CAPP

Estimation and Simulation of Earnings in IT-SILC

Emanuele Ciani, Marcello Morciano

CAPP*aper* n. 90

settembre 2011

Università di Modena e Reggio Emilia Facoltà di Economia Marco Biagi

Università di Bologna Dipartimento di Scienze Economiche

CAPP - Centro di Analisi delle Politiche Pubbliche
Dipartimento di Economia Politica - Università di Modena e Reggio Emilia
Ufficio 54 - Ala Ovest
Viale Berengario, 51 41100 Modena - ITALY
phone: +39 059 2056854 fax: +39 059 2056947
email capp@unimo.it

EUROPEAN COMMISSION
Employment, Social Affairs and Equal Opportunities

Social Protection and Integration
Social Protection, Social Services

COMMUNITY PROGRAMME FOR EMPLOYMENT AND SOCIAL SOLIDARITY (PROGRESS)

Action related to the development of administrative datasets and models for labour market and pensions analysis

Project ASSESSING ADEQUACY AND LONG TERM DISTRIBUTIVE EFFECTS OF THE ITALIAN PENSION SYSTEM.
A MICROSIMULATION APPROACH VS/2009/0498

This publication is supported under the European Community Programme for Employment and Social Solidarity – PROGRESS (2007-2013).

This programme is managed by the Directorate-General for Employment, social affairs and equal opportunities of the European Commission. It was established to financially support the implementation of the objectives of the European Union in the employment and social affairs area, as set out in the Social Agenda, and thereby contribute to the achievement of the Lisbon Strategy goals in these fields.

The seven years Programme targets all stakeholders who can help shape the development of appropriate and effective employment and social legislation and policies across the EU-27, EFTA- EEA and EU candidate and pre-candidate countries.

PROGRESS mission is to strengthen the EU contribution in support of Member States commitment. PROGRESS will be instrumental in:

- *providing analysis and policy advice on PROGRESS policy areas;*
- *monitoring and reporting on the implementation of EU legislation and policies in PROGRESS policy areas;*
- *promoting policy transfer, learning and support among Member States on EU objectives and priorities; and*
- *relaying the views of the stakeholders and society at large*

For more information see: http://ec.europa.eu/progress

Estimation and Simulation of Earnings in IT-SILC

Emanuele Ciani and Marcello Morciano
This version 07/09/2011

ABSTRACT: *This paper describes income distribution among workers in Italy using both the cross-sectional and panel component of IT-SILC. We highlight advantages and drawbacks of different econometric approaches, comparing standard OLS estimates with those obtained from Random Effects and Poisson Maximum Likelihood and assessing whether the results are sensitive to the different specification. Finally, we present the procedure in use in simulating future earnings in CAPP_DYN, the dynamic population-based microsimulation model of the CAPP.*

JEL codes: D3, J3.

Sommario

1. Introduction

One of the key components of a Dynamic Microsimulation Model (DMM) which focuses on the projection and analysis of the reformed Italian old-age pension system is the prediction of the individual labour market participation and earnings. This chapter focuses on the latter, while the module which simulates the individual position in regards to the labour market was already discussed in Flisi and Morciano (2011).

Ideally, we would like to use panel-data in order to estimate a model for mean earnings, conditional on a set of observable time-variant and time-invariant individual characteristics, and modelling the autoregressive component of the residuals. Then, by using the estimated parameters we could predict the evolution of an individual's earnings in future years, taking into account the likely evolution of her/his observable characteristics and making assumptions on the evolution of unobserved individual effect and how the expected increases in productivity in each of the simulated periods would be distributed among workers. In reality, limited long and up-to-date longitudinal data are available for Italy, which creates difficulties in estimating a satisfying and credible model for the error component of a wage-equation.[1] As a consequence, and in common with other DMMs, we focus here on cross-section estimates, but we also discuss how to make the earning module flexible enough to use information on the autocorrelation of earnings across time coming from other data studies/sources.

In this chapter, after describing income distribution among Italian workers, we report estimates of parsimonious models of earnings, aiming to be suitable for predictions in the DMM. We propose different econometric methods, applied on both the cross-sectional and longitudinal component of IT-SILC, paying particular attention to the assumptions regarding the unobservable heterogeneity. We also compare standard OLS estimates with those obtained from Random Effects and Poisson Maximum Likelihood, in order to assess whether the results are sensitive to the different specification.[2] To our knowledge, this is the first empirical analysis on earnings distribution using the Italian component of SILC. IT-SILC data on income have gone through a process of integration with the administrative archive, based on a one-to-one matching (see Ciani et al., 2011). This allow us to exploit the advantage of using (Italian) administrative archives which, nevertheless, have more precise information on income and— at the same time— to make use of a set of detailed socio-economic information (collected in IT-SILC) without employing statistical matching techniques.

[1] Moreover, it should be pointed out that in some cases, such as the Belgian MIDAS model (Dekkers, Desmet, & De Vil, 2010), the random effects panel data regression estimates resulted in very poor DMM simulation results, even though the regression results were by themselves credible.

[2] Estimates of a Poisson Maximum Likelihood model are given in the appendix whereas estimates of quintile regressions are available upon request.

The chapter is organized as follows. Section 2 provides a non-exhaustive synthesis of previous literature, highlighting advantages and drawbacks of the different approaches. Section 3 describes earnings distribution in Italy using IT-SILC data. Section 4 discusses OLS estimates, while Section 5 is focussed on longitudinal Random Effects results. Section 5 describes the method used in CAPP_DYN to simulate earnings in future years. The final section concludes.

2. Background

From the seminal contributions of Mincer (1974), a multitude of econometric models has been proposed to estimate earnings and the determinant of earnings.[3] While one stream of the literature aims at estimating the returns on schooling, here we are more interested in those studies focussed on forecasting earnings in a population-based DMM, such as Pudney (1992) and Bækgaard, King, & Robinson (1999), and on modelling the error component in longitudinal estimates, like Lillard & Willis (1978), Borella (2004) and Ramos (2003).

Generally, DMM are based on estimates of a model for log-earnings. Some DMM, such as the French DESTINIE, MIDAS (Dekkers, Desmet, & De Vil, 2010), the model for the Italian Tuscany region MIRTODIN (Maitino & Sciclone, 2009) and so on, make projections of individual earnings using essentially estimates obtained from cross-sectional data. Basically, the logarithm of annual, monthly or weekly earnings is regressed on a set of personal characteristics regarding, in particular, education, employment history and activity. In order to allow more flexibility, regressions are carried out separately on different groups of the population of workers. Estimates are then used to predict earnings in two directions. First, these DMMs need to build the earnings history for each individual up to the year of the survey, as the cross-sections clearly do not include a full record of previous employment. [4] Secondly, estimates are used to forecast earnings for future years, both for those who will still be working and for the individuals who will start a new job.

The main problem in using cross-sectional data is that differences across age groups, if observed at a single point in time, cannot be interpreted as cohort or time effects. Essentially, all

[3] Among others, see Spence (1973), Becker (1994), Weiss (1995).

[4] Harding (2007) wrote "*Where longitudinal earnings histories are not available, modellers have faced enormous challenges in attempting to 'back-cast' to simulate earnings (and other characteristics) earlier in life.".* According to O'Donoghue (2001), base data used by DMM can be divided into historical and current data. A number of models (CORSIM, DYNAMOD and DYNACAN) use historic data such as (a random sample of) census files from the 1960's. These models start their simulation at a point in the past, building up a sufficiently long work-history to the present day. The reason for this is that in order to simulate pensions, one needs information about work-histories since the year of entry in the labour market. Some models, such as MOSART or PENSIM, have base data sets that include work histories, so the early start date is not necessary. Other models use data fusion techniques (i.e. statistical matching procedures) for matching base year with other data sources containing more detailed retrospective information. Finally, other models (DESTINIE and CAPP_DYN) simulate both forwards as other models do, but also backwards to create work histories. Although statistical matching provides greater flexibility (Cohen, 1991) it might raise problems in the Conditional Independence Assumption (Ridder & Moffit, 2006). Instead, the backward simulation has the advantage of simplicity, albeit retrospective information is constructed using simplified assumptions.

DMM models face major methodological problems in attempting to disentangle age, cohort and period effects. As pointed out by Lillard & Willis (1978), the shape of a cross-sectional age-earnings distribution changes over time not only through the cohort and period effect. An important role is also played by changes in occupational composition and by changes in labour demand.

The use of longitudinal data can enable age, cohort and time effects to be disentangled. A number of DMM models, such as the British PenSim2 (Emmerson, Reed, & Shephard, 2004), the U.S CORSIM (Favreault & Caldweel, 1998), the Australian Dynamod (2002) and the Italian CeRPSIM (Borella & Coda Moscarola, (2006), (2009)), use panel data for the estimation of earnings. Another important advantage in using a longitudinal sample is the possibility to model the autocorrelation in the error component, allowing for more precise predictions about future wages. Lillard and Willis (1978), as well as Ramos (2003) and Borella (2004), discuss different models for the log-earnings residuals.

Nevertheless, the use of longitudinal estimates for prediction in a DMM might have some drawbacks. As pointed out in Harding (1990), if one uses coefficients estimated on panel data, then the cohort effects do not provide us with any information about future cohorts. For the MIDAS model, Dekkers, Desmet and Greet De Vil (2010, p. 34) pointed out that random- effects models resulted in very poor DMM simulations results.

In our case, the main problem is that the robustness of longitudinal estimates depends on the use of long panel data. In Italy, the longitudinal component of IT-SILC is still relatively short to provide reliable estimates of earnings autocorrelation. This is also due to the fact that Istat started to collect information on gross-earnings only as from 2007, so that previous waves cannot be employed for our estimates. Moreover, as discussed in section 5, the panel component does not contain all variables needed for the simulation and cannot be directly matched to single cross-sections.[5]

The longitudinal sample of the other most used survey with data on income, SHIW, is relatively long. However, results might be affected by its small size and by the absence of gross data. Differently, Borella & Coda Moscarola (2006) used administrative panel data from the National Social Security Institute (INPS), which cover a long time span (1985-1998). The earnings variables in this dataset are those which are currently used to calculate pensions, a clear advantage for the simulations. However, this administrative archive does not include individuals who worked in the public sector, which is a non-negligible fraction of employees. Moreover, focusing on the earnings distribution in the period before 2000, it may misrepresent the current situation (see

[5] Istat does not provide an identifier that allows this match, in order to comply with the Silc rules agreed at the European level, which were devised to protect respondents' privacy.

Harding's critique). Lastly, INPS administrative archive does not include any information on workers' education, which is an important variable in predicting the earnings distribution in future years, as we expect significant changes in the proportion of graduates (see Mazzaferro and Morciano, 2011).

As discussed, the IT-SILC panel is still not suitable for our purpose. Nonetheless, we decided to base the prediction of earnings on estimates carried out on its cross-sectional component. First of all, if we used a different dataset, such as the INPS administrative archive, we would need to impute earnings for all individuals, based on a limited set of characteristics that are observable in both samples. Differently, using estimates carried out on IT-SILC, we avoid the problems due to statistical matching. On the other hand, by employing regression coefficients estimated on the same sample we try to avoid the problems related to non-homogeneity between covariates in different datasets. Secondly, the method for predicting earnings in future years, described in section 6 and built on Pudney (1992), might be modified to use information on the autocorrelation of residuals estimated in other studies. Lastly, the SILC longitudinal component is growing at a fast rate, and we cannot neglect the benefit of setting up a DMM that can be very easily adapted to use it.

3. Descriptive analysis of gross earnings

This section discusses descriptive statistics on gross earnings in the sample corresponding to the initial population, drawn from IT-SILC 2006 as described in Ciani et al. (2011). We select only employed individuals, aged more than 20. The sample is also restricted to those below the legal State Pension Age (SPA) in force in 2006, in order to avoid the possible bias due to the substantial self-selection of individuals who choose to continue working after the SPA. Where not differently stated, all graphs and tables reported in this chapter refer to this sample, composed of 19,720 observations.[6]

Gross earnings include both employee cash income and cash benefits or losses from self-employment, plus the social contribution paid by the workers.[7] The measure we are using is the earning definition used by the Italian pension system in computing expected pension earnings.

Table 1 reports sample statistics on the annual gross earnings, in 2006 euro. Self-employed have, on average, higher earnings, but they display more dispersion. There are also few negative

[6] For reasons discussed in Ciani et al. (2011), we do not use sample weights. However, the use of sample weights makes little differences in the results reported in this chapter.

[7] The yearly gross earning is the sum of It-Silc variables py010g ("Employee cash or near cash income") and py050g ("Cash benefits or losses from self-employment"), both including the social contribution paid by the worker. For a full definition of both variables, we refer to the document "description of Silc user database variables, Version 2007.1 from 01-03-09".

values (44 observations), that we reclassify to be equal to one euro.[8] Among employees, annual gross earnings are higher in the public sector.[9] Atypical workers' earnings are similar to those of employees, even if they generally show smaller medians. The mean–median difference of the earnings distribution is higher for self-employed. The positive difference between mean and median earnings indicates that earnings are negatively skewed, in particular for self-employed. We therefore focus on the median value for the following distributive analysis.

Table 1 Sample statistics on annual gross earnings, by status in employment, euro 2006

Status in employment	Mean	Median	Minimum	Maximum	Standard Deviation	Observations
Employee, public sector, full-time	31,664	28,293	650	237,021	18,701	3,867
Employee, public sector, part-time	15,880	14,098	285	111,412	10,530	296
Employee, private sector, full-time	24,894	21,131	431	477,897	19,013	9,384
Employee, private sector, part-time	11,169	9,784	374	100,725	7,416	1,454
Self-employed, full-time	32,329	22,744	-60,000	641,600	37,167	3,943
Self-employed, part-time	20,510	12,300	-30,000	268,763	25,487	337
Atypical, public sector, full-time	23,917	21,772	3,266	87,569	14,931	67
Atypical, public sector, part-time	14,180	13,828	1,273	28,554	7,702	18
Atypical, private sector, full-time	23,465	17,652	510	415,160	35,299	234
Atypical, private sector, part-time	11,311	9,456	1,293	31,799	7,588	63
Total	26,417	21,753	-60,000	641,600	24,182	19,663

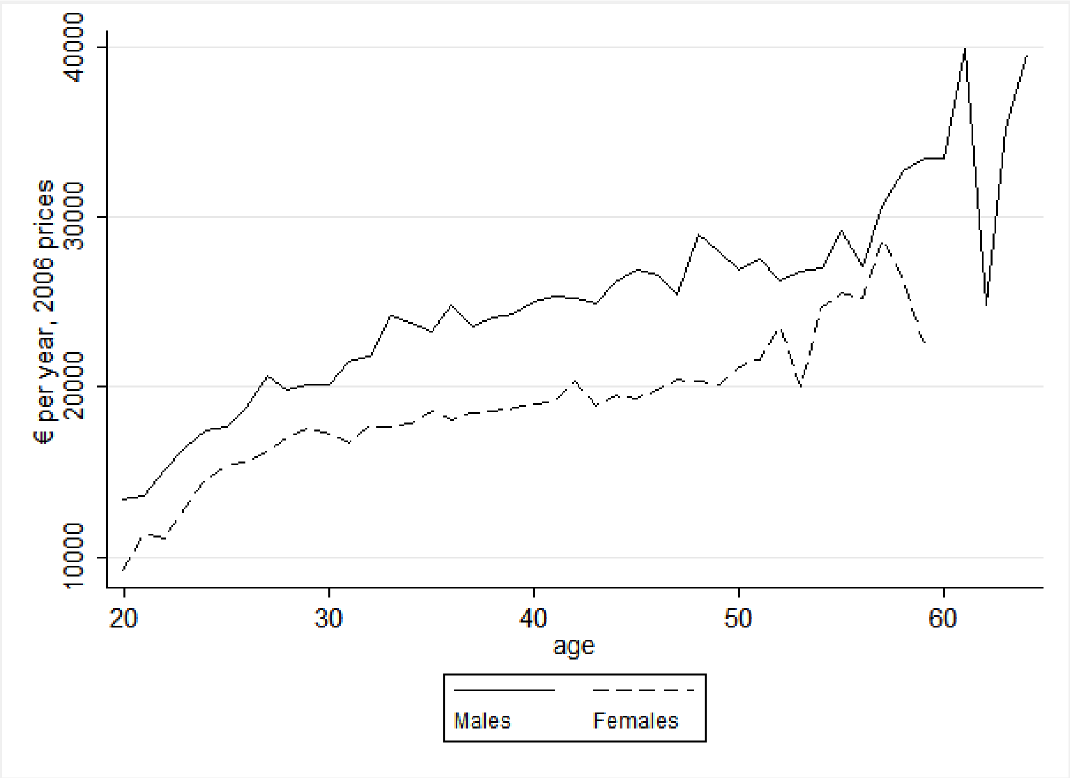
Graph 1 displays the medians annual gross earnings by age. The line relative to men always lies above that for women and the variability is larger near the official retirement age. We can clearly observe an increasing trend, even though the cross-section analysis might be confounding the age effect with the cohort one. A similar graph for the mean earnings, not reported, displays the same patterns, even if the trend looks more linear.

Graph 2 reports again the trend of the median earnings with respect to age, but disaggregating by level of education. Differences are quite small for workers aged less than 30, while for older individuals we observe an increasing educational premium. One reason for the limited differences for young workers might be the self-selection into post-graduate studies.

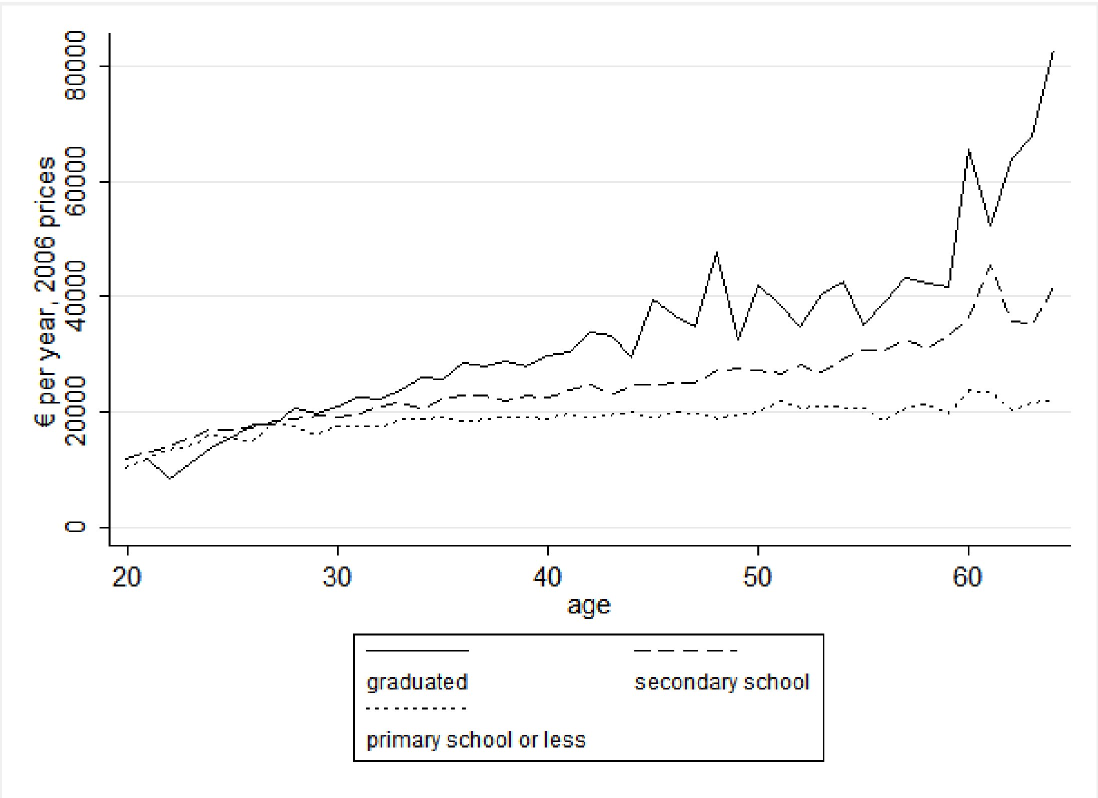
[8] We reclassify them to one euro in order to allow the logarithm transformation for regression estimates.

[9] Employees in the public sector show higher yearly net earnings also in the 2006 Survey of Households' Income and Wealth (SHIW) carried out by the Bank of Italy. Using SHIW microdata, we do not have gross variables, but we can define net earnings by summing variables *yl1* (annual net employee earnings, without non-cash employee income) and *ym* (annual net cash and losses from self-employment). Using sampling weights, the mean for employees in the public sector is 18,636 euro (s.d. 10,936), while it is 15,314 (s.d. 8,823) for employees in the private sector. Without sampling weight, similar results are obtained.

Graph 1 Median annual gross earnings, by age and sex, euro 2006



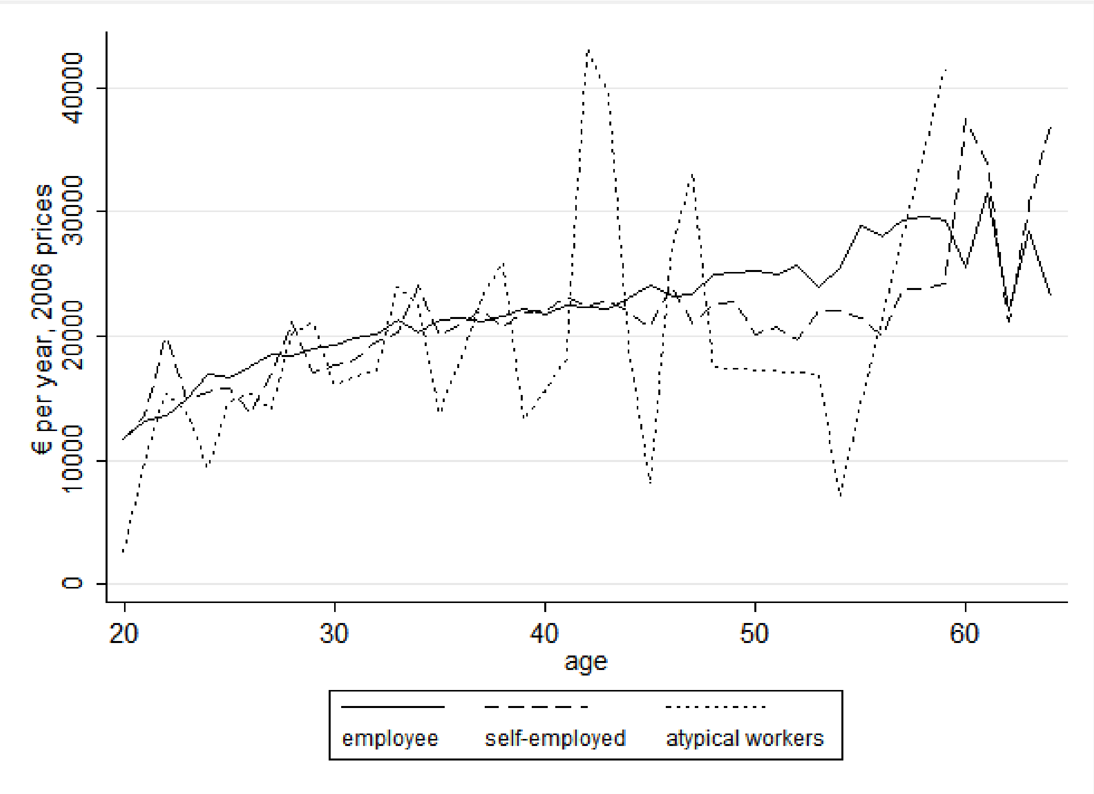
Graph 2 Median annual gross earnings by age and level of education, euro 2006



In Graph 3 we can observe that the age-trend is quite similar across different statuses in employment, even if the median earnings are larger for employees aged between 50 and 60.[10] Atypical workers display greater dispersion in their earnings.

Table 2 shows the presence of individuals with very low annual earnings among those who worked for less than 12 months in 2006, which are nearly 8% of the sample. Given that CAPP-DYN simulates economic and demographic transitions among states in discrete time (annual cycle), we should avoid this source of heterogeneity. On the one hand, we should take into account that these workers might have low annual earning levels as a consequence of within-year periods of unemployment.[11] On the other hand, we prefer not to exclude them from the analysis, in order to introduce further self-selection.

Graph 3 Median annual gross earnings by age and status in employment



A possible solution, apart from the one which requires the simulation of the conditional probability of being at work *m* months and the use of this covariate in the model, is to estimate the monthly gross earnings for those months in which these individuals were working.[12]

[10] In Graph 3 we excluded 541 (2.7%) observations with annual gross earnings higher than 80 thousand euro.
[11] It should be recalled that the way in which we define the economic status allows individuals who worked for two months to be workers. See Ciani et al. (2011) for a full discussion.
[12] The number of months worked in 2006 is the sum of variables pl070 and pl072. For 68 observations, corresponding to 0.34% of the sample, this sum is equal to zero. These are the persons who were reclassified as workers in the construction of the initial population, because their earnings were larger than their income from pensions. The median

Table 3 shows how this choice reduces the variability of mean earnings for different numbers of months in which the individuals worked. The means of both the 1st and 5th quintiles are more reasonable with respect to the limited period of one month, as it seems that there less outliers.

Table 2 Annual gross earnings, by months worked, euro 2006

Months worked in 2006	Mean	1st quintile	5th quintile	95th quintile	99th quintile	Per cent
2	9,122	755	1,273	23,723	67,073	0.6%
3	6,227	884	1,538	17,506	30,977	0.8%
4	7,659	732	1,878	23,222	44,499	0.8%
5	9,961	2,025	2,827	30,782	56,950	0.7%
6	12,913	1,346	3,554	31,341	117,033	1.0%
7	11,382	2,439	3,711	23,407	47,321	0.8%
8	15,738	2,923	4,868	40,043	72,554	0.9%
9	13,870	685	3,889	31,558	53,874	0.9%
10	15,778	1,542	3,253	36,565	57,231	1.1%
11	21,036	3,366	6,054	37,217	84,768	0.6%
12	27,619	3,526	7,791	62,334	116,998	92.1%

Table 3 Monthly gross earnings by months worked, euro 2006

Months worked in 2006	Mean	1st quintile	5th quintile	95th quintile	99th quintile	Per cent
2	4,561	378	637	11,862	33,536	0.6%
3	2,076	295	513	5,835	10,326	0.8%
4	1,915	183	470	5,806	11,125	0.8%
5	1,992	405	565	6,156	11,390	0.7%
6	2,152	224	592	5,224	19,506	1.0%
7	1,626	348	530	3,344	6,760	0.8%
8	1,967	365	609	5,005	9,069	0.9%
9	1,541	76	432	3,506	5,986	0.9%
10	1,578	154	325	3,657	5,723	1.1%
11	1,912	306	550	3,383	7,706	0.6%
12	2,302	294	649	5,195	9,750	92.1%

4. OLS estimates for monthly gross log-earnings

In CAPP-DYN we need a statistical model to forecast earnings for workers in future years of the simulation. First of all, earnings should change with experience and age. Secondly, we need to account for heterogeneity due to different socio-demographic conditions and statuses in employment.

of their annual gross earnings is 38,520, with a minimum of 4,320. For these individuals we set the number of months worked to 12, since we removed all monthly pension transfers.

As discussed, our prediction is based on cross-sectional estimates. The classical theoretical guide is the Mincer earnings function, where the logarithm of earnings *y* is a linear function of years of education *t* and of experience r, as described by Cahuc & Zylberberg (2004, p. 87)[13]

$$\ln y\,(t+x) = \alpha + \gamma_t t + \gamma_1 r + \gamma_2 r^2 + \ln[1 - s(r)]$$

where an individual with *t* years of schooling and experience *r* spends a share *s(r)* of his/her time in training. Generally, this last component is omitted in empirical analysis. Theory predicts a deterministic relation that, following the literature, we assume to hold on average:

$$E(\ln y\,|r,t) = \alpha + \gamma_t t + \gamma_1 r + \gamma_2 r^2$$

In order to allow for greater flexibility, instead of using the years of education *t* as a regressor, we decided to estimate different models for sub-groups of the working population, similarly to the Australian Microsimulation model NATSEM (Bækgaard, 2002, p. 39). In particular, we distinguished the seven groups reported in Table 4. We did not split the group of graduated self-employed by gender in order to maintain a larger number of observations.

Table 4 Groups for regressions

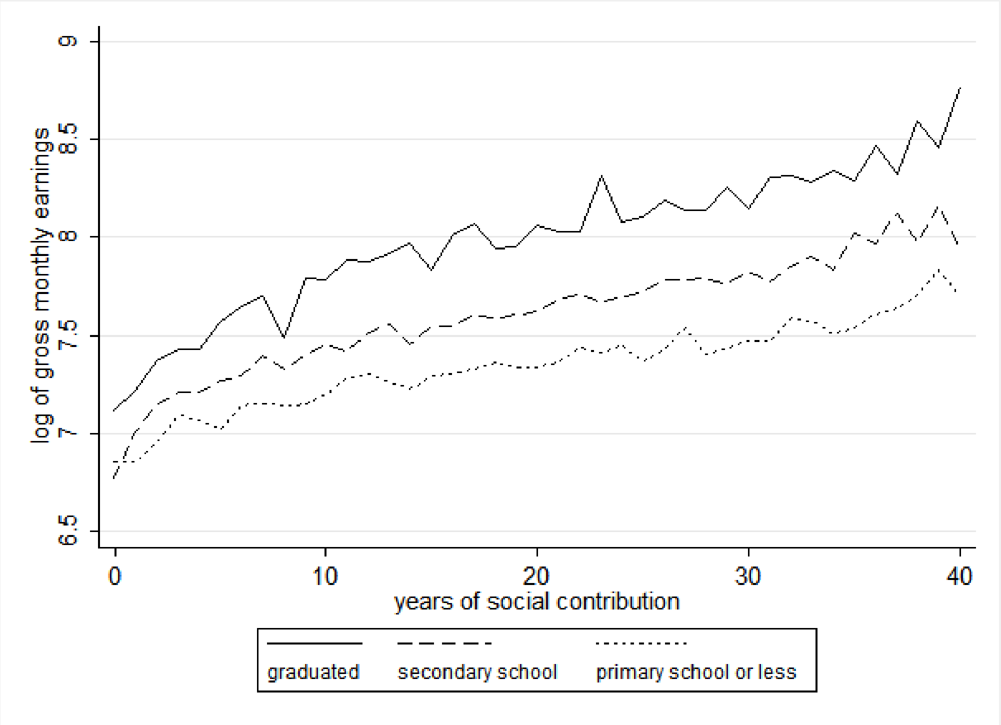
Group	Observations	Per cent
Men, not graduated, employees	7,478	38%
Men, graduated, employees	1,005	5%
Women, not graduated, employees	5,349	27%
Women, graduated, employees	1,169	6%
Graduated, self-employed	911	5%
Men, not graduated, self-employed	2,627	13%
Women, not graduated, self-employed	1,124	6%
Total	19,663	100%

As a proxy for years of experience we use the number of years in which the individual has paid social contributions, since in the CAPP_DYN model we keep track of this variable when we simulate the future employment history. The main limitation in using this variable is that it does not include years spent in paid work when the individual avoided paying social contributions for various reasons, such as participation in the grey or black market. Moreover, the variable is top-coded to the standard current requisite for retirement, which is 40 years. In order to check for its validity as a proxy, we compare it with a different It-Silc variable, where the respondents reported the number of years spent in paid work. In Graph 4 and Graph 5 we can observe that the relation between the mean logarithm monthly gross earnings and the years of social contributions is similar to the relation with respect to years spent in paid work, at least up to the top-coding for the former.

[13] A common procedure is to take the natural logarithm of earnings in adjusting for skewedness of the data. However, the logarithm transformation is not innocuous. See appendix A for a discussion and an alternative set up.

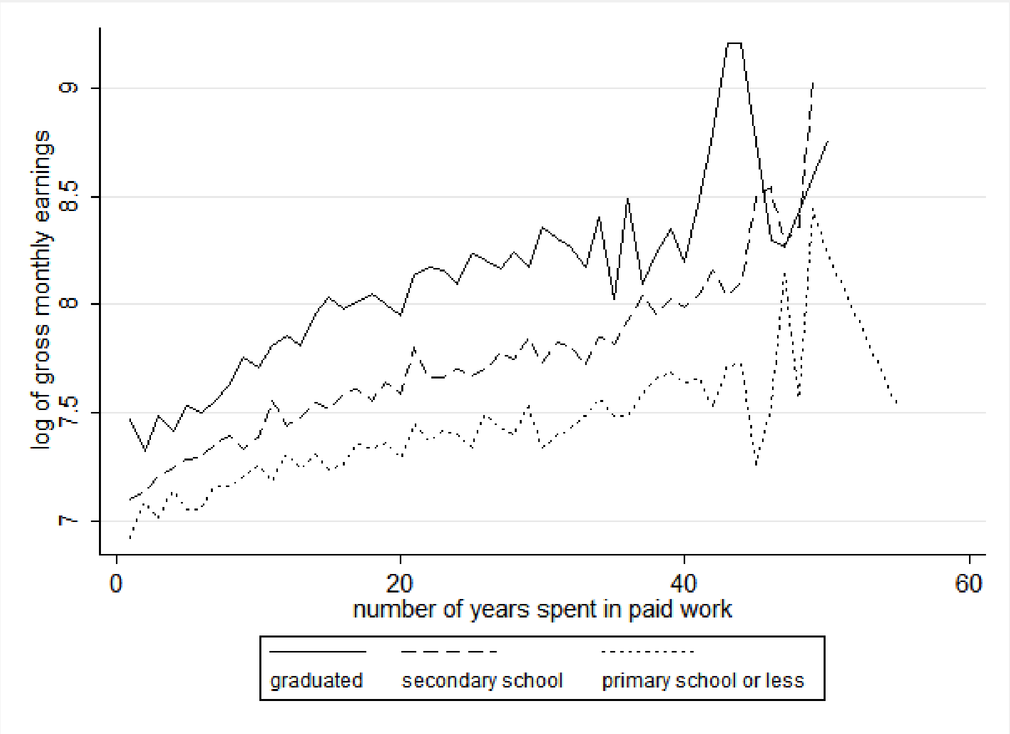
It is important to notice that Graph 5 shows a large dispersion after 40 years spent in paid job. Given that the estimated correlation between the two variables is quite high in the sample (0.8954), we prefer to use the years of social contribution, essentially because it is the variable that will keep track of the job history of each individual in the dynamic simulation.

Graph 4 Mean of the monthly logarithm gross earnings, by years of social contributions and level of education, euro 2006



We also added a vector of J regressors x to the basic model, as we expect them to have some effect on the earnings level: age and age squared; immigrant status; geographical area of residence; sector of employment; part-time or full-time job. In groups of non-graduated workers we also included a dummy indicating whether or not the individual completed secondary education. For self-employed, we control for a binary variable assuming value one if they have atypical contracts. Lastly, we add a dummy for females for the graduated self-employed. The description of the variables is reported in Table 10 in Appendix B. The gain in adding these variables is the increased amount of log-earnings variance explained by our model, even if we lose the correspondence with the theoretical Mincer equation.

Graph 5 Mean of the monthly gross earnings, by years spent in paid work and level of education, euro 2006



Furthermore, we added some interactions where required to improve the specification of the conditional mean. In order to choose which to include, we decided to proceed by adding interactions between the year of contributions on the one side and geographical and secondary education dummies on the other.[14] We did not include interactions with *age* in order to avoid picking up effects that are more likely to be related with cohort effects. The latter, as discussed in section 2, cannot be identifiable by using cross-sectional data.

Lastly, the high correlation between age, years of contributions and their squares causes estimates of the associated coefficients to show high variation. We removed the squares of either the age or social contributions dummies when they turned out to be not significant at the 10% level. For six out of seven groups we retain the square of age but not that of years of social contribution, consistently with previous works using Italian administrative data for the private sector (see Brugiavini & Peracchi (2003, p. 92), Giarda, (2007, p. 70)).[15]

If we always add the interactions, several coefficients turn out to be estimated with poor precision. Selection of the final model is a non-trivial issue. One solution could be to include all

[14] The original Mincer model predicts that "log-earnings experience profiles are parallel across schooling levels" (Heckman, Lochner, & Todd, 2003, p. 8). However, Heckman et al. (2003) strongly rejected this assumption with US data.

[15] Giarda (2007, p. 70) includes the logarithm of years of social contributions, but there is no specific reason to follow her choice.

interactions in a first step and then exclude those which are not significant at, say, the 5% level. However, it is not clear how to proceed in removing non-significant coefficients. As discussed by Goldberger (1991, p. 258-261), when selection is driven by data, standard t-test statistics are not valid. The risk of "data mining" (Lovell, 1983) is to select a few interactions that appear to be statistically significant only because we are using the wrong critical values. Furthermore, it is not clear whether or not we should remove regressors that were suggested by theory, such as experience, or that appear to be unavoidable, like the dummy for part-time. This is particularly true when their coefficients turned out insignificant from the statistical point of view but not from the economic perspective.

Therefore we proceed in a different way. We start with the basic set of regressors and test for correct specification using a RESET test (Ramsey, 1969), run by adding the squares of the fitted values in the OLS regression, and testing their significance with a robust F-test. If the model passes the test at the 5% level, we do not add any interaction. Otherwise, we first add the set of interactions previously discussed. We came up with the specifications displayed in Table 5 where all models have passed the RESET test at the 5% level. The group for non-graduated female employees passed the test with a p-value of only 0.0500, but the p-value improved to 0.0998 when we added the square of social contribution, whose coefficient is not significant at the 10% level. Other interactions turned out to be significant, in particular those among other dummies, such as private sector or immigrant. However, we chose not to add them, because we prefer to keep a parsimonious but possibly correctly specified model. Indeed, it is clear that adding interactions always increases the R^2, but it is well known that this statistics does not provide a good guidance for model choice. Anyway, we never observed any large increase in the explained variance when we included these additional regressors, while the RESET test result did not always improve.

If we define the $1\ X\ K$ vector of regressors as $z_i = (1, r_i, r_i^2, x_i)$, $K = J + 3$, the statistical model can be written as

$$\ln y_i = z_i\beta + \varepsilon_i$$
$$E(\varepsilon_i |\, z_i) = 0$$

where β is the $K\ X\ 1$ vector of parameters.[16] The model is therefore estimated using OLS, assuming linearity between log earning and its determinants, in line with all previous models used in a DMM. We pursued other approaches which relax this assumption. Poisson Maximum Likelihood model is discussed in the appendix.[17] OLS results are reported in Table 5.

[16] We bottom-coded $\ln y_i$ to 5.7 euro and we top-coded it at 9.2 euro, respectively the 1st and the 99th percentile, in order to avoid outliers.

[17] One critique might be that we can estimate the regression only for individuals who are actually working, because we do not know the wage for those currently unemployed. Given that the labor market status of an individual is

Geographical differences are highlighted by a positive North dummy coefficient, while the South dummy is usually negative. As discussed before, private sector employees tend to have lower gross earnings. Part-time workers show sensible lower earnings, even if the difference is lower among self-employed. The coefficient on the immigrant dummy is negative and quite large, as it might have been expected. Coefficients on age and years of social contributions are always positive, apart from the small and not significant coefficient for men, not graduated, self-employed. There seems to be decreasing returns on age or experience, as the quadratic term is generally negative. Among not-graduated employees, the interaction term between social contribution and secondary education is positive and significant, as well as the interaction with the North dummy, showing higher returns on experience for these groups. Lastly, the atypical dummy is negative and significantly different from zero only among men, not graduated, self-employed.

Results are in line with expectations and previous studies. However, one possible concern is that we still observe an increase in earnings with age even for individuals aged 50-64. Brugiavini & Peracchi (2003, p. 90-94), using administrative panel data for private sector non-agricultural employees between 1973 and 1997, argued that the annual earning profile is flatter after 50. We check whether the regression was able to smooth the steep increase in earnings in the last years of the active life that we observed in the descriptive analysis.[18] In Graph 6 the predicted age pattern for men seems slightly concave, and the steep though variable increase in years near to 65 seems to be smoothed with regression. For women we did not observe a similar pattern in actual values, and we observe an almost linear trend in predicted values (Graph 7). Employees (Graph 9) exhibited the steepest increases around age 60, which are again smoothed by regressions. We observe a similar result for self-employed (Graph 9). Lastly, it should be added that, when Brugiavini & Peracchi (2003) used the annualized monthly earnings, "defined as annual earnings divided by the fraction of months a person worked during the year" (idem, pg. 91), they also found that earnings increase both for men and for women.

endogenously determined in our model, we might need to use a selection correction in estimating earnings. As Emmerson, Reed, & Shephard (2004, p. 19) state "it would in theory be possible to estimate the entire model jointly via maximum." However, this would require labor market transitions and wage regressions to be estimated on the same dataset. We decided not to follow this route for the short length of It-Silc forced us to use the labor force survey (see Flisi and Morciano, 2011) in estimating conditional transition probabilities in the labor market. Furthermore, given the non-structural nature of our model, we would need a highly parameterized model in order to allow monte-carlo simulation. Lastly, reports from other dynamic microsimulation models do not currently mention any concern about this issue. See for example the French *DESTINIE* (Blanchet, Buffeteau, Crenner, & Le Minez, 2010, p. 12) and the model for the Italian Tuscany region *MIRTODIN* (Maitino & Sciclone, 2009).

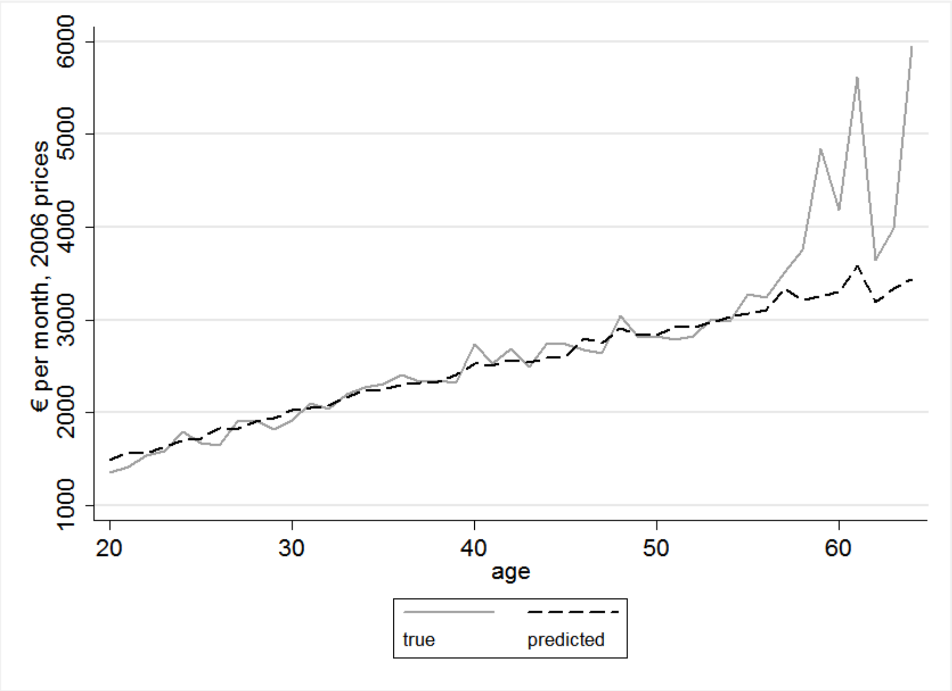
[18] In order to further address this concern, we repeated the estimates in Table 5 excluding first men aged 60 or more, and then restricting the sample to men and women aged 20 to 54.

Table 5 OLS estimates, dependent variable logarithm monthly gross earnings, euro 2006

	Men, not graduated, employees	Women, not graduated, employees	Men, graduated, employees	Women, graduated, employees	Graduated, self-employed	Men, not graduated, self-employed	Women, not graduated, self-employed
North	0.1006***	0.1187***	0.0651*	0.0595*	0.2425***	0.0646**	0.0442
	(0.023)	(0.027)	(0.039)	(0.033)	(0.057)	(0.032)	(0.050)
South	-0.1458***	-0.0961***	-0.1686***	-0.0772**	0.0023	-0.2655***	-0.3025***
	(0.027)	(0.034)	(0.046)	(0.038)	(0.070)	(0.036)	(0.059)
Private	-0.0845***	-0.1619***	-0.0694*	-0.0625*			
	(0.012)	(0.014)	(0.036)	(0.034)			
Part-time	-0.6440***	-0.5145***	-0.5826***	-0.5718***	-0.4604***	-0.2742***	-0.2830***
	(0.034)	(0.015)	(0.147)	(0.040)	(0.080)	(0.078)	(0.054)
Secondary	0.0507***	0.1402***				0.1926***	0.1990***
	(0.019)	(0.025)				(0.026)	(0.043)
Immigrant	-0.2228***	-0.2118***	-0.4556***	-0.3970***	-0.4060***	-0.0313	-0.2395**
	(0.020)	(0.032)	(0.074)	(0.081)	(0.139)	(0.079)	(0.110)
Age	0.0336***	0.0147***	0.0736***	0.0493***	0.0247***	-0.0029	0.0340**
	(0.004)	(0.005)	(0.015)	(0.012)	(0.005)	(0.010)	(0.017)
Age squared	-0.0004***	-0.0002***	-0.0006***	-0.0004***		0.0001	-0.0004**
	(0.000)	(0.000)	(0.000)	(0.000)		(0.000)	(0.000)
Contributions	0.0120***	0.0154***	0.0024	0.0111***	0.0396***	0.0060**	0.0082**
	(0.001)	(0.002)	(0.004)	(0.003)	(0.011)	(0.003)	(0.003)
Contributions squared					-0.0009***		
					(0.000)		
Secondary*Contrib.	0.0086***	0.0079***					
	(0.001)	(0.001)					
North*Contributions	-0.0032***	-0.0034**					
	(0.001)	(0.001)					
South*Contributions	-0.0003	0.0012					
	(0.001)	(0.002)					
Women					-0.1621***		
					(0.054)		
Atypical					-0.0002	-0.2127***	-0.1089
					(0.062)	(0.077)	(0.070)
Constant	6.7337***	6.9234***	6.0849***	6.4061***	6.5138***	7.3012***	6.4038***
	(0.076)	(0.100)	(0.330)	(0.244)	(0.169)	(0.209)	(0.332)
Observations	7478	5349	1005	1169	911	2627	1124
R^2	0.334	0.427	0.293	0.349	0.265	0.112	0.119
Adjusted R^2	0.333	0.426	0.287	0.344	0.258	0.109	0.111
Res. sum of squares	1154	915	258	247	477	1136	510
RESET (p-value)	0.155	0.171	0.259	0.213	0.529	0.434	0.458
RESET (p-value)	0.6852	0.4375	0.5726	0.4482	0.5171	0.8156	0.6057

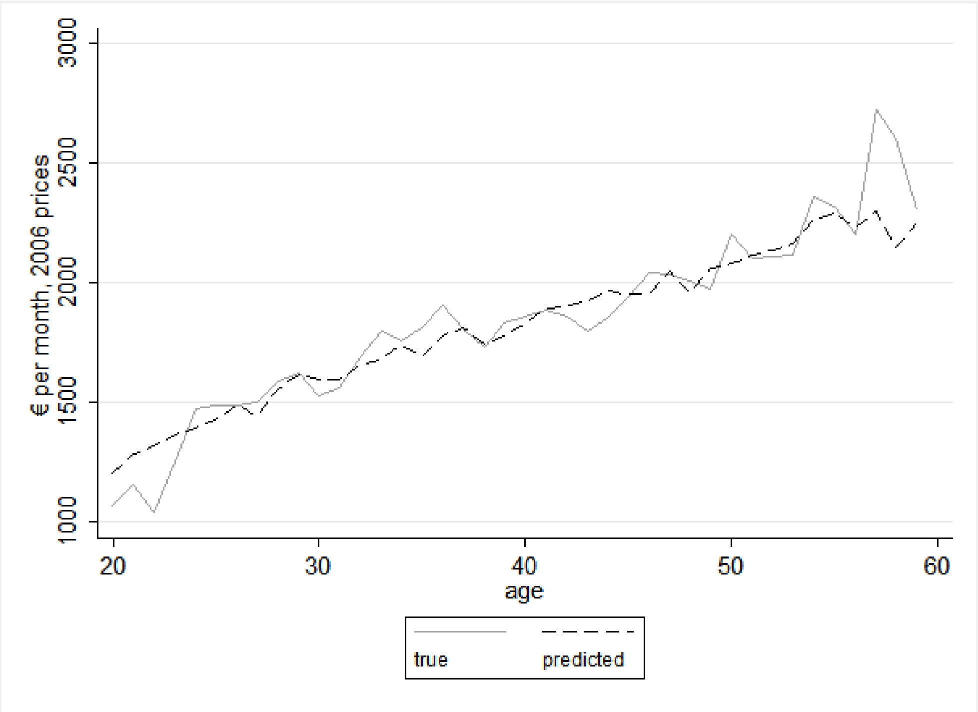
Note: Standard errors robust for heteroskedasticity in parentheses. * p<.10, ** p<.05, *** p<.01. Reference group: men (when both sexes are included), primary school, Italian citizen, working full-time, living in Central Italy, non-atypical worker, working in the public sector. The RESET test is conducted testing the joint significance of the square of fitted values, using a heteroskedasticity robust F test.

Graph 6 Monthly gross earnings for male workers, actual and predicted mean values, 2006



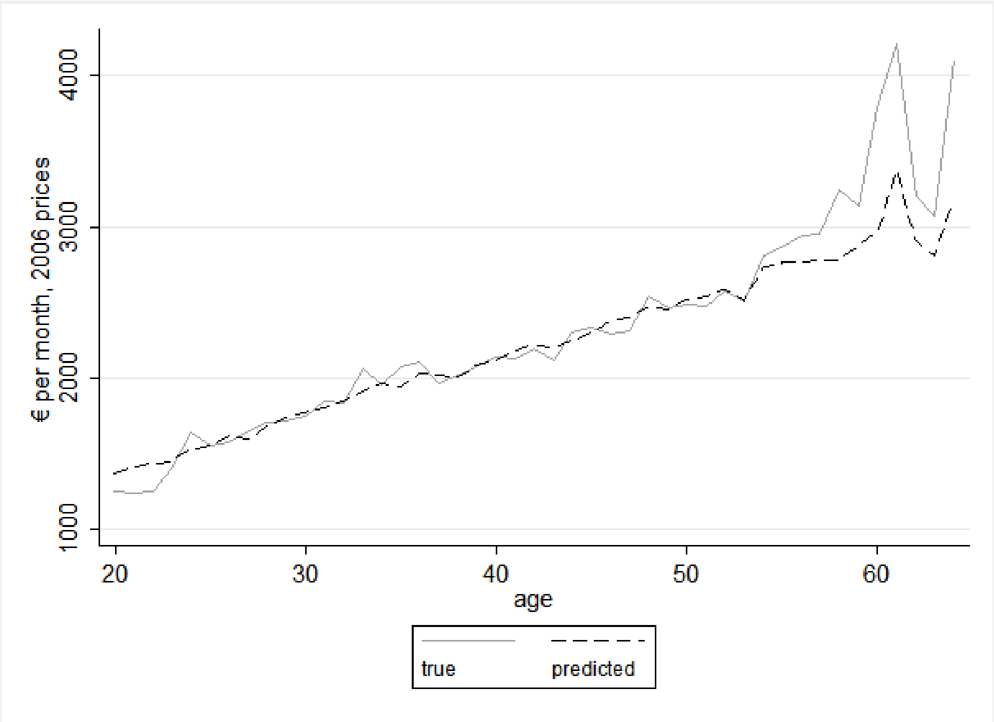
Note: the predicted values $z\beta$ are first obtained from the regression. To transform them in log, we use the transformation $y = e^{z\beta}e^{\sigma^2/2}$.

Graph 7 Monthly gross earnings for female workers, actual and predicted mean values, 2006



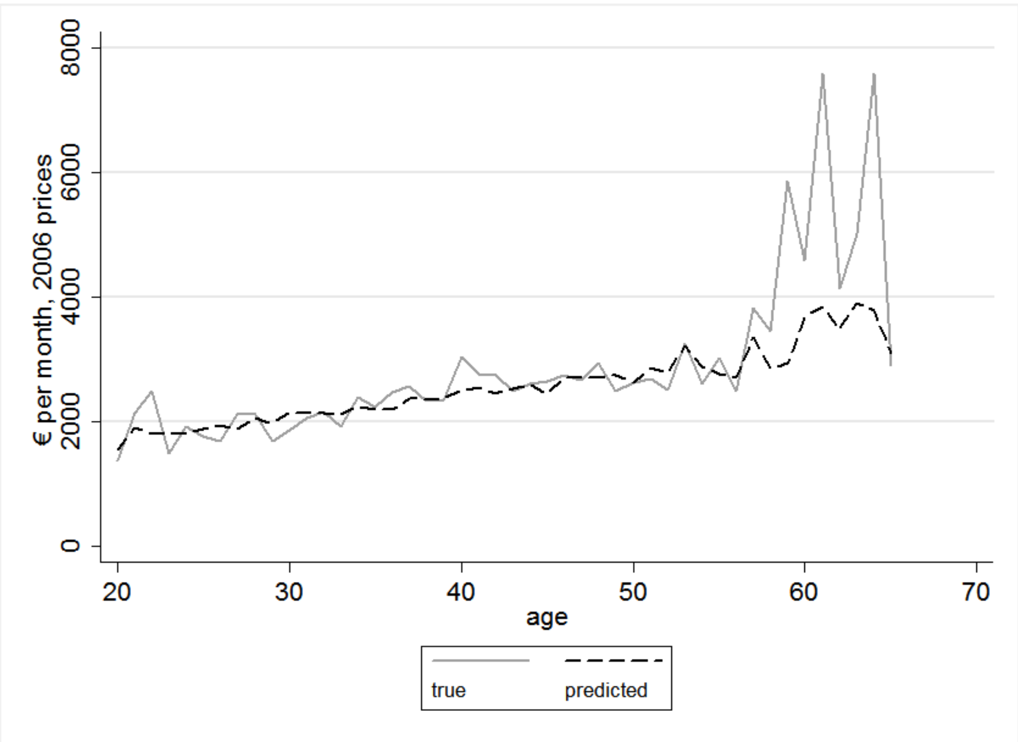
Note: the predicted values $z\beta$ are first obtained from the regression. To transform them in log, we use the transformation $y = e^{z\beta}e^{\sigma^2/2}$.

Graph 8 Monthly gross earnings for employees, actual and predicted mean values, 2006



Note: the predicted values $z\beta$ are first obtained from the regression. To transform them in log, we use the transformation $y = e^{z\beta}e^{\sigma^2/2}$.

Graph 9 Monthly gross earnings for self-employed, actual and predicted mean values, 2006



Note: the predicted values $z\beta$ are first obtained from the regression. To transform them in log, we use the transformation $y = e^{z\beta}e^{\sigma^2/2}$.

5. Panel estimates

We now estimate a similar model using the panel component of It-Silc. The main purpose of this exercise is to model the serial correlation in the earnings residuals, in order to allow for serial correlation due both to individual unobserved heterogeneity and to an autoregressive transitory component. In this way, following Pudney (1992), as explained in section 6, we were able to forecast future residual components.

Borella (2004) used data from the Bank of Italy Survey on Households' Income and Wealth to argue that a satisfactory model for the earnings residuals should include an individual effect plus a first-order autoregressive component:[19]

$$\begin{gathered}\ln y_{it} = z_{it}\beta + \varepsilon_{it} \\ \varepsilon_{it} = \mu_i + \xi_{it} \\ \xi_{it} = \rho\xi_{it-1} + \omega_{it},\ \omega_{it} \sim iid(0, \sigma_\omega^2),\ |\rho| < 1 \\ \mu_i \sim iid\left(0, \sigma_\mu^2\right)\end{gathered} \tag{1}$$

where z_{it} is the $1\ X\ K$ vector of regressors (including a constant) observed for individual *i* at time *t*. Actually, this is the model originally discussed in Lillard and Willis (1978). Baltagi and Li (1991) proposed a different and simpler estimation procedure, while Baltagi and Wu (1999) discussed the case of unequally spaced panels. It should be noted that, in order to recover consistent estimators for α, β, σ_ω^2 and σ_μ^2, we are assuming strict exogeneity, so that at each period *t* the time-varying error component is independent from the vector of covariates at all time periods.

Assuming that the individual effects μ_i are i.i.d. and independent from the regressors z_{it}, the model can be estimated using random effects. It may be argued that we should take into account the possibility that the individual effects are not independent of the exogenous regressors. However, this is not suitable for the setting up of the dynamic model, for the two reasons discussed by Emmerson, Reed and Shephard (2004, p. 33). First, each individual who enters paid employment or starts self-employment for the first time must be assigned an individual effect. However, fixed effect methods do not allow us to consistently estimate the variance of the fixed effect in order to predict it for these individuals. Secondly, fixed effect estimation does not allow for time-invariant characteristics, such as gender or education, which are important in explaining variation.

The main problem in estimating a model on the panel is that we do not have the information on gross earnings, because Italian National Institute of Statistics (ISTAT) started providing gross earnings only from the 2007 cross-section. Since then, ISTAT has started to recover information on

[19] Borella's model allows for the possibility that the variance of the error changes with age. We neglect this problem for the moment, assuming as in Lillard and Willis (1978, p. 6) that earnings in the initial condition display a shock $\xi_{it} = \omega_{i0}/\sqrt{(1-\rho^2)}$,, so that they can be treated as stationary.

gross-earnings also with respect to previous waves, as they come from one-to-one matching with administrative archives. However, the process is not yet concluded, and therefore we are not currently able to exploit the longitudinal data to analyse the gross earnings. We nevertheless carry out the estimation in order to provide some evidence on the autocorrelation of the residuals in net-earnings, assuming that this can approximately describe the covariance structure in gross-earnings. New waves available as from next year onward will enable us promptly to improve the model.

With this caveat in mind, we estimate this model on the first four rotational panels of It-Silc: 2004-2007, 2005-2008, 2006-2008, 2007-2008.[20] In order to make it comparable with CAPP_DYN initial population, we apply to each year the corrections of the sample that we described in Ciani et al (2011). Given the short length of the panel, we chose to remove observations where we observe a difference between the earnings in two subsequent waves larger in absolute value than the 95th percentile of the differences observed in all the years. This operation is conducted separately for the seven groups, so that the threshold varies from 953 euro per month for women, not graduated, employees, to 3835 per month for graduated self-employed.

In panel estimates it is not possible to include some variables used in the previous paragraph. For privacy reasons, we cannot match cross-sectional observations with longitudinal records. However, the latter does not include all variables provided with the Italian version of SILC. First of all, this means we cannot control directly for the private/public sector, because the European variables do not exactly correspond to the variable that we use in the Italian cross-section.[21] Secondly, we do not know whether the self-employed individual is working with an atypical contract. Thirdly, the information on years of social contribution is not available, hence we use the variable reporting the number of years spent in paid work, top-coded to 40. Lastly, the current Istat release of 2005-2008, 2006-2008, 2007-2008 panels does not contain the variable "pb220a: Citizenship" which is needed to identify the immigrant workers.

Some descriptive statistics are reported in Table 12 and Table 14. We do not observe appreciable differences in covariates means. However, Table 13 highlights the presence of some differences that are significantly different from zero in the statistical sense. The most significant are those arising in the variables *contributions*. This is essentially due to the fact that, in the panel, this

[20] Estimates are carried out using the Stata™ command *xtregar*, which follows Baltagi and Wu (1999). The four rotational panels are pooled in one single unbalanced panel. In order to adjust for inflation, we transform nominal values for earnings to express them in 2006 prices. For each year we use a coefficient equal to the average Italian consumer price index NIC (including tobacco) in that year divided by the average index in 2006. NIC time series are provided by Istat, and are available at http://www.istat.it/prezzi/precon/dati/ (last access: 18/04/2011). After adjusting for inflation, we bottom-coded $\ln y_{it}$ to 5.16 euro and top-coded it at 8.73 euro, respectively the 1st and the 99th percentile, in order to avoid outliers.

[21] This problem is exacerbated by the absence of the variable *pl110* in the current releases of the IT-SILC longitudinal dataset.

variable is measured as the number of years spent in paid work, and therefore it has to be understood as a proxy for the regressors used in cross-section estimates. The difference between the mean in the two samples is nevertheless quite small for employees, whereas it is approximately two years for the self-employed. With respect to individual categories, it seems that those displaying larger differences are the self-employed. In these groups we find sensible differences for *age*, which might be due to the retirement of older cohorts of individuals with lower education. In a nutshell, differences are small overall and might be explained by the fact the panel covers a period of three years before the cross-section from which the initial population was drawn.

Estimates are reported in Table 6. Note that the results are not directly comparable with Table 5, because in the set of regressors we include variables that are missing (private, atypical). Moreover, *contributions* is proxied by *years spent in paid work*, and we observe different estimates with respect to Table 5. The dummy *north* is generally smaller with respect to OLS, while *south*, *secondary* and *female* are similar. The coefficient on *age* and *age squared* is larger for all groups of employees. However, it is likely to include some of the effect of *contributions*, because the coefficient on its proxy *years spent in paid work* is smaller than what we observe for *contributions* in the cross-section. Among self-employed, the largest difference is observed for not-graduated men, because we find coefficients on *age* and *age squared*, which are respectively positive and negative.

The most significant finding is that the dummy for part-time exhibits smaller coefficients in all groups, in particular among employees. If we interpret it in terms of elasticities, the coefficient seems indeed quite small from the economic perspective. One possible reason is that the variable *part-time* does not satisfy the strict exogeneity restriction. Indeed, it should be recalled that, under the random effects assumptions previously stated, pooled OLS would be not efficient, but still consistent. Moreover, pooled OLS would still be consistent without strict exogeneity. We checked what happens for the employees group running pooled OLS, and we found estimated coefficients for *part-time* that are appreciably larger. Therefore we run a test for strict exogeneity proposed by Wooldridge (2002, p. 285), where we essentially test the significance of a lead of *part-time* as an additional regressor in random effects estimates. We fail to reject the null of its coefficient being equal to zero at the 1% level in all groups of employees. This finding suggests that the Random Effects panel estimates, although generally preferable, should be treated with caution, because strict exogeneity might be violated in modelling the mean of log-earnings.

Table 6 Panel estimates, dependent variable logarithm monthly net earnings, euro 2006

	Men, not graduated, employees	Women, not graduated, employees	Men, graduated, employees	Women, graduated, employees	Graduated, self-employed	Men, not graduated, self-employed	Women, not graduated, self-employed
North	0.0625***	0.0804***	0.0087	0.0555**	0.1049**	0.0535**	0.0532
	(0.014)	(0.016)	(0.028)	(0.023)	(0.049)	(0.023)	(0.039)
South	-0.1137***	-0.1029***	-0.1404***	-0.0393	-0.0543	-0.2483***	-0.2949***
	(0.015)	(0.019)	(0.031)	(0.026)	(0.056)	(0.026)	(0.047)
Part-time	-0.2892***	-0.2823***	-0.1701***	-0.2361***	-0.3922***	-0.1936***	-0.1835***
	(0.010)	(0.007)	(0.032)	(0.017)	(0.042)	(0.034)	(0.032)
Secondary	0.0692***	0.1634***				0.1164***	0.1196***
	(0.010)	(0.013)				(0.017)	(0.032)
Age	0.0504***	0.0270***	0.0890***	0.0368***	0.0162***	0.0561***	0.0190
	(0.002)	(0.003)	(0.009)	(0.008)	(0.003)	(0.007)	(0.014)
Age squared	-0.0005***	-0.0003***	-0.0008***	-0.0002**		-0.0006***	-0.0002
	(0.000)	(0.000)	(0.000)	(0.000)		(0.000)	(0.000)
Years spent in paid work	0.0002	0.0016**	0.0002	0.0001	0.0115*	0.0014	0.0028
	(0.001)	(0.001)	(0.001)	(0.001)	(0.006)	(0.001)	(0.002)
Years spent in paid work squared					-0.0003**		
					(0.000)		
Secondary*Years spent in paid work	0.0033***	0.0034***					
	(0.000)	(0.001)					
North*Years spent in paid work	-0.0009	-0.0012					
	(0.001)	(0.001)					
South*Years spent in paid work	0.0006	0.0019**					
	(0.001)	(0.001)					
Women					-0.1560***		
					(0.044)		
Year==2004	0.0166***	0.0212***	0.0169	0.0094	0.0031	0.0224	0.0186
	(0.005)	(0.007)	(0.016)	(0.016)	(0.043)	(0.019)	(0.032)
Year==2005	0.0036	-0.0068	0.0285*	0.0063	0.0261	0.0341*	-0.0053
	(0.005)	(0.008)	(0.016)	(0.017)	(0.043)	(0.019)	(0.033)
Year==2006	-0.0010	-0.0074	0.0594***	0.0353**	0.0711	0.0354*	-0.0239
	(0.006)	(0.008)	(0.017)	(0.017)	(0.043)	(0.019)	(0.034)
Year==2007	-0.0153**	-0.0151*	0.0212	-0.0044	0.0660	0.0612***	0.0178
	(0.006)	(0.009)	(0.019)	(0.019)	(0.047)	(0.021)	(0.037)
Constant	6.0714***	6.2791***	5.4221***	6.2621***	6.7279***	5.9161***	6.4171***
	(0.044)	(0.060)	(0.181)	(0.160)	(0.123)	(0.146)	(0.281)
Observations	21583	15335	2701	3286	1881	7317	2698
σ_μ	0.240	0.242	0.317	0.285	0.453	0.399	0.427
σ_ξ	0.185	0.222	0.194	0.213	0.410	0.369	0.396
ρ	0.322	0.346	0.249	0.282	0.123	0.171	0.237
$\sigma^2_\varepsilon = \sigma^2_{\mu +} \sigma^2_\xi$	0.092	0.108	0.138	0.127	0.373	0.295	0.339
RESET (p-value)	0.2508	0.6416	0.7302	0.3768	0.0047	0.231	0.4233

Note: Standard errors robust for heteroskedasticity in parentheses. * p<.10, ** p<.05, *** p<.01. Reference group: men (when both sexes are included), primary school, working full-time, living in Central Italy. The RESET test is conducted testing the joint significance of the square of fitted values, using an heteroskedasticity robust F test.

6. Simulation of the earnings profile in the dynamic model

Once a position within the labour force is simulated, the projections of individual annual earning in CAPP_DYN are obtained as follows. Estimates displayed in Table 5 or panel estimates presented in Table 6 are used to predict the deterministic component of the individual earnings in every year of the simulation, multiplying the monthly earnings by 12.[22] However individual income

[22] We assume that all the sample members, for whom an employment status is predicted in time $t + s$, will be in work for the entire year (12 months) without experiencing unemployment spells within the simulated year.

differs because of the presence of unobserved individual effect and a yearly component which can be considered as the increase in productivity distributed to all workers in each simulation period.

Unobserved individual effects are modelled using the procedure proposed in Pudney (1992). In paragraph 4 we decomposed the residual ε_{it} of the logarithm monthly earnings into two components: an individual specific effect μ_i and an orthogonal error term ξ_{it}, independent among themselves and mean independent with respect to the covariates z_{it}. In order to simulate values for future earnings, we further assume that the errors are normally distributed:

$$\varepsilon_{it} = \mu_i + \xi_{it} \qquad (1)$$

$$\xi_{it} = \rho\xi_{it-1} + \omega_{it},\ \omega_{it} \sim N(0, \sigma_\omega^2),\ |\rho| < 1$$

$$\mu_i \sim N\left(0, \sigma_\mu^2\right)$$

Both components are assumed to have zero mean and variance σ_μ^2 and σ_ξ^2 respectively. It should also be noted that the autoregressive component implies that

$$cov(\xi_{it}, \xi_{it-\mathrm{k}} | 1, z_i) = \rho^k \sigma_\xi^2 \qquad (2)$$

where $z_i = \left(z'_{i0}, z'_{i1}, \dots, z'_{iT_i}\right)'$. It is interesting to note that, in the population, we can identify the mean of individual's log-earnings y_{is} in period s conditional on log-earnings at a different period t, y_{it}, and on the set of covariates at both time periods , z_{it} and z_{is}. Assuming normality of both components of the error, and independence among them, the conditional expectation of y_{is},is:

$$E(y_{is} |\, y_{it}, z_{it}, z_{is}) = z_{is}\beta + \delta(s,t)(y_{it} - z_{it}\beta)(3)$$

The first term $z_{is}\beta$ can be interpreted as the deterministic part computed using coefficients in Table 6 or alternatively using coefficients obtained from the cross-sectional models (Table 5) by the vector of updated characteristic z_{is} whereas the second term is the product between the term $(y_{it} - z_{it}\beta)$, which is equal to the composite error term ε_{it} and a weighting factor

$$\delta(s,t) = \frac{\sigma_\mu^2 + \rho^{|s-t|}\sigma_\xi^2}{\sigma_\mu^2 + \sigma_\xi^2} \qquad (4)$$

The model implies that we can forecast the conditional mean of log-earnings in any future period s of the microsimulation model using the covariates x_{is} in that period, the residuals $\left(y_{i0} - x_{i0}\hat{\beta}\right)$ estimated for the worker in the initial population (denoted with subscript *0*), and the estimates for $\delta(s,t)$. Intuitively, when we predict log-earning in period *s* we take into account the individual residual, as estimated in period *t*, but we assign to it a weight that declines with the distance between *t* and *s*.

A practical problem is that parameters in eq (4) can be estimated only using panel data, whereas the deterministic component of the earning equation can be obtained using a cross-section, as we do in CAPP_DYN. From our panel estimates we found a value of ρ ranging from 0.123 to 0.346; σ_μ ranging from 0.240 to 0.427 and σ_ξ ranging from 0.185 to 0.410. It is a not straightforward task to assess to which extend these values are reliable or not. Comparable results can be found in Ramos (2003), who found for the UK (period 1991 to 2002) parameters quite similar to those founded by Lillard and Willis (1978) using the American PSID panel.

The most detailed and reliable analysis for Italy has been carried out by Borella (2004) and Borella and Coda-Moscarola (2009). We focus on the latter because the model is directly comparable to our result. From Table 7 and Table 8 we note that, using the panel component of It-SILC, we found slightly higher values for the variance of the residuals, in particular for the standard deviation of the time varying error component ξ. Interestingly, in our results for self-employed the correlation term ρ is quite consistent with the result from Borella and Coda-Moscarola (2009), whereas it seems to be underestimated in our case for employees. This is possibly due to the short length of the IT-SILC panel, as the fraction of individuals observed for the maximum number of four years is still quite limited.[23] Moreover, the uses of different samples representative of the Italian population in different points in time, and with different definition of earnings and set of observable socio-economic characteristics, may explain a not negligible part of the differences in the estimates.

Table 7 Results in our models (net monthly log earnings)

	Men, not graduated, employees	Women, not graduated, employees	Men, graduated, employees	Women, graduated, employees	Graduated, self-employed	Men, not graduated, self-employed	Women, not graduated, self-employed
σ_μ	0.240	0.242	0.317	0.285	0.453	0.399	0.427
σ_ξ	0.185	0.222	0.194	0.213	0.410	0.369	0.396
$\sigma^2_\mu/\sigma^2_\varepsilon$	0.627	0.543	0.728	0.642	0.550	0.539	0.538
ρ	0.322	0.346	0.249	0.282	0.123	0.171	0.237
δ(2,1)	0.747	0.701	0.795	0.743	0.605	0.618	0.647
$\sigma^2_\mu+\sigma^2_\xi$ da panel	0.092	0.108	0.138	0.127	0.373	0.295	0.339
$\sigma^2_\mu+\sigma^2_\xi$ da cross-section	0.155	0.171	0.259	0.213	0.529	0.434	0.458

[23] One solution would be to implement a more robust estimator, such as the minimum distance one proposed by Borella (2004). However, it is not worth developing the estimates in that direction until the IT-SILC panel has reached a sufficient size.

Table 8 Results in Borella and Coda Moscarola (2009), gross log earnings.

	Males			Females		
	Blue collar	White collar	Self-employed	Blue collar	White collar	Self-employed
σ_μ	0.242	0.335	0.263	0.332	0.360	0.229
σ_ξ	0.140	0.130	0.317	0.193	0.180	0.310
$\sigma^2_\mu/\sigma^2_\varepsilon$	0.750	0.870	0.407	0.748	0.799	0.353
ρ	0.432	0.529	0.165	0.419	0.440	0.070
$\delta(2,1)$	0.858	0.869	0.543	0.787	0.813	0.465
$\sigma^2_\mu+\sigma^2_\xi$ da panel	0.078	0.129	0.170	0.147	0.162	0.148

Source: Borella & Coda Moscarola (2009, p. 31). Notation is adapted to the present paper.
Note: σ_ξ is derived using the relation $\sigma_\xi^2 = \sigma_\omega^2/(1-\rho^2)$ (Lillard & Willis, 1978, p. 989)

Because of the uncertainty regarding these values, we decided to build an earning module which is fully flexible in the choice of these parameters. Essentially, it should be noted that, in order to predict future earnings using the procedure proposed by Pudney (1992), we can get an estimate for $\delta(s,t)$ using values for $\hat{\rho}$, $\hat{\sigma}_\mu^2$ and $\hat{\sigma}_\xi^2$ obtained from other sources. In this way we can also carry out sensitivity analysis using estimates coming from different sources.

A further problem in generating stochastically earnings for the simulated period is that the term $(y_{it} - z_{it}\beta)$, which is equal to the composite error term ε_{it}, is unavailable for those who the information on earning is not available at the time of the interview (in work and not respondent; temporarily not in work). Assuming normality we compute this term extracting a random number from a normally distributed function with mean zero and variance $(\sigma_\mu^2 + \sigma_\xi^2)$.

Finally, y_{is} is multiplied by a factor $(1+\tau_s)$ allowing the individual earning in s to be linked to the medium-long term productivity growth, calibrated through the “scenario” block. Again there is one point which needs to be made clear: the demographic evolution and the increase in the stock of human capital in the coming decades increase the average earning level, since age and education have a positive effect on average labour earnings[24]. However, in this model, endogenous growth is smaller than the growth forecasts according to RGS, since it does not allow for the expected increase in productivity. In order to avoid over/under-estimations of earnings growth rates for the coming decades, the following procedure is adopted: every year, a *pro-quota* growth factor τ is added to the endogenous growth due to the socio-demographic evolution. This factor is equal to the difference between the exogenous earning growth fixed in the “scenario” and the earnings growth estimated by the model.

The term τ_s is given by:

[24] Other factors could have a negative effect, for instance the increase of female participation in the labour market, the increase of immigrants and the diffusion of part-time contracts.

$$\tau_s = m_s - \left(\frac{E(y_s)}{E(y_{s-1})} - 1\right) \quad (4)$$

where m is exogenously determined in the "scenario"[25], while $E(y_s)/E(y_{s-1})$ describes the endogenous growth rate generated by the model.

Appendix A: A non-linear model for earnings

Suppose that theory predicts a non-linear relation between earnings, experience and socio-demographic variables. In particular, let us restrict the attention to exponential models such as:

$$y = exp(z_{it}\beta)$$

In this case we can again assume that the relation holds on average, adding an error term η_i for each individual $i=1,\dots,n$:

$$y_i = exp(z_{it}\beta)\eta_i$$

$$E(\eta_i|z_{it}) = 1$$

As discussed by Santos Silva & Tenreyro (2006), the choice of using the logarithm of y is far from being innocuous. In order to have consistent estimators for the parameters in the linear equation:

$$\ln y = z_{it}\beta + \ln \eta_i$$

we need $E(\ln \eta_i\,|z_{it}) = 0$. But this is not necessarily true given the previous assumption on η_i, because:

$$E(\ln \eta_i\,|z_{it}) \neq ln\, E(\eta_i|z_{it}) = 0$$

An alternative proposed by Santos Silva & Tenreyro (2006) is to use the independence condition between the error term and the vector of regressors $Z_i = (r_i, r_i^2, x_i)$:

$$E\big[\big(y_i - exp(z_{it}\beta)\big)z_i\big] = 0$$

whose sample analogue is

$$\frac{1}{n}\sum_{i=1}^{n}\big[y_i - exp\big(z_{it}\hat{\beta}\big)\big]z_i = 0$$

This condition is equivalent to that of the Poisson Maximum Likelihood estimator, that could be implemented in Stata™ using the command *poisson*.

Results are reported in Table 9. It is reassuring to observe that there are no sensible variations with respect to Table 5. One of the main differences is that the dummy for the private sector is generally not significant and smaller in magnitude. The dummy for immigrants is now slightly larger, even if differences with OLS are smaller than twice the standard errors of Poisson

[25] RGS projects yearly-increases in productivity of 1.1% until 2020; 1.6% in the period 2021-2030; 1.8% in the period 2031-2040 and 1.7% in the period 2041-2050 (RGS 2009).

estimates. We observe some changes in sign for the square of contributions and age, but always when the coefficient turns out to be not significantly different from zero at the 10% level. The RESET test now fails to reject the null at the 5% level for men, not graduated, self-employed. Overall, these results do not seem to suggest any significant problem in estimating the logarithmic specification with OLS.

Table 9 Poisson estimates, dependent variable monthly gross earnings, euro 2006

	Men, not graduated, employees	Women, not graduated, employees	Men, graduated, employees	Women, graduated, employees	Graduated, self-employed	Men, not graduated, self-employed	Women, not graduated, self-employed
North	0.0993***	0.1270***	0.1013**	0.0564	0.2677***	0.1076*	0.0351
	(0.038)	(0.039)	(0.049)	(0.042)	(0.089)	(0.058)	(0.076)
South	-0.1149***	-0.0612	-0.1941***	-0.0527	0.0856	-0.2657***	-0.3559***
	(0.040)	(0.041)	(0.052)	(0.047)	(0.104)	(0.054)	(0.087)
Private	-0.0036	-0.0967***	0.0014	-0.0021			
	(0.015)	(0.019)	(0.051)	(0.042)			
Part-time	-0.5996***	-0.4949***	-0.4155**	-0.6160***	-0.4726***	0.2351	-0.2834***
	(0.041)	(0.019)	(0.197)	(0.043)	(0.103)	(0.434)	(0.069)
Secondary	0.0542**	0.1793***				0.3319***	0.1818**
	(0.026)	(0.035)				(0.061)	(0.071)
Immigrant	-0.2595***	-0.1162	-0.5917***	-0.4284***	-0.4682***	-0.0968	-0.3385***
	(0.025)	(0.084)	(0.089)	(0.089)	(0.159)	(0.102)	(0.101)
Age	0.0210***	0.0064	0.0749***	0.0623***	0.0290***	-0.0520**	0.0401*
	(0.006)	(0.008)	(0.024)	(0.014)	(0.008)	(0.022)	(0.022)
Age squared	-0.0002**	-0.0001	-0.0006**	-0.0006***		0.0008***	-0.0004
	(0.000)	(0.000)	(0.000)	(0.000)		(0.000)	(0.000)
Contributions	0.0094***	0.0170***	0.0046	0.0084**	0.0343***	0.0063	0.0072
	(0.002)	(0.003)	(0.006)	(0.004)	(0.013)	(0.006)	(0.004)
Contributions squared					-0.0006*		
					(0.000)		
Secondary*Contributions	0.0099***	0.0065***					
	(0.001)	(0.002)					
North*Contributions	-0.0035*	-0.0044*					
	(0.002)	(0.003)					
South*Contributions	-0.0026	-0.0001					
	(0.002)	(0.003)					
Women					-0.1039		
					(0.075)		
Atypical					-0.0483	-0.2375	-0.1214
					(0.111)	(0.174)	(0.089)
Constant	6.9161***	7.0346***	6.0899***	6.2130***	6.6102***	8.2169***	6.4498***
	(0.110)	(0.177)	(0.547)	(0.290)	(0.305)	(0.374)	(0.432)
Observations	7478	5349	1005	1169	911	2627	1124
RESET (p-value)	0.0789	0.8477	0.2149	0.3525	0.4073	0.0332	0.2492

Note: robust standard errors in parentheses. * p<.10, ** p<.05, *** p<.01. Reference group: men (when both sex are included), primary school, Italian citizen, working full-time, living in the Centre Italy, non-atypical worker, working in the public sector. The RESET test is conducted testing the joint significance of the square of fitted values.

Appendix B: Additional results and tables

Table 10 Variables codebook

Variable	Description
Women	=1 if the individual is a women
North	=1 if the individual lives in North Italy
South	=1 if the individual lives in South Italy (including Sardinia and Sicilia)
Immigrant	=1 if the individual does not have Italian citizenship
Private	=1 for the private sector
Atypical	=1 if the individual has an atypical contract. Atypical refers to workers with particular contracts that are classified under self-employment, in particular those named "*Collaborazione coordinata e continuativa*"
Part-time	=1 if the individual has a part-time job
Secondary	=1 if the individual has completed the secondary level of education (high school)
Age	Age of the individual, in years
Contribution	Years of social contributions paid up to the current year

Table 11 Sample means of explanatory variables, by regression group, sample corresponding to the initial population

Group	Women	Nord	South	Immi-grant	Private	Atypical	Part-time	Secon-dary	Age	Contri-butions
Men, not graduated, employees	0.0%	47.5%	29.0%	5.5%	80.0%	0.0%	3.7%	54.0%	40.5	18.3
Men, graduated, employees	0.0%	48.5%	26.2%	3.0%	50.8%	0.0%	2.5%	0.0%	43.1	16.9
Women, not graduated, employees	100.0%	54.2%	20.8%	5.6%	72.9%	0.0%	24.1%	65.8%	39.8	15.8
Women, graduated, employees	100.0%	45.8%	27.0%	3.8%	38.2%	0.0%	13.8%	0.0%	40.3	14.1
Graduated, self-employed	41.2%	48.1%	26.0%	3.3%	94.0%	19.2%	12.5%	0.0%	40.2	12.8
Men, not graduated, self-employed	0.0%	47.0%	28.2%	2.9%	99.5%	3.5%	3.8%	56.1%	42.6	19.5
Women, not graduated, self-employed	100.0%	51.4%	23.7%	3.3%	98.5%	10.3%	18.1%	58.0%	41.2	16.5
Total	40.8%	49.4%	26.0%	4.7%	78.4%	1.9%	11.0%	49.2%	40.7	17.1

Table 12 Sample means of explanatory variables, by regression group, longitudinal sample

Group	Women	North	South	Part-time	Secon-dary	Age	Contri-butions
Men, not graduated, employees	0.0%	47.4%	28.9%	3.7%	53.0%	40.4	18.9
Men, graduated, employees	0.0%	48.5%	27.5%	3.2%	0.0%	42.6	16.0
Women, not graduated, employees	100.0%	54.7%	20.0%	24.5%	65.1%	39.8	16.3
Women, graduated, employees	100.0%	48.1%	25.7%	13.5%	0.0%	40.3	13.8
Graduated, self-employed	37.9%	48.7%	25.7%	12.7%	0.0%	42.0	14.9
Men, not graduated, self-employed	0.0%	49.0%	26.5%	3.6%	54.4%	43.4	21.0
Women, not graduated, self-employed	100.0%	51.1%	22.4%	16.1%	55.3%	42.5	18.6
Total	40.16%	49.94%	25.38%	10.92%	48.53%	40.9	17.9

Note: the estimates refer to the sample of those observed for at least one wave.

Table 13 P-values from the t-test of equal means in the panel and in the cross-section (initial population)

Group	Women	North	South	Part-time	Secondary	Age	Contri-butions / Years spent in paid work
Men, not graduated, employees	-	0.876	0.811	0.929	0.147	0.436	**0.000**
Men, graduated, employees	-	0.983	0.419	0.245		0.202	**0.030**
Women, not graduated, employees	-	0.517	0.191	0.598	0.394	0.913	**0.000**
Women, graduated, employees	-	0.170	0.358	0.820	-	0.947	0.389
Graduated, self-employed	0.086	0.733	0.876	0.875	-	**0.000**	**0.000**
Men, not graduated, self-employed	-	0.078	0.097	0.520	0.109	**0.000**	**0.000**
Women, not graduated, self-employed	-	0.863	0.373	0.140	0.109	**0.000**	**0.000**
Total	0.130	0.220	0.103	0.689	0.087	0.011	**0.000**

Note: the null hypothesis of the test is that the mean of the variable is equal in the two samples, against the alternative that it is different. The test is carried out using the Stata™ command *ttest*, allowing for unequal variance in the two samples. Estimates for the standard error should take into account that some individuals belongs to both the panel and the cross-section. However, privacy rules in Eu-Silc do not allow us to find them in the sample. Therefore this table ignores this issue.

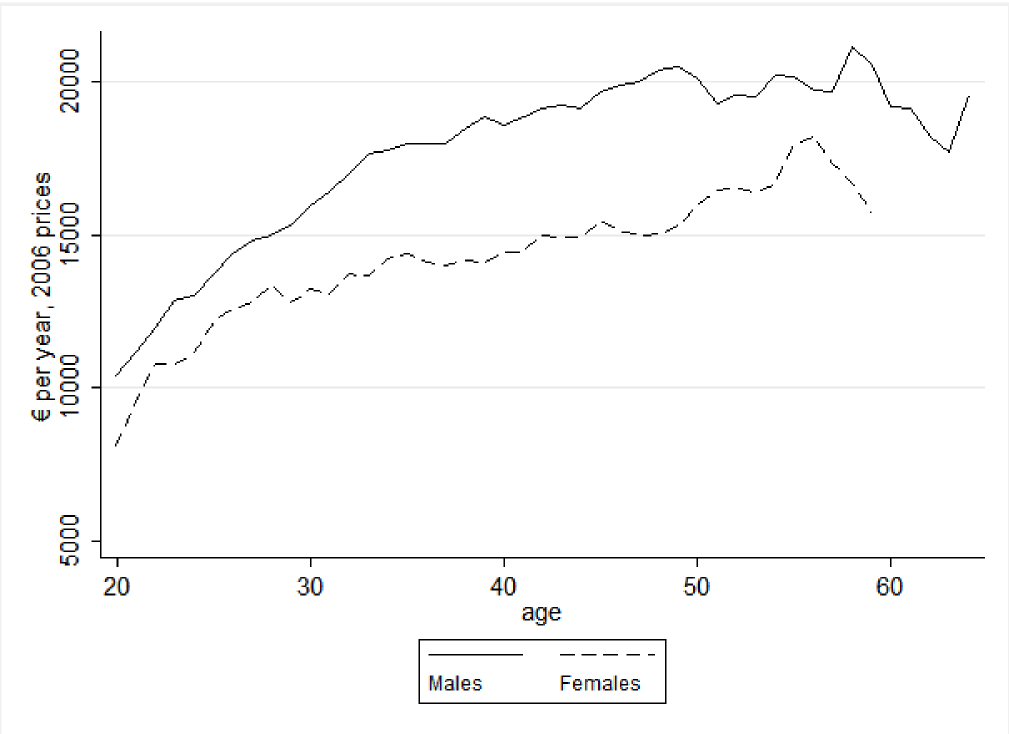
Table 14 Pattern for individuals included in the panel (employees)

Men, not graduated, employees			Women, not graduated, employees			Men, graduated, employees			Women, graduated, employees		
Pattern	Perc.	Obs.	Pattern	Perc.	Obs.	Pattern	Perc.	Obs.	Pattern	Perc.	Obs.
...11	18.4%	1909	...11	17.8%	261	...11	17.6%	1377	...11	18.0%	313
..111	13.4%	1389	..111	12.1%	177	..111	12.8%	1003	..111	14.2%	246
1111.	11.8%	1219	1111.	10.4%	153	1111.	10.9%	848	.1111	9.5%	164
.1111	11.5%	1187	.1111	9.4%	138	.1111	9.8%	767	1111.	9.5%	164
...1.	6.4%	663	...1.	8.4%	123	...1.	7.2%	565	...1.	8.9%	154
.1...	4.9%	502	1	7.4%	109	1	5.5%	429	1	7.7%	134
1....	4.8%	500	..1..	5.6%	82	..1..	5.2%	409	..1..	5.0%	87
..1..	4.8%	494	.1...	5.4%	79	.1...	5.2%	404	.1...	5.0%	86
....1	4.4%	458	..11.	5.1%	75	1....	5.1%	398	..11.	4.8%	83
..11.	4.3%	449	.111.	4.0%	59	..11.	4.8%	378	.111.	3.4%	59
.111.	3.5%	360	1....	3.5%	52	.111.	3.5%	276	1....	3.4%	59
.11..	2.9%	299	.11..	2.5%	37	.11..	3.3%	255	.11..	2.7%	46
11...	2.5%	258	111..	2.2%	32	11...	2.4%	190	11...	2.0%	35
111..	2.4%	245	11...	1.8%	27	111..	2.3%	182	111..	1.4%	24
..1.1	1.0%	108	..1.1	1.0%	14	..1.1	1.1%	87	..1.1	1.3%	22
.1.11	0.7%	67	.11.1	1.0%	14	.11.1	0.6%	47	.1.1.	0.8%	14
.11.1	0.6%	60	11.1.	0.6%	9	.1.11	0.6%	45	1.11.	0.6%	11
11.1.	0.5%	49	1.1..	0.5%	7	1.11.	0.6%	43	.1.11	0.5%	9
1.11.	0.5%	47	1.11.	0.5%	7	11.1.	0.5%	37	.11.1	0.5%	9
.1.1.	0.4%	37	.1.11	0.4%	6	.1.1.	0.4%	30	1..1.	0.3%	5
.1..1	0.2%	20	.1.1.	0.3%	4	1.1..	0.2%	16	.1..1	0.2%	4
1.1..	0.2%	20	.1..1	0.1%	2	.1..1	0.2%	14	1.1..	0.2%	4
1..1.	0.2%	17	1..1.	0.1%	1	1..1.	0.2%	13	11.1.	0.2%	4
All	100.0%	10357	All	100.0%	1468	All	100.0%	5871	All	100.0%	1736

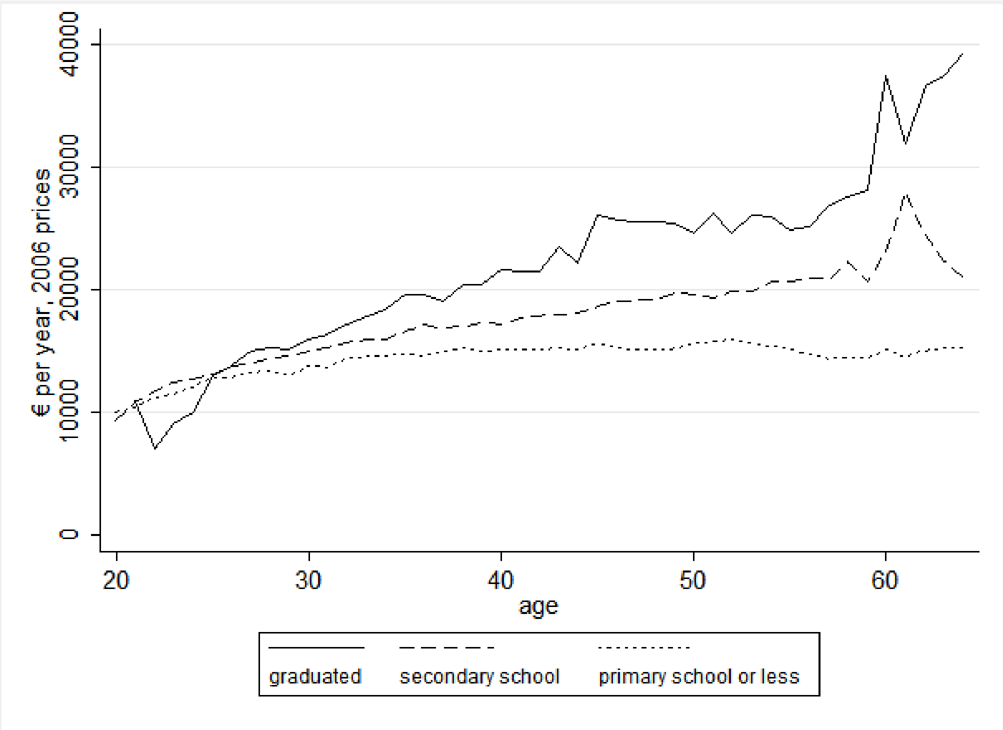
Table 15 Pattern for individuals included in the panel (self-employed)

Graduated, self-employed			Men, not graduated, self-employed			Women, not graduated, self-employed		
Pattern	Perc.	Obs.	Pattern	Perc.	Obs.	Pattern	Perc.	Obs.
...11	17.5%	236	...11	17.2%	698	...11	15.2%	284
...1.	12.8%	173	..111	12.2%	496	...1.	12.2%	227
..111	9.7%	131	...1.	10.0%	408	..111	8.5%	159
....1	9.6%	129	.1111	9.8%	399	.1...	7.9%	147
.1...	7.2%	97	1111.	9.8%	398	1111.	7.9%	147
1111.	7.1%	96	..1..	5.8%	236	..1..	7.8%	146
.1111	6.7%	90	1	5.6%	228	.1111	6.8%	127
..11.	6.2%	84	.1...	4.9%	197	1....	6.8%	127
..1..	5.6%	75	..11.	4.7%	190	1	6.0%	111
1....	4.3%	58	1....	4.4%	177	.111.	3.8%	71
.111.	3.2%	43	.111.	3.3%	135	..11.	3.7%	69
11...	2.3%	31	.11..	2.7%	111	.11..	2.6%	49
111..	1.7%	23	11...	2.5%	102	11...	2.6%	49
.11..	1.6%	22	111..	1.8%	72	..1.1	1.8%	34
..1.1	1.0%	14	..1.1	1.0%	41	111..	1.8%	33
.1.1.	0.8%	11	.1.11	0.8%	32	.1.1.	1.0%	18
.11.1	0.7%	9	.11.1	0.8%	31	.1.11	0.8%	15
1.11.	0.7%	9	1.11.	0.8%	31	11.1.	0.8%	15
.1.11	0.4%	5	11.1.	0.6%	23	1.1..	0.6%	12
11.1.	0.4%	5	.1.1.	0.5%	21	.11.1	0.6%	11
.1..1	0.3%	4	1.1..	0.4%	16	.1..1	0.4%	8
1..1.	0.2%	3	.1..1	0.4%	15	1..1.	0.2%	3
1.1..	0.2%	3	1..1.	0.2%	6	1.11.	0.1%	2
All	100.0%	1351	All	100.0%	4063	All	100.0%	1864

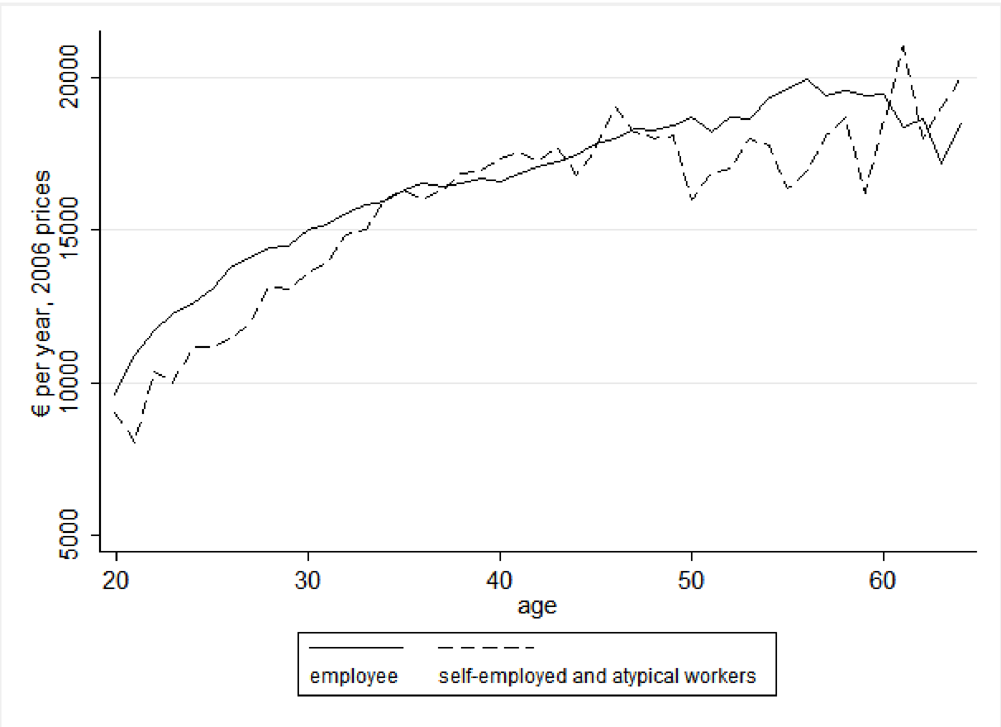
Graph 10 Median annual net earnings, by age and sex, euro 2006, longitudinal sample



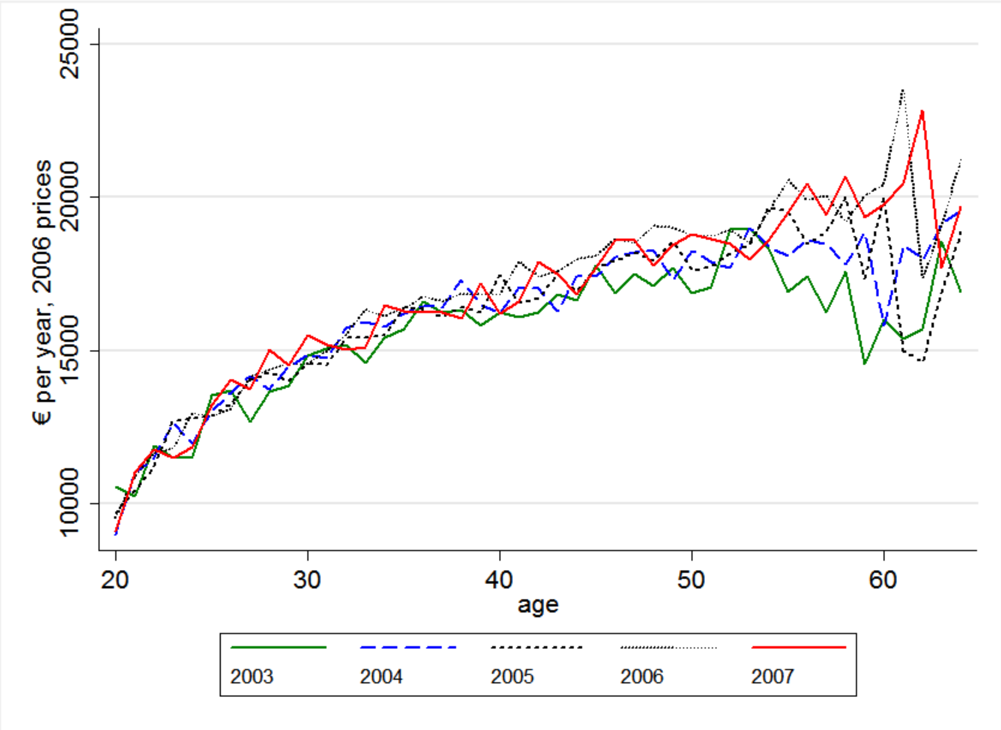
Graph 11 Median annual net earnings, by age and schooling, euro 2006, longitudinal sample



Graph 12 Median annual net earnings, by age and profession, euro 2006, longitudinal sample



Graph 13 Median annual net earnings by age and year, euro 2006, longitudinal sample



References

Bækgaard, H. (2002). Modelling the Dynamics of the Distribution of Earned Income. *NATSEM Technical Paper, no. 24*.

Bækgaard, H., King, A., & Robinson, M. (1999). Dynamod-2: an Overview. *NATSEM Technical Paper No. 19*.

Baltagi, B., & Li, Q. (1991). A transformation that will circumvent the problem of autocorrelation in an error-component model. *Journal of Econometrics, 48*, 385-393.

Baltagi, B., & Wu, P. (1999). Unequally Spaced Panel Data Regressions with AR(1) Disturbances. *Econometric Theory, 15*(6), 814-823.

Becker, G. (1994). *Human Capital: A Theoretical and Empirical Analysis with Special Reference to Education (3rd Edition).* The University of Chicago Press.

Blanchet, D., Buffeteau, S., Crenner, E., & Le Minez, S. (2010). Le modèle de microsimulation Destinie 2: principales caractéristiques et premiers résultats. *Institut National de la Statistique et des Études Économiques Document de travail, G 2010 / 13*.

Borella, M. (2004). The Error Structure of Earnings: an Analysis on Italian Longitudinal Data. *Rivista Internazionale di Scienze Economiche e Commerciali, 51*(3), 405-422.

Borella, M., & Coda Moscarola, F. (2006). Distributive Properties of Pensions Systems: a Simulation of the Italian Transition from Defined Benefit to Defined Contribution. *Giornale degli Economisti e Annali di Economia, 65*(1), 95-125.

Borella, M., & Coda Moscarola, F. (2009). Microsimulation of pension reforms: behavioural versus nonbehavioural approach. *Center for Research on Pensions and Welfare Policies Working Paper, 86/09*.

Brugiavini, A., & Peracchi, F. (2003). Social Security Wealth and Retirement Decisions in Italy. *LABOUR, 17*, 79-114.

Cohen, M. (1991). Statistical Matching and Microsimulation Models: Improving Information for Social Policy Decisions. *The Uses of Microsimulation Modeling, 2,* 62-85.

Dekkers, G., Desmet, R., & De Vil, G. (2010). The long-term adequacy of the Belgian public pension system: An analysis based on the MIDAS model. *Federal Planning Bureau Working Paper, 10/10*.

Emmerson, C., Reed, H., & Shephard, A. (2004). An Assessment of PenSim2. *The Institute of Fiscal Studies Working Paper, WP04/21*.

Favreault, M., & Caldweel, S. (1998). Assessing Distributional Impacts of Social Security by Cohort, Gender, Family Status, Lifetime Earnings and Race Using Dynamic Microsimulation. *Paper presented at the annual meeting of the Association for Public Policy and Management, New York City*.

Giarda, E. (2007). The worsening of wage expectations in Italy: a study based on administrative data. *International Journal of Manpower, 29*, 64-87.

Goldberger, A. (1991). *A Course in Econometrics.* Cambridge, Massachusetts: Harvard University Press.

Harding, A. (1990). Dynamic microsimulation models: problems and prospects. *STICERD Discussion Paper, WSP/48*.

Harding, A. (1996). Microsimulation and Public Policy. *Contributions to Economic Analysis, 232*.

Harding, A. (2007). Challenges and Opportunities of Dynamic Microsimulation Modelling. *Plenary paper presented to the 1st General Conference of the International Microsimulation Association, Vienna, 21 August 2007, available at http://www.euro.centre.org/ima2007/programme/papers/Challenges_Harding_Paper.pdf*.

Heckman, J., Lochner, L., & Todd, P. (2003). Fifty Years of Mincer Earnings Regressions. *IZA Discussion Paper, No. 775*.

Lillard, L. A., & Willis, R. J. (1978). Dynamic Aspects of Earning Mobility. *Econometrica, 46*(5), 958-1012.

Lovell, M. (1983). Data Mining. *The Review of Economics and Statistics, 1*, 1-12.

Maitino, M., & Sciclone, N. (2009). Assessing the Implications of Population Ageing on Tuscan Wellbeing: a Microsimulation Approach. *unpublished, available online at http://www.irpet.it/storage/eventoallegato/436_Relazione%20Londra%20Maitino_Sciclone%20271109.pdf (last access: 30/03/11)*.

Mincer, J. (1974). *Schooling, Experience, and Earnings.* Columbia University Press.

O'Donoghue, C. (2001). Dynamic Microsimulation: A Methodological Survey. *Brazilian Electronic Journal of Economics, 4*(2).

Pudney, S. (1992). Synamic Simulation of Pensioner's Incomes: Methodological Issues and a Model Design for Great Britain. *Dept. of Applied Economics Microsimulation Unit Discussion Paper, No. MSPMU 9201*.

Ramos, X. (2003). The Covariance Structure of Earnings in Great Britain: 1991-1999. *Economica, 70*(278), 353-374.

Ramsey, J. (1969). Tests for Specification Errors in Classical Linear Least-Squares Regression Analysis. *Journal of the Royal Statistical Society. Series B (Methodological), 31*, 350-371.

Ridder, G., & Moffit, R. (2006). The econometrics of data combination. *Handbook of Econometrics, 6*.

Santos Silva, J., & Tenreyro, S. (2006). The Log of Gravity. *The Review of Economics and Statistics, 88*(4), 641-658.

Spence, M. (1973). Job Market Signaling. *The Quarterly Journal of Economics*, *87*(3), 355-374.

Weiss, A. (1995). Human capital vs. signalling explanations of wages. *Journal of Economic Perspectives*, *9*(4), 133-154.

Wooldridge, J. (2002). *Econometric Analysis of Cross Section and Panel Data.* Cambridge, Massachusetts: The MIT Press.