Proc. 13th Econophysics Colloquium (EC) and 9th Symposium of Physics in Economy and Social Sciences (FENS), 2017

Measuring and Explaining Income Inequalities in Poland: an Estimation of Lorenz Curves using Hazard Function Approach

J.M. Landmesser and A.J. Orłowski*

Faculty of Applied Informatics and Mathematics, WULS-SGGW, Nowoursynowska 159, 02-776 Warsaw, Poland

In this study we compare the income distributions for men and women in Poland in 2014. To examine the differences in the entire range of income values we utilize the hazard function approach. A flexible hazard-function based estimator in the presence of covariates (education, age, etc.) is used to construct conditional density and cumulative distribution functions. Then, we decompose the differences between two distributions using the counterfactual distribution. We estimate also the Lorenz curves for incomes and decompose the differences between the values of the Gini coefficients.

DOI: 10.12693/APhysPolA.133.1445

PACS/topics: income inequalities, hazard function, Lorenz curve

1. Introduction

A common feature of labour markets around different world regions is a relatively large gender pay gap [1]. Most studies find some variation of the gender wage gap across the whole wage distribution (e.g. a small average wage gap might conceal *glass ceiling* effects at the top of the wage distribution; see [2] for Sweden, [3] for Spain, [4] for many other European countries). Additionally, the decompositions carried out show that a large part of this gap cannot be explained by differences in the labor-market skills of women and men (e.g. [5] for USA, [6] for Poland).

Recently statistical techniques of income inequalities decomposition (e.g. [7–10]) are becoming more popular. The idea behind these techniques is that the pay gap might vary across the income distribution and they go far beyond the simple comparison of average values.

The goal of the paper is to examine differences between the income distributions for men and women in Poland. We decompose the differences between two distributions using the counterfactual distribution, which is a mixture of a conditional distribution of the dependent variable and a distribution of the explanatory variables. Such a counterfactual distribution can be constructed in various ways (see [11]). To examine the differences in the entire range of income values we use the hazard function approach (we use a flexible hazard-function based estimator in the presence of covariates to construct conditional density and cumulative distribution functions). Then we estimate the Lorenz curves for incomes and decompose the differences between the values of the Gini coefficients.

2. Methodology of research

2.1. The idea of decomposing the income inequalities

Whenever is a need to explain the differences between expected values of dependent variable in two comparison groups we can apply the Oaxaca–Blinder decomposition method [12, 13]. Let variable Y_g be the personal income in group g, $g = M, W$. The standard assumption used in this decomposition is that the outcome variable Y_g is linearly related to the covariates X_g (the vector of individual characteristics of the person), and that the error term v_g is conditionally independent of X_g: $Y_g = X_g\beta_g + \nu_g, g = M, W$. Then, the expected value of Y_g is $E(Y_g\,|X_g) = X_g\beta_g$ and the Oaxaca–Blinder decomposition for the average income inequality between two groups is as follows:

$$\hat{\Delta}^{\mu} = \bar{Y}_M - \bar{Y}_W = \bar{X}_M\hat{\beta}_M - \bar{X}_W\hat{\beta}_W =$$
$$\underbrace{\bar{X}_M(\hat{\beta}_M - \hat{\beta}_W)}_{\text{unexplained part}} + \underbrace{(\bar{X}_M - \bar{X}_W)\hat{\beta}_W}_{\text{explained part}} \qquad (1)$$

The unexplained part in the equation is the "wage structure" effect and is result of differences in the "prices" of individual people's characteristics. It can be interpreted as the labor market discrimination. The explained part is an effect of characteristics and expresses the differences of the potentials of people in two groups. The approach presented later in this paper can be viewed as an extension of the Oaxaca–Blinder decomposition where the whole conditional income distribution are estimated using parametric methods.

Now, let $F_{Y_g}(y)$ be the distribution function for the variable Y_g:

$$F_{Y_g}(y) = \int F_{Y_g|X_g}(y|\,X)\,\mathrm{d}F_{X_g}(X),$$
$$g = M, W. \qquad (2)$$

We extend the mean decomposition analysis to the case of differences between the two income distributions.

[*] corresponding author; e-mail: arkadiusz_orlowski@sggw.pl

For this purpose, we construct a so-called counterfactual distribution $F_{Y_W^C}(y)$. This is the distribution which represents the hypothetical income distribution function that would prevail for people in group W if they had the distribution of characteristics of group M: $F_{Y_W^C}(y) = \int F_{Y_W|X_W}(y|X) dF_{X_M}(X)$. Then, the difference in income distributions for men and women can be calculated as [11]:

$$F_{Y_M}(y) - F_{Y_W}(y) = \underbrace{[F_{Y_M}(y) - F_{Y_W^C}(y)]}_{\text{unexplained part}} + \underbrace{[F_{Y_W^C}(y) - F_{Y_W}(y)]}_{\text{explained part}}. \quad (3)$$

2.2. The estimation of income distribution function in the presence of covariates

Following Donald et al. [14] we apply a hazard model to income data. Let the non-negative income variable Y have distribution function $F(y|X)$ conditional on a vector of covariates X. The probability that a person has at least income y is given by the survival function $S(y|X) = \Pr[Y \geq y|X] = 1 - F(y|X)$. The hazard function $h(y|X) = f(y|X)/S(y|X)$ gives the probability that the income equals y conditional on the income being at least as large as y.

A convenient model for the influence of covariates on the hazard function is the conditional piecewise-constant hazard model (exponential hazard with the hazard piece dummies) [15]. To allow for the flexible specification of the baseline hazard we divide the income distribution into P segments: $0 = c_0 < c_1 < ... < c_P = \infty$. Then the hazard function is given by

$$h(y|X) = h_{0k}(y) \exp(X\beta) \quad \text{for } y \in (c_{k-1}, c_k), k = 1, \ldots, P, \quad (4)$$

where $h_0(y)$ is the baseline hazard and the effect of X is constant within each segment. The survival function then becomes

$$S(y|X) = \exp\Big[-\sum_{j=1}^{k-1}(c_j - c_{j-1})h_{0j}(y)\exp(X\beta) - (y - c_{k-1})h_{0k}(y)\exp(X\beta)\Big] \text{ for } y \in (c_{k-1}, c_k). \quad (5)$$

An estimate of the survival function $\hat{S}(y|X)$ makes estimating the distribution function $F(y|X)$ very simple: $\hat{F}(y|X) = 1 - \hat{S}(y|X)$. The individual results are finally averaged at incomes values corresponding to each of the baseline segments giving $F_Y(y)$.

2.3. The Lorenz curve estimator and the Gini index calculation

Once a function $F_Y(y)$ is obtained, one can then estimate the Lorenz curve and the generalized Lorenz curve conditional upon the covariates. Let $\tau \in(0,1)$ and $Q_\tau = F_Y^{-1}(\tau)$ be the τ^{th} population quantile of Y. Thus, $\tau = \int_0^{Q_\tau} f(y)\,\mathrm{d}y = F(Q_\tau)$. The Lorenz curve of Y is then defined by

$$L(\tau) = \frac{\int_0^{Q_\tau} y f(y)\,\mathrm{d}y}{\int_0^{\infty} y f(y)\,\mathrm{d}y} = \frac{\int_0^{\tau} F_Y^{-1}(s)\,\mathrm{d}s}{E(Y)} \quad (6)$$

and plots the proportion of the total income held by the lowest ($\tau\times$ 100) percent of the population against τ [16, 17]. Note that $L(0) = 0$, $L(1) = 1$. The Lorenz curve equal to the 45° line denotes perfect equality and any inequality causes the Lorenz curve to fall below this line. It is also possible to extend Lorenz curve analysis to incorporate the mean, defining the generalized Lorenz curve as $GL(\tau) = E(Y)L(\tau)$. Comparing the generalized Lorenz curves for two income distributions one can see differences in incomes means and differences in dispersion. From the hazard model described above we obtain for $\tau \in(0,1)$ the following formulae:

$$L(\tau) = \frac{(1-\tau)\ln(1-\tau)}{1 + \lambda_{m+1}c_m - \sum_{j=1}^{m}(c_j - c_{j-1})\lambda_j} + \tau, \quad (7)$$

$$GL(\tau) = \frac{(1-\tau)\ln(1-\tau)}{\lambda_{m+1}} + \frac{\tau\Big(1 + \lambda_{m+1}c_m - \sum_{j=1}^{m}(c_j - c_{j-1})\lambda_j\Big)}{\lambda_{m+1}}, \quad (8)$$

where $\lambda_j = h_{0j}(y)\exp(X\beta)$.

The most common measure of inequality, the Gini index G, is the ratio of the area between the Lorenz curve and the 45° line (called the area of concentration) to the area under the 45° line. Having estimated the Lorenz curves for men's and women's incomes it is possible to calculate the Gini indices [17, 18]. Using the counterfactual distribution $F_{Y_W^C}(y)$, the counterfactual Lorenz curve can be derived and then the counterfactual Gini coefficient value can be calculated. Finally, we can form the decomposition of differences in Gini's indices as

$$G_{Y_M} - G_{Y_W} = \underbrace{[G_{Y_M} - G_{Y_W^C}]}_{\text{unexplained part}} + \underbrace{[G_{Y_W^C} - G_{Y_W}]}_{\text{explained part}}.$$

3. Data

Our analysis is based on data from the European Union Statistics on Income and Living Conditions project for Poland in 2014. It collects source and amount of incomes, labor force information, and general demographic characteristics. The sample size is 9,904 records (5,177 for males and 4,727 for females). The data concern annual net employee incomes in 2014, expressed in €(the outcome variable *income*). Each person is characterized by attributes such as *yearswork* — quasi-continuous variable from 0 to 54 years of work, *educlevel* — ordinal variable from 1 (lowest education level) to 5 (highest), *married* — binary variable, married (1) / unmarried (0), *parttime* — binary variable, part-time (1) / full-time job (0), *manager* — binary variable, supervisory managerial position (1) / non-supervisory position (0).

The selected features of the variables are presented in Table I.

TABLE I

The mean values and the share of categories for selected variables. Source: own calculation.

Variable	Men	Women	Variable		Men	Women
av. income	7,165.94	5,900.21	educlevel	= 1	4.91%	3.89%
av. yearswork	20.09	18.46		= 2	1.45%	0.55%
married = 1	71.53%	69.60%		= 3	68.57%	47.32%
parttime = 1	4.31%	10.09%		= 4	2.55%	7.91%
manager = 1	18.68%	15.74%		= 5	22.52%	40.32%

4. Results

In this section we analyse in detail, the difference in the income distributions for men and women in Poland using the decomposition methods described above.

In the first step the Oaxaca–Blinder decomposition has been applied for the average values. There was a positive difference between average values of incomes. The raw differential was equal to 1,265.73€. The explained effect was low and negative (–282.60 €), but the unexplained was huge and positive (1,548.33€). The inequalities examined should be assigned in the majority to the coefficients of estimated models rather than to the differentiation of individual characteristics.

Then, we estimated separately for men and women the coefficients for conditional piecewise-constant hazard models (not presented due to lack of space). We used 20 baseline segments with dividing points at the 20-quantiles of the unconditional pooled income distribution. The plots of hazard are presented in Fig. 1. The higher located graph of hazard for women indicates greater exposure of women to the loss of earnings than in the case of men.

Now we treat the hazard function as a flexible functional form that allows us to generate the estimates of the income distribution functions. First, the distributions $\hat{F}_{Y_g|X_g}(y|X), g = M, K$, were determined. Each of them gives the probability that incomes will take values lower than a certain level y (for fixed X and parameters β). To illustrate the variability of both income levels and people's characteristics along the income distribution, the results for each individual were averaged over the intervals $(c_{k-1}, c_k), k = 1, ..., 20$. The distribution functions averaged in this way are presented in the form of points connected by straight lines in Fig. 2. Since $\hat{F}_{Y_W|X_W}(y|X) > \hat{F}_{Y_M|X_M}(y|X)$ for all y, then for the woman, the probability of not exceeding the income level y is higher than for the male.

The counterfactual distribution $\hat{F}_{Y_W^C}(y)$ was determined by setting, first, the distribution of incomes that would prevail for women if they had the distribution of men's characteristics (the parameters β_W were taken from the hazard model for women and the values of explanatory variables X_M for men). Then, the results were averaged over the intervals $(c_{k-1}, c_k), k = 1, ..., 20$, gaining the curve Favg_C_(y) in Fig. 2.

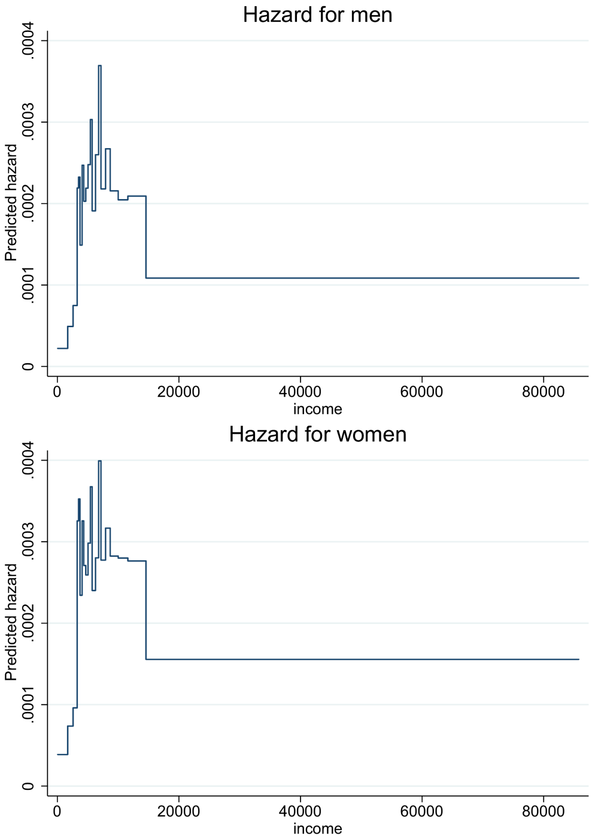

Fig. 1. The plots of hazard functions for men and women, respectively.

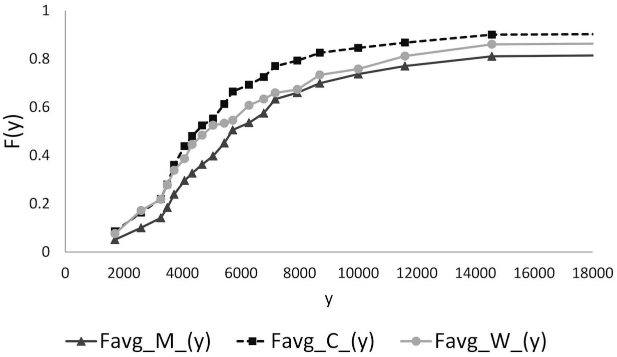

Fig. 2. The averaged cumulative distribution functions for incomes (for men — Favg_M_(y), for women — Favg_W_(y), counterfactual — Favg_C_(y)).

In the next step, the quantiles for income distributions are determined by inverting the estimated distribution functions. The precise values of $\hat{Q}_{g,\tau} = \hat{F}_{Y_g}^{-1}(\tau)$ and $\hat{Q}_{W,\tau}^{C} = \hat{F}_{Y_W^C}^{-1}(\tau)$ were computed using linear interpolations. This allowed to decompose the income gap for quantiles and to determine the explained and unexplained components of the difference in terms of quantiles. The results presented in Table II indicate positive differences between male and female incomes

TABLE II

Decomposition of difference in income distributions in terms of quantiles. Source: own calculation.

τ	$\hat{Q}_{M,\tau}$	$\hat{Q}_{W,\tau}$	$\hat{Q}^{C}_{W,\tau}$	Total difference	Unexplained part	Explained part
0.1	2,588.67	1,922.59	1,866.79	666.08	721.88	–55.80
0.2	3,553.51	2,999.38	3,026.52	554.13	526.99	27.14
0.3	4,109.81	3,571.89	3,544.49	537.92	565.32	-27.40
0.4	5,062.84	4,133.68	3,897.80	929.16	1,165.04	–235.88
0.5	5,688.47	4,823.51	4,488.98	864.96	1,199.48	–334.52
0.6	6,944.98	6,204.25	5,348.28	740.73	1,596.70	–855.97
0.7	8,714.88	8,262.12	6,385.50	452.76	2,329.39	–1,876.62
0.8	13,776.28	11,243.09	8,089.19	2,533.19	5,687.09	–3,153.90
0.9	85,765.00	63,218.94	14,509.47	22,546.06	71,255.53	–48,709.48

TABLE III

Decomposition of difference in Gini index values. Source: own calculation.

Scenario	G_{Y_M}	G_{Y_W}	$G_{Y_W^C}$	Total difference	Unexplained part	Explained part
X_M, X_W original	0.2906	0.2731	0.2366	0.0174	0.0539	–0.0365
yearswork_W 10%	0.2906	0.2784	0.2366	0.0122	0.0539	–0.0417
educlevel_W 10%	0.2906	0.2918	0.2366	–0.0012	0.0539	–0.0552
parttime_W 10%	0.2906	0.2726	0.2366	0.0180	0.0539	–0.0359
manager_W 10%	0.2906	0.2741	0.2366	0.0165	0.0539	–0.0374

at each level of income. These differences are non-monotonous: they are initially decreasing (among the poorest), for quantiles of the order 0.4–0.6 are higher again, then lower again, and on the right end of the income distribution grow stronger (among the richest).

The unexplained component of the income gap (associated with the "valuation" of the people's characteristics by the market) increases with the amount of income. This demonstrates that the discrimination is more evident for higher values of incomes. The negative values of the explained component are especially large in the groups of the best earning people. This reflects the reduction of wage inequality, probably due to "better" characteristics of women than men. Such a favorable reduction in the gap for women deepens as the higher income groups are considered (maybe women in the richest group should earn much more than men).

The last step of our analysis concerned the estimation of the Lorenz curves and the generalized Lorenz curves. From the formulae (7) and (8) we received the so-called pseudo-Lorenz curves (the intermediate results are averaged in the X-space over the intervals $(c_{k-1}, c_k), k = 1, ..., 20$) presented in Fig. 3 (the curves L_M, L_W). The counterfactual Lorenz curve was also derived (L_C). Comparing the generalized Lorenz curves for men's and women's income distributions (GL_M and GL_W in Fig. 3) we can see differences in incomes means (by their right most ordinates) and differences in dispersion (by how they are bowed out).

Finally, having estimated the pseudo-Lorenz curves for men's and women's incomes we calculated the Gini indices by interpolation (see Table III). According to the values obtained the income inequalities between men are bigger than between women (the difference equals 0.0174).

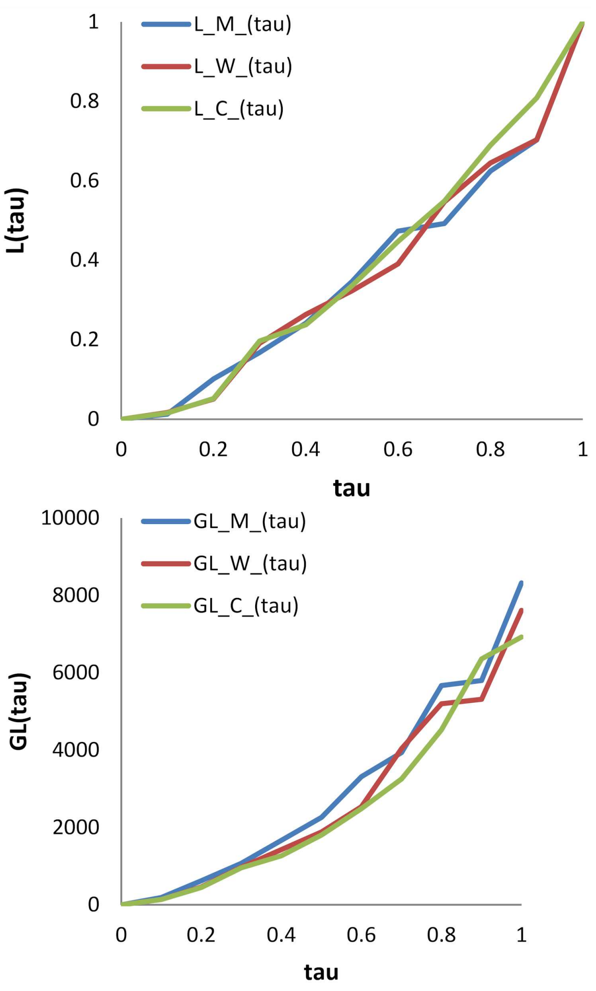

Fig. 3. The pseudo-Lorenz curves.

The Gini coefficient value for the counterfactual distribution was also calculated, which made it possible to decompose the observed difference.

Incorporating the effects of covariates into the Lorenz curve ordinate estimates makes it possible to infer the effects of various person's characteristics on the shape of this curve and on the Gini coefficient value. Especially, it is possible to calculate the effects on inequality of such scenarios like e.g. the increase in women's part-time employment by 10% or increase in the number of managerial positions held by women by 10% (Table III).

5. Conclusion

In this paper we applied the methods of the decomposition of differences between the income distributions. We started with the decomposition of the average values for incomes by using the Oaxaca–Blinder method. There was a positive difference between the mean values of log incomes. The explained effect was low and negative, but the unexplained was huge and positive.

Then we estimated two conditional piecewise-constant hazard models for men and women, separately. We also constructed the counterfactual distribution. Using the models estimated makes it possible to decompose the inequalities between incomes along the whole distribution. The total effect increases with income, the explained effect is lower and negative.

The method allows us to investigate the differences in incomes for two groups of people analyzing the Lorenz curves and the Gini indices for inequality.

Results obtained and presented here seem to be very promising and the subject is definitely worth of further explorations. We plan to develop some of the related ideas in a forthcoming paper.

References

[1] S. Pasqua, *Europ. J. Comparat. Econom.* **5**, 197 (2002).

[2] J. Albrecht, A. Björklund, S. Vroman, *J. Labor Econom.* **21**, 145 (2003).

[3] S. De la Rica, J.J. Dolado, V. Llorens, *J. Populat. Econom.* **21**, 751 (2008).

[4] W. Arulampalam, A.L. Booth, M.L. Bryan, *Industr. Labor Relat. Rev.* **60**, 163 (2007).

[5] K. Karpio, J. Landmesser, P. Łukasiewicz, A. Orłowski, *Acta Phys. Pol. A* **129**, 965 (2016).

[6] J.M. Landmesser, *Statist. Transit. New Series* **17**, 331 (2016).

[7] J. DiNardo, N.M. Fortin, T. Lemieux, *Econometrica* **64**, 1001 (1996).

[8] N.M. Fortin, T. Lemieux, *J. Human Resourc.* **33**, 610 (1998).

[9] J.F. Machado, J. Mata, *J. Appl. Econometr.* **20**, 445 (2005).

[10] B. Melly, *Labour Econom.* **12**, 577 (2005).

[11] N.M. Fortin, T. Lemieux, S. Firpo, "Decomposition Methods in Economics", *NBER Working Paper* **16045**, Cambridge 2010.

[12] R. Oaxaca, *Int. Econom. Rev.* **14**, 693 (1973).

[13] A. Blinder, *J. Human Resourc.* **8**, 436 (1973).

[14] S.G. Donald, D.A. Green, H.J. Paarsch, *Rev. Econom. Stud.* **67**, 609 (2000).

[15] P. Brochu, D.A. Green, T. Lemieux, J. Townsend, "The Minimum Wage, Turnover, and the Shape of the Wage Distribution", *Vancouver School of Economics Working Paper*, University of British Columbia, Vancouver (BC) 2015..

[16] S.G. Donald, D.A. Green, H.J. Paarsch, "Differences in Earnings and Wage Distributions between Canada and the United States: An Application of a Semi-Parametric Estimator of Distribution Functions with Covariates", Discussion Paper **95-34**, Department of Economics, The University of British Columbia, Vancouver 1995.

[17] J.L. Gastwirth, *Rev. Econom. Statist.* **54**, 306 (1972).

[18] M. Biernacki, *Math. Econom.* **3**, 125 (2006), (in Polish).