a) Current status in 2012						
All	2,591	100.0	1,250	100.0	1,341	100.0
"Employed"	1,196	46.2	565	45.2	631	47.0
"Self-employed"	139	5.4	40	3.2	99	7.4
"Unemployed"	537	20.7	271	21.7	266	19.8
"Pensioners"	308	11.9	112	9.0	195	14.6
"FDTCR"	109	4.2	108	8.6	1	0.1
"Unable to work"	21	0.8	7	0.6	14	1.0
"In education"	281	10.8	146	11.7	135	10.1
"Other inactive"	13	0.5	6	0.5	8	0.6
* *LFS unemployed*	370	14.3	176	14.1	194	14.5
Section (b) Status in 2011						
"Employed" for at least one month	1,293	100.0	611	100.0	682	100.0
12 months	1,110	85.9	524	85.7	587	86.0
7-11 months	72	5.6	34	5.5	39	5.7
1 to 6 months	110	8.5	53	8.8	57	8.3
"Self-employed" for at least one month	139	100.0	39	100.0	100	100.0
12 months	130	93.3	36	92.3	94	93.7
7-11 months	4	2.7	0	0.7	3	3.5
1 to 6 months	6	4.0	3	6.9	3	2.9
"Unemployed" for at least one month	617	100.0	307	100.0	310	100.0
12 months	414	67.1	212	68.9	202	65.2
7-11 months	90	14.6	45	14.5	45	14.6
1 to 6 months	113	18.4	51	16.6	62	20.1
"FDTCR" for at least one month	116	100.0	115	100.0	1	100.0
12 months	112	96.8	111	96.8	1	100.0
7-11 months	3	2.4	3	2.4	0	0.0
1 to 6 months	1	0.9	1	0.9	0	0.0

3.4 FORMING SUBSAMPLES OF EMPLOYED, UNEMPLOYED AND INACTIVE

This section describes a procedure that classifies SILC sample persons into one of three distinctive groups: *employed*, *unemployed* and *inactive*. We face two major problems here, both of them envisaged by the analysis in section 3.3. First, activity status is self-reported and for some persons does not correspond to the real one. Second, persons report their activity status in various time instances – on DIN and

in each month of IRY; for a significant number of persons, the status varies across the period (from January 2011 to DIN). Which time instance should be considered?

Concerning the latter issue, note that the working time-span of EUROMOD and MICROMOD is one year, i.e., these models consider incomes over the entire IRY. Therefore, the natural choice for definition of activity status is IRY and not DIN. Some persons change their status during IRY; in these cases, delineation rules must be provided. Concerning the issue of self-reported vs. real status, we use additional variables to determine the real status.

Our procedure is as follows. The starting sample, S0, captures the working age persons, defined as women aged 15 to 60 and men aged 15 to 65 years. Sample S1 is a subsample of S0 containing persons whose status was "employed", "self-employed", "unemployed" or "FDTCR" in at least one month during IRY. From sample S1, subsamples S2A and S2B are formed.

Subsample S2A consists of persons who were "employed" or "self-employed" for *9 months or more* in IRY. The members of S2A are then divided into two subgroups: (a) *employed* – containing persons whose prevalent status during IRY was "employed", and (b) *self-employed* – consisting of persons whose primary status was "self-employed".

Subsample S2B captures the remaining persons from S1 if their status is "unemployed" or/and "FDTCR" during IRY for at least one month. S2B is then divided into *unemployed* and *inactive* persons. *Unemployed* persons are members of S2B who satisfy any of the following conditions: (a) they are actively seeking a job on DIN, (b) they are not actively seeking a job on DIN, but have worked at least one month in IRY, or (c) they are "employed" or "self-employed" on DIN. *Inactive* persons are those members of S2B who do not belong to *unemployed*. The procedure used for forming of the subsamples of *employed*, *unemployed* and *inactive* is illustrated in figure A1 (appendix 3).

Note the following two features of the procedure:

(1) Persons are *unemployed* if they were "unemployed" or "FDTCR" even only one month during IRY. Furthermore, persons who worked during IRY (but not more than 8 months) can remain classified as *unemployed*. Table 5 shows the number of *unemployed* who have worked during part of the year.
(2) *Inactive* persons are those who (a) have not worked at all in IRY, (b) are out of work on DIN, and (c) were not actively seeking a job on DIN.

3.5 DIVISION INTO EXPERIENCED AND INEXPERIENCED

In comparison to the regular one-equation LRM, which contains the wage equation only, the Heckman selection model is much more complicated to build because it also introduces the selection equation. In choosing the variables for the selection equation, we adopted an exhaustive approach, including all possible

variables that were available in SILC that describe personal characteristics commonly included in estimations of this type (see section 3.2). One of these characteristics is occupation. The likelihood of employment of a person with a particular occupation depends upon the demand for and supply of this occupation in the labour market. The term "skill mismatch" describes the situation when the supply in a certain occupation is imbalanced with market demand. As Botrić (2009) shows for Croatia, occupations play a significant role in determining the risk of unemployment. Therefore, excluding occupation variables from the participation equation may lead to its misspecification.

The problem emerges because a portion of *unemployed* and *inactive* persons have never worked and hence lack information on occupation in SILC. For these persons, we cannot use occupation variables in the selection equation because they would be "perfect predictors" of non-employment. However, we have decided not to completely exclude from the analysis persons who have never worked. Therefore, *unemployed* and *inactive* are further divided into the subgroups *experienced* and *inexperienced*. *Experienced* are defined as persons who have previous work experience; these persons have either (a) worked before the beginning of IRY, i.e., for whom *we_yipw*>0, or (b) worked during IRY at least one month. *Inexperienced* are those *unemployed* and *inactive*, for which data on occupation do not exist. Thus, we obtain four subgroups: (a) *experienced unemployed*, (b) *inexperienced unemployed*, (c) *experienced inactive,* and (d) *inexperienced inactive*.

3.6 FINAL SUBSAMPLES

Table 2 presents the derivation of the research sample in terms of the number of sample observations. The total number of observations for *employed*, *unemployed* and *inactive* is 6,206, but the number is reduced to 5,877 after some observations are dropped from the analysis (see below).

TABLE 2
Derivation of the selected sample

Sample / Group	Observations	Final observations
S0	9,297	
S1:	6,727	
S2A:	4,136	
(a) *employed*	3,657	3,444
(b) *self-employed*	479	
S2B:	2,549	
(c) *unemployed:*	1,572	1,506
(c1) actively seeking a job on DIN	1,192	
(c2) not actively seeking a job on DIN, but have worked at least one month in IRY	263	
(c3) "employed" or "self-employed" on DIN	117	
(d) *inactive*	977	927
employed, *unemployed* and *inactive*	6,206	5,877

Thus, 329 observations are excluded; table 3 shows the summary. *Employed*, *experienced unemployed* and *experienced inactive* persons numbering 157 are excluded because they have no data on occupation. Next, we drop 38 observations without data on gross wage. Furthermore, 3 persons, whose gross hourly wage is below 5 HRK, are dropped from the sample.[15] Finally, we exclude 131 women who have newborn children and could potentially spend up to 12 months in maternity and parental leave during IRY.[16]

TABLE 3
Dropped observations

Type of person	Number of dropped observations					
	employed	***exp. unempl.***	***inexp. unempl.***	***exp. inactive***	***inexp. inactive***	***Total***
Without data on occupation	111	33	0	13	0	157
Without data on wage	38	0	0	0	0	38
With gross hourly wage below 5 HRK	3	0	0	0	0	3
Potential users of maternity and parental leave	61	22	11	21	16	131
Total	213	55	11	34	16	329

Table 4 presents the structure of the sample according to groups and subgroups. One-quarter of *unemployed* are *inexperienced unemployed*. Among *inactive*, the share of *inexperienced inactive* men is approximately one-third; the share of *inexperienced inactive* women is almost one-half.

TABLE 4
Subgroups of non-employed

	All		Women		Men	
	In thous.	**Share (%)**	**In thous.**	**Share (%)**	**In thous.**	**share (%)**
employed	1,071		485		586	
unemployed	427	100.0	196	100.0	231	100.0
experienced	324	75.9	148	75.6	176	76.2
inexperienced	103	24.1	48	24.4	55	23.8
inactive	225	100.0	169	100.0	56	100.0
experienced	122	54.2	87	51.5	35	62.4
inexperienced	103	45.8	82	48.5	21	37.6

It was indicated previously that the sample formation process enables some persons who have worked in IRY to enter the group of non-employed. It is therefore interesting to see the number of such people and the duration of their work. By definition, *inactive* persons are those who did not work at all in IRY. Additionally,

[15] We believe that these wages are misreported.
[16] SILC does not report data on months spent in maternity or parental leave. It only records the months spent in employment (which equals 12 for most *employed* mothers in the mentioned group). Therefore, we cannot calculate the months in which a person has earned a wage.

by definition, *inexperienced unemployed* are persons who have never worked. Therefore, table 5 shows the structure according to months spent in work only for *experienced unemployed*. Sixty-two per cent have been unemployed for the entire year; a further 29% have worked up to six months, and 9% worked 7 or 8 months.

Table 5

Months spent in work during 2011 for experienced unemployed

	All		Women		Men	
	In thous.	**Share (%)**	**In thous.**	**Share (%)**	**In thous.**	**Share (%)**
Unemployed experienced	324	100.0	148	100.0	176	100.0
0	201	61.8	91	61.5	109	62.1
1-3	43	13.3	20	13.5	23	13.0
4-6	51	15.6	24	16.0	27	15.3
7-8	30	9.3	13	9.0	17	9.6

3.7 AVERAGE CHARACTERISTICS OF EMPLOYED AND NON-EMPLOYED

Table A2 and table A3 (appendix 2) present the means and standard deviations of selected variables obtained for *employed* and four subgroups of non-employed, for women and men, respectively. Figure 1 provides an insight into differences among groups for several key characteristics: age, education, marital status, children, health, and area of living. In all of the graphs, subgroups are intentionally sorted in the following order: *employed*, *inexperienced unemployed*, *experienced unemployed*, *experienced inactive* and *inexperienced inactive*. A certain pattern can be observed for many variables, in which the mentioned groups are lying in an "employability spectrum"; adjacent groups on the graphs have similar personal characteristics.

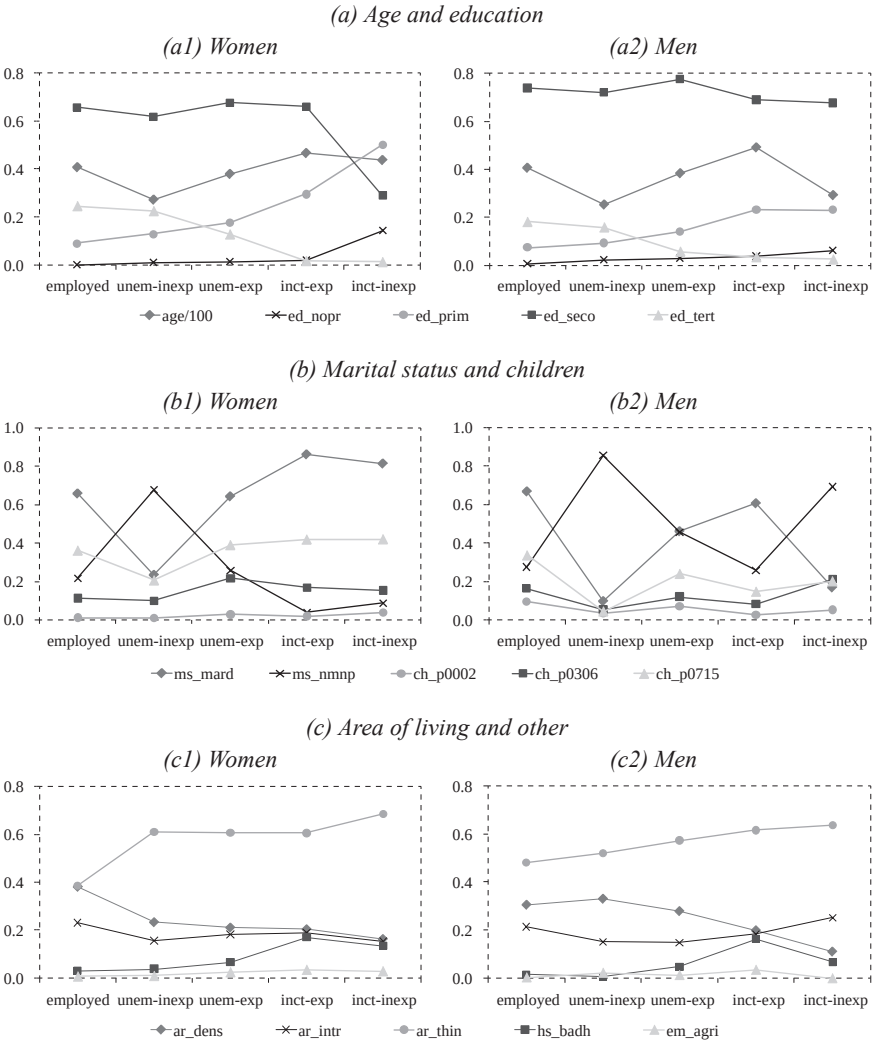
Age and education. As we move from left to right (figure 1a), the share of persons with primary education is increasing (*ed_prim*), whereas it is decreasing for tertiary education (*ed_tert*). The majority of persons have secondary education (*ed_seco*) and the share for women (men) is above 60% (70%). The exceptions are *inexperienced inactive* women, whose share in secondary educated is only 30%; for the same group, the share of tertiary educated is close to zero, and almost 70% of its members have a primary education or less. The youngest groups are *inexperienced unemployed* women and men (*age/100*). *Experienced unemployed* are of a similar average age as *employed*.

Marital status and children. Over 85% of *inactive* women are married (*ms_mard*), compared with 68% of *employed* and *experienced unemployed* women (figure 1b). *Inactive* and *experienced unemployed* women have somewhat more children aged 3 to 6 years than have *employed* and *inexperienced unemployed* women. *Inexperienced inactive* men and *inexperienced unemployed* women and men, are very similar in several respects; they are young people, mostly single, and still living in households with their parents.

Area of living. For women, there is a significant difference between *employed* and all other groups in terms of living area (figure 1c). Approximately 40% of *employed* women live in thinly populated areas (*ar_thin*) compared with over 60% of non-employed. A similar trend, but less pronounced, is observed for men. *Inactive* persons have a significantly higher average share of those with health problems than have *employed* and *unemployed* (*hs_badh*). Non-employed live more often in agricultural households than do *employed* (*em_agri*). For example, the share for *experienced inactive* men is 4%, compared with 0.3% for *employed* men.

Figure 1

Means of selected variables for different groups



Abbreviations: employed – employed; unem-inexp – inexperienced unemployed; unem-exp – experienced unemployed; inct-exp – experienced inactive; inct-inexp – inexperienced inactive.

3.8 "YEARS IN WORK" AND "YEARS OUT OF WORK"

Figure 2 shows scatter plots for the variables "years in work" (*we_yipw*) against age (*ag_year*) for the subgroups of *employed*, *experienced unemployed* and *experienced inactive* women and men. Each plot shows a quadratic polynomial fit of the data and the corresponding R2. A strong relationship between *we_yipw* and *ag_year* exists for *employed*, for which R2 is 0.75 for women and 0.88 for men. The correlation is also high for *experienced unemployed* but is lower than for *employed*; R2 for women and men are 0.56 and 0.66, respectively. *Experienced inactive* men are to some extent similar to *experienced unemployed*, with R2 of 0.46, but for *experienced inactive* women, the relationship is quite weak, with R2 of only 0.22.

FIGURE 2

"Years in work"

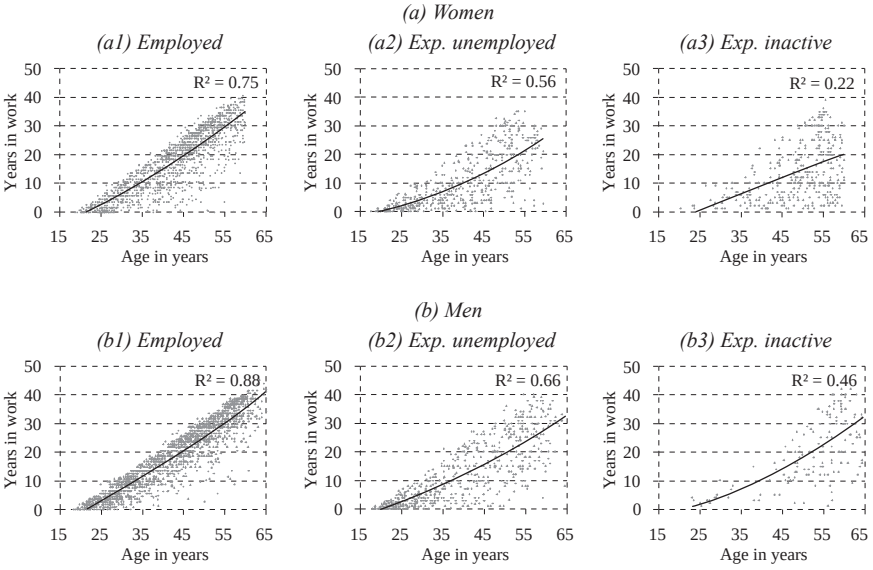
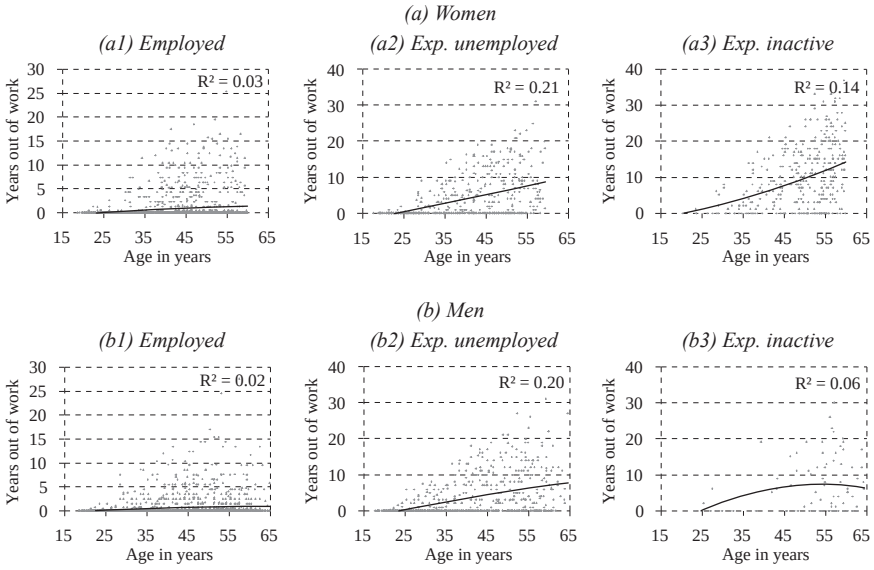


Figure 3 shows the same as figure 2, but for the variable "years out of work" (*we_yopw*). Recall that *we_yipw* and *we_yopw* stand in an inverse relationship. Accordingly, the correlations between *we_yopw* and *ag_year* show an opposite picture. For *employed*, R2 is close to zero, is approximately 0.2 for *experienced unemployed*, and is below 0.15 for *experienced inactive*.

The analysis based on figure 2 and figure 3 suggests that previous work experience is a very good predictor of current activity status. In other words, continuity of employment through the years – since the first job was taken – significantly increases the chances to be currently employed. Conversely, those who have worked little in the past, show much higher tendency to be non-employed in IRY.

FIGURE 3
"Years out of work"



Another point that can be made from this analysis is concerned with regression specifications. The presence of highly correlated regressors results in multicollinearity, which may cause inaccurate estimates of the coefficients and model instability. In our case, multicollinearity will emerge if we insert *we_yipw* and *ag_year* into the same equation because of their high correlation for *employed*, and to a somewhat lesser extent for *experienced unemployed*. The simplest cure for this problem would be to exclude one of the variables, either *we_yipw* or *ag_year*, from the models. However, both age and work experience appear to be important elements in explaining employment participation and wages. One means of keeping work experience in the models is to substitute the variable *we_yopw* for the variable *we_yipw*.

4 PARTICIPATION IN EMPLOYMENT AND NON-EMPLOYMENT

4.1 STRUCTURE OF EMPLOYED AND NON-EMPLOYED BY AGE, EDUCATION AND OCCUPATION

Figure 4, figure 5 and figure 6 show how age, education and occupation influence selection into employment and non-employment. Their review serves as an introduction to more formal analysis using probit models, presented in section 4.2.

The number of *employed* follows an inverted U-pattern (figure 4, graphs a1 and b1). The number is almost negligible under the age of 20. The number of employed women increases with age, and reaches the maximum for the age group 45 to 50 years. For men, the number of *employed* is relatively stable in the interval 25 to 50 years, but significantly falls above the age of 55; above the age of 60, there are few employed men. The number of *experienced unemployed* is relatively con-

stant in the interval 25-55 for both women and men; thereafter, it falls steeply. The numbers and shares of *experienced inactive* are significantly higher for women. The numbers of *inexperienced inactive* men are almost negligible. *Inexperienced unemployed* are primarily young people below the age of 30.

As seen previously, the prevalent education level is secondary (figure 5). Employment shares significantly increase with the level of education: 78% (84%) of women (men) with tertiary education are employed; conversely, among women (men) with primary education or less, only 28% (47%) are employed. Sixty-nine percent of *inexperienced inactive* women have primary or less education, whereas the same share for *employed* is only 9%.

Among women, the most frequent occupation is "service and sales workers" (figure 6, graph a1), whereas for men that category is "craft and related trades workers" (figure 6, graph b1). "Professionals and managers" have an employment share of over 90% for both women and men (figure 6, graphs a2 and b2). Women also have high employment shares in occupations "technicians and associate professionals" and "clerical support workers", but in all other occupations, their employment share is below 60% (figure 6, graph a2). The employment share is below 50% for men in "elementary occupations".

Figure 4

Structure of employed and non-employed by age

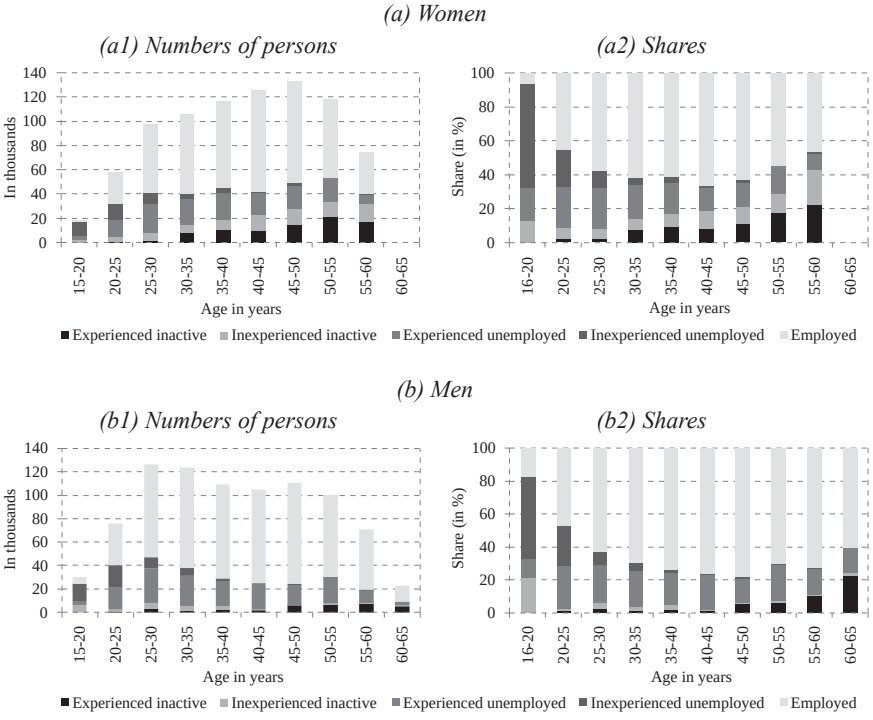


FIGURE 5

Structure of employed and non-employed by education

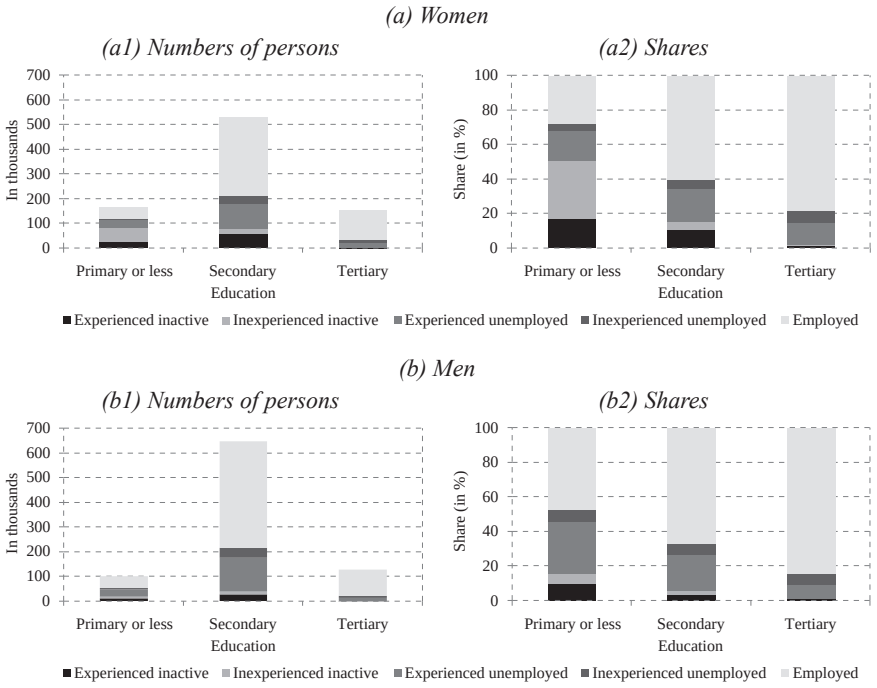
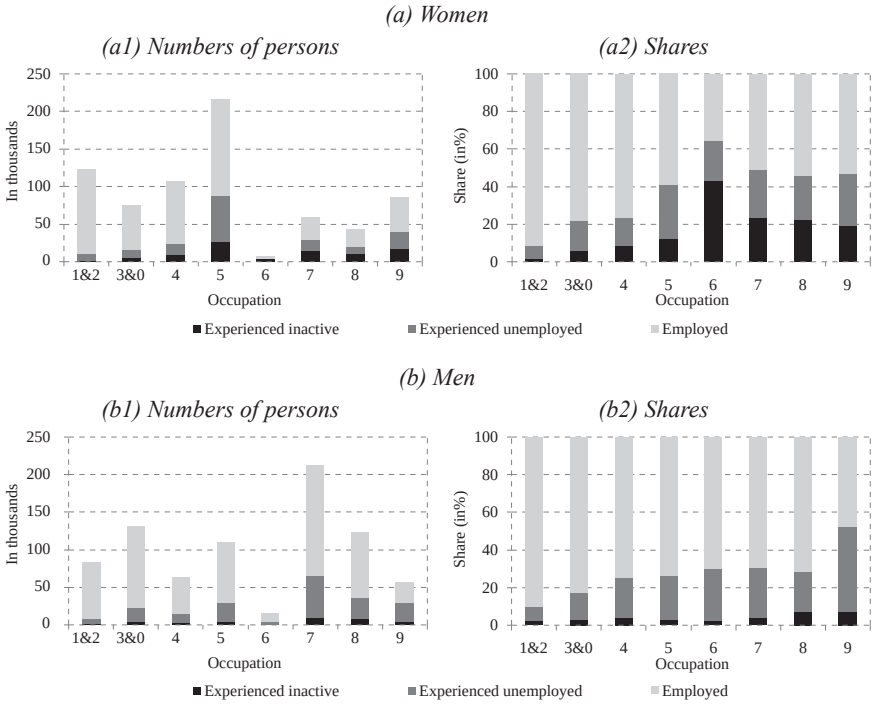


FIGURE 6

Structure of employed and non-employed by occupation



4.2 PROBIT MODEL ANALYSIS

In sections 3.7 and 4.1, descriptive statistics have indicated similarities and differences between *employed* and four groups of non-employed. In this section, the probit regression analysis is used to explore further the differences between various subgroups. Each of five subgroups is compared with one another, yielding 10 specifications each for women and men, which are shown in table 6.

The first four specifications (P1* to P4*) compare *employed* with the subgroups of non-employed. These specifications are relevant for further use as selection equations in HSM. The remaining six specifications (P5* to P10*) relate to the subgroups of non-employed between themselves. If these subgroups are different, they deserve separate analysis; otherwise, some of them could have been pooled together. In specifications P1*, P2* and P5*, which capture *employed*, *experienced unemployed* and *experienced inactive*, we use the "full" set of variables containing the variables on occupations because they are available for these groups of persons. In the remaining specifications, the "reduced" set of variables is used; they omit occupation variables and also "years out of work"; in particular, *inexperienced* have all zero values for *we_yopw*. The detailed results of probit regressions are presented in tables A4, A5, A6 and A7 (appendix 2).

TABLE 6
Probit specifications

Specification	"Positive" subgroup	"Negative" subgroup
P1*	*employed*	*experienced unemployed*
P2*	*employed*	*experienced inactive*
P3*	*employed*	*inexperienced unemployed*
P4*	*employed*	*inexperienced inactive*
P5*	*experienced unemployed*	*experienced inactive*
P6*	*experienced unemployed*	*inexperienced unemployed*
P7*	*experienced unemployed*	*inexperienced inactive*
P8*	*experienced inactive*	*inexperienced unemployed*
P9*	*experienced inactive*	*inexperienced inactive*
P10*	*inexperienced unemployed*	*inexperienced inactive*

Table 7 presents summary results for probit specifications involving *employed* persons, P1* to P4*. The detailed results of probit regressions are presented in table A4 and table A6 (appendix 2). Two standard measures of fit for probit models are presented ("Adjusted McFadden's pseudo R2" and "Adjusted count pseudo R2"), together with four additional indicators, also discussed in appendix 1.

Except for P2W and P3M, all models have relatively low values of ACPR2, particularly P1W (0.14) and P4M (0.09). The indicator $s_{0.5}^{NP}/n = 0.65$ for P1W implies that the probit model classifies 65% of *experienced unemployed* women as employed, whereas only 35% of these persons are correctly classified as non-employed. Conversely, $s_{0.5}^{PN}/p = 0.06$ indicates that only 6% of *employed* persons are

wrongly classified as non-employed. Recall that $s_{0.5}^{NP}/n$ = uses $\pi = 0.5$ as a cut-off probability point to classify a person as positive or negative.

Table 7

Measures of fit for probit models P1 to P4**

Spec.	AMFR2	ACPR2	$s_{0.5}^{PN}/p$	$s_{0.5}^{NP}/n$	s_{p}^{PN}/p	s_{p}^{NP}/n	p
P1W	0.20	0.14	0.06	0.65	0.25	0.26	0.77
P2W	0.39	0.33	0.03	0.49	0.16	0.17	0.85
P3W	0.31	0.22	0.02	0.63	0.20	0.15	0.91
P4W	0.33	0.26	0.04	0.53	0.18	0.21	0.86
P1M	0.17	0.18	0.04	0.68	0.28	0.30	0.77
P2M	0.31	0.16	0.01	0.71	0.16	0.23	0.94
P3M	0.35	0.25	0.01	0.63	0.20	0.14	0.91
P4M	0.27	0.09	0.00	0.87	0.20	0.17	0.97

Conversely, if $\pi = p$ is used as the cut-off probability point, the picture significantly changes (p represents the average probability of being employed in the overall sample). The indicator s_{p}^{NP}/n shows that 26% of *experienced unemployed* women are classified as employed; therefore, 74% of these women are correctly classified as non-employed. Additionally, indicator s_{p}^{PN}/p implies that 25% of *employed* women in P1W are classified as non-employed.

Thus, many groups overlap; some persons who have less-favourable personal characteristics are employed, and *vice versa*. This overlap manifests via the presence of unobservable characteristics, represented by the random term u_i (section 2). Relatively low values of AMFR2 and ACPR2 indicate that u_i plays an important role; in other words, we lack variables in the probit model that would better explain a participation mechanism.

Age variables (*ag_year* and *ag_ysqr*) are highly significant in all specifications, with positive and negative coefficients for *ag_year* and *ag_ysqr*, respectively. In P1* and P2*, which include the "years out of work" variables (*we_yopw*, *we_yosq*), these variables are highly significant and suggest a hyperbolic relationship; the likelihood of being currently non-employed increases with the length of period previously spent in non-employment.

Women and men living in thinly populated areas (*ar_thin*) and men living in "agricultural households" (*em_agri*) have a lower probability of being *employed*. Most "other income" types are not significant, except for family benefits (*oi_f*); the coefficient of family benefits is highly significant and negative for both women and men. In specification P1M, private transfers (*oi_c*) are negative and significant; one explanation is that *employed* men are net payers of transfers, simply because they have greater resources than non-employed. Health situation (*hs_badh*) is a very important factor in the selection process, as could be expected from section 3.7 (figure 1, graphs c1 and c2); all probit models indicate a significantly lower probability of employment for persons with health problems.

Concerning education and occupation variables in P1* and P2*, we have somewhat unforeseen results. For example, in P1W, the coefficient for tertiary education (*ed_tert*) is significant and negative, which is contrary to expectations (sections 3.7 and 4.1). The cause could be found in a high correlation with the "professionals and managers" variable (*oc_21*), whose coefficient is large, positive and significant in the same model. The majority of "professionals and managers" have tertiary education; because of multicollinearity, the model cannot properly estimate the effects of both variables. A similar but opposite situation can be seen in P2M, in which the coefficient for tertiary education is high and positive, but "professionals and managers" have a negative coefficient.

Proceeding with the analysis of specifications P5* to P10*, we again turn to measures of fit (the detailed results of probit regressions are presented in tables A4, A5, A6 and A7 (appendix 2)). The highest values of indicators AMFR2 and ACPR2 for both men and women are achieved for P8* specifications, which analyse *experienced inactive* vs. *inexperienced unemployed* subgroups. As seen in section 3.7, these two groups significantly differ in terms of age, marital status and education; these differences are confirmed by probit models.

The differences between *experienced unemployed* and *experienced inactive* are analysed with specification P5*. As probit models indicate, confirming the findings in section 3.7, the former group has better education and health. Additionally, widowed women and women with children are more likely to be inactive, rather than unemployed. For men, we find very low values of AMFR2 (0.11) and ACPR2 (0.05). According to indicator $s_{0.5}^{NP}/n$, 78% of negative are classified as positive, meaning that there is significant overlap between *experienced unemployed* and *experienced inactive* men.

Table 8

Measures of fit for probit models P5 to P10**

Spec.	AMFR2	ACPR2	$s_{0.5}^{PN}/p$	$s_{0.5}^{NP}/n$	s_{p}^{PN}/p	s_{p}^{NP}/n	p
P5W	0.17	0.29	0.16	0.43	0.32	0.23	0.63
P6W	0.18	0.27	0.07	0.50	0.24	0.27	0.76
P7W	0.18	0.32	0.16	0.40	0.21	0.30	0.64
P8W	0.55	0.73	0.04	0.20	0.09	0.14	0.65
P9W	0.10	0.38	0.31	0.30	0.32	0.28	0.51
P10W	0.41	0.62	0.21	0.10	0.15	0.14	0.37
P5M	0.11	0.05	0.03	0.78	0.25	0.24	0.83
P6M	0.25	0.29	0.09	0.42	0.25	0.19	0.76
P7M	0.18	0.10	0.01	0.81	0.26	0.12	0.89
P8M	0.55	0.74	0.14	0.08	0.13	0.10	0.39
P9M	0.36	0.68	0.11	0.14	0.17	0.13	0.62
P10M	0.12	0.18	0.06	0.65	0.27	0.19	0.72

Inexperienced inactive women represent a large group among women; therefore, it is interesting to analyse the differences between this and other groups of

non-employed, which is enabled by specifications P7W, P9W and P10W. Again, in line with the presentation in section 3.7, education is the most important factor (*ed_prnp*, *ed_tert*). Additionally, the likelihood of being *inexperienced inactive* is greater for women having a partner (*ms_mard*, *ms_nmhp*).

In summary, employed subgroups are significantly different from non-employed subgroups (specifications P1* to P4*). Several characteristics are identified that determine the likelihood of being employed vs. non-employed: age, education, occupation, years out of work, health status and family benefits. Among the non-employed themselves, the defined subgroups are significantly different, justifying their separate treatment (specifications P5* to P10*). Possible exceptions are *experienced inactive* and *experienced unemployed* men, who have relatively similar characteristics. Although all probit models are statistically significant, there is room for improvement, which would arrive from inclusion of additional variables such as region of living (which is not included in SILC).

5 ESTIMATION OF GROSS WAGES

5.1 GROSS WAGES IN SILC

Gross wage in SILC is obtained by grossing up net wage reported by surveyed persons. Imputation of personal income tax (PIT) and social insurance contributions (SIC), needed for net-to-gross conversion, is performed by CBS and considers all of the relevant factors that determine the amount of PIT (e.g., number of children, other dependents, and place of living).[17] Table 9 summarises the main indicators for gross monthly wage for different subgroups of *employed* persons.

Average monthly wage for *employed* persons equals 6,558 HRK, which is 16.5% below the official average monthly wage in 2011 (7,796 HRK), calculated for workers employed in legal entities (CBS, 2015). There could be more than one explanation for the relatively large discrepancy between the SILC and official data. First, unlike SILC data, the CBS indicator does not capture persons employed by self-employed persons (craftsmen, professionals, or small entrepreneurs in agriculture), who are likely to have lower average wages than will employees in legal entities. Second, and most importantly, SILC tends to underrepresent employees at the higher end of income distribution.[18] Third, SILC data will capture some "grey economy" wage payments, but the effect on the wage distribution is uncertain.

Table 9 indicates that gross wage increases with age and education. The average wage is higher in densely populated areas and for certain occupations such as "professionals and managers", "technicians and associate professionals" and

[17] Gross wage can be represented as the sum of net wage, PIT, surtax and employee's SIC. Surtax is a municipality tax determined as a percentage of PIT (this percentage varies by municipality from 0% to 18%). Employees' SIC are equal to 20% of gross wage.

[18] The analysis presented in the EUROMOD Country Report for Croatia (Urban and Bezeredi, 2015) compares wage distributions from SILC and Tax Administration. According to the Tax Administration data, 1.7% of employees have a gross wage above 300% of the official average gross wage, and their share in total gross wages is 9.7%. In SILC, the share of such employees is 0.6% and they obtain only 2.8% of total wage income.

"clerical support workers". The "raw gender gap", i.e., the difference between men and women's average gross wage, is approximately 14%.

Table 9
Mean monthly wage by groups

	All			Women			Men		
	Share (%)	**Mean monthly wage**	**St. err.**	**Share (%)**	**Mean monthly wage**	**St. err.**	**Share (%)**	**Mean monthly wage**	**St. err.**
Overall	100.0	6,558	76	100.0	6,080	112	100.0	6,955	100
Age in years									
16-25	18.9	5,608	133	17.1	5,339	200	20.5	5,794	175
25-40	28.2	6,445	155	28.2	5,907	216	28.3	6,891	212
40-55	31.0	6,569	136	34.3	6,006	193	28.2	7,138	189
55-64	21.9	7,504	165	20.4	7,058	273	23.1	7,832	202
Education									
Primary or less	8.7	4,604	126	9.4	3,885	118	8.1	5,307	194
Secondary	70.1	5,839	70	65.9	5,221	88	73.7	6,297	98
Tertiary	21.2	9,731	211	24.7	9,204	300	18.3	10,323	292
Area of living									
Densely	33.9	7,676	159	38.2	6,954	224	30.4	8,430	219
Intermediate	22.2	6,215	177	23.1	5,684	240	21.4	6,688	249
Thin	43.8	5,863	72	38.7	5,451	108	48.1	6,138	95
Occupation									
2&1	17.5	9,711	225	23.2	9,096	285	12.8	10,636	357
3	15.7	7,786	212	12.2	7,180	379	18.5	8,115	248
4	12.1	6,312	132	17.0	6,194	173	8.1	6,522	197
5	19.6	5,009	102	26.4	4,484	119	13.9	5,837	162
6	1.3	4,629	214	0.5	4,070	441	1.9	4,763	235
7	16.6	5,374	104	6.3	3,694	124	25.1	5,727	113
8	10.4	6,128	244	4.8	4,315	204	15.1	6,607	296
9	6.8	4,370	137	9.5	4,097	187	4.6	4,841	178

5.2 LRM REGRESSION ANALYSIS OF GROSS WAGES

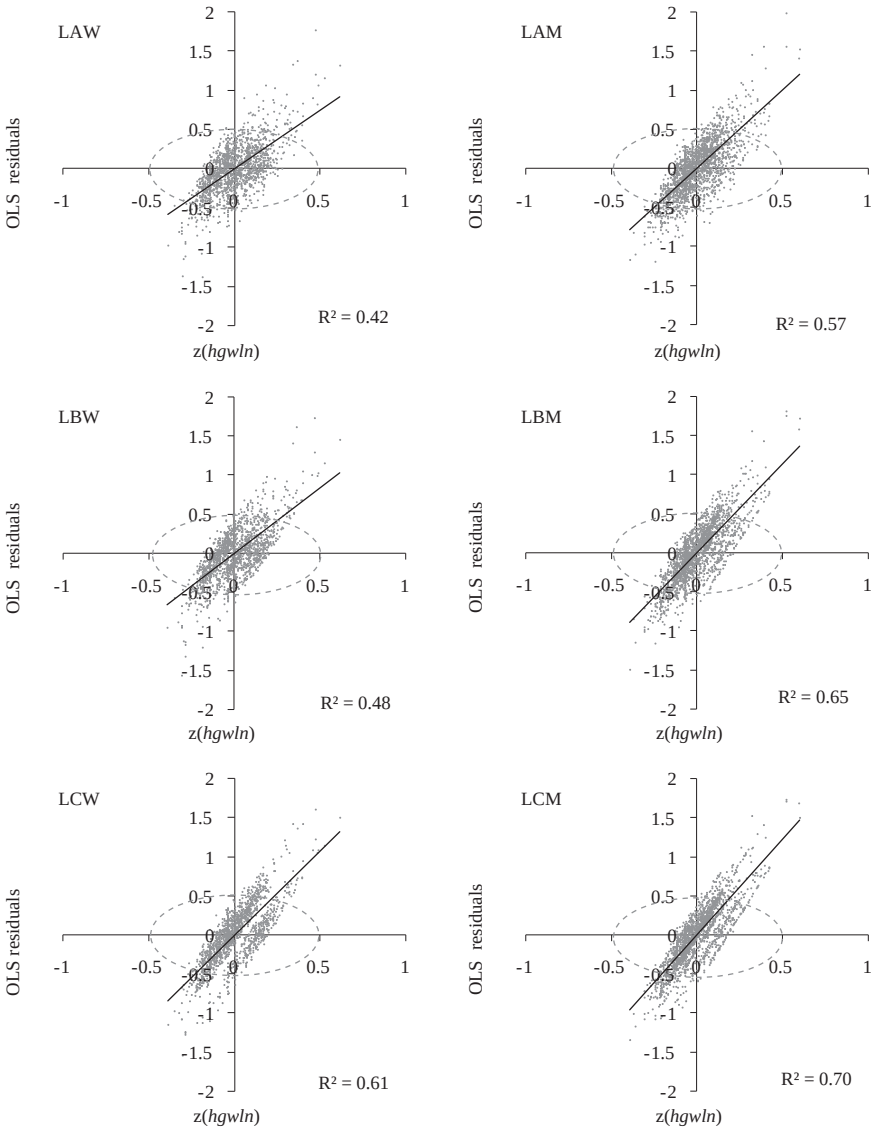
Section 3.2 presents the variables constructed using the SILC dataset. For some of these variables, data are available only for persons in employment. These variables address job characteristics (*em_locs*, *em_locl*, *em_perj*, *em_mana*) and industry of employment (e.g., *in_a*). An additional subset of variables has data for *employed* and *experienced* persons, but not for *inexperienced*; these subsets include "years out of work" variables (*we_yopw*, *we_yosq*) and occupation variables (e.g., *oc_21*).

If we were not interested in predicting wages of non-employed persons and if we had not considered the use of HSM, the natural step would be to use LRM and include all available variables into consideration. This is what specifications LA* do. However, in the current study, we must use those variables for which data are available both for *employed* and non-employed persons. Therefore, LBW and LBM specifications use all of the variables that have data for *employed*, *experi-*

enced unemployed and *experienced inactive*. Furthermore, specifications LC* contain only the variables that can be used for *inexperienced unemployed* and *inexperienced inactive*; occupation "years out of work" variables are excluded.

FIGURE 7

Residuals from LRM regressions



Abbreviations: LAW, LBW and LCW – LRM specifications for women; LAM, LBM and LCM – LRM specifications for men; z(hgwln) – standardised value of hgwln.

The results are presented in table A8 (appendix 2). Wage increases with age (*ag_year*); the quadratic term (*ag_ysqr*) is negative but not significant except in LCM. All specifications indicate a positive and statistically significant effect of living in urban areas (*ar_dens*). Married men have higher wages than men do in other marital statuses (*ms_mard*); additionally, men with small children (*ch_p0002*) have higher wages than others do. Persons with tertiary education (*ed_tert*) have significantly higher wages, which is confirmed by all models. The coefficients are smaller in LA* and LB* than in LC*; unlike LC*, the other two specifications contain occupation variables that take over the part of the positive influence of education. The variables describing job characteristics in LA* (*em_locs*, *em_locl*, *em_perj*, *em_mana*) are all highly significant.

From R2 statistics, we can see that the LA* specification has the greatest predictive power because it includes all of the relevant variables (R2 equals 0.55 for women and 0.44 for men). Conversely, LC* performs worst (R2 equals 0.38 for women and 0.31 for men) due to the lack of many important variables. The LB* specification is somewhere in between.

Figure 7 shows residuals from six LRM regressions presented in table A8, plotted against the standardised values of *hgwln*, denoted as z(*hgwln*). We note the same pattern in all six graphs: residuals, on average, increase with z(*hgwln*). The correlation between residuals and z(*hgwln*) is quite strong, as confirmed by R2s; it is lowest for LA*, and highest for LC* specifications. For smaller actual wages, the residuals tend to be negative, whereas they tend to be positive for higher actual wages. Because residuals are differences between actual and predicted wages, models tend to over-predict (under-predict) lower (higher) wages.

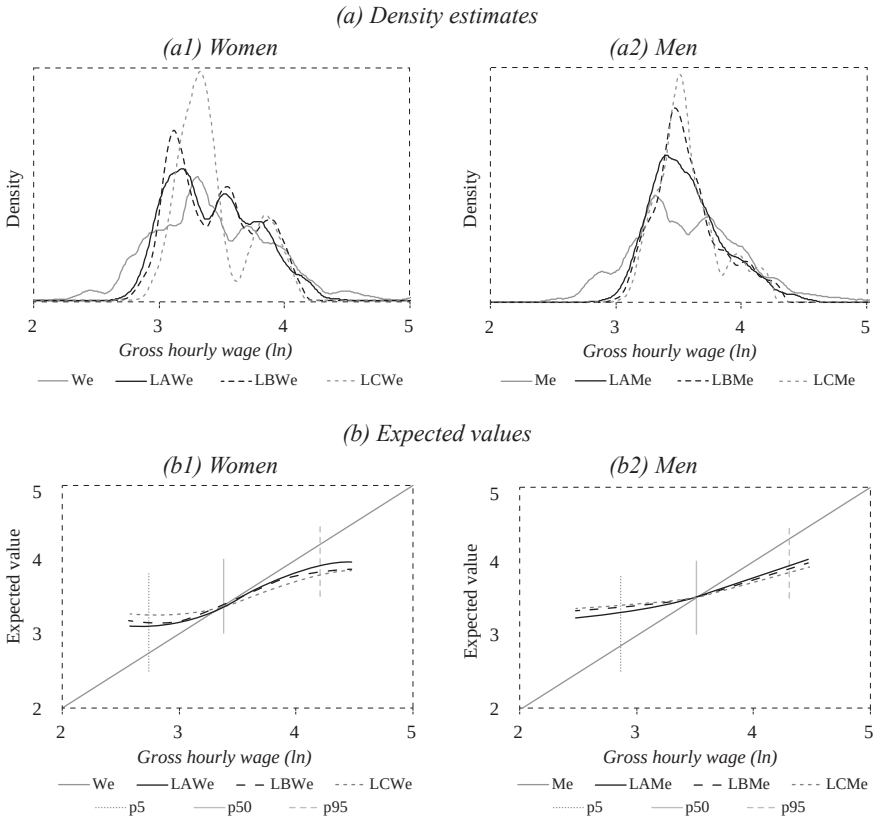
Let us further examine how successful LRM models are in prediction of gross wages. Graphs a1 and a2 of figure 8 show kernel density estimates of the distributions of actual sample wages of *employed* persons and predictions obtained by LA*, LB* and LC* models. The main conclusion is that all three LRM models fail to predict properly the correct number of persons at the right and the left tail of the wage distribution. This result is expected from the previous analysis of residuals. LA* fares slightly better in this respect than do the other two models. Graphs b1 and b2 of figure 8 show conditional expected values of predictions obtained by LA*, LB* and LC*.[19] Here, we can ascertain how much these predictions over- or under-estimate true wages for each level of actual gross wage. At the 5th percentile, the wage is over-predicted by more than 40%, whereas it is under-predicted by 30% at the 95th percentile.

[19] Kernel density estimates are obtained by the Stata program "Kernel density estimation" (command *kdensity*), using the Epanechikov kernel. Conditional expected values are obtained by the program "Local polynomial smoothing" (command *lpoly*), using the Epanechikov kernel and the 5th degree of polynomial.

FINANCIAL THEORY AND PRACTICE
40 (1) 1-61 (2016)

FIGURE 8

Gross wage predictions by LRM models



Abbreviations: We (Me) – actual wages of employed women (men); LAWe (LAMe) – wage predictions for employed women (men) based on LAW (LAM); LBWe (LBMe) – wage predictions for employed women (men) based on LBW (LBM); LCWe (LCMe) – wage predictions for employed women (men) based on LCW (LCM); p5, p50 and p95 – the 5th, 50th and 95th percentiles of actual wage distribution of employed, respectively.

5.3 QUANTILE REGRESSION ANALYSIS OF GROSS WAGES

The linear regression model, used in LA*, LB* and LC* specifications, provides us with a single coefficient for each variable in the wage equation; the set of these coefficients is denoted by $\tilde{\alpha}$ (recall section 2). For example, if the *k*th variable is tertiary education, $\tilde{\alpha}_k$ measures the effect on *hgwln* of having tertiary education compared with the benchmark education level (in our case, secondary education). This approach assumes that the effect of each variable is identical across the wage distribution. However, in reality, the influence of a certain variable on the wage may be different for persons with higher and lower incomes (see section 1 for reference to Nestić, 2005).

Therefore, we employed the quantile regressions model (QRM) to estimate the wage equations. Explanation of the method can be found in Cameron and Trivedi

(2005). For estimation, Stata program "Bootstrap quantile regression" was used (command *bsqreg*). One hundred bootstrap replications at each quantile are made to obtain standard errors of the coefficients. The specification contains the same variables as LA*. This part of the analysis does not use personal weights because the Stata program cannot properly calculate standard errors for QRM if sampling weights are used.[20]

Table A9 (appendix 2) presents the results of QRM for selected percentiles. For comparison, LRM results are shown in separate columns and represent LA* specifications, which are rerun without using the sampling weights. Note that the coefficients are slightly different from those obtained for LA* specifications in table A8.

The coefficients obtained by QRM change their magnitudes and statistical significance across different percentiles. Figure A2 (appendix 3) presents the QRM coefficients and their confidence intervals for several variables, estimated at the 5th, 10th, ..., 90th and 95th percentiles. They are compared with the LRM coefficients and their confidence intervals.

The QRM coefficients for tertiary education (*ed_tert*) increase with the percentiles. For women, the coefficients obtained at the 5th, 90th and 95th percentiles lie outside the LRM confidence interval. Living in densely populated areas (*ar_dens*) has different effects for women and men; the QRM coefficients decrease (increase) with percentiles for women (men). The positive effect of having a permanent job (*em_perj*) is much higher for low-wage than for high wage employed men; a similar trend, but less pronounced, can be observed for women. Industry sections O, P and Q fully belong to the public sector. For those employed in these sections, the "wage premium" is significantly higher for persons with lower wages, which conforms to the findings of Nestić (2005).

The results presented above indicate that the QRM-based model might be used to cure the problem of over-prediction (under-prediction) of wages at low (high) parts of wage distribution. Therefore, we perform an *ad hoc* exercise, using the QRM estimates to predict wages as follows.

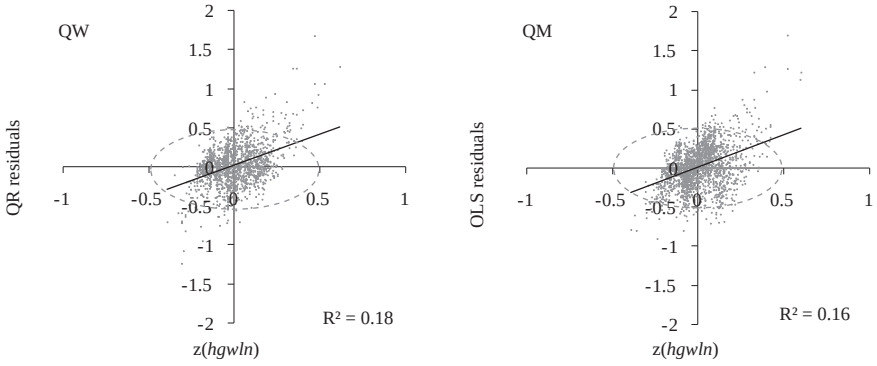
Denote with $\breve{\alpha}^q$ the set of QRM coefficients obtained at the q^{th} percentile. We focus only on the subsample of *employed* persons. The actual gross wage percentile of *employed* person *i* is denoted as q_i. The wage prediction for *employed* person *i* is obtained by application of the coefficients: (a) $\breve{\alpha}^{q=0.2}$, if $q_i \leq 0.2$, (b) $\breve{\alpha}^{q=0.5}$ if $0.2 < q_i < 0.8$, or (c) $\breve{\alpha}^{q=0.8}$, if $q_i > 0.8$.

New predictions and residuals are presented in figure 9. Again, there is a positive relationship between residuals and wages, but the problem appears much less severe than in the case of LRM models; namely, R2 is below 0.2 for both women and men.

[20] The effect of using sampling weights in this case is as though each observation is cloned *n* times, where *n* represents the sampling weight; a huge artificial population is obtained for which the standard errors appear negligibly small.

FIGURE 9

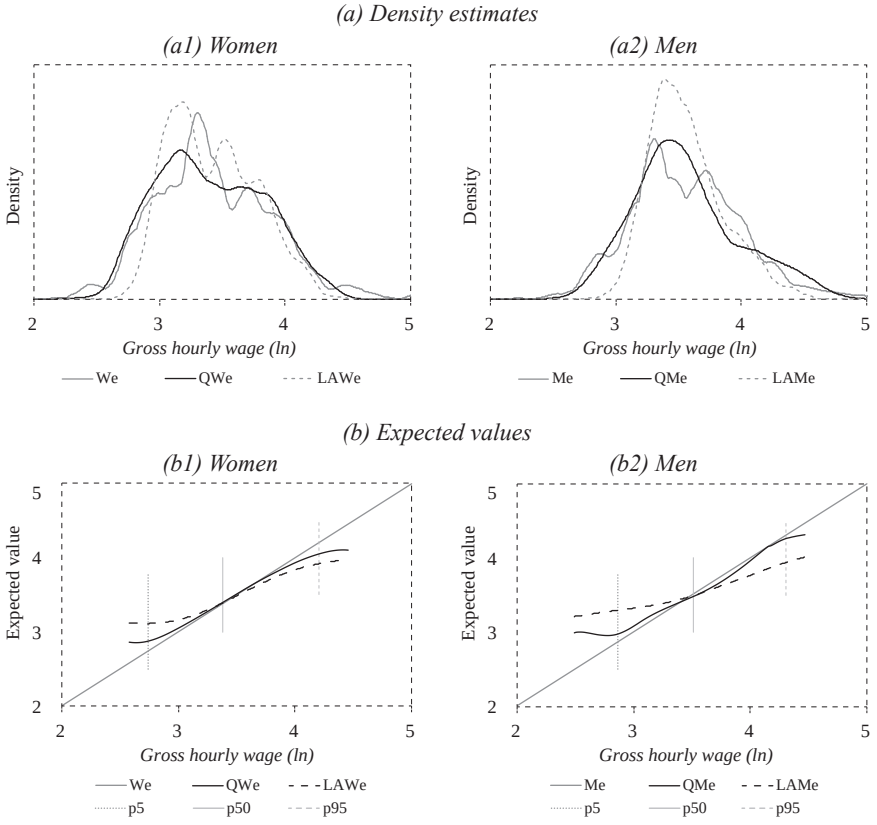
Residuals from QRM regressions



Abbreviations: QW (QM) – quantile regression models for women (men); z(hgwln) – standardised value of hgwln.

FIGURE 10

Gross wage predictions by quantile regression model



Abbreviations: We (Me) – actual wages of employed women (men); QWe (QMe) – wage predictions for employed women (men) based on QW (QM); LAWe (LAMe) – wage predictions for employed women (men) based on LAW (LAM); p5, p50 and p95 – the 5th, 50th and 95th percentiles of actual wage distribution of employed, respectively.

Graphs a1 and a2 of figure 10 show that the new model fits quite well the density of actual sample wages at the tails, significantly better than does LRM. Additional evidence of improvement is seen in graphs b1 and b2 of figure 10, in which for women the expected value of prediction lies very close to the line of equality, particularly at the bottom part of the wage distribution.

5.4 RESULTS OF THE HECKMAN SELECTION MODEL

Table 10 shows four specifications for HSM, which include *employed* and all subgroups of *unemployed* and *inactive* persons. Specifications H1* and H2* use the same variables for the wage equation as LB* does (section 5.2); the participation equations are equal to P1* and P2*, respectively (section 4.2). Conversely, specifications H3* and H4* use LC* specification variables for the wage equation; in participation equations, the variables from P3* and P4* are used, respectively. Following Verbeek's (2004) advice (section 2), all of the variables in wage equations are present in participation equations. Conversely, participation equations contain variables that are not included in wage equations: "other income" variables (*oi_a* to *oi_g*) and "agricultural household" variable (*em_agri*).[21]

TABLE 10

Heckman selection model specifications

Spec.	"Positive"	"Negative"	Participation equation as in:	Wage equation as in:
H1*	*employed*	*experienced unemployed*	P1*	LB*
H2*	*employed*	*experienced inactive*	P2*	LB*
H3*	*employed*	*inexperienced unemployed*	P3*	LC*
H4*	*employed*	*inexperienced inactive*	P4*	LC*

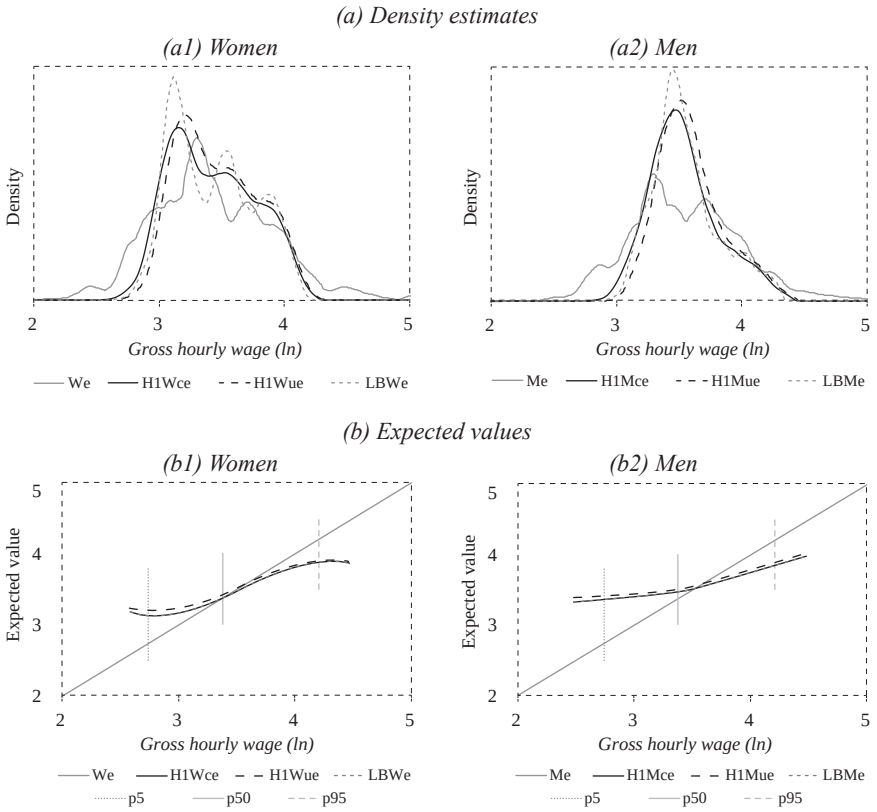
The results are presented in table A10 and table A11 (appendix 2). These tables consist of three parts. The first two parts contain coefficients and significance levels for the wage and participation equations. The third part contains various model indicators. *Sigma*, *rho* and *lambda* represent the estimates of coefficients $\hat{\sigma}_e$, $\hat{\rho}_{eu}$ and $\hat{\Lambda}$, respectively. */lnsigma* and */athrho* represent the estimate of the natural logarithm of σ_e and the estimate of the inverse hyperbolic tangent of ρ_{eu}; from these estimates, $\hat{\sigma}_e$ and $\hat{\rho}_{eu}$ are obtained by inversion. For *rho* and *lambda*, the Stata program does not obtain significance levels but only reports standard errors and confidence intervals. Therefore, for $\hat{\rho}_{eu}$ and $\hat{\Lambda}$, we assume the same significance level as for */athrho*. The presence of statistically significant $\hat{\rho}_{eu}$ manifests that the null hypothesis of no correlation between e_i and u_i cannot be rejected. At the 5% significance level, seven of eight specifications show the existence of such a correlation, i.e., that $\rho_{eu} \neq 0 \Rightarrow \Lambda \neq 0$. In H2M, $\hat{\Lambda}$ is significant only at the level of 0.15. In seven specifications, $\hat{\Lambda}$ is negative; it is positive only for H2W.

[21] Economic arguments tell us that these variables should not be included from wage equations, whereas regression analysis indicates their non-significance in wage equations.

As in section 2, the aim of HSM is to provide unbiased estimates of the wage equation coefficients, $\hat{\alpha}$. If $\rho_{eu} \neq 0$, coefficients $\hat{\alpha}$ will differ from coefficients $\tilde{\alpha}$, obtained by LRM. Therefore, for example, we can compare the wage equation coefficients obtained for H1W and H1M (table A10) with the coefficients $\tilde{\alpha}$, obtained for LBW and LBM (table A8), respectively. All of the coefficients that were significant in LRM are also significant in HSM. Comparing the magnitude of these coefficients, we note that the intercept increases – by 5.1% for women (H1W vs. LBW) and by 2.8% for men (H1M vs. LBM). The non-intercept coefficients are generally lower in HSM models (the exceptions are *ed_tert* for women and *ar_dens* for men).

FIGURE 11

*Gross wage predictions by the Heckman selection model – H1**



Abbreviations: We (Me) – actual wages of employed women (men); LBWe (LBMe) – wage predictions for employed women (men) based on LBW (LBM); H1Wce (H1Mce) – conditional wage predictions for employed women (men) based on H1W (H1M); H1Wue (H1Mue) – unconditional wage predictions for employed women (men) based on H1W (H1M). p5, p50 and p95 – the 5th, 50th and 95th percentiles of actual wage distribution of employed, respectively.

Figure 11 (graphs a1 and a2) shows the density estimates of predictions obtained by H1* models. Two sets of estimates are shown for HSM models: "conditional" and "unconditional" predictions, obtained according to equations (9) and (8), respectively. For comparison, graphs also show densities of actual wages of em-

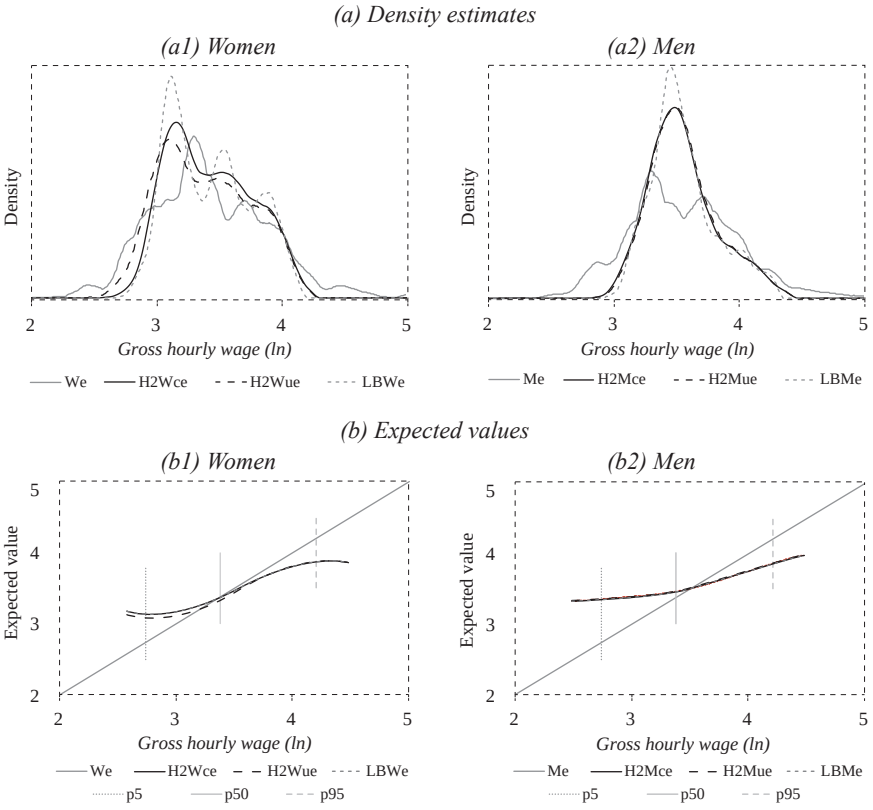
ployed and densities of LB*-based predictions. Conditional predictions are very similar to LB* predictions. Unconditional predictions are "scaled to the right" because $\hat{\Lambda}$ is negative for both women and men.

Consequently, expected values of conditional predictions overlap with those for LB* (figure 11, graphs b1 and b2). Expected values of unconditional predictions lie above those obtained for LB*.

Figure 12 shows the predictions based on H2* specifications. In H2W, $\hat{\Lambda}$ is positive; therefore, the density curve for unconditional predictions is situated to the left of the density curves obtained for conditional predictions and LBW (figure 12, graph a1). Additionally, expected values of unconditional predictions lie below those obtained for conditional predictions and LBW (figure 12, graph b1). In the case of H2M, $\hat{\Lambda}$ is negative but small and not statistically significant. Therefore, the density curves and expected values of conditional and unconditional predictions overlap.

Figure 12

*Gross wage predictions by the Heckman selection model – H2**



Abbreviations: We (Me) – actual wages of employed women (men); LBWe (LBMe) – wage predictions for employed women (men) based on LBW (LBM); H2Wce (H2Mce) – conditional wage predictions for employed women (men) based on H2W (H2M); H2Wue (H2Mue) – unconditional wage predictions for employed women (men) based on H2W (H2M). p5, p50 and p95 – the 5th, 50th and 95th percentiles of actual wage distribution of employed, respectively.

5.5 GROSS WAGE PREDICTIONS FOR NON-EMPLOYED

We now use the coefficient estimates of all models presented above to predict the wages of non-employed subgroups. Figure 13 shows the density estimates of wage predictions obtained by different models. The predictions based on LA*, LB*, LC* and Q* models are obtained for *experienced unemployed* persons. H1* and H2* predictions are obtained for *experienced unemployed* and *experienced inactive* persons, respectively.

In section 3.4, it was mentioned that a certain part of *experienced unemployed* persons have worked during IRY; for such persons, we have data on wages and show their distribution in figure 13.[22] Of course, these wages are not representative of all non-employed persons, but they can provide some illustration.

Making wage predictions for non-employed based on QRM is not fully straightforward. Specifically, we cannot use the same procedure as in section 5.3 because for non-employed persons, the quantiles q_i are unknown. Therefore, we first make "preliminary" wage predictions for non-employed, $\check{w}_i^0$, using the QRM coefficients $\check{\alpha}^{q=0.5}$. Denote with w_q the wage of an *employed* person at the q^{th} percentile. To obtain final predictions, we apply the following sets of QRM coefficients: (a) $\check{\alpha}^{q=0.2}$, if $\check{w}_i^0 \leq w_{q=0.2}$, (b) $\check{\alpha}^{q=0.5}$, if $w_{q=0.2} < \check{w}_i^0 \leq w_{q=0.8}$, or (c) $\check{\alpha}^{q=0.8}$, if $\check{w}_i^0 > w_{q=0.8}$.

Recall that LA* models include variables concerned with the current job characteristics, whose values are available for *employed* persons only. In making wage predictions for non-employed persons, we set the values of all of these variables to zero, which may be a reason why in figure 13 (graphs a1 and a2), LA*-based predictions for *experienced unemployed* show lower measures of central tendency than do those based on LB* and LC*. Q*-based predictions seem to provide better fit than do LRM models (graphs b1 and b2).

Figure 13 (graphs c1 and c2) presents the results for *experienced unemployed* made using H1*. Unconditional and conditional predictions are obtained using equations (8) and (10), respectively. There are large differences between LB*, unconditional and conditional H1*-based predictions, which are the consequence of the negative $\hat{\Lambda}$. The highest measures of central tendency are seen for unconditional H1*-based predictions, followed by conditional H1* and LB*-based predictions.

The case of H2M-based predictions for *experienced inactive* men is similar because of negative $\hat{\Lambda}$ (although not significant at usual levels) (figure 13, graph d2). Conversely, for H2W, the order of the three density curves is reversed. The modal values for unconditional H2W, conditional H2W and LBW-based predictions are 2.72, 2.90 and 3.05.

Figure 13 is illustrative and focuses only on *experienced* subgroups of non-employed. Table 11 and table 12 show the mean predicted values and standard errors

[22] Overall, 340 observations are used, 154 for women and 186 for men.

for all subgroups of non-employed and *employed* for all specifications. Within each subgroup, there are large differences in the means of predicted wages obtained by different models and indicators. For example, for *inexperienced inactive* women, the means range from 3.045 (QW) to 3.720 (LBW). In some cases, non-employed subgroups mean wage predictions are higher than the mean actual wage of actually *employed* persons. This situation specifically occurs for conditional predictions obtained by HSM models for inexperienced subgroups – H3* and H4*.

FIGURE 13

Predictions of gross wages for non-employed

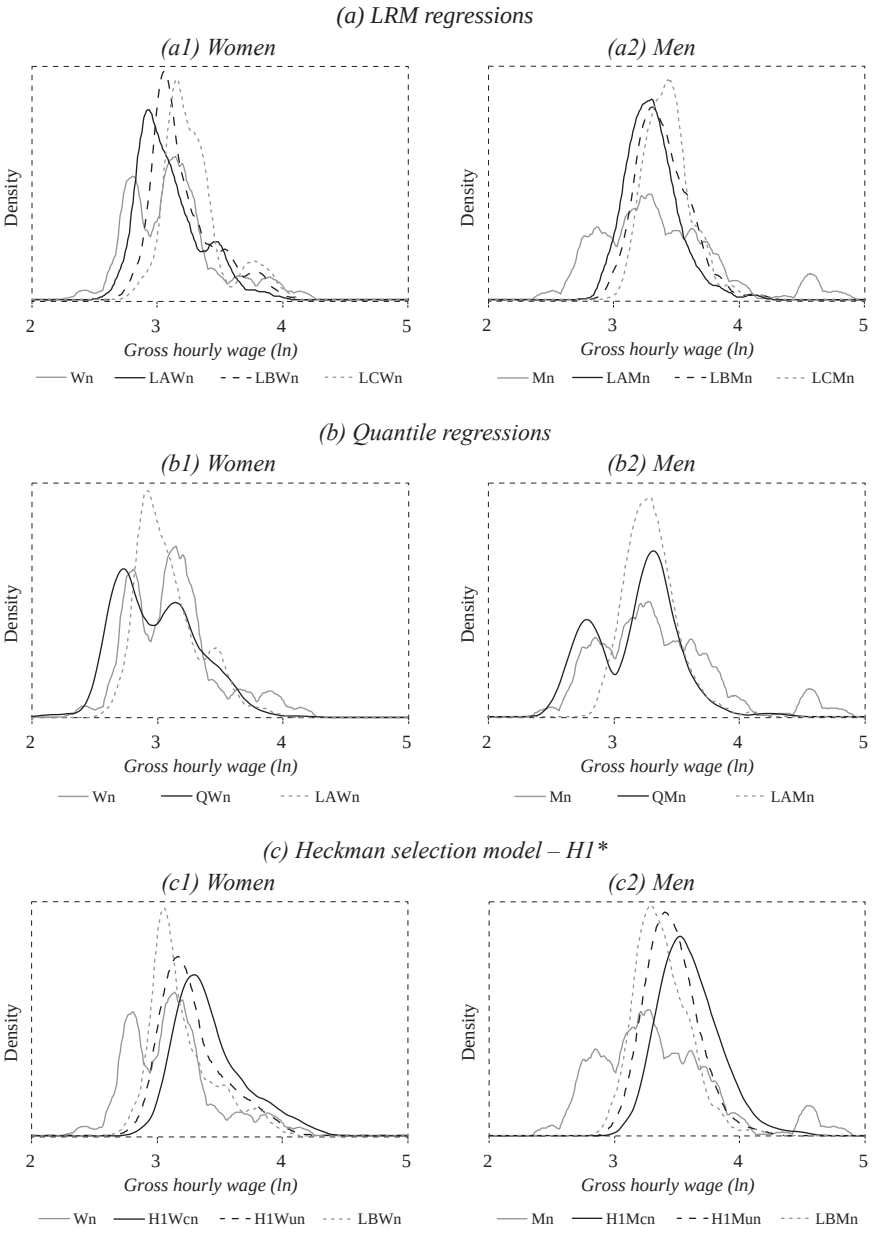
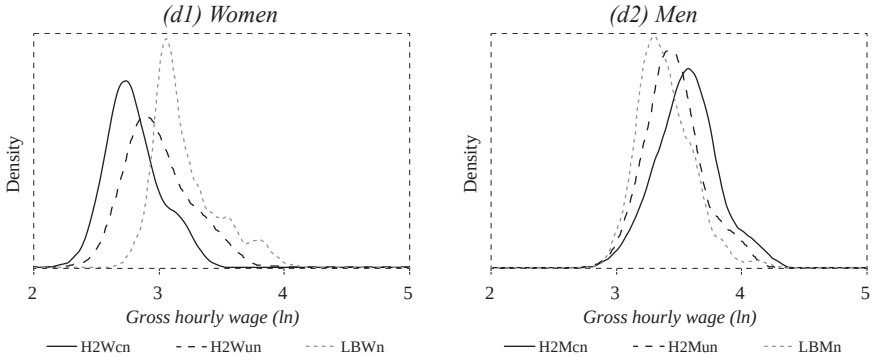


FIGURE 13 (continue)

Predictions of gross wages for non-employed

*(d) Heckman selection model – H2**



Abbreviations: Wn (Mn) – "pseudo-actual" wages of experienced unemployed women (men); LAWn (LAMn) – wage predictions for exp. unemployed women (men) based on LAW (LAM); LBWn (LBMn) – wage predictions for exp. unemployed women (men) based on LBW (LBM); LCWn (LCMn) – wage predictions for exp. unemployed women (men) based on LCW (LCM); QWn (QMn) – wage predictions for exp. unemployed women (men) based on QW (QM); H1Wcn (H1Mcn) – conditional wage predictions for exp. unemployed women (men) based on H1W (H1M); H1Wun (H1Mun) – unconditional wage predictions for exp. unemployed women (men) based on H1W (H1M); H2Wcn (H2Mcn) – conditional wage predictions for exp. inactive women (men) based on H2W (H2M); H2Wun (H2Mun) – unconditional wage predictions for exp. inactive women (men) based on H2W (H2M).

TABLE 11

Mean predicted gross wages for women

	employed		***experienced unemployed***		***experienced inactive***		***inexperienced unemployed***		***inexperienced inactive***	
	Mean	**St. err.**	**Mean**	**St. err.**	**Mean**	**St. err.**	**Mean**	**St. err.**	**Mean**	**St. err.**
LA	3.432	(0.011)	3.101	(0.013)	3.131	(0.016)	3.041	(0.021)	3.416	(0.033)
LB	3.432	(0.01)	3.210	(0.014)	3.210	(0.017)	3.186	(0.026)	3.720	(0.044)
LC	3.432	(0.009)	3.295	(0.013)	3.264	(0.012)	3.268	(0.026)	3.149	(0.011)
Q	3.407	(0.013)	3.003	(0.017)	2.982	(0.022)	2.912	(0.025)	3.045	(0.018)
H1c	3.431	(0.01)	3.429	(0.015)						
H2c	3.432	(0.01)			2.799	(0.013)				
H3c	3.431	(0.008)					3.684	(0.031)		
H4c	3.430	(0.009)							3.534	(0.015)
H1u	3.472	(0.009)	3.295	(0.013)						
H2u	3.396	(0.011)			2.999	(0.017)				
H3u	3.461	(0.008)					3.383	(0.024)		
H4u	3.472	(0.008)							3.296	(0.009)
actual	3.432	(0.015)	3.130	(0.033)						

Abbreviations: LA, LB and LC – wage predictions based on LAW, LBW and LCW models, respectively. Q – wage predictions based on the QW model. H1c, H2c, H3c and H4c – conditional wage predictions based on H1W, H2W, H3W and H4W, respectively. H1u, H2u, H3u and H4u – unconditional wage predictions based on H1W, H2W, H3W and H4W, respectively.

TABLE 12
Mean predicted gross wages for men

	employed		*experienced unemployed*		*experienced inactive*		*inexperienced unemployed*		*inexperienced inactive*	
	Mean	**St. err.**	**Mean**	**St. err.**	**Mean**	**St. err.**	**Mean**	**St. err.**	**Mean**	**St. err.**
LA	3.558	(0.008)	3.301	(0.009)	3.340	(0.02)	3.163	(0.013)	3.049	(0.017)
LB	3.558	(0.007)	3.391	(0.01)	3.426	(0.022)	3.270	(0.015)	3.145	(0.017)
LC	3.558	(0.007)	3.435	(0.009)	3.453	(0.016)	3.355	(0.02)	3.266	(0.024)
Q1	3.558	(0.012)	3.181	(0.015)	3.235	(0.03)	2.970	(0.026)	2.752	(0.022)
H1c	3.557	(0.007)	3.614	(0.011)						
H2c	3.558	(0.007)			3.569	(0.023)				
H3c	3.394	(0.007)					3.674	(0.028)		
H4c	3.558	(0.007)							3.644	(0.041)
H1u	3.601	(0.007)	3.472	(0.009)						
H2u	3.564	(0.007)			3.460	(0.021)				
H3u	3.428	(0.007)					3.368	(0.02)		
H4u	3.570	(0.007)							3.311	(0.022)
actual	3.558	(0.012)	3.328	(0.047)						

Abbreviations: LA, LB and LC – wage predictions based on LAM, LBM and LCM models, respectively. Q – wage predictions based on the QM model. H1c, H2c, H3c and H4c – conditional wage predictions based on H1M, H2M, H3M and H4M, respectively. H1u, H2u, H3u and H4u – unconditional wage predictions based on H1M, H2M, H3M and H4M, respectively.

The bottom rows of table 11 and table 12 show the means of actual wages (*hgwln*) for *employed*. They also present the means of actual wages for *experienced unemployed* who have worked during IRY (see footnote 21); these means are 3.130 and 3.328 for women and men, respectively. We can compare them with the means of conditional (unconditional) wage predictions based on H1W and H1M, which equalled 3.429 and 3.614 (3.295 and 3.472) for women and men, respectively. The conjecture arising from this comparison is that the use of HSM-based wage predictions, either unconditional or conditional, may lead us to significant over-prediction of wages of non-employed people. This over-prediction occurs when $\hat{\Lambda}$ is negative as in the majority of HSM specifications. Conversely, it seems much "easier to accept" HSM wage predictions when $\hat{\Lambda}$ is positive, the case which appears for *experienced inactive* women in the H2W specification.

6 CONCLUSION

Different micro-econometric models require information on wages of non-employed persons that could be earned by these persons if they were employed. Generally, such information does not exist in databases commonly used in empirical research, and these hypothetical wages must somehow be predicted from the available data.

This paper presents the findings of research devoted to wage predictions for selected subgroups of non-employed persons. The predictions will be used in future research, such as the calculation of METREM and in labour supply models. The database used is 2012 SILC for Croatia.

We have pursued two common methods in wage estimation – LRM and HSM. Our LRM results conform to usual findings; wage increases with education and work experience and depends on occupation, industry section and job characteristics. However, despite relatively high coefficients of determination, the results were not fully satisfactory because the residuals analysis has indicated that the models fail to predict the wages properly at the bottom and upper parts of wage distributions. Consequently, LRM over-predicts wages of low-wage earners, which are a focus of our future research. In an attempt to cure this problem, we introduced an *ad hoc* model that uses a quantile regression model. Such a model appears able to improve predictions significantly at the tails of the wage distribution, but further investigation is needed.

The use of HSM has indicated several difficulties. First, there is a general question whether HSM is appropriate for predicting wages of non-employed. According to Paci and Reilly (2004), HSM wage predictions do not reflect the wages that could actually be obtained in the market but rather the "wage offers" of persons based on their personal characteristics. Second, this problem is intensified in the case of negative correlation between random terms from participation and wage equations. In such cases, HSM predictions are generally higher than are those obtained by LRM. Statistically significant negative correlation occurs in most of our specifications.

Third, HSM is relatively complex to implement due to the existence of the participation equation and the assumptions required for correct estimation of the model. In modelling participation and wage equations, we have followed the usual recipes and included all of the common variables (that were available in SILC). Furthermore, non-employed are carefully divided into the subgroups of unemployed and inactive, which is another requirement for proper specification of the participation equation.

Thus, the study has left several questions open. However, the paper should provide a useful contribution for further investigation into predicting wages of non-employed persons and for the analysis of unemployment and inactivity in Croatia.

APPENDIX

APPENDIX 1: SELECTED MEASURES OF FIT FOR PROBIT MODEL

Various measures of fit are available for probit models (Veall and Zimmermann, 1992; Williams, 2015; UCLA, 2011). We use McFadden's pseudo R2, the Adjusted McFadden's pseudo R2, the Count pseudo R2 and the Adjusted count pseudo R2. McFadden's pseudo R2 is defined as follows:

$$\text{MFR2} = 1 - \frac{LL_M}{LL_0} \tag{A1}$$

where LL_M and LL_0 represent the log-likelihoods of the complete model and of the model that uses the intercept only, respectively. The Adjusted McFadden's pseudo R2 adjusts MFR2 for the number of regressors, H, plus one for the intercept, as follows:

$$\text{AMFR2} = 1 - \frac{LL_M - (H+1)}{LL_0} \tag{A2}$$

Because we use sampling weights whose average is approximately 300 in our estimations, LL_M and LL_0 are artificially inflated, and there is virtually no difference between MFR2 and AMFR2. Therefore, to produce the eligible estimate of AMFR2, we initially deflate LL_M and LL_0 by the mean of the sampling weights and then calculate AMFR2 using these deflated values.

The "Count pseudo R2" and the "Adjusted count pseudo R2" are based on the so-called classification tables, which are calculated as follows.

Assume that there are I persons in the sample, $i = \{1, ..., I\}$, of which K persons, $i = \{1, ..., K\}$, have the "positive" outcome, whereas $I-K$ persons, $i = \{K+1, ..., I\}$, have the "negative" outcome. For example, positive and negative outcomes can be *employed* and *experienced unemployed*, respectively.

The probit model calculates the estimate of the probability of each person i in the sample to have a positive outcome. This estimate, $\Phi(Z_i\hat{\beta})$, ranges from 0 to 1. If $\Phi(Z_i\hat{\beta}) \geq \pi$, i is classified as positive; otherwise, if $\Phi(Z_i\hat{\beta}) < \pi$, i is "classified as negative". We have four possibilities:

(a) Actual positive is classified as positive: if $i = \{1, ..., K\}$ and $\Phi(Z_i\hat{\beta}) \geq \pi$

(b) Actual positive is classified as negative: if $i = \{1, ..., K\}$ and $\Phi(Z_i\hat{\beta}) < \pi$

(c) Actual negative is classified as positive: if $i = \{K+1, ..., I\}$ and $\Phi(Z_i\hat{\beta}) \geq \pi$

(d) Actual negative is classified as negative: if $i = \{K+1, ..., I\}$ and $\Phi(Z_i\hat{\beta}) < \pi$

Let s_π^{PP}, s_π^{PN}, s_π^{NN} and s_π^{NP} denote the numbers of persons satisfying condition (a), (b), (c) and (d) in total number of persons, *I*. The following scheme is the *classification table*, containing the shares of persons falling into each of the four categories:

	Actual positive...	**Actual negative...**	**Total**
...classified as positive	s_π^{PP}	s_π^{NP}	$s_\pi^{PP} + s_\pi^{NP}$
...classified as negative	s_π^{PN}	s_π^{NN}	$s_\pi^{PN} + s_\pi^{NN}$
Total	$s_\pi^{PP} + s_\pi^{PN} = p$	$s_\pi^{NP} + s_\pi^{NN} = n$	$p + n = 1$

The number $s_\pi^{PP} + s_\pi^{NP}$ is the share of persons correctly classified by the model. Assuming that $\pi = 0.5$, we can obtain the Count pseudo R2 measure, as follows:

$$\text{CPR2} = s_{0.5}^{PP} + s_{0.5}^{NN} \tag{A3}$$

which represents the share of correctly classified in the total sample. The weakness of CPR2 is manifested in cases when one of the outcomes is much more frequent than is the other. For example, imagine the sample in which 90% of persons are employed and 10% are unemployed; the probit model yields the following results: $s_{0.5}^{PP} = 0.88$, $s_{0.5}^{NP} = 0.02$, $s_{0.5}^{PN} = 0.08$ and $s_{0.5}^{NN} = 0.02$. CPR2 would equal 0.9, which can be deemed very high, although the model has almost completely failed to classify the unemployed properly.

Therefore, "Adjusted count pseudo R2" is suggested and is obtained as follows:

$$\text{ACPR2} = \frac{s_{0.5}^{PP} + s_{0.5}^{NN} - \max(p,n)}{1 - \max(p,n)} \tag{A4}$$

ACPR2 corrects CPR2 by the share of persons with more frequent outcome. In our example, ACPR2=(0.9–0.9)/(1.0–0.9)=0, which better reflects the quality of the model.

Another interesting indicator is the ratio s_π^{NP} / n, which represents a share of incorrectly classified negative observations in the total number of negative observations. Similarly, s_π^{PN} / p can be defined. In calculating these indicators, we can choose $\pi = 0.5$, but such a choice may be too restrictive when s^N is relatively small. Namely, the probit model calculates coefficients $\hat{\beta}$, such that the mean of $\Phi(Z_i\hat{\beta})$ for all observations in the sample equals s^P. If the value $\pi = p$ is chosen, s_p^N / n measures the share of actual negative persons for whom $\Phi(Z_i\hat{\beta}) > p$, i.e., for whom the probability of employment is greater than the overall sample probability of employment.

APPENDIX 2: OTHER TABLES

TABLE A1

Variables and their descriptions

Variable	Description
Age	
ag_year	age in years, in the middle of IRY
ag_ysqr	*ag_year* squared / 100
ag_1525	[bin] aged 16-25 years
ag_2540	[bin] aged 25-40 years
ag_4055	[bin] aged 40-55 years
ag_5565	[bin] aged 55-65 years
Marital status	
ms_mard	[bin] married
ms_nmnp	[bin] never married, does not live with a partner in a household
ms_nmhp	[bin] never married, lives with a partner in a household
ms_divo	[bin] divorced
ms_widw	[bin] widowed
Children	
ch_p0002	number of own parent's children aged 0-2 years
ch_p0306	number of own parent's children aged 3-6 years
ch_p0715	number of own parent's children aged 7-15 years
ch_o0015	number of non-parent's children aged 0-15 years
Education	
ed_nopr	[bin] unfinished primary education
ed_prim	[bin] primary education
ed_seco	[bin] secondary education
ed_tert	[bin] tertiary education
ed_prnp	[bin] primary or unfinished primary education
Area of living	
ar_dens	[bin] inhabitant of densely populated areas
ar_intr	[bin] inhabitant of intermediately populated areas
ar_thin	[bin] inhabitant of thinly populated areas
Health	
hs_badh	[bin] bad or very bad health (self-perceived)
hs_lima	[cat] limitation in activities because of health problems (2 if "strongly limited"; 1 if "limited"; 0 if no limitation)
Wage and income	
hgw	a ratio between yearly gross employment income (gross wage; *bruto plaća*) and yearly working hours
hgwln	(ln) *hgw*
oi_a	(ln) employment and self-employment income (net), earned by other household members
oi_b	(ln) rental and capital income (net), obtained by a household
oi_c	(ln) private transfers received minus private transfers paid, by a household (note: for negative amounts, -ln(-amount) is obtained)
oi_d	(ln) pension income (net), received by other household members

Variable	Description
oi_e	(ln) unemployment and sickness benefits (net), received by other household members
oi_f	(ln) child benefits, maternity and parental leave benefits
oi_g	(ln) imputed rent, obtained by a household (net of interest on mortgage and actual rent paid)
Employment	
we_yipw	work experience: years spent in paid work before the beginning of IRY
we_yisq	*we_yipw* squared / 100
we_yopw	"years out of work": years in which person was not working, measured from the date when the first job was taken, until the beginning of IRY
we_yosq	*we_yopw* squared / 100
em_locs	[bin] works in enterprise local unit with up to 10 employees
em_locl	[bin] works in enterprise local unit with 50 or more employees
em_perj	[bin] has permanent job contract (as opposed to temporary contract)
em_mana	[bin] managerial position at work
em_agri	[bin] agricultural household, defined as a household in which more than 50% of employment/self-employment income is derived from self-employment income in agriculture
Occupation	(according to ISCO-08)
oc_0	[bin] armed forces occupations
oc_1	[bin] managers
oc_2	[bin] professionals
oc_3	[bin] technicians and associate professionals
oc_4	[bin] clerical support workers
oc_5	[bin] service and sales workers
oc_6	[bin] skilled agricultural, forestry and fishery workers
oc_7	[bin] craft and related trades workers
oc_8	[bin] plant and machine operators, and assemblers
oc_9	[bin] elementary occupations
oc_21	[bin] professionals and managers (*oc_2+oc_1*)
oc_30	[bin] technicians, associate professionals, and armed forces occupations (*oc_3+oc_0*)
Industry	(industry sections according to NACE Rev. 2)
in_a	[bin] employed in section A (agriculture, forestry and fishing)
in_bcde	[bin] employed in sections B (mining and quarrying), C (manufacturing), D (electricity, gas, steam and air conditioning supply), or E (water supply, sewerage, waste management and remediation activities)
in_f	[bin] employed in section F (construction)
in_g	[bin] employed in section G (wholesale and retail trade; repair of motor vehicles and motorcycles)
in_h	[bin] employed in section H (transportation and storage)
in_i	[bin] employed in section I (accommodation and food service activities)
in_j	[bin] employed in section J (information and communication)
in_k	[bin] employed in section K (financial and insurance activities)
in_lmn	[bin] employed in section L (real estate activities), M (professional, scientific and technical activities), or N (administrative and support service activities)

Variable	Description
in_o	[bin] employed in section O (public administration and defence; compulsory social security)
in_p	[bin] employed in section P (education)
in_q	[bin] employed in section Q (human health and social work activities)
in_rstu	[bin] employed in section R (arts, entertainment and recreation), S (other service activities), T (activities of households as employers; undifferentiated goods- and services-producing activities of households for own use), or U (activities of extraterritorial organisations and bodies)
in_gi	[bin] the sum of *in_g* and *in_i*
in_jk	[bin] the sum of *in_j*, *in_k*
in_opq	[bin] the sum of *in_o*, *in_p and* in_q

Notes: [bin] denotes binary variable; [cat] denotes categorical variable; and (ln) denotes the natural logarithm of the amount.

TABLE A2

Means and standard deviations of variables – women

	employed		*experienced unemployed*		*inexperienced unemployed*		*experienced inactive*		*inexperienced inactive*	
Variable	**Mean**	**S. d.**	**Mean**	**S. d.**	**Mean**	**S. d.**	**Mean**	**S. d.**	**Mean**	**S. d.**
ag_year	40.82	9.82	38.04	10.61	27.44	9.30	46.65	9.04	43.77	11.23
ms_mard	0.66	0.47	0.64	0.48	0.23	0.42	0.86	0.34	0.81	0.39
ms_nmnp	0.22	0.41	0.26	0.44	0.68	0.47	0.04	0.19	0.09	0.28
ms_nmhp	0.01	0.11	0.03	0.16	0.04	0.19	0.03	0.16	0.05	0.22
ms_divo	0.07	0.25	0.06	0.24	0.04	0.19	0.02	0.15	0.02	0.12
ms_widw	0.05	0.21	0.01	0.10	0.01	0.11	0.05	0.21	0.04	0.18
ch_p0002	0.01	0.11	0.03	0.18	0.01	0.10	0.02	0.17	0.04	0.21
ch_p0306	0.11	0.36	0.22	0.51	0.10	0.33	0.17	0.45	0.15	0.45
ch_p0715	0.36	0.68	0.39	0.69	0.21	0.53	0.42	0.72	0.42	0.77
ch_o0015	0.09	0.38	0.14	0.44	0.30	0.73	0.13	0.56	0.22	0.66
ed_nopr	0.00	0.04	0.01	0.12	0.01	0.11	0.02	0.14	0.15	0.35
ed_prim	0.09	0.29	0.18	0.38	0.13	0.34	0.30	0.46	0.50	0.50
ed_seco	0.66	0.47	0.68	0.47	0.62	0.49	0.66	0.47	0.29	0.45
ed_tert	0.25	0.43	0.13	0.34	0.23	0.42	0.02	0.14	0.02	0.12
ed_prnp	0.09	0.29	0.19	0.39	0.14	0.35	0.32	0.47	0.65	0.48
ar_dens	0.38	0.49	0.21	0.41	0.23	0.42	0.20	0.40	0.16	0.37
ar_intr	0.23	0.42	0.18	0.39	0.16	0.36	0.19	0.39	0.15	0.36
ar_thin	0.39	0.49	0.61	0.49	0.61	0.49	0.60	0.49	0.68	0.47
hs_badh	0.03	0.17	0.07	0.25	0.04	0.19	0.17	0.38	0.13	0.34
hs_lima	0.08	0.29	0.11	0.33	0.09	0.36	0.27	0.52	0.19	0.43
hgw	34.79	19.95	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
hgwln	3.43	0.47	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
oi_a	8.35	4.81	8.15	4.72	8.47	4.63	8.82	4.50	7.92	4.94
oi_b	0.97	2.70	0.62	2.26	0.48	1.97	0.96	2.70	0.38	1.74
oi_c	0.08	2.60	0.46	2.57	0.25	2.21	0.16	2.82	0.24	1.85
oi_d	3.73	4.98	4.08	5.04	3.73	4.92	3.56	4.91	4.06	5.06
oi_e	0.46	1.99	0.67	2.32	0.35	1.71	0.53	2.08	0.65	2.37
oi_f	1.36	3.14	3.54	4.31	3.10	4.23	3.01	4.26	3.26	4.35
oi_g	9.71	1.68	9.45	1.88	9.34	1.88	9.61	1.65	9.62	0.64
we_yipw	16.13	10.28	9.60	8.82	0.00	0.00	13.33	9.76	0.00	0.00
we_yopw	0.79	2.41	3.50	5.48	6.01	9.90	7.97	8.05	24.39	14.42
em_locs	0.26	0.44	0.11	0.31	0.07	0.26	0.00	0.00	0.00	0.00
em_locl	0.34	0.47	0.04	0.20	0.01	0.10	0.00	0.00	0.00	0.00
em_perj	0.88	0.32	0.33	0.47	0.03	0.16	0.62	0.49	0.01	0.09
em_mana	0.11	0.31	0.04	0.18	0.00	0.00	0.02	0.15	0.00	0.00
em_agri	0.01	0.09	0.03	0.16	0.01	0.10	0.04	0.18	0.03	0.17
oc_21	0.23	0.42	0.05	0.23	0.07	0.25	0.02	0.15	0.00	0.00
oc_30	0.12	0.33	0.08	0.27	0.00	0.00	0.05	0.22	0.00	0.00
oc_4	0.17	0.38	0.11	0.31	0.01	0.09	0.11	0.31	0.00	0.00
oc_5	0.26	0.44	0.42	0.49	0.10	0.30	0.31	0.46	0.01	0.09
oc_6	0.01	0.07	0.01	0.10	0.01	0.08	0.04	0.19	0.00	0.00
oc_7	0.06	0.24	0.10	0.30	0.01	0.10	0.16	0.37	0.01	0.11
oc_8	0.05	0.21	0.07	0.25	0.00	0.00	0.11	0.31	0.00	0.06
oc_9	0.09	0.29	0.16	0.37	0.01	0.12	0.19	0.39	0.00	0.00
in_a	0.01	0.12	0.01	0.08	0.00	0.06	0.00	0.00	0.00	0.00
in_bcde	0.17	0.38	0.04	0.20	0.01	0.10	0.00	0.00	0.00	0.00
in_f	0.01	0.11	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
in_gi	0.24	0.43	0.09	0.29	0.06	0.23	0.00	0.00	0.00	0.00
in_h	0.02	0.15	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
in_jk	0.06	0.23	0.01	0.07	0.00	0.00	0.00	0.00	0.00	0.00
in_lmn	0.06	0.24	0.05	0.22	0.03	0.16	0.00	0.00	0.00	0.00
in_opq	0.32	0.47	0.03	0.18	0.04	0.20	0.00	0.00	0.00	0.00
in_rstu	0.04	0.19	0.02	0.13	0.02	0.15	0.00	0.00	0.00	0.00

TABLE A3

Means and standard deviations of variables – men

	employed		*experienced unemployed*		*inexperienced unemployed*		*experienced inactive*		*inexperienced inactive*	
Variable	**Mean**	**S. d.**	**Mean**	**S. d.**	**Mean**	**S. d.**	**Mean**	**S. d.**	**Mean**	**S. d.**
ag_year	40.66	10.80	38.30	11.43	25.30	7.48	49.02	10.98	29.25	10.56
ms_mard	0.67	0.47	0.46	0.50	0.10	0.30	0.61	0.49	0.17	0.38
ms_nmnp	0.28	0.45	0.46	0.50	0.86	0.35	0.26	0.44	0.69	0.46
ms_nmhp	0.02	0.14	0.03	0.18	0.02	0.15	0.06	0.23	0.08	0.27
ms_divo	0.03	0.16	0.04	0.20	0.02	0.14	0.06	0.23	0.06	0.23
ms_widw	0.00	0.07	0.00	0.06	0.00	0.00	0.02	0.14	0.00	0.00
ch_p0002	0.10	0.30	0.07	0.27	0.04	0.18	0.03	0.16	0.05	0.23
ch_p0306	0.16	0.44	0.12	0.39	0.06	0.26	0.08	0.28	0.21	0.54
ch_p0715	0.34	0.67	0.24	0.58	0.05	0.24	0.15	0.48	0.20	0.64
ch_o0015	0.07	0.33	0.11	0.42	0.27	0.68	0.14	0.52	0.52	1.02
ed_nopr	0.01	0.08	0.03	0.17	0.02	0.15	0.04	0.19	0.06	0.24
ed_prim	0.07	0.26	0.14	0.35	0.09	0.29	0.23	0.42	0.23	0.42
ed_seco	0.74	0.44	0.77	0.42	0.72	0.45	0.69	0.46	0.68	0.47
ed_tert	0.18	0.39	0.06	0.23	0.16	0.36	0.03	0.18	0.03	0.16
ed_prnp	0.08	0.27	0.17	0.38	0.11	0.32	0.27	0.44	0.29	0.45
ar_dens	0.30	0.46	0.28	0.45	0.33	0.47	0.20	0.40	0.11	0.31
ar_intr	0.21	0.41	0.15	0.36	0.15	0.36	0.18	0.39	0.25	0.43
ar_thin	0.48	0.50	0.57	0.49	0.52	0.50	0.62	0.49	0.64	0.48
hs_badh	0.02	0.12	0.05	0.21	0.01	0.08	0.16	0.37	0.07	0.25
hs_lima	0.07	0.28	0.10	0.32	0.03	0.18	0.34	0.50	0.09	0.29
hgw	38.94	20.58	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
hgwln	3.56	0.44	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
oi_a	7.63	5.07	6.50	5.33	7.81	5.09	5.96	5.41	7.43	5.31
oi_b	0.86	2.56	0.55	2.09	0.97	2.75	0.82	2.56	0.18	1.29
oi_c	-0.24	1.98	0.48	2.76	0.18	2.77	0.04	1.75	-0.15	1.15
oi_d	3.20	4.77	3.92	4.96	5.19	5.12	3.39	4.62	4.43	5.14
oi_e	0.63	2.31	0.58	2.20	1.17	3.01	0.30	1.62	0.65	2.32
oi_f	2.05	3.81	2.55	4.06	2.09	3.72	2.31	3.89	2.98	4.41
oi_g	9.51	2.05	9.60	1.51	9.66	1.29	9.70	0.55	9.82	0.67
we_yipw	16.96	10.80	11.54	10.20	0.00	0.00	18.63	11.66	0.00	0.00
we_yopw	0.57	1.93	2.88	4.95	4.72	7.24	5.64	7.04	10.91	10.64
em_locs	0.18	0.39	0.07	0.26	0.03	0.17	0.00	0.00	0.00	0.00
em_locl	0.36	0.48	0.05	0.22	0.03	0.16	0.00	0.00	0.00	0.00
em_perj	0.90	0.29	0.35	0.48	0.01	0.10	0.64	0.48	0.00	0.00
em_mana	0.14	0.35	0.06	0.23	0.00	0.00	0.06	0.24	0.00	0.00
em_agri	0.00	0.06	0.01	0.11	0.02	0.15	0.04	0.18	0.00	0.00
oc_21	0.13	0.33	0.04	0.19	0.03	0.18	0.06	0.24	0.00	0.00
oc_30	0.19	0.39	0.11	0.31	0.01	0.11	0.11	0.31	0.00	0.00
oc_4	0.08	0.27	0.08	0.27	0.00	0.00	0.08	0.26	0.00	0.00
oc_5	0.14	0.35	0.14	0.35	0.04	0.19	0.11	0.31	0.01	0.12
oc_6	0.02	0.14	0.03	0.16	0.00	0.07	0.01	0.10	0.00	0.00
oc_7	0.25	0.43	0.31	0.46	0.05	0.21	0.27	0.45	0.00	0.00
oc_8	0.15	0.36	0.15	0.36	0.03	0.18	0.25	0.43	0.00	0.00
oc_9	0.05	0.21	0.15	0.35	0.01	0.08	0.11	0.32	0.00	0.00
in_a	0.04	0.19	0.00	0.07	0.00	0.07	0.00	0.00	0.00	0.00
in_bcde	0.30	0.46	0.06	0.24	0.03	0.18	0.00	0.00	0.00	0.00
in_f	0.11	0.32	0.03	0.17	0.02	0.12	0.00	0.00	0.00	0.00
in_gi	0.15	0.36	0.07	0.25	0.01	0.10	0.00	0.00	0.00	0.00
in_h	0.10	0.30	0.03	0.17	0.00	0.00	0.00	0.00	0.00	0.00
in_jk	0.05	0.21	0.00	0.06	0.00	0.00	0.00	0.00	0.00	0.00
in_lmn	0.05	0.21	0.03	0.18	0.01	0.08	0.00	0.00	0.00	0.00
in_opq	0.14	0.35	0.01	0.12	0.01	0.09	0.00	0.00	0.00	0.00
in_rstu	0.02	0.12	0.01	0.10	0.00	0.06	0.00	0.00	0.00	0.00

TABLE A4

Results of probit models P1 to P5 for women

	P1W		P2W		P3W		P4W		P5W	
	Coeff.	**Sig. l.**	**Coeff.**	**Sig. l.**	**Coeff.**	**Sig. l.**	**Coeff.**	**Sig. l.**	**Coeff.**	**Sig. l.**
ag_year	**0.121**	(0)	0.098	(0.07)	**0.351**	(0)	**0.213**	(0)	-0.019	(0.74)
ag_ysqr	**-0.133**	(0)	**-0.161**	(0.01)	**-0.383**	(0)	**-0.274**	(0)	-0.033	(0.63)
ar_dens	0.147	(0.23)	0.154	(0.37)	-0.121	(0.56)	-0.040	(0.8)	-0.136	(0.52)
ar_thin	**-0.256**	(0.02)	-0.096	(0.49)	**-0.455**	(0.01)	**-0.522**	(0)	0.088	(0.58)
ms_mard	0.251	(0.1)	0.193	(0.48)	0.428	(0.05)	-0.215	(0.34)	-0.303	(0.28)
ms_nmhp	-0.090	(0.74)	**-1.047**	(0.03)	-0.143	(0.71)	**-0.822**	(0.01)	-0.619	(0.19)
ms_divo	0.084	(0.69)	**0.765**	(0.04)	0.039	(0.91)	0.507	(0.18)	0.460	(0.21)
ms_widw	**1.067**	(0)	0.343	(0.32)	0.485	(0.27)	0.302	(0.32)	**-0.962**	(0.02)
ch_p0002	-0.210	(0.44)	-0.145	(0.71)	0.437	(0.46)	0.039	(0.89)	-0.536	(0.09)
ch_p0306	**-0.256**	(0.03)	**-0.722**	(0)	-0.077	(0.69)	-0.225	(0.11)	-0.286	(0.09)
ch_p0715	0.047	(0.52)	-0.141	(0.09)	-0.138	(0.22)	-0.082	(0.33)	**-0.271**	(0.01)
ch_o0015	0.130	(0.23)	0.222	(0.08)	0.228	(0.07)	0.035	(0.72)	-0.009	(0.95)
ed_prnp	-0.218	(0.1)	-0.289	(0.07)	**-0.891**	(0)	**-1.387**	(0)	-0.006	(0.97)
ed_tert	**-0.403**	(0.01)	**0.782**	(0)	**-0.309**	(0.05)	**0.690**	(0)	**1.112**	(0)
hs_badh	**-0.457**	(0.01)	**-0.970**	(0)	**-0.668**	(0.01)	**-0.544**	(0)	**-0.403**	(0.03)
we_yopw	**-0.166**	(0)	**-0.209**	(0)					-0.031	(0.15)
we_yosq	**0.369**	(0.01)	**0.398**	(0)					0.006	(0.95)
oc_21	**0.981**	(0)	0.406	(0.15)					-0.395	(0.31)
oc_30	0.213	(0.19)	0.009	(0.97)					-0.062	(0.81)
oc_4	**0.301**	(0.02)	0.082	(0.6)					-0.142	(0.46)
oc_6	0.309	(0.45)	-0.564	(0.14)					**-1.143**	(0)
oc_7	0.004	(0.98)	**-0.481**	(0.01)					-0.304	(0.11)
oc_8	0.086	(0.61)	-0.173	(0.34)					-0.200	(0.32)
oc_9	0.034	(0.81)	0.176	(0.43)					-0.150	(0.44)
oi_a	0.004	(0.62)	-0.023	(0.07)	0.003	(0.82)	0.006	(0.6)	**-0.037**	(0.01)
oi_b	0.013	(0.43)	-0.041	(0.09)	0.028	(0.25)	0.030	(0.19)	-0.047	(0.05)
oi_c	-0.024	(0.14)	-0.025	(0.27)	-0.007	(0.78)	-0.033	(0.08)	0.015	(0.55)
oi_d	-0.003	(0.73)	0.013	(0.25)	0.011	(0.38)	0.003	(0.8)	-0.007	(0.59)
oi_e	-0.007	(0.69)	0.010	(0.67)	0.046	(0.16)	0.013	(0.49)	0.029	(0.26)
oi_f	**-0.071**	(0)	**-0.061**	(0)	**-0.065**	(0)	**-0.050**	(0)	0.019	(0.32)
oi_g	0.021	(0.37)	0.022	(0.61)	0.029	(0.33)	**-0.049**	(0.04)	-0.003	(0.94)
em_agri	-0.158	(0.6)	0.011	(0.97)	0.177	(0.6)	-0.401	(0.17)	-0.400	(0.21)
_cons	**-1.865**	(0.01)	0.597	(0.56)	**-5.777**	(0)	-1.368	(0.13)	**2.786**	(0.01)
Observations	2,034		1,880		1,699		1,883		860	
MFR2	0.229		0.435		0.360		0.357		0.225	
AMFR2	0.200		0.394		0.313		0.326		0.167	
CPR2	0.799		0.898		0.930		0.893		0.738	
ACPR2	0.142		0.326		0.221		0.257		0.290	
p	0.766		0.848		0.910		0.856		0.630	
n	0.234		0.152		0.090		0.144		0.370	
s(PP, 0.5)	0.717		0.820		0.897		0.824		0.529	
s(PN, 0.5)	0.049		0.028		0.014		0.031		0.102	
s(NN, 0.5)	0.083		0.078		0.034		0.068		0.209	
s(NP, 0.5)	0.151		0.074		0.056		0.076		0.161	
s(PN, 0.5) / p	0.064		0.033		0.015		0.036		0.162	
s(NP, 0.5) / n	0.647		0.488		0.625		0.528		0.435	
s(PP, p)	0.572		0.708		0.726		0.705		0.429	
s(PN, p)	0.194		0.140		0.185		0.151		0.202	
s(NN, p)	0.172		0.126		0.076		0.114		0.283	
s(NP, p)	0.062		0.026		0.014		0.031		0.086	
s(PN, p) / p	0.253		0.165		0.203		0.176		0.320	
s(NP, p) / n	0.264		0.173		0.153		0.214		0.234	

Notes: Specifications are explained in section 4.2 (table 6). Coefficients significant at the 5% level are marked as bold.

Abbreviations: coeff. – coefficient, sig. l. – significance level; see appendix 1 for explanation of the measures of fit. s(PP, 0.5) denotes $s^{PP}_{0.5}$ *(subsequent indicators are denoted analogously).*

TABLE A5

Results of probit models P6 to P10 for women

	P6W		P7W		P8W		P9W		P10W	
	Coeff.	**Sig. l.**	**Coeff.**	**Sig. l.**	**Coeff.**	**Sig. l.**	**Coeff.**	**Sig. l.**	**Coeff.**	**Sig. l.**
ag_year	**0.253**	(0)	**0.144**	(0)	**0.229**	(0)	**0.148**	(0.01)	0.000	(1)
ag_ysqr	**-0.264**	(0)	**-0.193**	(0)	-0.149	(0.12)	**-0.127**	(0.05)	-0.071	(0.35)
ar_dens	-0.184	(0.44)	-0.183	(0.38)	-0.332	(0.29)	0.058	(0.78)	0.336	(0.29)
ar_thin	-0.039	(0.84)	-0.099	(0.55)	**-0.562**	(0.04)	-0.093	(0.57)	0.250	(0.29)
ms_mard	0.393	(0.11)	-0.466	(0.06)	0.645	(0.11)	-0.254	(0.49)	**-0.810**	(0.01)
ms_nmhp	0.112	(0.76)	**-0.871**	(0.02)	**1.747**	(0)	-0.240	(0.65)	**-0.995**	(0.02)
ms_divo	-0.053	(0.9)	0.481	(0.22)	-0.755	(0.17)	0.025	(0.96)	0.682	(0.25)
ms_widw	-0.368	(0.47)	**-0.941**	(0.02)	0.233	(0.69)	0.021	(0.97)	-0.315	(0.62)
ch_p0002	0.302	(0.57)	-0.131	(0.62)	1.044	(0.06)	**0.742**	(0.05)	-0.451	(0.37)
ch_p0306	0.164	(0.39)	-0.027	(0.86)	0.470	(0.06)	0.265	(0.21)	-0.394	(0.13)
ch_p0715	-0.141	(0.38)	**-0.243**	(0.02)	0.108	(0.57)	0.077	(0.44)	-0.001	(1)
ch_o0015	0.160	(0.23)	-0.163	(0.17)	0.247	(0.3)	-0.118	(0.41)	-0.098	(0.4)
ed_prnp	**-0.565**	(0.01)	**-1.074**	(0)	-0.391	(0.14)	**-0.984**	(0)	**-0.751**	(0)
ed_tert	**-0.408**	(0.04)	0.509	(0.06)	**-1.164**	(0.01)	-0.189	(0.65)	**0.834**	(0.01)
hs_badh	-0.124	(0.68)	-0.078	(0.69)	0.312	(0.35)	0.291	(0.07)	0.215	(0.53)
we_yopw										
we_yosq										
oc_21										
oc_30										
oc_4										
oc_6										
oc_7										
oc_8										
oc_9										
oi_a	0.007	(0.65)	0.003	(0.82)	0.044	(0.08)	**0.032**	(0.02)	0.003	(0.89)
oi_b	0.006	(0.85)	-0.007	(0.77)	0.057	(0.12)	0.048	(0.08)	0.028	(0.42)
oi_c	0.006	(0.83)	0.014	(0.55)	-0.014	(0.69)	-0.021	(0.39)	0.019	(0.67)
oi_d	0.022	(0.12)	0.011	(0.36)	**0.045**	(0.05)	0.006	(0.64)	0.013	(0.48)
oi_e	0.055	(0.1)	0.029	(0.16)	0.018	(0.76)	0.004	(0.86)	0.016	(0.71)
oi_f	-0.008	(0.7)	0.028	(0.11)	-0.009	(0.81)	0.019	(0.36)	0.021	(0.39)
oi_g	0.042	(0.25)	**-0.067**	(0.04)	0.084	(0.07)	-0.037	(0.42)	**-0.371**	(0)
em_agri	**0.894**	(0.05)	-0.023	(0.95)	0.914	(0.12)	0.319	(0.33)	-0.404	(0.36)
_cons	**-4.978**	(0)	-0.586	(0.53)	**-7.406**	(0)	**-3.342**	(0.01)	**4.606**	(0)
Observations	679		863		525		709		528	
MFR2	0.248		0.222		0.619		0.146		0.478	
AMFR2	0.184		0.179		0.548		0.097		0.409	
CPR2	0.823		0.757		0.904		0.698		0.861	
ACPR2	0.274		0.318		0.729		0.379		0.622	
p	0.756		0.644		0.645		0.515		0.368	
n	0.244		0.356		0.355		0.485		0.632	
s(PP, 0.5)	0.701		0.544		0.620		0.358		0.291	
s(PN, 0.5)	0.055		0.100		0.026		0.157		0.078	
s(NN, 0.5)	0.122		0.213		0.284		0.341		0.570	
s(NP, 0.5)	0.122		0.143		0.071		0.145		0.061	
s(PN, 0.5) / p	0.073		0.155		0.040		0.305		0.211	
s(NP, 0.5) / n	0.500		0.401		0.199		0.298		0.097	
s(PP, p)	0.574		0.506		0.588		0.351		0.312	
s(PN, p)	0.182		0.138		0.057		0.164		0.056	
s(NN, p)	0.177		0.251		0.304		0.348		0.545	
s(NP, p)	0.067		0.105		0.051		0.138		0.087	
s(PN, p) / p	0.241		0.215		0.089		0.319		0.152	
s(NP, p) / n	0.273		0.295		0.143		0.284		0.137	

Notes: see notes and abbreviations for table A4.

TABLE A6

Results of probit models P1 to P5 for men

	P1M		P2M		P3M		P4M		P5M	
	Coeff.	**Sig. l.**	**Coeff.**	**Sig. l.**	**Coeff.**	**Sig. l.**	**Coeff.**	**Sig. l.**	**Coeff.**	**Sig. l.**
ag_year	**0.062**	(0.01)	**0.147**	(0)	**0.283**	(0)	**0.188**	(0)	0.060	(0.13)
ag_ysqr	**-0.058**	(0.04)	**-0.204**	(0)	**-0.289**	(0)	**-0.202**	(0)	**-0.121**	(0.01)
ar_dens	**-0.255**	(0.02)	0.053	(0.77)	**-0.422**	(0.02)	0.334	(0.18)	0.307	(0.18)
ar_thin	**-0.193**	(0.04)	0.017	(0.91)	-0.171	(0.28)	0.025	(0.9)	0.032	(0.86)
ms_mard	**0.366**	(0)	**0.370**	(0.04)	**0.552**	(0.01)	**0.985**	(0)	0.147	(0.48)
ms_nmhp	-0.016	(0.95)	**-0.711**	(0.05)	0.232	(0.42)	-0.081	(0.76)	-0.508	(0.19)
ms_divo	-0.163	(0.4)	0.240	(0.49)	-0.255	(0.51)	-0.058	(0.89)	0.403	(0.22)
ms_widw										
ch_p0002	**0.325**	(0.03)	0.557	(0.15)	0.517	(0.06)	0.177	(0.57)	0.215	(0.58)
ch_p0306	0.142	(0.2)	-0.029	(0.89)	-0.083	(0.62)	**-0.444**	(0.03)	-0.048	(0.83)
ch_p0715	**0.159**	(0.03)	0.230	(0.1)	0.013	(0.92)	**-0.313**	(0.03)	0.165	(0.38)
ch_o0015	0.105	(0.29)	0.156	(0.27)	0.125	(0.31)	**-0.289**	(0.04)	-0.012	(0.94)
ed_prnp	-0.074	(0.52)	-0.230	(0.12)	**-0.555**	(0)	**-0.698**	(0)	-0.121	(0.45)
ed_tert	0.250	(0.12)	**1.441**	(0)	-0.175	(0.25)	**0.563**	(0.05)	**0.777**	(0.05)
hs_badh	**-0.438**	(0.01)	**-1.189**	(0)	-0.214	(0.68)	**-1.357**	(0)	-0.407	(0.12)
we_yopw	**-0.200**	(0)	**-0.243**	(0)					-0.014	(0.59)
we_yosq	**0.506**	(0)	**0.605**	(0)					-0.002	(0.99)
oc_21	0.344	(0.1)	-0.674	(0.08)					-0.807	(0.06)
oc_30	0.079	(0.56)	-0.275	(0.22)					-0.135	(0.62)
oc_4	0.017	(0.91)	-0.165	(0.51)					-0.109	(0.7)
oc_6	0.224	(0.32)	**1.062**	(0.02)					0.552	(0.11)
oc_7	-0.129	(0.24)	-0.188	(0.33)					0.039	(0.86)
oc_8	-0.053	(0.67)	-0.386	(0.07)					-0.240	(0.31)
oc_9	**-0.559**	(0)	0.121	(0.64)					0.330	(0.2)
oi_a	0.005	(0.53)	0.021	(0.08)	0.012	(0.29)	0.000	(0.99)	0.011	(0.43)
oi_b	0.029	(0.07)	-0.001	(0.98)	-0.003	(0.9)	0.033	(0.42)	-0.022	(0.42)
oi_c	**-0.083**	(0)	-0.014	(0.59)	-0.023	(0.41)	0.025	(0.49)	**0.057**	(0)
oi_d	-0.002	(0.76)	0.015	(0.25)	**-0.023**	(0.04)	-0.017	(0.34)	0.014	(0.35)
oi_e	0.011	(0.45)	0.030	(0.33)	-0.012	(0.57)	0.020	(0.52)	0.031	(0.4)
oi_f	**-0.047**	(0)	**-0.053**	(0.01)	**-0.045**	(0.03)	-0.005	(0.85)	-0.024	(0.32)
oi_g	-0.032	(0.14)	**-0.116**	(0.01)	-0.020	(0.53)	-0.129	(0.17)	-0.092	(0.05)
em_agri	**-0.700**	(0.04)	**-1.383**	(0)	-0.628	(0.09)	-0.849	(0.56)	-0.598	(0.15)
_cons	-0.242	(0.64)	0.591	(0.46)	**-4.141**	(0)	-0.828	(0.25)	1.566	(0.09)
Observations	2,552		2,073		2,109		1,979		791	
MFR2	0.190		0.365		0.377		0.327		0.198	
AMFR2	0.171		0.307		0.347		0.268		0.114	
CPR2	0.811		0.953		0.936		0.969		0.843	
ACPR2	0.183		0.157		0.251		0.090		0.049	
p	0.769		0.944		0.914		0.965		0.835	
n	0.231		0.056		0.086		0.035		0.165	
s(PP, 0.5)	0.738		0.936		0.904		0.964		0.807	
s(PN, 0.5)	0.031		0.008		0.010		0.001		0.028	
s(NN, 0.5)	0.073		0.017		0.032		0.005		0.036	
s(NP, 0.5)	0.158		0.040		0.054		0.030		0.129	
s(PN, 0.5) / p	0.040		0.008		0.011		0.001		0.034	
s(NP, 0.5) / n	0.683		0.706		0.632		0.868		0.782	
s(PP, p)	0.552		0.791		0.735		0.770		0.625	
s(PN, p)	0.217		0.153		0.179		0.195		0.210	
s(NN, p)	0.163		0.043		0.074		0.029		0.126	
s(NP, p)	0.068		0.013		0.012		0.006		0.039	
s(PN, p) / p	0.282		0.162		0.196		0.202		0.251	
s(NP, p) / n	0.295		0.229		0.140		0.174		0.236	

Notes: see notes and abbreviations for table A4.

TABLE A7

Results of probit models P6 to P10 for men

	P6M		P7M		P8M		P9M		P10M	
	Coeff.	**Sig. l.**	**Coeff.**	**Sig. l.**	**Coeff.**	**Sig. l.**	**Coeff.**	**Sig. l.**	**Coeff.**	**Sig. l.**
ag_year	**0.240**	(0)	**0.173**	(0)	**0.137**	(0.04)	**0.223**	(0)	-0.022	(0.75)
ag_ysqr	**-0.237**	(0)	**-0.191**	(0)	-0.029	(0.74)	**-0.193**	(0.03)	-0.022	(0.83)
ar_dens	-0.061	(0.77)	**0.777**	(0.01)	-0.390	(0.43)	0.260	(0.61)	**1.196**	(0)
ar_thin	0.093	(0.62)	0.204	(0.38)	0.062	(0.89)	-0.223	(0.61)	0.287	(0.32)
ms_mard	0.218	(0.44)	**0.820**	(0.01)	0.042	(0.92)	**1.287**	(0.01)	**1.392**	(0.02)
ms_nmhp	0.019	(0.96)	-0.239	(0.55)	0.456	(0.46)	0.966	(0.12)	1.369	(0.08)
ms_divo	-0.202	(0.62)	0.122	(0.73)	-0.473	(0.41)	0.461	(0.4)	0.067	(0.91)
ms_widw										
ch_p0002	0.175	(0.6)	0.404	(0.47)	-0.261	(0.69)	1.183	(0.16)	-0.632	(0.24)
ch_p0306	-0.135	(0.54)	-0.513	(0.06)	0.113	(0.81)	-0.644	(0.09)	**-1.015**	(0.02)
ch_p0715	-0.042	(0.79)	**-0.528**	(0.02)	-0.090	(0.77)	**-0.968**	(0.01)	-0.599	(0.14)
ch_o0015	0.029	(0.84)	**-0.547**	(0)	-0.407	(0.17)	**-0.746**	(0)	**-0.377**	(0.01)
ed_prnp	-0.127	(0.49)	-0.392	(0.1)	-0.446	(0.22)	-0.361	(0.31)	-0.650	(0.07)
ed_tert	**-0.658**	(0)	0.518	(0.19)	**-1.338**	(0)	-0.527	(0.36)	**0.912**	(0.01)
hs_badh	0.168	(0.77)	-0.856	(0.08)	1.092	(0.09)	-0.927	(0.06)	**-1.704**	(0.02)
we_yopw										
we_yosq										
oc_21										
oc_30										
oc_4										
oc_6										
oc_7										
oc_8										
oc_9										
oi_a	0.008	(0.59)	-0.012	(0.59)	0.003	(0.94)	-0.043	(0.17)	-0.018	(0.57)
oi_b	-0.035	(0.13)	0.032	(0.53)	0.051	(0.25)	0.003	(0.96)	0.073	(0.13)
oi_c	0.009	(0.73)	0.075	(0.14)	-0.020	(0.53)	0.063	(0.42)	0.051	(0.27)
oi_d	-0.021	(0.11)	-0.016	(0.42)	**-0.072**	(0.01)	-0.050	(0.17)	0.023	(0.31)
oi_e	-0.008	(0.77)	0.045	(0.19)	-0.054	(0.46)	0.024	(0.7)	0.035	(0.42)
oi_f	0.000	(1)	0.057	(0.07)	-0.003	(0.95)	0.047	(0.47)	0.061	(0.1)
oi_g	0.023	(0.58)	-0.240	(0.2)	0.433	(0.12)	-0.098	(0.72)	-0.427	(0.05)
em_agri	-0.078	(0.83)			0.596	(0.18)				
_cons	**-4.398**	(0)	0.049	(0.98)	**-8.475**	(0)	-3.606	(0.21)	5.117	(0.05)
Observations	827		697		348		218		254	
MFR2	0.294		0.262		0.649		0.505		0.242	
AMFR2	0.249		0.181		0.555		0.359		0.121	
CPR2	0.830		0.904		0.897		0.878		0.773	
ACPR2	0.286		0.099		0.736		0.676		0.179	
p	0.762		0.894		0.388		0.624		0.724	
n	0.238		0.106		0.612		0.376		0.276	
s(PP, 0.5)	0.693		0.884		0.333		0.555		0.677	
s(PN, 0.5)	0.069		0.010		0.055		0.069		0.046	
s(NN, 0.5)	0.137		0.020		0.565		0.323		0.096	
s(NP, 0.5)	0.100		0.086		0.047		0.053		0.181	
s(PN, 0.5) / p	0.091		0.011		0.142		0.110		0.064	
s(NP, 0.5) / n	0.422		0.808		0.077		0.142		0.654	
s(PP, p)	0.574		0.658		0.338		0.520		0.526	
s(PN, p)	0.188		0.235		0.050		0.104		0.198	
s(NN, p)	0.193		0.093		0.551		0.327		0.223	
s(NP, p)	0.045		0.013		0.062		0.049		0.053	
s(PN, p) / p	0.246		0.263		0.129		0.167		0.273	
s(NP, p) / n	0.190		0.124		0.101		0.130		0.193	

Notes: see notes and abbreviations for table A4.

TABLE A8

LRM wage regressions (dependent variable is hgwln)

	LAW		LBW		LCW		LAM		LBM		LCM	
	Coeff.	**Sig. l.**	**Coeff.**	**Sig. l.**	**Coeff.**	**Sig. l.**	**Coeff.**	**Sig. l.**	**Coeff.**	**Sig. l.**	**Coeff.**	**Sig. l.**
ag_year	0.011	(0.24)	0.017	(0.08)	0.019	(0.11)	0.011	(0.1)	**0.016**	(0.02)	**0.020**	(0)
ag_ysqr	-0.006	(0.59)	-0.011	(0.33)	-0.012	(0.39)	-0.009	(0.28)	-0.012	(0.13)	**-0.017**	(0.05)
ar_dens	**0.070**	(0.02)	**0.088**	(0.01)	**0.116**	(0)	**0.123**	(0)	**0.129**	(0)	**0.140**	(0)
ar_thin	0.003	(0.91)	-0.005	(0.86)	-0.002	(0.24)	-0.025	(0.29)	-0.048	(0.07)	**-0.060**	(0.03)
ms_mard	-0.011	(0.78)	0.014	(0.74)	**-0.047**	(0)	**0.076**	(0.01)	**0.087**	(0)	**0.072**	(0.02)
ms_nmhp	-0.057	(0.47)	-0.082	(0.35)	-0.108	(0.81)	-0.035	(0.6)	-0.029	(0.64)	-0.020	(0.78)
ms_divo	-0.049	(0.36)	-0.031	(0.57)	-0.046	(0.74)	-0.021	(0.67)	-0.016	(0.76)	-0.037	(0.51)
ms_widw	-0.081	(0.19)	-0.046	(0.5)	-0.100	(0.36)	-0.141	(0.18)				
ch_p0002	-0.050	(0.69)	-0.001	(0.99)	-0.011	(0.19)	**0.087**	(0.03)	**0.085**	(0.05)	**0.095**	(0.03)
ch_p0306	0.006	(0.86)	-0.003	(0.94)	0.004	(0.63)	-0.011	(0.69)	-0.009	(0.75)	-0.011	(0.73)
ch_p0715	-0.011	(0.53)	-0.018	(0.38)	-0.017	(0.48)	0.013	(0.43)	0.030	(0.12)	0.029	(0.16)
ed_prnp	**-0.076**	(0.02)	**-0.102**	(0)	-0.254	(0.23)	-0.052	(0.1)	**-0.082**	(0.01)	**-0.170**	(0)
ed_tert	**0.202**	(0)	**0.237**	(0)	**0.524**	(0)	**0.187**	(0)	**0.237**	(0)	**0.451**	(0)
hs_badh	-0.022	(0.66)	-0.028	(0.6)	-0.096	(0.47)	-0.044	(0.46)	-0.036	(0.53)	-0.066	(0.26)
we_yopw	**-0.019**	(0.02)	**-0.027**	(0)			0.002	(0.82)	-0.004	(0.64)		
we_yosq	0.112	(0.06)	**0.154**	(0.01)			-0.044	(0.36)	-0.030	(0.53)		
em_locs	**-0.125**	(0)					**-0.069**	(0.01)				
em_locl	**0.053**	(0.02)					**0.138**	(0)				
em_perj	**0.161**	(0)					**0.087**	(0.02)				
em_mana	**0.201**	(0)					**0.216**	(0)				
oc_21	**0.440**	(0)	**0.481**	(0)			**0.308**	(0)	**0.341**	(0)		
oc_30	**0.255**	(0)	**0.311**	(0)			**0.138**	(0)	**0.180**	(0)		
oc_4	**0.233**	(0)	**0.277**	(0)			**0.085**	(0.04)	**0.115**	(0)		
oc_6	-0.264	(0.09)	-0.225	(0.2)			-0.078	(0.25)	**-0.135**	(0.02)		
oc_7	**-0.171**	(0)	**-0.169**	(0)			0.028	(0.41)	-0.027	(0.38)		
oc_8	0.026	(0.68)	0.022	(0.64)			0.021	(0.58)	0.035	(0.36)		
oc_9	-0.053	(0.32)	-0.016	(0.71)			**-0.124**	(0.01)	**-0.150**	(0)		
in_a	0.088	(0.09)					0.017	(0.74)				
in_f	-0.012	(0.84)					-0.005	(0.88)				
in_gi	0.068	(0.18)					-0.028	(0.37)				
in_h	**0.157**	(0.02)					**0.178**	(0)				
in_jk	**0.208**	(0)					**0.142**	(0)				
in_lmn	0.080	(0.11)					-0.031	(0.44)				
in_opq	0.052	(0.21)					**0.115**	(0)				
in_rstu	**0.112**	(0.04)					-0.074	(0.22)				
_cons	**2.672**	(0)	**2.691**	(0)	**2.761**	(0)	**2.890**	(0)	**2.914**	(0)	**2.879**	(0)
Observations	1,527		1,527		1,527		1,917		1,917		1,917	
F	43.6		46.1		42.4		34.5		42.4		54.4	
R2	0.550		0.492		0.361		0.442		0.361		0.310	
Root MSE	0.317		0.335		0.375		0.333		0.355		0.368	

Notes: specifications are explained in section 5.2. Coefficients significant at the 5% level are marked as bold.

Abbreviations: coeff. – coefficient, sig. l. – significance level.

TABLE A9

Quantile regressions (dependent variable is hgwln)

	Women					Men				
	LAW'	**p10**	**p40**	**p60**	**p90**	**LAM'**	**p10**	**p40**	**p60**	**p90**
ag_year	**0.015**	*0.028*	0.007	0.009	**0.034**	0.008	0.008	*0.010*	0.005	-0.006
ag_ysqr	-0.010	*-0.030*	-0.001	-0.004	**-0.033**	-0.004	-0.007	-0.007	-0.001	0.012
ar_dens	**0.074**	**0.121**	**0.058**	**0.062**	0.055	**0.144**	*0.080*	**0.133**	**0.174**	**0.136**
ar_thin	-0.015	*0.055*	-0.036	-0.022	-0.035	-0.014	-0.026	-0.022	-0.008	0.001
ms_mard	*-0.055*	-0.024	-0.039	-0.054	**-0.155**	**0.057**	0.062	0.007	0.045	**0.104**
ms_nmhp	*-0.147*	-0.181	-0.203	-0.074	**-0.244**	-0.072	-0.051	-0.085	**-0.146**	-0.015
ms_divo	**-0.086**	-0.080	-0.044	-0.046	**-0.220**	-0.038	0.061	-0.081	-0.048	-0.113
ms_widw	-0.048	-0.075	-0.040	0.012	-0.072	*-0.181*	-0.211	-0.134	*-0.213*	**-0.266**
ch_p0002	-0.094	-0.149	-0.033	-0.050	-0.043	0.051	-0.068	0.044	0.080	**0.164**
ch_p0306	0.022	0.011	0.029	0.033	0.029	0.000	-0.033	0.002	0.011	-0.030
ch_p0715	-0.004	-0.023	-0.006	-0.006	0.004	0.018	0.004	0.024	0.017	0.003
ed_prnp	**-0.062**	-0.046	-0.053	**-0.091**	-0.080	-0.047	0.051	-0.017	-0.057	*-0.107*
ed_tert	**0.173**	**0.122**	**0.173**	**0.177**	**0.229**	**0.217**	**0.179**	**0.218**	**0.198**	**0.265**
hs_badh	-0.017	-0.018	0.033	0.009	-0.010	-0.073	-0.148	-0.100	-0.060	0.124
we_yopw	**-0.017**	-0.005	-0.013	**-0.018**	-0.021	-0.001	-0.008	-0.002	-0.007	0.010
we_yosq	0.087	0.062	0.056	0.068	0.093	-0.026	0.030	-0.045	0.014	-0.107
em_locs	**-0.104**	**-0.135**	**-0.098**	**-0.069**	**-0.074**	**-0.089**	**-0.131**	**-0.094**	-0.031	-0.024
em_locl	**0.046**	**0.087**	**0.036**	*0.042*	0.047	**0.117**	**0.078**	**0.122**	**0.137**	**0.132**
em_perj	**0.090**	**0.112**	**0.104**	**0.080**	0.055	**0.099**	**0.172**	**0.161**	**0.097**	0.048
em_mana	**0.197**	**0.172**	**0.164**	**0.183**	**0.308**	**0.199**	**0.201**	**0.220**	**0.212**	**0.216**
oc_21	**0.437**	**0.430**	**0.426**	**0.442**	**0.406**	**0.294**	**0.262**	**0.253**	**0.258**	**0.329**
oc_30	**0.252**	**0.163**	**0.245**	**0.291**	**0.295**	**0.134**	**0.156**	**0.108**	**0.122**	**0.194**
oc_4	**0.235**	**0.185**	**0.222**	**0.241**	**0.248**	0.036	0.037	0.042	0.013	0.010
oc_6	-0.135	-0.330	-0.073	-0.005	0.070	-0.058	-0.101	-0.048	-0.041	*-0.173*
oc_7	**-0.141**	-0.084	**-0.155**	**-0.138**	**-0.149**	0.041	0.048	**0.035**	0.031	0.055
oc_8	0.004	-0.006	-0.043	-0.013	0.034	0.019	0.011	-0.005	-0.012	0.075
oc_9	**-0.077**	-0.091	**-0.095**	-0.039	**-0.133**	**-0.105**	-0.088	-0.100	-0.072	**-0.165**
in_a	0.087	0.107	0.175	0.080	-0.059	0.024	0.064	*0.087*	0.017	0.016
in_f	0.013	**0.232**	0.013	-0.104	**-0.176**	0.004	0.002	0.048	-0.006	-0.014
in_gi	**0.067**	**0.125**	0.063	0.066	0.031	-0.029	0.011	-0.009	*-0.063*	-0.058
in_h	**0.171**	**0.224**	0.231	**0.217**	0.134	**0.176**	**0.158**	**0.203**	**0.171**	**0.215**
in_jk	**0.199**	**0.259**	**0.196**	**0.191**	**0.202**	**0.195**	**0.262**	**0.212**	**0.217**	0.102
in_lmn	*0.081*	**0.179**	0.037	*0.116*	0.135	-0.027	0.058	-0.014	-0.067	-0.018
in_opq	**0.090**	**0.204**	**0.141**	**0.075**	0.051	**0.142**	**0.226**	**0.190**	**0.161**	0.070
in_rstu	**0.121**	**0.260**	**0.146**	0.081	0.044	-0.054	0.048	0.030	*-0.128*	-0.022
_cons	**2.683**	**2.015**	**2.755**	**2.866**	**2.742**	**2.943**	**2.521**	**2.782**	**3.054**	**3.601**

Notes: specifications are explained in section 5.3. Coefficients significant at the 5% (10%) level are marked bold (italic).

Abbreviation: coeff. – coefficient; LAW' and LAM' – LRM estimates based on LAW and LAM, without using the sampling weights; p10, p40, p60 and p90 – estimates at the 10th, 40th, 60th and 90th percentile.

TABLE A10

Heckman selection model – H1 and H2 (dependent variable is hgwln)

	H1W		H1M		H2W		H2M	
Wage equation	**Coeff.**	**Sig. l.**	**Coeff.**	**Sig. l.**	**Coeff.**	**Sig. l.**	**Coeff.**	**Sig. l.**
ag_year	0.010	(0.35)	0.013	(0.07)	**0.022**	(0.03)	**0.015**	(0.03)
ag_ysqr	-0.004	(0.75)	-0.009	(0.29)	-0.019	(0.11)	-0.010	(0.23)
ar_dens	**0.084**	(0.01)	**0.141**	(0)	**0.090**	(0.01)	**0.129**	(0)
ar_thin	0.011	(0.72)	-0.037	(0.16)	-0.012	(0.68)	-0.048	(0.07)
ms_mard	0.000	(0.99)	**0.069**	(0.02)	0.021	(0.63)	**0.083**	(0)
ms_nmhp	-0.078	(0.37)	-0.027	(0.67)	-0.141	(0.14)	-0.018	(0.78)
ms_divo	-0.029	(0.59)	-0.010	(0.85)	-0.007	(0.9)	-0.015	(0.76)
ms_widw	-0.082	(0.22)			-0.028	(0.67)		
ch_p0002	0.027	(0.83)	0.080	(0.07)	-0.040	(0.74)	**0.084**	(0.05)
ch_p0306	0.016	(0.69)	-0.011	(0.7)	-0.042	(0.3)	-0.007	(0.8)
ch_p0715	-0.014	(0.52)	0.028	(0.15)	-0.029	(0.16)	0.029	(0.12)
ed_prnp	**-0.089**	(0.01)	**-0.076**	(0.02)	**-0.124**	(0)	**-0.079**	(0.01)
ed_tert	**0.250**	(0)	**0.226**	(0)	**0.254**	(0)	**0.230**	(0)
hs_badh	-0.008	(0.88)	-0.004	(0.95)	-0.102	(0.07)	-0.009	(0.88)
we_yopw	-0.018	(0.06)	0.008	(0.37)	**-0.038**	(0)	0.000	(0.97)
we_yosq	**0.144**	(0.02)	-0.057	(0.28)	**0.108**	(0.04)	-0.036	(0.47)
oc_21	**0.437**	(0)	**0.329**	(0)	**0.490**	(0)	**0.345**	(0)
oc_30	**0.298**	(0)	**0.176**	(0)	**0.312**	(0)	**0.182**	(0)
oc_4	**0.259**	(0)	**0.115**	(0)	**0.281**	(0)	**0.116**	(0)
oc_6	-0.245	(0.17)	**-0.141**	(0.01)	-0.285	(0.09)	**-0.142**	(0.01)
oc_7	**-0.164**	(0)	-0.018	(0.57)	**-0.205**	(0)	-0.026	(0.39)
oc_8	0.017	(0.71)	0.040	(0.29)	0.012	(0.81)	0.039	(0.3)
oc_9	-0.021	(0.64)	**-0.113**	(0.02)	-0.003	(0.96)	**-0.152**	(0)
_cons	**2.883**	(0)	**3.044**	(0)	**2.595**	(0)	**2.946**	(0)
Participation eq.	**Coeff.**	**Sig. l.**	**Coeff.**	**Sig. l.**	**Coeff.**	**Sig. l.**	**Coeff.**	**Sig. l.**
ag_year	**0.120**	(0)	**0.060**	(0.01)	0.092	(0.09)	**0.146**	(0)
ag_ysqr	**-0.132**	(0)	**-0.057**	(0.04)	**-0.150**	(0.02)	**-0.204**	(0)
ar_dens	0.140	(0.25)	**-0.234**	(0.03)	0.165	(0.31)	0.102	(0.61)
ar_thin	**-0.257**	(0.01)	-0.173	(0.06)	-0.110	(0.4)	0.051	(0.74)
ms_mard	0.208	(0.18)	**0.360**	(0)	0.260	(0.3)	**0.370**	(0.04)
ms_nmhp	-0.134	(0.63)	-0.013	(0.96)	**-0.994**	(0.04)	-0.700	(0.05)
ms_divo	0.043	(0.84)	-0.160	(0.41)	**0.688**	(0.04)	0.251	(0.46)
ms_widw	**1.064**	(0)			0.411	(0.22)		
ch_p0002	-0.217	(0.42)	**0.379**	(0.02)	-0.059	(0.88)	0.601	(0.13)
ch_p0306	-0.234	(0.05)	0.143	(0.19)	**-0.776**	(0)	-0.020	(0.93)
ch_p0715	0.059	(0.42)	**0.174**	(0.02)	**-0.187**	(0.02)	0.236	(0.09)
ch_o0015	0.133	(0.21)	0.135	(0.17)	0.055	(0.68)	0.183	(0.21)
ed_prnp	-0.225	(0.09)	-0.068	(0.54)	**-0.334**	(0.03)	-0.233	(0.12)
ed_tert	**-0.420**	(0.01)	0.233	(0.15)	**0.939**	(0)	**1.456**	(0)
hs_badh	**-0.451**	(0.01)	**-0.428**	(0.02)	**-0.946**	(0)	**-1.181**	(0)
we_yopw	**-0.163**	(0)	**-0.195**	(0)	**-0.214**	(0)	**-0.237**	(0)
we_yosq	**0.359**	(0.01)	**0.485**	(0)	**0.429**	(0)	**0.581**	(0)
oc_21	**0.941**	(0)	0.329	(0.11)	0.521	(0.09)	-0.658	(0.1)
oc_30	0.198	(0.23)	0.081	(0.55)	0.047	(0.84)	-0.237	(0.29)
oc_4	**0.268**	(0.04)	0.012	(0.94)	0.069	(0.65)	-0.156	(0.53)
oc_6	0.277	(0.48)	0.208	(0.34)	-0.575	(0.16)	**1.028**	(0.02)
oc_7	0.006	(0.97)	-0.117	(0.28)	**-0.498**	(0)	-0.149	(0.45)

	H1W		H1M		H2W		H2M	
Participation eq.	**Coeff.**	**Sig. l.**	**Coeff.**	**Sig. l.**	**Coeff.**	**Sig. l.**	**Coeff.**	**Sig. l.**
oc_8	0.085	(0.61)	-0.024	(0.85)	-0.146	(0.4)	-0.335	(0.13)
oc_9	0.034	(0.81)	**-0.551**	(0)	0.149	(0.48)	0.172	(0.51)
oi_a	0.004	(0.67)	0.005	(0.51)	-0.014	(0.28)	0.021	(0.06)
oi_b	0.024	(0.21)	**0.033**	(0.03)	-0.036	(0.13)	0.002	(0.93)
oi_c	-0.023	(0.14)	**-0.078**	(0)	-0.010	(0.66)	-0.011	(0.66)
oi_d	-0.005	(0.57)	-0.006	(0.44)	0.018	(0.09)	0.014	(0.27)
oi_e	-0.008	(0.66)	0.010	(0.5)	0.018	(0.41)	0.031	(0.32)
oi_f	**-0.073**	(0)	**-0.058**	(0)	**-0.045**	(0.01)	**-0.062**	(0.01)
oi_g	0.026	(0.25)	-0.032	(0.12)	-0.001	(0.98)	**-0.110**	(0.01)
em_agri	-0.355	(0.22)	**-0.715**	(0.03)	0.424	(0.23)	**-1.398**	(0)
_cons	**-1.850**	(0.01)	-0.207	(0.68)	0.655	(0.52)	0.502	(0.53)
/lnsigma	**-1.075**	(0)	**-1.017**	(0)	**-1.059**	(0)	**-1.039**	(0)
/athrho	**-0.415**	(0.03)	**-0.393**	(0.01)	**0.799**	(0)	-0.242	(0.15)
sigma	**0.341**	(0)	**0.362**		**0.347**		**0.354**	
rho	**-0.393**		**-0.374**		**0.664**		-0.237	
lambda	**-0.134**		**-0.135**		**0.230**		-0.084	
Total obs.	2,034		2,552		1,880		2,073	
"Negative" obs.	507		635		353		156	

Notes: see notes for table A11.

TABLE A11

Heckman selection model – H3 and H4 (dependent variable is hgwln)

	H3W		H3M		H4W		H4M	
Wage eq.	**Coeff.**	**Sig. l.**	**Coeff.**	**Sig. l.**	**Coeff.**	**Sig. l.**	**Coeff.**	**Sig. l.**
ag_year	-0.004	(0.73)	0.008	(0.33)	0.007	(0.54)	0.014	(0.07)
ag_ysqr	0.014	(0.32)	-0.004	(0.7)	0.004	(0.75)	-0.009	(0.28)
ar_dens	**0.121**	(0)	**0.149**	(0)	**0.119**	(0)	**0.135**	(0)
ar_thin	0.017	(0.6)	**-0.055**	(0.04)	0.027	(0.39)	**-0.061**	(0.03)
ms_mard	-0.059	(0.21)	0.058	(0.06)	-0.045	(0.34)	0.057	(0.07)
ms_nmhp	-0.093	(0.35)	-0.034	(0.62)	-0.065	(0.52)	-0.023	(0.74)
ms_divo	-0.044	(0.49)	-0.035	(0.55)	-0.061	(0.34)	-0.029	(0.62)
ms_widw	-0.109	(0.12)			-0.122	(0.08)		
ch_p0002	-0.030	(0.85)	0.081	(0.08)	-0.001	(1)	**0.089**	(0.05)
ch_p0306	0.002	(0.97)	-0.012	(0.71)	0.022	(0.57)	-0.007	(0.83)
ch_p0715	-0.009	(0.7)	0.032	(0.13)	-0.006	(0.8)	0.033	(0.11)
ed_prnp	**-0.220**	(0)	**-0.157**	(0)	**-0.116**	(0)	**-0.152**	(0)
ed_tert	**0.524**	(0)	**0.453**	(0)	**0.500**	(0)	**0.444**	(0)
hs_badh	-0.074	(0.21)	-0.063	(0.28)	-0.055	(0.36)	-0.015	(0.81)
_cons	**3.281**	(0)	**3.180**	(0)	**2.995**	(0)	**3.044**	(0)
Participation eq.	**Coeff.**	**Sig. l.**	**Coeff.**	**Sig. l.**	**Coeff.**	**Sig. l.**	**Coeff.**	**Sig. l.**
ag_year	**0.347**	(0)	**0.269**	(0)	**0.192**	(0)	**0.172**	(0)
ag_ysqr	**-0.374**	(0)	**-0.270**	(0)	**-0.245**	(0)	**-0.183**	(0)
ar_dens	-0.062	(0.75)	**-0.381**	(0.03)	-0.031	(0.84)	0.368	(0.12)
ar_thin	**-0.428**	(0.01)	-0.148	(0.33)	**-0.496**	(0)	0.094	(0.6)
ms_mard	0.357	(0.1)	**0.541**	(0.02)	-0.251	(0.21)	**1.017**	(0)
ms_nmhp	-0.091	(0.83)	0.249	(0.38)	**-0.797**	(0.01)	-0.127	(0.64)
ms_divo	-0.070	(0.83)	-0.223	(0.58)	0.486	(0.16)	0.012	(0.98)
ms_widw	0.565	(0.25)			0.350	(0.24)		
ch_p0002	0.613	(0.28)	**0.589**	(0.04)	0.012	(0.97)	0.410	(0.3)
ch_p0306	-0.123	(0.51)	-0.088	(0.57)	-0.169	(0.23)	**-0.433**	(0.03)
ch_p0715	-0.097	(0.4)	0.046	(0.71)	-0.037	(0.66)	-0.284	(0.07)
ch_o0015	0.178	(0.14)	0.120	(0.32)	0.030	(0.76)	**-0.289**	(0.03)
ed_prnp	**-0.977**	(0)	**-0.515**	(0)	**-1.367**	(0)	**-0.670**	(0)
ed_tert	**-0.472**	(0)	-0.235	(0.12)	**0.601**	(0.01)	0.482	(0.1)
hs_badh	**-0.664**	(0.01)	-0.282	(0.55)	**-0.520**	(0)	**-1.291**	(0)
oi_a	0.003	(0.81)	0.016	(0.16)	0.009	(0.4)	0.007	(0.74)
oi_b	0.051	(0.06)	0.007	(0.71)	**0.055**	(0.04)	0.055	(0.2)
oi_c	0.001	(0.97)	-0.015	(0.61)	-0.034	(0.07)	0.033	(0.28)
oi_d	0.010	(0.41)	**-0.028**	(0.01)	-0.002	(0.83)	-0.022	(0.2)
oi_e	0.022	(0.41)	-0.007	(0.75)	0.008	(0.65)	0.031	(0.34)
oi_f	**-0.077**	(0)	**-0.056**	(0.01)	**-0.061**	(0)	-0.020	(0.48)
oi_g	0.015	(0.63)	-0.011	(0.69)	-0.036	(0.11)	-0.070	(0.22)
em_agri	-0.222	(0.53)	**-0.748**	(0.04)	**-0.756**	(0)		
_cons	**-5.533**	(0)	**-4.025**	(0)	-1.179	(0.19)	-1.251	(0.3)
/lnsigma	**-0.952**	(0)	**-0.988**	(0)	**-0.947**	(0)	**-0.990**	(0)
/athrho	**-0.816**	(0)	**-0.506**	(0)	**-0.718**	(0)	**-0.658**	(0)
sigma	**0.386**		**0.372**		**0.388**		**0.372**	
rho	**-0.673**		**-0.467**		**-0.616**		**-0.577**	
lambda	**-0.260**		**-0.174**		**-0.239**		**-0.214**	
Total obs.	1,699		2,109		1,883		1,979	
"Negative" obs.	172		192		356		62	

Notes: specifications are explained in section 5.4 (table 10). Coefficients significant at the 5% level are marked as bold.

Abbreviations: coeff. – coefficient, sig. l. – significance level.

APPENDIX 3: OTHER FIGURES

FIGURE A1

Formation of subsamples

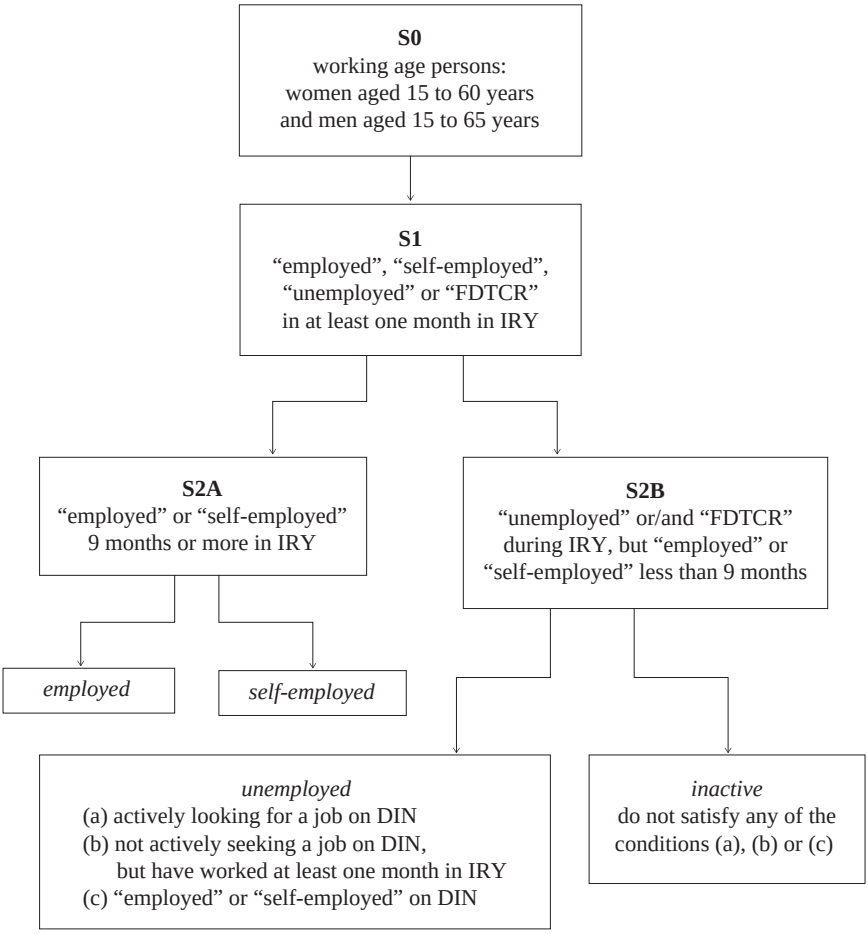
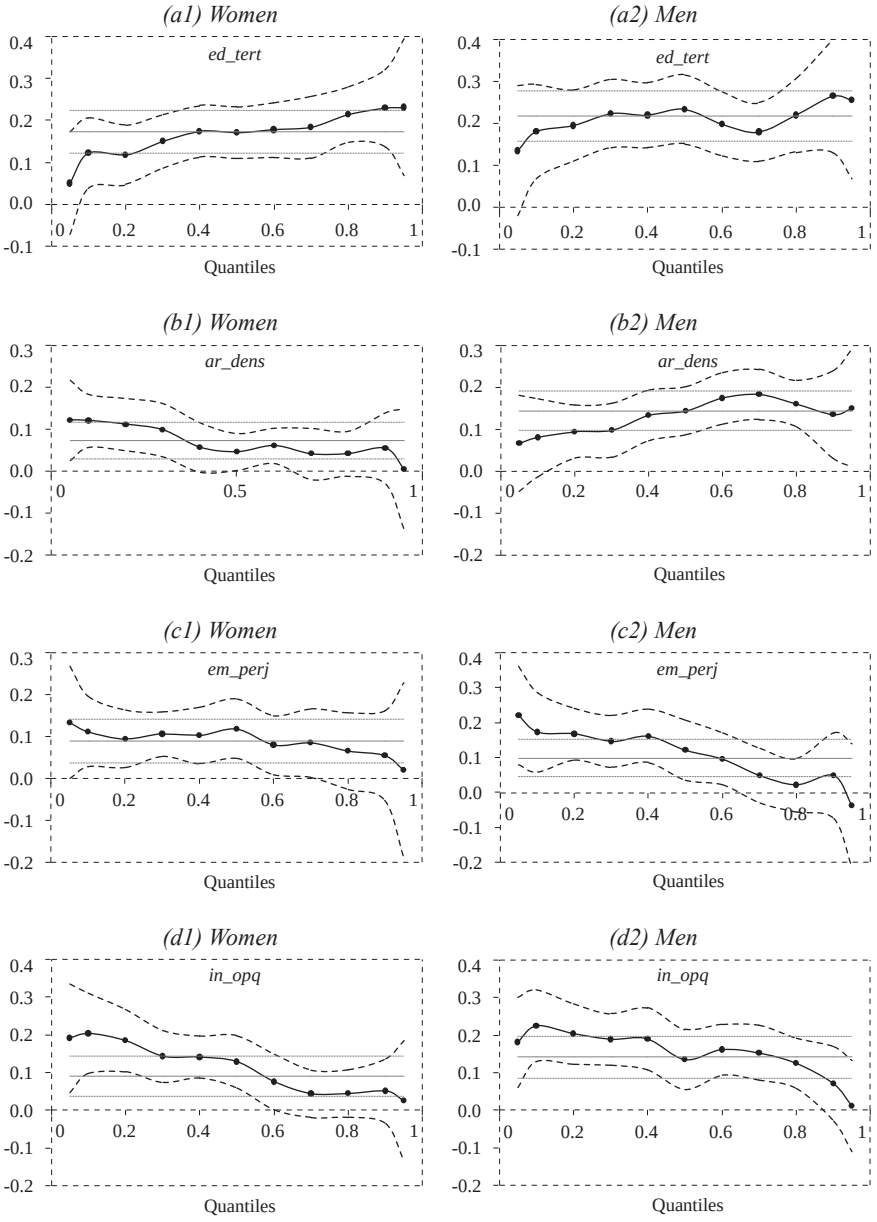


FIGURE A2

Coefficients from QRM and LRM regressions, with confidence intervals



REFERENCES

1. Amemiya, T., 1985. *Advanced econometrics*. Cambridge Mass.: Harvard University Press.
2. Anon, 2011. *What are pseudo R-squareds?* [online]. UCLA: Statistical Consulting Group. Available at: <http://www.ats.ucla.edu/stat/mult_pkg/faq/general/Psuedo_RSquareds.htm>.
3. Avlijaš, S. [et al.], 2013. *Gender Pay Gap in the Western Balkan Countries: Evidence From Serbia, Montenegro and Macedonia*. FREN – Foundation for the Advancement of Economics.
4. Bejaković, P [et al.], 2012. Isplati li se raditi u Hrvatskoj? *Revija za socijalnu politiku*, 19(1), pp. 83-92. doi: 10.3935/rsp.v19i1.1054
5. Berger, F., Islam, N. and Liégeois, P., 2011. Behavioural microsimulation and female labour supply in Luxembourg. *Brussels Economic Review*, 54(4), pp. 389-420.
6. Bičáková, A., Slačálek, J. and Slavík, M., 2011. Labor Supply after Transition: Evidence from the Czech Republic. *Czech Journal of Economics and Finance*, 61(4), pp. 327-347.
7. Botrić, V., 2009. Unemployed and long-term unemployed in Croatia: Evidence from Labour Force Survey. *Revija za socijalnu politiku*, 16(1), pp. 25-44. doi: 10.3935/rsp.v16i1.807
8. Breunig, R. and Mercante, J., 2010. The Accuracy of Predicted Wages of the Non-Employed and Implications for Policy Simulations from Structural Labour Supply Models. *Economic Record*, 86(272), pp. 49-70. doi: 10.1111/j.1475-4932.2009.00619.x
9. Bushway, S., Johnson, B. D. and Slocum, L. A., 2007. Is the Magic Still There? The Use of the Heckman Two-Step Correction for Selection Bias in Criminology. *Journal of Quantitative Criminology*, 23(2), pp. 151-178. doi: 10.1007/s10940-007-9024-4
10. Cameron, A. C. and Trivedi, P. K., 2005. *Microeconometrics: methods and applications*. Cambridge; New York: Cambridge University Press. doi: 10.1017/CBO9780511811241
11. Carone, G. [et al.], 2004. *Indicators of Unemployment and Low-Wage Traps: Marginal Effective Tax Rates on Employment Incomes*. Paris: OECD. doi: 10.1787/137550327778
12. CBS, 2013a. *Anketa o dohotku stanovništva* [online]. Available at: <http://www.dzs.hr/Hrv/important/Obrasci/14-Potrosnja/Obrasci/ADS-1.pdf>.
13. CBS, 2013b. *Poverty indicators, 2011* [online]. Available at: <http://www.dzs.hr/Hrv_Eng/publication/2012/14-01-03_01_2012.htm>.
14. CBS, 2015. *Active population (labour force) in Republic of Croatia, according to results of Labour Force Survey* [online]. Available at: <http://www.dzs.hr/Hrv/publication/StatisticsInLine.htm>.
15. Eurostat, 2015. *EU-SILC: List of variables* [online]. Available at: <http://ec.europa.eu/eurostat/web/income-and-living-conditions/methodology/list-variables>.
16. Figari, F., Paulus, A. and Sutherland, H., 2014. Microsimulation and Policy Analysis. In: A.B. Atkinson and F. Bourguignon, eds., *Handbook of Income Distribution*. Elsevier.

17. Greene, W. H., 2008. *Econometric analysis*. Upper Saddle River, N. J.: Prentice Hall.
18. Heckman, J. J., 1976. The Common Structure of Statistical Models of Truncation, Sample Selection and Limited Dependent Variables and a Simple Estimator for Such Models. *Annals of Economic and Social Measurement*, 5(4), pp. 475-492.
19. Heckman, J. J., 1979. Sample Selection Bias as a Specification Error. *Econometrica*, 47(1), pp. 153-161. doi: 10.2307/1912352
20. Immervoll, H. and O'Donoghue, C., 2002. *Welfare benefits and work incentives: an analysis of the distribution of net replacement rates in Europe using EUROMOD, a multi-country microsimulation model. EUROMOD Working Paper No. EM4/01.*
21. Khitarishvili, T., 2009. *Explaining the Gender Wage Gap in Georgia*. *The Levy Economics Institute Working Paper,* No. 577.
22. Labeaga, J. M., Oliver, X. and Spadaro, A., 2008. Discrete choice models of labour supply, behavioural microsimulation and the Spanish tax reforms. *The Journal of Economic Inequality*, 6(3), pp. 247-273. doi: 10.1007/s10888-007-9057-9
23. Mojsoska-Blazevski, N., Petreski, M. and Petreska, D., 2013. *Increasing labour market activity of poor and female: Let's make work pay in Macedonia*. MPRA Working Papers Nr. 57228. University Library of Munich, Germany.
24. Nestić, D., 2005. The Determinants of Wages in Croatia: Evidence from Earnings Regressions. In: *Proceedings of 65th Anniversary Conference of the Institute of economics*. Zagreb: The Institute of Economics, pp. 131-162.
25. Nestić, D., Rubil, I. and Tomić, I., 2015. Analysis of the Difference in Wages between the Public Sector, State-Owned Enterprises and the Private Sector in Croatia in the Period 2000-2012. *Privredna kretanja i ekonomska politika*, 24(1), pp. 7-51.
26. Nicaise, I., 2001. Human capital, reservation wages and job competition: Heckman's lambda re-interpreted. *Applied Economics*, 33(february), pp. 309-315. doi: 10.1080/00036840121810
27. Paci, P. and Reilly, B., 2004. *Does economic liberalization reduce gender inequality in the labor market: The experience of the transition economies of Europe and Central Asia*. World Bank: Working paper.
28. Pacifico, D., 2009. A behavioral microsimulation model with discrete labour supply for Italian couples. *CAPPaper,* No. 65.
29. Pastore, F. and Verashchagina, A., 2008. *The Determinants of Female Labour Supply in Belarus. IZA Working Paper*, No. 3457.
30. Puhani, P., 2000. The Heckman Correction for Sample Selection and Its Critique. *Journal of Economic Surveys*, 14(1), pp. 53-68. doi: 10.1111/1467-6419.00104
31. Schaffner, J. A., 1998. Generating conditional expectations from models with selectivity bias: comment. *Economics Letters*, 58, pp. 255-261. doi: 10.1016/S0165-1765(98)00004-4

32. van Soest, A., 1995. Structural Models of Family Labor Supply: A Discrete Choice Approach. *The Journal of Human Resources*, 30(1), pp. 63-88. doi: 10.2307/146191
33. Urban, I. and Bezeredi, S., 2015. *EUROMOD Country Report: Croatia*. Institute for Social and Economic Research.
34. Veall, M. R. and Zimmermann, K. F., 1992. Pseudo R2's in the ordinal probit model. *The Journal of Mathematical Sociology*, 16(4), pp. 333-342. doi: 10.1080/0022250X.1992.9990094
35. Vella, F., 1998. Estimating Models with Sample Selection Bias: A Survey. *The Journal of Human Resources*, 33(1), pp. 127-169. doi: 10.2307/146317
36. Verbeek, M., 2004. *A guide to modern econometrics*. Chichester: Wiley.
37. Williams, R., 2015. *Scalar Measures of Fit: Pseudo R2 and Information Measures (AIC & BIC)* [online]. Available at: <https://www3.nd.edu/~rwilliam/xsoc73994/L05.pdf>.
38. Winship, R. and Mare, D., 1992. Models for Sample Selection Bias. *Annual Review of Sociology*, pp. 327-350. doi: 10.1146/annurev.so.18.080192.001551