

08

WORKING
PAPERS 2020



MEASURING WAGE
INEQUALITY UNDER
RIGHT CENSORING

João Nicolau | Pedro Raposo
Paulo M. M. Rodrigues



BANCO DE
PORTUGAL
EUROSYSTEM

08

WORKING PAPERS 2020

MEASURING WAGE INEQUALITY UNDER RIGHT CENSORING

João Nicolau | Pedro Raposo
Paulo M. M. Rodrigues

MAY 2020

The analyses, opinions and findings of these papers represent
the views of the authors, they are not necessarily those of the
Banco de Portugal or the Eurosystem

Please address correspondence to
Banco de Portugal, Economics and Research Department
Av. Almirante Reis, 71, 1150-012 Lisboa, Portugal
Tel.: +351 213 130 000, email: estudos@bportugal.pt



BANCO DE PORTUGAL
EUROSYSTEM

Lisboa, 2020 • www.bportugal.pt

Measuring wage inequality under right censoring

João Nicolau

ISEG-Universidade de Lisboa and
CEMAPRE

Pedro Raposo

Católica Lisbon School of Business and
Economics

Paulo M. M. Rodrigues

Banco de Portugal and Nova School of
Business and Economics

May 2020

Abstract

In this paper we investigate 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. In specific, 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. Thus, having a tail index estimator that explicitly accounts for this sample characteristic is of importance to better understand and compute the tail index dynamics in the censored right tail of the wage distribution. 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; ii) we highlight that the factor values used to adjust the top-coded wage have changed over time and depend on the characteristics of individuals, occupations and industries, and propose suitable values; and iii) we provide 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.

JEL: C18, C24, E24, J11, J31

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

E-mail: nicolau@iseg.utl.pt; pedro.raposo@ucp.pt; pmrodrigues@bportugal.pt

The opinions expressed in this article are those of the authors and do not necessarily coincide with those of Banco de Portugal or the Eurosystem. Any errors and omissions are the sole responsibility of the authors.

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.

Existing 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 *et al.* (2006)).² 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)).

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. Parker (1999) developed a model of wages in which wages follow a Generalized Beta Distribution of the second kind (GB2). Bordley *et al.* (1995) show that GB2 exhibits better fit to US wage data than alternative distributions. Because the authors are modelling the distribution of total wage and not its components, they only need to know whether total wages are censored or not, and therefore do not need to be concerned with consistency problems in categories as in Burkhauser *et al.* (2004).

While some authors have tried to address the top-coding issues by restricting the sample under analysis, the method presented in this article 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 on the level of wage inequality than other available measures. In specific, a conditional tail index regression specifically designed to account for right censoring is used. The tail index (sometimes also referred to as Pareto coefficient) 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 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; Section 4 describes and discusses the results of an empirical analysis 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, in 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 interested in measuring long-term trends in wages 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) multiplied top-coded hourly wages by 1.4, and Autor *et al.* (2008) multiplied top-coded weekly wages by 1.5); (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, *inter alia*, 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 first the survival function, \bar{F} , of a Pareto distribution³

$$\bar{F}(y) := P(Y_i > y) = \left(\frac{y_0}{y_i}\right)^\alpha, \text{ where } y_i \geq y_0 > 0 \text{ and } \alpha > 0 \quad (1)$$

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),

$$\hat{\alpha}_{Hill} := \left[\frac{1}{m} \sum_{j=1}^m \log y_{(j)} - \log y_{(0)} \right]^{-1} \quad (2)$$

where m is the number of largest order statistics used in the estimation of α , $y_{(j)}$, $j = 1, \dots, m$ 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,

$$\omega_i = \begin{cases} y_i & \text{if } y_0 \leq y_i < y_c \\ y_c & \text{if } y_i \geq y_c \end{cases}, \quad (3)$$

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,

$$g_Y(\omega_i) = \left(\frac{\alpha y_0^\alpha}{\omega_i^{\alpha+1}}\right)^{I_{(y_0 \leq \omega_i < y_c)}} \left[\left(\frac{y_0}{y_c}\right)^\alpha\right]^{I_{(\omega_i \geq y_c)}} \quad (4)$$

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

and the respective log-likelihood function,

$$\begin{aligned} \log L &= \sum_{i=1}^m g_Y(\omega_i) \\ &= \sum_{i=1}^m I_{(y_0 \leq \omega_i < y_c)} \left(\log(\alpha) + \alpha \log(y_0) - (\alpha + 1) \log(\omega_i) \right) + \\ &\quad \sum_{i=1}^m I_{(\omega_i \geq y_c)} \left(\alpha \log(y_0) - \alpha \log(y_c) \right). \end{aligned} \quad (5)$$

Consequently, the conditional MLE estimator proposed by Armour *et al.* (2016) computed from (5) is,

$$\hat{\alpha}_{Hill}^c = \frac{n_0}{\sum_{i=1}^m I_{(y_0 \leq y_i < y_c)} \log(y_i) + n_c \log(y_c) - (n_0 + n_c) \log(y_0)} \quad (6)$$

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) , where $Y_i \in \mathbb{R}^1$ is the response of interest, and $\mathbf{X}_i := (x_{1i}, \dots, x_{pi})' \in \mathbb{R}^p$ is an associated p -dimensional vector of predictors with $1 \leq i \leq n$. In addition, let $F(y|\mathbf{x}; \boldsymbol{\theta}) := P[Y_i \leq y | \mathbf{X}_i = \mathbf{x}]$ be the cumulative distribution function of Y_i conditional on \mathbf{X}_i , and assume that the corresponding survival function (under no censoring) is,

$$\bar{F}(y|\mathbf{x}; \boldsymbol{\theta}) := 1 - F(y|\mathbf{x}; \boldsymbol{\theta}) = y^{-\alpha(\mathbf{x})} \mathcal{L}(y; \mathbf{x}), \quad (7)$$

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 $\mathcal{L}(yk; \mathbf{x})/\mathcal{L}(y; \mathbf{x}) \rightarrow 1$ for any $k > 0$ as $y \rightarrow \infty$. Specifically, following Hall (1982) we characterize the

slowly varying function as,

$$\mathcal{L}(y; \mathbf{x}) := c_0(\mathbf{x}) + c_1(\mathbf{x})y^{-\beta(\mathbf{x})} + o(y^{-\beta(\mathbf{x})}) \quad (8)$$

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 $y \rightarrow \infty$, $\mathcal{L}(y; \mathbf{x}) \rightarrow c_0(\mathbf{x})$ and $\dot{\mathcal{L}}(y; \mathbf{x}) \rightarrow 0$, where $\dot{\mathcal{L}}(y; \mathbf{x}) = \partial \mathcal{L}(y; \mathbf{x}) / \partial y$.

From (7), it follows that the probability density function of Y_i conditional on \mathbf{X}_i is,

$$f(y|\mathbf{x}; \theta) = \alpha(\mathbf{x})y^{-\alpha(\mathbf{x})-1} \mathcal{L}(y; \mathbf{x}) - y^{-\alpha(\mathbf{x})} \dot{\mathcal{L}}(y; \mathbf{x}). \quad (9)$$

Considering (8) and assuming that y is sufficiently large, it follows that the density in (9) can be approximated as,

$$f(y|\mathbf{x}; \theta) \approx c_0(\mathbf{x})\alpha(\mathbf{x})y^{-\alpha(\mathbf{x})-1}; \quad (10)$$

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,

$$f(y|\mathbf{x}; \theta) \approx \alpha(\mathbf{x})(y/y_0)^{-\alpha(\mathbf{x})-1}, \quad (11)$$

where y_0 is the threshold that controls the sample fraction used for estimation. Note that (11) is the approximate conditional Pareto density function of an unrestricted random variable and thus, its use when some form of censoring (such as right censoring⁴ 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 , as in the previous section, we effectively observe w_i as defined in (3). In this context, the adequately adjusted conditional Pareto density function is,

$$g(w_i|\mathbf{x}_i, y_c, \theta) := f(w_i|\mathbf{x}_i, y_c, \theta)^{I(y_0 \leq w_i < y_c)} [1 - F(y_c|\mathbf{x}_i, y_c, \theta)]^{I(w_i \geq y_c)} \quad (12)$$

where $I(\cdot)$ is the indicator function and $f(\cdot|\mathbf{x})$ and $F(\cdot|\mathbf{x})$ 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,

$$\mathcal{K}_n^c(\theta; y_c) := \sum_{i=1}^n \log g(w_i|\mathbf{x}_i; y_c; \theta) \quad (13)$$

where $g(w_i|\mathbf{x}_i; y_c; \theta)$ is as defined in (12), w_i is given in (3) and y_c is the censoring threshold.

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 .

Moreover, since

$$\begin{aligned}
\log [g(w_i|\mathbf{x}_i, y_c, \boldsymbol{\theta})] &= I(y_0 \leq w_i < y_c) \log f(w_i|\mathbf{x}_i, y_c, \boldsymbol{\theta}) \\
&\quad + I(w_i \geq y_c) \log [1 - F(y_c|\mathbf{x}_i, y_c, \boldsymbol{\theta})] \\
&= I_{\{y_0 \leq w_i < y_c\}} \{ \log \alpha(\mathbf{x}_i) + \alpha(\mathbf{x}_i) \log y_0 - [\alpha(\mathbf{x}_i) + 1] \log w_i \} \\
&\quad + I_{\{w_i = y_c\}} \{ \alpha(\mathbf{x}_i) [\log(y_0) - \log(y_c)] \}, \tag{14}
\end{aligned}$$

if we replace $\alpha(\mathbf{x}_i) = \exp(\mathbf{x}_i' \boldsymbol{\theta})$ the approximate negative log-likelihood function in (13) (omitting for simplicity of notation the terms not related to $\boldsymbol{\theta}$) becomes,

$$\begin{aligned}
\mathcal{K}_n^c(\boldsymbol{\theta}; y_c) &= \sum_{i=1}^n I_{\{y_0 \leq w_i < y_c\}} \left(\exp(\mathbf{x}_i' \boldsymbol{\theta}) \log \left(\frac{w_i}{y_0} \right) - \mathbf{x}_i' \boldsymbol{\theta} \right) \\
&\quad - \sum_{i=1}^n I_{\{w_i = y_c\}} \exp(\mathbf{x}_i' \boldsymbol{\theta}) \log \left(\frac{y_0}{y_c} \right). \tag{15}
\end{aligned}$$

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:

(A1) $n_0^{-1} \sum_{i=1}^n \mathbf{Z}_{ni} \mathbf{Z}_{ni}' I(w_i \geq y_0) = \Sigma_{y_0}^{-1/2} \widehat{\Sigma}_{y_0} \Sigma_{y_0}^{-1/2} \xrightarrow{p} \mathbf{I}_p$, where $\mathbf{Z}_{ni} := \Sigma_{y_0}^{-1/2} \mathbf{x}_i$,

\mathbf{I}_p is a $p \times p$ identity matrix and $\widehat{\Sigma}_{y_0} := n_0^{-1} \sum (\mathbf{x}_i \mathbf{x}_i') I(w_i \geq y_0)$.

(A2) (*Slowly varying function*) We assume that the remainder term $o(y^{-\beta(\mathbf{x})})$ satisfies

$$\sup_{\mathbf{x}} y^{\beta(\mathbf{x})} o(y^{-\beta(\mathbf{x})}) \rightarrow 0 \text{ as } y \rightarrow \infty.$$

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

Theorem 1. *Under Assumptions (A1) - (A2) it follows that*

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

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

Corollary 1. *Under the same conditions of Theorem 2.1 as $n \rightarrow \infty$ it follows that,*

$$T_j = n^{-1/2} \left(\Sigma_{y_0, jj}^{-1/2} \right) \Lambda^{-1/2} \widehat{\boldsymbol{\theta}}_j \xrightarrow{d} N(0, 1). \tag{16}$$

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

$$\bar{F}(y|x; \boldsymbol{\theta}) := P(Y > y|x; \boldsymbol{\theta}) = \left(\frac{y_0}{y_i} \right)^{\alpha(x)}.$$

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 $\bar{F}(y|x; \boldsymbol{\theta})$, consider

$$\delta := \frac{\bar{F}(y|\Delta x + x; \boldsymbol{\theta}) - \bar{F}(y|x; \boldsymbol{\theta})}{\bar{F}(y|x; \boldsymbol{\theta})} \times 100 \quad (17)$$

where y is an extreme value, say the $(1 - u)$ quantile, with $u \in (0, 1)$, such that,

$$y = (1 - u)^{\frac{1}{\alpha(x)}} y_0. \quad (18)$$

Hence, δ in (17) 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 $P(Y > y_0)$, 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 $\alpha(x) := \exp(\varphi(x))$, and $\varphi(x)$ is some function of x , we show in the appendix that

$$\delta(u) = \left[(1 - u)^{\varphi'(x)\Delta x} - 1 \right] \times 100. \quad (19)$$

For instance, in the multivariate case, $\varphi(\mathbf{x}) := \mathbf{x}'\boldsymbol{\beta}$, where \mathbf{x} is a $p \times 1$ vector of covariates, $\alpha(\mathbf{x}) = \exp(\mathbf{x}'\boldsymbol{\beta})$, and the impact of x_j on $\alpha(\mathbf{x})$ is,

$$\delta_j(u) = \left[(1 - u)^{\beta_j \Delta x} - 1 \right] \times 100. \quad (20)$$

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 $u = 0.15$ and $u = 0.20$, we are actually analysing the 96th and 97th quantile, respectively, of the unconditional distribution.

Remark 2. When the covariate considered is discrete (e.g. a dummy variable) a simple adaptation of (20) 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.

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,

$$y_i \sim D(\alpha(\mathbf{x}_i)) \quad (21)$$

$$\alpha(\mathbf{x}_i) = \exp(\beta_1 + \beta_2 x_i), \quad x_i \sim U(0, 1) \quad (22)$$

where $\beta_1 = \beta_2 = 1$ and the $k100\%$, $k \in (0, 1)$, largest observations of the empirical distribution $D(\cdot)$ 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 $D(\cdot)$ in (21) is either a Pareto or a Burr distribution⁵ 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 $[kn]$ 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.

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

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 1 provides the bias and RMSEs associated with the estimates of β_1 and β_2 in (22) 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 the these estimations approaches. In specific, the ratios are,

$$Ratio\ 1 = \frac{RMSE(\hat{\alpha}^c(\mathbf{x}_i))}{RMSE(\hat{\alpha}(\mathbf{x}_i))}, \quad Ratio\ 2 = \frac{RMSE(\tilde{\alpha}(\mathbf{x}_i))}{RMSE(\hat{\alpha}(\mathbf{x}_i))}.$$

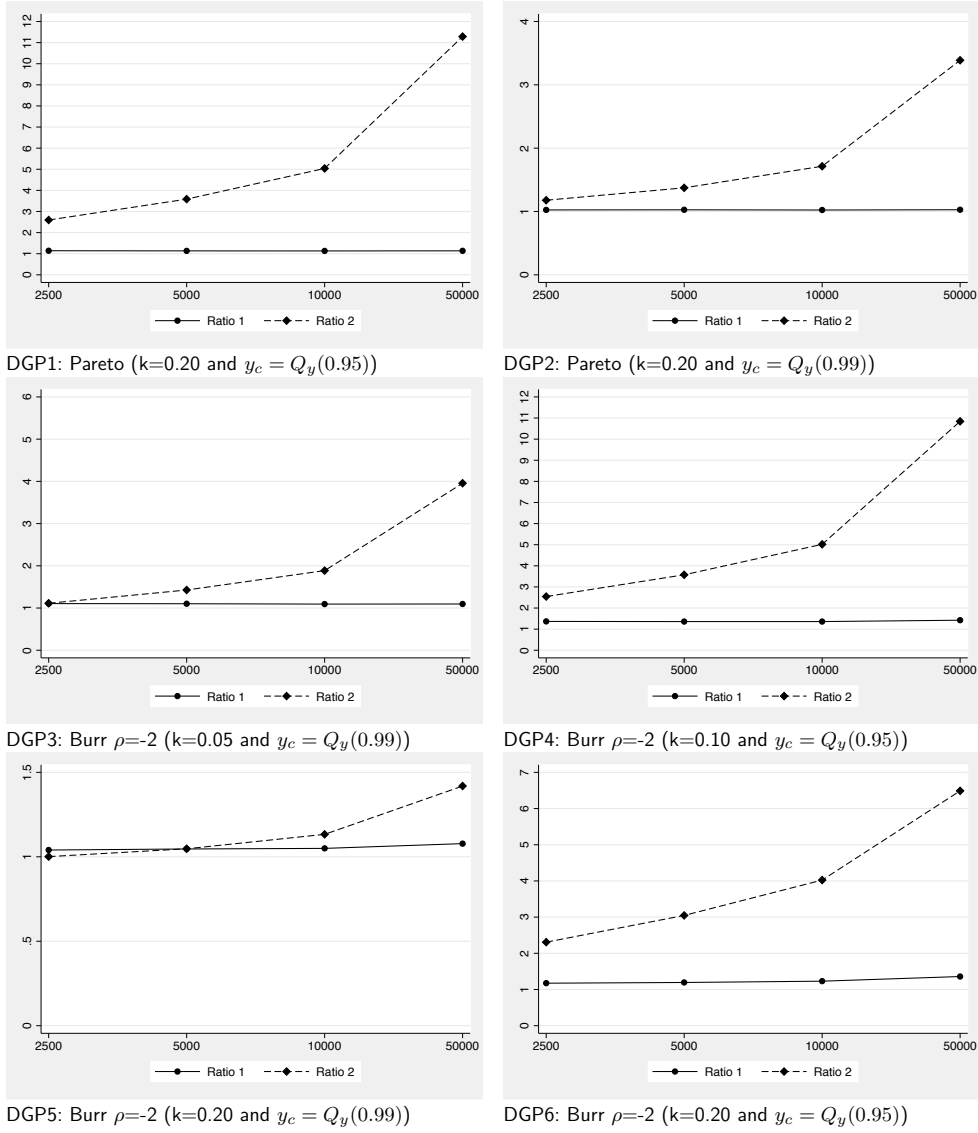
Since $\hat{\alpha}$, obtained as described in i), is the best estimator, *Ratio 1* and *Ratio 2* are larger than 1, across the different values of n . However, *Ratio 1* is just slightly above 1, which means that the censored estimator performs very well and mimics closely the behavior of the best estimator, $\hat{\alpha}$, although the former is based on the censored data. On the contrary, *Ratio 2* 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 *Ratio 2* tends to be larger (see, for example, the results for Case 4 in Table 1 and Figure 1).

Case 1: DGP Pareto (k=0.20 and $y_c = Q_y(0.95)$)												
	$\hat{\alpha}(x_i)$	$\hat{\alpha}^c(x_i)$	$\tilde{\alpha}(x_i)$	$\hat{\alpha}(x_i)$	$\hat{\alpha}^c(x_i)$	$\tilde{\alpha}(x_i)$	$\hat{\alpha}(x_i)$	$\hat{\alpha}^c(x_i)$	$\tilde{\alpha}(x_i)$	$\hat{\alpha}(x_i)$	$\hat{\alpha}^c(x_i)$	$\tilde{\alpha}(x_i)$
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 $y_c = Q_y(0.99)$)												
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 $\rho=-2$ (k=0.05 and $y_c = Q_y(0.99)$)												
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 $\rho=-2$ (k=0.10 and $y_c = Q_y(0.95)$)												
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 $\rho=-2$ (k=0.20 and $y_c = Q_y(0.99)$)												
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 $\rho=-2$ (k=0.20 and $y_c = Q_y(0.95)$)												
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: $\hat{\alpha}(x_i)$ is the tail index regression estimate considering the complete sample of data (with no censoring); $\hat{\alpha}^c(x_i)$ is the censored tail index regression estimate; and $\tilde{\alpha}(x_i)$ 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 1. Bias and RMSE of estimators



Note: $Ratio\ 1 = RMSE(\hat{\alpha}^c)/RMSE(\hat{\alpha})$ and $Ratio\ 2 = RMSE(\tilde{\alpha})/RMSE(\hat{\alpha})$. Ratio 1 compares the censored estimator with the best estimator $\hat{\alpha}$, while Ratio 2 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.

Figure 1: Ratios of the tail index estimates' RMSEs

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 ,

$$\hat{\tau}_1(y_c) = \frac{\hat{\alpha}_1}{\hat{\alpha}_1 - 1} y_c \quad (23)$$

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)

$$\hat{\tau}_2(y_c) = \frac{\hat{\alpha}_{Hill}^c}{\hat{\alpha}_{Hill}^c - 1} y_c \quad (24)$$

where $\hat{\alpha}_{Hill}^c$ is a consistent estimate of the tail index parameter computed as in (6). Note that $\tau_2(y_c) := E(y_i | y_i > y_c)$;

- iii) the imputed mean wage based on the method introduced in this paper

$$\hat{\tau}_3(y_c) = \frac{\hat{\alpha}(\mathbf{x}_i)}{\hat{\alpha}(\mathbf{x}_i) - 1} y_c \quad (25)$$

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

The main difference between $\hat{\tau}_3(y_c)$ and the other two approaches ($\hat{\tau}_1(y_c)$ and $\hat{\tau}_2(y_c)$) 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 $\hat{\tau}_3(y_c)$ corresponds to the optimal mean square predictor because $\tau_3(y_c)$ is the conditional expectation of y_i given \mathbf{x}_i and $y_i > y_c$. This follows from the well known result $E(y_i - E(y_i | \mathbf{x}_i, y_i > y_c))^2 \leq E(y_i - g(\cdot))^2$ where $g(\cdot)$ 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 $\tau_3(y_c)$ coincide.

Thus, having established the superiority of $\tau_3(y_c)$, it remains to be shown how much improvement is provided by $\tau_3(y_c)$ compared to $\tau_1(y_c)$ and $\tau_2(y_c)$ 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 , $i = 1, 2, \dots, n$ 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 $i = 1, 2, \dots, n$ simulate $\alpha(\mathbf{x}_i)$ 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 $\{w_i, i = 1, 2, \dots, n\}$ where $w_i = \min(y_i, y_c)$, from which we estimate $\hat{\tau}_1(y_c)$, $\hat{\tau}_2(y_c)$ and $\hat{\tau}_3(y_c)$.
4. The estimators $\hat{\tau}_1(y_c)$, $\hat{\tau}_2(y_c)$ and $\hat{\tau}_3(y_c)$ 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 $\hat{\tau}_1(y_c)$, $\hat{\tau}_2(y_c)$ and $\hat{\tau}_3(y_c)$ are 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 we present results only for the case $\alpha(\mathbf{x}_i) = \exp(1 + 2x_i)$ (note that $E(\alpha(\mathbf{x}_i)) = 3.7$). However, $\alpha(\mathbf{x}_i)$ should be set so that $\alpha(\mathbf{x}_i) > 1$, otherwise the conditional and marginal expected values do not exist, and consequently none of the above estimators will be well defined. The case $0 < \alpha(\mathbf{x}_i) < 1$ should be dealt with using other estimators, such as, for example, the median, $\hat{\alpha}(\mathbf{x}_i)\sqrt{2}y_c$.

Figure 2 illustrates our results. The lines represent two MSE ratios, *Ratio 1* := $MSE(\hat{\tau}_1)/MSE(\hat{\tau}_3)$ and *Ratio 2* := $MSE(\hat{\tau}_2)/MSE(\hat{\tau}_3)$, computed over different sample sizes ($n = (250, 500, 1000, 2000, 5000, 10000)$).

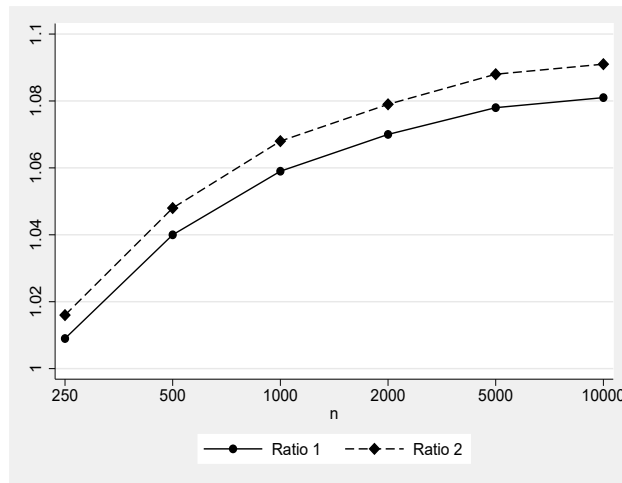


Figure 2: MSE Ratios Computed Over Different Sample Sizes

As can be observed from Figure 2, the $\hat{\tau}_3(y_c)$ estimator introduced in this paper produces the best results as all MSE ratios are larger than one. The gains are modest (between 1% and 2%) when the sample size is small ($n=250$), but they increase steadily as the sample size increases. Another conclusion, is that $\hat{\tau}_2$ (Armour *et al.*

(2016)) is better than the (naive) $\hat{\tau}_1$ estimator that does not accommodate the censoring.

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.^{6,7}

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 3 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.

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

7. 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.

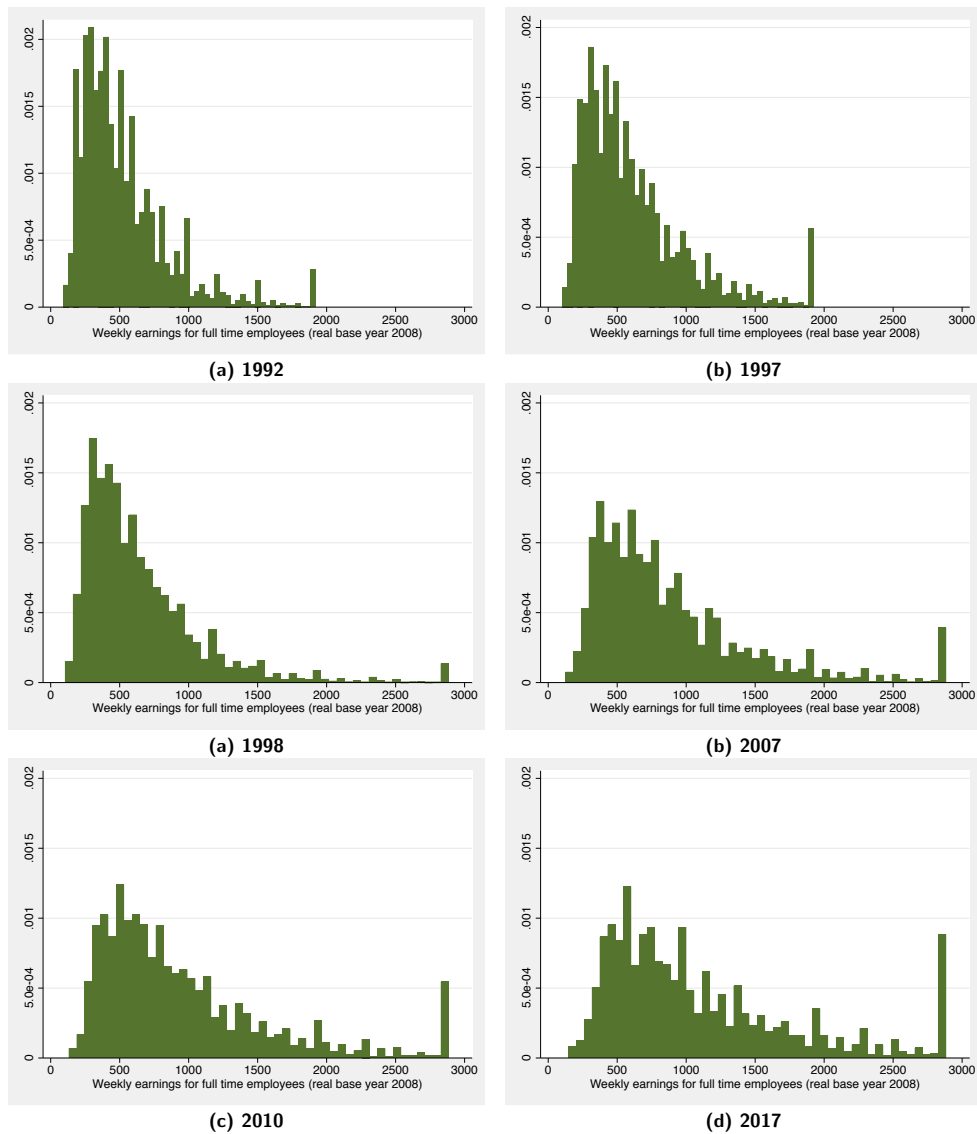


Figure 3: Annual unconditional histograms of the weekly wage.

The means and proportions of workers according to their characteristics, occupations and industry, for observations above the 80th percentile, are presented in Table 2. 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).

Mean	Year	
	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

^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.

Table 2. **Characteristics of individuals: means and proportions**

The results in Table 2 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).

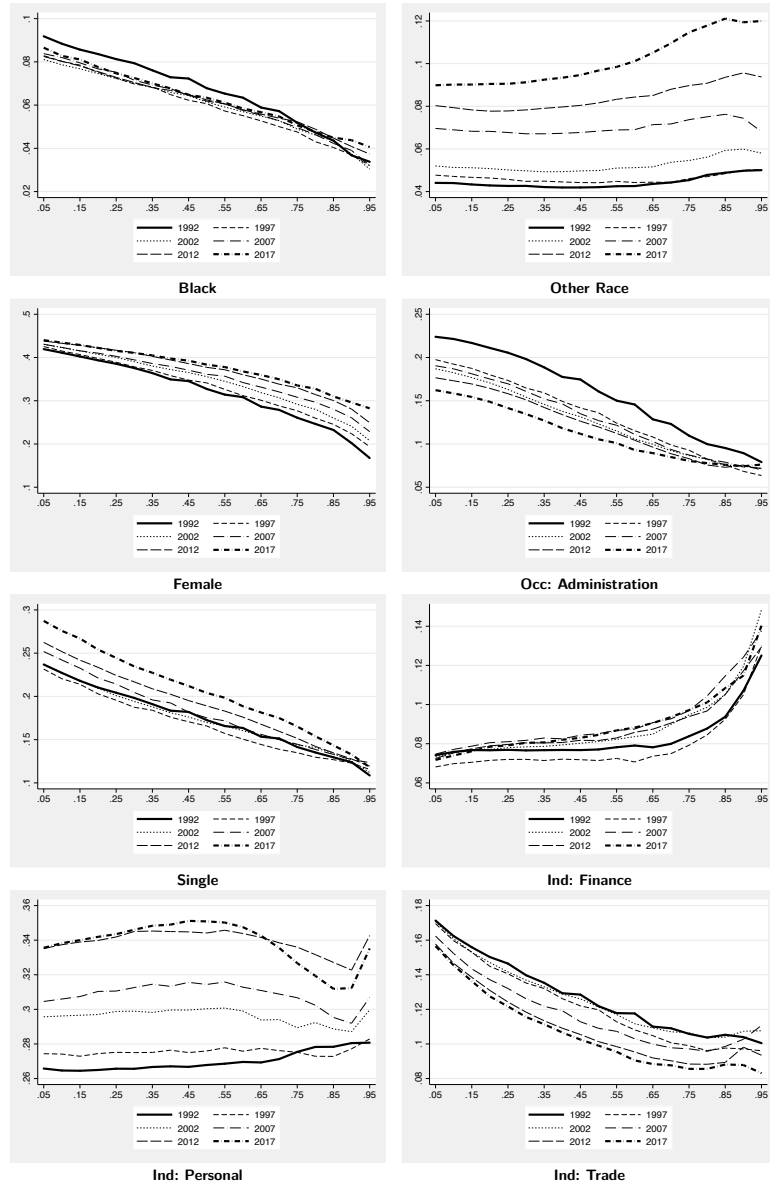
A further observation that can be made from the results in Table 2 is job polarization. A significant growth in employment in the right tail for non-routine cognitive tasks is observed in detriment of routine occupations (see also Autor (2019) and Goos and Manning (2007)). The decrease in the share of 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 2). 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.



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.

Figure 4: Proportion of individuals in different wage percentiles

4.2. Conditional tail index estimation results

Table 3 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.,

$$\hat{\alpha}_t = \frac{1}{n} \sum_{i=1}^n \hat{\alpha}(\mathbf{x}_{t,i}), \quad t = 1992, \dots, 2017 \quad (26)$$

where $\hat{\alpha}(\mathbf{x}_{t,i}) = \exp(\mathbf{x}'_{t,i} \hat{\theta}_t)$ (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 (26), $\hat{\alpha}_t$, provides an estimate of the unconditional tail index after considering the characteristics of all individuals in the sample for each year.

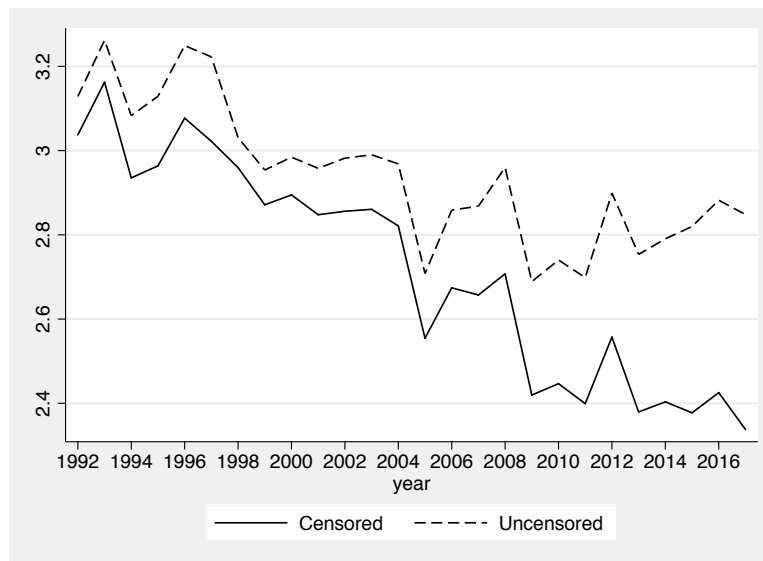
The results in Table 3 show that in general, the estimates based on the method that ignores censored data underestimate the true effects of the variables. This is especially clear in the estimates for 2017 (e.g. female and finance). It is a consequence of the potential inconsistency of the uncensored 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.

	Uncensored		Censored	
	1992	2017	1992	2017
Constant	2.597 *** (0.066)	2.321 *** (0.076)	2.723 *** (0.075)	2.812 *** (0.121)
Age	-0.011 *** (0.001)	-0.007 *** (0.001)	-0.013 *** (0.001)	-0.012 *** (0.001)
Female	0.272 *** (0.014)	0.134 *** (0.013)	0.309 *** (0.016)	0.229 *** (0.022)
Education	-0.076 *** (0.004)	-0.061 *** (0.004)	-0.085 *** (0.004)	-0.100 *** (0.006)
Race				
Black	0.085 *** (0.029)	0.004 (0.034)	0.098 *** (0.031)	0.010 *** (0.054)
Other race	-0.023 *** (0.025)	0.004 (0.016)	-0.031 (0.030)	0.015 (0.033)
Marital Status				
Married no spouse	0.075 (0.063)	-0.036 (0.046)	0.095 (0.070)	-0.100 (0.091)
Separated	0.035 (0.042)	0.051 (0.059)	0.044 (0.047)	0.093 (0.096)
Divorced	0.018 (0.019)	0.066 *** (0.022)	0.025 (0.022)	0.119 *** (0.036)
Widowed	0.003 (0.083)	-0.070 (0.052)	0.022 (0.077)	-0.083 (0.099)
Single	0.009 (0.018)	0.066 *** (0.018)	0.015 (0.021)	0.114 *** (0.029)
Occupations				
Administrative	0.113 *** (0.020)	0.021 (0.022)	0.125 *** (0.023)	0.006 (0.047)
Low Skill	0.098 (0.064)	0.093 (0.060)	0.111 * (0.067)	0.153 * (0.091)
Craft	0.289 *** (0.033)	0.069 (0.056)	0.324 *** (0.035)	0.154 * (0.086)
Operators	0.394 *** (0.050)	0.148 *** (0.062)	0.416 *** (0.053)	0.217 *** (0.090)
Transportation	0.366 *** (0.023)	0.184 *** (0.030)	0.400 *** (0.025)	0.264 *** (0.045)
Industry				
Construction	-0.202 *** (0.039)	-0.117 *** (0.042)	-0.234 *** (0.047)	-0.173 *** (0.071)
Manufacturing	0.003 (0.028)	-0.011 (0.029)	0.004 (0.031)	0.026 (0.052)
Transports	0.050 *** (0.021)	-0.016 (0.027)	0.051 ** (0.024)	-0.001 (0.040)
Trade	0.002 (0.020)	-0.052 ** (0.024)	0.001 (0.025)	-0.052 (0.043)
Finance	-0.126 *** (0.020)	-0.099 *** (0.022)	-0.184 *** (0.026)	-0.219 *** (0.043)
Repair	-0.059 *** (0.028)	-0.098 *** (0.022)	-0.071 ** (0.033)	-0.142 *** (0.041)
Personal	0.125 *** (0.016)	0.048 *** (0.018)	0.127 *** (0.019)	0.060 * (0.033)
Public	0.238 *** (0.022)	0.120 *** (0.024)	0.272 *** (0.025)	0.265 *** (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.

Table 3. Uncensored and Censored Tail Index Regression Results



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

Figure 5: Uncensored and censored right tail index estimates from 1992 to 2017

Figure 5 plots the tail index estimate computed as suggested in (26) 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 5 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

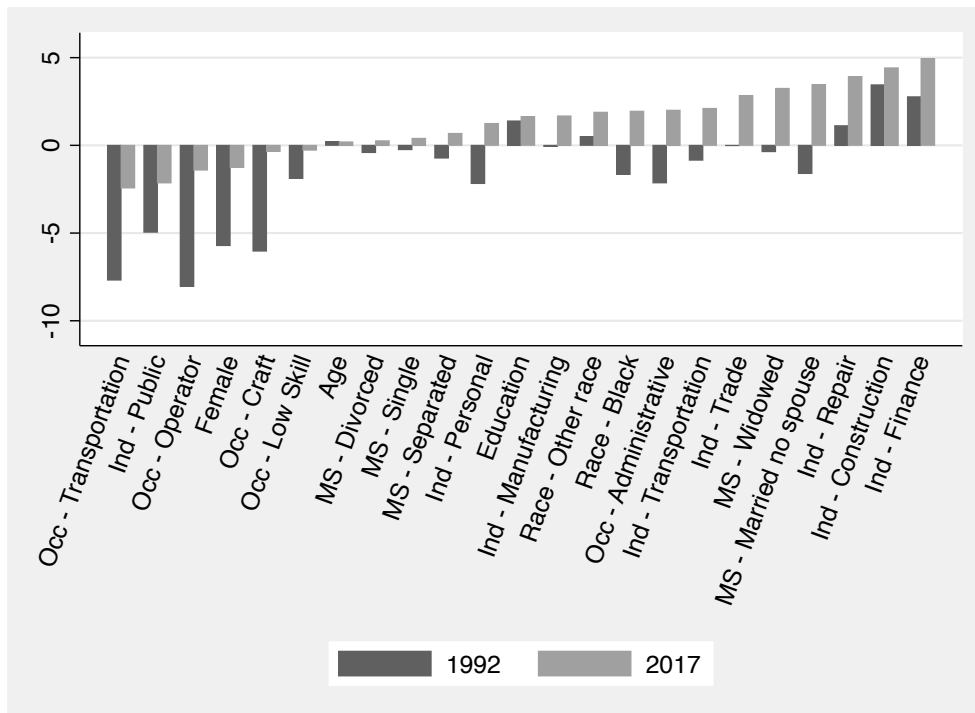
provides additional flexibility by allowing researchers to evaluate which determinants impact the probability of being in the right tail of the wage distribution and which determinants do not. Thus, permits a more detailed analysis on how inequality has spread across industries, occupation, gender and 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 6 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 2) 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).



Note: For a description of the occupations see note under Table 2.

Figure 6: Partial Effects of Variables when $u=0.15$

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 $E(y_i | y_i > y_c; \mathbf{x}_i)$ given by $\hat{\tau}_3(\mathbf{x}_i, y_c)$ in (25). However, for some individuals in the CPS database, especially those in the highest wage groups, $\hat{\alpha}(\mathbf{x}_i)$ can be in the neighborhood of 1, or even lower than 1, which implies that $E(y_i | y_i > y_c; \mathbf{x}_i)$ does not exist and, therefore, the estimate $\hat{\tau}_3(\mathbf{x}_i, y_c)$ is inadequate. For these cases, we use the conditional median of y_i given $y_i > y_c$ and \mathbf{x}_i , which is $2^{1/\hat{\alpha}(\mathbf{x}_i)}y_c$. In specific, to accommodate all situations (i.e. low and high values of $\hat{\alpha}(\mathbf{x}_i)$) we propose the following estimator,

$$\hat{\tau}_4(\mathbf{x}_i, y_c) = \begin{cases} 2^{1/\hat{\alpha}(\mathbf{x}_i)}y_c & 0 < \hat{\alpha}(\mathbf{x}_i) \leq c \\ \frac{\hat{\alpha}(\mathbf{x}_i)}{\hat{\alpha}(\mathbf{x}_i)-1}y_c & \hat{\alpha}(\mathbf{x}_i) > c \end{cases}. \quad (27)$$

In the empirical application of this statistic we set $c = 1.5$, since using $c = 1$ may lead to explosive estimates as $\frac{\hat{\alpha}(\mathbf{x}_i)}{\hat{\alpha}(\mathbf{x}_i)-1} \rightarrow \infty$ as $\hat{\alpha}(\mathbf{x}_i) \rightarrow 1^+$. 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 $\hat{\alpha}(\mathbf{x}_i) > 1.5$. Thus, for most cases, the estimate $\hat{\tau}_4(\mathbf{x}_i, y_c)$ coincides with the second branch of (27), which is the $\hat{\tau}_3(\mathbf{x}_i, y_c)$ estimator in (3.3). Hence, we use $\hat{\tau}_4(\mathbf{x}_i, y_c)$ to impute wages above the top-code over all individuals of the sample across time (see Figure 7).

Figure 7 illustrates the estimates of the imputed wages above the top-code computed from the different approaches discussed in (23), (24) and (27). The $\hat{\tau}_2$ and $\hat{\tau}_4$ estimates are similar. This result is expected given that for the overall analysis $\hat{\tau}_4$ is based on the values of $\hat{\alpha}_t$ computed as in (26), 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 $\hat{\tau}_3$ estimates will certainly be different from the $\hat{\tau}_2$ estimates (see next section).

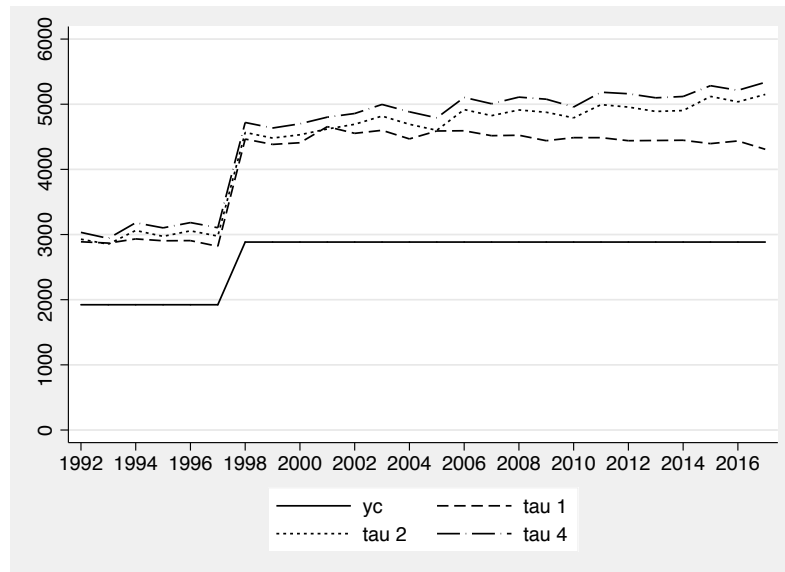


Figure 7: Prediction of Topcoded Wages

When the proportion of wages above the top-code is relatively small (as for example, from 1992 to 2005), the difference between $\hat{\tau}_1$, $\hat{\tau}_2$ and $\hat{\tau}_4$ 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 ($\hat{\tau}_1(y_c)$).

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 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 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 1.85 in 2017.

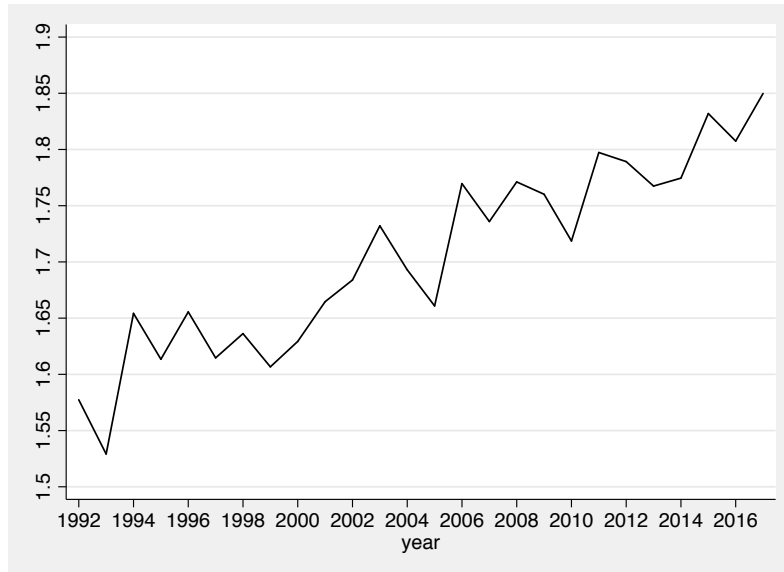


Figure 8: Top-coded wage adjustment factors between 1992 and 2017

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.

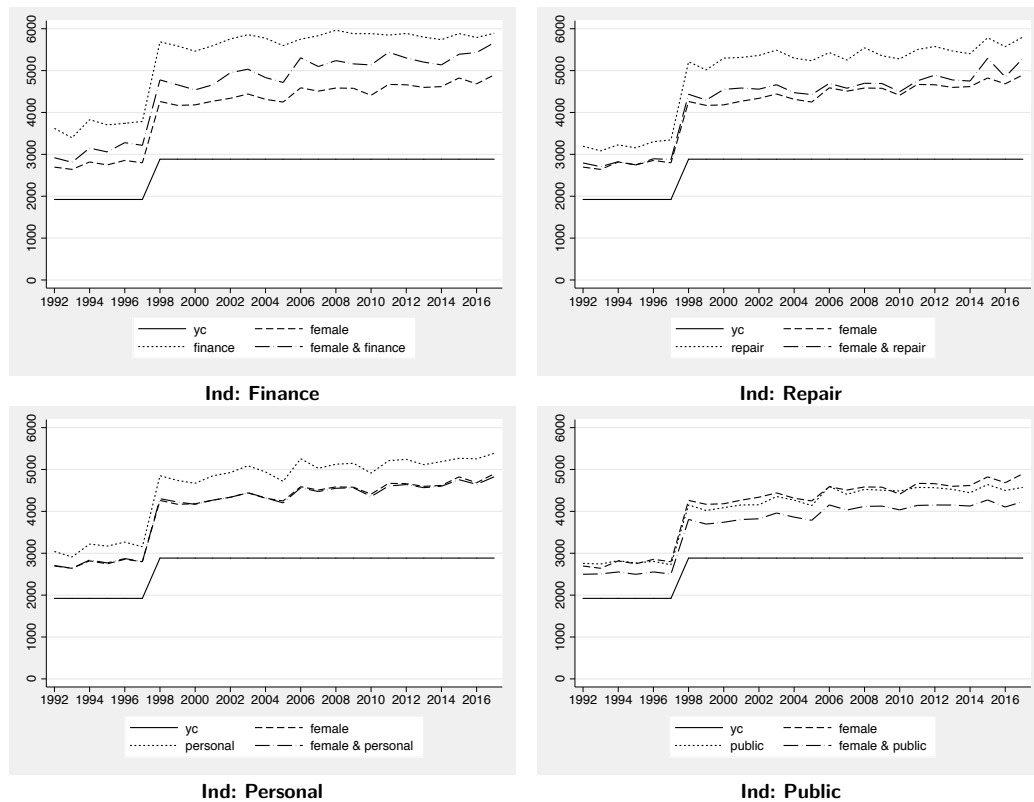


Figure 9: Imputed mean wages between 1992 and 2017 by industry

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.

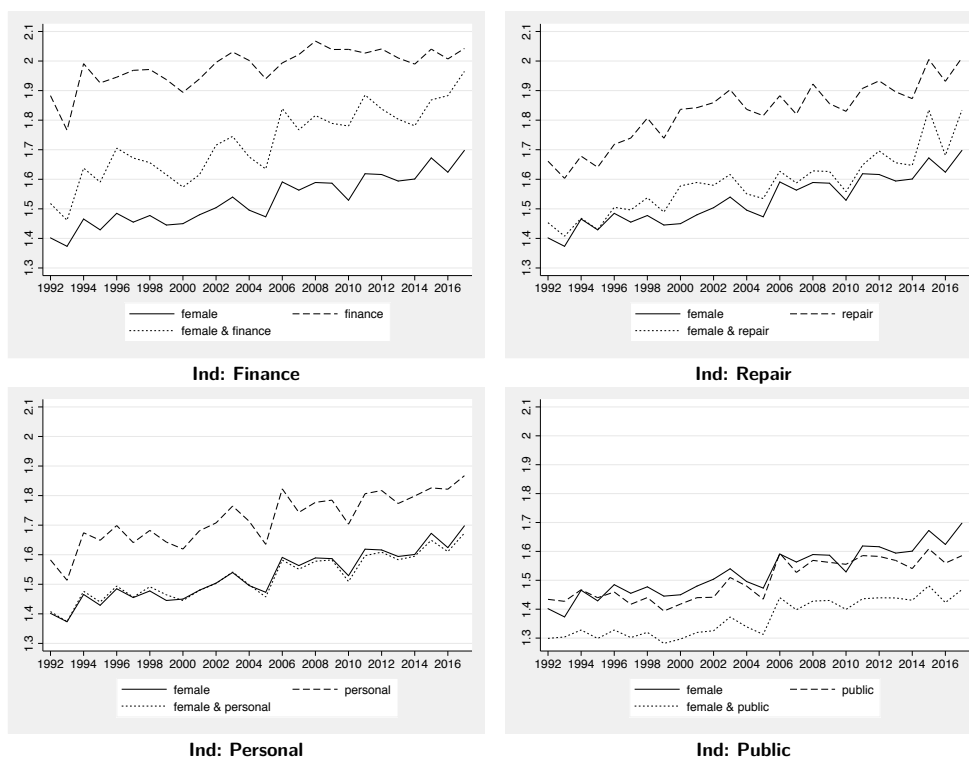


Figure 10: Adjustment factors between 1992 and 2017 by industry to impute topcoded wages

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 indepth 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.

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 between 1992 and 2017, but the impact is still towards an increase of alpha (although smaller in absolute values than in 1992) i.e. a decrease in the probability of an extreme wage. Furthermore, it is also noted that in comparison to an individual working in a non-routine cognitive occupation (managers) all occupations observed a positive contribution to observe an extreme value (although smaller in 2017). However, conclusions are different across occupations.

Our analysis also showed that women working in finance (public) would need a higher (lower) adjustment factor to impute the top-coded wages. Furthermore, 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.

References

- Acemoglu, Daron and David Autor (2011). *Skills, tasks and technologies: Implications for employment and earnings, Handbook of Labor Economics*, vol. 4, chap. 12, pp. 1043–1171. Elsevier, URL <https://ideas.repec.org/h/eee/labchp/5-12.html>.
- Armour, Philip, Richard V. Burkhauser, and Jeff Larrimore (2016). "Using the Pareto distribution to improve estimates of top-coded earnings." *Economic Inquiry*, 54(2), 1263–1273.
- Autor, D. and D. Dorn (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, David, Lawrence Katz, and Melissa Kearney (2006). "The Polarization Of The US Labor Market." *The American Economic Review*, 96(2), 189–194.
- Autor, David, Lawrence Katz, and Melissa Kearney (2008). "Trends in US Wage Inequality: Revisioning the Revisionists." *The Review of Economics and Statistics*, 90(2), 300–323.
- Autor, David H. (2019). "Work of the Past, Work of the Future." *AEA Papers and Proceedings*, 109, 1–32.
- Bernstein, Jared and Lawrence Mishel (1997). "Has wage inequality stopped growing?" *Monthly Labor Review*, pp. 3–16.
- Bordley, R.F., J.B. McDonald, and A. Mantrala (1995). "Something new, something old: parametric models for the size distribution of income." *Journal of Income Distribution*, 6, 91–103.

- Burkhauser, R. V. and J. Larrimore (2009). "Using internal CPS data to reevaluate trends in labor-earnings gaps." *Monthly Labor Review*, 132(8), 3–18.
- Burkhauser, Richard V., J. S. Butler, Shuaizhang Feng, and Andrew J. Houtenville (2004). "Long term trends in earnings inequality: what the CPS can tell us." *Economics Letters*, 82(2), 295–299.
- Card, David and John E. DiNardo (2002). "Skill-biased technological change and rising wage inequality: Some problems and puzzles." *Journal of Labor Economics*, 20(4), 733–783.
- Feng, Shuaizhang, Richard V. Burkhauser, and J. S. Butler (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, Maarten and Alan Manning (2007). "Lousy and Lovely Jobs: The Rising Polarization of Work in Britain." *The Review of Economics and Statistics*, 89(1), 118–133.
- Goos, Maarten, Alan Manning, and Anna Salomons (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.T. and S.H. Shore (2015). "Changes in the distribution of earnings volatility." *Journal of Human Resources*, 3(50), 811–836.
- Kalleberg, Arne L. (2011). *Good jobs, bad jobs: The rise of polarized and precarious employment systems in the United States, 1970s-2000s*. Russell Sage Foundation.
- Katz, Lawrence and Kevin Murphy (1992). "Changes in relative wages, 1963 – 1987: Supply and demand factors." *The Quarterly Journal of Economics*, 107(1), 35–78.
- Larrimore, J., R.V. Burkhauser, S. Feng, and L. Zayatz (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, Thomas (2006). "Increasing residual wage inequality: Composition effects, noisy data, or rising demand for skill?" *American Economic Review*, 96(3), 461–498.
- Levy, Frank and Richard Murnane (1992). "US Earnings Levels and Earnings Inequality: A Review of Recent Trends and Proposed Explanations." *Journal of Economic Literature*, 30(3), 1333–81.

- Ma, Yaolan, Yuexiang Jiang, and Wei Huang (2019). "Tail index varying coefficient model." *Communications in Statistics - Theory and Methods*, 48(2), 235–256.
- Macpherson, David A and Barry T Hirsch (1995). "Wages and gender composition: Why do women's jobs pay less?" *Journal of Labor Economics*, 13(3), 426–471.
- Mishel, L., J. Bernstein, and H. Shierholz (2013). *The State of Working America*. Ithaca, NY, 12th edition ed., Cornell University Press.
- Nicolau, J. and P. M. M. Rodrigues (2019). "A New Regression-Based Tail Index Estimator." *The Review of Economics and Statistics*, 101(4), 667–680.
- Parker, S.C. (1999). "The generalized beta as a model for the distribution of earnings." *Economics Letters*, 62, 197–200.
- Piketty, Thomas and Emmanuel Saez (2003). "Income inequality in the United States, 1913–1998." *The Quarterly Journal of Economics*, 118(1), 1–41.
- Wang, Hansheng and Chih-Ling Tsai (2009). "Tail Index Regression." *Journal of the American Statistical Association*, 104(487), 1233–1240.

Appendix A.: Technical Appendix

A.1 Proof of Theorem 2.1

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

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

where

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

Thus, considering (A.1) we can show that,

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

Moreover, it follows that

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

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

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

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

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

Given these results, considering that $\hat{\theta}$ 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

$$\frac{\dot{\mathcal{K}}^c(\theta, y_c)}{n} = \frac{\dot{\mathcal{K}}^c(\theta_0, y_c)}{n} + \frac{\ddot{\mathcal{K}}^c(\theta_1, y_c)}{n} (\hat{\theta} - \theta_0) \quad (\text{A.2})$$

for some $\theta_1 \in [\hat{\theta}, \theta_0]$. ■

A.2 Computation of the partial effects

We have

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

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

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

Now and given the specification $\alpha(x) = \exp(\varphi(x))$ where $\varphi(x)$ is usually of type $\varphi(x) = \mathbf{x}'\beta$ it follows that,

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

Therefore

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

In conclusion

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

■

Appendix B: Additional figures on proportion of individuals, according to characteristics, occupation and industry

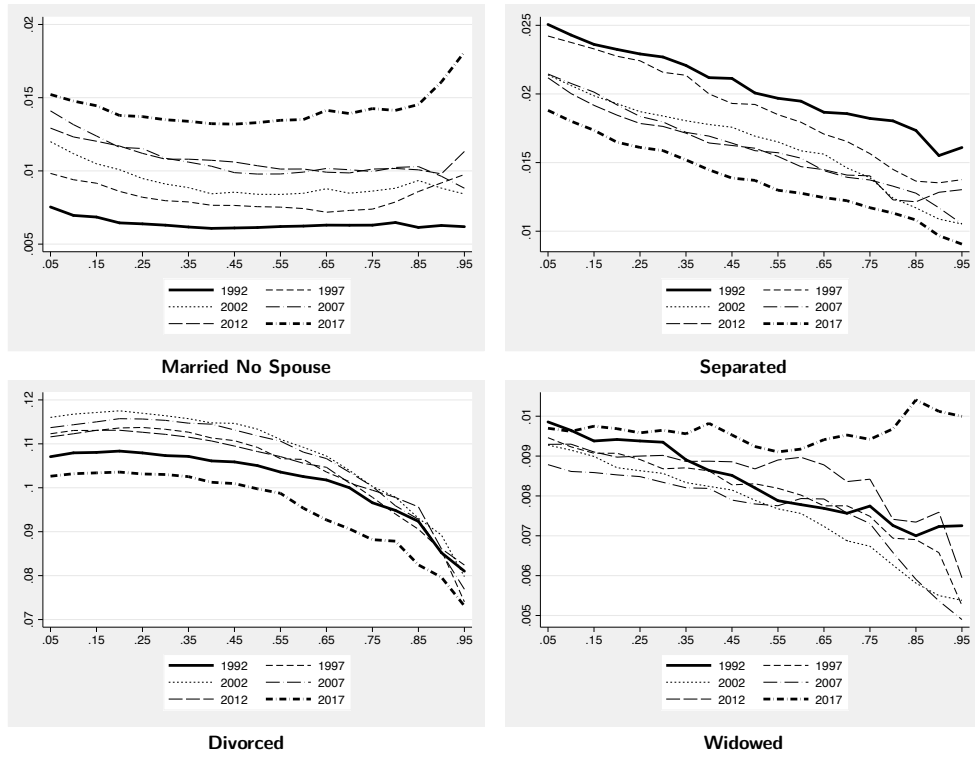


Figure B.1: Marital Status

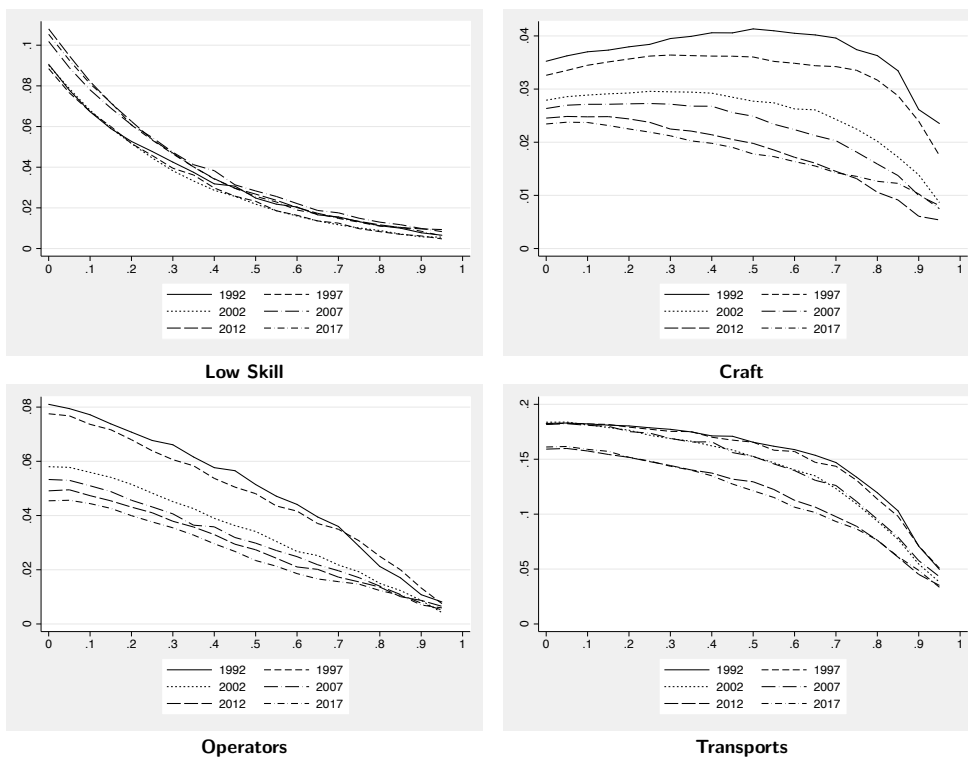


Figure B.2: Occupation

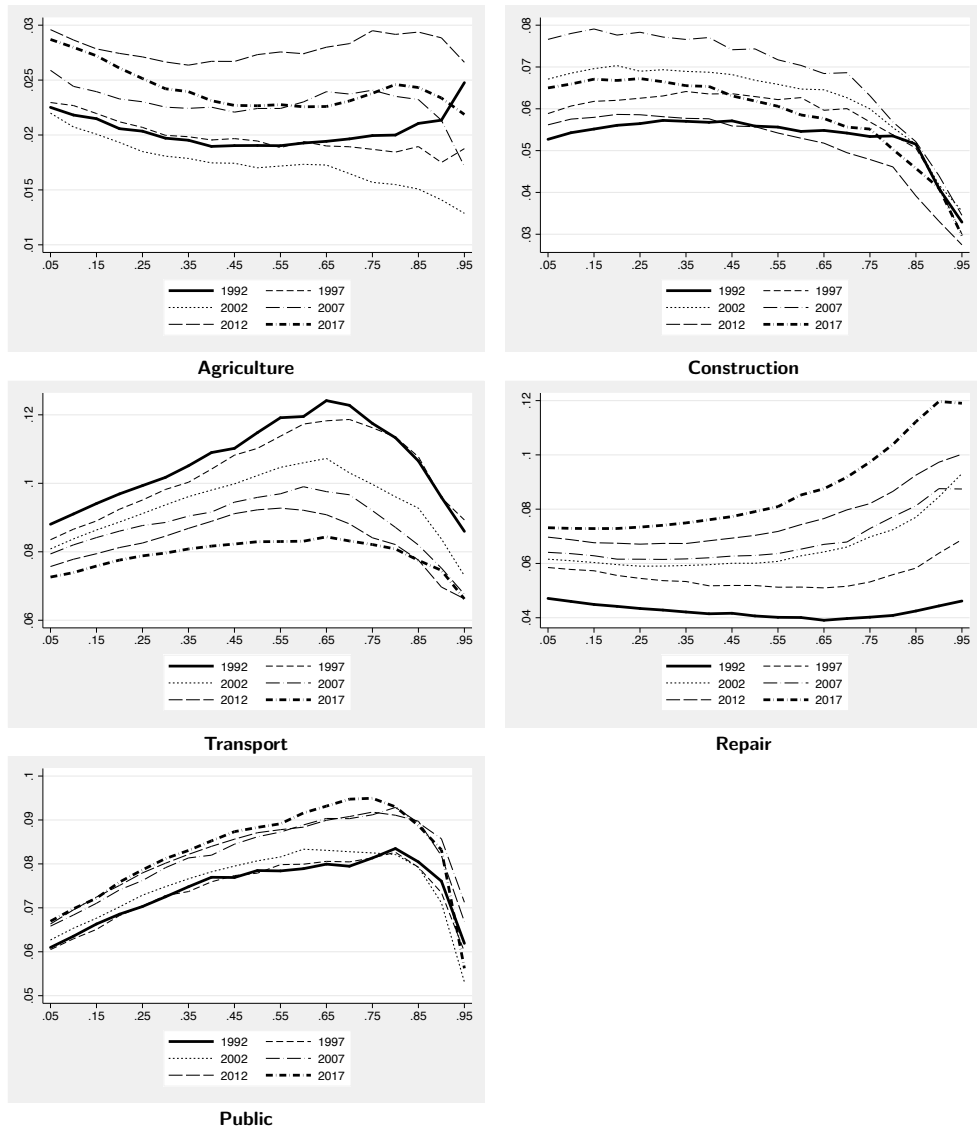


Figure B.3: Industry

Appendix C.: Industry classification

Table C.1 - 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.

Working Papers

2017

- 1|17 The diffusion of knowledge via managers' mobility
Giordano Mion | Luca David Opromolla | Alessandro Sforza
- 2|17 Upward nominal wage rigidity
Paulo Guimarães | Fernando Martins | Pedro Portugal
- 3|17 Zooming the ins and outs of the U.S. unemployment
Pedro Portugal | António Rua
- 4|17 Labor market imperfections and the firm's wage setting policy
Sónia Félix | Pedro Portugal
- 5|17 International banking and cross-border effects of regulation: lessons from Portugal
Diana Bonfim | Sónia Costa
- 6|17 Disentangling the channels from birthdate to educational attainment
Luís Martins | Manuel Coutinho Pereira
- 7|17 Who's who in global value chains? A weighted network approach
João Amador | Sónia Cabral | Rossana Mastrandrea | Franco Ruzzenenti
- 8|17 Lending relationships and the real economy: evidence in the context of the euro area sovereign debt crisis
Luciana Barbosa
- 9|17 Impact of uncertainty measures on the Portuguese economy
Cristina Manteu | Sara Serra
- 10|17 Modelling currency demand in a small open economy within a monetary union
António Rua
- 11|17 Boom, slump, sudden stops, recovery, and policy options. Portugal and the Euro
Olivier Blanchard | Pedro Portugal
- 12|17 Inefficiency distribution of the European Banking System
João Oliveira
- 13|17 Banks' liquidity management and systemic risk
Luca G. Deidda | Ettore Panetti
- 14|17 Entrepreneurial risk and diversification through trade
Federico Esposito
- 15|17 The portuguese post-2008 period: a narrative from an estimated DSGE model
Paulo Júlio | José R. Maria
- 16|17 A theory of government bailouts in a heterogeneous banking system
Filomena Garcia | Ettore Panetti
- 17|17 Goods and factor market integration: a quantitative assessment of the EU enlargement
FLorenzo Caliendo | Luca David Opromolla | Fernando Parro | Alessandro Sforza

2018

- 1|18 Calibration and the estimation of macro-economic models
Nikolay Iskrev
- 2|18 Are asset price data informative about news shocks? A DSGE perspective
Nikolay Iskrev
- 3|18 Sub-optimality of the friedman rule with distorting taxes
Bernardino Adão | André C. Silva
- 4|18 The effect of firm cash holdings on monetary policy
Bernardino Adão | André C. Silva
- 5|18 The returns to schooling unveiled
Ana Rute Cardoso | Paulo Guimarães | Pedro Portugal | Hugo Reis
- 6|18 Real effects of financial distress: the role of heterogeneity
Francisco Buera | Sudipto Karmakar
- 7|18 Did recent reforms facilitate EU labour market adjustment? Firm level evidence
Mario Izquierdo | Theodora Kosma | Ana Lamo | Fernando Martins | Simon Savsek
- 8|18 Flexible wage components as a source of wage adaptability to shocks: evidence from European firms, 2010–2013
Jan Babecký | Clémence Berson | Ludmila Fadejeva | Ana Lamo | Petra Marotzke | Fernando Martins | Pawel Strzelecki
- 9|18 The effects of official and unofficial information on tax compliance
Filomena Garcia | Luca David Opromolla Andrea Vezulli | Rafael Marques
- 10|18 International trade in services: evidence for portuguese firms
João Amador | Sónia Cabral | Birgitte Ringstad
- 11|18 Fear the walking dead: zombie firms, spillovers and exit barriers
Ana Fontoura Gouveia | Christian Osterhold
- 12|18 Collateral Damage? Labour Market Effects of Competing with China – at Home and Abroad
Sónia Cabral | Pedro S. Martins | João Pereira dos Santos | Mariana Tavares
- 13|18 An integrated financial amplifier: The role of defaulted loans and occasionally binding constraints in output fluctuations
Paulo Júlio | José R. Maria
- 14|18 Structural Changes in the Duration of Bull Markets and Business Cycle Dynamics
João Cruz | João Nicolau | Paulo M.M. Rodrigues
- 15|18 Cross-border spillovers of monetary policy: what changes during a financial crisis?
Luciana Barbosa | Diana Bonfim | Sónia Costa | Mary Everett
- 16|18 When losses turn into loans: the cost of undercapitalized banks
Laura Blattner | Luísa Farinha | Francisca Rebelo
- 17|18 Testing the fractionally integrated hypothesis using M estimation: With an application to stock market volatility
Matei Demetrescu | Paulo M. M. Rodrigues | Antonio Rubia

- 18|18 Every cloud has a silver lining: Micro-level evidence on the cleansing effects of the Portuguese financial crisis
Daniel A. Dias | Carlos Robalo Marques
- 19|18 To ask or not to ask? Collateral versus screening in lending relationships
Hans Degryse | Artashes Karapetyan | Sudipto Karmakar
- 20|18 Thirty years of economic growth in Africa
João Amador | António R. dos Santos
- 21|18 CEO performance in severe crises: the role of newcomers
Sharmin Sazedj | João Amador | José Tavares
- 22|18 A general equilibrium theory of occupational choice under optimistic beliefs about entrepreneurial ability
Michele Dell’Era | Luca David Opromolla | Luís Santos-Pinto
- 23|18 Exploring the implications of different loan-to-value macroprudential policy designs
Rita Basto | Sandra Gomes | Diana Lima
- 24|18 Bank shocks and firm performance: new evidence from the sovereign debt crisis
Luísa Farinha | Marina-Eliza Spaliara | Serafem Tsoukas
- 25|18 Bank credit allocation and productivity: stylised facts for Portugal
Nuno Azevedo | Márcio Mateus | Álvaro Pina
- 26|18 Does domestic demand matter for firms’ exports?
Paulo Soares Esteves | Miguel Portela | António Rua
- 27|18 Credit Subsidies
Isabel Correia | Fiorella De Fiore | Pedro Teles | Oreste Tristani

2019

- 1|19 The transmission of unconventional monetary policy to bank credit supply: evidence from the TLTRO
António Afonso | Joana Sousa-Leite
- 2|19 How responsive are wages to demand within the firm? Evidence from idiosyncratic export demand shocks
Andrew Garin | Filipe Silvério
- 3|19 Vocational high school graduate wage gap: the role of cognitive skills and firms
Joop Hartog | Pedro Raposo | Hugo Reis
- 4|19 What is the Impact of Increased Business Competition?
Sónia Félix | Chiara Maggi
- 5|19 Modelling the Demand for Euro Banknotes
António Rua
- 6|19 Testing for Episodic Predictability in Stock Returns
Matei Demetrescu | Iliyan Georgiev
Paulo M. M. Rodrigues | A. M. Robert Taylor
- 7|19 The new ESCB methodology for the calculation of cyclically adjusted budget balances: an application to the Portuguese case
Cláudia Braz | Maria Manuel Campos
Sharmin Sazedj
- 8|19 Into the heterogeneities in the Portuguese labour market: an empirical assessment
Fernando Martins | Domingos Seward
- 9|19 A reexamination of inflation persistence dynamics in OECD countries: A new approach
Gabriel Zsurkis | João Nicolau | Paulo M. M. Rodrigues
- 10|19 Euro area fiscal policy changes: stylised features of the past two decades
Cláudia Braz | Nicolas Carnots
- 11|19 The Neutrality of Nominal Rates: How Long is the Long Run?
João Valle e Azevedo | João Ritto | Pedro Teles
- 12|19 Testing for breaks in the cointegrating relationship: on the stability of government bond markets' equilibrium
Paulo M. M. Rodrigues | Philipp Sibbertsen
Michelle Voges
- 13|19 Monthly Forecasting of GDP with Mixed Frequency Multivariate Singular Spectrum Analysis
Hossein Hassani | António Rua | Emmanuel Sirimal Silva | Dimitrios Thomakos
- 14|19 ECB, BoE and Fed Monetary-Policy announcements: price and volume effects on European securities markets
Eurico Ferreira | Ana Paula Serra
- 15|19 The financial channels of labor rigidities: evidence from Portugal
Edoardo M. Acabbi | Ettore Panetti | Alessandro Sforza
- 16|19 Sovereign exposures in the Portuguese banking system: determinants and dynamics
Maria Manuel Campos | Ana Rita Mateus | Álvaro Pina
- 17|19 Time vs. Risk Preferences, Bank Liquidity Provision and Financial Fragility
Ettore Panetti

- 18|19 Trends and cycles under changing economic conditions
Cláudia Duarte | José R. Maria | Sharmin Sazedj
- 19|19 Bank funding and the survival of start-ups
Luísa Farinha | Sónia Félix | João A. C. Santos
- 20|19 From micro to macro: a note on the analysis of aggregate productivity dynamics using firm-level data
Daniel A. Dias | Carlos Robalo Marques
- 21|19 Tighter credit and consumer bankruptcy insurance
António Antunes | Tiago Cavalcanti | Caterina Mendicino | Marcel Peruffo | Anne Villamil

2020

- 1|20 On-site inspecting zombie lending
Diana Bonfim | Geraldo Cerqueiro | Hans Degryse | Steven Ongena
- 2|20 Labor earnings dynamics in a developing economy with a large informal sector
Diego B. P. Gomes | Felipe S. Iachan | Cezar Santos
- 3|20 Endogenous growth and monetary policy: how do interest-rate feedback rules shape nominal and real transitional dynamics?
Pedro Mazedo Gil | Gustavo Iglésias
- 4|20 Types of International Traders and the Network of Capital Participations
João Amador | Sónia Cabral | Birgitte Ringstad
- 5|20 Forecasting tourism with targeted predictors in a data-rich environment
Nuno Lourenço | Carlos Melo Gouveia
António Rua
- 6|20 The expected time to cross a threshold and its determinants: A simple and flexible framework
Gabriel Zsurkis | João Nicolau | Paulo M. M. Rodrigues
- 7|20 A non-hierarchical dynamic factor model for three-way data
Francisco Dias | Maximiano Pinheiro; António Rua
- 8|20 Measuring wage inequality under right censoring
João Nicolau | Pedro Raposo | Paulo M. M. Rodrigues

