Conditional Income Distribution Across Age and Race in the U.S. in 2016

Dingyi Lai[1]

Dept. of Statistics, Humboldt University
dingyi.lai@student.hu-berlin.de

Abstract. Conditional income distribution is critical to understand how income varies in different backgrounds. Such panorama is beneficial for decomposition of income inequality, and can offer a better guide for policy-making accordingly in a further step. In this term paper, theoretical background of generalized additive models for location, scale and shape (GAMLSS) and generalized beta distribution of the second kind (GB2) is introduced, followed by an empirical estimation based on SCF+ data in 2016 in the United States. The parameters are selected via optimization under GAIC criterion. Furthermore, model is evaluated by analysis of quantile residual analysis. Fitted parameters of model are transformed to a grouped histogram across 4 typical ages and 3 races and a 3D income density plot across all ages and races. In conclusion, racial discrepancy widens as age increases, and generally, the probability of earning higher income grows at first and then drops over the age.

Keywords: conditional income inequality · distributional regression model

1 Introduction

Inequality is always a subject of intense scholarly debate. Beyond expectation, Piketty et al. (2014) conclude that two world wars and the following public policies help reduce inequalities in the twentieth century, and inequalities rise again since 1970s and 1980s, which is confirmed by Moritz et al. (2020). Unlike wars, financial crisis does not prevent structural increase of inequality in the United States, but affects stock market and slowers the development of inequality (Piketty et al., 2014). Here comes a critical question: How inequality is structured? In other words, in which dimensions the inequality exists? Generally speaking, Pyatt (1976) gives a simple interpretation of Gini coefficient, and Theil (1967) introduces Theil index as its alternative. Other original works related to decomposition of inequality are explored by Shorrocks (1980, 1982, 1983, 1984). Tausch et al. (2007) bring forward that income inequality can be decomposed in terms of subgroups, income sources, causal factors and others, while Piketty et al. (2014) take another perspective which are of labor, capital and transfer. Besides income, wealth is another significant index to measure economic resources owned by individual or household, estimated by capitalization of income (Saez and Zucman, 2016).

The SCF+ data from Moritz et al. (2020) sheds new light on further studies of inequality. The racial income gap is observed via descriptive analysis. Moreover, growing age has effect on Gini coefficient, but both racial effect and age effect have not been examined thoroughly in this paper. As for racial effect, Chetty et al. (2020) study sources of racial inequality across generation, Emmons (2020) find out that housing wealth gain does not narrow disparities across races, and Alina et al. (2021) conclude that monetary policies aiming to increase employment of black households have little effect in reducing racial inequalities. Nevertheless, racial gap in the U.S increasingly arouse research attention. As for the age effect, Ozhamaratli et al. (2022) firstly study the joint distribution of age and income based on a panel data via Generalized Method of Moments (GMM) and Least Squares Method (LSM).

In order to depict an overall picture of decomposition of inequality, income inequality in 2016 in terms of age and race is studied in this term paper. Starting from a cross-sectional data, panel data and difference across year could be studied in the near future.

A similar idea has been tried out by Sohn et al. (2014). They apply structured additive distributional regression to compare conditional income distributions in Germany between 1992 and 2010. Distributional regression is recognized as a powerful tool to study different dimensions of distribution since many years. It considers a joint additive model and determines the amount of smoothness and nonlinearity from the data. Moreover, it can be interpreted and estimated under both a frequentist and Bayesian context, and it can be extended to time series model (Rigby and Stasinopoulos, 2005; Kneib, 2013). Therefore, to achieve the above research goal, generalized additive models for location, scale and shape (GAMLSS) is applied for estimation and prediction.

In this term paper, income inequality across races is discussed through decomposing income distribution by both age and race based on distributional regression. In the following sections, the detail and derivation of GAMLSS is presented, and SCF+ data is introduced along with some characteristics that could be observed from a descriptive analysis; Next, estimation result is given and illustrated with a thorough interpretation; In the end, some conclusions are draw and further potential research topics are listed.

2 The Model and Data

2.1 Generalized Additive Models for Location, Scale and Shape (GAMLSS)

Many empirical phenomena are intrinsically linked to distributions rather than only to means. Therefore, to research an empirical topic under a more realistic and richer framework is proposed by means of distributional regression (Kneib et al., 2021). While Generalized Linear Model (GLM), Generalized Additive Model (GAM) and Generalized Linear Mixed Model (GLMM) all assume an exponential family conditional distribution for y, and are not or rarely modelled

variance, skewness and kurtosis explicitly in terms of the explanatory variables, but implicitly through their dependence on μ, GAMLSS relaxes the condition by assuming a very general distribution family and models location, scale and shape separately.

Generally, given observations $(y_i.\mathbf{x}_i), i = 1, ..., n$, where y_i is response and $\mathbf{x}_i$ are covariates featuring random effects, spatial effects, or other regression effects going within and beyond classical linear regression predictors, (conditional) independence of the responses $y_i|\mathbf{x}_i$ with the covariates $\mathbf{x}_i$ is assumed. Their distribution can be described by a parametric density $p(y_i|\vartheta(\mathbf{x}_i))$, where $\vartheta(\mathbf{x}_i) = (\theta_1(\mathbf{x}_1), ..., \theta_K(\mathbf{x}_i))^T$ is a K-dimensional vector of distributional parameters. Usually, the first two population parameters θ_1 and θ_2 are characterized as location and scale parameters μ and σ, and the remaining are shape parameters, but it is not always the case. Let $g_k(\cdot)$ be a known monotonic link function relating $\theta_k(\mathbf{x}_i)$ to explanatory variables and random effects through an additive model given by (1).

$$g_k(\theta_k(\mathbf{x}_i)) = \eta_k(\mathbf{x}_i) = \mathbf{X}_k\beta_k + \sum_{j=1}^{J_k} \mathbf{Z}_{jk}\gamma_{jk} \tag{1}$$

where $\theta_k(\mathbf{x}_i)$ and $\eta_k(\mathbf{x}_i)$ are vectors of length n, $\mathbf{X}_k$ is a known design matrix of order $n \times J_k$, β_k is a vector of corresponding parameters, $\mathbf{Z}_{jk}$ is a fixed known $n \times q_{jk}$ design matrix and γ_{jk} is a q_{jk}-dimensional random variable. It indicates that $g_k(\theta_k(\mathbf{x}_i))$ can be modeled by a J_k-dimensional parametric model and a q_{jk}-dimensional random model. If $\mathbf{Z}_{jk} = \mathbf{I}_n$, where $\mathbf{I}_n$ is an $n \times n$ identity matrix, and $\gamma_{jk} = \mathbf{h}_{jk} = h_{jk}(\mathbf{x}_j k)$ for all combinations of j and k in model (1), this leads to

$$g_k(\theta_k(\mathbf{x}_i)) = \eta_k(\mathbf{x}_i) = \mathbf{X}_k\beta_k + \sum_{j=1}^{J_k} h_{jk}(\mathbf{x}_{jk}) \tag{2}$$

If μ, σ, ν and τ are considered simultaneously, the following model is derived as

$$\begin{cases} g_1(\mu) = \eta_1(\mathbf{x}_i) = \mathbf{X}_1\beta_1 + \sum_{j=1}^{J_k} h_{j1}(\mathbf{x}_{j1}) \\ g_2(\sigma) = \eta_2(\mathbf{x}_i) = \mathbf{X}_2\beta_2 + \sum_{j=1}^{J_k} h_{j2}(\mathbf{x}_{j2}) \\ g_3(\nu) = \eta_3(\mathbf{x}_i) = \mathbf{X}_3\beta_3 + \sum_{j=1}^{J_k} h_{j3}(\mathbf{x}_{j3}) \\ g_4(\tau) = \eta_4(\mathbf{x}_i) = \mathbf{X}_4\beta_4 + \sum_{j=1}^{J_k} h_{j4}(\mathbf{x}_{j4}) \end{cases} \tag{3}$$

After the general setup of GAMLSS is introduced, the specific type of population distribution in this study need to be chose. Hajargasht et al. (2012) investigate income distribution using grouped data in detail and conclude that generalized beta distribution of the second kind (GB2) and its special-case distribution, including Dagum that is tested to be better than Log-normal distribution by Sohn et al. (2014), is recommended for describing income distribution. Clementi et al. (2016) provide another option for both income and wealth distribution, which stems from exactly the GB2. Additionally, Stasinopoulos et al.

(2007) offer a guidance of how `GAMLSS` package in R could be used under various family distribution.

Based on the above resources (Rigby and Stasinopoulos, 2005) and inspired by Sohn et al. (2014), the predictors for the conditional income distribution are estimated in a structured additive framework such that

$$log(\hat{\mu}) = \eta_\mu = s_\mu(age) + Rs_\mu(age) \tag{4}$$

$$log(\hat{\sigma}) = \eta_\sigma = s_\sigma(age) + Rs_\sigma(age) \tag{5}$$

$$log(\hat{\nu}) = \eta_\nu = s_\nu(age) + Rs_\nu(age) \tag{6}$$

$$log(\hat{\tau}) = \eta_\tau = s_\tau(age) + Rs_\tau(age) \tag{7}$$

where s denotes a smooth function such that the effect of age can be modeled in a non-linear way. According to the research by Ozhamaratli et al. (2022), s can be modeled as age^ξ. Cubic splines additive terms can also implemented on age^ξ in R via `cs()`. Although it is still under risk of misspecifying the response model, scrutinizing quantile residuals, see 3.1) is a possible solution.

2.2 Generalized Beta Distribution of the Second Kind (GB2)

Duangkamon et al. (2016) introduce GB2 in detail. The probability density function (pdf) of GB2 is given by

$$f(y|\mu,\sigma,\nu,\tau) = \frac{\mu y^{\mu p-1}}{\sigma^{\mu\nu} B(\nu,\tau)\left(1+\left(\frac{y}{\sigma}\right)^{\mu}\right)^{\nu+\tau}} \tag{8}$$

where $y>0, \mu>0, \sigma>0, \nu>0, \tau>0$ and $B(\nu,\tau)$ is the beta function. The corresponding cumulative distribution function (cdf) is denoted by $F(y|\mu,\sigma,\nu,\tau)$. The k-th moment of the GB2 exists for $-\mu\nu < k < \mu\nu$ and is given by

$$E(Y^k) = \frac{b^k B(\nu+k/\mu, \tau-k/\mu)}{B(\nu,\tau)} = \frac{b^k \Gamma(\nu+k/\mu)\Gamma(\tau-k/\mu)}{\Gamma(\nu)\Gamma\tau)} \tag{9}$$

The k-th moment distribution function for the GB2 is given by

$$F_k(y|\mu,\sigma,\nu,\tau) = \frac{1}{E(Y^k)}\int_0^y t^k f(t)dt = F(t|\mu,\sigma,\nu+k/\mu,\tau-k/\mu) \tag{10}$$

2.3 SCF+ Data

As SCF+ data is clearly described by Kuhn et al. (2020), codes and document provided by Kuhn in his personal website are studied carefully. Figures and tables in appendix of the paper are reproduced successfully, which ensures a proper use of data. The key points of processing SCF+ is

1. For each observation, 1 imputed value should be randomly selected out of 5
2. For the indication of time, 'yearmerge' should be used because it's 3-year window, which is not used at all in this term paper, but this tip is useful for further research
3. Weights 'wgtI95W95' should be used for each calculation since the adjustment makes the data more representative
4. Income and wealth etc. are already adjusted in 2016 dollars
5. Growth of income and wealth is already built relative to 1971

Particularly, weights are standardized when model is estimated for the compatibility to `GAMLSS` in R. The frequency of log income based on different age and race of head is presented in Figure 1.

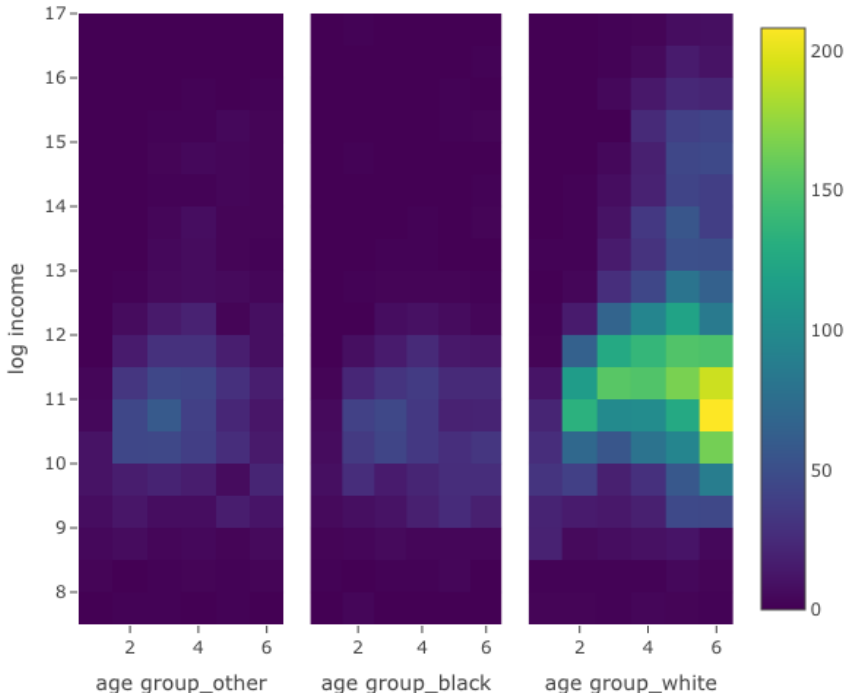


Fig. 1: Log Income Histogram of Different age and Race of Head

We can roughly observe that the frequency of high log income is overall higher if head of household is white. On the contrary, black and other races achieve acme in income in a relative young age. Additionally, variance of income from white head of household seems to be larger as age increases. Generally, conditional income goes up and down. Since the R package used in Figure 1 does not support adjustment by weights, the frequencies across races can not be

compared properly. Moreover, if weights cannot be applied adequately, deviation within a particular range might be tolerated (See 3.1).

The pairwise point plot confirms the above conjecture to some extent. In addition, covariates seem not to have strong correlation to each other in Figure 2

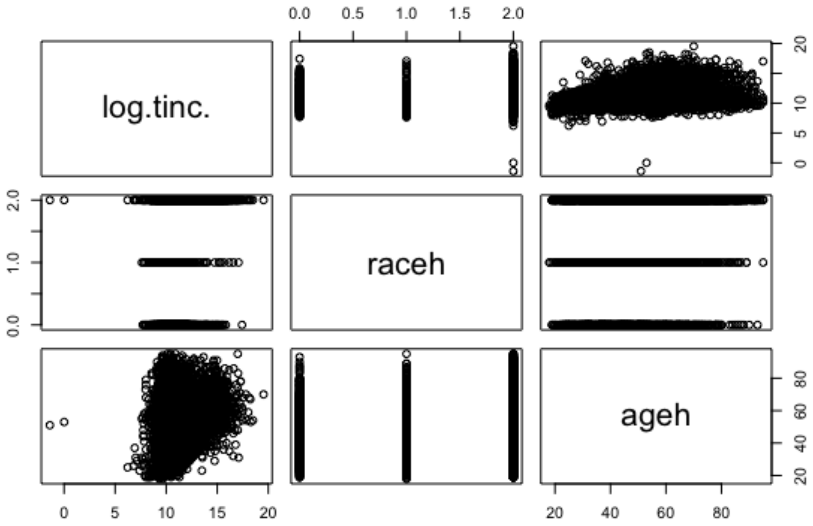


Fig. 2: Wealth of Different Age, Race and Income Groups of Head

3 Estimation and Results

3.1 Estimation

Nichols (2008) analyses the reason of zero and negative income and wealth based PSID data. In his discovery, many of the measured income zeros are real, and Moritz et al. (2020) also point out in the appendix of their paper that negative values do not change the overall trend of Gini coefficients for both income and wealth. Therefore, as for income, only positive value as majority is used for estimation, while wealth, if further research considers, should be analysed separately as Clementi et al. (2016) did in their study.

The default estimation algorithm in `GAMLSS()` in R is `RS()`, which is a generalization of the algorithm that is suitable not only when parameters ϑ in $p(y_i|\vartheta(\mathbf{x}_i))$ are information orthogonal, but also for most of densities in `GAMLSS()`. `CG()` is another option for optimization, particularly when orthogonality is violated.

`find.hyper()` and `optim` can be used to select the values of hyper parameters and/or non-linear parameters in a GAMLSS model by minimizing the generalized Akaike information criterion (GAIC) with a user defined penalty. The penalty is set to 2 as default. Different possible powers of age and cubic spines with various optimization methods are searched, e.g. `RS()`, `CG()` and `mixed()`. Cubic splines additive terms are fitted using a backfitting algorithm. (Stasinopoulos and Rigby, 2005 and 2007)

Through 20 GAMLSS-RS iterations, the estimation results of GAMLSS for GB2 function is listed in Table 1.

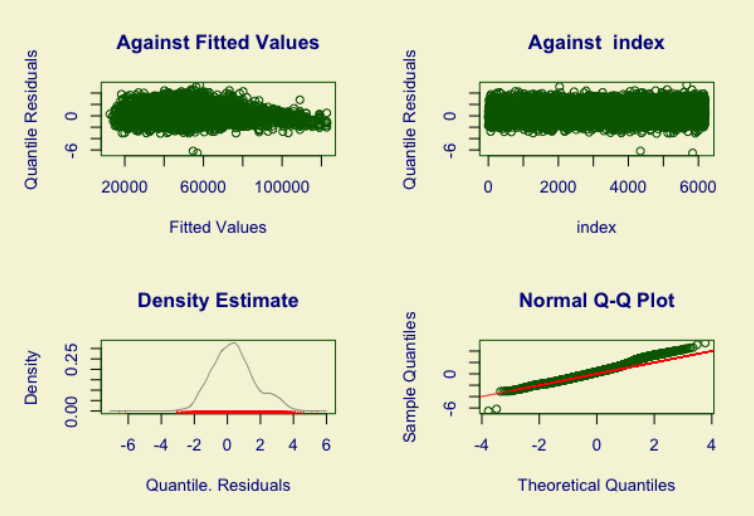
Table 1: Estimation of Parameters of Income Distribution

	μ	σ	ν	τ
$age^{0.55}$	-0.2339^{***}	0.0686^{***}	0.0804^{***}	-0.2339^{***}
$age^{0.55} : raceh1$	0.0234^{*}	-0.0037	0.0065	0.0234^{*}
$age^{0.55} : raceh2$	0.0008	0.0011	-0.0207^{**}	0.0008
Intercept	2.0311^{***}	-0.0130	-0.4393^{***}	2.0311^{***}
Global Deviance	47903.15			
AIC	47935.15			
SBC	48042.85			

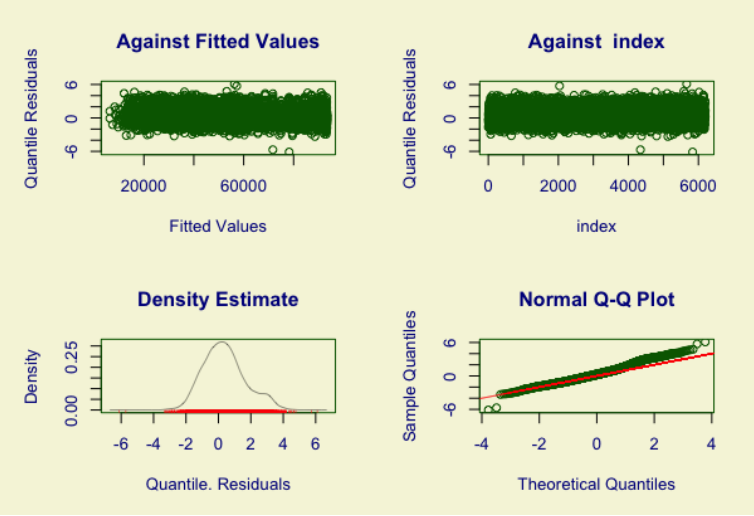
Note: $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$

To evaluate the model, a diagnosis analysis of quantile residual is given in Figure 3. For each subplot, top left and right figures plot the residuals against the fitted values of μ and against age respectly, whereas bottom left and right figures offer a kernel density estimate and normal QQ-plot for them respectively. Residuals against the fitted values of μ appears not random if age is modeled without cubic spines, while the second subplot improve the result. However, both models do not have an ideal QQ-plot that ought to be similar to Box-Cox t-distribution. One possible reason is that weights are here not been applied to residuals because weights are not rigid frequencies. In addition, because the coefficient estimates of the second optional model (b) are rarely significant, which is not reliable, the first model (a) above is kept.

Apart from quantile residual analysis, worm plot and Q statistics of residuals are both tested for evaluation. As they function as QQ-plot to check normality of residual in a similar way, concrete plots are abbreviated. In addition, `fittedPlot` can only present the fitted value for parameters, embodying no practical meaning. In other words, though the estimated GAMLSS model with GB2 function is still valuable for further analysis, lots of visual descriptions of fitted parameters are abbreviated due to its lack of interpretability.



(a) Modeling Age Without Cubic Spines



(b) Modeling Age With Suitable Cubic Spines

Fig. 3: Quantile Residual Analysis

3.2 Results

Since μ, σ, ν and τ are not corresponding explicable statistics directly, transformation needs to be done before presenting income distribution. Ideally, for each combination of covariates, there are 4 distribution of μ, σ, ν and τ with their own point estimates and standard deviation; And for each combination of μ, σ, ν and τ, a simulation of income distribution can be plotted, one of all potential income distributions which this specific combination of covariates might have. To generate a readable figure, μ, σ, ν and τ are predicted based on 4 typical ages (30, 40, 50, 60) at first and all 3 races. Then, histograms are built based on these predictions of GB2 parameters with number of simulation being 1000. Relevant figures are combined in Figure 4.

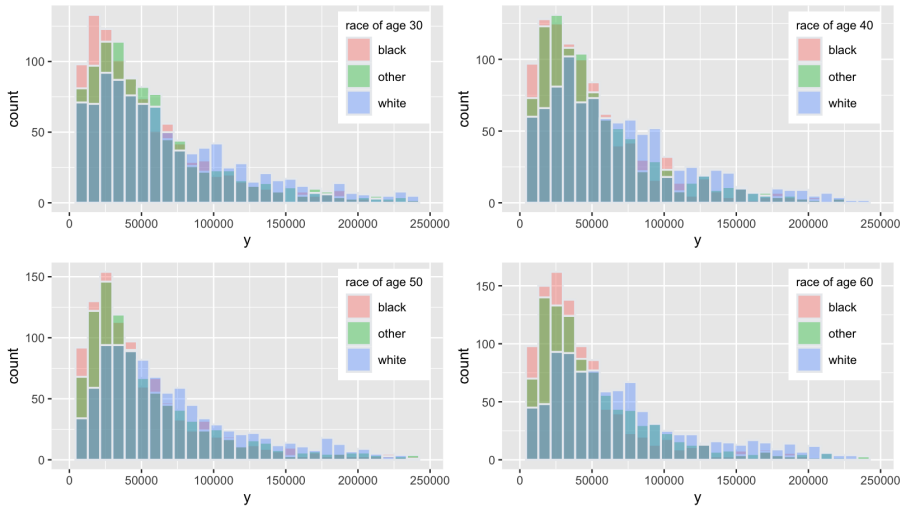


Fig. 4: Income Histogram in Terms of Age and Race

Through comparison, income distribution of black head of household is most right tailed, followed by that of other and white subsequently. Income distributions of both black and other races remain almost heavy right tailed over ages, while the skewness of that of white is eased as age increases. A 3D plot is derived to present the simulation of income distribution in a more comprehensive way.

Figure 5 gives a elegant overview of the conditional income distribution in terms of age and race. p is the estimated probability given particular age, race and income (y) based on estimated model parameters. The original 3D plot can be spanned manually. Four snapshot is presented side by side. From Figure 5(a)(b) we can see that black head of household is most likely have higher probability in lower income, while white head of household more likely to gain higher income. Figure 5(c)(d) clearly show that as age goes up, the probability of earning no income ($y = 0$) drops at first but then it grows up as people grow

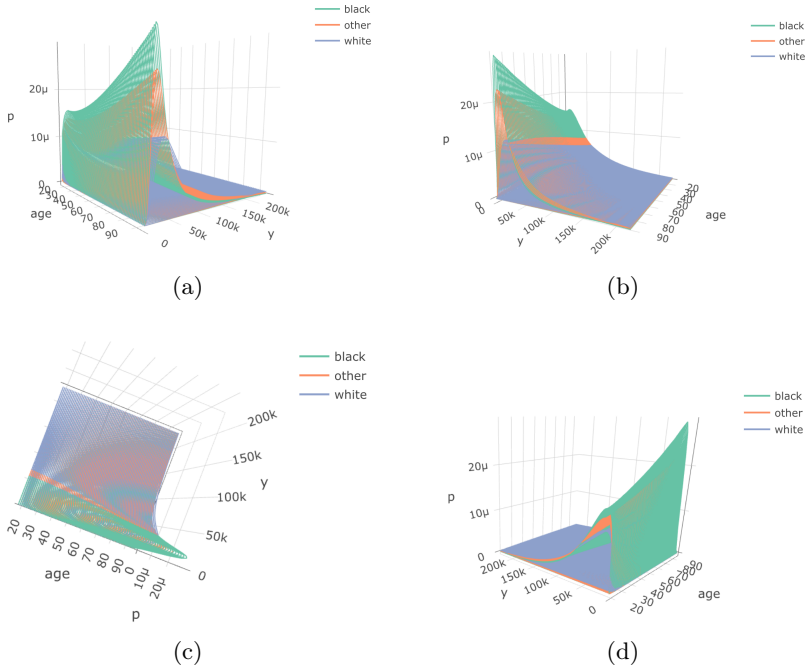


Fig. 5: 3D Income Density Plot Across Age and Race

older. The variance of income is larger when age is lower if head of household is black, but the trend is reversed if head is white. A plausible explanation is that the elder of black seldom receive income from both labor and capital market, but elder of white might have more capital gain by contrast.

The general trend in Figure 5 is reasonable, so the estimated GAMLSS model for income distribution is recommended for further research. However, since it is only one of all inferred income distribution, the visualization of all is limited.

4 Conclusion

As a more general regression framework, distributional regression can depict a panorama of conditional income distribution via estimation and simulation, which is a basis for study of income inequality. Based on SCF+ Data, a conditional income distribution can be modeled via GAMLSS with its population distribution set to be GB2. A simulation of income distribution in terms of age and race is given by a grouped histogram and a 3D density plot. From these simulation plots, income is gathered to be low if head of household is black compared to white, and the discrepancy is larger as age increases. Although the model simulates a satisfactory conditional income distribution, the absence of weight while analysing quantile residual is still a problem that needs to be

solved. Extension to wealth distribution and panel data is what could be done in the future.

5 Appendix

See attachment DingyiLai.R

References

1. Alina K. Bartscher, Moritz K., Moritz S., Paul W., 2021. "Monetary Policy and Racial Inequality". *ECONtribute Discussion Papers Series 061*, University of Bonn and University of Cologne, Germany.
2. Chetty R., Hendren N., Jones R.M., Porter R.S. (2020). "Race and Economic Opportunity in the United States: an Intergenerational Perspective". *The Quarterly Journal of Economics*, Volume 135, Issue 2, May 2020, Pages 711–783, https://doi.org/10.1093/qje/qjz042
3. Clementi, F., Gallegati, M., Kaniadakis, G., Landini, S. (2016). "k-generalized models of income and wealth distributions: A survey." *The European Physical Journal*. ST, Special Topics, 225(10), 1959-1984.
4. Duangkamon C., William E. G., Gholamreza H., Wasana K., Rao D. S. P. (2018). "Using the GB2 Income Distribution". *Econometrics*, MDPI, vol. 6(2), pages 1-24, April.
5. Emmons, R.W. (2020). "Housing Wealth Climbs for Hispanics and Blacks, Yet Racial Wealth Gaps Persist". *Federal Reserve Bank of St. Louis*, April 1, 2020. Working Paper.
6. Hajargasht, G., Griffiths, W., Brice, J., Rao, D., Chotikapanich, D. (2012). "Inference for Income Distributions Using Grouped Data". *Journal of Business Economic Statistics*, 30(4), 563-575.
7. Kneib, T. (2013). "Beyond mean regression". *Statistical Modelling*, 13(4), 275–303. https://doi.org/10.1177/1471082X13494159
8. Kneib, T., Silbersdorff, A., Säfken, B. (2021). "Rage Against the Mean – A Review of Distributional Regression Approaches". *Econometrics and Statistics*.
9. Moritz K., Moritz S., Ulrike I. Steins, (2020). "Income and Wealth Inequality in America, 1949–2016". *Journal of Political Economy*, University of Chicago Press, vol. 128(9), pages 3469-3519.
10. Nichols, A.L. (2008). "Measuring Trends in Income Variability".
11. Ozhamaratli, F., Kitov, O., Barucca, P. (2022). "A generative model for age and income distribution". *EPJ Data Sci.* 11, 4 . https://doi.org/10.1140/epjds/s13688-022-00317-x
12. Piketty, T., Goldhammer, A., Ganser, L. J. (2014). *Capital in the twenty-first century*. Unabridged. Grand Haven, MI: Brilliance Audio, Inc.
13. Pyatt, G. (1976). "On the Interpretation and Disaggregation of Gini Coefficients". *Economic Journal*, 86, issue 342, p. 243-55.
14. 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. http://www.jstor.org/stable/3592732
15. Saez, E. and Zucman, G., (2014). "Wealth Inequality in the United States since 1913: Evidence from Capitalized Income Tax Data". No 20625, *NBER Working Papers*, National Bureau of Economic Research, Inc.

16. Shorrocks, A.F, (1980). "The Class of Additively Decomposable Inequality Measures". *Econometrica*, 48, issue 3, p. 613-25.
17. Shorrocks, A.F., (1982). "Inequality Decomposition by Factor Components". *Econometrica*, 50, issue 1, p. 193-211.
18. Shorrocks, A.F., (1983). "The Impact of Income Components on the Distribution of Family Incomes". *The Quarterly Journal of Economics*, 98, issue 2, p. 311-326.
19. Shorrocks, A.F. (1984). "Inequality Decomposition by Population Subgroups", *Econometrica*. 52, issue 6, p. 1369-85.
20. Sohn, A., Klein, N., Kneib, T. (2014). "A new semiparametric approach to analysing conditional income distributions". *University of Göttingen Working Papers in Economics 192*, University of Goettingen, Department of Economics.
21. Stasinopoulos, D. M., Rigby, R. A. (2007). Generalized Additive Models for Location Scale and Shape (GAMLSS) in R. *Journal of Statistical Software*, 23(7), 1–46. https://doi.org/10.18637/jss.v023.i07
22. Tausch, A., Heshmati, A. (2007). *Global Trends in Income Inequality*, Nova Science Publishers, 27-48.
23. Theil, H. (1967) *Economics and Information Theory*. North-Holland Publishing Company, Amsterdam.
24. Moritz Schularick Personal Website, https://www.moritzschularick.com/. Last accessed 4 Mar 2022