QUANTITATIVE METHODS IN ECONOMICS
Vol. XVI, No. 2, 2015, pp. 43 – 52

DECOMPOSITION OF DIFFERENCES BETWEEN PERSONAL INCOMES DISTRIBUTIONS IN POLAND

Joanna Małgorzata Landmesser
Department of Econometrics and Statistics, WULS-SGGW
joanna_landmesser@sggw.pl
Krzysztof Karpio
Department of Informatics, WULS-SGGW
krzysztof_karpio@sggw.pl
Piotr Łukasiewicz
Department of Informatics, WULS-SGGW
piotr_lukasiewicz@sggw.pl

Abstract: In this paper we study differences between personal incomes distributions in Poland in 2002 and 2012. The empirical data have been collected within the Household Budget Survey project. We used the Machado & Mata decomposition, which utilizes quantile regression. This method allowed us to investigate differences between income distributions in the whole range of values, going beyond simple average value decomposition. We evaluated influence of person's attributes on the differences of incomes distributions in 2002 and 2012. By decomposing the differences into the explained and unexplained components we got information about their causes. The differences described by the explained part are caused by different characteristics of samples. The unexplained part shows differences caused by the changes of attribute importance.

Keywords: quantile regression, Machado & Mata decomposition, counterfactual distribution

INTRODUTION

Nowadays one can observe significant development of various microeconomic decomposition methods. Based on the works [Oaxaca 1973] and [Blinder 1973] one elaborated techniques which went far beyond simple comparison of average values, for example decomposition of the variances or the

whole distributions. New techniques allowed to discover various factors influencing incomes distributions, as minimal wage [DiNardo et al. 1996]. They also have been useful in studying differences of incomes distributions for various group of people [Albrecht et al. 2003].

During the decomposition of differences between the distributions one utilizes so called counterfactual distributions. They are a mixture of an conditional distribution of the dependent variable and various distributions of the explanatory variables [see Juhn et al. 1993, DiNardo et al. 1996]. One of them, proposed in Machado & Mata [2005] decomposes differences of distributions using a quantile regression.

In this paper we compared incomes of the employees running the one-person households in 2012 with those in 2002. The data have been collected in the Household Budget Survey project in Poland. The aim of the work is to study differences between income distributions in year 2002 and 2012. By use of the [Machado & Mata 2005] decomposition method we investigated differences in the whole range of income values. The past studies in Poland were mostly focused on the decomposition of the average values by using the Oaxaca & Blinder method [e.g. Śliwicki & Ryczkowski 2014]. On the other hand, the studies of [Newell & Socha 2005] showed that many factors influence only high wages, localized in the high quatiles on the wages distribution. Similarly [Rokicka & Ruzik 2010] showed that differences between wages of men and women are the biggest in the right part of the distributions.

DECOMPOSITION METHODS

Oaxaca & Blinder decomposition of average incomes differences

We consider two groups of one-person households. The first one contains data for 2002, the second one – for 2012, denoted by T_1 and T_2 respectively. We also deal with the outcome variable y, and a set of predictors X. The variable y is individual income and predictors X are individual sociodemographic characteristics of households (people) such as sex, age, education and others. The idea of Oaxaca & Blinder decomposition can be applied whenever we need to explain the differences between the expected values of dependent variable y in two comparison groups [Oaxaca 1973, Blinder 1973]. We assume that the expected value of y conditionally on X is a linear function of X:

$$y_i = X_i \beta_i + v_i, \qquad i = T_1, T_2, \tag{1}$$

where X_i are characteristics of objects in the year i and β_i is the vector of parameters. The equation (1) can be estimated for both years:

$$\hat{y}_i = X_i \hat{\beta}_i, \qquad i = T_1, T_2. \tag{2}$$

The difference between the expected values of y in both years is as follows:

$$\Delta^{\mu} = E(y_{T_2}) - E(y_{T_1}) = E(X_{T_2})\beta_{T_2} - E(X_{T_1})\beta_{T_1}. \quad (3)$$

Based on (2) and (3) the decomposition of the difference Δ^{μ} has the following form:

$$\hat{\Delta}^{\mu} = \overline{X}_{T_2}\hat{\beta}_{T_2} - \overline{X}_{T_1}\hat{\beta}_{T_1} = \underbrace{(\overline{X}_{T_2} - \overline{X}_{T_1})\hat{\beta}_{T_2}}_{\hat{\Delta}^{\mu}_{\text{explained}}} + \underbrace{\overline{X}_{T_1}(\hat{\beta}_{T_2} - \hat{\beta}_{T_1})}_{\hat{\Delta}^{\mu}_{\text{unexplained}}} \quad (4)$$

The expression (4) is named Oaxaca & Blinder decomposition of average incomes differences [Oaxaca 1973, Blinder 1973]. The first component gives the effect of characteristics and expresses the difference of the potentials of both groups. The second component represents the effect of coefficients, typically interpreted as discrimination in numerous studies.

Decomposition of differences between distributions

The mean decomposition analysis may be extended to the case of differences along the whole distribution. Let $f^{T_1}(y)$ and $f^{T_2}(y)$ be the density functions for the variable *y* in 2002 and 2012, respectively. The distribution $f^{i}(y)$, $i = T_1, T_2$, is the marginal distribution of the joint distribution $\varphi^{i}(y, X)$:

$$f^{i}(y) = \int_{C(X)} \ldots \int \varphi^{i}(y, X)dX, \quad (5)$$

where *X* is a vector of individual characteristics observed [Bourguignon & Ferreira 2005]. Let $g^{i}(y|X)$ be the conditional distribution of *y*. Then one can (5) express as:

$$f^{i}(y) = \int_{C(X)} \ldots \int g^{i}(y|X)h^{i}(X)dX, \quad (6)$$

where $h^{i}(X)$ is the joint distribution of all elements of *X* in year *i*. The difference between the two distributions may be decomposed onto

$$f^{T_2}(y) - f^{T_1}(y) = [f^{T_2}(y) - f^{C}(y)] + [f^{C}(y) - f^{T_1}(y)], \quad (7)$$

where $f^{C}(y)$ is the counterfactual distribution, which can be constructed as

$$f^{C}(y) = \int_{C(X)} \ldots \int g^{T_2}(y|X)h^{T_1}(X)dX. \quad (8)$$

The first component in (7) gives the effect of the unequal different personal characteristic's distributions in 2012 and 2002. The second component describes the inequalities between two distributions of *y* conditional on *X*. The difference with respect to the Oaxaca & Blinder decomposition is that this decomposition refer to full distributions, rather than just to their means.

Quantile regression

The linear regression assumes the relationship between the regressors and the outcome variable based on the conditional mean function. This gives only a partial insight into the relation. The quantile regression allows the description of the relationship at different points in the conditional distribution of *y* [Koenker & Bassett 1978].

We consider the relationship between the regressors and the outcome using the conditional quantile function:

$$Q_\theta(y|X) = \Phi^{-1}_{y|X}(\theta, X) = X\beta(\theta), \tag{9}$$

where $Q_\theta(y|X)$ – the θ^{th} quantile of a variable *y* conditional on covariates *X*, $\theta \in (0,1)$; $\Phi_{y|X}$ – the cumulative distribution of the conditional variable $y|X$. We assume that all quantiles of *y* conditional on *X* are linear in *X*. The quantile regression estimator for quantile θ minimizes the sum:

$$\sum_{i=1}^{n} \rho_\theta(y_i - X_i\beta) \to \min, \tag{10}$$

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

The sum (10) gives asymmetric penalties for over and under prediction. For each quantile other parameters are estimated. We interpret these coefficients as the returns to different characteristics *X* at given quantiles of the distribution of *y*. The standard errors of parameters are calculated using bootstrap method [Gould 1992].

Machado & Mata decomposition of differences in distributions

Machado & Mata [2005] have used quantile regression in order to estimate counterfactual unconditional income distributions. The unconditional quantile is not the same as the integral of the conditional quantiles. Therefore, authors provide a simulation based estimator where the counterfactual distribution is constructed from the generation of a random sample. The approach is the as follows:

(1) generate a random sample of size *m* from a $U[0,1]$: $\theta_1, \theta_2, \ldots, \theta_m$;

(2) using the dataset for T_1 estimate *m* different quantile regression $Q_{\theta_i}(y|X_{T_1})$, obtaining coefficients $\hat{\beta}_{T_1}(\theta_i), i = 1,...,m$;

(3) generate a random sample of size *m* with replacement from X_{T_1}, denoted by $\{X^*_{T_1 i}\}, i = 1,...,m$;

(4) $\{y^*_{T_1 i} \equiv X^*_{T_1 i}\hat{\beta}_{T_1}(\theta_i)\}, i = 1,...,m$ is a random sample from the unconditional distribution $f^{T_1}(y)$.

Alternative distributions could be estimated by drawing X from another distribution and using different coefficient vectors. To generate a random sample from the income density that would prevail in group T_2 and covariates were distributed as $h^{T_1}(X)$, we follow the steps (1), (2) from the previous procedure for T_2 and then:

(3) generate a random sample of size m with replacement from X_{T_1}, denoted by $\left\{X^*_{T_1 i}\right\},\ i=1,...,m$;

(4) $\left\{y^{C*}_{T_2 i} \equiv X^*_{T_1 i}\hat{\beta}_{T_2}(\theta_i)\right\},\ i=1,...,m$ is a random sample from the counterfactual distribution $f^C(y)$.

The Machado & Mata decomposition of the difference between the income densities in two years for each quantile is as follows:

$$\begin{aligned}\hat{\Delta}^{\theta} &= Q_\theta(y^*_{T_2}\left|X^*_{T_2}\right.) - Q_\theta(y^*_{T_1}\left|X^*_{T_1}\right.) = \\ &= Q_\theta(y^*_{T_2}\left|X^*_{T_2}\right.) - Q_\theta(y^{C*}_{T_2}\left|X^*_{T_1}\right.) + Q_\theta(y^{C*}_{T_2}\left|X^*_{T_1}\right.) - Q_\theta(y^*_{T_1}\left|X^*_{T_1}\right.) = \\ &= \underbrace{(X^*_{T_2} - X^*_{T_1})\hat{\beta}_{T_2}(\theta)}_{\hat{\Delta}^{\theta}_{\text{explained}}} + \underbrace{X^*_{T_1}(\hat{\beta}_{T_2}(\theta) - \hat{\beta}_{T_1}(\theta))}_{\hat{\Delta}^{\theta}_{\text{unexplained}}}\end{aligned} \tag{11}$$

To estimate standard errors for the estimated densities we repeat the Machado & Mata procedure many times and generate a set of estimated densities.

EMPIRICAL DATA

The data were collected in the Household Budget Survey project for 2002 and 2012, group T_1 and T_2 respectively. The analyzed data regards households running by one person whose main source of earning comes from a work as an employee. The annual disposable incomes (variable *INC*, in thousands of PLN) in 2012 were compared with those obtained in 2002. The incomes in 2002 during the analysis were expressed in prices in 2012 (variable *INCREAL*). The sample consisted of 834 and 1594 persons in 2002 and 2012 respectively. For each person: sex, age, education, place of residence, type of labor position have been obtained. Based on the obtained attributes one defined the following describing variables:

SEX (0 – woman, 1 – man),
AGE (years),
EDU (education, 1–9, 1 – primary, . . ., 9 – tetriary),
RES (residence, 1–6, 1 – village, . . ., 6 – town ≥ 500k of inhabitants),
POS (0–1, 0 – manual labor position, 1 – non-manual labor position).
Features of the variables have been collected in the Table 1.

Table 1. The mean values and the standard deviations for the selected variables

	Whole sample	2002	2012
Number of observation	2428	834	1594
INC	27.92 (21.32)	19.76 (17.92)	32.19 (21.71)
INCREAL	30.09 (22.58)	26.07 (23.64)	32.19 (21.71)
SEX (% men)	40.28	38.49	41.22
AGE	41.65 (12.12)	40.65 (11.29)	42.16 (12.51)
POS (% non-manuals)	67.26	64.51	68.70

Source: own calculations

RESULTS

We compared the personal incomes distributions for years 2002 and 2012. In the first step of the analysis the Oaxaca & Blinder decomposition has been applied for the average values. The results are listed in the Table 2.

Table 2. The Oaxaca & Blinder decomposition of the average incomes differences

Average *INCREAL* in 2002			26.072
Average *INCREAL* in 2012			32.189
Raw gap			6.117
Aggregate decomposition			
Explained effect			1.302
Unexplained effect			4.816
% explained			21.3
% unexplained			78.7
Detailed decomposition			
explained component		unexplained component	
SEX	0.217	*SEX*	6.251
AGE	0.199	*AGE*	0.445
EDU	1.038	*EDU*	10.852
RES	–0.369	*RES*	1.462
POS	0.216	*POS*	–6.804
const	0.000	*const*	–7.390
Total	1.302	Total	4.816

Source: own calculations

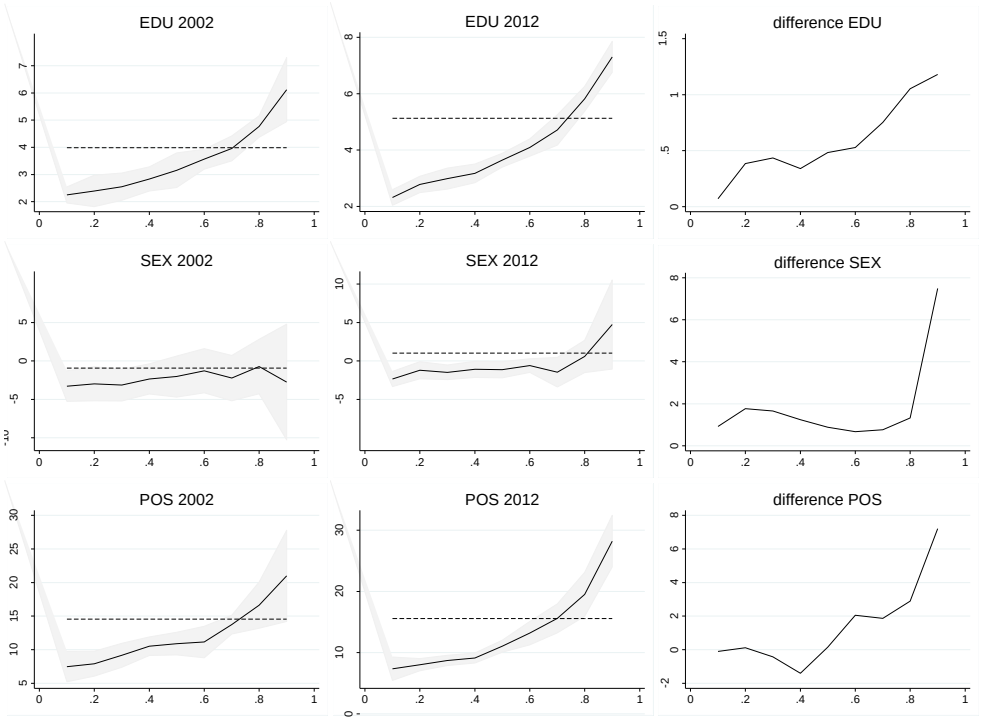
One can observe the positive difference between average values of the real incomes in 2012 and 2002. In the next step one tried to explain the observed difference. Using the decomposition method, one evaluated strength of the influence of the analyzed factors onto the average incomes. Generally, the differences are explained by the factors being studied in 21.3%. The *SEX*, *AGE*, *EDU*, *POS* were positively correlated with the change of the average value

of incomes. However the biggest influence exhibited the *EDU* attribute. The increase of the average incomes are explained the most by the big increase of the education level in 2002 to 2012. On the other hand the *RES* exhibits negative correlation with the change of the average income.

The remaining 78% of average income changes are unexplained by the regression models being used. The unexplained part is assigned to the changes of the estimated parameters' influence onto the average income between 2002 and 2012. Such a different "labor market value" of parameters in the two years is the main source of the unexplained part of the model.

In the next step of the analysis the quantile regression models have been estimated. The influence of the selected attributes (measured by the model coefficients) onto the various quantiles of the income distribution are summarized in the Figure 1. The strength of the influence is presented as a function of the quantile range for both years. The differences of the results for both years are also presented.

Figure 1. The influence (vertical axis) of the selected attributes on the income distribution as a function of the quantile range (horizontal axis)



* The shaded areas represent 95% confidence intervals. The dashed lines represent results of the classical linear regression model.

Source: own preparation

The education has the positive impact on the income distribution in the whole range of values. However its importance increases with the quantile range, being the biggest for the highest incomes. The same behavior is observed for both years. The bigger influence on the income distribution is exhibited by the *POS* variable. This is directly related to the bigger incomes gained by the persons on the non-manual positions. However, instead of such a big overall influence we observe its rise along with the income distribution. On the other hand for the *SEX* variable we observe constant, negligible influence on the income for both years, especially in the right part of income distributions.

Studying the unexplained components of the quantile differences, related to the different parameters values in both years, one can conclude that for *EDU* as well as for *POS* they are quite similar. They increase with the values of the quantiles. However the differences for the *POS* parameters are relatively small for the left part of the distribution rising strongly for the higher quantiles.

On the other hand the unexplained part of the *SEX* parameters is negligible. The big rise of the differences for the last quantile is within the statistical error.

In the last step of the analysis one performed the decomposition of differences between income distributions. The differences are expressed as the sum of the explained and unexplained components along the whole income distributions. The Machado & Mata method has been used to estimate quantile regressions for 19 percentiles. The results for deciles are presented in the Table 3. The errors have been evaluated using the bootstrap method.

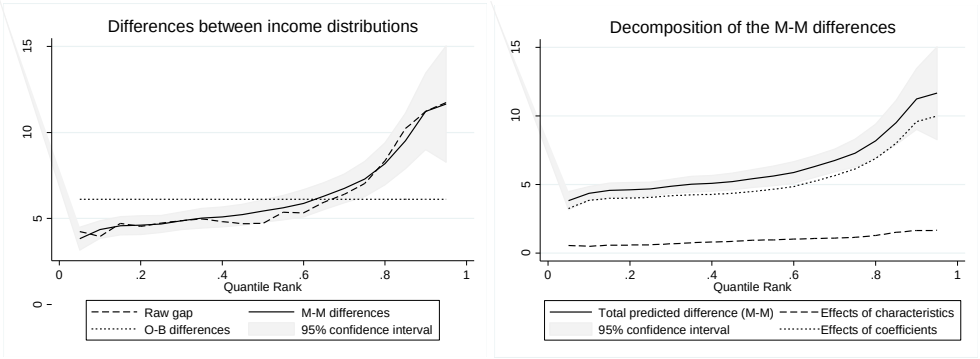
Table 3. The results of the Machado & Mata decomposition of the differences of the incomes distributions for 2002 and 2012

Quantile	Raw gap	Diference M-M	Explained part	Unexplained part	Explained %	Unexplained %
0.10	3.9567	4.3506 (0.2590)	0.5032 (0.3057)	3.8473 (0.2559)	12%	88%
0.20	4.5382	4.6094 (0.2784)	0.5934 (0.2674)	4.0160 (0.2707)	13%	87%
0.30	4.8726	4.8689 (0.2608)	0.6819 (0.2748)	4.1870 (0.2473)	14%	86%
0.40	4.8204	5.0937 (0.2955)	0.8079 (0.3076)	4.2858 (0.2825)	16%	84%
0.50	4.7140	5.4314 (0.3305)	0.9372 (0.3436)	4.4942 (0.3103)	17%	83%
0.60	5.3322	5.8808 (0.4041)	1.0257 (0.4118)	4.8551 (0.3791)	17%	83%
0.70	6.4007	6.7484 (0.4282)	1.0990 (0.4992)	5.6494 (0.4517)	16%	84%
0.80	8.3482	8.1739 (0.6231)	1.2908 (0.6968)	6.8831 (0.5822)	16%	84%
0.90	11.2275	11.228 (1.1327)	1.6515 (1.1115)	9.5765 (1.0202)	15%	85%

Source: own calculations

The results in Table 3 are also presented in the Figure 2. The left plot contains the raw differences between income distributions and the model predictions. The right plot shows the decomposition of the differences onto the explained and unexplained parts.

Figure 2. The results of the Machado & Mata decomposition of the differences of the incomes distributions for 2002 and 2012



Source: own preparation

The whole model approximates the data well. The differences between income distributions are rising with incomes. Their decomposition onto the explained and unexplained parts indicate low share of its explained part (12% to 17%). The share of the model's unexplained part is relatively high (83% to 88%) and increasing what indicates on the increase of the "labor market value" of the households' attributes. However, the explained part of the model also increases with income level.

SUMMARY

In this paper one studied differences between personal's income distributions in Poland in 2002 and 2012. The households of the single employers have been taken into account. The Oaxaca & Blinder and Machado & Mata decompositions of the average values and the whole incomes distributions respectively have been used. The Oaxaca & Blinder decomposition showed the positive influence of the most analyzed variables (*SEX*, *AGE*, *EDU*, *POS*) on the average income differences. The only variable with negative impact was *RES* what indicates on a "shift of big incomes towards smaller town". The Machado & Mata decomposition showed the increase of the differences between income distributions with the value of income. The differences were mostly caused by change of "labor market values" of the households' characteristics, described by the unexplained part of the model. These changes were greater when going towards big incomes.

The observed change might be caused by global processes in the European economy. It is known that such events took place in the past (e.g. financial crisis 2007-2009). Of course such a hypothesis needs to be confirmed through further studies. The future studies can also cover a detailed decomposition, which may exhibit the influence of the attributes on the whole income distribution.

REFERENCES

Albrecht J., Björklund A., Vroman S. (2003) Is There a Glass Ceiling in Sweden? Journal of Labor Economics, 21, pp. 145-177.

Blinder A. (1973) Wage Discrimination: Reduced Form and Structural Estimates, Journal of Human Resources, 8, pp. 436-455.

DiNardo J., Fortin N. M., Lemieux T. (1996) Labor Market Institutions and the Distribution of Wages, 1973-1992: A Semiparametric Approach, Econometrica, 64, pp. 1001-1044.

Gould W.W. (1992) Quantile Regression with Bootstrapped Standard Errors, Stata Technical Bulletin, 9, pp. 19-21.

Juhn Ch., Murphy K. M., Pierce B. (1993) Wage Inequality and the Rise in Returns to Skill, Journal of Political Economy, 101, pp. 410-442.

Koenker R., Bassett G. (1978) Regression Quantiles, Econometrica, 46, pp. 33-50.

Machado J. F., Mata J. (2005) Counterfactual Decomposition of Changes in Wage Distributions Using Quantile Regression Journal of Applied Econometrics, 20, pp. 445-465.

Newell A., Socha M. (2005) The Distribution of Wages in Poland, IZA Discussion Paper, 1485.

Oaxaca R. (1973) Male-Female Wage Differentials in Urban Labor Markets, International Economic Review, 14, pp. 693-709.

Rokicka M., Ruzik A. (2010) The Gender Pay Gap in Informal Employment in Poland, CASE Network Studies and Analyses, 406.

Śliwicki D., Ryczkowski M. (2014) Gender Pay Gap in the micro level – case of Poland, Quantitative Methods in Economics, Vol. XV, 1, pp. 159-173.