Schmollers Jahrbuch 135 (2015), 13 – 22
Duncker & Humblot, Berlin

A Semiparametric Analysis of Conditional Income Distributions

By Alexander Sohn, Nadja Klein, and Thomas Kneib*

Abstract

We explore the application of structured additive distributional regression for the analysis of conditional income distributions in Germany following the reunification using the German Socio Economic Panel (SOEP) database. This methodology allows us to explore both between and within income inequality at a highly disaggregated level. Using a bootstrapped version of the Kolmogorov-Smirnov test, we find that conditional personal income distributions can generally be modelled using a mixture distribution entailing the three parameter Dagum distribution.

JEL-Classification: C13, C21, D31, J31

1. Introduction

The SOEP panel database has been used extensively to inquire into the development of inequality in Germany (see among others Biewen, 2000; Bach et al., 2009; Grabka / Kuhn 2012). However, this literature has a dichotomic streak. On the one hand, the rising inequality has been analysed at a disaggregated level using regression techniques mostly focussing on the divergence of conditional means. On the other hand, the overall divergence of incomes has been observed in widening aggregate income distributions, mostly conducted at the national level. Yet, we have to admit that "we know relative little about the determinants of residual inequality" (Acemoglu, 2002), i.e. the inequality of income distributions at the disaggregated level, which has been noted to have increased "very much in tandem with overall inequality" (ibid.).

One reason for this dichotomy is found in the focus of conventional regression techniques which allow for analyses at a highly disaggregated level but which are geared towards point estimates (means, quantiles, etc.) and not to-

* Funding from the German Research Foundation through the projects KN 922/4-1/2 is gratefully acknowledged. We thank the DIW for the data and their support by expert knowledge on the SOEP data.

wards estimating whole conditional distributions of the variable under consideration. By contrast, the studies which focus on distributional aspects are constrained to a highly aggregated level due to the size of the SOEP whose sample size is insufficient to allow for the independent estimation of a large number of income distributions for different covariate sets.

In this shortened version of our paper (Sohn et al., 2014), we discuss the possibility to bridge the analytical gap by considering whole conditional income distributions (CIDs) in a regression framework, which allow for the contemplation of income distributions at the disaggregated level. For this purpose, we introduce structured additive distributional regression (see Klein et al., 2013) to the analysis of conditional income distributions. Specifically, we investigate whether CIDs can be modelled by a mixture distribution entailing the Dagum distribution and fitting into the framework of generalised additive models of location, scale and shape (Rigby / Stasinopoulos, 2005) with estimation based on penalised maximum likelihood approaches. Our study thus adds to recent advances in the literature on the estimation of CIDs, like Biewen / Jenkins (2005), Quintano / D'Agostino (2006) and Chernozhukov et al., (2013).

The structure of this paper is as follows: In the next two sections, we introduce structured additive distributional regression for CIDs with a special focus on the estimation of a mixed discrete-continuous version of the Dagum distribution. Thereby we estimate the CIDs for males with respect to the three explanatory variables age, educational attainment and region. In Section 4, we employ a bootstrapped Kolmogorov-Smirnov test to check whether the proposed mixture distribution provides an adequate fit for the analysis of our CIDs. Subsequently, we go on discussing some aspects of inference in a distributional regression framework in our application with a specific focus on the relation of skill-biased technological change to the variables age and region. In the last section we conclude.

2. Conditional Income Distributions

Using the data available in the German Socio-Economic Panel (SOEP) database (see Wagner et al., 2007) we consider the personal labour income for the years 1992 and 2010 as defined in the gross market income definition from Bach et al. (2009). Thereby our income definition entails wage income (including social security contributions) both from the private and the public sector as well as business income from agriculture and forestry, unincorporated enterprise and self-employment. However, contrary to Bach et al. (2009) we exclude capital income. Our labour income definition thus entails practically all income types derived from the factor labour. Consequently, we implicitly incorporate both changes in wage rates and changes in working time. Since we aim to analyse the evolution of labour related income inequality at large, this seems the

most appropriate definition to use. Following one of the most popular decomposition categories, namely decomposition by population groups, we will condition our income distributions on three demographic variables – namely region, education and age. We consider region as a binary variable differentiating between the geographical region of the former Federal Republic of Germany and the former German Democratic Republic (entailing both former East and West Berlin). Following Acemoglu (2002), we consider education as a binary variable which is unity for everybody who has obtained at least a university degree and zero otherwise. Conversely to the standard approach in the literature, age is not considered in a highly discretised manner but as a continuous variable. Thereby we evade implicit homogeneity assumptions within artificially constructed age groups which may cover up important dynamics within the groups.

For the estimation of the CIDs we employ structured additive distributional regression, which models the parametric distributions with respect to a selected set of covariates. While we acknowledge that "the use of the parametric approach to distributional analysis runs counter to the general trend towards the pursuit of non-parametric methods, […]" (Cowell, 2000, p. 145) we perceive the parametric approach as a form of regularisation itself which by imposing a structure lends stability to the estimation process. Moreover, we concur with Morduch / Sicular (2002, p. 93) that it is often "necessary to impose more structure in order to draw sharp conclusion". And last but not least it should be noted that parametric models are better suited for robustness checks (see Silber, 1999, p. 8). Naturally, the applicability of any parametric approach hinges on the "agreement between the model being identified and the actual observations" (Dagum, 1977). In other words it is critical to find a parametric model which is able to provide a sufficiently "good fit of the whole range of the distribution" (ibid.) for all the covariate sets of interest.

To ensure an adequate fit that also captures zero- and precarious incomes, we found that a mixture distribution consisting of two probability masses for zero-incomes and precarious incomes (which we define as an annual income below the 4800 €) and a Dagum distribution provides the most reasonable fit, out of a wide list of distributions suggested for aggregate income distributions (based on Kleiber / Kotz, 2003; Chotikapanich, 2008). Each CID thus takes the following form:

$$f(y \mid \pi_0, \pi_{pr}, \theta_1, \ldots, \theta_K) = \mathbb{1}_{\{y=0\}}\pi_0 + \mathbb{1}_{\{0<y\leq 4800\}}\pi_{pr} + (1 - \pi_0 - \pi_{pr})t(y - 4800 \mid a, b, p), \tag{1}$$

where $\mathbb{1}_{\{y=0\}}$ is an indicator function which takes unity if the income is zero, while $\mathbb{1}_{\{0<y\leq 4800\}}$ takes unity if the person receives a precarious income. The corresponding probabilities are π_0 and π_{pr}. The truncated conditional density

function is denoted by $t(y - 4800 \mid a, b, p)$ and assumed to be a Dagum distribution with parameters a, b and p. To improve the fit of the two distributions and to evade problems with identification, we shifted the truncated income distribution to the right such that their support is restricted to the domain $(4800, \infty)$. For notational brevity we use $\tilde{y} = y - 4800$ in the following.

3. Estimating Conditional Income Distribution

The parametric CID is described by five parameters (π_0, π_{pr}, a, b and p). The parameters π_0 and π_{pr} can be estimated by simple sequential logit (see Fahrmeir et al., 2013, Ch. 6). For estimating the truncated continuous part of the income distribution, we use the `gamlss` package in `R`, which employs backfitting algorithms for the maximisation of the (penalised) likelihood (see Rigby / Stasinopoulos, 2005). In case of the Dagum distribution, we obtain the following set-up:

$$t(\tilde{y} \mid a, b, p) = \frac{ap\tilde{y}^{ap-1}}{b^{ap}[1 + (\tilde{y}/b)^a]^{p+1}}. \tag{2}$$

Each parameter is estimated in an additive manner:

$$\log(a) = \eta_a = s_{1a}(age) + Hs_{2a}(age) + Os_{3a}(age) + HOs_{4a}(age), \tag{3}$$

$$\log(b) = \eta_b = s_{1b}(age) + Hs_{2b}(age) + Os_{3b}(age) + HOs_{4b}(age), \tag{4}$$

$$\log(p) = \eta_p = s_{1p}(age) + Hs_{2p}(age) + Os_{3p}(age) + HOs_{4p}(age), \tag{5}$$

where s denotes a smooth function modelling the effect of age in a non-linear way. We thereby follow the notion of Lemieux (2003) that the relation between earnings and experience (or age) is not linear. The variable H is binary and unity if we consider the CID for people with higher education. The variable O is also binary unity if the CID is for people living in East Germany. The fourth term is an interaction between education, region and age. Hence, we allow for differing effects of age for all four combinations of education and region.

Using the five parameter estimates, we are able to obtain a fully specified CID. From there it is straightforward to get estimates for any desired distribution measure, like the mean, standard deviation, skewness, etc. Also estimates for other economic measures of interest like inequality measures such as the Gini coefficient or the Theil Index can easily be calculated. See Sohn et al., (2014) for a more detailed discussion on this issue.

4. Assessing Conditional Income Distributions

As noted previously, our estimation strategy hinges on the assumption that our parametrically specified CIDs provide an adequate fit to the data. Before going on to interpreting the estimates, we test the hypothesis that the parametric fit obtained by structured additive distributional regression is sufficiently close to the true distribution. More formally we test the hypothesis

$$H_0 : f(y) = f_0(y, \theta), \tag{6}$$

where $f(y)$ is the observation-generating probability density function (p.d.f.) and $f_0(y, \theta)$ is the parametric mixture distribution thought to model the data for every possible combination of our covariates.

We test the hypothesis by using a bootstrapped version of the Kolmogorov-Smirnov test and Monte Carlo simulations to obtain the distribution of the test statistic as suggested by Andrews (1997). The test statistic for this test is given by

$$D_n = \sup_y \mid F_0(y, \theta) - S_n(y) \mid, \tag{7}$$

where $F_0(y, \theta)$ denotes the cumulative density function (c.d.f.) of our parametric fit for the CIDs, constructed from our estimates from Equations 3–5. The empirical cumulative distribution function for observations $y_1, \ldots, y_n$ is denoted by $S_n(y)$, with n being the sample size of the given subpopulation under consideration. For the subpopulations, we considered each education-region combination at all forty possible 40 years of age in our sample, such that we consider 160 subpopulations overall. Naturally this implies that for several subpopulations we only have very few observations available and consequently a high degree of statistical uncertainty attached to our hypothesis test.

The distribution of the test statistic was obtained by parametric bootstrap with 100,000 simulation samples of size n for each subpopulation yielding a distribution of the test statistic under the null hypothesis. Using this procedure we obtained the p-values for each subpopulation which are provided in Sohn et al., (2014).

For the p-values, we expect to see a 5 % share of observations to show a test statistic with a corresponding p-value smaller than 0.05. On average over both time periods we get an average rejection rate of 0.056, which is just above the 0.05 we would expect. While it must be noted that for the males without higher education in the West and in the East in 1992, we generally have slightly too high shares of rejections, the results do not imply that our model of the distributions must be rejected on grounds of the empirical observations. We therefore conclude that structured additive regression with a mixed discrete-continuous version of the Dagum distribution adequately models both the continuous and

the truncated parts of the CIDs under consideration and consequently allows us to model whole conditional income distributions.

5. Analysing Within-Group Inequality

From the estimated CIDs, a variety of distribution measures can be deduced and analysed. In this short paper, we focus on the matter of residual inequality and thus concentrate on the within-group inequality as measured by the Theil index of the CIDs.

Figure 1 displays the Theil indices deduced from the CIDs for labour incomes in 1992. The solid line marks the penalised maximum likelihood estimate, while the dotted lines are bootstrapped 95 % pointwise confidence intervals. As we can see, the within-group inequality changes over the age-span for both education levels and in West and East. Generally, we observe a U-shaped relation, so that within-group inequality is markedly higher for men below 30 and above 55. This is hardly surprising, since at a young age (due to education / vocational training) zero- and low incomes are common. At a higher age, retirement rates and reduced work-schemes increase substantially causing a wider dispersion of the CID and hence inflating the Theil index. The differences in the Theil index for both region and education levels and a given age are slim and non-significant at the 5 % level.

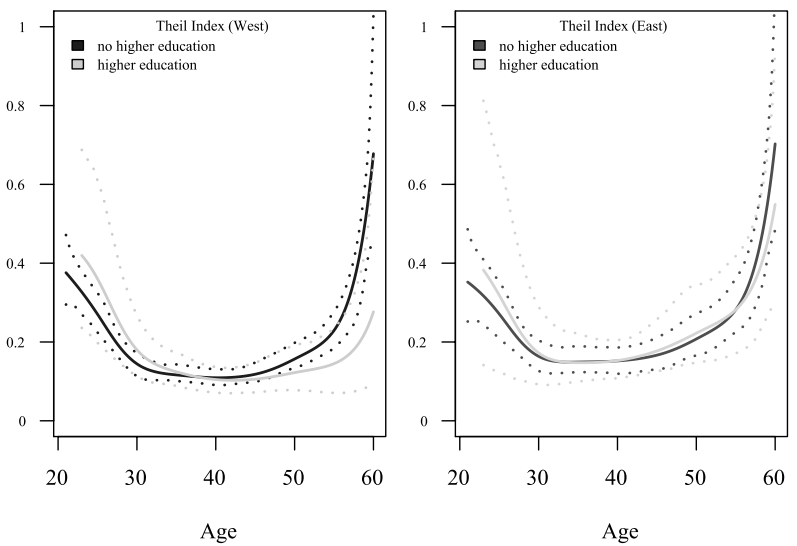

Figure 1: Theil Indices of CIDs in 1992

Figure 2 displays the corresponding Theil indices for labour incomes in 2010. As can be observed, the general U-shape over the age-span has persisted over time. In the West of Germany slight increases in the inequality can be observed, which are largest among young men with higher education. This highlights that inequality among university graduates has increased in recent years, indicating that the demarcation lines drawn between winners and losers of the much discussed skill-biased technological change cross the threshold of the university campus with graduates faced by increasingly precarious employment prospects. In the East, men without higher education see an increase in within-group inequality which is much more drastic. The main thrust of this increased inequality is the risen share of zero- and precarious incomes. This rise on the lower range of the distribution induces a much more pronounced positive skew to the CID and increases the Theil index. Although less pronounced, a similar rise is observed for young men without higher education in the West. By contrast, at the other side of the age-span our estimation results indicate an opposite dynamic of decreasing within-group inequality.

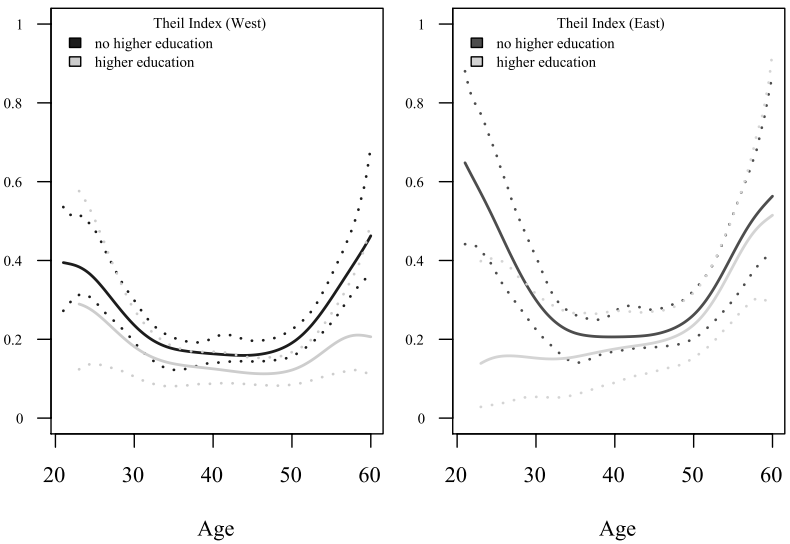

Figure 2: Theil Indices of CIDs in 2010

A comprehensive discussion of the consequences of this increase of within-group inequality is beyond the scope of this paper. As an exemplary analysis of the consequences of our findings, it may be noted that in-group inequality plays an important part for group-identity and alienation found within a society (see Duclos et al., 2004). A rising within-group inequality as observed in particular

for the young men without higher education in the East would thereby erode group-identity and lead to growing isolation within that cohort. Indeed it could be argued that especially the growth of within-group inequality rather than inequality at large undermines the very fabric of our society, as it leads to the erosion of group-identity at the disaggregated level. Thus, the analysis of within-group inequality could shed some light on the phenomenon of decreasing trust in modern societies (see Misztal, 2013).

For such far reaching inferences from the data it is obviously of importance, whether the changes are statistically significant at the usual levels. As the confidence intervals for the Theil index indicate, considerable uncertainty remains. Given the complexity of the estimation of whole conditional distributions and the relative scarcity of data available, this is hardly surprising. Nonetheless, thanks to the regularisation induced by regression, a multitude of effects can be considered allowing for the analysis of aspects like inequality at a highly disaggregated level. This may yield new important insights on the nature of the change of inequality that cannot be observed by conventional regression or aggregate distribution analysis.

6. Conclusion and Outlook

At the outset of this article, we highlighted the need for the analysis of CIDs. Using the SOEP database and structured additive distributional regression, we showed that it is possible to estimate CIDs with respect to a set of variables, both continuous and discrete in an additive set-up. Specifically, we regressed German labour incomes on the continuous variable age, and two discrete binary variables for education and region and found that a mixture consisting of two discrete probability masses and a Dagum distribution provides an appropriate fit to the data. Subsequently, we considered the development of residual inequality with respect to the three explanatory variables. We found conditional inequality, as measured by the Theil index of the CIDs, to be especially high among young men in the East for whom inequality also seems to have increased considerably between 1992 and 2010. While further research on the consequences of the growing within-group inequality needs to be conducted, this finding may contribute to explain the decline of trust and solidarity in modern societies.

Clearly much work remains to be done in the field of modelling conditional income distributions / residual inequality and this working paper is no more than a first attempt at addressing the issue. Yet, this analysis shows that the SOEP allows not only for the analysis of incomes using conventional regression techniques or aggregate distribution analysis but also the analysis of conditional income distributions which shed a new perspective on what Paul Krugman (2007) called the great divergence of incomes.

References

Acemoglu, D. (2002): Technical Change, Inequality and the Labor Market, Journal of Economic Literature, 40(1), 7 – 72.

Andrews, D. W. K. (1997): A Conditional Kolmogorov Test, Econometrica, 65(5), 1097 – 1128.

Bach, S. / *Corneo*, G. / *Steiner*, V. (2009): From Bottom to Top: the Entire Income Distribution in Germany, 1992 – 2003, Review of Income and Wealth, 55(2), 303 – 330.

Biewen, M. (2000): Income Inequality in German during the 1980s and 1990s, Review of Income and Wealth, 46(1), 1 – 19.

Biewen, M. / *Jenkins*, S. P. (2005): A framework for the decomposition of poverty differences with an application to poverty differences between countries, Empirical Economics, 30(2), 331 – 358.

Chernozhukov, V. / *Fernandez-Val*, I. / *Melly*, B. (2013): Inference on Counterfactual Distributions, arXiv:, 0904.0951v6[stat.ME].

Chotikapanich, D. (ed.) (2008): Modeling Income Distributions and Lorenz Curves, Springer, New York.

Cowell, F. A. (2000): Measurement of Inequality, in: A. B. Atkinson / F. Bourguignon (eds.), Handbook of income distribution, pp. 87 – 166, Elsevier, Amsterdam.

Dagum, C. (1977): A New Model of Personal Income Distribution: Specification and Estimation, Economie Applicée, 30, 413 – 437.

Duclos, J. Y. / *Esteban*, J. / *Ray*, D. (2004): Polarization: Concepts, Measurement, Estimation, Econometrica, 72(6), 1737 – 1772.

Fahrmeir, L. / *Kneib*, T. / *Lang*, S. / *Marx*, B. D. (2013): Regression: Models, Methods and Applications, Springer, Berlin and New York.

Grabka, M. M. / *Kuhn*, U. (2012): The Evolution of Income Inequality in Germany and Switzerland since the turn of the millenium, Swiss Journal of Sociology, 38(2), 311 – 334.

Kleiber, C. / *Kotz*, S. (2003): Statistical Size Distributions in Economics and Actuarial Sciences, Wiley, Hoboken.

Klein, N. / *Kneib*, T. / *Lang*, S. (2013): Bayesian Structured Additive Distributional Regression, Working Papers in Economics and Statistics, 2013 – 23.

Krugman, P. R. (2007): The conscience of a liberal, W. W. Norton & Co., New York, 1st edition.

Lemieux, T. (2003): The "Mincer Equation" Thirty Years after Schooling, Experience, and Earnings, Center for Labor Economics, University of California, Working Paper No. 62.

Misztal, B. (2013): Trust in Modern Societies: The Search for the Bases of Social Order, Wiley, Hoboken.

Morduch, J. / *Sicular*, T. (2002): Rethinking Inequality Decomposition, with Evidence from Rural China, The Economic Journal, 112(476), 93 – 106.

Quintano, C. / *D'Agostino*, A. (2006): Studying Inequality in Income Distribution of Single-Person Households in Four Developed Countries, Review of Income and Wealth, 52(4), 525 – 546.

Rigby, R. A. / *Stasinopoulos*, D. M. (2005): Generalized Additive Models for Location, Scale and Shape, Journal of the Royal Statistical Society: Series C (Applied Statistics), 54(3), 507 – 554.

Silber, J. (1999): Introduction – Thirty Years of Intensive Research on Income Inequality Measurement, in: J. Silber (ed.), Handbook of Income Inequality Measurement, volume 1-18, Kluwer Academic, Boston.

Sohn, A. / *Klein*, N. / *Kneib*, T. (2014): A New Semiparametric Approach to Analysing Conditional Income Distributions, SOEPpapers on Multidisciplinary Panel Data Research, No. 676.

Wagner, G. G. / *Frick*, J. R. / *Schupp*, J. (2007): The German Socio-Economic Panel Study (SOEP) – Scope Evolution and Enhancements, Schmollers Jahrbuch, 127(1), 139 – 169.