Measuring wage inequality under right censoring*

João Nicolau[a], Pedro Raposo[b] and Paulo M. M. Rodrigues[c]

[a] ISEG-Universidade de Lisboa and CEMAPRE

[b] Católica Lisbon School of Business and Economics

[c] Banco de Portugal and Nova School of Business and Economics

June 25, 2022

Abstract

A conditional tail index estimator is introduced which explicitly allows for right-tail censoring (top-coding), which is a feature of the widely used current population survey (CPS), as well as of other surveys. We show that the factor values used to adjust top-coded wages have changed over time and depend on individuals' characteristics, occupations and industries, and propose suitable values. Specifically, we observe that the factor values used to adjusted the top-coded wages for women in the Finance industry increased from 1.5 in 1992 to around 2 in 2017 (and similarly in other industries although of smaller magnitude). We also observe that inequality in the US weekly wage distribution increased between 1992 and 2017 more than what had previously been described in the literature. Specifically, contrasting the results of our approach with those of a conservative fixed adjustment factor of 1.5 (used in the literature), our procedure indicates that inequality in 2017 is 5% larger than that suggested by the fixed adjustment factor.

Keywords: Wage inequality, tail index, top-coding, current population survey, weekly wage distribution, Pareto, occupations
JEL classification: C18, C24, E24, J11, J31

*The authors thank two anonymous referees, the Co-Editor (Shankha Chakraborty), Daron Acemoglu, David Autor, Manuel Arellano, participants of the Workshop on *New Frontiers in Statistics of Extremes* and of the IX_t Workshop in *Time Series Econometrics* for their helpful and constructive feedback on an earlier version of this paper, and Filipe Caires for excellent research assistance. This work was funded by Fundação para a Ciência e Tecnologia through project number UID/Multi/00491/2019, PTDC/EGE-ECO/28924/2017, UID/ECO/00124/2013, UID/ECO/00124/2019, UID/GES/00407/2013 and Social Sciences DataLab (LISBOA-01-0145-FEDER-022209), POR Lisboa (LISBOA-01-0145-FEDER-007722, LISBOA-01-0145-FEDER-022209), and POR Norte (LISBOA-01-0145-FEDER-022209).

1 Introduction

The sharp rise in overall wage inequality in the second half of the 20th century has become a stylized fact (Autor, 2019 and Goos et al., 2014). Wage inequality growth in the 1980s was followed by a slowdown in the 1990s as a result of divergent trends in the bottom and top of the wage distribution. Both the 90/50 and 50/10 indexes grew rapidly in the early 1980s, and although lower tail inequality virtually came to an halt after 1987 upper-tail inequality kept rising. According to Autor et al. (2008) between 1963 and 2005, the 90th percentile wage rose, by more than 55% relatively to the 10th percentile for both men and women.

The monotonic increase of inequality until the late 1980s followed by the divergent evolution in the top and lower half of the distribution is robust to different measures and samples.[1] Steady growth in the upper-tail inequality can also be seen from the rising share of wages paid to the top 10% and 1% earners (Piketty and Saez, 2003). For instance, Burkhauser et al. (2012) using March CPS and IRS tax returns data observe that income inequality changes since 1993 are largely driven by changes in the incomes of the top 1%. However, literature based on public-use CPS data has produced a less than perfect picture of the right tail of the wage distribution because of the top-coding (Armour et al., 2016). CPS wage data has historically been censored at the top (top-coded) and ignoring this fact or not adequately handling it may result in inconsistent tail index estimates, lead to understatements of inequality and affect the estimates of its dynamics (Feng et al., 2006). In addition, top-coding has changed over time. For instance, the weekly wage measure was top-coded at $1923 between 1989 and 1997, and at $2884 between 1998 and 2017. But even during periods of constant nominal top-coding the data may hide changes in inequality (Levy and Murnane, 1992).

Hlasny and Verme (2021) using CPS data show that income inequality in the US between 1979 and 2014 has been consistently underestimated by several percentage points and Hlasny and Verme (2018) using EU SILC data also find downward biased inequality estimates for several European countries. In both papers the authors use two alternative correction methods of the right censoring, a stochastic approach based on reweighing and a semi-parametric approach based on replacing observations.

While some authors have tried to address the top-coding issues by restricting the sample under analysis, the method presented in this paper makes use of the complete set of information available from the public use CPS data, for every year, in a time-consistent fashion, arguably providing better estimates of the level of weekly wage inequality than other available measures. Our procedure provides additional flexibility by allowing researchers to evaluate which determinants impact the probability of being in the right tail of the weekly wage distribution and the corresponding Gini coefficient. The approach allows for a detailed analysis on how inequality has spread across industries, occupation, gender and other population characteristics.

Our paper relies on the tail index parameter to assess wage inequality. The tail index characterizes the rate of decay of the tail of a power law distribution, i.e., the likelihood of observing extreme wages. The smaller the values of the tail index, the larger the likelihood of observing extreme wages and large fluctuations in the right tail of the distribution. To relate heavy-tailedness and upper-tail inequality, it is worth noting that for a Pareto distribution with a tail index of α, the Gini coefficient is given by $G = 1/\left(2\alpha - 1\right), \alpha > 1$ (see Ibragimov and Ibragimov, 2018). Hence, since the Gini coefficient[2] is directly related to the tail index, the additional information that we obtain from the estimation of the conditional tail index does therefore extend to the Gini coefficient. Thus, the tail index estimates of a Pareto-type distribution can be used as an additional measure of inequality. Specifically, if a variable follows approximately a Pareto distribution for large wages, the value of the right tail index and corresponding Gini coefficient may be considered as measures of upper tail inequality, that is, inequality between moderate or high wages and extremely high wages. Our paper shows that this tail index and Gini coefficient vary across age, education, gender, race, marital status, occupation and industry, which means that the likelihood of observing extreme weekly wages, and hence different wage inequality patterns, varies across these groups.

Several estimation approaches have recently been proposed which consider either non-random or random covariates; see e.g. Ma et al. (2019) (and references therein). Our contribution falls into the latter class and provides a tail index estimator which takes the top-coding explicitly into consideration, providing in this way more efficient and consistent estimates than methods currently available in the literature. The superior performance of the new approach is illustrated. It is shown, using the public-use CPS database from 1992 to 2017, that the factor values used for the adjustment of the top-coded weekly wages changed over time and across individuals' characteristics, occupations and industries, which highlights the heterogenous nature of inequality; moreover results also show that the tail index has been decreasing since 1992, which corresponds to an increasing Gini coefficient, and which suggests an increase in the right-tail inequality.

The contribution of this paper is threefold: First, a conditional tail index estimator is introduced that explicitly handles the top-coding problem, and its finite sample performance is evaluated and compared to competing methods; second, it is shown that the factor values necessary to adjust the top-coded wages has change over time and across individuals' characteristics, occupations and industries, and suitable values are proposed; and third, an in-depth analysis of the dynamics of the US weekly wage distribution's right tail using the public-use CPS database from 1992 to 2017 is provided.

The remainder of the paper is organized as follows. Section 2 introduces the methodology of analysis, the new tail index estimator and a detailed description of the computation of the partial effects; Section 3 presents the results of an in-depth Monte Carlo analysis on the finite sample properties of the new approach and a comparison to existing procedures as well as an analysis of their performance for the imputation of weekly wages using non-censored data; Section 4 describes and discusses the results of the right tail characteristics of the weekly wage distribution and weekly wage inequality in the US using the CPS database from 1992 to 2017; and finally, Section 5 presents the main conclusions of the paper. A technical appendix collects the proofs of the results put forward throughout the paper.

2 Methodology

To reduce the top-coding bias researchers typically impute top-coded values to create consistent series. Until recently one of four approaches has been adopted in the literature: (1) the top-coding problem is ignored i.e., top-coded observations are dropped (see e.g. Jensen and Shore, 2015); (2) an *ad hoc* adjustment of the top-coded wages is made (e.g. Lemieux, 2006 and Autor et al., 2008 multiplied top-coded hourly wages by 1.4, and top-coded weekly wages by 1.5, respectively); (3) a Pareto distribution is used to estimate wages at the top of the distribution (e.g. Bernstein and Mishel, 1997, Piketty and Saez, 2003); and (4) cell means or rank-proximity swapped data based on the still-censored internal CPS data is used (e.g. Larrimore et al., 2008 and Burkhauser et al., 2008); for a discussion and shortcomings of these approaches see, e.g., Burkhauser et al. (2010) and Armour et al. (2016).

In a recent contribution Armour et al. (2016) proposed an alternative approach which consists of the estimation of the tail index of a censored Pareto distribution. To briefly illustrate the procedure consider first the survival function, $\overline{F}$, of an uncensored Pareto distribution,[3] $\overline{F}(y) := P\big(Y > y\big) = \big(y_0/y\big)^{\alpha}$ with $y \geq y_0 > 0,\ \alpha > 0$, and corresponding density function $f_Y(y) = \alpha y_0^{\alpha}/y^{\alpha+1}$. A large number of tail index estimators is available in the literature. One widely used approach is the conditional maximum likelihood estimator (MLE) proposed by Hill (1975),

$$\widehat{\alpha}_{Hill} := m\left[\sum_{j=1}^{m} \log y_{(j)} - \log y_{(0)}\right]^{-1} \tag{2.1}$$

where m is the number of largest order statistics, y_0 is the tail cut off point and $y_{(j)}, j = 1, ..., m$, is the j^{th} largest value used in the estimation of α.

However, recognizing the limitations of $\widehat{\alpha}_{Hill}$ when the data is top-coded, Armour et al. (2016) proposed an alternative approach, which consists of an adaptation of the Hill estimator taking the censoring into consideration. This approach provides an unbiased estimate of the censored Pareto parameter, α, while using all available information.

In the case of a censored sample the outcome variable is,

$$\omega_i = \begin{cases} y_i & if \quad y_0 \le y_i < y_c \\ y_c & if \qquad y_i \ge y_c \end{cases}, \tag{2.2}$$

where y_0 is the tail cut off point and y_c is the censoring threshold. The density function of the censored Pareto distribution is,

$$g_Y(\omega_i) = \left(\frac{\alpha y_0^{\alpha}}{\omega_i^{\alpha+1}}\right)^{I_{(y_0 \le \omega_i < y_c)}} \left[\left(\frac{y_0}{y_c}\right)^{\alpha}\right]^{I_{(\omega_i \ge y_c)}} \tag{2.3}$$

and the respective log-likelihood function,

$$\begin{aligned} \sum_{i=1}^{m} \log\left[g_Y(\omega_i)\right] &= \sum_{i=1}^{m} I_{(y_0 \le \omega_i < y_c)}\Big(\log(\alpha) + \alpha \log(y_0) - (\alpha+1)\log(\omega_i)\Big) \\ &\quad + \sum_{i=1}^{m} I_{(\omega_i \ge y_c)}\Big(\alpha \log(y_0) - \alpha \log(y_c)\Big). \end{aligned} \tag{2.4}$$

Consequently, the conditional MLE estimator proposed by Armour et al. (2016) computed from (2.4) is,

$$\widehat{\alpha}^{c}_{Hill} = n_0\Bigg(\sum_{i=1}^{m} I_{(y_0 \le y_i < y_c)} \log(y_i) + n_c \log(y_c) - (n_0 + n_c)\log(y_0)\Bigg)^{-1} \tag{2.5}$$

where n_0 is the number of individuals with wages between y_0 and y_c, n_c is the number of individuals with wages at or above y_c, and $n_0 + n_c = m$.

2.1 The conditional tail index estimator and properties

In this paper, we introduce a conditional tail index estimator which also takes the right censoring of the data into account and uses covariates in the estimation process. The procedure has the advantage of allowing for an in depth analysis of the determinants that impact the tail index according to individuals' characteristics, occupations and industries, which will be useful for a better understanding of the heterogenous nature of inequality.

To introduce the approach consider observations $(\mathbf{X}_i, Y_i)$, $i = 1, ..., n$, where $Y_i \in \mathbb{R}^1$ is the response of interest, and $\mathbf{X}_i := (x_{1i}, ..., x_{pi})' \in \mathbb{R}^p$ is an associated p-dimensional vector of predictors. In addition, let $F(y|\mathbf{x}; \boldsymbol{\theta}) := P[Y_i \le y|\mathbf{X}_i = \mathbf{x}]$ be the cumulative distribution function of Y_i conditional on $\mathbf{X}_i$, and assume that the corresponding uncensored survival function is,

$$\overline{F}(y|\mathbf{x};\boldsymbol{\theta}) := 1 - F(y|\mathbf{x};\boldsymbol{\theta}) = y^{-\alpha(\mathbf{x})}\mathcal{L}(y;\mathbf{x}), \tag{2.6}$$

where $\alpha(\mathbf{x}) := \exp(\mathbf{x}'\boldsymbol{\theta})$, $\boldsymbol{\theta} \in \mathbb{R}^p$ is the unknown vector of coefficients and $\mathcal{L}(y;\mathbf{x})$ is some predictor-dependent slowly varying function, such that $\mathcal{L}(yk;\mathbf{x})/\mathcal{L}(y;\mathbf{x}) \to 1$ for any $k > 0$ as $y \to \infty$. Following Hall (1982) we characterize the slowly varying function as,

$$\mathcal{L}(y;\mathbf{x}) := c_0(\mathbf{x}) + c_1(\mathbf{x})y^{-\beta(\mathbf{x})} + o(y^{-\beta(\mathbf{x})}) \tag{2.7}$$

where $c_0(\mathbf{x})$ and $c_1(\mathbf{x})$ are functions in $\mathbf{x}$ with $c_0(\mathbf{x}) > 0$, $\beta(\mathbf{x}) > \alpha(\mathbf{x})$ a positive function and $o(y^{-\beta(\mathbf{x})})$ is the higher-order remainder term. As a result, as $y \to \infty$, $\mathcal{L}(y;\mathbf{x}) \to c_0(\mathbf{x})$ and $\dot{\mathcal{L}}(y;\mathbf{x}) := \partial\mathcal{L}(y;\mathbf{x})/\partial y \to 0$.

From (2.6), it follows that the probability density function of Y_i conditional on $\mathbf{X}_i$ is,

$$f(y|\mathbf{x};\theta) = \alpha(\mathbf{x})y^{-\alpha(\mathbf{x})-1}\mathcal{L}(y;\mathbf{x}) - y^{-\alpha(\mathbf{x})}\dot{\mathcal{L}}(y;\mathbf{x}). \tag{2.8}$$

Considering (2.7) and assuming that y is sufficiently large, it follows that the density in (2.8) can be approximated as, $f(y|\mathbf{x};\theta) \approx c_0(\mathbf{x})\alpha(\mathbf{x})y^{-\alpha(\mathbf{x})-1}$; see also Wang and Tsai (2009). Thus, the conditional probability function of Y_i given $\mathbf{X}_i$ and $Y_i > y_0$ can be approximated as,

$$f(y|\mathbf{x};\theta) \approx \alpha(\mathbf{x})(y/y_0)^{-\alpha(\mathbf{x})-1}, \tag{2.9}$$

where y_0 is the threshold that controls the sample fraction used for estimation. (2.9) is the approximate conditional Pareto density function of an unrestricted random variable and its use when some form of censoring (such as right censoring[4] in the CPS database) is imposed on the data will originate inconsistent tail index parameter estimates.

In the censored case, rather than observing the outcome y_i we effectively observe w_i as defined in (2.2). The adequately adjusted conditional Pareto density function is,

$$g(\omega_i|\mathbf{x}_i, y_c, \boldsymbol{\theta}) := f(\omega_i|\mathbf{x}_i, y_c, \boldsymbol{\theta})^{I(y_0 \leq w_i < y_c)}\left[1 - F(y_c|\mathbf{x}_i, y_c, \boldsymbol{\theta})\right]^{I(w_i \geq y_c)} \tag{2.10}$$

where $I(.)$ is the indicator function and $f\left(.|\,\mathbf{x}\right)$ and $F\left(.|\,\mathbf{x}\right)$ are the conditional Pareto density function and the conditional cumulative Pareto distribution function, respectively. The negative log-transformed likelihood function for the top-coded data is,

$$\mathcal{K}_n^c(\boldsymbol{\theta}; y_c) := \sum_{i=1}^{n} \log g\left(w_i|\,\mathbf{x}_i; y_c; \boldsymbol{\theta}\right) \tag{2.11}$$

where $g\left(w_i|\,\mathbf{x}_i; y_c; \boldsymbol{\theta}\right)$ is as in (2.10), w_i is given in (2.2) and y_c is the censoring threshold.

Since

$$\begin{aligned} \log\left[g(\omega_i|\mathbf{x}_i, y_c, \boldsymbol{\theta})\right] &= I(y_0 \leq w_i < y_c)\log f(\omega_i|\mathbf{x}_i, y_c, \theta) + I(w_i \geq y_c)\log\left[1 - F(y_c|\mathbf{x}_i, y_c, \theta)\right] \\ &= I_{(y_0 \leq w_i < y_c)}\left(\log\alpha\left(\mathbf{x}_i\right) + \alpha\left(\mathbf{x}_i\right)\log y_0 - \left[\alpha\left(\mathbf{x}_i\right) + 1\right]\log w_i\right) \\ &\quad + I_{(w_i = y_c)}\left(\alpha\left(\mathbf{x}_i\right)\left[\log\left(y_0\right) - \log(y_c)\right]\right), \end{aligned} \tag{2.12}$$

using $\alpha\left(\mathbf{x}_i\right) = \exp\left(\mathbf{x}_i'\boldsymbol{\theta}\right)$ the approximate negative log-likelihood function in (2.11) (omitting for simplicity of notation the terms not related to $\boldsymbol{\theta}$) becomes,

$$\mathcal{K}_n^c(\boldsymbol{\theta}; y_c) = \sum_{i=1}^{n} I_{(y_0 \leq w_i < y_c)}\left(\exp\left(\mathbf{x}_i^{'}\boldsymbol{\theta}\right)\log\left(\frac{w_i}{y_0}\right) - \mathbf{x}_i^{'}\boldsymbol{\theta}\right) - \sum_{i=1}^{n} I_{\{w_i = y_c\}}\exp\left(\mathbf{x}_i^{'}\boldsymbol{\theta}\right)\log\left(\frac{y_0}{y_c}\right).$$

Hence, we see from this approximate log-likelihood function that censoring the data imposes a penalty term, which the unrestricted estimator does not take into consideration.

To derive the limit distribution of the parameter estimators and corresponding test statistics we consider, as in Wang and Tsai (2009), the following assumptions:

Assumption A:

A.1: $n_0^{-1}\sum\limits_{i=1}^{n}\mathbf{Z}_{ni}\mathbf{Z}_{ni}'I\left(w_i \geq y_0\right) = \boldsymbol{\Sigma}_{y_0}^{-1/2}\widehat{\boldsymbol{\Sigma}}_{y_0}\boldsymbol{\Sigma}_{y_0}^{-1/2} \xrightarrow{p} \mathbf{I}_p$, where $\mathbf{Z}_{ni} := \boldsymbol{\Sigma}_{y_0}^{-1/2}\mathbf{x}_i$, $\mathbf{I}_p$ is a $p \times p$ identity matrix and $\widehat{\boldsymbol{\Sigma}}_{y_0} := n_0^{-1}\sum(\mathbf{x}_i\mathbf{x}_i')I(w_i \geq y_0)$.

A.2: (*Slowly varying function*) We assume that the remainder term $o(y^{-\beta(\mathbf{x})})$ satisfies $\sup_{\mathbf{x}} y^{\beta(\mathbf{x})}o(y^{-\beta(\mathbf{x})}) \to 0$ as $y \to \infty$.

A.3: (*Convergence rate*) Assume that $n_0 \to \infty$ and $\frac{n}{n_0}E\big[\mathbf{Z}_{ni}c_1(\mathbf{x}_i)\frac{\beta(\mathbf{x}_i)}{\alpha(\mathbf{x}_i)+\beta(\mathbf{x}_i)}y_0^{-\alpha(\mathbf{x}_i)-\beta(\mathbf{x}_i)}\big]$ $\to \mathbf{h}$, for some non-zero constant vector $\mathbf{h} \in R^p$.

As indicated by Wang and Tsai (2009) Assumption A.1 enforces the weak law of large numbers for the standardized covariate $\mathbf{Z}_{ni}$; Assumption A.2 regularizes the extreme behavior of the slowly varying function $\mathcal{L}(y;\mathbf{x})$; and Assumption A.3 implicitly specifies the optimal convergence rate of y_0. A rate faster than that in A.3 mitigates the asymptotic bias of the parameter estimates but increases its variability. In contrast, a rate slower than that in Assumption A.3 reduces variability but increases bias; see Wang and Tsai (2009) for further details.

The following theorem characterizes the limit distribution of the MLE estimators of $\boldsymbol{\theta}$.

Theorem 2.1 *Under Assumptions A.1 - A.3 as $n \to \infty$ it follows that*

$$n^{-1/2}\Sigma_{y_0}^{-1/2}\Lambda^{-1/2}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} N(\mathbf{0}, \mathbf{I}_p),$$

where $\Lambda := E(e_i^2|\mathbf{x}_i)$ *and* $e_i = \begin{cases} \exp\left(\mathbf{x}_i'\boldsymbol{\theta}\right)\log\left(\frac{w_i}{y_0}\right) - 1, & \textit{for} \quad I_{\left(y_0 \leq w_i < y_c\right)} \\ -\exp\left(\mathbf{x}_i'\boldsymbol{\theta}\right)\log\left(\frac{y_0}{y_c}\right), & \textit{for} \quad I_{(w_i = y_c)} \end{cases}.$

Corollary 1 *Under the same conditions of Theorem 2.1 it follows as $n \to \infty$ that,* $T_j = n^{-1/2}\left(\Sigma_{y_0,jj}^{-1/2}\right)\Lambda^{-1/2}\widehat{\theta}_j \xrightarrow{d} N(0,1)$, *where* $\Sigma_{y_0,jj}^{-1/2}$ *corresponds to the* $(j,j)^{th}$ *element of the* $\boldsymbol{\Sigma}_{y_0}^{-1/2}$ *matrix.*

2.2 Computation of partial effects

A further important and not immediately obvious aspect of the methodology just described relates to the computation of the partial effects of the covariates used in the conditional tail index regression. In specific, for ease of presentation consider $\overline{F}(y|x;\boldsymbol{\theta}) := P(Y > y|x;\boldsymbol{\theta}) = (y_0/y)^{\alpha(x)}$, where for the sake of simplicity, but with no loss of generality, x is a scalar and continuous. Thus, to measure the impact of x on $\alpha(x)$ and subsequently on $\overline{F}(y|x;\boldsymbol{\theta})$, consider

$$\delta := \left(\frac{\bar{F}\left(y|\,\Delta x + x;\boldsymbol{\theta}\right) - \bar{F}\left(y|\,x;\boldsymbol{\theta}\right)}{\bar{F}\left(y|\,x;\boldsymbol{\theta}\right)}\right) \times 100 \qquad (2.13)$$

where y is an extreme value, say the $(1-u)$ quantile, with $u \in (0,1)$, such that, $y = (1-u)^{\frac{1}{\alpha(x)}} y_0$. The expression in (2.13) measures the percentage of variation in the probability of an extreme value due to a variation of x, Δx. For example, considering $u = 0.15$ and $\Delta x = 1$, if $\delta = 20\%$ then $\mathrm{P}(Y > y_0)$, where y_0 is the 0.85 quantile, increases by 20% as a result of $\Delta x = 1$. Therefore, the variation of x increases the likelihood of observing extreme values by 20%.[5]

For computational purposes, assuming that $\alpha(x) := \exp(\phi(x))$, and $\phi(x)$ is some function of x, we show in the appendix that $\delta(u) = \left[(1-u)^{\phi'(x)\Delta x} - 1\right] \times 100$. For instance, in the multivariate case, $\phi(\mathbf{x}) := \mathbf{x}'\beta$, where $\mathbf{x}$ is a $p \times 1$ vector of covariates and $\alpha(\mathbf{x}) = \exp(\mathbf{x}'\beta)$, the impact of x_j on $\overline{F}(y|\mathbf{x};\boldsymbol{\theta})$ is,

$$\delta_j(u) = \left[(1-u)^{\beta_j \Delta x} - 1\right] \times 100. \tag{2.14}$$

A negative coefficient, $\beta_j < 0$, which implies $\delta_j > 0$, increases the likelihood of having more extreme values. This is also obvious from the impact on $\alpha(\mathbf{x})$ since $\alpha(\mathbf{x})$ decreases and the right tail becomes heavier. Similarly, $\beta_j > 0$ implies $\delta_j < 0$, and the likelihood of having more extreme values decreases.[6]

A further interesting way to discuss partial effects is to consider the right-tail Gini coefficient. Considering for the sake of simplicity that x is a scalar, the right tail Gini coefficient is,

$$G(x) = (2\alpha(x) - 1)^{-1}, \tag{2.15}$$

where $\alpha(x) = \exp(x\beta)$. Thus, to measure the impact of a change in x on $\alpha(x)$ and consequently on $G(x)$ we take the derivative of $G(x)$ with respect to x which, after some simplifications (noting that $\alpha(x) = [1 + G(x)]/2G(x))$, is $G'(x) = -G(x)(1+G(x))\beta$. The sign of β indicates whether an increase of x leads to an increase or decrease in $G(x)$. For example, $\beta < 0$ or $\beta > 0$ lead to an increase or a decrease of inequality in the tail, respectively, since $G'(x) < 0$ for the former or $G'(x) > 0$ for the latter. The extend of this increase or decrease depends on the magnitude of $G(x)$. From the previous expression, we establish that $(\Delta G(x)/G(x))100\% \simeq -((1+G(x))\beta\Delta x)100\%$, which represents the percentage of change in $G(x)$, when x changes by Δx (ceteris paribus). Interestingly, this result shows that the higher the inequality, the greater the impact of a change in x. Since these effects depend on x (or on $\mathbf{x}_i \in \mathbb{R}^K$ in the general case) we compute the partial effects of x_j as the average of the partial effects at every observation (average partial effect). For example, to estimate $(\Delta G(x)/G(x))\,100\%$ we consider

$$\left(-\frac{1}{n}\sum_{i=1}^{n}\left(1+\widehat{G}(\mathbf{x}_i)\right)\widehat{\beta}_j\right)100\%, \text{ where } \widehat{G}(\mathbf{x}_i) = \left(2\widehat{\alpha}(\mathbf{x}_i) - 1\right)^{-1}. \tag{2.16}$$

If the covariate under analysis is a dummy variable, say d, then the partial effect of group $d = 1$ over $d = 0$ is obtained as,

$$\frac{1}{n}\sum_{i=1}^{n}\left(\widehat{G}(\mathbf{x}_i, d_i = 1) - \widehat{G}(\mathbf{x}_i, d_i = 0)\right). \tag{2.17}$$

3 Monte Carlo simulation

In this section we evaluate the finite sample properties of the procedures and their performance in imputing mean wages above the top-code.

3.1 Finite sample performance of the tail index estimators

To evaluate the finite sample performance of the conditional tail index estimator introduced in the previous section, we conduct an in-depth Monte Carlo analysis using several data generation processes (DGPs). Specifically, data is generated from the general framework,

$$\begin{aligned} y_i &\sim D(\alpha(\mathbf{x}_i)) && (3.1)\\ \alpha(\mathbf{x}_i) &= \exp(\beta_1 + \beta_2 x_i), \qquad x_i \sim U(0,1) && (3.2)\end{aligned}$$

where $\beta_1 = \beta_2 = 1$ and the $k100\%$, $k \in (0,1)$, largest observations of the empirical distribution $D(.)$ closely follow a Pareto distribution. We consider the case of right censoring given by the censoring threshold y_c so that the sequence $\{y_i\}$ is not completely observed. Instead, we observe $w_i = \min(y_i, y_c)$.

To be more precise about the framework used to generate the data, we consider that $D(.)$ in (3.1) is either a Pareto or a Burr distribution[7] and generate samples of size $n \in \{2500, 5000, 10000, 50000\}$. The samples are censored using $y_c = \{\widehat{q}^y_{0.95}, \widehat{q}^y_{0.99}\}$ which corresponds to the 95th and 99th empirical quantile of y. For estimation of the tail index we use the $\lfloor kn \rfloor$ largest observations, with $k = 0.2$ when the Pareto distribution is considered and $k = \{0.05, 0.10, 0.20\}$ for the Burr. Note that in the case of samples generated from a Pareto distribution we could have set $k = 1$, however, in order to mimic the conditions typically found in empirical analysis a lower value was considered.

Based on the specifications described above 10,000 sequences of $\{y_i\}$ and $\{w_i\}$ of size n are generated and in each iteration three approaches are used to compute the tail index:

i) the tail index regression of Wang and Tsai (2009) applied to the sequence of $\{y_i\}$. We define the resulting estimator as $\widehat{\alpha}$. This method should provide the best results since it is applied to the original uncensored data.

ii) the censored tail index regression introduced in this paper applied to the sequence $\{w_i\}$. The resulting estimator is denoted as $\widehat{\alpha}^c$.

iii) the tail index regression of Wang and Tsai (2009) applied to the censored data $\{w_i\}$. The resulting tail index is defined as $\tilde{\alpha}$. This approach will be useful in providing information on the impact of neglecting the censoring on the tail index estimates.

Table S1 in the Supplementary Appendix provides the bias and root mean square errors (RMSEs) associated with the estimates of β_1 and β_2 in (3.2) computed based on the three approaches described in i), ii) and iii). The first observation we can make is that, in general, the largest bias and RMSEs (regardless of considering β_1 or β_2) result from the use of the approach described in iii), i.e., when censoring is ignored. Additionally, it is interesting to observe that the difference in the bias and RMSEs obtained from the approaches described in i) and ii) are relatively small, which suggest that the estimation approach which accounts for the censoring produces results close to those obtained when the sample without censoring is used for estimation as is the case in i).

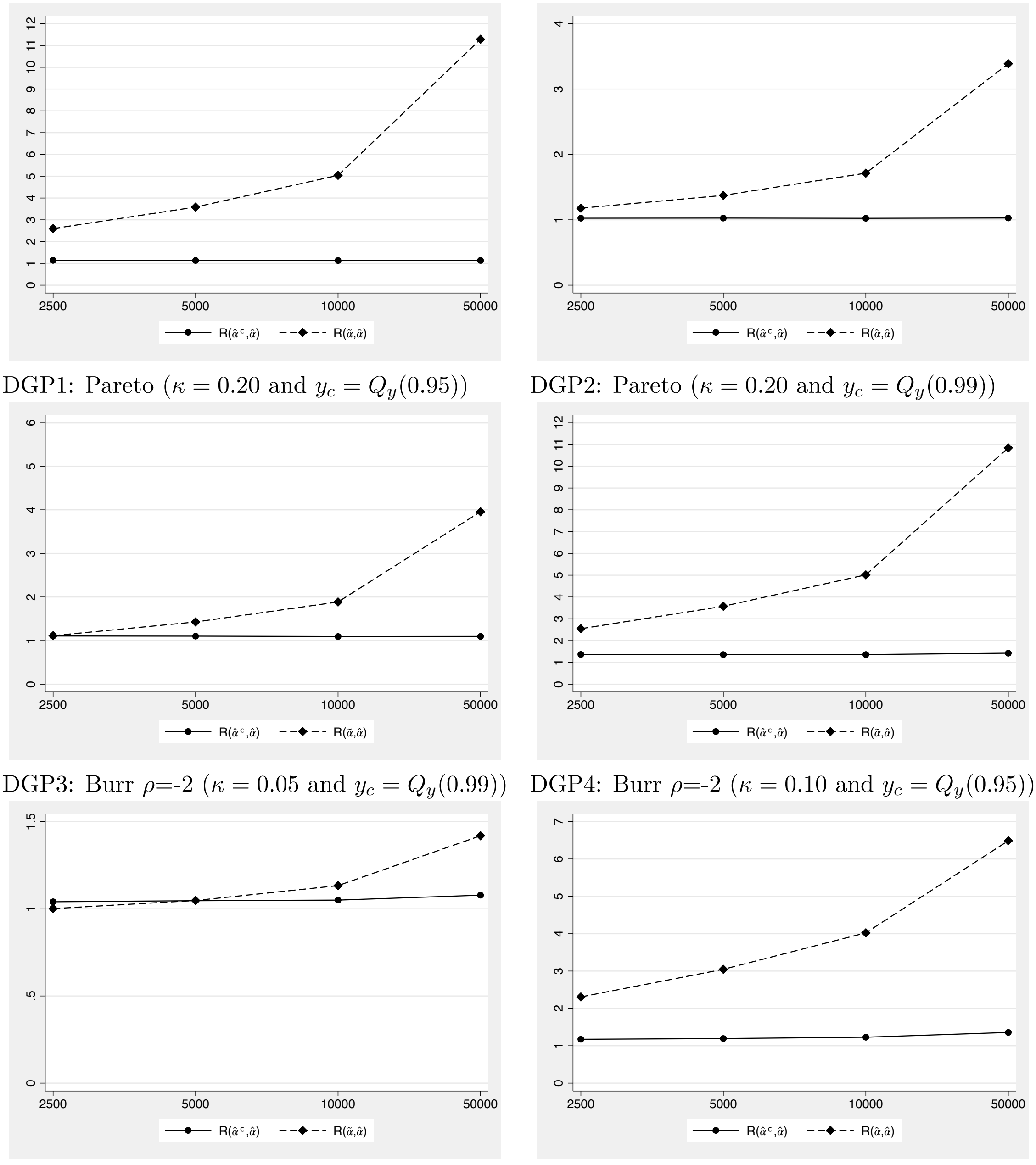
This Table also shows that the bias remains relatively stable and does not decrease as n increases. There are however different patterns according to the values of k and y_c. For instance, in Cases 3 to 6, which use the Burr distribution as DGP, a small value of k tends to improve the estimation results given that the tail of the Burr distribution approximates the tail of a Pareto distribution, this is also illustrated in Figure 1.

Figure 1 plots the ratios of the RMSEs of the α estimates obtained under i), ii) and iii), i.e.,

$$\mathsf{R}(\widehat{\alpha}^c, \widehat{\alpha}) = \frac{RMSE\left(\widehat{\alpha}^c\left(\mathbf{x}_i\right)\right)}{RMSE\left(\widehat{\alpha}\left(\mathbf{x}_i\right)\right)}, \qquad \mathsf{R}(\tilde{\alpha}, \widehat{\alpha}) = \frac{RMSE\left(\tilde{\alpha}\left(\mathbf{x}_i\right)\right)}{RMSE\left(\widehat{\alpha}\left(\mathbf{x}_i\right)\right)}.$$

Since $\widehat{\alpha}$, obtained as described in i), is in this Monte Carlo exercise the best estimator by design, $\mathsf{R}(\widehat{\alpha}^c\widehat{\alpha})$ and $\mathsf{R}(\tilde{\alpha}\widehat{\alpha})$ are larger than 1 across the different values of n. However, $\mathsf{R}(\widehat{\alpha}^c, \widehat{\alpha})$ is just slightly above 1, which means that the censored estimator, $\widehat{\alpha}^c$, performs very well and mimics closely the behavior of the best estimator, $\widehat{\alpha}$, although the former is based on the censored data. The relative performance of $\tilde{\alpha}$ worsens as n increases (Figure 1). This occurs because $\tilde{\alpha}$ is inconsistent since it neglects the right-censoring of the data. So, the bias remains relatively unchanged as n increases. The asymptotic variance of $\tilde{\alpha}$ may decrease, but the RMSE decreases only slightly since the bias does not converge to zero. On the contrary, $\widehat{\alpha}$ and $\widehat{\alpha}^c$ are consistent and, therefore, their statistical properties improve as n increases. This is reflected in a sharp decrease of the RMSE statistics (see Table S1 in the Supplementary Appendix). As a consequence the ratio $R\left(\tilde{\alpha}, \widehat{\alpha}\right)$ exhibits an increasing trend as n increases.

The censoring threshold y_c also impacts the estimation results. The lower its value, the greater the impact of censoring on estimation will be, and $\mathsf{R}(\tilde{\alpha}, \widehat{\alpha})$ tends to be larger (see, for example, the results for Case 4 in Table S1 in the Supplementary Appendix and Figure 1).



DGP1: Pareto ($\kappa = 0.20$ and $y_c = Q_y(0.95)$)

DGP2: Pareto ($\kappa = 0.20$ and $y_c = Q_y(0.99)$)

DGP3: Burr ρ=-2 ($\kappa = 0.05$ and $y_c = Q_y(0.99)$)

DGP4: Burr ρ=-2 ($\kappa = 0.10$ and $y_c = Q_y(0.95)$)

DGP5: Burr ρ=-2 ($\kappa = 0.20$ and $y_c = Q_y(0.99)$)

DGP6: Burr ρ=-2 ($\kappa = 0.20$ and $y_c = Q_y(0.95)$)

Figure 1: Ratios of the tail index estimates' RMSEs

Note: $\mathsf{R}(\widehat{\alpha}^c, \widehat{\alpha}) = RMSE\,(\widehat{\alpha}^c)/RMSE\,(\widehat{\alpha})$ and $\mathsf{R}(\tilde{\alpha}, \widehat{\alpha}) = RMSE\,(\tilde{\alpha})/RMSE\,(\widehat{\alpha})$. $\mathsf{R}(\widehat{\alpha}^c, \widehat{\alpha})$ compares the censored estimator with the best estimator $\widehat{\alpha}$ in the Monte Carlo design, while $\mathsf{R}(\tilde{\alpha}, \widehat{\alpha})$ compares the estimator that ignores the censoring with the best estimator $\widehat{\alpha}$. The results reported are based on the Pareto and the Burr distributions as DGPs.

3.2 Imputing mean wages

To provide further insight on the usefulness of the procedure introduced in this paper we compare next the performance of the different methods for imputing mean wages. Specifically, we compare,

i) the Pareto-imputed mean wage above the top-code y_c,

$$\widehat{\tau}_1 (y_c) = \frac{\widehat{\alpha}_1}{\widehat{\alpha}_1 - 1} y_c, \tag{3.3}$$

where $\widehat{\alpha}_1$ is the tail index estimate obtained assuming an uncensored Pareto distribution as in Section 2 (see e.g. Hill, 1975 and Nicolau and Rodrigues, 2019 for tail index estimators);

ii) the imputed mean wage based on the approach suggested by Armour et al. (2016),

$$\widehat{\tau}_2 (y_c) = \frac{\widehat{\alpha}^c_{Hill}}{\widehat{\alpha}^c_{Hill} - 1} y_c, \tag{3.4}$$

where $\widehat{\alpha}^c_{Hill}$ is a consistent estimate of the tail index parameter computed as in (2.5). Note that $\tau_2 (y_c) := E (y_i | y_i > y_c)$;

iii) two approaches to impute wages based on the method introduced in this paper:

a) the mean predictor,

$$\widehat{\tau}_3 (\mathbf{x}_i, y_c) = \frac{\widehat{\alpha} (\mathbf{x}_i)}{\widehat{\alpha} (\mathbf{x}_i) - 1} y_c, \tag{3.5}$$

b) the median predictor,

$$\widehat{\tau}_4 (\mathbf{x}_i, y_c) = 2^{1/\widehat{\alpha}(\mathbf{x}_i)} y_c, \tag{3.6}$$

where $\widehat{\alpha} (\mathbf{x}_i) = \exp \left(\mathbf{x}_i' \widehat{\boldsymbol{\theta}} \right)$.

Note that $\tau_3 (\mathbf{x}_i, y_c) := E (y_i | \, \mathbf{x}_i, y_i > y_c)$ and that $\widehat{\tau}_4 (\mathbf{x}_i, y_c)$ is particularly useful when $\widehat{\alpha} (\mathbf{x}_i) \leq 1$.

The main difference between $\widehat{\tau}_k (\mathbf{x}_i, y_c)$, $k = 3, 4$, and $\widehat{\tau}_1 (y_c)$ and $\widehat{\tau}_2 (y_c)$ is that in the former the particular characteristics of the individuals whose wage is above the threshold are explicitly taken into account through $\mathbf{x}_i$. Interestingly, $\widehat{\tau}_3(\mathbf{x}_i, y_c)$ corresponds to the optimal mean square predictor since it is the conditional expectation of y_i given $\mathbf{x}_i$ and $y_i > y_c$. This follows from the well known result $E\left(y_i - E\left(y_i \middle| \mathbf{x}_i, y_i > y_c\right)\right)^2 \leq E\left(y_i - g(.)\right)^2$, where $g(.)$ is any other predictor of y_i given $y_i > y_c$. It turns out that $\tau_2(y_c)$ is optimal only if y_i is mean-independent of $\mathbf{x}_i$, in which case both $\tau_2(y_c)$ and $\tau_3(\mathbf{x}_i, y_c)$ coincide.

Thus, having established the superiority of $\tau_3(\mathbf{x}_i, y_c)$, it remains to be shown how much improvement is obtained from $\tau_3(\mathbf{x}_i, y_c)$ when compared to $\tau_1(y_c)$ and $\tau_2(y_c)$ in computing the mean wage above y_c. The following Monte Carlo study looks to answer this question. Our experiment is based on the following steps:

1. Simulate y_i, $i = 1, 2, ..., n$, with $n \in \{250, 500, 1000, 2000, 5000, 20000\}$, according to a conditional Pareto distribution $P(\alpha(\mathbf{x}_i))$ where $\alpha(\mathbf{x}_i) = \exp(1 + 2x_i)$ and $x_i \sim U(0, 1)$. For each $i = 1, 2, ..., n$ simulate $\alpha(\mathbf{x}_i)$ and then the corresponding y_i;

2. All observations above quantile 0.95, which corresponds to y_c, are censored, but their original values are saved for comparison purposes (these values are used to assess the estimators' predictive precision). The data used for estimation are $\{w_i, i = 1, 2, ..., n\}$, where $w_i = \min(y_i, y_c)$ and from which we compute $\widehat{\tau}_1(y_c)$, $\widehat{\tau}_2(y_c)$, $\widehat{\tau}_3(\mathbf{x}_i, y_c)$ and $\widehat{\tau}_4(\mathbf{x}_i, y_c)$.

3. Predict the imputed mean value above the threshold y_c based on $\widehat{\tau}_1(y_c)$, $\widehat{\tau}_2(y_c)$, $\widehat{\tau}_3(\mathbf{x}_i, y_c)$ and $\widehat{\tau}_4(\mathbf{x}_i, y_c)$.

4. Steps 1 to 3 are repeated 2000 times and the mean square errors (MSE) of the estimators computed.

Other combinations of $\alpha(\mathbf{x}_i)$ produce essentially the same results as long as $\alpha(\mathbf{x}_i) \geq 1$, and for this reason only results for $\alpha(\mathbf{x}_i) = \exp(1 + 2x_i)$ (note that $E(\alpha(\mathbf{x}_i)) = 3.7$) are presented. Note that, $\alpha(\mathbf{x}_i)$ should be set so that $\alpha(\mathbf{x}_i) > 1$, otherwise the conditional and marginal expected values do not exist, and consequently $\widehat{\tau}_1(y_c)$, $\widehat{\tau}_2(y_c)$ and $\widehat{\tau}_3(\mathbf{x}_i, y_c)$

are not well defined. The case $0 < \alpha(\mathbf{x}_i) < 1$ can be dealt with using $\widehat{\tau}_4(\mathbf{x}_i, y_c)$, as indicated above. Figure 2 illustrates our results.

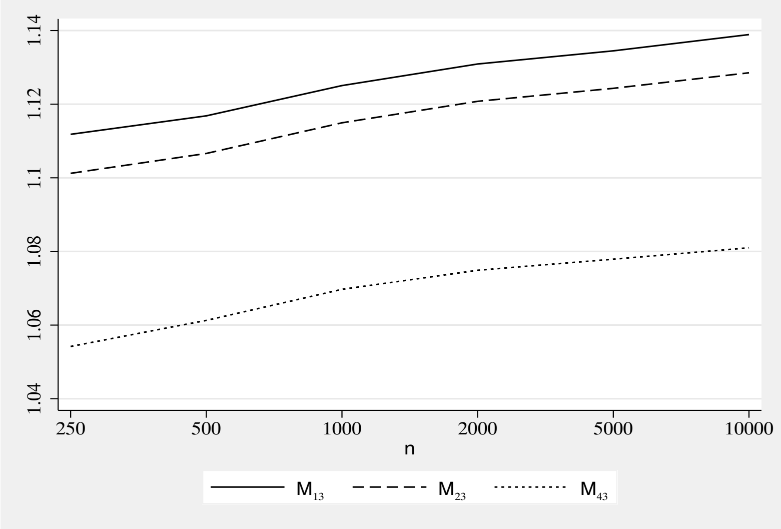


Figure 2: MSE ratios computed for sample sizes $n = (250, 500, 1000, 2000, 5000, 10000)$

Note: The lines represent the MSE ratios, $\mathsf{M}_{13} := \frac{MSE(\widehat{\tau}_1(y_c))}{MSE(\widehat{\tau}_3(\mathbf{x}_i,y_c))}$, $\mathsf{M}_{23} := \frac{MSE(\widehat{\tau}_2(y_c))}{MSE(\widehat{\tau}_3(\mathbf{x}_i,y_c))}$ and $\mathsf{M}_{43} := \frac{MSE(\widehat{\tau}_4(\mathbf{x}_i,y_c))}{MSE(\widehat{\tau}_3(\mathbf{x}_i,y_c))}$.

Figure 2 shows that $\widehat{\tau}_3(\mathbf{x}_i, y_c)$ and $\widehat{\tau}_4(\mathbf{x}_i, y_c)$ produce the best results. M_{13} and M_{23} are larger than one and larger than M_{43}, and the gains increase steadily as the sample size grows. Moreover, $\widehat{\tau}_2(y_c)$ (Armour et al., 2016) is better than $\widehat{\tau}_1(y_c)$ and $\widehat{\tau}_3(\mathbf{x}_i, y_c)$ is slightly better than $\widehat{\tau}_4(\mathbf{x}_i, y_c)$.

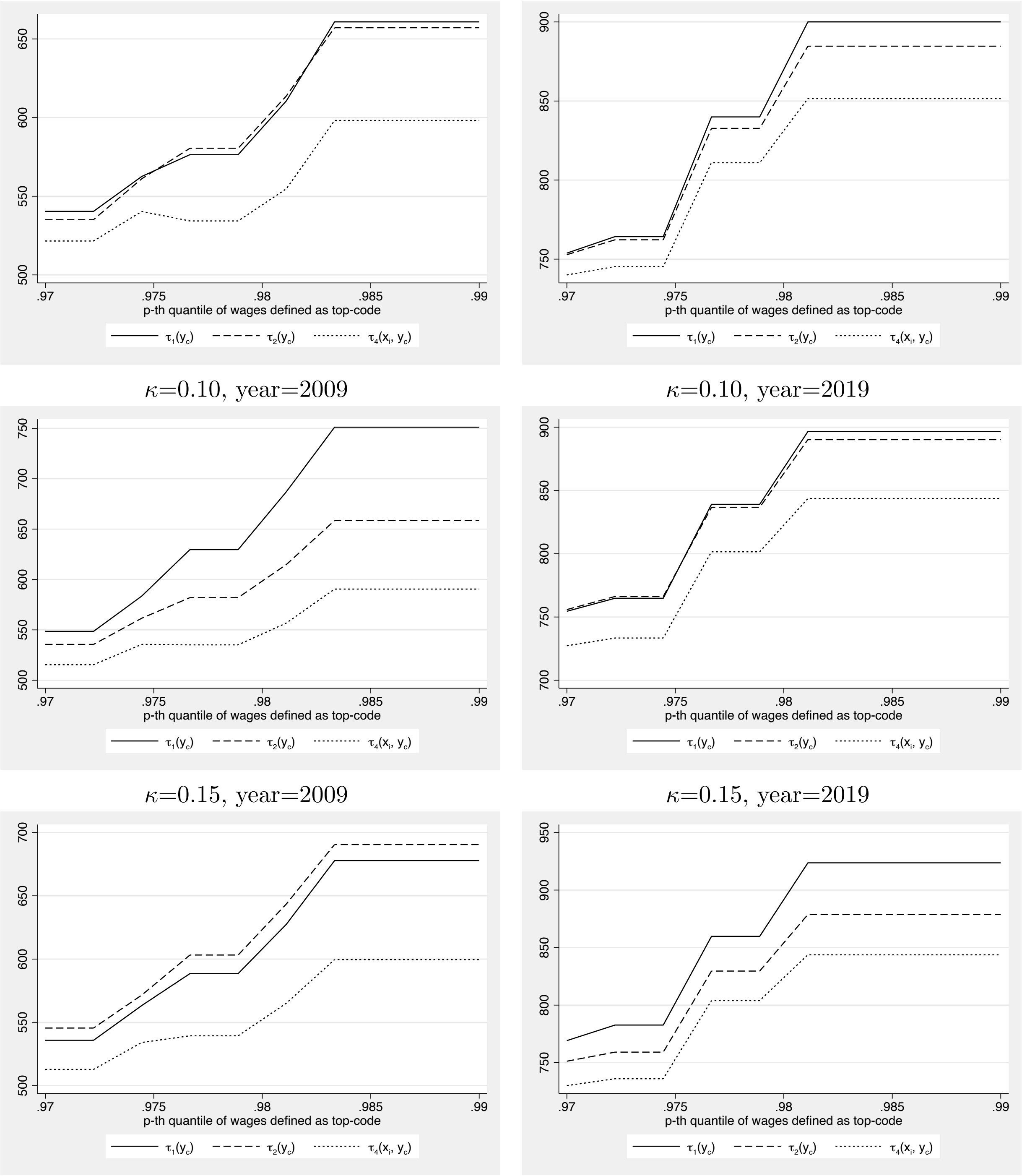
3.3 Robustness testing using non-censored data

To evaluate the performance of the conditional tail index estimator in a real situation we use the Portuguese labor force survey. As wages in this survey are not censored, we can simulate different top-codes and thus accurately measure the potential gains of our methodology (as an ordinary out-of-sample forecast exercise). We censor the sample considering $y_c \in \{\widehat{q}^y_{0.97}, \widehat{q}^y_{0.975}, ..., \widehat{q}^y_{0.99}\}$, where the generic element $\widehat{q}^y_p$ corresponds to the *pth* empirical quantile of wages. The order of these quantiles is approximately the same as the top-codes in the CPS database since 1992 (they have been decreasing from 0.99 in 1992 to 0.95 in 2017).

To estimate the conditional tail index, we use the $\lfloor \kappa n \rfloor$ largest observations, where $\kappa \in \{0.1, 0.15, 0.2, 0.25\}$. We start by computing $\widehat{\alpha}(\mathbf{x}_i)$ from the censored tail index regression introduced in this paper, and use $\widehat{\tau}_4(\mathbf{x}_i, y_c)$ to predict the censored wage values. At this point we should add two remarks. The first is related to the choice of the covariates $\mathbf{x}_i$. We have more than 30 variables to choose from in our data set, such as, age, gender, tenure, education, job, region and industry. However, instead of searching for the covariates with greatest explanatory power in each year, for simplicity of analysis, we use the same basic covariates throughout all combinations of year, κ and y_c. Specifically, $\mathbf{x}_i = [gender_i, age_i, education_i, tenure_i]$. Although we do not fully demonstrate the potential of our estimator over the competing estimators (as likely we do not use all relevant explanatory variables), the results still clearly show that the tail index regression is preferable. The second remark is related to the use of the conditional median, $\widehat{\tau}_4(\mathbf{x}_i, y_c)$. Since, some estimates of $\widehat{\alpha}(\mathbf{x}_i)$, $i = 1, 2, ..., n$, are close to 1, and given that for these cases, the conditional mean $\widehat{\tau}_3(\mathbf{x}_i, y_c)$ may not be defined, we opt for the use of $\widehat{\tau}_4(\mathbf{x}_i, y_c)$, which, in most cases, performs very well (see results in Section 3.2). For comparison purposes we also use $\widehat{\tau}_1(y_c)$ in (3.3) and $\widehat{\tau}_2(y_c)$ in (3.4) to predict the censored wages.

The forecast performance is assessed with the RMSE statistics, $RMSE(j) = \sqrt{\frac{1}{s}\sum_{i=1}^{s}\left[\widehat{\tau}_j(y_c) - y_i\right]^2}$, 1, 2 and $RMSE(4) = \sqrt{\frac{1}{s}\sum_{i=1}^{s}\left[\widehat{\tau}_4(\mathbf{x}_i, y_c) - y_i\right]^2}$, where s is the number of predicted wages above the top-code; results are presented in Figure 3. $\widehat{\tau}_2(y_c)$ generally performs better than $\widehat{\tau}_1(y_c)$ in 2009 and 2019. This is especially clear when the top-code is defined for high or very high wages. In these cases, the number of wages above the top-code is relatively small, but their magnitude is large (forecast errors are particularly high in these cases).

$\widehat{\tau}_4(\mathbf{x}_i, y_c)$ outperforms the other estimators, especially when the top-code is defined for high or very high wages. Adding more information about the individuals whose salary is coded significantly improves the quality of the forecast, compared to other non-conditional methods, especially when the top-code is relatively high. We have also observed that by adding more information related to profession, industry and region, further improves the quality of the forecasts (results are available upon request).



κ=0.10, year=2009

κ=0.10, year=2019

κ=0.15, year=2009

κ=0.15, year=2019

κ=0.20, year=2009

κ=0.20, year=2019

Figure 3: RMSE of imputed wages based on $\widehat{\tau}_1(y_c)$, $\widehat{\tau}_2(y_c)$ and $\widehat{\tau}_4(\mathbf{x_i}, y_c)$, computed from non-censored wage data from the Portuguese Labor force survey.

4 Empirical analysis of the CPS data

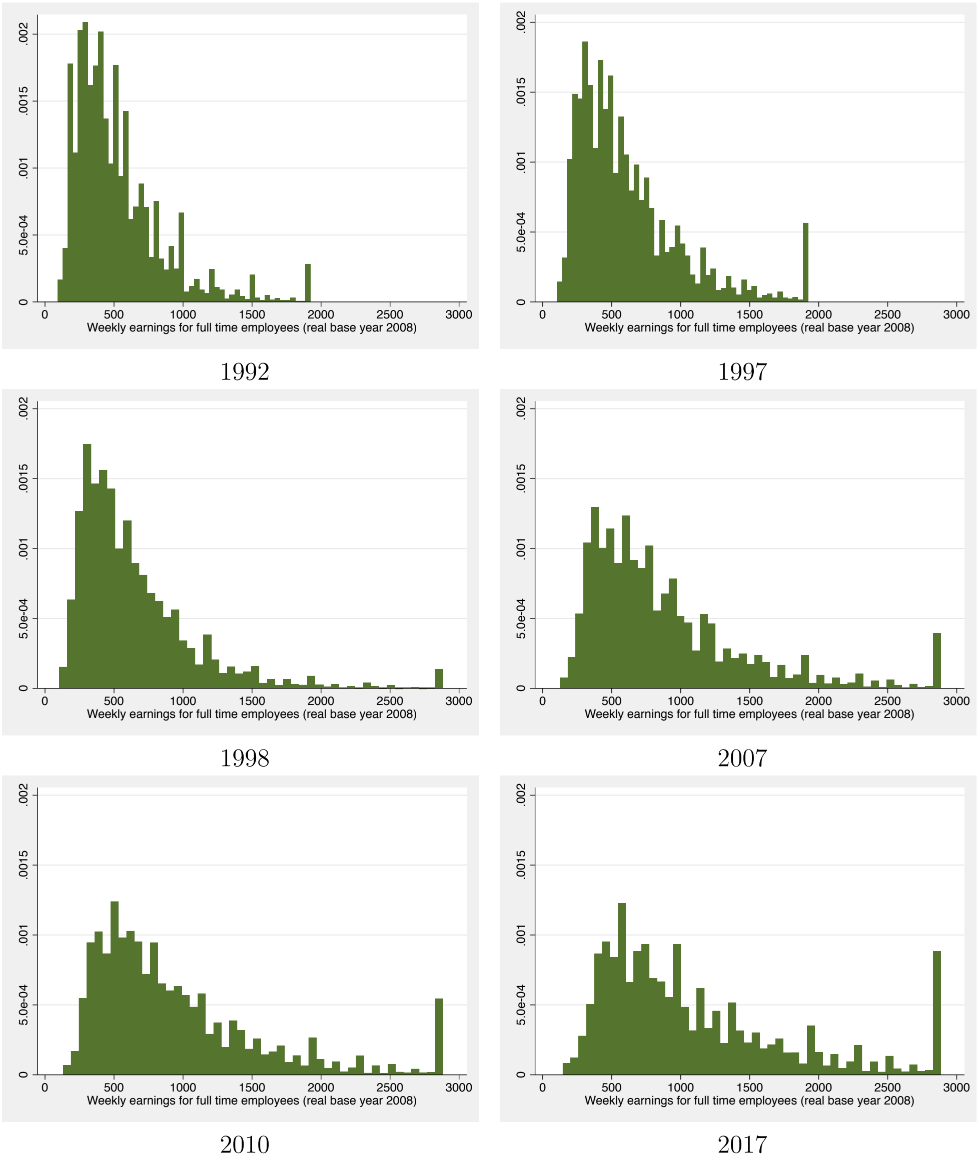
In this section, we use the censored publicly available CPS data to evaluate how the right tail inequality of the US weekly wage distribution has changed over time and how these changes may differ across the characteristics of individuals, occupations and industries. In specific, we show that the new tail index estimator introduced provides very rich and detailed insights about the right tail distribution of weekly wages and inequality. We also assess the sensitivity of the adjustment of the top-coded weekly wage to changes over time and across the characteristics of individuals.

4.1 Data

For the empirical analysis the March CPS files from IPUMS for the period between 1992 and 2017 are used. The weekly wage measure is top-coded at \$1923 between 1989 and 1997, and at \$2884 between 1998 and 2017. The sample is restricted to workers between 16 and 64 year-old on full-time full year basis employed during the CPS sample survey reference week (35+ hours per week, 40+ weeks per year). Following Autor and Dorn (2013) the real weekly wage data are weighted by the appropriate CPS weight to provide a measure of the full distribution of the weekly wages paid.[8,9]

Occupations are defined as job task requirements of the US Department of Labor Dictionary of Occupational Titles (DOT, 1977) and Census occupation classifications for routine, abstract and manual task classifications (Autor and Dorn, 2013).

In Figure 4 we present the distribution of the weekly wages for 1992, 1997, 1998, 2007, 2010 and 2017. From 1992 to 2017 the concentration of wages has become more skewed to the right. Between 1992 and 1997 the mass points around the top-code increased and with the relaxing of the top-code in 1998 real values beyond that top-code are potentially observable. The same phenomenon also occurs in the most recent period. In 2017 the mass point around the current top-code used in the CPS data is much larger.



1992

1997

1998

2007

2010

2017

Figure 4: Annual unconditional histograms of the weekly wage.

The means and proportions of workers according to their characteristics, occupations and industry, for observations above the 80th percentile, are presented in Table 1. In contrast to 1992, in 2017 the population in the right tail is older (41.03 years on average in 1992 and 43.83 years in 2017), the percentage of women is greater (25% in 1992 increased to 33% in 2017), and is about one year more educated (15.20 years of education in 1992 and 16.06 years in 2017).

Table 1: **Characteristics of individuals in the right tail (observations above percentile 80) of the wage distribution: means and proportions**

	Year	
Mean	1992	2017
Age	41.03	43.83
Female	0.25	0.33
Education	15.20	16.06
Race		
White	0.90	0.83
Black	0.05	0.05
Other race	0.05	0.12
Marital Status		
Married	0.83	0.81
Married no spouse	0.01	0.02
Separated	0.02	0.01
Divorced	0.09	0.09
Widowed	0.01	0.01
Single	0.14	0.16
Occupations		
Managers	0.72	0.81
Administrative	0.10	0.08
Low skill	0.01	0.01
Craft	0.04	0.01
Operators	0.02	0.01
Transportation	0.11	0.08
Industry[a]		
Agriculture	0.02	0.03
Construction	0.05	0.05
Manufacturing	0.22	0.14
Transports	0.11	0.08
Trade	0.11	0.09
Finance	0.09	0.10
Repair	0.04	0.10
Personal	0.28	0.32
Public	0.08	0.09
Observations	112,960	64,002
Observations above Percentile 80	22,485	12,791
Observations censored	1,110	3,197

Notes: This table reports the means for the variables used in the analysis for both 1992 and 2017. These statistics were calculated for the right tail (for observations above percentile 80). All variables are reported on a scale between 0 and 1 with the exception of age and education which are reported in years. The occupation dummies using 6 aggregate occupation groups are based on the International Standard Classification of occupations (ISCO) as used in Autor and Dorn (2013). The category Managers includes management, professional, technical, financial sales and public security occupations. The category Administrative consists of routine non cognitive occupations and includes administrative support and retail sales occupations. The category Low-skill includes low-skill services, such as cleaning, guard, food, health, janitors, beauty, recreation, working with children and other personal low-skill services. The category Craft aggregates precision production and craft occupations. The category Operators refers to machine operators, assemblers and inspectors. Finally, the category Transportation includes transportation, construction, mechanics, mining and agricultural occupations. In Table S2 of the Supplementary Appendix we provide a detailed description of the industry classification.

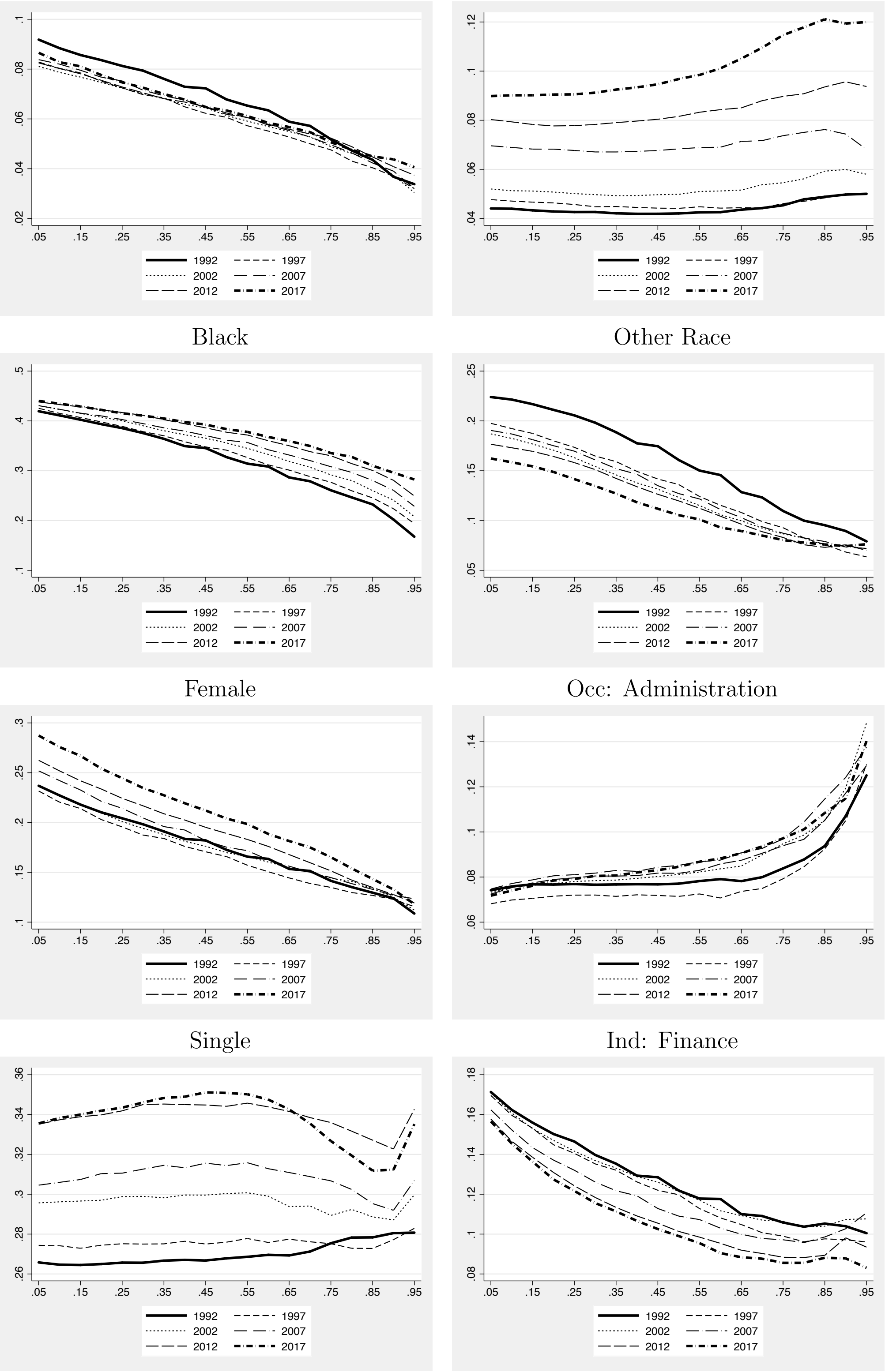
The results in Table 1 show that the 7% decrease of white individuals in the right tail in 2017, when compared to 1992, seems to be compensated by a 7% increase of individuals from other races (non-white nor non-black). There is a 3% change in the composition of the sample, where the reduction of married individuals is compensated by an increase in individuals that are single. However, married still represents the marital status of the majority of individuals in the right tail (83% in 1992 and 81% in 2017).

A further observation that can be made from the results in Table 1 is job polarization. A significant growth in employment in the right tail for non-routine cognitive tasks is observed in detriment of routine occupations (see also Autor, 2019 and Goos and Manning, 2007). The decrease in the share of employment is even more significant for non-routine manual tasks (individuals in occupations associated with transportation, construction and mechanics (Transportation) decreased their share of employment in the right tail, from 11% in 1992 to 8% in 2017). The percentage of individuals in occupation Managers, which is the occupation with the largest proportion of individuals, increased from 72% in 1992 to 81% in 2017.

In 1992, the proportion of individuals, working in manufacturing (transports) was 22% (12%) and this proportion decreased to 14% (8%). This decrease in the share of employment was compensated by an increase in the repair (+6% between 1992 and 2017) and finance, personal and public industries (+6% between 1992 and 2017). Note that the industries with the largest number of individuals in the right tail in (1992, 2017) are Personal (28%, 32%), Manufacturing (22%, 14%), Finance (9%, 10%), Repair (4%, 10%) and Public (8%, 9%). However, Manufacturing (22%, 14%), Transport (11%, 8%) and Trade (11%, 9%) see their weight decrease in 2017.

To illustrate the evolution of the proportions of individuals in the different percentiles of the overall weekly wage distribution, Figure 5 plots the proportions in percentiles 0.05 to 0.95 considering different attributes, occupations and industries (in the Supplementary Appendix we present additional plots for all other cases analyzed in Table 1). From this Figure we distinguish two patterns from 1992 to 2017: Other Race, Female, Single and Personal increase in proportion from 1992 to 2017 across all percentiles, whereas Administration and Trade decrease across all percentiles. The number of Black individuals seems to decrease up to around percentile 80 and increases thereafter.

Moreover, this Figure also shows that individuals that are Black, Female or Single, as well as individuals working in Administration and Trade display a downward trend over the percentiles, whereas the number of individuals of Other Races and individuals working in the Personal industry display a different pattern. The former is relatively constant across all percentiles in 1992, but increases for percentiles larger than the median in 2017, and the latter is relatively constant across all percentiles in 1992 and 2017. Interestingly, Finance shows a different pattern when compared to all other covariates. In specific, the largest proportions are observed at the higher percentiles (i.e. from the median onward). This pattern is very similar across all years analyzed.



Black

Other Race

Female

Occ: Administration

Single

Ind: Finance

Ind: Personal

Ind: Trade

Figure 5: Proportion of individuals in different weekly wage percentiles

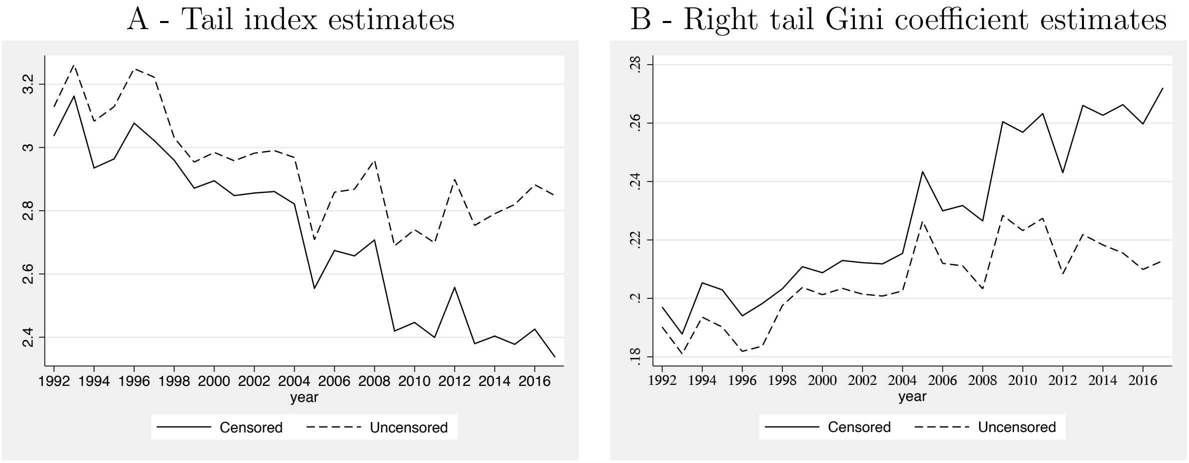
Note: The graphs represent the proportion of individuals in different categories for percentiles 0.05, 0.1, 0.15,..., 0.95, computed for the years of 1992, 1997, 2002, 2007, 2012 and 2017.

4.2 Tail index and Gini coefficient estimation results

Figure 6 plots the unconditional right-tail index estimates and corresponding Gini coefficients. The tail index is computed as an average of the conditional tail index estimates obtained from uncensored and censored tail index regressions, i.e.,

$$\widehat{\alpha}_t = \frac{1}{n}\sum_{i=1}^{n} \widehat{\alpha}\left(\mathbf{x}_{t,i}\right), \ t = 1992, ..., 2017 \tag{4.1}$$

where $\widehat{\alpha}\left(\mathbf{x}_{t,i}\right) = \exp\left(\mathbf{x}'_{t,i}\widehat{\theta}_t\right)$. The right tail Gini coefficient estimates are computed as indicated in (2.15).


A - Tail index estimates

B - Right tail Gini coefficient estimates

Figure 6: Uncensored and censored right tail index estimates (left-plot) and corresponding right tail Gini coefficient estimates (right-plot) from 1992 to 2017

This figure shows, on the one hand, that the uncensored approach misrepresents the true unconditional tail index, as it ignores the top-coded wages, leading to overestimation of the true values of α_t, which corresponds to an under estimation of inequality as can be seen from the Gini coefficient plot.[10] On the other hand, it also shows that the censored estimates of $\widehat{\alpha}_t$ have decreased over the last 20 years, which corresponds to an increase of the Gini coefficient value. In other words, this Figure shows that the probability of observing an extreme value today is higher compared to the 90s or even in the more recent past. This finding, which is also illustrated through the Gini coefficient, supports the idea that upper-tail inequality has increased since the 90s and has become more pronounced over the last 20 years. Autor et al. (2008) observed that the 90th percentile wage rose by more than 55% relative to the 10th percentile between 1963 and 2005, which represents a significant increase. However, our approach suggests that the increase in inequality found by Autor et al. (2008) although significant may correspond to a lower bound of the true increase in inequality.

Figure 6 also highlights the link between the tail index and the Gini coefficient. The Gini coefficient plot illustrates the increase in inequality detected by the censored tail index estimator. The uncensored estimates seem to suggest a stabilization of inequality in the right tail starting in 2005, and the Gini coefficient estimates suggest that inequality in recent years is relatively low.

Complementary to Figure 6, Table 2 presents the censored and uncensored tail index regression results for 1992 and 2017. As discussed in Section 2.2, a negative regression coefficient corresponds to a decrease in the right tail index (ceteris paribus) and hence a larger number of extreme values may result as a consequence of changes in the specific variable associated to such a coefficient. Similarly, a positive regression coefficient leads to an increase in the tail index which suggests that a smaller number of extreme values may occur.

Table 2 shows that, in general, estimates based on the method that ignores censored data underestimate the true effects of the variables (which is in line with Figure 6). This is especially clear in the estimates for 2017 (e.g. female and finance). However, the direction of the impact of the covariates is consistent with the results obtained from the censored tail index regression.

Comparing the censored estimation results for 1992 and 2017, we generally observed a decrease in the estimates for 2017 (e.g. Transportation and Craft and Precision). In some cases, although in general not statistically significant (see e.g. other races, married no spouse, widowed, Transports and Trade), positive estimates in 1992 become negative in 2017 and vice versa. Considering only the statistically significant covariates it can be observed that Female, Black, Divorced, Single, Low Skill, Craft, Operators, Transportation and Public have a positive impact leading to a reduction in the probability of individuals with these characteristics being in the right tail (i.e. a decrease in the right tail Gini coefficient), whereas Age, Education, Construction, Finance and Repair have a negative impact, originating an increase of the probability of individuals with these characteristics being in the right tail (i.e. an increase in the right tail Gini coefficient).

It is clear from Table 2 that industries such as Finance, Construction and Repair are the ones with more extreme wages. The probability of observing an extreme value increased in 2017 by 4.94% for an individual working in the Finance industry. The probability of an individual working in the public industry having a wage higher than the 96th percentile decreased in 2017 by 2.14%. The biggest increase from 1992 to 2017 was observed for individuals that are black, married without a spouse, or widowed. Older and more educated workers continued to have a significant probability of observing an extreme wage but there was no relevant change between 1992 and 2017. Women observed a positive increase between 1992 and 2017, but the impact is still towards an increase in the tail index (although smaller in absolute value than in 1992), i.e., a decrease in the probability of being in the right tail. In specific, women in 2017 have a partial effect on the tail index of -1.26% which corresponds to a decrease in the probability of a wage being in the right tail.

The impact in terms of occupation is very interesting. For instance, in contrast to an individual working in a non-routine cognitive occupation (managers) all other occupations display a positive contribution to observe an extreme value (although smaller in 2017). However, the picture is very different across occupations. Individuals working in routine occupations such as administrative workers reduced their presence in the right tail from 10% to 8% (see Table 1) but the probability to observe an extreme value for these occupations increased in 2017 (-2.14% in 1992 and changed to 2.00% in 2017). The contribution to the probability of observing an extreme value for an individual working as an operator was significantly higher in 1992 than in 2017 (the partial effect was -8.04% in 1992 and it reduced to -1.41% in 2017).

The right tail Gini coefficient's partial effects confirm the general picture already presented. First, the groups where the upper-tail inequality (within groups) increased the most, from 1992 to 2017, were Married no spouse, Repair, Widowed and Finance; second, the groups where the upper-tail inequality decreased the most from 1992 to 2017 were Agriculture, Public, Low Skill and Transportation. Third, the percentage change in the right tail Gini coefficient when education increases by one year (ceteris paribus) was 10.4% in 1992 and 13% in 2017.

Table 2: Uncensored and censored tail index regression results

					partial effects		gini partial effects	
	Uncensored		Censored		Censored		Censored	
	1992	2017	1992	2017	1992	2017	1992	2017
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Constant	2.597***	2.321***	2.723***	2.812***				
	(0.066)	(0.076)	(0.075)	(0.121)				
Age†	−0.011***	−0.007***	−0.013***	−0.012***	0.211	0.195	1.55%	1.53%
	(0.001)	(0.001)	(0.001)	(0.001)				
Female	0.272***	0.134***	0.309***	0.229***	−5.714	−1.262	−0.043	−0.047
	(0.014)	(0.013)	(0.016)	(0.022)				
Education†	−0.076***	−0.061***	−0.085***	−0.100***	1.391	1.638	10.36%	12.97%
	(0.004)	(0.004)	(0.004)	(0.006)				
Race								
Black	0.085***	0.004	0.098***	0.010***	−1.659	1.945	−0.044	−0.031
	(0.029)	(0.034)	(0.031)	(0.054)				
Other race	−0.023***	0.004	−0.031	0.015	0.497	1.888	0.015	0.014
	(0.025)	(0.016)	(0.030)	(0.033)				
Marital Status								
Married no spouse	0.075	−0.036	0.095	−0.1	−1.607	3.468	−0.018	0.067
	(0.063)	(0.046)	(0.070)	(0.091)				
Separated	0.035	0.051	0.044	0.093	−0.728	0.68	−0.025	−0.038
	(0.042)	(0.059)	(0.047)	(0.096)				
Divorced	0.018	0.066***	0.025	0.119***	−0.411	0.26	−0.019	−0.044
	(0.019)	(0.022)	(0.022)	(0.036)				
Widowed	0.003	−0.070	0.022	−0.083	−0.361	3.248	−0.010	0.031
	(0.083)	(0.052)	(0.077)	(0.099)				
Single	0.009	0.066***	0.015	0.114***	−0.245	0.4	−0.023	−0.061
	(0.018)	(0.018)	(0.021)	(0.029)				
Occupations								
Administrative	0.113***	0.021	0.125***	0.006	−2.141	2.001	−0.023	−0.007
	(0.020)	(0.022)	(0.023)	(0.047)				
Low Skill	0.098	0.093	0.111*	0.153*	−1.89	−0.276	−0.044	−0.083
	(0.064)	(0.060)	(0.067)	(0.091)				
Craft	0.289***	0.069	0.324***	0.154*	−6.029	−0.353	−0.049	−0.064
	(0.033)	(0.056)	(0.035)	(0.086)				
Operators	0.394***	0.148***	0.416***	0.217***	−8.042	−1.412	−0.075	−0.100
	(0.050)	(0.062)	(0.053)	(0.09)				
Transportation	0.366***	0.184***	0.400***	0.264***	−7.682	−2.43	−0.084	−0.122
	(0.023)	(0.030)	(0.025)	(0.045)				
Industry								
Construction	−0.202***	−0.117***	−0.234***	−0.173***	3.449	4.417	−0.051	−0.086
	(0.039)	(0.042)	(0.047)	(0.071)				
Manufacturing	0.003	−0.011	0.004	0.026	−0.065	1.677	0.007	0.008
	(0.028)	(0.029)	(0.031)	(0.052)				
Transports	0.050***	−0.016	0.051**	−0.001	−0.847	2.106	−0.039	−0.043
	(0.021)	(0.027)	(0.024)	(0.04)				
Trade	0.002	−0.052**	0.001	−0.052	−0.016	2.842	0.001	0.008
	(0.020)	(0.024)	(0.025)	(0.043)				
Finance	−0.126***	−0.099***	−0.184***	−0.219***	2.769	4.943	0.090	0.120
	(0.020)	(0.022)	(0.026)	(0.043)				
Repair	−0.059***	−0.098***	−0.071**	−0.142***	1.12	3.921	0.024	0.065
	(0.028)	(0.022)	(0.033)	(0.0419				
Personal	0.125***	0.048***	0.127***	0.060*	−2.177	1.248	0.015	0.012
	(0.016)	(0.018)	(0.019)	(0.033)				
Public	0.238***	0.120***	0.272***	0.265***	−4.953	−2.142	−0.028	−0.068
	(0.022)	(0.024)	(0.025)	(0.04)				

Note: This table reports the tail index regression results for 1992 and 2017. The first two columns present the uncensored results while the last two columns contain the censored results. The omitted categories are white, married, working as manager and the agriculture industry. Standard errors in parentheses and *, **, *** indicate significance at the 10%, 5% and 1% significance levels.
† It is important to note that the partial effects of the continuous and dummy variables is computed as in (2.16) and (2.17), respectively.

4.3 Imputed mean wages

4.3.1 Imputing wages above the top-code

A further important contribution of the approach introduced in this paper is its use for the imputation of mean weekly wages above the top-code as described in Section 3.2. Some authors use values to adjust the top-coded weekly wage which are time-varying and differ by group. For instance, Macpherson and Hirsch (1995) provide separate Pareto estimates according to gender and by year from 1973 to 2014 using public CPS-Merged Outgoing Rotation Groups (CPS-MORG). These authors indicate that these values increase over time and are higher for men than for women (e.g. for 2014 the adjustment coefficient is 2.06 for men and 1.81 for women). In contrast, our analysis is based on the March CPS (outgoing rotation groups 4 and 8), and on the weekly wages instead of annual earnings, however, we also find evidence in favor of changing adjustment parameters.

This renders support to the observation that imputed wages above the top-code, based on a fixed value may lead to misstatement of results, given that this approach considers wages above the top-code to be independent of time, age, gender, race and other personal characteristics; as well as industry and occupation.

To impute wages above the top-code, we consider the estimate of $E\left(y_i \middle| y_i > y_c; \mathbf{x}_i\right)$ given by $\widehat{\tau}_3\left(\mathbf{x}_i, y_c\right)$ in (3.5). However, for some individuals in the CPS database, especially those in the highest wage groups, $\widehat{\alpha}\left(\mathbf{x}_i\right) \leq 1$, which implies that $E\left(y_i \middle| y_i > y_c; \mathbf{x}_i\right)$ does not exist and, $\widehat{\tau}_3\left(\mathbf{x}_i, y_c\right)$ is inadequate. For these cases, we use the conditional median of y_i given $y_i > y_c$ and $\mathbf{x}_i$, i.e., the $\widehat{\tau}_4\left(\mathbf{x}_i, y_c\right)$ statistic in (3.6). In specific, to accommodate all situations (i.e. low and high values of $\widehat{\alpha}\left(\mathbf{x}_i\right)$) we use $\widehat{\tau}_4\left(\mathbf{x}_i, y_c\right)$ when $0 < \widehat{\alpha}\left(\mathbf{x}_i\right) \leq c$ and $\widehat{\tau}_3\left(\mathbf{x}_i, y_c\right)$ for $\widehat{\alpha}\left(\mathbf{x}_i\right) > c$. In the empirical application we set $c = 1.5$, since using $c = 1$ may lead to explosive estimates as $\frac{\widehat{\alpha}(\mathbf{x}_i)}{\widehat{\alpha}(\mathbf{x}_i)-1} \to \infty$ as $\widehat{\alpha}\left(\mathbf{x}_i\right) \to 1^+$.[11] In the application to the CPS data we observed that for the overwhelming majority of estimates $\widehat{\alpha}\left(\mathbf{x}_i\right) > 1.5$, and therefore $\widehat{\tau}_3\left(\mathbf{x}_i, y_c\right)$ is mostly used. For simplicity of presentation, in the discussion below we will just refer to $\widehat{\tau}_3\left(\mathbf{x}_i, y_c\right)$ although imputed wages were computed as described above.

Figure 7 illustrates the estimates of the imputed wages above the top-code computed from the different approaches discussed in (3.3), (3.4) and (3.5). The $\widehat{\tau}_2(y_c)$ and $\widehat{\tau}_3\left(\mathbf{x}_i, y_c\right)$ estimates are similar. This result is expected given that for the overall analysis $\widehat{\tau}_3\left(\mathbf{x}_i, y_c\right)$ is based on the values of $\widehat{\alpha}_t$ computed as in (4.1), which provides an estimate of the unconditional tail index after considering the characteristics of all individuals in the sample. However, in the case where estimates for a particular group, occupation or industry are considered, the $\widehat{\tau}_3\left(\mathbf{x}_i, y_c\right)$ estimates will certainly differ from $\widehat{\tau}_2(y_c)$ (see next section).

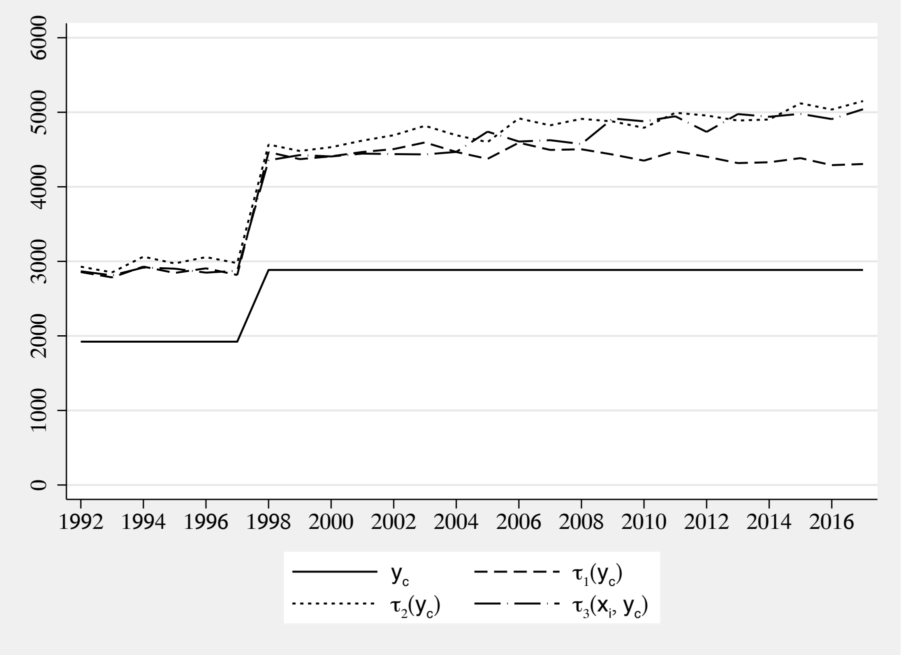
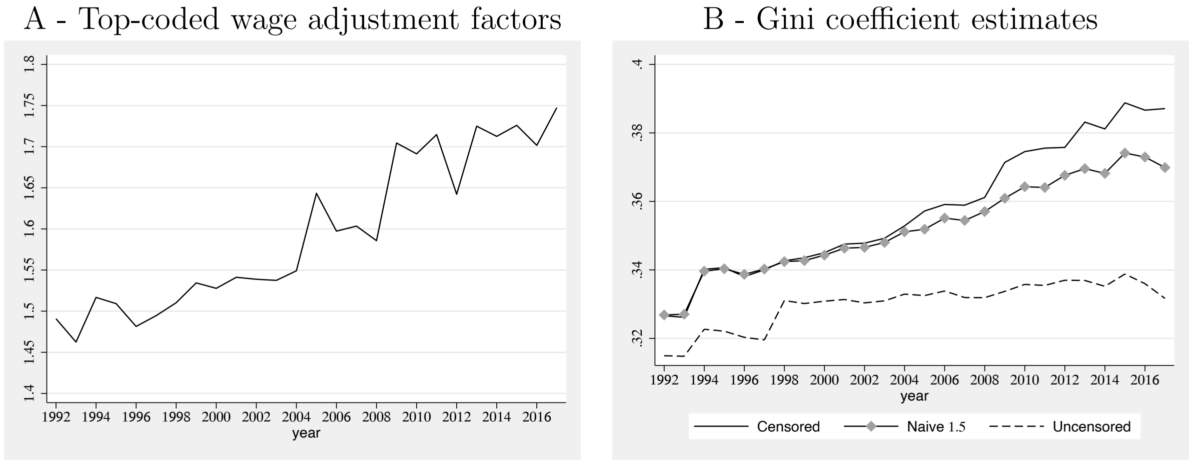


Figure 7: Prediction of top-coded weekly wages

When the proportion of wages above the top-code is relatively small (as for example, from 1992 to 2005), the difference between $\widehat{\tau}_1\left(y_c\right)$, $\widehat{\tau}_2\left(y_c\right)$ and $\widehat{\tau}_3\left(\mathbf{x}_i, y_c\right)$ is relatively small; however, as more wages are located in the top-coded category (as for example, in the years following 2007), the effect of censored data becomes stronger and the bias (underestimation) produced by the Hill estimator more pronounced ($\widehat{\tau}_1\left(y_c\right)$).

Figure 8 illustrates the time varying nature of the factor necessary to compute the imputed mean wages. Recall that to overcome the top-coding bias, in the literature, a constant value is frequently used to adjust the top-coded wages. For instance, Autor and Dorn (2013) consider an adjustment factor of 1.5 on a sample between 1950-2005; Acemoglu and Autor (2011), 1.5 on a sample between 1973-2009; Autor et al. (2008) 1.5 on a sample from 1963 to 2005; Katz and Murphy (1992) 1.45 on a sample between 1963 and 1987; Lemieux (2006) 1.4 on a sample from 1973 to 2003; and Beaudry et al. (2016) 1.4 on a sample from 1979 to 2011. Figure 8 shows that using a fixed value may have been adequate for pre-1992 data, but that the adjustment factor has increased over time reaching an overall value around 1.75 in 2017.



A - Top-coded wage adjustment factors

B - Gini coefficient estimates

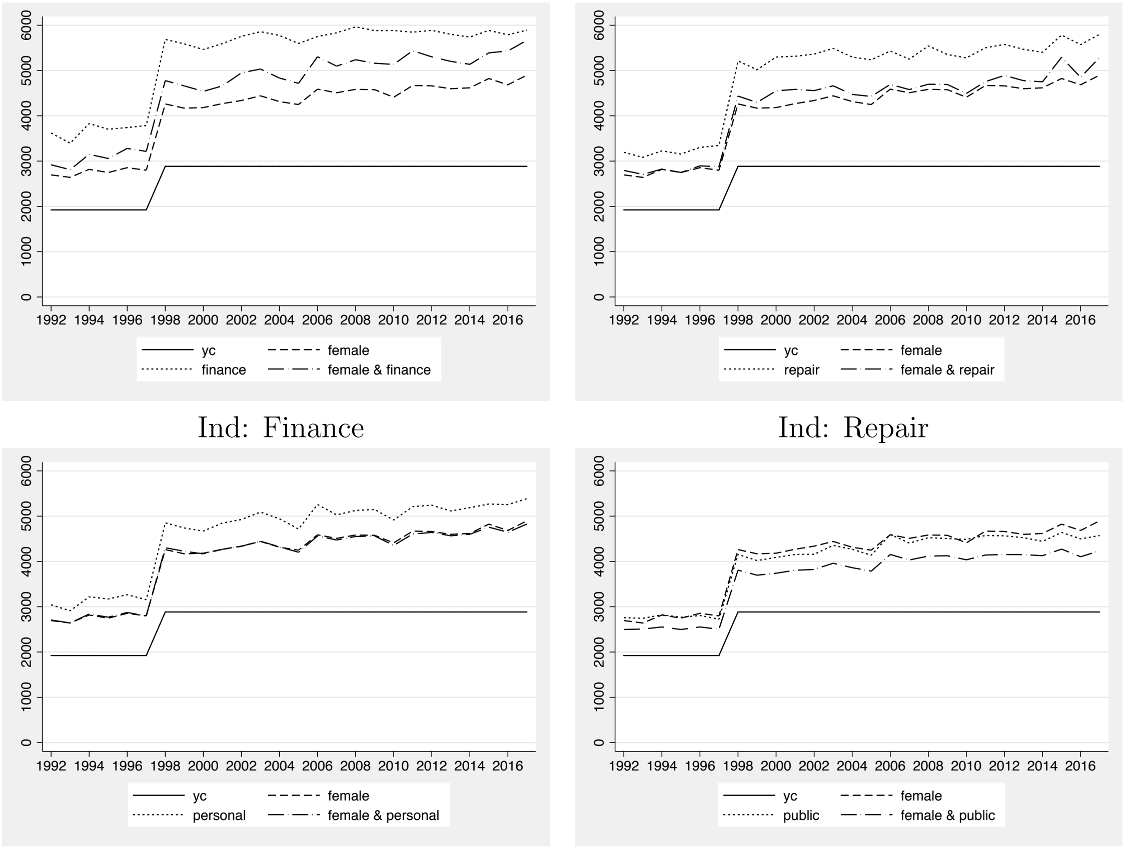
Figure 8: Top-coded weekly wage adjustment factors between 1992 and 2017 (left-plot) and corresponding Gini coefficient estimates for the overall weekly wage distribution (right-plot) from 1992 to 2017

Figure 8 also plots the Gini coefficient for the overall distribution which is computed based on a sample in which the top-coded values have been imputed using the approach developed in this paper, in order to illustrate the importance of the adjustment factors proposed and the use of the censored estimates. From this Figure we observe that the 1.5 adjustment factor reports a Gini coefficient which is aligned with the one computed based on our procedure from 1992 to 2000, but under evaluates inequality from 2000 onward. We observe that this undervaluation is around 5% in 2017.

4.3.2 Imputed wages above the top-code by gender and industry

Figure 9 illustrates the difference of the imputed mean values for individuals in the Finance, Repair, Personal and Public industries, as well as for women and women working in those industries. The purpose of these graphs is to further highlight the importance of allowing for different scaling factors depending on individuals characteristics and industry, but other graphs considering other characteristics can be plotted using our approach.

The first noticeable result is that the imputed wage of individuals decreases when we compare the wages for individuals in the Finance, Repair, Personal and Public industries. Finance displays the largest and Public the lowest imputed wages of the four industries. With the exception of the Public industry, women's imputed wages are lower for the other three industries and this observation also holds when we condition women's imputed wages on the industry they are in. A further interesting result is that the imputed mean wages display an increasing trend over time in all industries, for women and for women in those industries, which is an indication that the adjustment factors used to compute the imputed wages also changes over time.



Ind: Finance

Ind: Repair

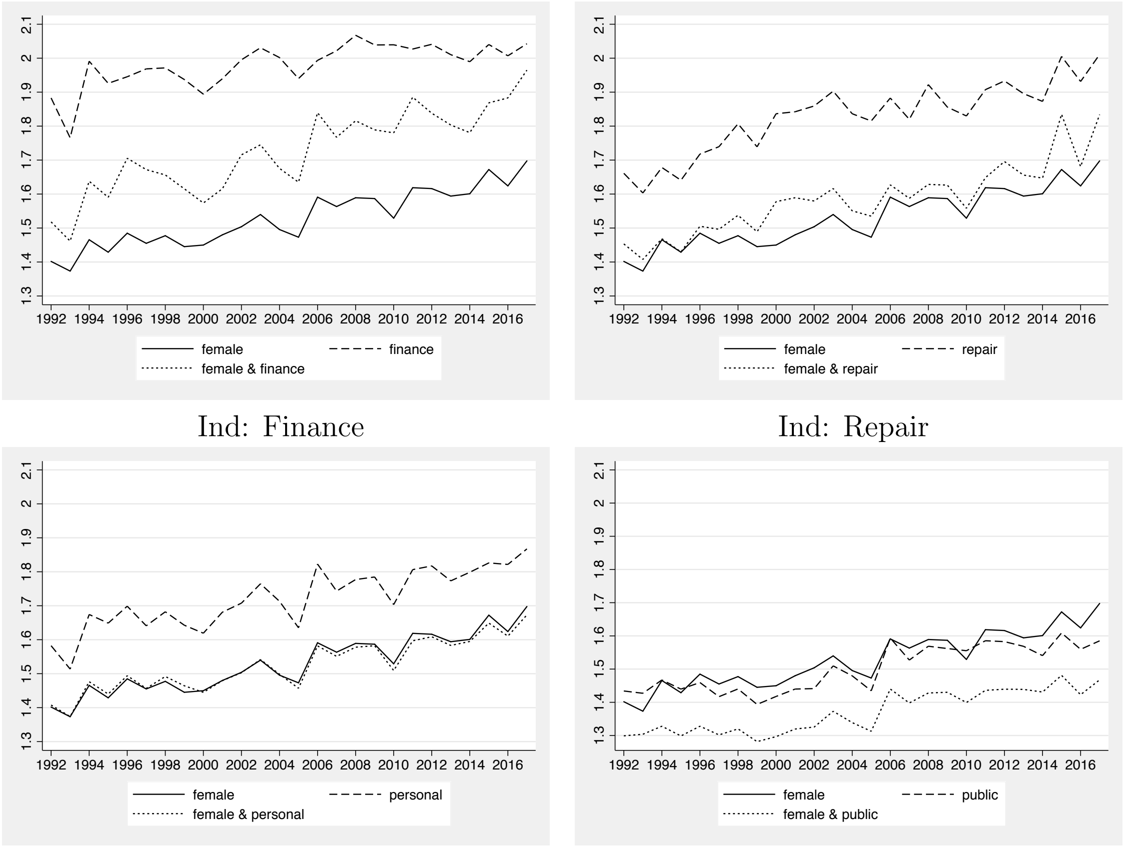
Ind: Personal

Ind: Public

Figure 9: Imputed mean wages between 1992 and 2017 by industry

This is further highlighted in Figure 10, where the graphs show the top-coded wage adjustment factors between 1992 and 2017, for different combinations of women working in different industries. Women's top-coded adjustment factor is always smaller than the top-coded adjustment factor that would be applied to males in any industry with the exception of public. This implies that the public industry is the less heavy tailed. On top of that women working in the public industry earn less in the right tail than women in the right tail working in other industries.

While women working in personal and repair are not earning much more nor much less than in other industries we find that women working in finance would need a much higher adjustment factor. This means that this is the industry in which they have been earning more and this result is reinforced with a clear positive trend between 1992 and 2017.



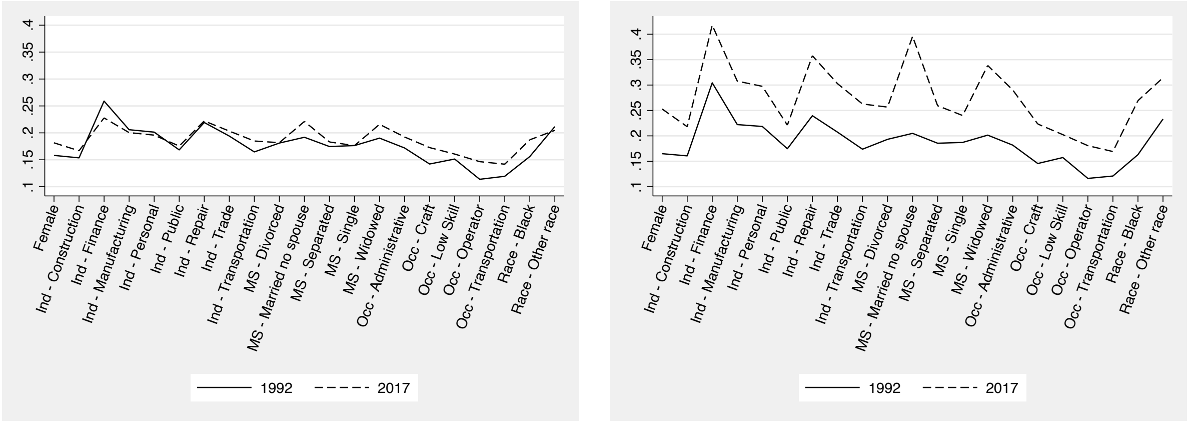
Ind: Finance

Ind: Repair

Ind: Personal

Ind: Public

Figure 10: Adjustment factors between 1992 and 2017 by industry to impute top-coded wages



Uncensored

Censored

Figure 11: Right-tail Gini coefficients in 1992 and in 2017 by individual, occupation and industry

Figure 11 shows the right-tail Gini coefficients in 1992 and in 2017 using both the uncensored and the censored estimators. The censored right-tail estimator shows that inequality is higher for all specific characteristics of the individuals in 1992 and in 2017. Moreover, inequality increased more between 1992 and 2017 than we would consider if we were to neglect the fact that the sample is censored.

For some characteristics of the individuals inequality rose more than what was suggested by the uncensored estimates. For instance, for female the right-tail Gini coefficient rose less than .05 when the uncensored estimates are considered while when we consider the censored right-tail index estimates the change is larger than .10. For other characteristics, assuming censored or uncensored samples in the estimation provides opposing directions of change in the Gini coefficient. For instance, individuals working in the finance industry observed a small decrease in inequality (uncensored graph on the left) when in fact, when censoring is taken into account inequality increased for these individuals by .10 points in the Gini coefficient.

5 Conclusions

This paper provides three important contributions to the literature. The first corresponds to the introduction of a conditional tail index estimator which explicitly handles the top-coding problem and an in-depth evaluation of its finite sample performance and comparison with competing methods. The Monte Carlo simulation exercise shows that the method proposed to estimate the tail index performs well in terms of estimation of the tail index and when used in the imputation of wages above the top-code when the sample is censored, which is an intrinsic feature of the public-use CPS database.

Second, evidence is provided which shows that the factor values used to adjust the top-coded weekly wages have changed over time and across the characteristics of individuals, occupations and industries and an indication of suitable values is proposed. Interestingly, the empirical results show that the upper-tail inequality has increased since the 90s and has become more pronounced over the last 20 years. Our analysis showed that women working in finance would need a higher adjustment factor to impute the top-coded weekly wages, whereas for women in public industry the change in the adjustment factor required is relatively small. For instance, in the period between 1992 and 2017, the adjustment factor for women working in the finance industry changed from 1.5 in 1992 to 2 in 2017, whereas that for women in the public industry the change was only from 1.3 in 1992 to 1.45 in 2017. Moreover, we also observe that the adjustment factor used to impute top-coded wages should be adapted over time and across characteristics of the individuals.

Third, an in depth empirical analysis of the dynamics of the US weekly wage distribution's right tail using the public-use CPS database from 1992 to 2017 is provided. The application of the procedure to the CPS data reveals that individuals working in industries such as finance, construction and repair are the ones with the more extreme wages. Moreover, it is also observed that the biggest increases in the probability of observing an extreme wage between 1992 and 2017 (which corresponds to an increase in the Gini coefficient) was for individuals that are black, married without a spouse, or widowed. Older and more educated workers continued to have a significant probability of observing an extreme wage but there was no relevant change between 1992 and 2017. Women observed a positive increase in alpha between 1992 and 2017, i.e. a decrease in the probability of an extreme wage of -6% in 1992 to -1% in 2017, and a decrease in the Gini coefficient of 0.04 in 1992 and 0.05 in 2017.

In our analysis we also used our approach to impute mean wages in order to obtain a consistent series of uncensored weekly wages in order to compute the overall Gini coefficient. This analysis revealed that the Gini coefficient of the US weekly wages increased from around 0.32 in 1992 to close to 0.39 in 2017. Furthermore, contrasting the results of our approach with those of a conservative fixed adjustment factor of 1.5 (as used in the literature), our procedure indicates that inequality in 2017 is 5% larger than that suggested by the fixed adjustment factor.

Additionally, although our empirical analysis has focused on the right tail of the US weekly wage distribution, the approaches proposed in this paper are also useful in other contexts, as adequately handling top-coding is of importance in a variety of other applications. For instance, it can prove helpful in the analysis of returns to education (Hubbard, 2011) and in the study of differences in gender and race (Burkhauser and Larrimore, 2009). Moreover, it could also be of value for welfare analysis of government policies such as tax cuts/increases (Hendren and Sprung-Keyser, 2020).

Notes

[1] The result holds for male and female samples separately, considering weekly wages of full-time workers as well as for the March CPS samples (Autor et al., 2006).

[2] The Gini coefficient takes values between 0 and 1, where 0 refers to perfect equality and 1 to the most extreme case of inequality. Hence, the lower the value of the Gini coefficient the more equal a society is, but a Gini above 0.5 is already an indication of considerable inequality.

[3] This distribution was used, for instance, by Harrison (1981) to analyze earnings by size in the UK.

[4] An observation is said to be right censored at y_c if the exact value of the observation is not known except that it is greater than or equal to y_c.

[5] If only a portion of the sample to estimate the model is used, say for example, all observations larger than y_0, then y is the quantile of order $1-u$ of the conditional distribution $P\left(Y<y \middle| Y>y_0\right)$, i.e. $P\left(Y<y \middle| Y>y_0\right)=1-u$. To determine the quantile order of the unconditional distribution $P\left(Y<y\right)$ we use the relation $P\left(Y<y \middle| Y>y_0\right)=1-u \Rightarrow P\left(Y<y\right)=P\left(Y<y_0\right)+\left(1-u\right)P\left(Y>y_0\right)$. In the empirical analysis we use all observations larger than the empirical quantile of 0.80 (see also Mishel et al., 2013), so that $P\left(Y>y_0\right)=0.20$ in the above formula. Hence, for $u=0.15$ or $u=0.20$, we are actually analyzing the 96th and 97th quantile, respectively, of the unconditional distribution.

[6] When the covariate considered is discrete (e.g. a dummy variable) an adaptation of (2.14) leads to the following formula, which measures the impact of group $d=1$ over $d=0$, $\delta\left(u\right):=\left[\left(1-u\right)^{\frac{\alpha\left(\mathbf{x};d=1\right)}{\alpha\left(\mathbf{x};d=0\right)}-1}-1\right]\times 100$. Given that $\alpha\left(\mathbf{x};d=1\right)$ and $\alpha\left(\ \mathbf{x};d=0\right)$ depend on $\mathbf{x}$, we also need to provide values for $\mathbf{x}$. One possible solution is to replace $\mathbf{x}$ by its respective averages.

[7] The cumulative Burr distribution function considered in the simulations is $F(x):=1-(1+x^{-\alpha\rho})^{\frac{1}{\rho}}$ and the corresponding probability density function $f(x):=x^{-1-\alpha\rho}(1+x^{-\alpha\rho})^{-1\frac{1}{\rho}}\alpha$.

[8] Wages are converted to 2012 values using the GDP personal consumption expenditure deflator.

[9] Workers in our sample come from outgoing rotation groups 4 and 8 and according to Unicon: When the Outgoing Rotation files are produced, two rotations are extracted from each of the twelve months and gathered into a single annual file. The weights on the file must be modified by the user before giving reliable counts. Since the final weight is gathered from 12 months but only 2/8 rotations, the weight on the outgoing file should be divided by 3 (12/4) before it is applied. The earner weight is gathered from

12 months from the 2 rotations. Since those two rotations were originally weighted to give a full sample, the earner weight must be divided by 12, not 3.

[10] Note that the Gini coefficient that we present is a measure of upper tail inequality and not a measure of inequality over the whole wages distribution as is usually computed. For this reason, these results are not directly comparable with the Gini estimates for the whole distribution found in the literature.

[11] Other values of c in the neighborhood of $c = 1.5$ yield basically the same results.

References

Acemoglu, D. and Autor, D. (2011). *Skills, tasks and technologies: Implications for employment and earnings*, volume 4 of *Handbook of Labor Economics*, chapter 12, pages 1043–1171. Elsevier.

Armour, P., Burkhauser, R. V., and Larrimore, J. (2016). Using the *p*areto distribution to improve estimates of top-coded earnings. *Economic Inquiry*, 54(2):1263–1273.

Autor, D. and Dorn, D. (2013). The growth of low-skill service jobs and the polarization of the US labor market. *The American Economic Review*, 103(5):1553–1597.

Autor, D., Katz, L., and Kearney, M. (2006). The polarization of the US labor market. *The American Economic Review*, 96(2):189–194.

Autor, D., Katz, L., and Kearney, M. (2008). Trends in US wage inequality: Revisioning the revisionists. *The Review of Economics and Statistics*, 90(2):300–323.

Autor, D. H. (2019). Work of the Past, Work of the Future. *AEA Papers and Proceedings*, 109:1–32.

Beaudry, P., Green, D. A., and Sand, B. M. (2016). The great reversal in the demand for skill and cognitive tasks. *Journal of Labor Economics*, 34(S1):S199–S247.

Bernstein, J. and Mishel, L. (1997). Has wage inequality stopped growing? *Monthly Labor Review*, pages 3–16.

Burkhauser, R. V., Feng, S., Jenkins, S. P., and Larrimore, J. (2012). Recent trends in top income shares in the United States: Reconciling estimates from march CPS and IRS tax return data. *The Review of Economics and Statistics*, 94(2):371–388.

Burkhauser, R. V. and Larrimore, J. (2009). Using internal CPS data to reevaluate trends in labor-earnings gaps. *Monthly Labor Review*, 132(8):3–18.

Feng, S., Burkhauser, R. V., and Butler, J. S. (2006). Levels and long-term trends in earnings inequality: Overcoming current population survey censoring problems using the GB2 distribution. *Journal of Business & Economic Statistics*, 24(1):57–62.

Goos, M. and Manning, A. (2007). Lousy and lovely jobs: The rising polarization of work in Britain. *The Review of Economics and Statistics*, 89(1):118–133.

Goos, M., Manning, A., and Salomons, A. (2014). Explaining job polarization: routine-biased technological change and offshoring. *The American Economic Review*, 104(8):2509–2526.

Hall, P. (1982). On some simple estimates of an exponent of regular variation. *Journal of the Royal Statistical Association (Series B)*, 44:37–42.

Harrison, A. (1981). Earnings by size: A tale of two distributions. *The Review of Economic Studies*, 48(4):621–631.

Hendren, N. and Sprung-Keyser, B. (2020). A Unified Welfare Analysis of Government Policies. *The Quarterly Journal of Economics*, 135(3):1209–1318.

Hill, B. (1975). A simple general approach to inference about the tail of a distribution. *The Annals of Statistics*, 3:1163–1174.

Hlasny, V. and Verme, P. (2018). Top incomes and inequality measurement: A comparative analysis of correction methods using the *eu silc* data. *Econometrics*, 6(2).

Hlasny, V. and Verme, P. (2021). The impact of top incomes biases on the measurement of inequality in the *u*nited *s*tates. *Oxford Bulletin of Economics and Statistics*, forthcoming.

Hubbard, W. (2011). The phantom gender difference in the college wage premium. *Journal of Human Resources*, 46(3):568–86.

Ibragimov, M. and Ibragimov, R. (2018). Heavy tails and upper-tail inequality: The case of russia. *Empirical Economics*, 54(2):823–837.

Jensen, S. and Shore, S. (2015). Changes in the distribution of earnings volatility. *Journal of Human Resources*, 3(50):811–836.

Katz, L. and Murphy, K. (1992). Changes in relative wages, $1963-1987$: Supply and demand factors. *The Quarterly Journal of Economics*, 107(1):35–78.

Larrimore, J., Burkhauser, R., Feng, S., and Zayatz, L. (2008). Consistent cell means for top coded incomes in the public-use march CPS (1976-2007). *Journal of Economic and Social Measurement*, 33(2-3):89–128.

Lemieux, T. (2006). Increasing residual wage inequality: Composition effects, noisy data, or rising demand for skill? *American Economic Review*, 96(3):461–498.

Levy, F. and Murnane, R. (1992). US earnings levels and earnings inequality: A review of recent trends and proposed explanations. *Journal of Economic Literature*, 30(3):1333–81.

Ma, Y., Jiang, Y., and Huang, W. (2019). Tail index varying coefficient model. *Communications in Statistics - Theory and Methods*, 48(2):235–256.

Macpherson, D. A. and Hirsch, B. T. (1995). Wages and gender composition: Why do women's jobs pay less? *Journal of Labor Economics*, 13(3):426–471.

Mishel, L., Bernstein, J., and Shierholz, H. (2013). *The State of Working America*. Ithaca, NY. Cornell University Press, 12th edition edition.

Nicolau, J. and Rodrigues, P. M. M. (2019). A new regression-based tail index estimator. *The Review of Economics and Statistics*, 101(4):667–680.

Piketty, T. and Saez, E. (2003). Income inequality in the United States, 19131998. *The Quarterly Journal of Economics*, 118(1):1–41.

Wang, H. and Tsai, C.-L. (2009). Tail index regression. *Journal of the American Statistical Association*, 104(487):1233–1240.

On-Line Supplementary Appendix:

Measuring wage inequality under right censoring

by

João Nicolau, Pedro Raposo and Paulo M. M. Rodrigues

S.1 Technical Appendix

Proof of Theorem 2.1

For the proof of Theorem 2.1, let us show first that $\boldsymbol{\Sigma}_{y_0}^{-1/2}\frac{\dot{\mathcal{K}}^c(\theta,y_c)}{n} \xrightarrow{p} 0$. Consider that the sequence $\{(y_i,\mathbf{x}_i)\}$ is independently distributed.

$$\begin{aligned}
\boldsymbol{\Sigma}_{y_0}^{-1/2}\dot{\mathcal{K}}\left(\theta,y_c\right) &= \sum_{t=1}^{n}\left\{I_{\left\{y_0\le w_i<y_c\right\}}\left[\exp\left(\mathbf{x}_i'\theta\right)\log\left(\frac{w_i}{y_0}\right)-1\right]\right. \\
&\quad \left. -I_{\{w_i=y_c\}}\exp\left(\mathbf{x}_i'\theta\right)\log\left(\frac{y_0}{y_c}\right)\right\}\boldsymbol{\Sigma}_{y_0}^{-1/2}\mathbf{x}_i \\
&= \sum_{t=1}^{n} e_i\mathbf{Z}_{ni}
\end{aligned}$$

where

$$e_i = \begin{cases} \exp\left(\mathbf{x}_i'\theta\right)\log\left(\frac{w_i}{y_0}\right)-1, \text{ for} & I_{\left\{y_0\le w_i<y_c\right\}} \\ -\exp\left(\mathbf{x}_i'\theta\right)\log\left(\frac{y_0}{y_c}\right), \text{ for} & I_{\{w_i=y_c\}} \end{cases} \tag{S.1}$$

Thus, considering (S.1) we cans how that,

$$\begin{aligned}
E\left(e_i|\,\mathbf{x}_i\right) &= \int_{y_0}^{y_c}\left(\exp\left(\mathbf{x}_i'\theta\right)\log\left(\frac{y}{y_0}\right)-1\right)f\left(y|\,\mathbf{x}_i;\theta\right)dw - \exp\left(\mathbf{x}_i'\theta\right)\log\left(\frac{y_0}{y_c}\right)P\left(I_{\{w_i=y_c\}}\right) \\
&= \int_{y_0}^{y_c}\left(\exp\left(\mathbf{x}_i'\theta\right)\log\left(\frac{y}{y_0}\right)-1\right)f\left(y|\,\mathbf{x};\theta\right)dw - \exp\left(\mathbf{x}_i'\theta\right)\log\left(\frac{y_0}{y_c}\right)\left(1-F\left(y_c|\,\mathbf{x}_i;\theta\right)\right) \\
&= \left(\frac{y_0}{y_c}\right)^{\exp\left(\mathbf{x}_i'\theta\right)}\exp\left(\mathbf{x}_i'\theta\right)\log\left(\frac{y_0}{y_c}\right) - \exp\left(\mathbf{x}_i'\theta\right)\log\left(\frac{y_0}{y_c}\right)\left(\frac{y_0}{y_c}\right)^{\exp\left(\mathbf{x}_i'\theta\right)} \\
&= 0
\end{aligned}$$

Moreover, it follows that

$$\begin{aligned}
E\left(e_i^2|\,\mathbf{x}_i\right) &= \int_{y_0}^{y_c}\left(\exp\left(\mathbf{x}_i'\theta\right)\log\left(\frac{y}{y_0}\right)-1\right)f\left(y|\,\mathbf{x}_i;\theta\right)dw - \exp\left(\mathbf{x}_i'\theta\right)\log\left(\frac{y_0}{y_c}\right)P\left(I_{\{w_i=y_c\}}\right) \\
&= 1-\left(\frac{y_0}{y_c}\right)^{\exp\left(\mathbf{x}_i'\theta\right)} := \Lambda.
\end{aligned}$$

From Chebychev's weak law of large numbers, we have that

$$\boldsymbol{\Sigma}_{y_0}^{-1/2}\frac{\dot{\mathcal{K}}^c\left(\theta,y_c\right)}{n} = \frac{1}{n}\sum_{t=1}^{n}e_i\mathbf{Z}_{ni} \xrightarrow{p} E\left(e_i\mathbf{Z}_{ni}\right)$$

and $E\left(e_i\mathbf{Z}_{ni}\right)$ is zero since

$$E\left(e_i\mathbf{Z}_{ni}\right)=E\left(E\left(e_i\mathbf{Z}_{ni}\right|\mathbf{Z}_{ni}\right))=E\left(E\left(e_i\right|\mathbf{Z}_{ni}\right)\mathbf{Z}_{ni})=0.$$

Given these results, considering that $\hat{\boldsymbol{\theta}}$ minimizes the log-likelihood function, $\frac{\mathcal{K}^c(\theta,y_c)}{n}$ such that $\frac{\dot{\mathcal{K}}^c(\theta,y_c)}{n}=0$, using the Mean Value Theorem it follows that

$$\frac{\dot{\mathcal{K}}^c\left(\theta,y_c\right)}{n}=\frac{\dot{\mathcal{K}}^c\left(\theta_0,y_c\right)}{n}+\frac{\ddot{\mathcal{K}}^c\left(\theta_1,y_c\right)}{n}\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\right) \tag{S.2}$$

for some $\boldsymbol{\theta}_1\in[\hat{\boldsymbol{\theta}},\boldsymbol{\theta}_0]$. ■

Computation of the partial effects

We have

$$\begin{aligned}\delta &= \frac{\bar{F}\left(y\right|\Delta x+x)-\bar{F}\left(y\right|x)}{\bar{F}\left(y\right|x)}\times 100\\ &= \frac{\left(\frac{y}{y_0}\right)^{\alpha(\Delta x+x)}-\left(\frac{y}{y_0}\right)^{\alpha(x)}}{\left(\frac{y}{y_0}\right)^{\alpha(x)}}\times 100\\ &= \left[\left(\frac{y}{y_0}\right)^{\alpha(\Delta x+x)-\alpha(x)}-1\right]\times 100 \end{aligned}\tag{S.3}$$

since $y=(1-u)^{\frac{1}{\alpha(x)}y_0}$, it follows that,

$$\delta=\left((1-u)^{\frac{\alpha(\Delta x+x)}{\alpha(x)}-1}-1\right)\times 100. \tag{S.4}$$

Now and given the specification $\alpha\left(x\right)=\exp\left(\phi\left(x\right)\right)$ where $\phi\left(x\right)$ is $\phi\left(x\right)=\mathbf{x}'\beta$ it follows that,

$$\alpha\left(\Delta x+x\right)\simeq\alpha\left(x\right)+\frac{d\alpha\left(x\right)}{dx}\Delta x=\alpha\left(x\right)+\phi'\left(x\right)\exp\left(\phi\left(x\right)\right)\Delta x.$$

Therefore

$$\frac{\alpha\left(\Delta x+x\right)}{\alpha\left(x\right)}=\frac{\alpha\left(x\right)+\phi'\left(x\right)\exp\left(\phi\left(x\right)\right)\Delta x}{\alpha\left(x\right)}=1+\phi'\left(x\right)\Delta x.$$

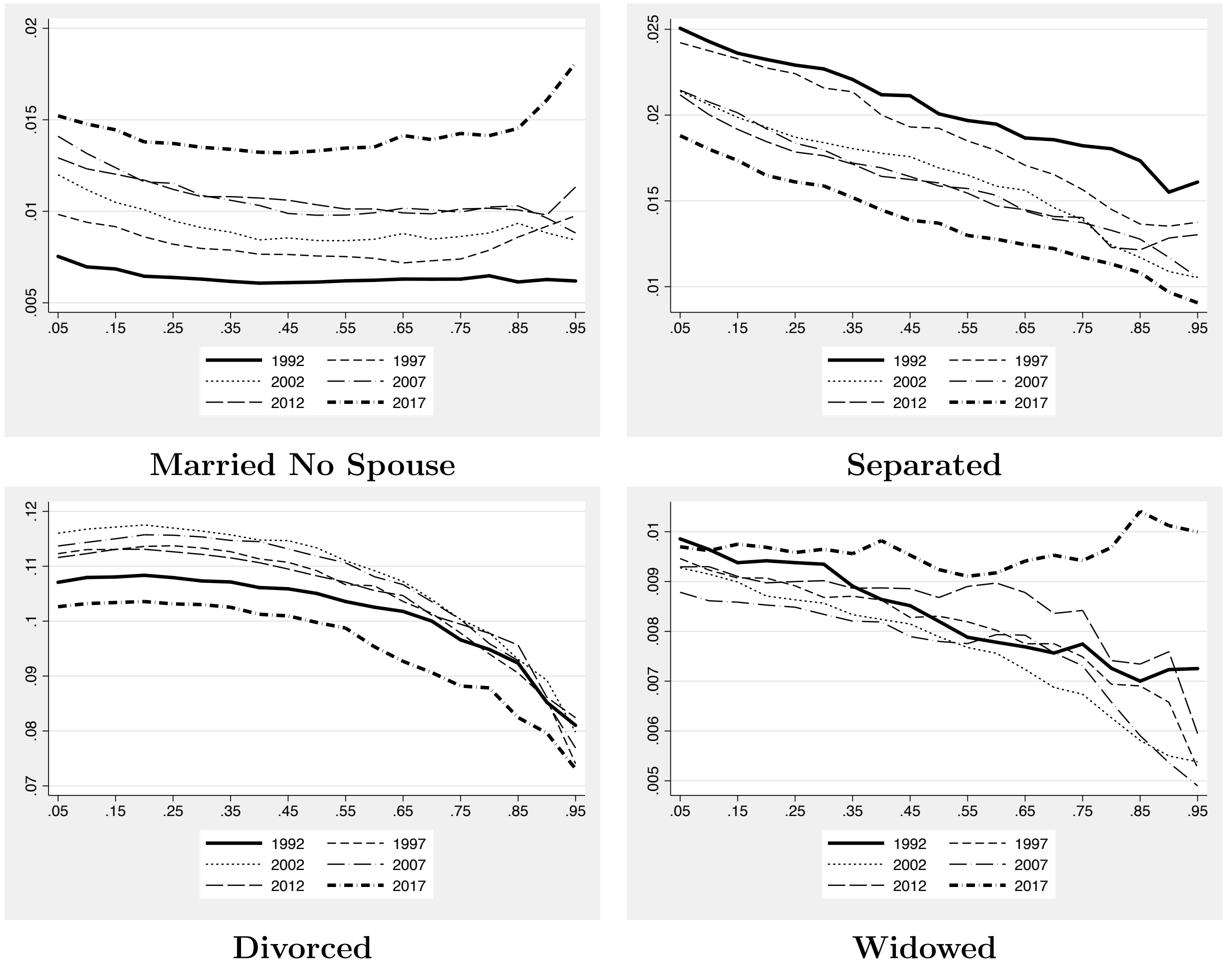
In conclusion

$$\delta=\frac{\bar{F}\left(y\right|\Delta x+x)-\bar{F}\left(y\right|x)}{\bar{F}\left(y\right|x)}\times 100=\left((1-u)^{\phi'(x)\Delta x}-1\right)\times 100.$$

■

S.2 Additional Figures

Figure S.1: Proportion of individuals according to Marital Status by wage percentile



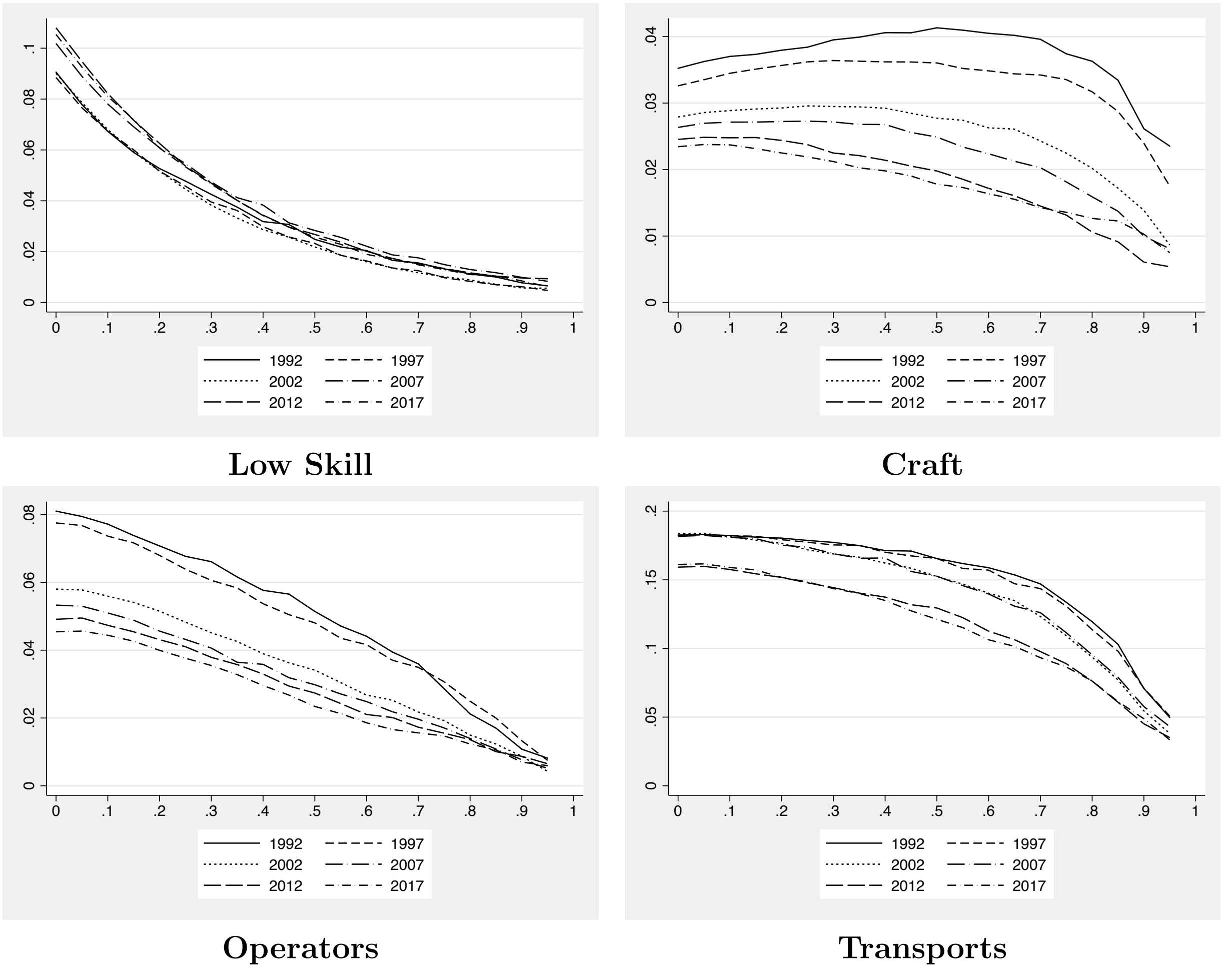
Married No Spouse

Separated

Divorced

Widowed

Figure S.2: Proportion of individuals according to Occupation by wage percentile



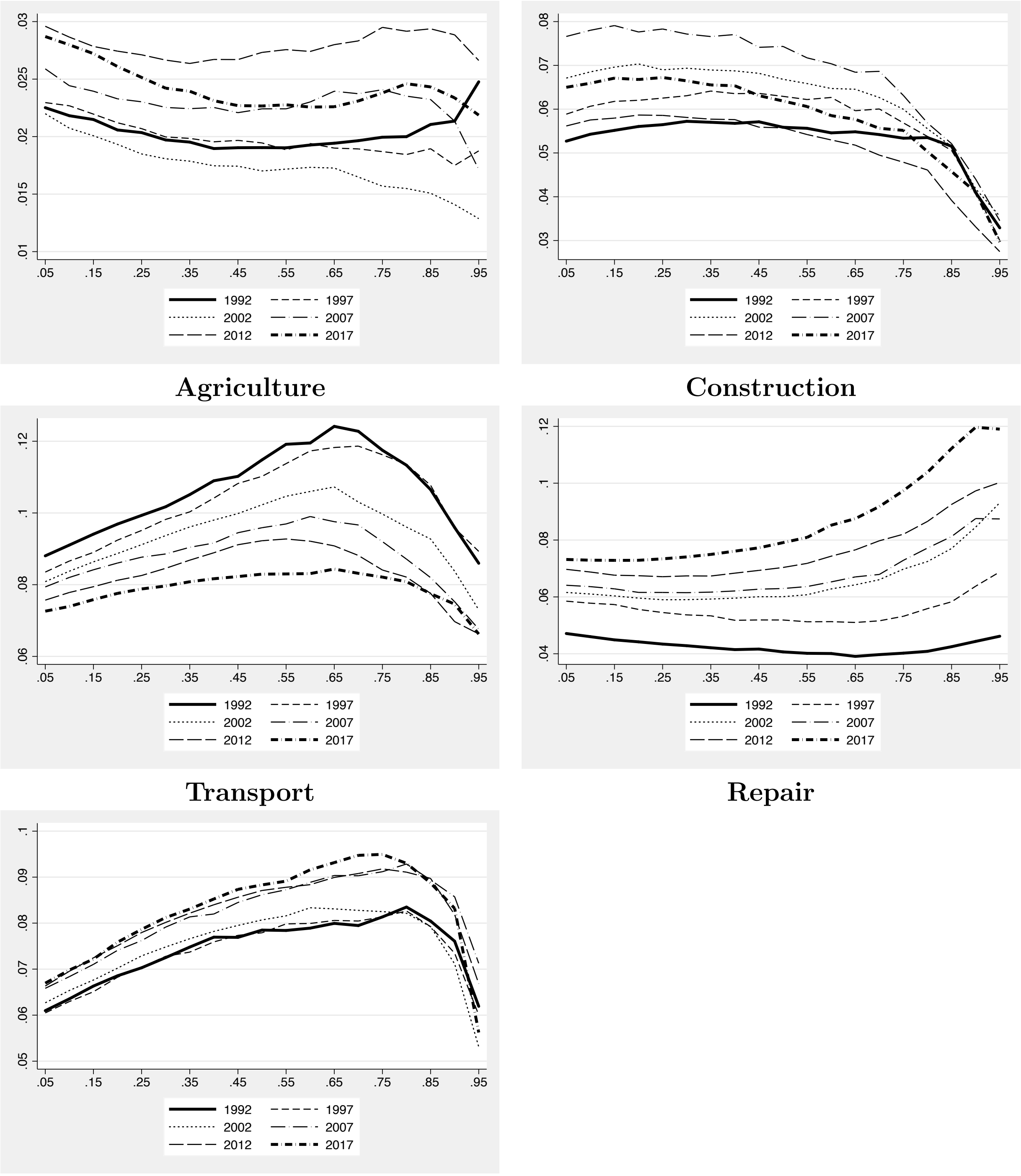
Low Skill

Craft

Operators

Transports

Figure S.3: Proportion of individuals according to Industry by wage percentile



Agriculture

Construction

Transport

Repair

Public

S.3 Additional Tables

Table S1: Bias and RMSE of estimators

Case 1: DGP Pareto (k=0.20 and $\mathbf{y_c} = \mathbf{Q_y(0.95)}$)												
	$\hat{\alpha}(\mathbf{x}_i)$	$\hat{\alpha}^c(\mathbf{x}_i)$	$\tilde{\alpha}(\mathbf{x}_i)$	$\hat{\alpha}(\mathbf{x}_i)$	$\hat{\alpha}^c(\mathbf{x}_i)$	$\tilde{\alpha}(\mathbf{x}_i)$	$\hat{\alpha}(\mathbf{x}_i)$	$\hat{\alpha}^c(\mathbf{x}_i)$	$\tilde{\alpha}(\mathbf{x}_i)$	$\hat{\alpha}(\mathbf{x}_i)$	$\hat{\alpha}^c(\mathbf{x}_i)$	$\tilde{\alpha}(\mathbf{x}_i)$
n		2500			5000			10000			50000	
bias(β_1)	0.0025	0.0004	0.4553	0.0011	−0.0002	0.4539	0.0004	−0.0005	0.4528	0.0001	−0.0002	0.4526
bias(β_2)	0.0029	0.0024	−0.4241	0.0027	0.0034	−0.4237	0.0014	0.0020	−0.4246	0.0003	0.0007	−0.4257
RMSE(β_1)	0.0775	0.0937	0.4611	0.0552	0.0660	0.4567	0.0386	0.0465	0.4542	0.0174	0.0208	0.4529
RMSE(β_2)	0.1703	0.1942	0.4422	0.1208	0.1369	0.4329	0.0852	0.0963	0.4292	0.0378	0.0430	0.4267
Case 2: DGP Pareto (k=0.20 and $\mathbf{y_c} = \mathbf{Q_y(0.99)}$)												
n		2500			5000			10000			50000	
bias(β_1)	0.0025	−0.0012	0.1000	0.0011	−0.0010	0.0984	0.0004	−0.0005	0.0980	0.0001	−0.0001	0.0977
bias(β_2)	0.0029	0.0070	−0.1202	0.0027	0.0050	−0.1202	0.0014	0.0023	−0.1219	0.0003	0.0005	−0.1229
RMSE(β_1)	0.0775	0.0807	0.1258	0.0552	0.0575	0.1126	0.0386	0.0401	0.1051	0.0174	0.0182	0.0992
RMSE(β_2)	0.1703	0.1745	0.2006	0.1208	0.1240	0.1659	0.0852	0.0871	0.1460	0.0378	0.0389	0.1281
Case 3: DGP Burr ρ=-2 (k=0.05 and $\mathbf{y_c} = \mathbf{Q_y(0.99)}$)												
n		2500			5000			10000			50000	
bias(β_1)	0.0015	−0.0118	0.3277	−0.0006	−0.0084	0.3241	−0.0025	−0.0065	0.3229	−0.0042	−0.0061	0.3210
bias(β_2)	0.0388	0.0488	−0.3189	0.0186	0.0242	−0.3356	0.0134	0.0161	−0.3409	0.0073	0.0097	−0.3452
RMSE(β_1)	0.1450	0.1683	0.3562	0.1021	0.1180	0.3387	0.0719	0.0821	0.3301	0.0323	0.0372	0.3224
RMSE(β_2)	0.4088	0.4518	0.4549	0.2833	0.3121	0.4043	0.1991	0.2176	0.3754	0.0891	0.0976	0.3523
Case 4: DGP Burr ρ=-2 (k=0.10 and $\mathbf{y_c} = \mathbf{Q_y(0.95)}$)												
n		2500			5000			10000			50000	
bias(β_1)	−0.0099	−0.0239	0.9137	−0.0120	−0.0258	0.9098	−0.0129	−0.0259	0.9079	−0.0137	−0.0259	0.9071
bias(β_2)	0.0297	0.0380	−0.6485	0.0250	0.0402	−0.6489	0.0216	0.0372	−0.6512	0.0200	0.0371	−0.6519
RMSE(β_1)	0.1055	0.1596	0.9196	0.0754	0.1129	0.9127	0.0537	0.0820	0.9094	0.0271	0.0434	0.9074
RMSE(β_2)	0.2609	0.3563	0.6637	0.1835	0.2492	0.6561	0.1305	0.1773	0.6548	0.0602	0.0857	0.6527
Case 5: DGP Burr ρ=-2 (k=0.20 and $\mathbf{y_c} = \mathbf{Q_y(0.99)}$)												
n		2500			5000			10000			50000	
bias(β_1)	−0.0364	−0.0435	0.0602	−0.0378	−0.0433	0.0586	−0.0385	−0.0428	0.0582	−0.0386	−0.0421	0.0581
bias(β_2)	0.0487	0.0578	−0.0733	0.0486	0.0558	−0.0732	0.0474	0.0531	−0.0748	0.0460	0.0510	−0.0762
RMSE(β_1)	0.0847	0.0909	0.0966	0.0665	0.0716	0.0798	0.0542	0.0584	0.0693	0.0422	0.0458	0.0606
RMSE(β_2)	0.1737	0.1807	0.1739	0.1283	0.1342	0.1344	0.0961	0.1009	0.1089	0.0592	0.0638	0.0840
Case 6: DGP Burr ρ=-2 (k=0.20 and $\mathbf{y_c} = \mathbf{Q_y(0.95)}$)												
n		2500			5000			10000			50000	
bias(β_1)	−0.0364	−0.0551	0.4134	−0.0378	−0.0557	0.4119	−0.0385	−0.0559	0.4108	−0.0386	−0.0553	0.4109
bias(β_2)	0.0487	0.0702	−0.3813	0.0486	0.0713	−0.3808	0.0474	0.0699	−0.3817	0.0460	0.0682	−0.3831
RMSE(β_1)	0.0847	0.1084	0.4196	0.0665	0.0864	0.4150	0.0542	0.0727	0.4124	0.0422	0.0591	0.4112
RMSE(β_2)	0.1737	0.2040	0.4010	0.1283	0.1532	0.3908	0.0961	0.1182	0.3866	0.0592	0.0804	0.3841

Note: $\hat{\alpha}(\mathbf{x}_i)$ is the tail index regression estimate considering the complete sample of data (with no censoring); $\hat{\alpha}^c(\mathbf{x}_i)$ is the censored tail index regression estimate; and $\tilde{\alpha}(\mathbf{x}_i)$ is the uncensored tail index regression estimate computed from censored data. n corresponds to the total sample size, k to the % of observations used for the tail index estimation and $\lfloor kn \rfloor$ is the effective number of observations used in the estimation of the tail index. y_c is the censoring value used and $Q_y(\tau)$ corresponds to the τ^{th} quantile of y.

Table S2: Industry classification

Agriculture	agriculture, forestry, fishing, hunting, mining and utilities;
Construction	only construction;
Manufacturing	manufacturing of non-durable and durable goods and wood;
Transports	transportation, warehousing, utilities electric light;
Trade	wholesale and retail trade;
Finance	finance and insurance;
Repair	business and repair;
Personal	personal services, entertainment and recreation, professional and related services;
Public	public administration and armed forces.

Note: This classification is based on the harmonized variable ”ind1990” from IPUMS.