I Z A Institute
of Labor Economics
Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 13459

Exploratory Data Analysis on Large Data Sets: The Example of Salary Variation in Spanish Social Security Data

Catia Nicodemo
Albert Satorra

JULY 2020

IZA Institute of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 13459

Exploratory Data Analysis on Large Data Sets: The Example of Salary Variation in Spanish Social Security Data

Catia Nicodemo
University of Oxford and IZA

Albert Satorra
Universitat Pompeu Fabra and BGSE

JULY 2020

ISSN: 2365-9793

IZA – Institute of Labor Economics

Schaumburg-Lippe-Straße 5–9
53113 Bonn, Germany

Phone: +49-228-3894-0
Email: publications@iza.org

www.iza.org

ABSTRACT

Exploratory Data Analysis on Large Data Sets: The Example of Salary Variation in Spanish Social Security Data[*]

New challenges arise in data visualization when a sizable database is used in the analysis. With many data points, classical scatterplots are non-informative due to the cluttering of points. On the contrary, simple plots such as the boxplot that are of limited use in small samples, offer great potential to facilitate group comparison in the case of an extensive sample. This paper presents Exploratory Data Analysis (EDA) methods that are useful when a large dataset is involved. The EDA methods, (introduced by Tukey in his seminal book of 1977) encompass a set of statistical tools aimed to extract information from data using simple graphical tools. In this paper, some of the EDA methods like the Boxplot and Scatterplot are revisited and enhanced using modern graphical computational devices (as, e.g., the heat-map) and their use illustrated with Spanish Social Security data. We explore how earnings vary across several factors like age, gender, type of occupation and contract and in particular, the gender gap in salaries is visualized in various dimensions relating to the type of occupation. The EDA methods are also applied to assessing competing regressions with earnings as the dependent variable. The methods discussed should be useful to researchers to assess heterogeneity in data, across group-variation, and classical diagnostic plots of residuals from alternative models fits.

JEL Classification: C55, J01, J08, Y10, C80
Keywords: EDA Analysis, large dataset, ggplot, heat-maps, R

Corresponding author:
Catia Nicodemo
University of Oxford
Centre for Health Service and Economics Organisation
Nuffield Department of Primary Health Care Science
Radcliffe Observatory Quarter
Woodstock Road, Oxford, OX2 6GG
United Kingdom
E-mail: catia.nicodemo@economics.ox.ac.uk

[*] We are grateful to the editor, the reviewers and Stuart Redding for his insightful comments, and the participants at the Big data conferences at UAB 2017 and in Lisbon 2017.

1 Introduction

The topic of exploratory data analysis (EDA) as a distinctive tool in applied statistics was created by John W. Tukey in his 1977 book "Exploratory Data Analysis". That book renewed the topic of descriptive statistics and enlightened three main strategies that have become crucial in modern data sciences: 1) graphical presentation; 2) flexibility in viewpoint; and 3) intensive search for simplicity. The methods of EDA do not present p-values or standard errors, rather they focus on the sharp visualization of key aspects of the data at hand. Tukey's developments of EDA focused on robust statistics and strategic graphical displays. In this paper we focus on use of the boxplot and the scatterplot in combination with new computational tools for graphical display (heat-maps).

A major feature of modern applied statistics work is the widespread use of graphical displays of the data. This has been grounded on the methods of EDA developed by Tukey in the late 80s, together with the advancement of graphical capabilities in computer sciences. The discipline of graphical displays has evolved as an entire discipline of statistics (e.g. Chamber et al. (1983), Cleveland (1993), Downey (2014), Myatt and Johnson (2014), Hoaglin et al. (1983), Healy and Enns (2002)). All current statistical software (SPSS, Stata, etc.) has ways to display data that were not present just a few years ago.

This paper revisits EDA tools, applying the boxplot and the scatterplot to databases with many observations. We focus on methods that assess variation across groups, point to outliers, disclose clustering of cases, and highlight non-linearities in relationships between variables. The role of EDA is to explore graphically data in ways that could reveal structural secrets, and to gain new, often unsuspected, insight into the data. The EDA methods discussed align with recent arguments that graphs usually create "pre-attentive" visual processing in the brain, helping to focus on the important message (Hussain and Prieto (2016), Camacho et al. (2015), Healey and Ennes 2002). Wattenberg et al. (2011) suggest that an ideal visualization should not only communicate key facets of the data, but also stimulate viewer engagement and attention. Cheng et al. (2013) and Schwabish (2014) suggest that researchers who want to disseminate their research to a wider audience of non-specialists, should think carefully about how to construct effective graphics. Similar recommendations are made in the papers of Jebb et al. (2017), and Varian (2014). In reflecting on statistics for management, George et al. (2014) conclude that with large datasets it is too easy to get false correlations when using typical statistical tools. The EDA methods can produce a set of visualizations that simplify the understanding of complex data, which will be increasingly useful to researchers in management, business and social sciences in general, that often face the challenges associated with using large databases in their research.

The classical inferential methods of statistics become less useful in very large samples, since test results are nearly always significant, and the standard errors are very small. EDA could help in this case by emphasizing data visualization and dimension reduction methods. Traditional descriptive statistical analysis faces problems with large datasets. For example, a bi-variate relationship that would be clear in a basic scatterplot with relatively few observations may not be readily apparent in a dataset containing a million observations (for example, see Figure 1 where a simple scatterplot is used). A large part of the literature on big data concentrates on computer-intensive methods, such as machine learning or regression trees, that emphasize prediction (for more details about these tools see Qiu et al. (2016), Al-Jarrah et al. (2015). In our study rather than big data we concentrate on the case of large database (large sample size), see De Mauro et al. (2017) for terminology on big data. In contrast, EDA methods focus on methods that serve to explain and describe data, and should be among the first steps when analysts want to explore large datasets. We believe there is gap in the literature of big data analysis on the subject of exploring and presenting statistical features of large datasets. This paper focuses on methods that can fill part of this gap. Our main contribution is to show how traditional simple methods should be used in the context of the new paradigm of big data. It should be recognized that descriptive tools do not give yes/no answer to basic research questions, unlike the classical testing used in econometrics. However, the EDA methods have the ability to show shifts, heterogeneity, outliers and non-linearities among variables, without the requirement of model assumptions. EDA methods should be seen as complementary, not only prior to the classical econometric methods, but also to give support or rejection to a priori posed econometric models. We show how EDA methods are useful also after a regression fit in residual diagnostic plots.

To better convey the practicalities of the methods proposed, we illustrate them by analysing variation of salaries in the Spanish Social Security database. In particular, we address issues such as whether the observed difference in wages are explained by: the rigidity of labour market (temporary versus permanent contract); discrimination (women versus men); and age (old versus young)? Previous work involving the Spanish labour market has focused on fitting models to test aspects of labour economic theories (e.g., Dolado et al. 2002, Gehrke and Weber (2018)). In this paper we deviate from this testing approach in favour of descriptive methods that permit direct, visual assessments of the questions explored. The structure of the paper is as follows. Section 2 describes the database. In Section

3 we discuss group variation in wages using the boxplots. Section 4 discusses the use of heat-map scatterplots. In Section 5 we apply the heat-map approach to a classical residual diagnostic plot, and the comparison between OLS and Tobit regression. Finally, in the last Section conclusions are presented.

——— FIGURE 1 HERE ———

2 Social Security Data

To illustrate the EDA methods we use the Spanish Social Security (SS) data on labour, specifically, the "Muestra Continua de Vidas Laborales" (hereafter, MCVL) in 2010.[1] The data comes from the register of the Social Security System (SS) for people active in the labour market. Starting in 2004, social security records have been released for a 4% non-stratified random sample of the population who in that year have had any relationship with Spanish SS (individuals who are working, receiving unemployment benefits, or receiving a pension). Given the structure and magnitude of the MCVL, this is a useful dataset with which to illustrate the EDA approach we advocate. The rest of this section is devoted to describing briefly the statistics and economic labour context of this database. The data set gives information regarding historical relationships of individuals with the SS relating to work, unemployment benefits and pensions, for around one million observations each year. It also contains information regarding the type of contract, sector of activity, qualifications, earnings, date of entering or leaving the job market, part-time or full-time status and firm size. The MCVL also provides individual characteristics such as age, gender, residence, country of birth and level of education. Information on educational attainment has improved in recent editions of the MCVL. The main outcome variable of interest in our data set is the earnings of Spanish people. The MCVL only provides information on the "social contribution base" (censored earnings), which captures monthly labour earnings plus 1/12 of bonuses received over the year. The censored earnings variable has minimum and maximum values, which vary over time.[2]

We use the wave of 2010, where 722,957 individuals were included. To make easy the representation of this database we consider only wages for people aged between

[1] More information here http://www.seg-social.es/prdi00/groups/public/documents/binario/190489.pdf

[2] For more details see http://www.empleo.gob.es/es/guia/pdfs/EVOLUCIxN_DE_LAS_BASES_MxNIMAS_Y_MxXIMAS_DE_COTIZACIxN_2015.pdf

15 and 65 years old. We exclude pensioners, the self-employed, individuals receiving unemployment benefits, individuals who report strange earning values (outside the minimum and maximum values), and individuals with missing information in relevant variables (age, education, type of contract, occupation). This leaves a sample of 541,457 individuals. We calculate the daily wage as our main earnings measure, computed as the ratio between the monthly contribution base and the number of days worked in that particular month. If the individual records more than one job at the same time we sum the earnings. Others years and variables of course could be consider, but this will be beyond the scope of this paper which focus is not to explore the Spanish labour market.

We analyse the variation of wage with respect to variables: age, gender, type of contract (fixed (permanent) or temporary), education (primary, secondary and tertiary) and occupation (high, medium and low skilled). These variables are traditionally considered in the literature of labor markets (Heckman et al. (2006)). These are variables that are found to be significant in a regression analysis of wages reported below. Note however that significance within such a large sample does not provide reliable evidence on the substantive relevance of these variables. As such, visual inspection of the variation of wage with respect to those variables is necessary. We extract information about qualified and non-qualified employees, splitting the sample by qualification: high, middle and low skill level, according to the ISCED classification and following the classification proposed by García-P èrez (2008) for the MCVL data. We consider the type of contract held (fixed or temporary), and we construct five age groups defined by the following intervals: (15, 25], (25, 35], (35, 45], (45, 55] and (55, 65].[3]

Table 1 reports descriptive statistics of the wage by gender, occupation, education, type of contract and age.

Older people generally experience higher wages. Workers with a fixed contract and highly skilled jobs also have increased wages, as do workers who have reached higher education levels. This table shows a wage gap between male and female workers of around 17.50%. This observation of a gender gap in the Spanish labour market is well documented, (see e.g. Amuedo-Dorantes and De La Rica (2006)).

The next sections illustrate the use of EDA for assessing the variation of wages across workers' characteristics.

[3] We grouped the age using 10 years interval. Usually this correspond also to the structure of labour market as we can see trough the graphs.

——— Table 1 HERE ———

3 The comparative boxplots: log-wage by age groups

The boxplot was introduced by Tukey (1977) as a way to present graphically the distribution of a continuous variable, and also to display the variation of a continuous variables across groups. Boxplots display the first, second and third quartile, the interquartile range and outliers of a database. The information displayed by the boxplot, and most of its variations, is based on the data's median. The 50% central bulk of the data is represented by a box; two whiskers, one at side of the box, represent the two 25% extremes of the data distribution. A line dividing the box represents the median. Symmetry/skewness of the distribution is visualized graphically by the position of the dividing line of the box: symmetry when the line divides the box in two equal halves, and skewness when one half is larger than the other. The boxplot is very useful for visualizing group variation in a variable: the boxplot of each group is set in parallel (vertically, or horizontally), one besides the other. The boxplot is of limited use when sample size is small, thus displaying variation across groups using the boxplot requires a large sample.

Note that in the case of a very large sample, the F-test of an ANOVA table will tend to be significant (i.e., p-values below the significance level), since any small difference of the population means will be detected as significant due to the large sample size increasing the power of the test. EDA methods and in particular the display of parallel boxplots offer researchers direct visualization of the variation across groups by showing the median and quartile values.

Wages can be presented using raw values or after a log transformation. The advantage of the log transformation for group comparison is the approximate normality of the distribution, hence the comparative boxplots will have eliminated the skewness of the wage distribution and make it easier to compare the ends of the wage scale (see Hubert and Vandervieren (2008)). We use the log base 10 transformation. It should be noted, that once a researcher finds interesting variation at the log scale, they can easily display the same data in the original scale.

Figure 2 shows parallel boxplots of log-wage for different age groups. For each group, the box and the whiskers span the whole variation of log-wage in the group.

The edges of the box are the first and third quartile; the median, is the dividing line of the box. Points exceeding the solid line of each of the whiskers are cases that could be categorized as outliers.[4] The y-axis is the variable, common to all the boxplots, thus the variation on the level of the dividing line shows graphically the variation of the median across groups. We see that for young people, the distribution of log-wages is not symmetric (the median line does not divide the box in two halves), but near symmetry is observed in the other groups.

A feature we have added to the standard boxplot display is to make the width of the box proportional to the size of the group (number of individuals in each group) in the whole sample. This makes it easy to assess the relevance of a specific group; for example, we see that the age group (55,65) represents a smaller proportion of workers than the age group (35,45).

—— FIGURE 2 HERE ———

The following findings can be drawn from this diagram:

1. There is a slight curvilinear increase of the median level of log-wage with age, with the first age group (15,25) showing the lowest median log-wage and different from the other groups.

2. There is slight variation across the groups in the dispersion of log-wage, as measured by the inter-quartile range (the length of the box). The highest dispersion arises in the two most extreme age groups.

3. The width of the boxplots shows the proportion of the population in the group, and we see that the groups (25,35) and (35,45) are the widest.

4. While there is a significant increase of wage when going from group (15,25) to group (25,35) (the youngest groups), there is a slight decrease in the median of log-wage when moving from (45,55) to (55,65).

5. The log transformation normalizes (symmetrizes) the distribution for all groups, except for the youngest group that still shows a long-tail on the right. This asymmetry for the salary of young could raise policy discussion.

[4] Several approaches for points to be declared the status of outliers have been in the literature (see e.g., Bruffaerts et al. (2014).)

These findings are common to previously established economic empirical research (Cabrales et al. (2017)), which states that skills and experience (age would be a surrogate) impact positively on earnings. We now consider group variation when we cross two or more categorical variables.

3.1 Variation of log-wage by age groups and a third categorical variable

Log-wage by age can vary when controlling for different additional variables, such as occupation, level of education, type of employment, and gender. Figure 3 displays variation of log-wage on age for each of those mentioned variables.

——— FIGURES 3 HERE ———

Panel *a* shows that there is a substantial gender gap in wages for all age groups except for the two youngest. The gap reaches a maximum of 24 points (of log-wage) for those aged 45-55. The graph shows that the median (log) salary gap between men and women increases dramatically with age, although it is important to note that we do not control for other variables.

Panel *b* of Figure 3 depicts the variation of log-wage on age for occupations classified as low, middle and highly skilled workers. The proportion of people in the different groups is shown at the top of the graph. We observe the middle skills group is the largest and accounts for 63% of all workers. The width of the boxplots represents the relative size of the group. The group of young people (ages 15-25) is the smallest and the one with lowest log-wage, and this happens for the three skill levels. As expected, the group of highly skilled workers is the one with highest earnings.

Panel *c* shows the variation of log-wage on age for the three educational levels: primary, secondary and tertiary. As expected, people with a tertiary level of education receive the highest wages (see the median levels of each group). Note that, the size of the group decreases with age for all education levels (as is illustrated by the width of the box). Furthermore, the plot shows that the highest variation of wage by age occurs in the secondary education level.

Panel *d* shows the variation of log-wage on age for the two types of contract: permanent and temporary. We see that the highest proportion of contracts are permanent, with the highest salaries. In both groups of temporary and fixed contracts, median salaries remain fairly homogeneous across age, though in temporary contracts we see a slight decrease of salary as age increases.

Figure 3 describes, in a four-display graph, the variation of wage conditional on age and one additional variable (specifically, gender, occupation, education or type of contract). Note that one could also consider a five-display graph where each display shows the boxplot relative to one or two of the covariates for each age group. The necessity of such additional graphs would arise from the inspection of the ones at hand. As the main purpose of the paper is to highlight useful EDA devices, we do not comment further on these additional graphs.

This graph allows researchers to assess directly how the distribution of salaries changes with type of contract in age group, or between men and women. Note that this type of comparison, considering variation of wages in groups defined by crossing many categorical variables, will be possible only in the context of large data. Otherwise, the sub-samples will be small and the boxplot will become non-informative. Variation conditional on a third variable will be introduced in the next sub-section.

3.2 Exploring the gender gap in log-wage

In this section we explore the gender wage gap. Reducing the gender wage gap is an important topic on the European political agenda. The persistence of the gender wage gap is the result of direct discrimination against women and/or a structural inequality, such as segregation in sectors or occupations, access to education and training, or biased evaluation and/or pay systems. We present evidence showing the difference between male and female earnings considering several factors that could explain not only the wage gender gap but also the selection of women into jobs with certain characteristics.

In Panel *a* of Figure 3 we see a substantial gender gap in wages for all age groups. The log-wage gaps between males and females persist after controlling for age. The proportion of workers who are young is larger for males than females, whilst among females there is a smaller proportion of older workers (this is seen from the width of the boxplot). There is a wage gap between younger females' wages than wages of all other female age groups. This does not occur for males. The dispersion of wages is similar across age and gender.

—— FIGURE 4 HERE ———

Panels *a* to *c* of Figure 4 shows the variation of log-wage by gender, controlling for additional characteristics of workers: occupational skills, level of education and type of contract.

Panel *a* shows that while the gender wage gap is not apparent in young groups with temporary contracts, it is clear amongst workers with permanent contracts. We also see that few women in the older age group hold permanent contracts.

In panel *b*, we can see that the boxplots of high skill groups are thinner. This reflects official statistics which report a lower proportion of workers in the high skill group. Among high skill workers, the gender wage gap is minimal. The maximum gender gap arises in the middle skills group and is largest among the young. The gender wage gap is also apparent amongst low skill workers but is smaller.

Finally, Panel *c* shows that when controlling for the level of education, a gender gap on wage is still visible at all levels. The gap is at its highest for the second level of education, and greatest for the younger women at this level. There are more women than men in the group defined by lowest age and higher level of education (see the thickness of the boxplot for the tertiary education level for females aged 15 to 25).

Panel *a* shows that while the gender wage gap is not apparent in young groups with temporary contracts, it is clear amongst workers with permanent contract. We also see that few of the women in the higher age group hold permanent contracts.

Panel *b* shows that the high skill groups have thinner boxplots. This reflects official statistics which report a lower proportion of workers in the high skill group. Among high skill workers, the gender wage gap is minimal. The maximum gender gap arises in the middle skills group, and is largest among the young. The gender wage gap persists amongst low skill workers but is smaller.

Finally, Panel *c* shows that when controlling for the level of education, a gender gap on wage is still visible at all levels. The gap is at its highest for the second level of education, and more for the younger women at this level. It is interesting to see that there are more women than men in the group who have the higher level of education and lowest age (see the thickness of the boxplot for the tertiary education level for females aged 15 to 25).

4 Scatterplot: The heat-map

So far, we have assessed the variation of a continuous variable Y (the log of daily wage) by considering groupings based on several categorical variables Xs. One of the Xs we considered was age, split into several age groups. In our database, however, age is a numerical variable, so a simple scatterplot of daily wage Y [5] on the variable age Xcould be applied. This is attempted in Figure 1, where, as commented earlier (in Section 1), the large sample sizes produce a cluttering of points and a non-informative graph. By using modern computational tools of density mapping - possible only in the case of a large sample- this data can be presented in a more informative heat-map scatterplot (for more details about heat-map see Barter and Yu (2018) or Wilkinson and Friendly (2009)). The heat-map scatterplot of daily wage on age is shown in Figure 5. [6] The plot shows that the highest density of observations occurs at age around 30 and salaries around 50 euros, and that there is also a ceiling effect on the daily wage variable that gives rise to a concentration of points. This ceiling effect group will be further discussed below.

——— FIGURE 5 HERE ———

Using the heat-map approach, we can assess the variation of daily wage on age when controlling for other variables, as shown in Figure 6.

——— FIGURE 6 HERE ———

Looking at panel a of Figure 6 we can see that women have lower wages than men. Yet, we observe that there are few older women working. The modal concentration [7] at the top wage level is of lower intensity for women, showing that there is a smaller proportion of women in this ”ceiling effect” group.

Panels *b* and *d* show daily wage variation on age when controlling for occupation and education levels. They show that high skilled workers, and those with tertiary

[5] In the previous section, we used the log transformation of daily wage to symmetrize the distribution. This log transformation is the one mostly used in labor and has the effect of improving the comparativeness of the boxplots displays (Hubert and Vandervieren, 2008).

[6] This is a graph that has been produced using the ggplot2 function of the free software R.

[7] By modal concentration we mean a local high concentration of cases in a given point; note that it does not refer to any summary level measure unlike the median or the mean.

level of education, have the largest earnings, many in the ceiling group. The variation of wages across skill and education levels show, as expected, that wages increase with skill and education. Individuals with primary level of education, or in the low skill group, present larger heterogeneity, not only across wage but also across age. Fitting a nonparametric regression line to each of the scatterplots, we see a steady, near-linear increase of wage with age.

Panel *c* shows the variation of salary by age when controlling for the type of contract. We see that in contrast to permanent (fixed) contracts, workers with temporary contracts show significant wage dispersion, concentrated in a small range of age variation. Furthermore, the ceiling effect (of high salary workers) appears only to impact those with permanent contracts. The modal wage for temporary contracts is lower than the modal wage for permanent contract workers (if we discount the ceiling group, then the modal wage for the permanent contracts intersects with the highest modal wage salaries of the nonpermanent contracts). The results in this section are in line with many studies on the Spanish labour market. For more details see, for example Cabrales et al. (2017).

5 The heat-map scatterplot in regression diagnostics

The relationship between wage and age is now explored using a regression approach. We assume that the expected value of the dependent variables Y is linear on a set of covariates, $X_1, \ldots X_n$. Violation of the assumption of linearity, however, can invalidate the results of regression analysis. A key diagnostic plot used to test this assumption is the residuals-vs-fitted-y plot. In the context of a very large sample, this diagnostic plot will be cluttered by too many points.

In this section we illustrate the use of regression residual plots in the context of our labour data, especially when assessing the impact of age, gender, education and other worker characteristics on our dependent variable, namely the log of the daily wage.

We fit three regression models using the log daily wage as the dependent variable and the covariates listed in the first column Table 2. In the covariates we have categorical variables (represented int he regression by dummies of category) with reference categories detailed in the footnote of the Table. The second column of this table shows the estimates for the fitted OLS regression using the all sample, the third column reports the Tobit regression, and the fourth column reports the OLS regression excluding all the cases where the log daily wage at the ceiling point 4.506. As expected, given the large sample, all the regression coefficients are all statistically significant. We have not reported in the table the standard errors since for all the coefficients estimates are 0.00 when rounded at three decimal digits (the rounding we use for all the numbers f the table). The significance has been computed using the robust standard errors to protect for unknown heteroscedasticity.

——— TABLE 2 HERE ———

After a regression fit, it is useful to inspect the residuals-vs-fitted-y-plot. This is the way to asses possible no-linearity and other types of misspecification of the regression equation. Figure 7 reports residuals-vs-fitted y-plot for the three regression model.

——— FIGURE 7 HERE ———

In Panel *a* we note a bimodality of the bivariate distribution, with a high density of points on the right-hand side of fitted *ys* and positive residuals. This is illustrated in the graph by a high density-mass of points in the upper right section of the graph. The nonparametric curve that fits the mean level of the residuals across the fitted values of Y is shown in the graph by a black line. Strict linearity implies that this line should be a horizontal line at $Y = 0$. The simple scatterplot would be very hard to detect problems in the regression, given a large number of points in the graph. In addition to the right censored observations, we see an unimodal distribution with density mass in the center of the bivariate graph.

As mentioned in the section above, earnings in the MCVL data suffer from censoring: any individual with a daily wage above the threshold of 106 euros is recorded to have a wage equal to this threshold. If we eliminate individuals from the regression who have been censored, we observe how the positive residuals disappear in the heat-map of residuals-versus-fitted-y (Panel *b* of Figure 7). Finally, Panel *c* reports the heat-map of fitted versus residuals generated using a Tobit model. Note that a researcher will always be faced with the issue of which of the three fitted models to trust. Standard econometrics would recommend the Tobit regression. The problems of OLS analysis when data is censored have been widely reported (see Blundell and Meghir (1987)). Note that using EDA methods, we identified modal concentrations among the residuals of the fitted OLS regression, which suggests violation of OLS assumptions and potential problems with the estimation.

The alternative estimates shown in Table 2 should be complemented with the residuals shown in Figure 7. The full picture of the fitted model cannot be assessed without the estimated residuals. Note that OLS estimates do not account for the censoring shown in the residuals of Figure 7 (a) (where we see that a line of censored residuals on the top right). The residuals from a Tobit regression shown in Figure 7 (c) have already accounted for the censoring. The censoring is not visible in Figure 7 (b) because the censored cases have been removed from the estimation.

6 Conclusions

The analysis of large data sets is increasingly popular in business and social science research, and there are many new opportunities to extract useful empirical evidence from data. Extensive research in economics has been devoted to the econometrics of regression (or extended regression) models. In this paper we present how EDA methods that focus on graphical displays of data could help to inspect large databases and explore heterogeneity across groups of variables. In particular we focus on the use of boxplots and heat-maps. EDA methods are useful not only as ways to present descriptive statistics, but but can also help as a robustness check in selecting appropriate models to use in regression analysis.

Using Social Security data from Spain, this paper illustrates how EDA can be used to highlight group variation of critical variables in the labour market and how they interact. We use boxplots and heat-map scatterplots to assess the heterogeneity of earnings on basic covariates, such as gender, contract status, experience, skills, and others. We also showed how the heat-map scatterplot can be used for the basic diagnostic plots on regression analysis when the researcher is confronted with a very large data set.

With large databases, there are problems associated with the use of traditional statistical models and EDA offers instruments that can help overcome some of these issues. These methods also complement predictive approaches to big data, such as machine learning, random forest, etc. Although it should be noted that EDA methods have limitations, as they do not permit statistical confirmation or refusal of a causal hypothesis. It is well known, however, that EDA methods have been proved useful in the discovery of later confirmed causal hypothesis.

Many are the possible extensions of the EDA methods discussed in this paper for large database. We have concentrated just on few features that we feel are immediately available to practitioners using the current free software. [8] One area where we are currently working and that we feel is also a promising avenue for research in the case of large samples, is the visual exploration of longitudinal data.

Declaration of interests: No competing interests from all the authors, no financial relationships with any organisations that might have an interest in the submitted work in the previous three years; no other relationships or activities that could appear to have influenced the submitted work. Role of the funding source: This project was not funded. No ethical approval is need. All the authors give the informed consent.

7 References

Al-Jarrah, O.Y., Yoo, P.D., Muhaidat, S., Karagiannidis, G.K. and Taha, K., 2015. *Efficient machine learning for big data: A review.* Big Data Research, 2(3), pp.87-93.

Amuedo-Dorantes, C. and De la Rica, S., 2006. The role of segregation and pay structure on the gender wage gap: evidence from matched employer-employee data for Spain. The BE Journal of Economic Analysis and Policy, 5(1).

Barter, R.L. and Yu, B., 2018. *Superheat: an R package for creating beautiful and extendable heat-maps for visualizing complex data.* Journal of Computational and Graphical Statistics, 27(4), pp.910-922.

Blundell, R. and Meghir, C., 1987. *Bivariate alternatives to the Tobit model.* Journal of Econometrics, 34(1-2), pp.179-200.

Bruffaerts, C., Verardi, V. and Vermandele, C., 2014. *A generalized boxplot*

[8] The code of the implementation in R of the methods of the paper is provided in the supplementary material to this paper.

for skewed and heavy–tailed distributions. Statistics and probability letters, 95, pp.110-117.

Cabrales, A., Dolado, J.J. and Mora, R., 2017. *Dual Labour Markets and (Lack of) On-the-Job Training: Evidence for Spain Using PIAAC Data.* SERIEs, Journal of the Spanish Economic Association, 8, pp.345-371.

Camacho, J., Prez-Villegas, A., Rodrguez-Gmez, R.A. and Jimnez-Maas, E., 2015. *Multivariate exploratory data analysis (MEDA) toolbox for Matlab.* Chemometrics and Intelligent Laboratory Systems, 143, pp.49-57. Chambers JM, Cleveland WS, Kleiner B, Tukey PA. *Graphical Methods for Data Analysis.* Belmont: Wadsworth International Group, pp.94-104.

Cheng, S., Shi, Y., Qin, Q. and Bai, R., 2013, October. *Swarm intelligence in big data analytics. In International Conference on Intelligent Data Engineering and Automated Learning* (pp. 417-426). Springer, Berlin, Heidelberg.

Cleveland WS. Visualizing data. Hobart Press; 1993 Sep 1.

Cleveland, W.S., 1993. *Visualizing data.* Hobart Press.

De Mauro, A., Greco, M. and Grimaldi, M., 2016. *A formal definition of Big Data based on its essential features.* Library Review 65(3):122-135

Downey A. *Think stats: exploratory data analysis.* " O'Reilly Media, Inc."; 2014 Oct 16.

Dolado, J.J., Garcia-Serrano, C. and Jimeno, J.F., 2002. *Drawing lessons from the boom of temporary jobs in Spain.* The Economic Journal, 112(480), pp.F270-F295.

Garcia-Perez, J.I., 2008. *La muestra continua de vidas laborales: una guia de uso para el analisis de transiciones.* Revista de Economia Aplicada, 16(1), pp.5-28.

Gehrke, B. and Weber, E., 2018. *Identifying asymmetric effects of labor market reforms.* European Economic Review, 110, pp.18-40.

George, G., Haas, M.R. and Pentland, A., 2014. *Big data and management.* Academy Management Journal, 321-326.

Healey, C. G., and Enns, J. T., (2002). *Perception and painting: A search for* *effective, engaging visualizations.* IEEE Computer Graphics and Applications, 22(2), pp.10-15.

Heckman, J.; Lochner, L.; Todd, P. (2006). *Earnings Functions, Rates of Return and Treatment Effects: The Mincer Equation and Beyond.* Handbook of the Economics of Education. 1. Amsterdam: North-Holland. pp. 307458.

Hoaglin, D.C., Mosteller, F. and Tukey, J.W. eds., 1983. *Understanding robust and exploratory data analysis* (Vol. 3). New York: Wiley.

Horton JJ, Tambe P. (2015). *Labor economists get their microscope: big data and labor market analysis.* Big data, (3):130-7.

Hubert, M. and Vandervieren, E., 2008. *An adjusted boxplot for skewed distributions.* Computational statistics and data analysis, 52(12), pp.5186-5201.

Hussain, K. and Prieto, E., 2016. *Big data in the finance and insurance sectors. In New Horizons for a Data-Driven Economy* (pp. 209-223). Springer, Cham.

Jebb, A.T., Parrigon, S. and Woo, S.E., 2017. *Exploratory data analysis as a foundation of inductive research.* Human Resource Management Review, 27(2), pp.265-276.

Lazer, D., Kennedy, R., King, G. and Vespignani, A., (2014). *The parable of Google Flu: traps in big data analysis.* Science, 343(6176), pp.1203-1205.

Myatt, G.J., and Johnson W. P. 2014. *Making Sense of Data I: A Practical Guide to Exploratory Data Analysis and Data Mining*, 2nd Edition, John Wiley and Sons

Mayer-Schnberger, V. and Cukier, K., (2013). *Big data: la revolucion de los datos masivos.* Turner

Qiu, J., Wu, Q., Ding, G., Xu, Y. and Feng, S., 2016. *A survey of machine learning for big data processing.* EURASIP Journal on Advances in Signal Processing, 2016(1), p.67.

Rodgers, G.B., 1975. Nutritionally based wage determination in the low-income labour market. Oxford Economic Papers, 27(1), pp.61-81.

Schwabish, J.A., 2014. *An economist's guide to visualizing data.* Journal of Economic Perspectives, 28(1), pp.209-34.

Taylor, L., Schroeder, R. and Meyer, E., (2014). *Emerging practices and perspectives on Big Data analysis in economics: Bigger and better or more of the same?* . Big Data & Society, 1(2),p.205

Tukey, J., (1977). *Exploratory Data Analysis*, Vol. 2. Addisson-Wesley Publishing Company, Inc, USA

Hoaglin, D.C., Mosteller, F. and Tukey, J.W. eds., 1983. Understanding robust and exploratory data analysis (Vol. 3). New York: Wiley.

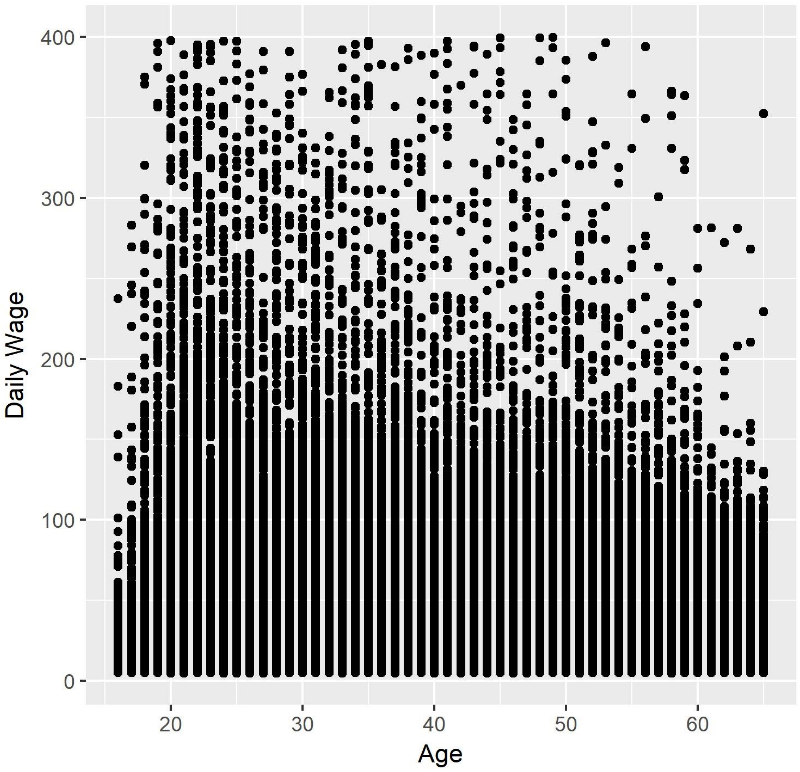
Varian, H.R., 2014. *Big data: New tricks for econometrics.* Journal of Economic Perspectives, 28(2), pp.3-28.

Wattenberg MM, Kriss JH, Viigas FB, (2011) International Business Machines Corp, assignee. *System and method for visually analyzing geographic data.* United States patent US 7,917,852. Mar 29.

Wilkinson, L. and Friendly, M., 2009. *The history of the cluster heat-map.* The American Statistician, 63(2), pp.179-184.

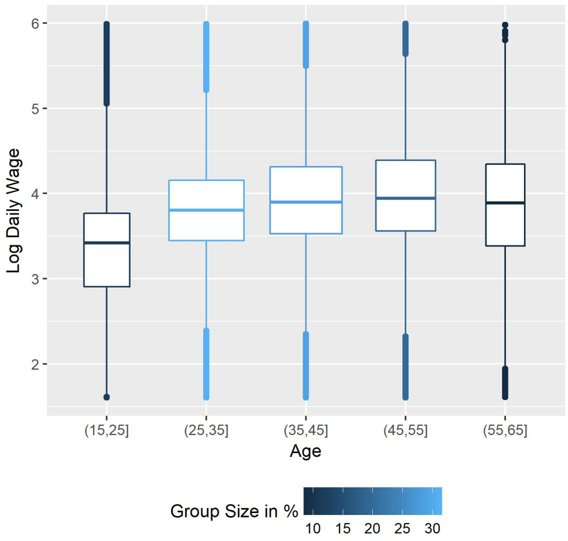
8 Tables and Figures

Figure (1) Example of scatter plot of log-wage by age



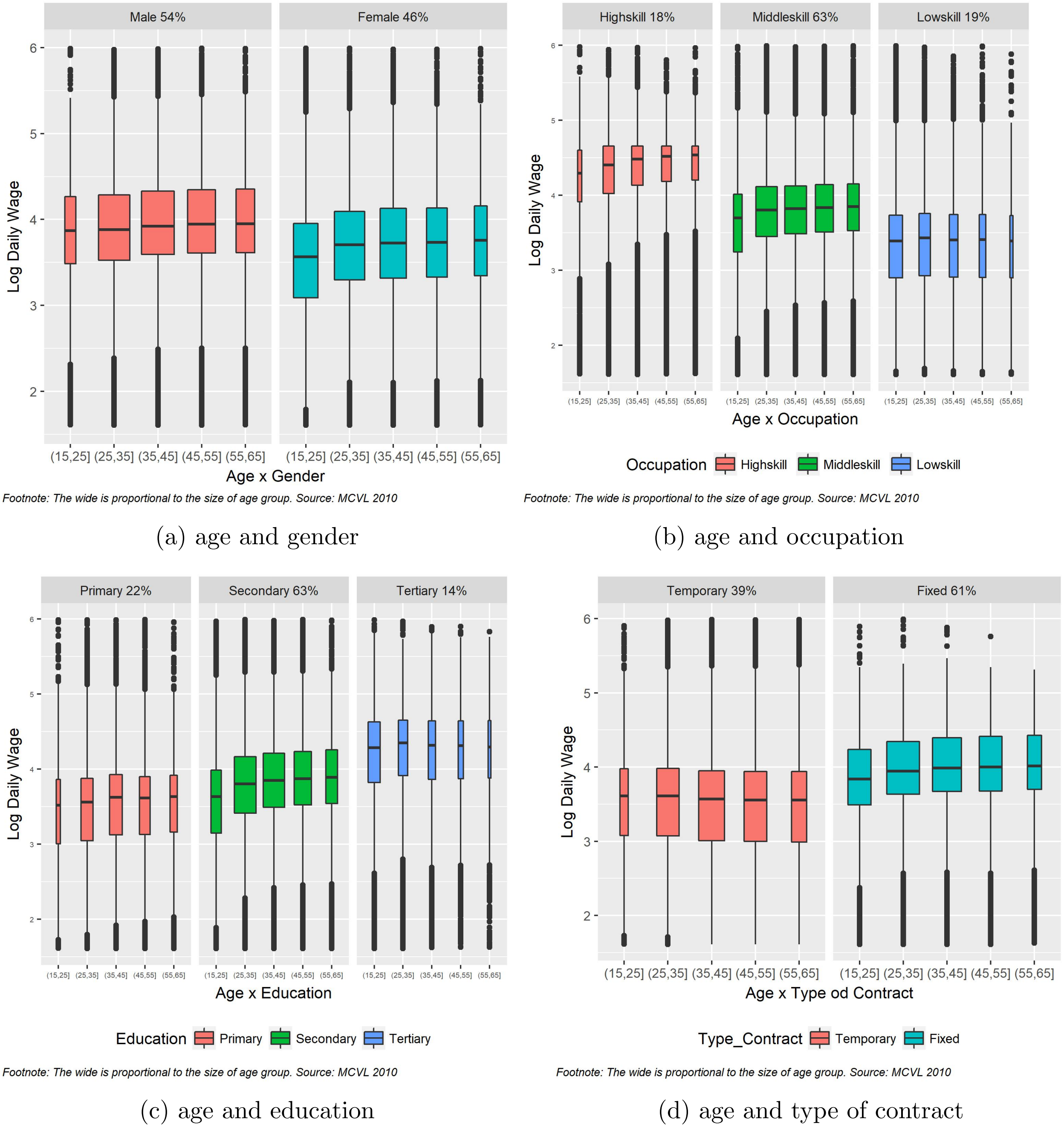
Source: MCVL 2010

Figure (2) Boxplot of log-wage by age



Footnote: The wide is proportional to the size of age group. Source: MCVL 2010

Figure (3) Boxplot of log-wage



Footnote: The wide is proportional to the size of age group. Source: MCVL 2010

(a) age and gender

Footnote: The wide is proportional to the size of age group. Source: MCVL 2010

(b) age and occupation

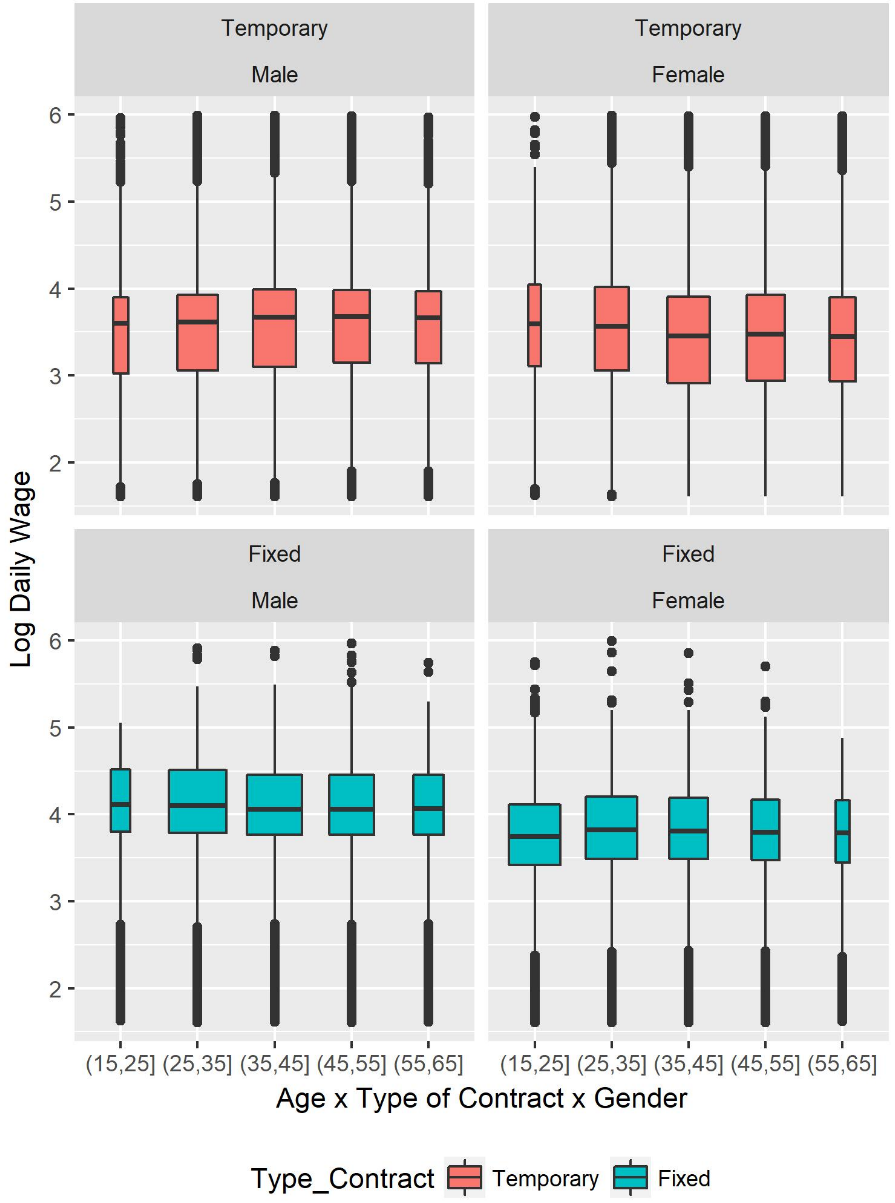
Footnote: The wide is proportional to the size of age group. Source: MCVL 2010

(c) age and education

Footnote: The wide is proportional to the size of age group. Source: MCVL 2010

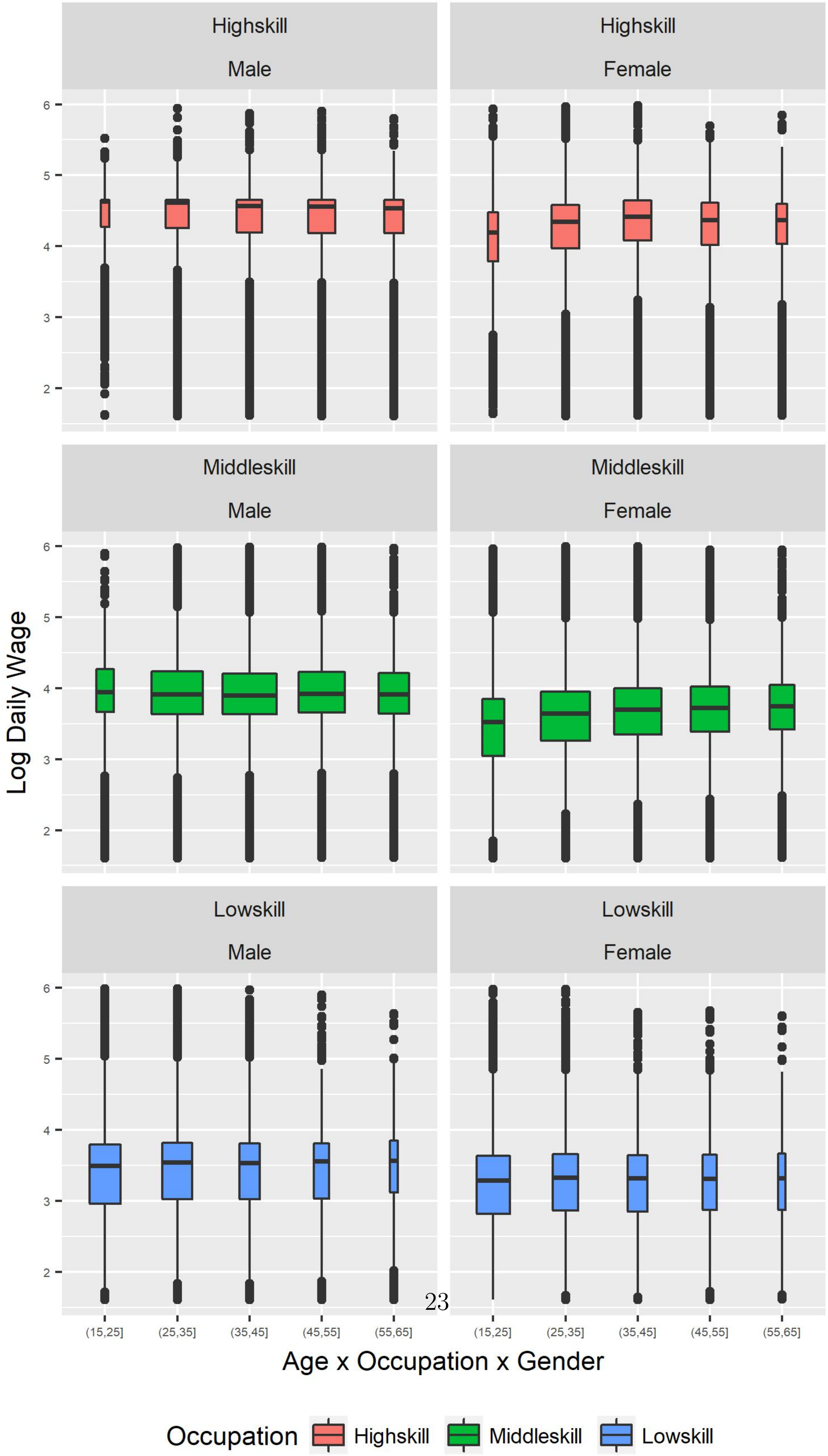
(d) age and type of contract

Figure (4) Boxplot of log-wage and gender



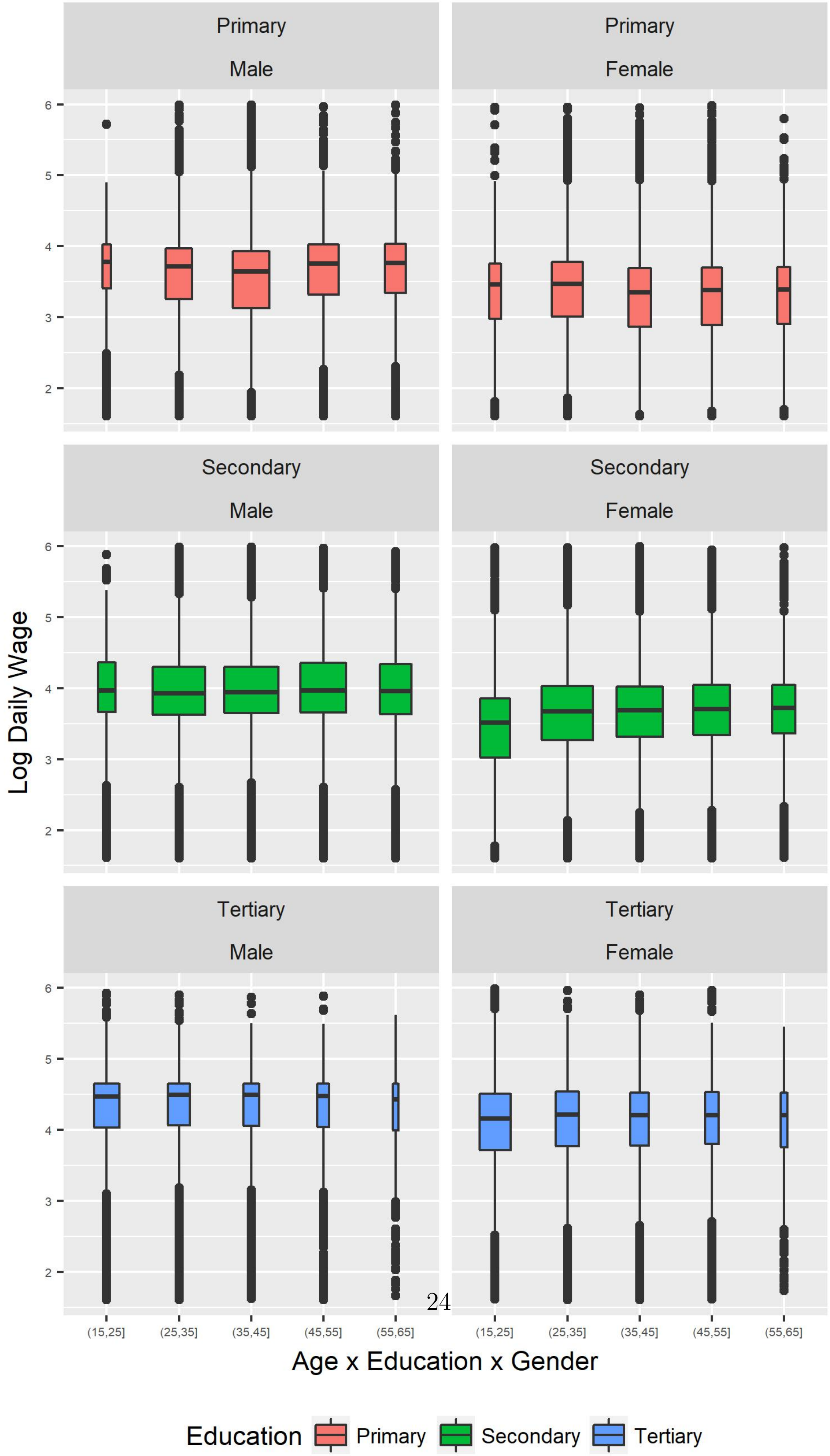
Footnote: The wide is proportional to the size of age group. Source: MCVL 2010

(a) age, gender and type of contract



Footnote: The wide is proportional to the size of age group. Source: MCVL 2010

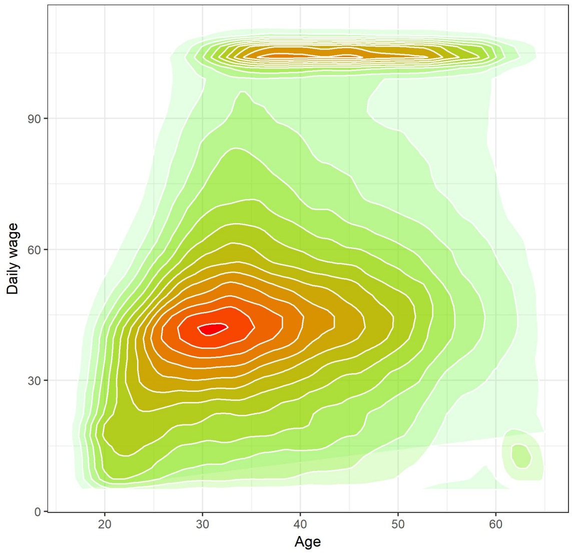
(b) age, gender and occupation



Footnote: The wide is proportional to the size of age group. Source: MCVL 2010

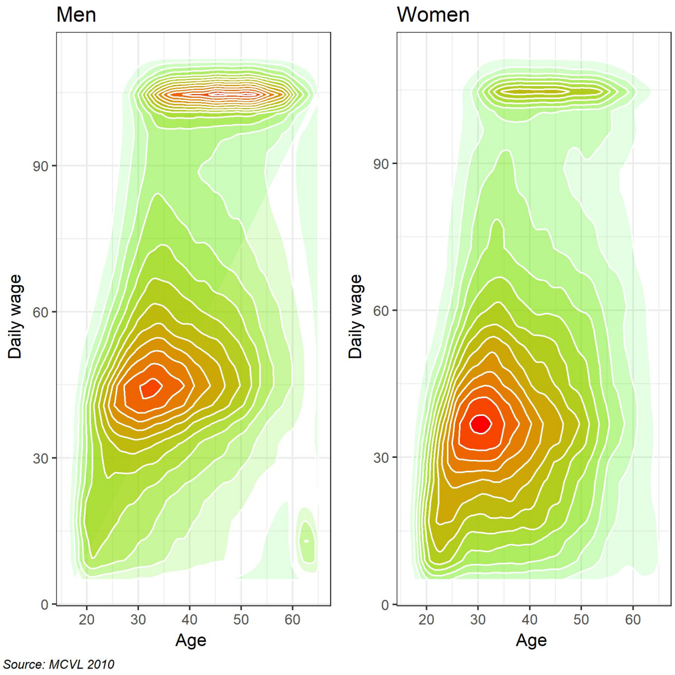
(c) age, gender and education

Figure (5) Heat-map of density plot for daily wage by age

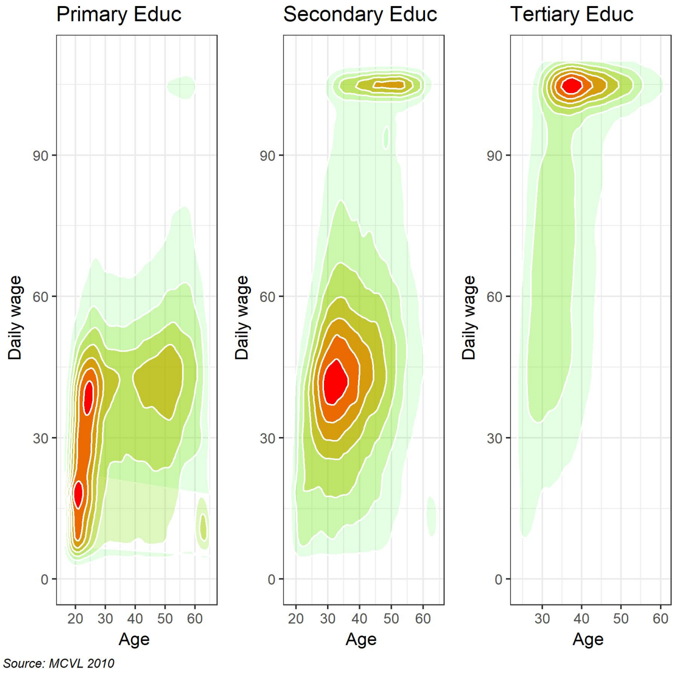


Source: MCVL 2010

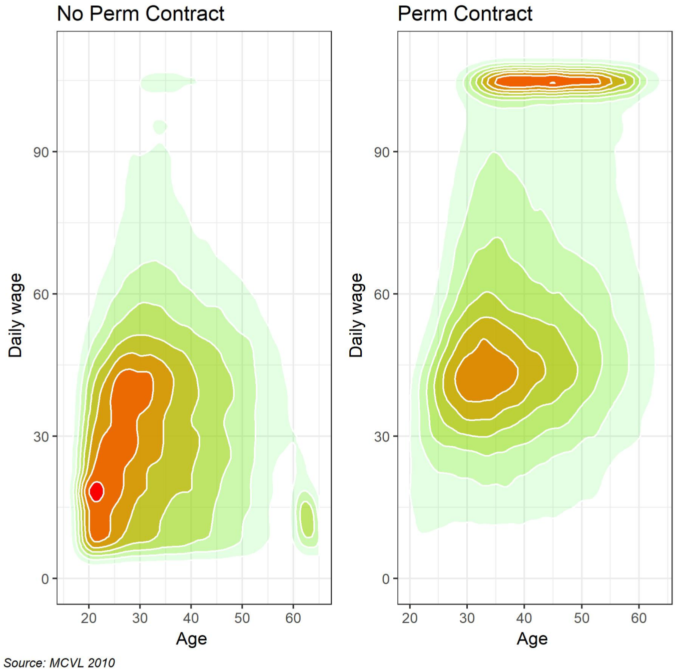
Figure (6) Heat-map of density plot for daily wage and gender



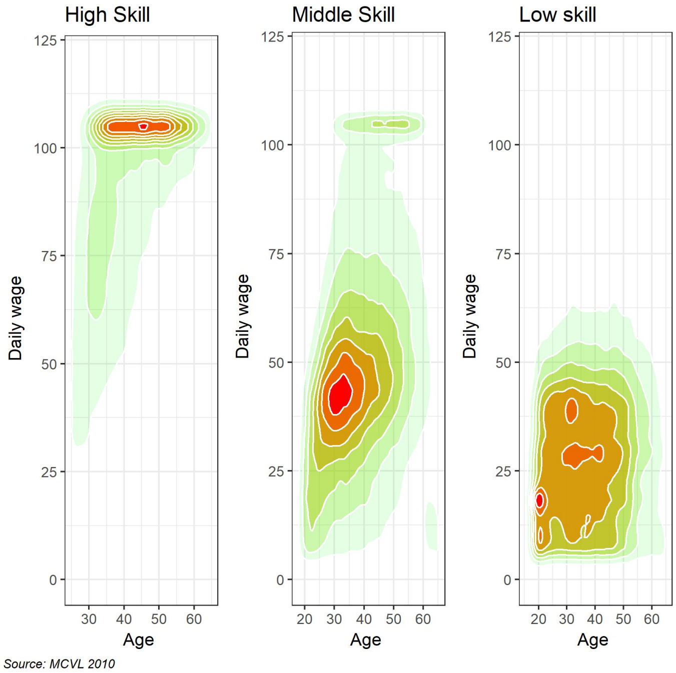
(a) Heat-map of density plot for daily wage by gender and age



(b) Heat-map of density plot for daily wage by education and age

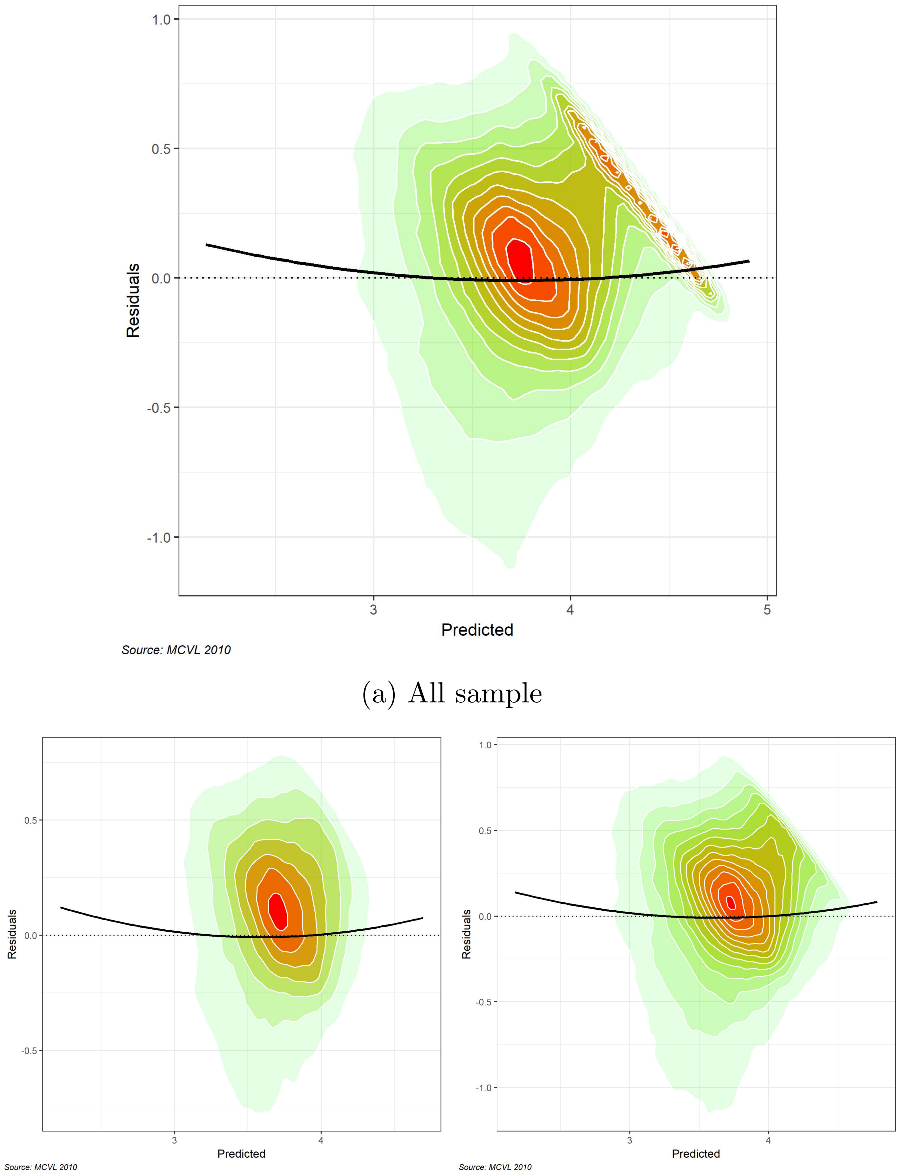


(c) Heat-map of density plot for daily wage by type of contract and age



(d) Heat-map of density plot for daily wage by occupation and age

Figure (7) Heat-map of regression analysis for daily wage



Source: MCVL 2010

(a) All sample

Source: MCVL 2010

Source: MCVL 2010

(b) Without censored data

(c) Tobit model (all sample)

Figure (8) Heat-map of the residuals-vs-fitted-y

Appendix

Table (1) Descriptive Statistics of Daily Wage

	Mean	**Median**	**Std**	**N**	%
Total	45.63	51.59	41.27	541457	100%
Gender					
Male	56.07	50.23	30.39	294132	54.32
Female	46.26	39.89	29.47	247325	45.68
Age					
(15,25]	35.31	30.52	29.07	64011	11.82
(25,35]	49.89	44.90	27.99	167927	31.01
(35,45]	55.20	49.29	29.85	156692	28.94
(45,55]	57.68	51.47	30.81	106941	19.75
(55,65]	54.06	48.81	31.81	45886	8.47
Type of contract					
Temporary	41.41	35.73	31.69	208954	38.59
Fixed	57.99	52.06	27.65	332503	61.41
Occupation					
High Skill	80.53	87.12	29.78	94431	17.44
Medium Skill	49.25	45.35	26.61	343481	63.44
Low Skill	32.94	30.16	23.10	103545	19.12
Education					
Primary	39.23	36.61	25.24	117967	21.79
Secondary	51.26	45.91	29.13	347429	64.17
Tertiary	72.26	74.46	32.19	76061	14.05

Table (2) Estimates of the three regression models. The dependent variable in all the models is the log daily wage[a]

Covariates	OLS (all sample)	Tobit	OLS (sample truncated)[b]
Age	0.052[c]	0.073	0.051
Age Square	−0.001	−0.001	−0.001
Education (Secondary)	0.059	0.054	0.043
Education (Tertiary)	0.165	0.124	0.131
Occupation (Mid Skill)	−0.405	−0.376	−0.311
Occupation (Low Skill)	−0.621	−0.632	−0.509
Gender (Female)	−0.193	−0.184	−0.170
Contract (Fixed)	0.263	0.271	0.252
Industry	0.434	0.512	0.415
Building	0.371	0.423	0.382
Trade	0.175	0.226	0.166
Transport	0.331	0.415	0.319
Hotel	0.043	0.053	0.061
Tele Communication	0.313	0.332	0.263
Finance	0.403	0.387	0.296
Service Intellectual	0.277	0.207	0.218
Service Manual	0.148	0.153	0.164
Public Admin.	0.361	0.306	0.382
Education	0.094	0.072	0.143
Health	0.357	0.253	0.339
Others	0.058	0.061	0.065
# Obs	553,103	553,103	486,398
R2	0.398	0.402	0.304
Adjusted R2	0.398	0.402	.304

[a] The reference categories for the dummies of education, occupation, gender, type of contract and sector are respectively: primary education, high skill occupation, male and temporary contract, and agriculture. The municipalities fixed effects are also included in all the regressions.

[b] We have excluded in the sample all the cases when the log daily wage it is at the ceiling value at the log of 106 euros (4.663).

[c] All the coefficients of the three regressions are significant at p-value level 0.005 (significance computed using robust standard errors). Standard errors are not in display in the table since they are all equal to 0.000 when rounded at three decimal digits.