Proceedings of the 8th Polish Symposium of Physics in Economy and Social Sciences FENS, Rzeszów, November 4–6, 2015

# The Quantile Decomposition of Personal Income Distributions in the USA

K. Karpio, J.M. Landmesser, P. Łukasiewicz and A.J. Orłowski
Faculty of Applied Informatics and Mathematics, WULS-SGGW, Nowoursynowska 159, 02-776 Warsaw, Poland

In this study we compared incomes distributions in the USA for two subgroups (defined according to sex or race). We utilized the quantile decomposition method to describe differences between the two distributions as a function of their quantiles. The analyzed objects are characterized by the set of attributes (education, age, etc.). We evaluate strength of the influence of the attributes onto the various parts of the incomes distributions. In such a way we evaluate income inequalities and their causes in two subgroups of people.

DOI: 10.12693/APhysPolA.129.965

PACS/topics: 89.65.Gh, 88.05.Lg

## 1. Introduction

The U.S. economy is very often seen as "the engine" of the world economy and a driver of global growth. Therefore, it is important to observe the state of this economy and to analyze its particular market conditions. One can note an increase of interest towards studies of income (wages) inequities related to various economics–social aspects. Because there is a need for continuous updating of research in this area, also the authors of this paper focus their interest on the study of the inequalities observed in the U.S.

The U.S. Government Accountability Office study, released in 2003, showed that women earned, on average, 20% less than men during the period 1983 to 2000 [1]. Also Wood et al. [2] examined the gender wage gap in USA and found that women were paid 81.5% of what men "with similar demographic characteristics, family situations, work hours, and work experience" were paid. Many of the studies found that the gender wage gap can only be partially explained by human capital factors. Using the current population survey (CPS) data Altonji and Blank [3] found that only 27% of the gender wage gap is explained by differences in people's characteristics, whereas Boraas and Rodgers [4] reported that 39% of the gap is explained.

Many studies carried out in the U.S. confirm that the gender wage gap increases with age. According to the U.S. Bureau of Labor Statistics young women earn 92.3% of men's earnings while older women earn just 76.4% of men's earnings [5]. Moreover, in some large urban centers women in their twenties earn even more than their male counterparts [6]. On the other hand, it is also known that women earn in general less than men at all educational levels but a gender pay gap widens with level of education [7].

Majority of studies focuses on the explanation of wages differences by differences in people's characteristics. Blau and Kahn [8] noted when many human capital variables are taken into account then the gender wage gap gets smaller but still a substantial portion of the pay gap remains unexplained. In our paper we focus on the unexplained part of the pay gap (that means, unexplained by differences in characteristics). This part is usually attributed to wage (income) discrimination. The same people's characteristics usually lead to different wages in two groups. We evaluate such a "unit prices" and analyze people's characteristics to find out how much they are discriminant.

In this paper we study personal income distribution in the U.S. The distribution itself has been investigated many times and discussed in the econophysics literature [9–13]. Those papers were mainly focused on studying incomes dynamics and modeling the income distribution. Besides studying the income distribution itself we also take into account the impact of personal characteristics on incomes. Persons gaining income are characterized by the set of attributes (sex, education, age, race and origin). The relation between attributes values and incomes has been already investigated by means of a multidimensional analysis using decision trees [14, 15]. However this time we evaluate attributes 'unit prices' on the labor market and an impact of changes of attributes values on incomes. We study differences between incomes of two subpopulations (*Men* vs. *Women* and *Whites* vs. *Others*) based on the latest accessible data collected by survey of income and program participation (SIPP-2008) project. SIPP is the premier source of information for income and well-being of U.S. households. The household's sample consists of twice more records than is in the mentioned above CPS data.

The main aim of this paper is to study size of the income gap between two subpopulations along the whole income distribution. According to the idea of Oaxaca and Blinder [16, 17] an income gap can be decomposed into two parts, see formula (2). The first part is related to differences of attributes values while the second one exhibits differences between "unit prices" of the attributes on the labor market. By decomposing income gap we search for answers for the following questions: (i) what is the size of the income gap vs. income, (ii) what is a share of each part in the gap and how big is a level of discrimination, (iii) which attributes are the most discriminatory.

Let $y_{n\times 1} = [y_1, y_2, \ldots, y_n]'$ denote vector of incomes and $X_{n\times k} = [1, A_1, \ldots, A_k]$ — matrix of attributes values. Using a linear regression model the incomes can be described

$$\widehat{y} = X\widehat{\beta}, \tag{1}$$

where $\widehat{y}_{n\times 1} = E(y|X)$ — vector of mean conditional incomes, $\widehat{\beta}_{k\times 1}$ — vector of model coefficients. The coefficients of the model (1) are estimated by the minimization of $\sum\limits_{i=1}^{n}(y_i - X_i\beta_j)^2$.

Of course, this means that the unconditional income is equal to an average of the mean conditional incomes, $E(\widehat{y}) = E(X\widehat{\beta}) = \bar{X}\widehat{\beta}$, where $\bar{X} = E(X)$ — vector of mean values of attributes.

Let $\mathit{\Delta}(\mu)$ denote difference between unconditional means of the distributions of incomes for two groups $A$ and $B$. Then a decomposition of $\mathit{\Delta}(\mu)$ into two components can be expressed [16, 17]:

$$\mathit{\Delta}(\mu) = \bar{X}_A\widehat{\beta}_A - \bar{X}_B\widehat{\beta}_B =$$

$$(\bar{X}_A - \bar{X}_B)\widehat{\beta}_A + \bar{X}_B(\widehat{\beta}_A - \widehat{\beta}_B), \tag{2}$$

The first component in (2) explains the difference of the means by differences between the attributes values. The second term describes the difference caused by different values of the models coefficients. Such an approach gives only the overall indicator of differences between the incomes distributions. In our work we used quantile regression models in order to study differences at any level of incomes. It is particularly important because attributes influence different parts of the income distributions in different way. The approach utilized will also allow us to decompose the observed differences into two classes as in (2).

## 2. Quantile decomposition method

We describe a relationship between the attributes and the income distribution using the quantile regression (QR):

$$\widehat{q}(\theta) = X\widehat{\beta}(), \tag{3}$$

where $\widehat{q}(\theta) = Q_\theta(y|X)$ — the vector of the conditional quantiles of $y$ for the fixed $\theta \in \widehat{I}(0,1)$.

The quantile regression estimator for the quantile $q$ minimizes $\sum\limits_{i=1}^{n}\rho_\theta(y_i - X_i\beta_j(\theta))$, where

$$\rho_\theta(u) = \begin{cases} \theta u & \text{for} \quad u \geq 0 \\ (\theta - 1)u & \text{for} \quad u < 0 \end{cases}$$

The above sum contains the penalty asymmetric function $\rho_\theta$ for over and under prediction [18]. Similarly to (1), the coefficients of the quantile regression can be interpreted as “unit prices” of the attributes-skills on the labor market.

Because of a mean of the conditional quantiles is not equal to the unconditional quantile, in order to evaluate the unconditional quantiles we use a bootstrap method. The evaluated coefficients are used to generate a random sample of the incomes based on the random set of persons. This procedure [19, 20] is as follows:

1. generate a random sample of size $m$ from an $U$[0,1]: $\theta_1, \theta_2, \ldots, \theta_m$;
2. using the dataset $X$ estimate $m$ quantile regressions $Q_{\theta_i}(y\,|X)$, obtaining coefficients $\widehat{\beta}(\theta_i)$, $i = 1, \ldots, m$;
3. generate a random sample $\{X_i^*\}$, $i = 1, \ldots, m$ with replacement from rows of $X$;
4. then $\{y_i^* \equiv X_i^*\beta(\theta_i)\}$, $i = 1, \ldots, m$ is a random sample from the unconditional distribution of incomes.

We employ the above procedure to both subgroups $A$ and $B$ using the common value of $m$.

Performing the decomposition of the differences between the distributions one utilizes so-called counterfactual distributions. They are a mixture of a conditional distribution of a dependent variable and various distributions of explanatory variables [21]. In order to perform the decomposition of differences between the income distributions for groups $A$ and $B$ we need to generate a random sample from the income distribution that would have prevailed in $A$ group if all attributes had been distributed as in the group $B$. In other words we use the model for the group $A$ with a sample of data $X$ from the group $B$.

We study differences between incomes distributions for the groups $A$ and $B$ by calculating: $\Delta^X(\theta) = X_A\beta_A(\theta) - X_B\beta_A(\theta)$, shows the contribution of the attributes; $\Delta^\beta(\theta) = X_B\beta_A(\theta) - X_B\beta_B(\theta)$, shows the contribution of the model coefficients.

The decomposition of the differences between the income distributions for the groups $A$ and $B$ is as follows:

$$\widehat{\mathit{\Delta}}(\theta) = Q_\theta(y_A^*\,|X_A^*) - Q_\theta(y_B^*\,|X_B^*) = Q_\theta(y_A^*\,|X_A^*)$$

$$-Q_\theta(y_A^{C*}\,|X_B^*) + Q_\theta(y_A^{C*}\,|X_B^*) - Q_\theta(y_B^*\,|X_B^*) =$$

$$\underbrace{(X_A^* - X_B^*)\widehat{\beta}_A(\theta)}_{\widehat{\mathit{\Delta}}^X(\theta)} + \underbrace{X_B^*(\widehat{\beta}_A(\theta) - \widehat{\beta}_B(\theta))}_{\widehat{\mathit{\Delta}}^\beta(\theta)}. \tag{4}$$

## 3. Data

The data were collected in the SIPP project. The SIPP is a statistical survey conducted by the United States Census Bureau [22]. It collects source and amount of incomes, labor force information, and general demographic characteristics. The analyzed data concern personal annual incomes in 2008, expressed in k$ and consist of 287,298 records. Each person is characterized by AGE, RACE, SEX, ORIGIN and EDUCATE.

AGE — quasi-continuous variable from 14 to 84,
RACE — binary variable, Whites(1)/Others(0),
SEX — binary variable, Men(1)/Women(0),

ORIGIN — binary variable, Spanish(1)/Others(0),
EDUCATE — ordinal variable from 1 (lowest) to 16 (highest).

During the analysis the data were split among two subgroups. In the first step we compared the income distribution for Men (group $A$) with the same distribution for Women (group $B$). The second step concerned a comparison of the income distributions according to race: Whites (group $A$) and Others (group $B$). Details of data for all subgroups are summarized in Table I.

TABLE I

The number of records and mean annual incomes for the studied groups of people.

| Group | Men | Women | Whites | Others |
|---|---|---|---|---|
| # of records | 138 077 | 149 221 | 232 576 | 54 722 |
| average income | 45.46 k$ | 27.70 k$ | 37.35 k$ | 31.51 k$ |

## 4. Results

### *4.1. Decomposition of differences between the income distributions for men and women*

The first step of the analysis concerned differences between the personal income distributions for men and women. Figure 1 (left plot) contains the differences vs. quantile rank for raw data as well as the results of the quantile regression models. The results for the models do not differ significantly from the empirical data. The positive values indicate on higher incomes for men than for women. The differences between the income distributions increase linearly with income but for the highest incomes the rate of changes gets higher. The difference starts at about 2 k$ for the 0.05 quantile and grows up to 43.6 k$ for the last quantile.

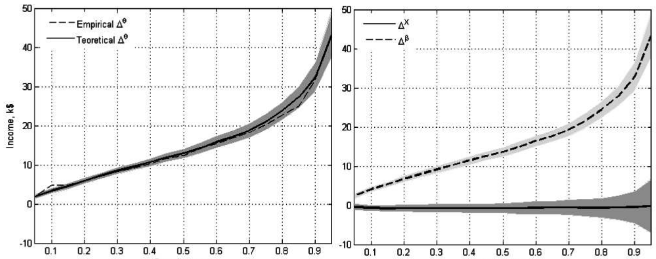

Fig. 1. Differences of income distributions for Men and Women vs. quantile rank. Shaded areas indicate 95% confidence intervals. Left plot: total differences $\widehat{\Delta}(\theta)$. Dashed line: empirical data, solid line: theoretical results. Right plot: decomposition $\widehat{\Delta}(\theta)$ into attribute part $\widehat{\Delta}^{X}(\theta)$ and QR coefficient part $\widehat{\Delta}^{\beta}(\theta)$.

The theoretical differences were further decomposed into the two components (Fig. 1, right plot): the first one explaining the contribution of the attributes differences and the second one explaining the contribution of the different values of models coefficients. The first component is negligible in the whole range of the incomes. It implies that the observed differences can be fully assigned to differences of the models coefficients. In other words, the differences of the incomes of men and women are exclusively due to the different "unit prices" of the person's attributes.

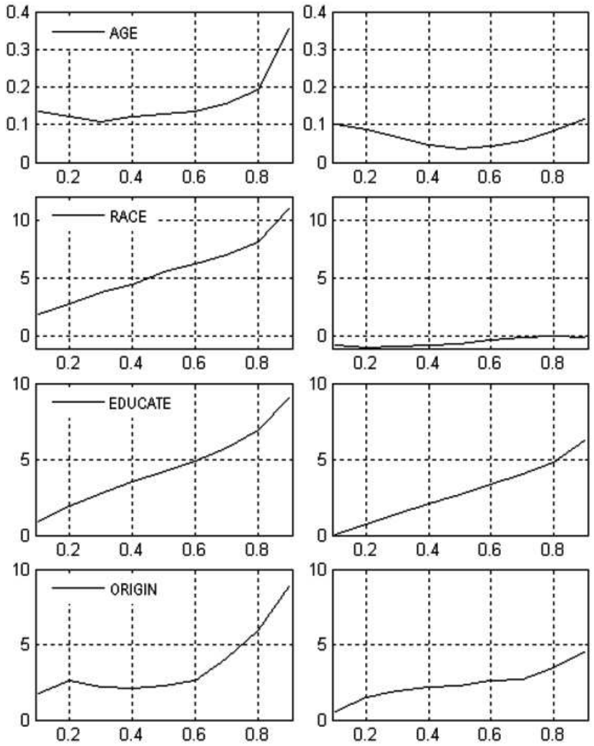

Fig. 2. Quantile regression coefficients for Men (left column) and Women (right column). Horizontal axes — quantile rank.

Figure 2 contains values of the QR coefficients for each of the analyzed attributes. Dispersion of the coefficients values is smaller for Women than for Men for every attribute. It is clearly visible in the case of RACE. White men gain higher incomes than others and these differences increase with income. On the other hand, the influence of RACE is close to zero for women.

In the next step the differences of the QR coefficients for men and women were calculated for each attribute. Then they were multiplied by the mean values of the corresponding attributes. The results are presented in Fig. 3. This allowed us to compare a relative impact of the coefficients differences on the total differences between the distributions. EDUCATION has the strongest impact. The influence of the AGE and RACE is of the same level but is significantly weaker than for EDUCATION. The significance of each of the three attributes increases with income. However the influence of ORIGIN on the total differences is very small and is negligible in the center part of the income distributions.

### *4.2. Decomposition of differences between the income distributions for Whites and Others*

The same procedure has been performed to analyze differences between incomes of Whites and Others. The re-

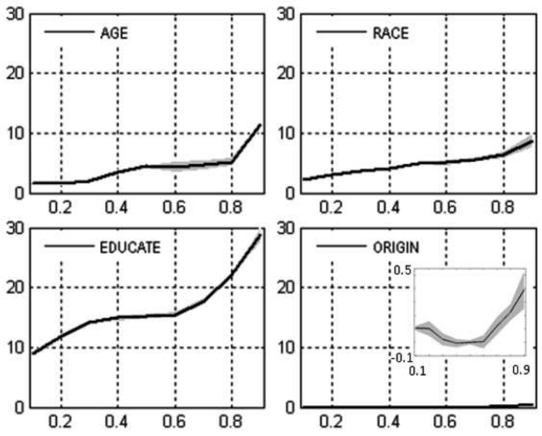

Fig. 3. Differences between incomes due to the different values of QR coefficients for Men and Women vs. quantile rank. Differences of coefficients (Fig. 2) were multiplied by the mean values of the corresponding attributes.

sults are presented in Figs. 4–6. In this case both components are important. The first one (related to attributes) is positive in almost whole range of incomes. This means that Whites are characterized by "better" values of the attributes, i.e. the values that lead to higher incomes. The interesting behavior is associated with the second component (related to models coefficients). For small incomes ($\theta < 0.1$) its value is negative, then it is consistent with zero ($0.1 < \theta < 0.3$) and is positive afterwards. In the first range of incomes Whites have lower "unit prices" of the attributes than Others but they have "better" values of the attributes. Both components cancel each other out that is why there are no observed differences of the distributions for small incomes. In the third region values of the second component are positive and rising, which means that Whites become increasingly favored. One should note that both components are of the same value within the errors in the third region of income distribution (see Fig. 4, right plot).

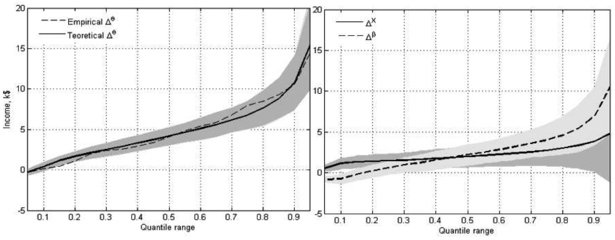

Fig. 4. Differences of income distributions for Whites and Others vs. quantile rank. Shaded areas indicate 95% confidence intervals. Left plot: total differences $\widehat{\Delta}(\theta)$. Dashed line: empirical data, solid line: theoretical results. Right plot: decomposition $\widehat{\Delta}(\theta)$ into attribute part $\widehat{\Delta}^X(\theta)$ and QR coefficient part $\widehat{\Delta}^\beta(\theta)$.

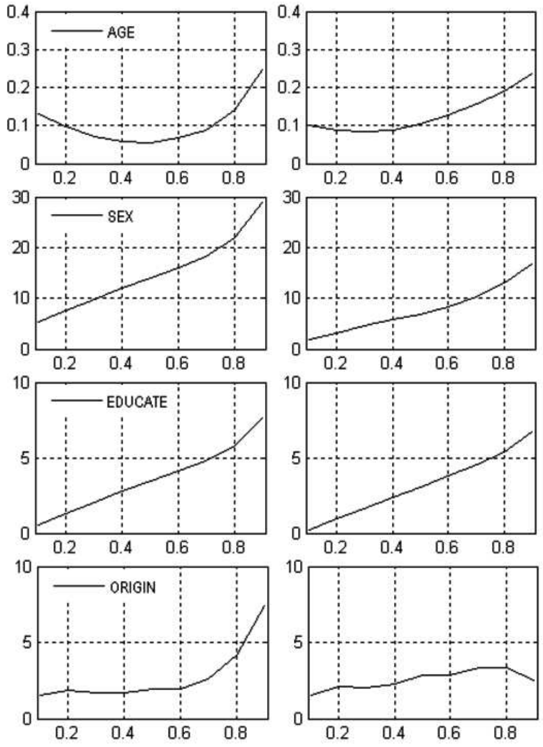

Fig. 5. Quantile regression coefficients for Whites (left column) and Others (right column). Horizontal axis — quantile rank.

Values of the QR coefficients are similar to each other in both groups (Fig. 5). Note that men gain higher "unit prices" than women in the case of both Whites and Others. That differences increase with income. Returns ("unit prices") from EDUCATION are similar to each other in both groups, otherwise as previously.

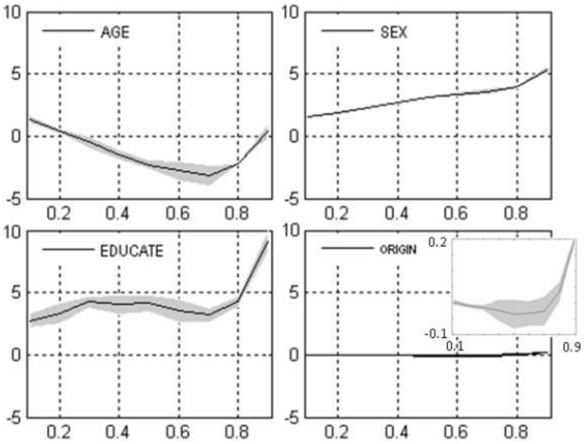

Fig. 6. Differences between incomes due to the different values of quantile regression coefficients for Whites and Others vs. quantile rank. Differences of coefficients (Fig. 5) were multiplied by the mean values of the corresponding attributes.

The results for each attribute are presented in Fig. 6. As previously the influence of ORIGIN on the differences between the income distributions is very small and is negligible in comparison with the remaining attributes. The most important attributes are

EDUCATION and SEX. The influence of SEX increases in the whole range of incomes. EDUCATION is at the similar level and strongly rises for the last two deciles of incomes.

## 5. Conclusions

In this paper we applied the Machado–Mata method of the decomposition of differences between the income distributions. This method was applied to analyze personal incomes in the USA in 2008. We analyzed differences between the income distributions for the two groups of people. The elaborated procedure was applied to Men and Women and repeated for Whites and Others. The goal of the decomposition was to split the differences between the income distributions into two components. The first component shows the contribution of the attributes values and the second one is related to the contribution of the model coefficients. The second component is attributed to income discrimination.

By comparing both analyzed splits we conclude that differences between income distributions for Men and Women are significantly bigger than for Whites and Others. Values of income gender gap and race gap are shown in Table II.

TABLE II

Personal income distribution in USA, 2008. Predicted gender and race gaps.

| Quantile | 0.05 | 0.10 | 0.25 | 0.50 | 0.75 | 0.90 | 0.95 |
|---|---|---|---|---|---|---|---|
| gender gap, k$ ($\approx \widehat{\Delta}^{\beta}$) | 1.94 | 3.41 | 7.21 | 12.92 | 20.98 | 32.42 | 43.34 |
| race gap, k$ | −0.22 | 0.40 | 2.11 | 4.19 | 6.77 | 10.78 | 15.46 |
| attribute part $\widehat{\Delta}^{X}$ | 0.58 | 1.15 | 1.48 | 1.98 | 2.74 | 3.85 | 4.84 |
| coefficient part $\widehat{\Delta}^{\beta}$ | −0.80 | −0.75 | 0.63 | 2.21 | 4.03 | 6.93 | 10.62 |

The component related to the attributes is irrelevant for Men and Women. The observed differences between the distributions of incomes can be fully assigned to the differences of the models coefficients. Hence gender gap is determined only by "unit prices" of the attributes. The differences related to EDUCATION are the most important. Less important are AGE and RACE. The influence of ORIGIN is negligible. The differences increase with income and favor the first group (Men). This suggests the existence of *glass ceilings* phenomenon for women (cf. findings in [23]). The lower "evaluation" of personal characteristics for Women than for Men across the whole income distribution allows the conclusion that women are discriminated against men.

Our results cannot be directly compared to the results quoted in Introduction. We analyzed the full range of incomes sources, not just wages. According to our findings median income of women is 60% much as the median income of men. This ratio is the smallest being about 20% for very low incomes, 50% for 0.15 quantile and increases gradually with income up to 65%.

For Whites and Others both components are important. However in the case of the lowest incomes they cancel each other out. Values of the attributes are "better" for Whites but returns of the attributes ("unit prices") are higher for Others. Starting from 0.3 quantile both components are statistically significant and favor Whites. A share of the second component in the race gap increases from 40% to 70%. Therefore the level of discrimination of Others increases significantly with income.

The EDUCATION has the greatest positive influence on the differences between the income distributions, while being the most discriminatory attribute. For the comparison: the author of [23] found that an education is the primary contributor to differences in endowments and favors white workers. The SEX is also important and indicates on the better situation of white men. On the other hand, an influence of the AGE is negative in almost whole range of incomes, which means the return of AGE is bigger for Others than for Whites. Thus the AGE is attribute which discriminates Whites.

In the case of Whites and Others changes in the individuals' attributes and in the returns to these attributes contribute in the same direction to the observed increase in income inequities for average and high incomes. The contribution of the second component increases with incomes. Similar results were obtained in [23] on income inequality in Brazil. The author stated that racial wage differences tend to widen at higher wage quantiles, due to both larger differences in characteristics in favor of white workers and higher returns to those characteristics (the existence of glass ceilings for non-white workers). In Ref. [24] the author also applied quantile regression to issues concerning racial discrimination in the USA. He found that the differences in basic human capital characteristics explain about one-third of the differences in the level of wages and suggested that the amount of discrimination depends on the quantile at which it is evaluated (but he did not interpret the results as a glass ceiling effect).

We note finally that in the lowest part of the income distribution incomes of Others are higher and are equal to 167% of the incomes of Whites. This relations is reversed at 0.1 quantile and remains at about the same level of 85%. For comparison, one showed in [25] that a median black male worker earns 74% much as a median white male worker.

## References

[1] GAO, Women's Earnings: *Federal Agencies Should Better Monitor Their Performance in Enforcing Anti-Discrimination Laws*, GAO-08-799 (2008).

[2] R.G. Wood, M.E. Corcoran, P. Courant, *J. Labor Econ.* **11**, 417 (1993).

[3] J.G. Altonji, R.M. Blank, in: *Handbook of Labor Economics*, Eds. O.C. Ashenfelter, D. Card, Vol. 3, 1999, p. 3143.

[4] S. Boraas, W.M. Rodgers, *Monthly Labor Rev.* **126**, 9 (2003).

[5] *Highlights of Women's Earnings in 2014*, Bureau of Labor Statistics, 2015.

[6] S. Roberts, "For Young Earners in Big City, a Gap in Women's Favor", *The New York Times*, August 3, 2007.

[7] F.D. Blau, L.K. Kahn, *Acad. Manag. Perspect.* **21**, 7 (2007).

[8] F.D. Blau, L.M. Kahn, *J. Labor Econ.* **15**, 1 (1997).

[9] A. Drăgulescu, V.M. Yakovenko, *Europ. Phys. J. B* **20**, 585 (2001).

[10] A. Drăgulescu, V.M. Yakovenko, *Physica A* **299**, 213 (2001).

[11] M. Jagielski, R. Kutner, M. Pęczkowski, *Acta Phys. Pol. A* **121**, B-47 (2012).

[12] P. Łukasiewicz, A. Orłowski, *Physica A* **344**, 146 (2004).

[13] P. Łukasiewicz, K. Karpio, A. Orłowski, *Acta Phys. Pol. A* **121**, B-82 (2012).

[14] K. Karpio, G. Koszela, P. Łukasiewicz, A. Orłowski, *Quantitat. Meth. Econ.* **XV**, 403 (2014).

[15] K. Gajowniczek, K. Karpio, P. Łukasiewicz, A. Orłowski, T. Ząbkowski, *Acta Phys. Pol. A* **127**, A-38 (2015).

[16] R. Oaxaca, *Int. Econ. Rev.* **14**, 693 (1973).

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

[18] R. Koenker, G. Bassett, *Econometrica* **46**, 33 (1978).

[19] J.F. Machado, J. Mata, *J. Appl. Econom.* **20**, 445 (2005).

[20] J.M. Landmesser, "Decomposition of Differences in Income Distributions Using Quantile Regression", *Statist. Transit. New Series*, 2016, in press.

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

[22] `www.census.gov/`.

[23] P. Salardi, *Wage Disparities and Occupational Intensity by Gender and Race in Brazil, An Empirical Analysis using Quantile Decomposition Techniques*, Job Market Paper, University of Sussex, Brighton 2012.

[24] F.D. Blau, L.M. Kahn, *Am. Econ. Rev.* **82**, 533 (1992).

[25] J.D. Gwartney, J.E. Long, *Industrial. Labor Relat. Rev.* **31**, 336 (1978).