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

July 30, 2021

Abstract

This paper investigates potential changes which may have occurred, over the last two decades, in the probability mass of the right tail of the wage distribution, through the analysis of the corresponding tail index. 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. Ignoring the top-coding may lead to inconsistent estimates of the tail index and to under or over statements of inequality and of its evolution over time. The contribution of this paper is threefold: i) We introduce a conditional tail index estimator that explicitly handles the top-coding problem, and evaluate its finite sample performance and compare it with competing methods. Based on this estimator we observe that inequality increased more than what was previously described in the literature; ii) We show that the factor values used to adjust the top-coded wages have changed over time and depend on the characteristics of individuals, 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); and iii) an in-depth empirical analysis of the dynamics of the US wage distribution's right tail using the public-use CPS database from 1992 to 2017 is provided. It is shown that the probability of observing an extreme value increased in 2017 by 4.94% for an individual in the Finance industry, which is an indication of an increase in concentration in the right-tail of the wage distribution in this industry (an increase, although of smaller magnitude, of the partial effects is also observed for all other industries except for Public sector which is still negative).

Keywords: Wage inequality, tail index, top-coding, current population survey, wage distribution, Pareto, occupations

JEL classification: C18, C24, E24, J11, J31

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 stopped growing after 1987 upper-tail inequality kept rising. The deceleration in inequality growth observed in the 1990s resulted mainly from polarization, i.e., from an abrupt stop or reversal of inequality growth in the lower-tail coupled with a sustained secular rise of the upper-tail inequality. 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.

Empirical evidence suggests that the rise in wage inequality is largely explained by shifts in the supply and demand for skills (Goos and Manning, 2007), and by the erosion of labour market institutions, e.g. unions and minimum wage (Kalleberg, 2011). It is documented that the increase in inequality in the 1980s was the result of a secular rise in the demand for skill which faced an abrupt slowdown in the relative supply of high-skilled workers (college or equivalent) in the form of lower attainment and of smaller labor-entering cohorts which originated expanding wage differentials (Autor et al., 2008; Katz and Murphy, 1992; Card and DiNardo, 2002; and Acemoglu and Autor, 2011).

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). 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).

The 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

1

et al., 2006). In addition, top-coding has changed over time. For instance, the top-coded wage was set at \$1923 in 1997 and changed to \$2884 from 1998 onward. But even during periods of constant nominal top-coding the data may hide changes in inequality (Levy and Murnane, 1992).

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 wage inequality than other available measures. A conditional tail index regression specifically designed to account for right censoring is used. The tail index is an important indicator, as it can be interpreted as an inverse measure of concentration of top wages. The lower the value of the index the more concentrated the distribution is.

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, and 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 wages changed over time and across the characteristics of individuals, occupations and industries; moreover it is also shown using the new estimator that the tail index has been decreasing since 1992 suggesting increased concentration in the right tail.

The contribution of this paper is threefold: i) we introduce a conditional tail index estimator that explicitly handles the top-coding problem, evaluate its finite sample performance and compare it to competing methods; ii) we show that the factor values used for the adjustment of top-coded wages has change over time and across the characteristics of individuals, occupations and industries, and suitable values are proposed; and iii) we provide an in-depth analysis of the dynamics of the US wage distribution's right tail using the public-use CPS database from 1992 to 2017.

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 wages using non-censored data; Section 4 describes and discusses the results of the right tail characteristics of the wage distribution and 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 proofs of the results put forward throughout the paper.

2   Methodology

To reduce the top-coding bias, researchers typically impute top-code values to create a consistent series. Until recently one of four approaches has in general 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 in estimating the tail index of a censored Pareto distribution. To briefly illustrate the procedure consider the survival function, $\overline{F}$, of a Pareto distribution[2] $\overline{F}(y) := P(Y_i > y) = \left(\frac{y_0}{y_i}\right)^\alpha$, where $y_i \geq y_0 > 0$, $\alpha > 0$ and the corresponding density function, $f_Y(y) = (\alpha y_0^\alpha)/(y_i^{\alpha+1})$. A large number of tail index estimation procedures is available in the literature. One widely used approach is the conditional maximum likelihood estimator

(MLE) proposed by Hill (1975),

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

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

However, recognizing the limitations of this approach when the data is top-coded, Armour et al. (2016) proposed an alternative method, which consists of an adaptation of the Hill estimator taking into consideration the censoring. This approach provides an unbiased estimate of the censored Pareto parameter, α, while using all available information. In specific, in the case of a censored sample the outcome variable is,

$$\omega_i = \begin{cases} y_i & if \quad y_0 \leq y_i < y_c \\ y_c & if \quad\quad y_i \geq y_c \end{cases}, \tag{2.2}$$

where y_0 is the tail cut off point and y_c the top-coded value. Hence, the density function of the censored Pareto distribution is,

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

and the respective log-likelihood function,

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

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

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

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, a new approach, also designed to overcome the top-coding bias, is proposed. In specific, a conditional tail index estimator which explicitly takes the right censoring of the data into account and uses covariates in the estimation process is introduced. Correctly estimating this tail index is of importance as it is used, for instance, for the imputation of wages above the top-code. Furthermore, the procedure has the additional advantage of allowing for an in-depth analysis of the determinants that impact the tail index the strongest according to the characteristics of individuals, occupations and industries and whether these impacts have changed over time. The use of different scaling factors to impute wages depending on the different categorizations of individuals has been used previously in the literature, see e.g., Macpherson and Hirsch (1995) who allow the scaling factors to vary according to gender and over the years.

To introduce the conditional tail index estimation approach we 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 survival function (under no censoring) 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 $y \to \infty$. Specifically, 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

5

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 $\dot{\mathcal{L}}(y; \mathbf{x}) \to 0$, where $\dot{\mathcal{L}}(y; \mathbf{x}) = \partial \mathcal{L}(y; \mathbf{x})/\partial y$.

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[3] 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,

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

where $I(.)$ is the indicator function and $f(.|\mathbf{x})$ and $F(.|\mathbf{x})$ correspond to the conditional Pareto density function and the conditional cumulative Pareto distribution function, respectively. Hence, 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 defined in (2.10), w_i is given in (2.2) and y_c is the censoring threshold.

Moreover, since

$$log\left[g(\omega_i|\mathbf{x}_i, y_c, \boldsymbol{\theta})\right] = I(y_0 \le w_i < y_c)\log f(\omega_i|\mathbf{x}_i, y_c, \theta) + I(w_i \ge y_c)\log\left[1 - F(y_c|\mathbf{x}_i, y_c, \theta)\right]$$

$$= I_{\{y_0 \le w_i < y_c\}}\{\log \alpha\left(\mathbf{x}_i\right) + \alpha\left(\mathbf{x}_i\right)\log y_0 - \left[\alpha\left(\mathbf{x}_i\right) + 1\right]\log w_i\}$$

$$+ I_{\{w_i = y_c\}}\{\alpha\left(\mathbf{x}_i\right)\left[\log\left(y_0\right) - \log(y_c)\right]\}, \tag{2.12}$$

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

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

Hence, we see from this approximate log-likelihood function that censuring 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:

(A1) $n_0^{-1} \sum_{i=1}^n \mathbf{Z}_{ni}\mathbf{Z}_{ni}'I\left(w_i \ge y_0\right) = \boldsymbol{\Sigma}_{y_0}^{-1/2}\widehat{\boldsymbol{\Sigma}}_{y_0}\boldsymbol{\Sigma}_{y_0}^{-1/2} \overset{p}{\to} \mathbf{I}_p$, where $\mathbf{Z}_{ni} := \boldsymbol{\Sigma}_{y_0}^{-1/2}\mathbf{x}_i$, $\mathbf{I}_p$ is a $p \times p$ identity matrix and $\widehat{\boldsymbol{\Sigma}}_{y_0} := n_0^{-1}\sum(\mathbf{x}_i\mathbf{x}_i')I(w_i \ge y_0)$.

(A2) (*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 $y \to \infty$.

(A3) (*Convergence rate*) Assume that $n_0 \to \infty$ and $(\frac{n}{n_0})E\{\mathbf{Z}_{ni}c_1(\mathbf{x}_i)\frac{\beta(\mathbf{x}_i)}{\alpha(\mathbf{x}_i)+\beta(\mathbf{x}_i)}y_0^{-\alpha(\mathbf{x}_i)-\beta(\mathbf{x}_i)}\} \to \mathbf{h}$, for some nonzero constant vector $\mathbf{h} \in R^p$.

Note that Assumption (A3) implicitly specifies the optimal convergence rate of y_0. A rate faster than that in (A3) mitigates the asymptotic bias of the parameter estimates but increases its variability. In contrast, a rate slower than that in (A3) reduces variability but increases bias; see Wang and Tsai (2009) for details.

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

Theorem 2.1 *Under Assumptions (A1) - (A3) it follows that* $n^{-1/2}\Sigma_{y_0}^{-1/2}\Lambda^{-1/2}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \overset{d}{\to} N(\mathbf{0}, \mathbf{I}_p)$, *where* $\Lambda := E(e_i^2|\mathbf{x}_i)$ *and* $e_i = \begin{cases} \exp\left(\mathbf{x}_i'\boldsymbol{\theta}\right)\log\left(\frac{w_i}{y_0}\right) - 1, & for \quad I_{\{y_0 \le w_i < y_c\}} \\ -\exp\left(\mathbf{x}_i'\boldsymbol{\theta}\right)\log\left(\frac{y_0}{y_c}\right), & for \quad I_{\{w_i = y_c\}} \end{cases}$.

Corollary 1 *Under the same conditions of Theorem 2.1 as $n \to \infty$ it follows that,* $T_j = n^{-1/2} \left(\Sigma_{y_0,jj}^{-1/2} \right) \Lambda^{-1/2} \widehat{\theta}_j \overset{d}{\to} N(0,1)$, *where $\Sigma_{y_0,jj}^{-1/2}$ corresponds to the $(j,j)^{th}$ element of the $\Sigma_{y_0}^{-1/2}$ 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 $\overline{F}(y|x;\boldsymbol{\theta}) := P(Y > y|x;\boldsymbol{\theta}) = \left(\frac{y_0}{y_i} \right)^{\alpha(x)}$, where for the sake of simplicity, but with no loss of generality, we consider x to be a scalar and continuous. Thus, to measure the impact of x on $\alpha(x)$ and subsequently on $\overline{F}(y|x;\boldsymbol{\theta})$, consider

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

where y is an extreme value, say the $(1-u)$ quantile, with $u \in (0,1)$, such that, $y = (1-u)^{\frac{1}{\alpha(x)}} y_0$. Hence, δ in (2.13) measures the probability's percentage variation 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%.

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, $\alpha(\mathbf{x}) = \exp(\mathbf{x}'\beta)$, and the impact of x_j on $\overline{F}(y|\mathbf{x};\boldsymbol{\theta})$ is,

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

Thus, a negative (positive) coefficient increases (decreases) the likelihood of having more extreme values, i.e. $\beta_j < 0$ ($\beta_j > 0$) implies $\delta_j > 0$ ($\delta_j < 0$) (this is also obvious from the impact on $\alpha(\mathbf{x})$ since if $\alpha(\mathbf{x})$ decreases (increases), the right tail becomes (less)

heavier).

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

Remark 2. 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 $d = 0$, $\delta(u) := \left[(1 - u)^{\frac{\alpha(\mathbf{x}; d=1)}{\alpha(\mathbf{x}; d=0)} - 1} - 1 \right] \times 100$. Given that $\alpha(\mathbf{x}; d = 1)$ and $\alpha(\mathbf{x}; d = 0)$ 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. $\diamond$

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 tail index estimators

To evaluate the 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). In specific, data is simulated from the general framework,

$$y_i \sim D(\alpha(\mathbf{x}_i)) \tag{3.1}$$

$$\alpha(\mathbf{x}_i) = \exp(\beta_1 + \beta_2 x_i), \qquad x_i \sim U(0, 1) \tag{3.2}$$

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[4] and generate samples of size $n \in \{2500, 5000, 10000, 50000\}$. Moreover, we censor the sample considering $y_c = \{\hat{q}_{0.95}^y, \hat{q}_{0.99}^y\}$ 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. In the case of samples generated from a Pareto distribution we could have set $k = 1$, however, a lower value was considered in order to mimic the conditions typically found in empirical analysis.

Based on the specifications described above we generated 10,000 sequences of $\{y_i\}$ and $\{w_i\}$ of size n and use in each iteration three estimation methods:

i) The tail index regression of Wang and Tsai (2009) applied to the sequence of $\{y_i\}$. We define the resulting estimator as $\hat{\alpha}$. 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 $\hat{\alpha}^c$.

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 A1 (see Supplementary material) 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 the censoring is ignored. On the other hand, 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).

Moreover, 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 gets closer to the tail of a Pareto distribution.

To further evaluate the estimation performance of the three estimation approaches in i), ii) and iii), Figure 1 plots the ratios of the RMSEs of the α estimates obtained under these estimations approaches. In specific, the ratios are,
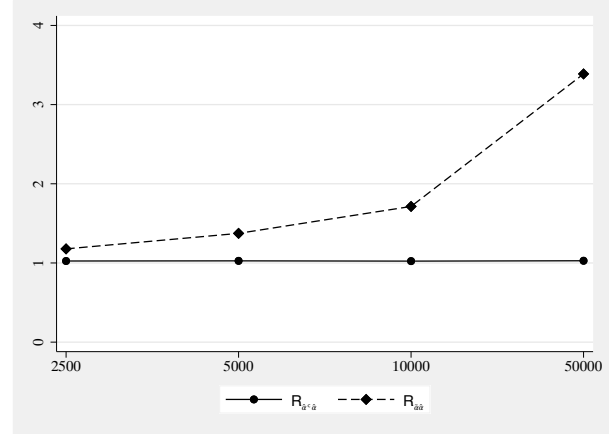
$$\mathsf{R}_{\hat{\alpha}^c\hat{\alpha}} = \frac{RMSE\left(\hat{\alpha}^c\left(\mathbf{x}_i\right)\right)}{RMSE\left(\hat{\alpha}\left(\mathbf{x}_i\right)\right)}, \qquad \mathsf{R}_{\tilde{\alpha}\hat{\alpha}} = \frac{RMSE\left(\tilde{\alpha}\left(\mathbf{x}_i\right)\right)}{RMSE\left(\hat{\alpha}\left(\mathbf{x}_i\right)\right)}.$$

Since $\hat{\alpha}$, obtained as described in i), is the best estimator, $\mathsf{R}_{\hat{\alpha}^c\hat{\alpha}}$ and $\mathsf{R}_{\tilde{\alpha}\hat{\alpha}}$ are larger than 1, across the different values of n. However, $\mathsf{R}_{\hat{\alpha}^c\hat{\alpha}}$ is just slightly above 1, which means that the censored estimator performs very well and mimics closely the behavior of the best estimator, $\hat{\alpha}$, although the former is based on the censored data. On the contrary, $\mathsf{R}_{\tilde{\alpha}\hat{\alpha}}$ is substantially higher than 1, which means that, ignoring the censoring when estimating the tail index produces an inconsistent estimator; see Figure 1.
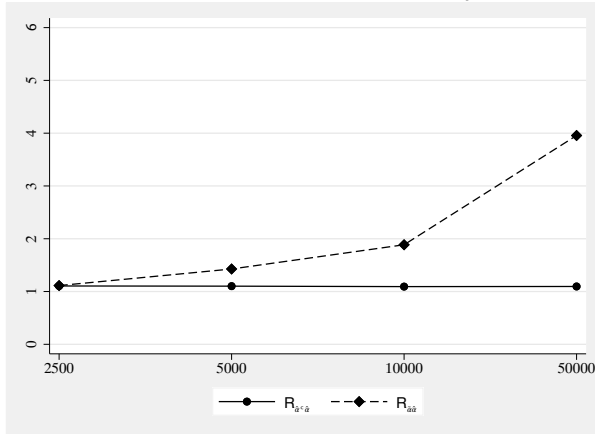
The censoring threshold y_c also impacts the estimation results, i.e., the lower its value, the greater is the impact of censoring on estimation, and $\mathsf{R}_{\tilde{\alpha}\hat{\alpha}}$ tends to be larger (see, for example, the results for Case 4 in Table A1 (see Supplementary material) and Figure 1).
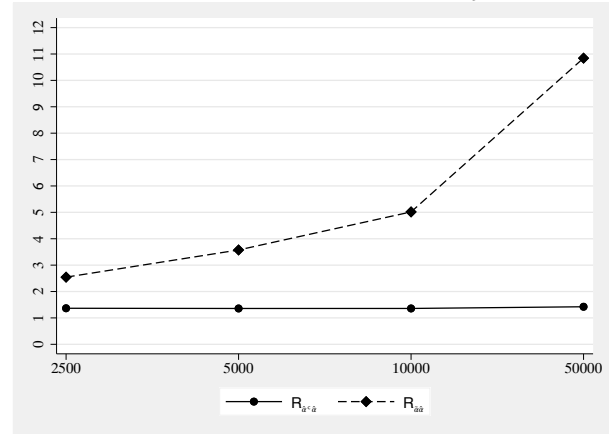
11
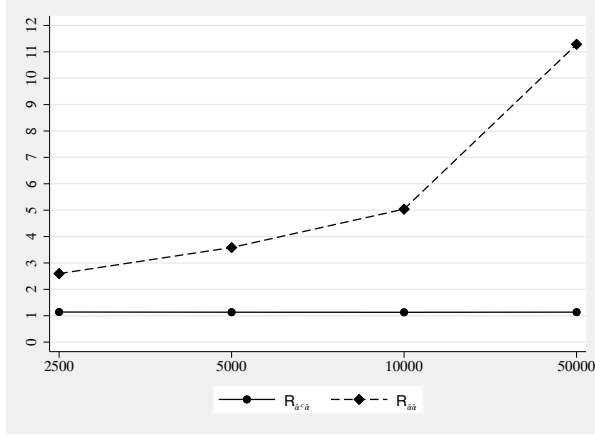
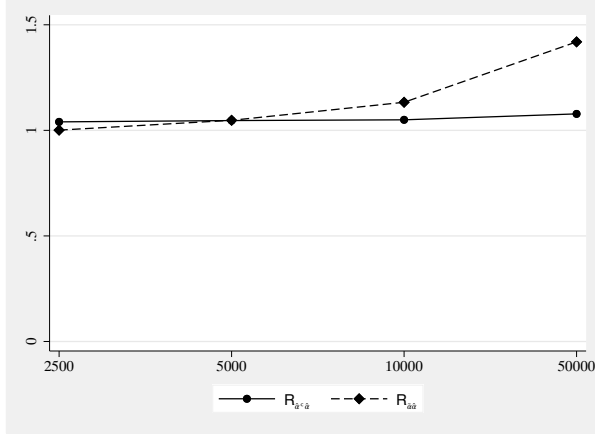DGP1: Pareto ($\kappa = 0.20$ and $y_c = Q_y(0.95)$)   DGP2: Pareto ($\kappa = 0.20$ and $y_c = Q_y(0.99)$)
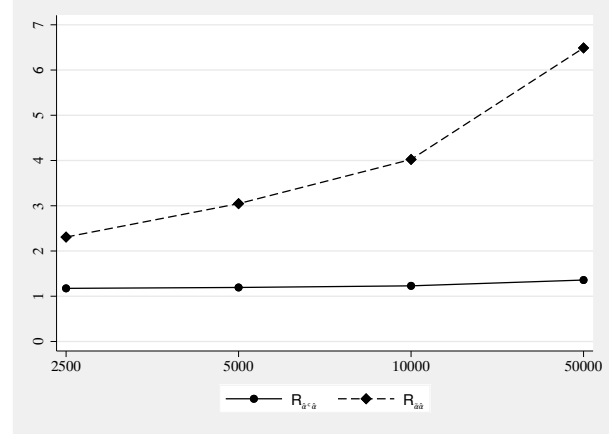
DGP3: Burr ρ=-2 ($\kappa = 0.05$ and $y_c = Q_y(0.99)$) DGP4: Burr ρ=-2 ($\kappa = 0.10$ and $y_c = Q_y(0.95)$)

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: $\mathsf{R}_{\hat{\alpha}^c\hat{\alpha}} = RMSE\left(\hat{\alpha}^c\right)/RMSE\left(\hat{\alpha}\right)$ and $\mathsf{R}_{\tilde{\alpha}\hat{\alpha}} = RMSE\left(\tilde{\alpha}\right)/RMSE\left(\hat{\alpha}\right)$. $\mathsf{R}_{\hat{\alpha}^c\hat{\alpha}}$ compares the censored estimator with the best estimator $\hat{\alpha}$, while $\mathsf{R}_{\tilde{\alpha}\hat{\alpha}}$ compares the estimator that ignores the censoring with the best estimator $\hat{\alpha}$. We report the results using the Pareto and the Burr distributions as DGP.

3.2 Imputing mean wages

To provide further insights on the usefulness of the procedure introduced in this paper we provide next an analysis of the performance of the different methods for imputing mean wages. In specific, we compare the following three methods:

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

$$\hat{\tau}_1(y_c) = \frac{\hat{\alpha}_1}{\hat{\alpha}_1 - 1} y_c \qquad (3.3)$$

where $\hat{\alpha}_1$ is the tail index estimate considering an uncensored Pareto distribution as in Section 2 (see e.g. Hill, 1975 and Nicolau and Rodrigues, 2019 for tail index estimators) and y_c is the top-code threshold;

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

$$\hat{\tau}_2(y_c) = \frac{\hat{\alpha}_{Hill}^c}{\hat{\alpha}_{Hill}^c - 1} y_c \qquad (3.4)$$

where $\hat{\alpha}_{Hill}^c$ 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 wage based on the method introduced in this paper

   a) The first is the mean predictor,

$$\hat{\tau}_3(\mathbf{x}_i, y_c) = \frac{\hat{\alpha}(\mathbf{x}_i)}{\hat{\alpha}(\mathbf{x}_i) - 1} y_c; . \qquad (3.5)$$

   where $\hat{\alpha}(\mathbf{x}_i) = \exp\left(\mathbf{x}_i'\hat{\boldsymbol{\theta}}\right)$. Note that $\tau_3(\mathbf{x}_i, y_c) := E(y_i | \mathbf{x}_i, y_i > y_c)$.

   b) The second is the median estimator,

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

   where $\hat{\alpha}(\mathbf{x}_i) = \exp\left(\mathbf{x}_i'\hat{\boldsymbol{\theta}}\right)$. This estimator is particularly useful when $\hat{\alpha}(\mathbf{x}_i) \leq 1$.

The main difference between $\hat{\tau}_3\left(\mathbf{x}_i, y_c\right)$ (and $\hat{\tau}_4\left(\mathbf{x}_i, y_c\right)$) and the other two approaches $\left(\hat{\tau}_1\left(y_c\right)\text{ and }\hat{\tau}_2\left(y_c\right)\right)$ is that in the former the particular characteristics of the individuals whose wage is above the threshold are taken into account through $\mathbf{x}_i$. Interestingly, the estimator $\hat{\tau}_3\left(\mathbf{x}_i, y_c\right)$ corresponds to the optimal mean square predictor because $\tau_3\left(\mathbf{x}_i, y_c\right)$ is the conditional expectation of y_i given $\mathbf{x}_i$ and $y_i > y_c$. This follows from the well known result $E\left(y_i - E\left(y_i | \mathbf{x}_i, y_i > y_c\right)\right)^2 \leq E\left(y_i - g\left(.\right)\right)^2$, where $g\left(.\right)$ is any other predictor of y_i given $y_i > y_c$. It turns out that $\tau_2\left(y_c\right)$ is optimal only if y_i is mean-independent of $\mathbf{x}_i$, in which case both $\tau_2\left(y_c\right)$ and $\tau_3\left(\mathbf{x}_i, y_c\right)$ coincide.

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

1. Select a sample size from $n \in \{250, 500, 1000, 2000, 5000, 20000\}$;

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

3. 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 $w_i = \min\left(y_i, y_c\right)$, from which we estimate $\hat{\tau}_1\left(y_c\right)$, $\hat{\tau}_2\left(y_c\right)$, $\hat{\tau}_3\left(\mathbf{x}_i, y_c\right)$ and $\hat{\tau}_4\left(\mathbf{x}_i, y_c\right)$.

4. The estimators $\hat{\tau}_1\left(y_c\right)$, $\hat{\tau}_2\left(y_c\right)$, $\hat{\tau}_3\left(\mathbf{x}_i, y_c\right)$ and $\hat{\tau}_4\left(\mathbf{x}_i, y_c\right)$ are used to predict the imputed mean value above the threshold y_c.

5. Steps 2 to 4 are repeated 2000 times and the mean square errors (MSE) of $\hat{\tau}_1\left(y_c\right)$, $\hat{\tau}_2\left(y_c\right)$, $\hat{\tau}_3\left(\mathbf{x}_i, y_c\right)$ and $\hat{\tau}_4\left(\mathbf{x}_i, y_c\right)$ are computed.

Other combinations of $\alpha\left(\mathbf{x}_i\right)$ produce essentially the same results as long as $\alpha\left(\mathbf{x}_i\right) \geq 1$, and for this reason we present results only for the case $\alpha\left(\mathbf{x}_i\right) = \exp\left(1 + 2x_i\right)$ (note

that $E\left(\alpha\left(\mathbf{x}_i\right)\right) = 3.7$). However, $\alpha\left(\mathbf{x}_i\right)$ should be set so that $\alpha\left(\mathbf{x}_i\right) > 1$, otherwise the conditional and marginal expected values do not exist, and consequently $\hat{\tau}_1\left(y_c\right)$, $\hat{\tau}_2\left(y_c\right)$ and $\hat{\tau}_3\left(\mathbf{x}_i, y_c\right)$ are not well defined. The case $0 < \alpha\left(\mathbf{x}_i\right) < 1$ can be dealt with using the $\hat{\tau}_4\left(\mathbf{x}_i, y_c\right)$ estimator, as indicated above.

Figure 2 illustrates our results. The lines represent the MSE ratios, $\mathsf{M}_{13} := \frac{MSE(\hat{\tau}_1(y_c))}{MSE(\hat{\tau}_3(\mathbf{x}_i, y_c))}$, $\mathsf{M}_{23} := \frac{MSE(\hat{\tau}_2(y_c))}{MSE(\hat{\tau}_3(\mathbf{x}_i, y_c))}$ and $\mathsf{M}_{43} := \frac{MSE(\hat{\tau}_4(\mathbf{x}_i, y_c))}{MSE(\hat{\tau}_3(\mathbf{x}_i, y_c))}$, computed over different sample sizes (n = (250, 500, 1000, 2000, 5000, 10000)).
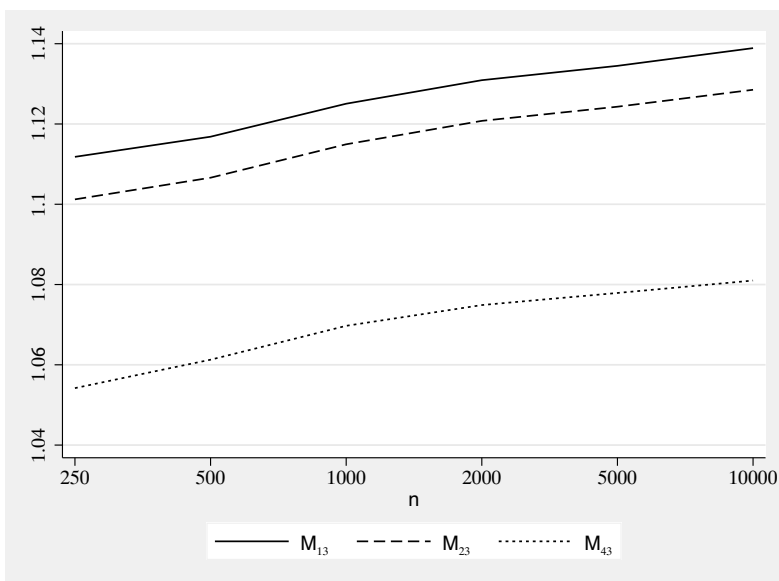

Figure 2: MSE ratios computed over different sample sizes

Note: The lines represent the the MSE ratios, $\mathsf{M}_{13} := \frac{MSE(\hat{\tau}_1(y_c))}{MSE(\hat{\tau}_3(\mathbf{x}_i, y_c))}$, $\mathsf{M}_{23} := \frac{MSE(\hat{\tau}_2(y_c))}{MSE(\hat{\tau}_3(\mathbf{x}_i, y_c))}$ and $\mathsf{M}_{43} := \frac{MSE(\hat{\tau}_4(\mathbf{x}_i, y_c))}{MSE(\hat{\tau}_3(\mathbf{x}_i, y_c))}$ computed over different sample sizes ($n = (250, 500, 1000, 2000, 5000, 10000)$).

As can be observed from Figure 2, the $\hat{\tau}_3(\mathbf{x}_i, y_c)$ and $\hat{\tau}_4(\mathbf{x}_i, y_c)$ estimators, introduced in this paper produce the best results. The MSE ratios, M_{13} and M_{13} are larger than one and larger than M_{43}. The gains increase steadily as the sample size grows. Notice further that $\hat{\tau}_2(y_c)$ (Armour et al., 2016) is better than the (naive) $\hat{\tau}_1(y_c)$ estimator that does not accommodate the censoring and that $\hat{\tau}_3(\mathbf{x}_i, y_c)$ is slightly better than $\hat{\tau}_4(\mathbf{x}_i, y_c)$.

3.3 Robustness testing our method using non-censored data

To further 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 that can be obtained from our methodology (as an ordinary out-of-sample forecast exercise). We censor the sample considering $y_c \in \{\hat{q}^y_{0.97}, \hat{q}^y_{0.975}, ..., \hat{q}^y_{0.99}\}$, where the generic element $\hat{q}^y_p$ 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 $\kappa \in \{0.1, 0.15, 0.2, 0.25\}$. For comparison purposes we study three estimation methods applied to the censored data:

1. We start by computing $\hat{\alpha}(\mathbf{x}_i)$ from the censored tail index regression introduced in this paper. Then, we use $\hat{\tau}_4(\mathbf{x}_i, y_c)$ 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 dataset, such as, age, gender, tenure, education, job, region and industry. However, instead of searching for the covariates with greatest explanatory power, for each year we take a simpler approach. We just use the same basic covariates throughout all combinations of year, κ and y_c. Specifically, we consider $\mathbf{x}_i = [gender_i, age_i, education_i, tenure_i]$. 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, $\hat{\tau}_4(\mathbf{x}_i, y_c)$. Actually, some estimates of $\hat{\alpha}(\mathbf{x}_i)$, $i = 1, 2, ..., n$, are close to 1, and for these cases, the conditional mean $\hat{\alpha}(\mathbf{x}_i) / (\hat{\alpha}(\mathbf{x}_i) - 1) y_c$ is not defined. A remedy is to always use the conditional median, which, in most of the cases, performs very well (see results in Section 3.2).

2. The Hill estimator $(\hat{\tau}_1(y_c))$, which ignores censored data.

3. The Armour estimator $(\hat{\tau}_2(y_c))$, that takes into consideration censored data.

To assess the quality of the various estimators we use the RMSE statistics,

$$RMSE\left(j\right) = \sqrt{\frac{1}{s}\sum_{i=1}^{s}\left(\hat{\tau}_j\left(y_c\right) - y_i\right)^2}$$

where $j = 1, 2, 4$, and $\hat{\tau}_j\left(y_c\right)$ is the estimator used to predict the imputed mean value above the top-code y_c. The parameter s represents the number of wages above the top-code.

$\hat{\tau}_2(y_c)$ generally performs better than $(\hat{\tau}_1(y_c))$ in both of the years. This is an expected result given that $(\hat{\tau}_2(y_c))$, which is an adaptation of $(\hat{\tau}_1(y_c))$, takes the censoring into consideration. 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).

The tail index regression 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 forecast (results are available upon request).
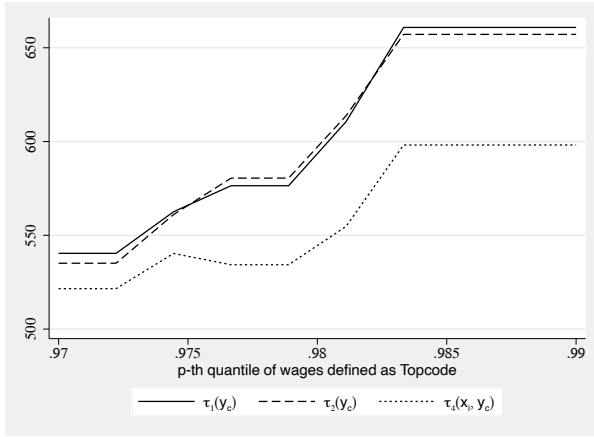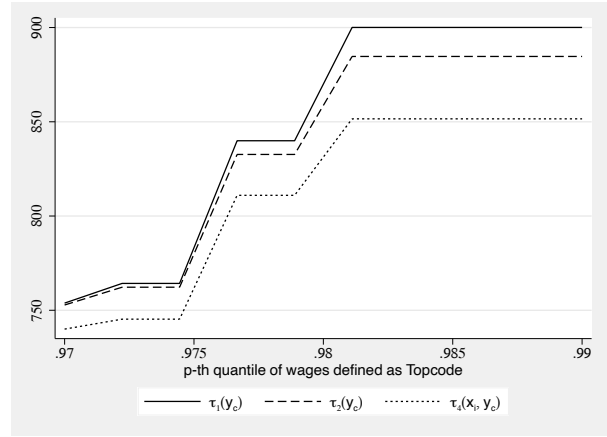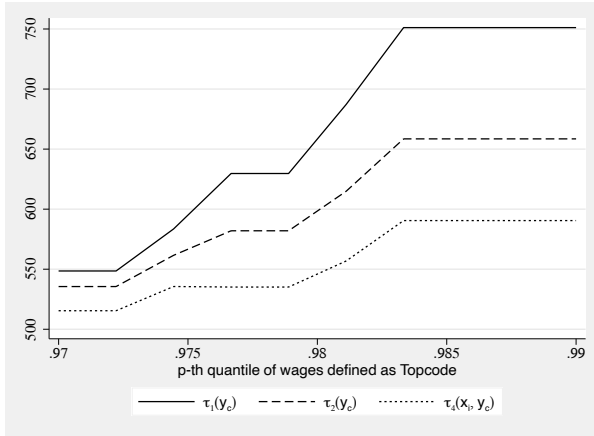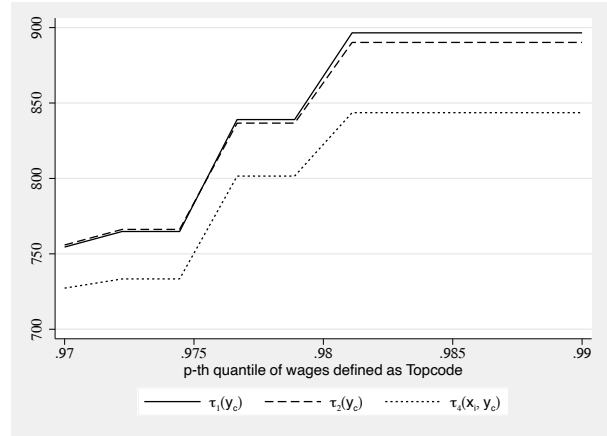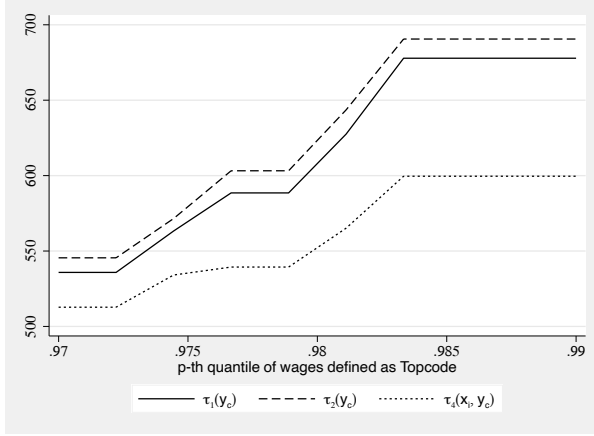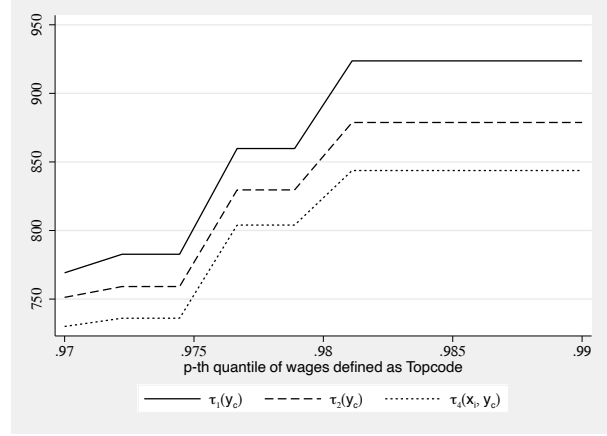
Figure 3: RMSE of imputed wages based on $\hat{\tau}_1(y_c)$, $\hat{\tau}_2(y_c)$ and $\hat{\tau}_4(\mathbf{x_i}, y_c)$, computed from non censored wage data from the Portuguese Labor force survey.

4   Empirical analysis

In this section, we use censored publicly available CPS data to evaluate how the right tail index of the US 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 wages. We also assess the sensitivity of the adjustment of the top coded 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 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 to 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 weekly wages paid.[5,6]

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.
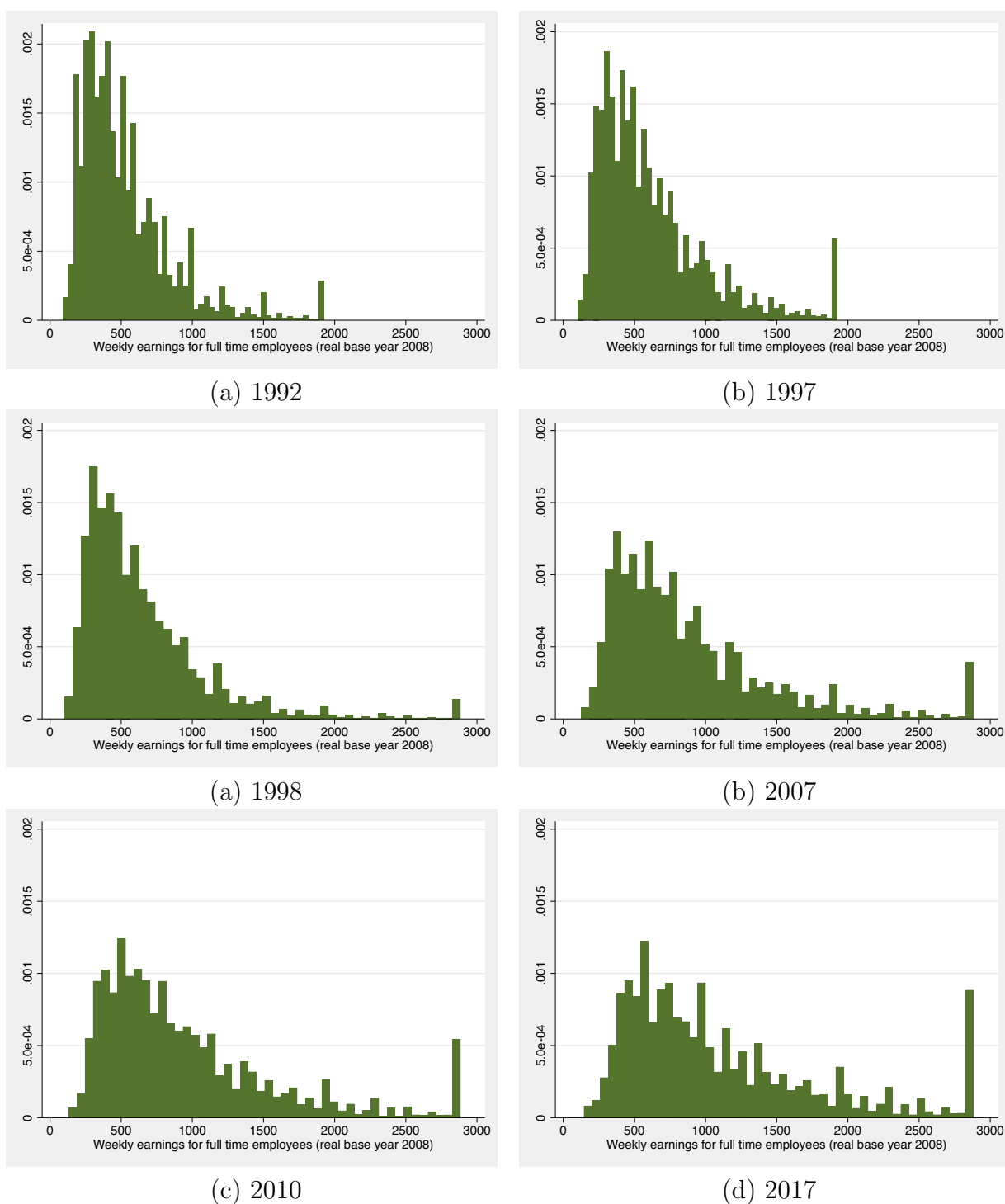
Figure 4: Annual unconditional histograms of the weekly wage.

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 larger (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: 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

[a] In Appendix C we provide a detailed description of the industry classification.

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.

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 represent 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 wage distribution, Figure 4 plots the proportions in percentiles 0.05 to 0.95 considering different attributes, occupations and industries (in the appendix we present additional plots for all other cases analysed 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 analysed.
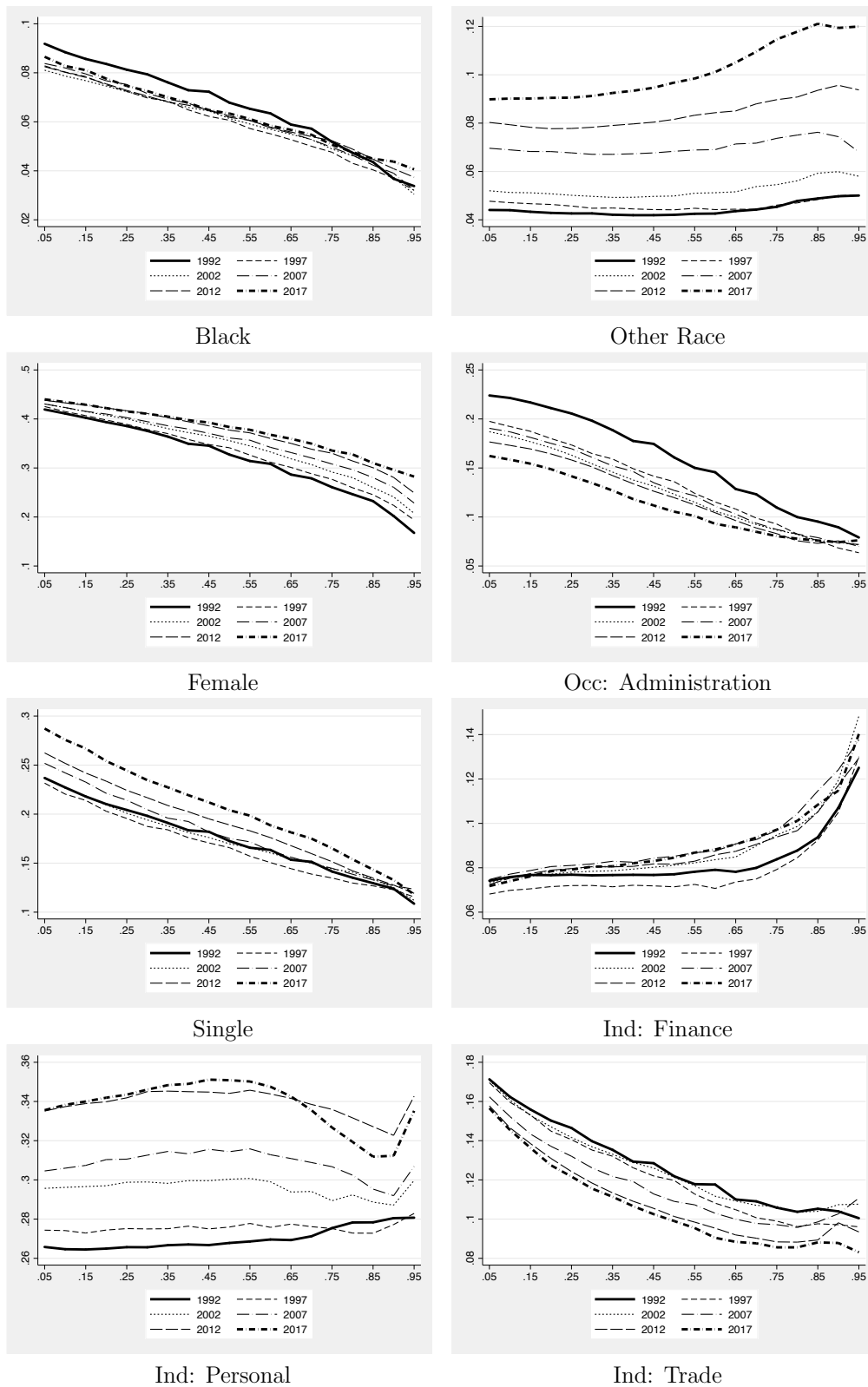
Figure 5: Proportion of individuals in different wage percentiles

Notes: The graphs presented in this figure 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 Conditional tail index estimation results

Table 2 presents the censored and uncensored tail index regression results for 1992 and 2017. A negative (positive) regression coefficient corresponds to a decrease (increase) in the tail index (ceteris paribus) and hence a larger (smaller) number of extreme values may result as a consequence of changes in the specific variable associated to such a coefficient. Before analyzing the partial effects, as described in Section 2.2, the behavior of a tail index estimate computed as an average of the conditional tail indexes is examined, i.e.,

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

where $\hat{\alpha}\left(\mathbf{x}_{t,i}\right) = \exp\left(\mathbf{x}'_{t,i}\hat{\theta}_t\right)$ (we only report results for 1992 and 2017, however the tail index regression estimation results from 1993 to 2016 can be obtained upon request). The tail index estimate in (4.1), $\hat{\alpha}_t$, provides an estimate of the unconditional tail index after considering the characteristics of all individuals in the sample for each year.

The results in Table 2 show that in general, the estimates based on the method that ignores censored data underestimate the true effects of the variables. This is especially clear in the estimates for 2017 (e.g. female and finance). It is a consequence of the potential inconsistency of the uncensored estimates as highlighted in the Monte Carlo section above. However, the direction of the impact of the covariates suggested by the uncensored estimation is consistent with the results obtained from the censored tail index regression.

Comparing 1992 and 2017 (censored estimation), 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, 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.

Table 2: Uncensored and censored tail index regression results

		Uncensored		Censored	
		1992	2017	1992	2017
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***
		(0.001)	(0.001)	(0.001)	(0.001)
Female		0.272***	0.134***	0.309***	0.229***
		(0.014)	(0.013)	(0.016)	(0.022)
Education		−0.076***	−0.061***	−0.085***	−0.100***
		(0.004)	(0.004)	(0.004)	(0.006)
Race					
	Black	0.085***	0.004	0.098***	0.010***
		(0.029)	(0.034)	(0.031)	(0.054)
	Other race	−0.023***	0.004	−0.031	0.015
		(0.025)	(0.016)	(0.030)	(0.033)
Marital Status					
	Married no spouse	0.075	−0.036	0.095	−0.100
		(0.063)	(0.046)	(0.070)	(0.091)
	Separated	0.035	0.051	0.044	0.093
		(0.042)	(0.059)	(0.047)	(0.096)
	Divorced	0.018	0.066***	0.025	0.119***
		(0.019)	(0.022)	(0.022)	(0.036)
	Widowed	0.003	−0.070	0.022	−0.083
		(0.083)	(0.052)	(0.077)	(0.099)
	Single	0.009	0.066***	0.015	0.114***
		(0.018)	(0.018)	(0.021)	(0.029)
Occupations					
	Administrative	0.113***	0.021	0.125***	0.006
		(0.020)	(0.022)	(0.023)	(0.047)
	Low Skill	0.098	0.093	0.111*	0.153*
		(0.064)	(0.060)	(0.067)	(0.091)
	Craft	0.289***	0.069	0.324***	0.154*
		(0.033)	(0.056)	(0.035)	(0.086)
	Operators	0.394***	0.148***	0.416***	0.217***
		(0.050)	(0.062)	(0.053)	(0.090)
	Transportation	0.366***	0.184***	0.400***	0.264***
		(0.023)	(0.030)	(0.025)	(0.045)
Industry					
	Construction	−0.202***	−0.117***	−0.234***	−0.173***
		(0.039)	(0.042)	(0.047)	(0.071)
	Manufacturing	0.003	−0.011	0.004	0.026
		(0.028)	(0.029)	(0.031)	(0.052)
	Transports	0.050***	−0.016	0.051**	−0.001
		(0.021)	(0.027)	(0.024)	(0.040)
	Trade	0.002	−0.052**	0.001	−0.052
		(0.020)	(0.024)	(0.025)	(0.043)
	Finance	−0.126***	−0.099***	−0.184***	−0.219***
		(0.020)	(0.022)	(0.026)	(0.043)
	Repair	−0.059***	−0.098***	−0.071**	−0.142***
		(0.028)	(0.022)	(0.033)	(0.041)
	Personal	0.125***	0.048***	0.127***	0.060*
		(0.016)	(0.018)	(0.019)	(0.033)
	Public	0.238***	0.120***	0.272***	0.265***
		(0.022)	(0.024)	(0.025)	(0.040)

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.
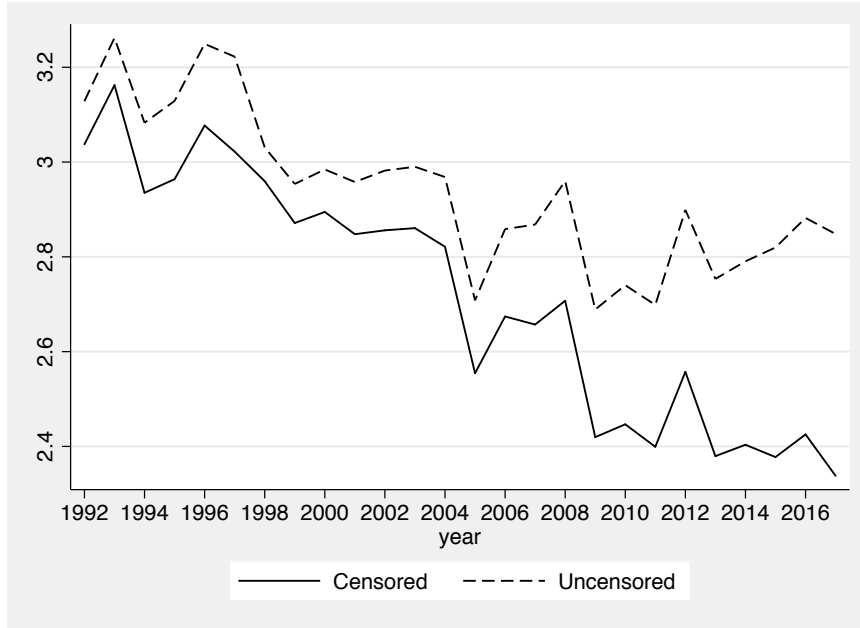
Figure 6: Uncensored and censored right tail index estimates from 1992 to 2017

Note: This figure reports the yearly right tail index estimates computed from both the uncensored and the censored regressions.

Figure 6 plots the tail index estimate computed as suggested in (4.1) for each year from 1992 to 2017, based on uncensored and censored tail index regression estimates. As discussed above, the uncensored approach misrepresents the true unconditional tail index, as it ignores the top-coded wages, leading to the overestimation of the true values of α_t (where the bias/inconsistency is due to the fact that the top extreme values are simply not included in the estimation). On the contrary, the censored estimates take the information of the individuals of the top-coded wages, although the true wages are not known, into account. Figure 6 shows that the censored estimates of $\hat{\alpha}_t$ have declined over the last 20 years. 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 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) may be a lower bound of the true increase in inequality.

Not adequately handling the top-coding may also lead to overestimation of the tail index in a variety of other applications (this has been recognized in e.g. the analysis of returns to education by Hubbard (2011), and differences in gender and race by Burkhauser and Larrimore (2009) who already use approaches to accommodate for the fact that the wage data is top-coded). Our procedure provides additional flexibility by allowing researchers to evaluate which determinants impact the probability of being in the right tail of the wage distribution and which determinants do not. Thus, permits a more detailed analysis on how inequality has spread across industries, occupation, gender and other population characteristics.

In what follows, we focus on the (conditional) tail index regression, and especially on the partial effects computed as discussed in section 2.2. Figure 7 groups the partial effect estimates and rearranges them from the lowest effect to the highest. A negative (positive) regression coefficient translates into a positive (negative) partial effect, which is associated with an increase (decrease) in the likelihood of having more extreme values (the right tail becomes (less) heavier).

In the discussion that follows on the partial effects of the covariates, whenever we refer to an extreme event, we are referring to observations which are larger than the 96th quantile ($u = 0.15$); see Remark 2 for details.

It is clear 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 alpha (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 an extreme value.

The impact in terms of occupation is very interesting. In comparison 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).
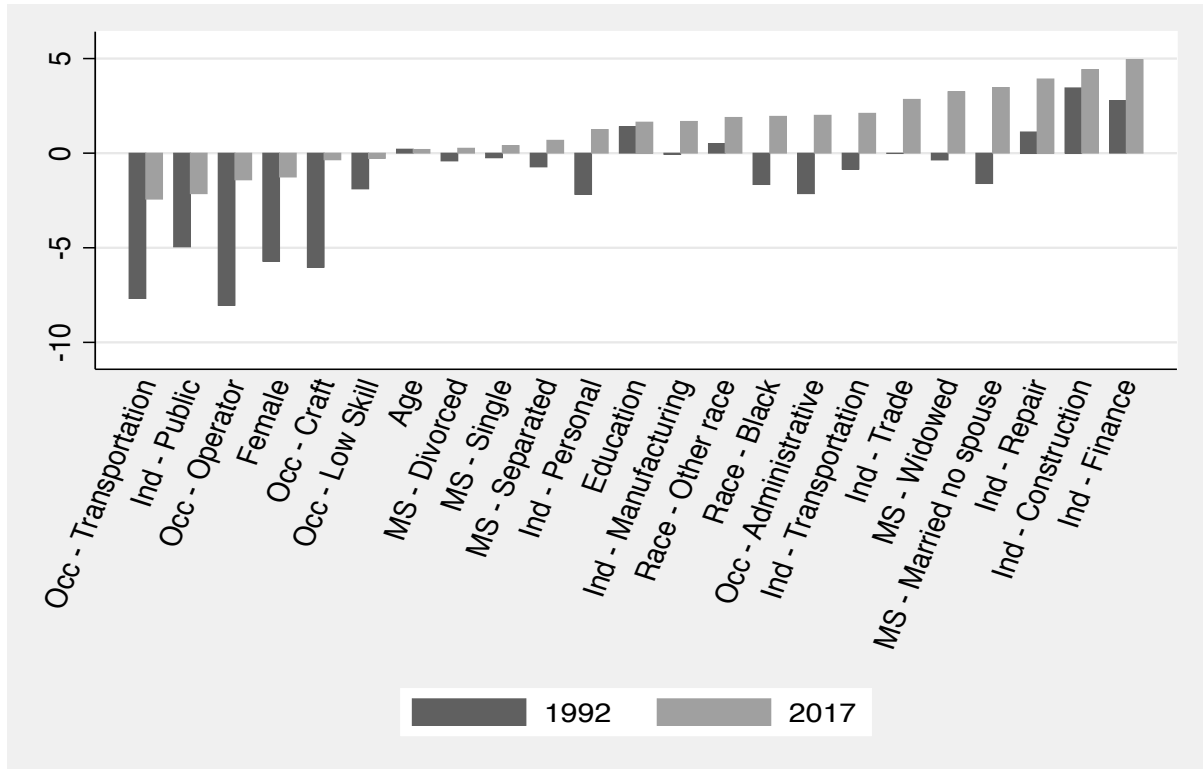
Figure 7: Partial effects of variables ($u = 0.15$)

Note: For a description of the occupations see note under Table 1.

4.3 Imputed mean wages

4.3.1 Imputing wages above the top-code

A further important contribution of the approach introduced is its use for the imputation of mean wages above the top-code as described in Section 3.2. Some authors use values to adjust the top-coded 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 favour 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 $E\left(y_i \mid y_i > y_c; \mathbf{x}_i\right)$ given by $\hat{\tau}_3\left(\mathbf{x}_i, y_c\right)$ in (3.5). However, for some individuals in the CPS database, especially those in the highest wage groups, $\hat{\alpha}\left(\mathbf{x}_i\right)$ can be in the neighborhood of 1, or even lower than 1, which implies that $E\left(y_i \mid y_i > y_c; \mathbf{x}_i\right)$ does not exist and, therefore, the estimate $\hat{\tau}_3\left(\mathbf{x}_i, y_c\right)$ 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 $\hat{\tau}_4\left(\mathbf{x}_i, y_c\right)$ statistic in (3.6). In specific, to accommodate all situations (i.e. low and high values of $\hat{\alpha}\left(\mathbf{x}_i\right)$) we use $\hat{\tau}_4\left(\mathbf{x}_i, y_c\right)$ when $0 < \hat{\alpha}\left(\mathbf{x}_i\right) \leq c$ and $\hat{\tau}_3\left(\mathbf{x}_i, y_c\right)$ for $\hat{\alpha}\left(\mathbf{x}_i\right) > c$. In the empirical application we set $c = 1.5$, since using $c = 1$ may lead to explosive estimates as $\frac{\hat{\alpha}(\mathbf{x}_i)}{\hat{\alpha}(\mathbf{x}_i)-1} \to \infty$ as $\hat{\alpha}\left(\mathbf{x}_i\right) \to 1^+$. Other values of c in the neighborhood of $c = 1.5$ yield basically the same results. In the application to the CPS data we observed that for the overwhelming majority of estimates $\hat{\alpha}\left(\mathbf{x}_i\right) > 1.5$. Thus, the $\hat{\tau}_3\left(\mathbf{x}_i, y_c\right)$ estimator in (3.5) is mostly used. For simplicity of presentation, in the discussion below we will just refer to $\hat{\tau}_3\left(\mathbf{x}_i, y_c\right)$ although imputed wages were computed as described above.

Figure 8 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 $\hat{\tau}_2(y_c)$ and $\hat{\tau}_3\left(\mathbf{x}_i, y_c\right)$ estimates are similar. This result is expected given that for the overall analysis $\hat{\tau}_3\left(\mathbf{x}_i, y_c\right)$ is based on the values of $\hat{\alpha}_t$ 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 $\hat{\tau}_3\left(\mathbf{x}_i, y_c\right)$ estimates will certainly be different from the $\hat{\tau}_2(y_c)$ estimates (see next section).
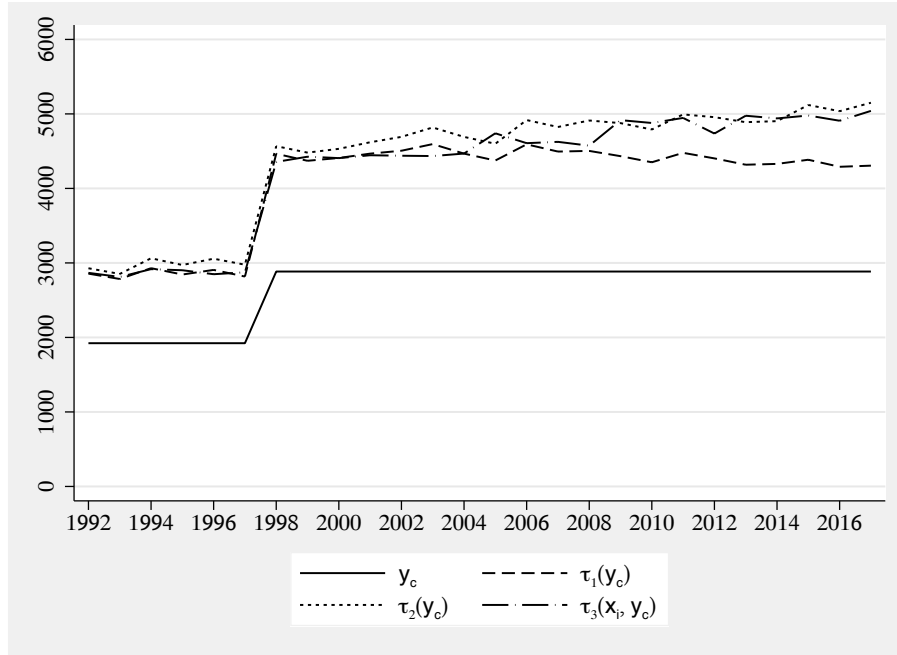
Figure 8: Prediction of top-coded wages

When the proportion of wages above the top-code is relatively small (as for example, from 1992 to 2005), the difference between $\hat{\tau}_1\left(y_c\right)$, $\hat{\tau}_2\left(y_c\right)$ and $\hat{\tau}_3\left(\mathbf{x}_i, y_c\right)$ 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 $\left(\hat{\tau}_1\left(y_c\right)\right)$.

Figure 9 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 a sample between 1980-2005; Acemoglu and Autor (2011), between 1973-2009; Autor et al. (2008) from 1963 to 2005; Katz and Murphy (1992) between 1963 and 1987; Autor and Dorn (2013) between 1950 and 2005; Lemieux (2006) from 1973 to 2003; and Beaudry, Green and Sand (2013) from 1979 to 2011. Figure 9 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.
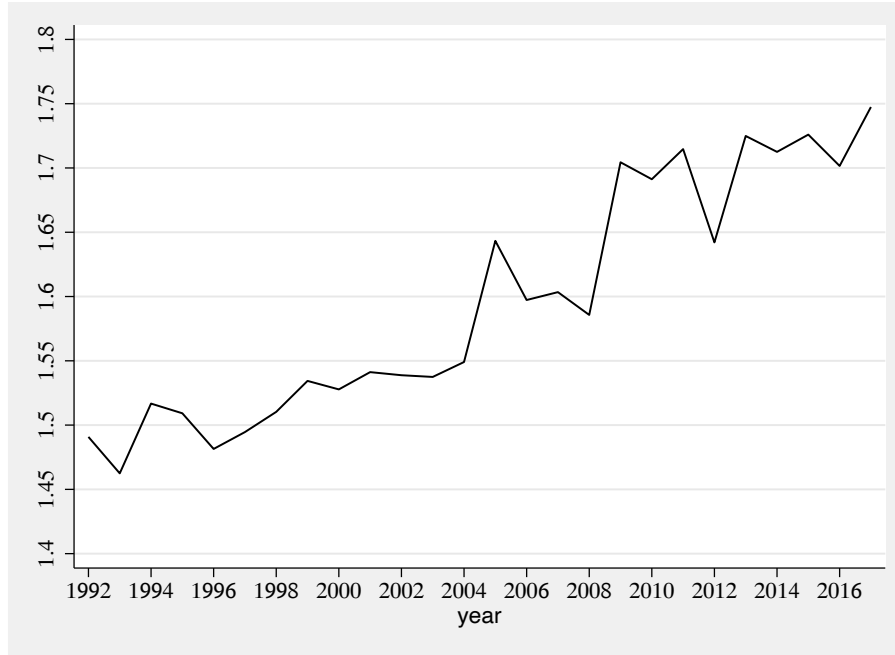
Figure 9: Top-coded wage adjustment factors between 1992 and 2017

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

Figure 10 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.
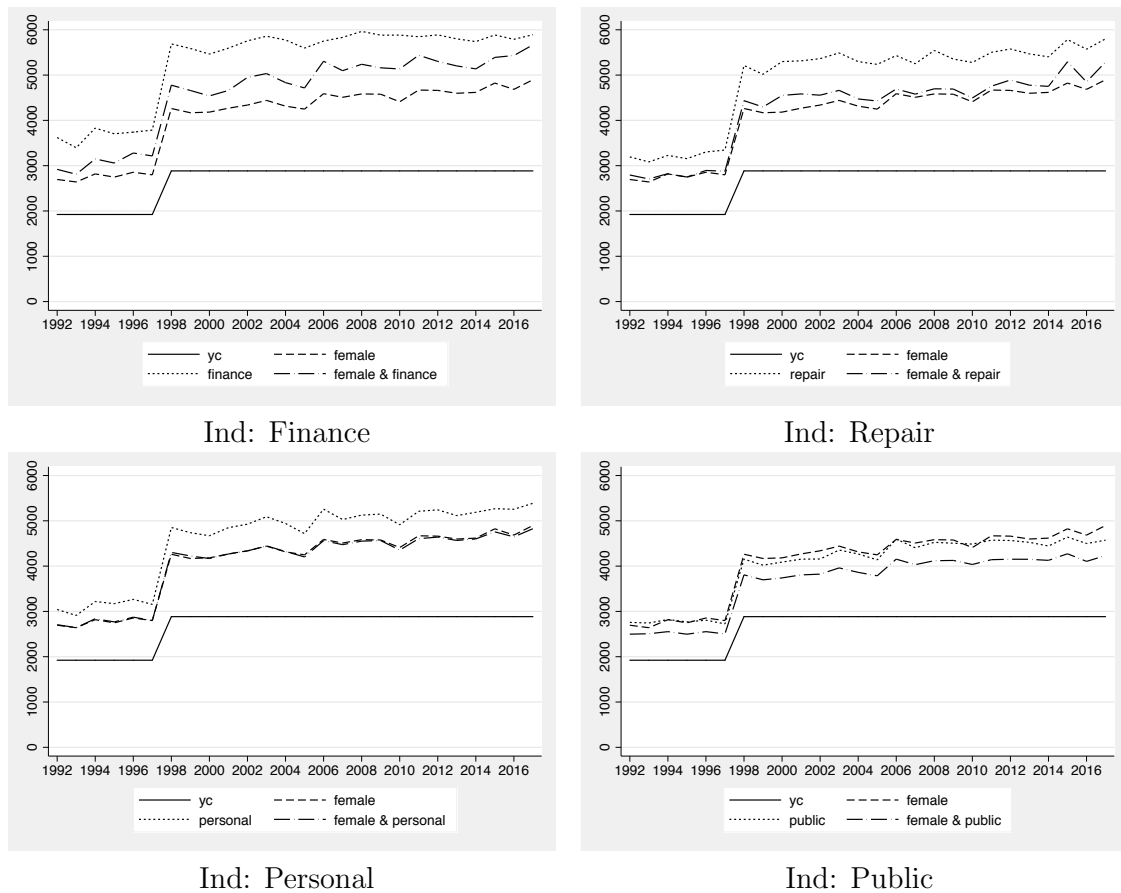
34

Figure 10: 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. Womens top-coded adjustment factor is always smaller than the topcoded 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.
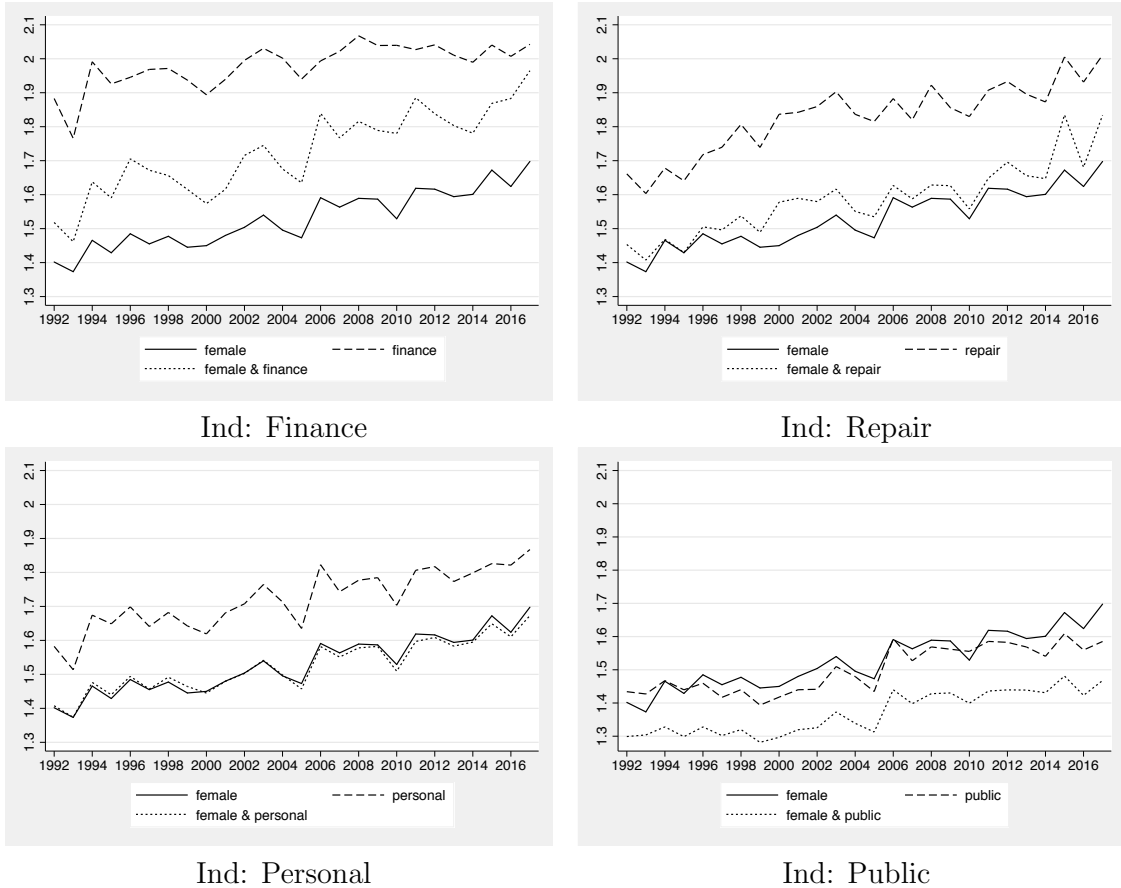
Figure 11: Adjustment factors between 1992 and 2017 by industry to impute top-coded wages

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 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 (public) would need a higher (lower) adjustment factor to impute the top-coded wages. In the period 1992-2017, the adjustment factor changed for women working in the finance (public) industry between 1.5 (1.3) in 1992 to 2 (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 especially when using censored data.

Third, an in depth empirical analysis of the dynamics of the US 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 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. It is also noted that in comparison to an individual working in a non-routine cognitive occupation (managers) all occupations displayed a positive contribution to observe an extreme value (although smaller in 2017). However, conclusions are different across occupations.

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. This distribution was used, for instance, by Harrison (1981) to analyse earnings by size in the UK.

3. 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.

4. 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 $f(x) := x^{-1-\alpha\rho}(1 + x^{-\alpha\rho})^{-1\frac{1}{\rho}}\alpha$.

5. Wages are converted to 2012 values using the GDP personal consumption expenditure deflator.

6. 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.

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.

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

Burkhauser, R. V. and Larrimore, J. (2009). Using internal CPS data to reevaluate trends in labor-earnings gaps. *Monthly Labor Review*, 132(8):318.

Card, D. and DiNardo, J. E. (2002). Skill-biased technological change and rising wage inequality: Some problems and puzzles. *Journal of Labor Economics*, 20(4):733–783.

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.

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

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

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

Kalleberg, A. L. (2011). *Good jobs, bad jobs: The rise of polarized and precarious employment systems in the United States, 1970s-2000s*. Russell Sage Foundation.

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.