School of Economics Working Paper
2023-01

Copula-based estimation of health inequality measures with an application to COVID-19

Taoufik Bouezmarni *
Mohamed Doukali **
Abderrahim Taamouti ***

* Université de Sherbrooke
** School of Economics, University of East Anglia
*** University of Liverpool

University of East Anglia

SCHOOL OF ECONOMICS

School of Economics
University of East Anglia
Norwich Research Park
Norwich NR4 7TJ
United Kingdom
www.uea.ac.uk/economics

Copula-based estimation of health inequality measures with an application to COVID-19[*]

Taoufik Bouezmarni[†]
Université de Sherbrooke

Mohamed Doukali[‡]
University of East Anglia

Abderrahim Taamouti[§]
University of Liverpool

January 26, 2023

[*] The authors would like to thank very much Serena Ng and the participants at the 2022 CISEM International Symposium on Statistics and Econometrics, 2022 Africa Meeting of the Econometric Society, and INSEA for their very useful comments.

[†] Département de Mathématiques, Université de Sherbrooke, CIREQ, CREAS. Email addresses: Taoufik.Bouezmarni@USherbrooke.ca

[‡] University of East Anglia, School of Economics, Norwich Research Park, NR4 7TJ, Norwich, United Kingdom email: m.doukali@uea.ac.uk

[§] *Corresponding author.* Department of Economics, University of Liverpool Management School. Address: Chatham St, Liverpool L69 7ZH. E-mail: Abderrahim.Taamouti@liverpool.ac.uk.

Copula-based estimation of health inequality measures with an application to COVID-19

ABSTRACT

This paper aims to use copulas to derive alternative estimators of Health Concentration Curve [hereafter CH] and Gini coefficient for health distribution. We motivate the importance of expressing health inequality measures in terms of copula, which we in turn use to build copula-based semi- and non-parametric estimators of the above measures. Thereafter, we study the asymptotic properties of these estimators. In particular, we establish their consistency and asymptotic normality. We provide expressions for their variances, which can be used to construct confidence intervals and build tests for health concentration curve and Gini health coefficient. A Monte-Carlo simulation exercise shows that the semiparametric estimator outperforms the smoothed nonparametric estimator, and that the latter does better than the empirical estimator in terms of Mean Squared Error. We also run an extensive empirical study where we apply our CH and Gini health coefficient estimators to show that the inequalities across U.S. states' socioeconomic variables like income/poverty and race/ethnicity explain the observed inequalities in the U.S. COVID-19's infections and deaths.

Keywords: Health concentration curve, Gini health coefficient, inequality, copula, semi- and non-parametric estimators, COVID-19 infections and deaths.

Journal of Economic Literature classification: C13, C14, I14.

1 Introduction

COVID-19 has created an unprecedented global health crisis that caused millions of deaths worldwide. However, many argue that pre-existing social inequalities led to inequalities in the number of COVID-19's infections and deaths across social classes, with the most-deprived classes are worst hit. In a 2020 report on unequal risks of infection and severe illness, World Health Organization (WHO) European Region pointed out that COVID-19 exposure risk and the severity of its health, social and economic impacts are not being felt equally.[1] Policymakers were urged to identify those classes that are most at risk of infection and dying from the virus and set social and economic measures to deal with these inequities in COVID-19's infections and deaths. In this paper, we aim to develop alternative estimators that can help quantify health inequalities caused by socioeconomic factors such as income, poverty, race, etc. These estimators will be used to check if real data on COVID-19's infection and deaths supports the claims regarding the impact of socioeconomic factors on COVID-19 exposure risk and deaths.

Chen and Krieger (2021) argue that reporting disaggregated COVID-19 cases by race, ethnicity and other socioeconomic factors is vital to informing efforts to distribute resources, develop treatments, and coordinate public policy. Using data from Public Health Disparities Geocoding Project, Chen and Krieger (2021) utilize descriptive statistics to report disparities in COVID-19 death rate in the US by county level sociodemographic attributes using available surveillance and US Census data. Chin, et al. (2020) argue that preliminary data from the pandemic in the US have shown that demographic and socioeconomic issues make low-income communities and people of color more vulnerable to COVID-19 than others. To assess each county's risk of high COVID-19 medical burden, Chin, et al. (2020) examine available data for a range of characteristics for all U.S. counties or county equivalents.

Moreover, Knittel and Ozaltun (2020) use linear regressions and negative binomial mixed models to study the correlation between counties COVID-19 death rates and some socioeconomic variables across U.S. states as well as within states. Their analysis shows that higher shares of African American and elderly residents across states are correlated with higher death rates. However, within a given state, this correlation becomes statistically insignificant, while remaining positive. Unexpectedly, they did not find any correlation between poverty rates and COVID-19 death rates.

[1] See the WHO's report here: https://apps.who.int/iris/bitstream/handle/10665/338199/WHO-EURO-2020-1744-41495-56594-eng.pdf

Using linear regressions and monthly county-level mortality data, McLaren (2021) shows that there is a strong positive correlation across counties between the minority's population share and COVID-19 deaths. For further empirical evidence on the impact of socioeconomic variables on COVID-19 infections and deaths in countries other than U.S., readers can consult Ehlert (2021), Bermudi et al. (2021), Lassale et al. (2020), among others.

The above-mentioned studies, however, are based on simple correlations and regressions, which could be misleading in terms of detecting health inequalities in COVID-19's infections and deaths [see Section (3)]. To measure health inequality with respect to socioeconomic variables, Wagstaff et al. (1989) introduced a concentration index that depends on a weighting function that represents the aversion to socioeconomic health inequality. This concentration index is an extension of the Gini index that is widely adopted in the income inequality literature. Since the work of Wagstaff et al. (1989), several alternative indices based on concentration curves have been developed using different weighting functions that correspond to different judgements of inequality aversion, see Wagstaff et al. (1991), Wagstaff (2002, 2005), Allison and Foster (2004), Erreygers and Van Ourti (2011), Zheng (2011) among others.

In this paper, we use parametric and nonparametric copulas to develop semiparametric and nonparametric estimators of health concentration curves and Gini coefficient for health distribution [hereafter Gini health coefficient] that are designed to quantify health inequalities. These estimators will be utilized to study health inequalities in COVID-19's infections and deaths across and within U.S. states. Health concentration curves are used to plot the cumulative percentage of the health variable (e.g. COVID-19's infection and death rates) against the cumulative percentage of the population, ranked by socioeconomic covariates such as living standards (e.g. beginning with the poorest and ending with the richest), race, etc. The plot helps visualize inequalities in health variables by observing the position of the health concentration curve with respect to the 45-degree line (known as the line of equality) in a two-dimensional space. If individuals in the dataset had equal chance of contracting/dying from COVID-19 regardless of their socioeconomic position, then as we move from lowest socioeconomic position to highest socioeconomic position, the proportion of individuals with/died of COVID-19 should remain the same, and in this case the health concentration curve matches the 45-degree line. Otherwise, we say that there is inequality not in favour of individuals with lowest socioeconomic position if the curve is above the 45-degree line, or not in favour of individuals with highest socioeconomic position if the curve is below the 45-degree line. The farther the curve is above/below the line of equality, the more concentrated the COVID-19's infections/deaths is among the individuals with lowest/highest socioeconomic position, respectively.

We first motivate the importance of expressing health inequality measures in terms of copula of health and socioeconomic variables. We re-formulate health concentration curve as a function of copula and derive expressions of CH for some specific copula functions that represent independence case, and lower and upper Fréchet–Hoeffding dependence cases, which help illustrate the link between health concentration curve and Lorenz curve of the health variable.

Thereafter, we use the copula-based re-formulation of CH to build semi- and nonparametric estimators of health concentration curves. To derive the semi-parametric estimator, a parametric copula is used to model the dependence between health and socioeconomic variables. The parameter of the copula function is estimated using the maximum pseudo-likelihood estimation method after replacing the cumulative distribution of the health variable by its empirical estimator. Once the maximum likelihood estimator of the copula's parameter is obtained, we use it to calculate the semi-parametric estimator of health concentration curve. For the non-parametric estimator, we replace the copula function by its smoothed non-parametric estimator that represent a Bernstein copula. Bernstein estimator helps avoid the misspecification problem that might face the semiparametric estimator. Furthermore, we use the above estimators of health concentration curves to derive semi-parametric and nonparametric estimators of Gini health coefficient.

Moreover, we study the asymptotic properties of the above estimators. We establish their consistency and asymptotic normality. We provide expressions for their variances, which can be used to construct confidence intervals and build tests for health concentration curve and Gini health coefficient. A Monte-Carlo simulation exercise based on several data-generating processes and sample sizes shows that the semiparametric estimator outperforms the smoothed nonparametric estimator, and that the latter does better than the empirical estimator in terms of Integrated Mean Squared Error. Furthermore, we run an extensive empirical study to illustrate the importance of CH and Gini health index estimators for quantifying and investigating inequalities in COVID-19's infections and deaths in the U.S. Our results show that socioeconomic variables like income, poverty, and race might explain the observed inequalities in COVID-19's infections and deaths across U.S. states.

The rest of the paper is organized as follows. In Section 2, we provide some examples to motivate the importance of expressing health inequality measures in terms of copulas. In Section 3, we introduce the notations, we re-formulate the health concentration curve as a function of copula and we derive its expression for some specific copula functions. In Section 4, we develop semi-parametric and nonparametric estimators of health concentration curve using parametric and non-parametric copulas, which we in turn use to derive estimators of Gini health index. In Section 5, we study the asymptotic properties of the semi- and nonparametric estimators. In Section 6, we run Monte Carlo simulations to assess the performances of the semi- and nonparametric estimators and compare them with that of the empirical estimator of the concentration health curve. Section 7 contains an extensive empirical study where we employ our estimators to study health inequality in the U.S. COVID-19 data, and Section 8 concludes. The proofs and the simulation and empirical results can be found in a separate companion appendix.

2 Motivation

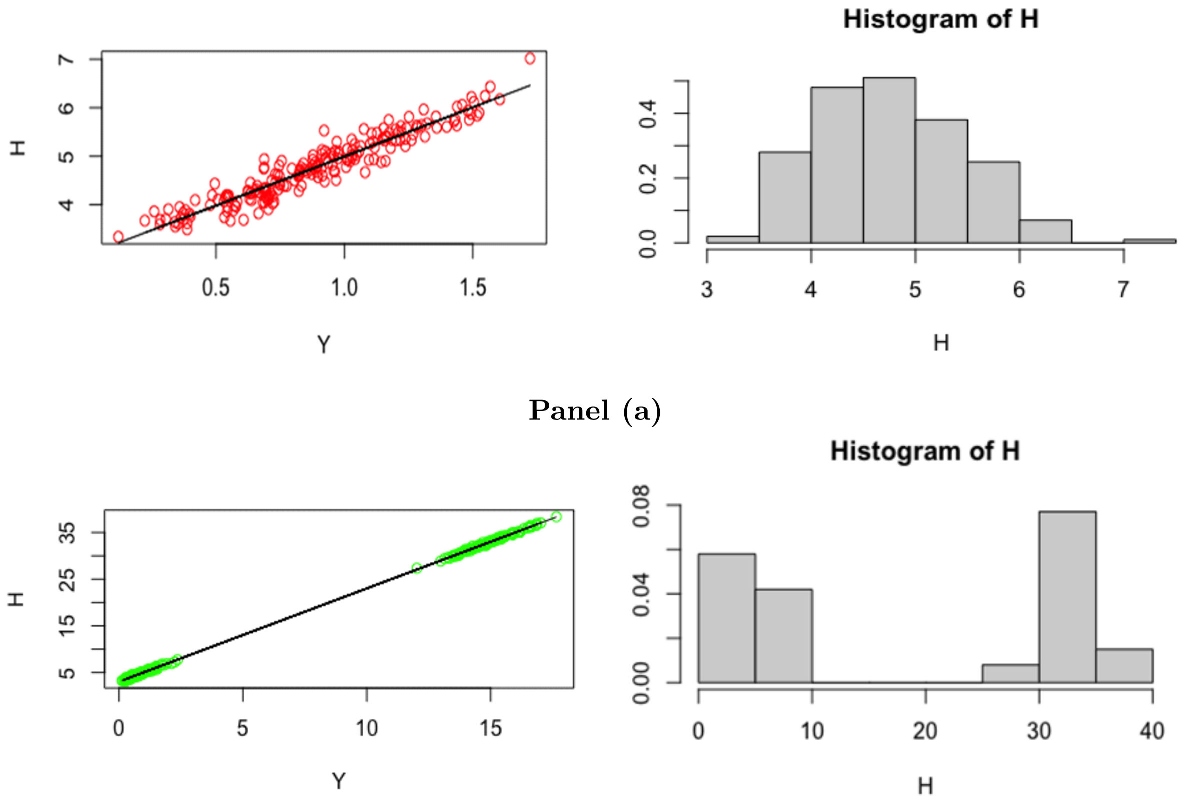
This section aims to motivate the importance of expressing health inequality measures in terms of copula. This helps illustrate the role that play the marginal distribution of the health variable, H, and the dependence structure between this variable and the socioeconomic variable, Y, for the estimation of the health inequality measures. Based on some examples we consider in this section, we show that using simple regressions to study health inequality -as some papers do- might lead to misleading conclusions [see the discussion in the introduction]. For example, when Y and H are linearly dependent, we find that health inequality might or might not be present depending on the shape of the marginal distribution of H. Furthermore, when Y and H follow similar marginal distributions, we find that the existence of health inequality depends on the nature of the dependence between Y and H. We consider the following examples.

Example 1. *Suppose that the health variable H is generated according to the following process:*

$$H_i = 3 + 2Y_i + \varepsilon_i, \text{ for } i = 1...n,$$

where n is the sample size, ε_i is an error term that follows a normal distribution $N(0, 0.2)$*, and the socioeconomic variable Y follows either: **(i)** a Weibull distribution* $W(3, 1)$ [hereafter ***Case 1***] *or **(ii)** a mixture of Weibull distributions* $0.5 * W(2, 1) + 0.5 * W(15, 1)$ [hereafter ***Case 2***].

Using the data we generated in Example 1 [see Figure 1], we calculated the health concentration curves for **Case 1** and **Case 2** using Formula (1) in Section (3)**,** and the results are reported in Figure 2. From this, we see that the marginal distribution of the health variable plays a key role in detecting health inequalities. Panel (a) of Figure 1 and plot (a) of Figure 2 show that when Y and H are linearly dependent, health inequality might not exist if the marginal distribution of H is unimodal and the data are well concentrated around the mean. However, under the same linear dependence between Y and H, we find that if the marginal distribution of H is bimodal as shown in Panel (b) of Figure 1, then we might find that there is a health inequality as illustrated by the plot (b) of Figure 2. The latter case could represent the situation where the high blood pressure is present in poor and rich people, but not among middle income groups.



Panel (a)

Panel (b)

Figure 1: This figure illustrates different marginal distributions of H and linear dependence structure between H and Y.

Example 1 indicates that employing linear regressions might not be informative about the presence or absence of health inequalities. Furthermore, this shows that a good estimator of the marginal distribution of H is necessarily for obtaining a good estimator of health concentration curve. For example, an estimator of the health concentration curve that is based on a parametric estimator of the marginal distribution of H might lead to misleading results if the assumed parametric distribution is misspecified [e.g. the true distribution of H is a mixture of Weibull distributions as in **Case 2**, but the assumed one is a Weibull as in **Case 1**] as shown in Figures 1 and 2. In this case, we would recommend to use a nonparametric estimator of the marginal distribution of H as we do in this paper.

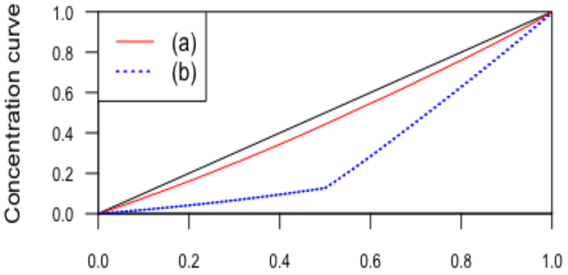
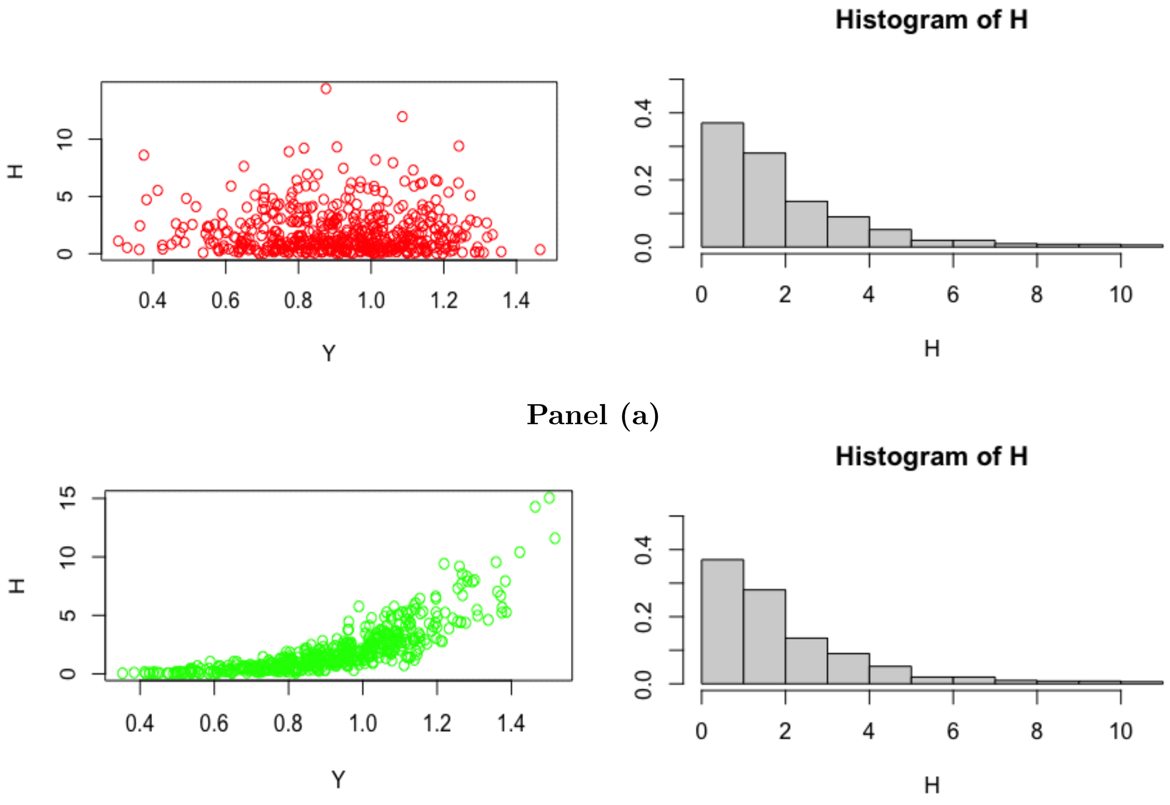


Figure 2: This figure illustrates health concentration curves for the marginal distributions of H and linear dependence between H and Y reported in Figure 1; see panels (a) and (b) of Figure 1.

Example 2. *Suppose that the health variable H and the socioeconomic variable Y follow the Weibull distributions $W(1,2)$ and $W(5,1)$*, respectively, *and consider a Guassian copula for the dependence between Y and H with two different correlation coefficients: (i) $\rho = 0$* [hereafter ***Case 1***] *and (ii) $\rho = 0.9$* [hereafter ***Case 2***].

As in Example 1, we use the data that we generated in Example 2 [see Figure 3] to calculate the health concentration curves for **Case 1** and **Case 2,** and the results are reported in Figure 4. As we can expect, we find that the nature of dependence between Y and H can inform us about the presence or absence of health inequality. On the one hand, as shown in Panel (a) of Figure 3 and plot (a) of Figure 4, the absence of dependence between Y and H implies the absence of health inequality. On the other hand, for the same marginal distribution of H as the one in Panel (a) of Figure 3, Panel (b) of Figure 3 and plot (b) of Figure 4 show that health inequality appears when Y and H are dependent (non-linearly dependent). Example 2 indicates that knowing the true marginal distribution of H or having a good estimator of it is not enough to obtain a good estimator of the health concentration curve. In other words, good estimators of both marginal distribution of H and dependence between Y and H are needed to obtain a good estimator of health concentration curve, hence the focus of this paper on improving the estimation of health concentration curve by providing good estimators of the copula function that captures the dependence between Y and H, in addition to using a nonparametric estimator of the marginal distribution of H.



Panel (a)

Panel (b)

Figure 3: This figure illustrates different marginal distributions of H and different structures for the dependence between H and Y.

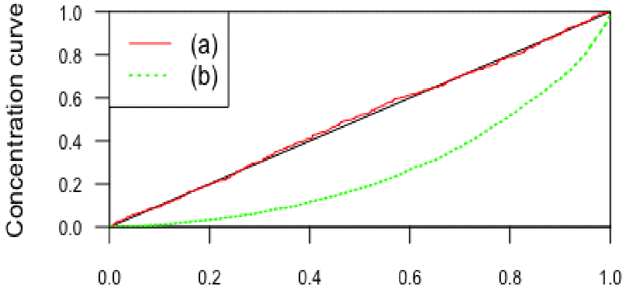


Figure 4: This figure illustrates health concentration curves for the marginal distributions of H and structures of dependence between H and Y reported in Figure 3; see panels (a) and (b) of Figure 3.

3 Copula-based health concentration curve

The health concentration curve plots the cumulative percentage of the health variable against the cumulative percentage of the population, ranked by socioeconomic covariates such as living standards, beginning with the poorest and ending with the richest. The curve illustrates the effect of a socioeconomic variable on concentration of a health variable such as COVID-19 infections and deaths.

Formally, consider H representing the health variable of interest and Y a socioeconomic random variable such as income. Let f be the joint density of the vector (H, Y), with f_H and f_Y the marginal densities of H and Y, $f_{H|Y}$ the conditional density of H given Y, and F_H and F_Y the marginal distributions of H and Y. For $p \in (0,1)$, the health concentration curve is defined as

$$CH(p) = \frac{\int_0^p \mathbb{E}\left[H \mid Y = F_Y^{-1}(u)\right] du}{\int_0^1 \mathbb{E}\left[H \mid Y = F_Y^{-1}(u)\right] du}, \tag{1}$$

where, for $u \in (0,1)$, $F_Y^{-1}(u) = \inf\{t : F_Y(t) \geq u\}$ is the quantile function of Y. The values of $F_Y^{-1}(u)$ at $u = 0,\, 1$ can be set to arbitrary finite real numbers. For example, if Y is the income and H is the rate of COVID-19 mortality, $CH(p)$ is the percentage of the cumulative rate of COVID-19 mortality rate of the $(100 * p)\,\%$ of the poorest population.

Note that the dependence between H and Y plays a crucial role in the calculation of health concentration curve $CH(p)$. For instance, if H and Y are independent, then $CH(p) = p$, i.e., we are in the presence of perfect equality of health variable H across the socioeconomic variable Y. However, this does not mean a perfect equality of the health concentration. As shown in the following proposition, the concentration health curve can be derived through the dependence structure of the vector (H, Y) using copula function [see the proof of Proposition 1 in the appendix of this paper].

Proposition 1 *Let C and c be the copula function and the copula density function of the vector (H, Y), respectively. The health concentration curve CH in Equation (1) can be rewritten as a function of copula:*

$$CH(p) = \frac{\int_0^1 F_H^{-1}(u) C_u(u,p) du}{\mathbb{E}(H)}, \text{ for } p \in (0,1), \tag{2}$$

where $C_u(u,p)$ is the partial derivative of the copula function of the vector (H, Y) :

$$C_u(u,v) = \frac{\partial C(u,v)}{\partial u} = \int_0^v c(u,z) dz.$$

The result in (2) is a copula-based representation of CH. We can now use this representation to derive expressions of the health concentration curve for some specific copula functions. We consider independence and lower and upper Fréchet–Hoeffding dependence structures. As shown below, the latter dependence structures can help illustrate the link between the concentration curve and the Lorenz curve of the health variable H.

Example 1 (Independence): *If H and Y are independent, then $C_u(u, v) = v$. Therefore, $CH(p) = p$, and in this case we obtain a perfect equality, i.e. the socioeconomic variable has no effect on the concentration of the health variable.*

Example 2 (Lower Fréchet–Hoeffding): Suppose that H and Y are countermonotonic variables, i.e., the dependence between the two variables can be modelled using Fréchet–Hoeffding lower bound copula $W(u, v) = \max\{u + v - 1, 0\}$. In this case, $W_u(u, p) = \mathbb{I}(u > 1 - p)$, where $\mathbb{I}(.)$ is an indicator function that takes value one if $u > 1 - p$ and zero otherwise. Hence,

$$CH(p) = \frac{\int_0^1 F_H^{-1}(u) W_u(u, p) du}{\mathbb{E}(H)} = \frac{\int_{1-p}^1 F_H^{-1}(u) du}{\mathbb{E}(H)} = 1 - L_H(1 - p),$$

where L_H is the Lorenz curve of H. For example, if H represents overweightness, then $CH(p)$ is the percentage of the total overweightness of the $(100 * p)\,\%$ of the richest population.

Example 3 (Upper Fréchet–Hoeffding): Suppose that H and Y are comonotone variables, i.e., the dependence between the two variables can be modelled using Fréchet–Hoeffding upper bound copula $M(u, v) = \min\{u, v\}$. In this case, $M_u(u, p) = \mathbb{I}(u < p)$, where $\mathbb{I}(.)$ is an indicator function that takes value one if $u < p$ and zero otherwise. Hence,

$$CH(p) = \frac{\int_0^1 F_H^{-1}(u) M_u(u, p) du}{\mathbb{E}(H)} = \frac{\int_0^p F_H^{-1}(u) du}{\mathbb{E}(H)} = L_H(p),$$

For example, if H represents the payments people made for health care and if we assume a perfect positive correlation between Y and H, then $CH(p)$ is the percentage of the total payments spent by the $(100 * p)\,\%$ of the poorest population.

4 Copula-based estimation

The copula-based representation in Equation (2) will be used to derive semi-parametric and nonparametric estimators of health concentration curve and Gini health coefficient. For the semiparametric estimator, we use a parametric copula function to model the dependence between health and socioeconomic variables. The copula's parameter is estimated using the maximum pseudolikelihood estimator after replacing the cumulative distribution of health variable by its empirical analogue. For the non-parametric estimator, we replace the copula function by a Bernstein copula estimator. Furthermore, we use the above estimators to derive semi-parametric and nonparametric estimators of Gini health coefficient.

Let us first set some notations. We denote by $\{(H_i, Y_i), i = 1, \ldots, n\}$ an independent and identically distributed sample of n copies of the vector (H, Y). From the representation in (2), the health concentration curve CH can be estimated using different approaches that correspond to different ways of estimating the expectation and distribution of H, say $\mathbb{E}(H)$ and F_H, but also to different ways of estimating copula function $C_u(u, p)$. In the following, $\mathbb{E}(H)$ will be estimated using its empirical analogue (sample average of H), and F_H will be estimated using the rescaled empirical distribution:

$$\widehat{F}_H(h) = \frac{1}{n+1} \sum_{i=1}^{n} \mathbb{I}(H_i \leq h), \tag{3}$$

where $\mathbb{I}(.)$ is an indicator function that takes the value one if $H_i \leq h$ and zero otherwise. Note that alternative estimators of F_H - such as a parametric estimator or a smoothed nonparametric estimator like Kernel, Bernstein, etc.- can be considered.

In the next two sub-sections, we consider two different estimators of copula $C_u(u, p)$. In Sub-section 4.1, we introduce a semiparametric estimator of $CH(p)$ in which only realizations of Y in $C_u(u, p)$ will be used for estimating the copula function, i.e., the information on Y are required for modelling the dependence structure of the vector (H, Y). In Sub-section 4.2, we derive a nonparametric estimator of $CH(p)$ using a nonparametric estimator of copula $C_u(u, p)$; hereafter Bernstein estimator.

4.1 Semiparametric estimation

We consider that the copula C of (H, Y) belongs to a parametric family $\{C(., ., \theta), \theta \in \Theta\}$, with an unknown parameter θ that is in the set Θ, which is a compact subset of $\mathbb{R}^q$. Denote by θ_0 the true value of θ. There exist several estimators of θ_0 and a most popular one is given by the following maximum pseudo likelihood estimator:

$$\hat{\theta}_n = \underset{\theta}{\operatorname{argmax}} \sum_{i=1}^{n} \log c \left(\widehat{F}_H(H_i), \widehat{F}_Y(Y_i), \theta \right), \tag{4}$$

where c is the copula density of C, $\widehat{F}_H(h)$ is defined in Equation (3) and $\widehat{F}_Y(y) = (n+1)^{-1} \sum_{i=1}^{n} \mathbb{I}(Y_i \leq y)$. Note that $\widehat{\theta}_n$ is the estimator proposed by Shih and Louis (1995) and Genest, Ghoudi and Rivest (1995), and its asymptotic representation can straightforwardly be obtained from the proof of Theorem 1 in Tsukuhara (2005).

Using the maximum likelihood estimator of θ_0 in (4) and the nonparametric estimator of F_H in (3), a semiparametric estimator of $CH(p)$ is defined as follows:

$$\widehat{CH}(p) = \frac{\sum_{i=1}^{n} H_i \, C_u(\widehat{F}_H(H_i), p, \widehat{\theta}_n)}{\sum_{i=1}^{n} H_i}, \tag{5}$$

where $C_u(u, v, \theta) = \frac{\partial C(u,v,\theta)}{\partial u}$. We next use simulations based on examples 1-3 of Section 3 to empirically illustrate the semiparametric estimator in (5). We also compare this estimator with the following empirical estimator of $CH(p)$ [see Wagstaff (2002)],

$$\widehat{CH}_n(p) = \frac{\sum_{i=1}^{n} H_i \; \mathbb{I}(Y_i \le \widehat{F}_Y^{-1}(p))}{\sum_{i=1}^{n} H_i}, \tag{6}$$

where $\mathbb{I}(.)$ is an indicator function that takes the value one if $Y_i \le \widehat{F}_Y^{-1}(p)$ and zero otherwise. The simulation results are reported in Figure A1 of the separate companion appendix, which we obtain after generating $n = 100$ observations of (H_i, Y_i) from a Gaussian copula with correlation coefficients $\rho = 0$ (Example 1), $\rho = -0.99$ (Example 2) and $\rho = 0.99$ (Example 3), and using as a marginal distribution of H and Y an exponential distribution with a parameter $\lambda = 1$. To avoid numerical problems, we excluded the cases $\rho = \pm 1$. Thereafter, we report the results for both known and unknown copulas. When the copula is unknown, we use the R package `VineCopula` to select a copula function that fits better our generated data [see curves of the unknown copula in Figure A1]. Furthermore, Figure A1 illustrates the estimators of $1 - L_H(1-p)$ and $L_H(p)$, where L_H is the Lorenz curve of H; see subfigures (b) and (c) of Figure A1.

Subfigure (a) of Figure A1 shows that the copula-based estimator of $CH(p)$ is much closer to the 45-degree line (true curve of $CH(p)$ under independence) than the empirical estimator, and that the former is much smoother than the latter. This result holds for both known and unknown copulas, which might suggest a high performance in favour of the semi-parametric estimator. We also see that the estimators of $CH(p)$ for known and unknown copulas are very similar, which might indicate that the R package `VineCopula` performs well in terms of copula selection. Moreover, we find that the Lorenz curve's estimator is quite distant from the 45-degree line, which means that perfect equality using $CH(p)$ does not imply perfect equality using L_H. Subfigures (b) and (c) illustrate the semiparametric and empirical estimators of $CH(p)$ and $1 - L_H(1-p)$ (or L_H) when the dependence between H and Y is given by lower and upper Fréchet–Hoeffding, respectively. From these, we see that the estimators of $CH(p)$ and $1 - L_H(1-p)$ are quite similar, but again the semi-parametric estimator is smoother than the empirical estimator of $CH(p)$.

4.2 Nonparametric estimation

The semiparametric estimation of CH assumes that the copula C of (H, Y) belongs to a parametric family of copulas. In practice, however, the copula family is unknown and consequently the semiparametric estimator in (5) might be biased if the chosen parametric copula is misspecified. To overcome this issue, we propose a nonparametric estimation of CH based on the following nonparametric estimator of copula. Formally, we consider the following Bernstein estimator of $C(u, v)$:

$$C_{m,n}(u,v) = \sum_{k_0=0}^{m} \sum_{k_1=0}^{m} C_n\left(\frac{k_0}{m}, \frac{k_1}{m}\right) P_{m,k_0}(u) P_{m,k_1}(v), \tag{7}$$

where m is an integer that plays the role of bandwidth, C_n is the empirical copula, and

$$P_{m,k}(z) = \begin{pmatrix} m \\ k \end{pmatrix} z^k (1-z)^{m-k},$$

is the binomial distribution function with the parameters (m, k). For independent and identically distributed $(i.i.d.)$ data, Sancetta and Satchell (2004) introduced a Bernstein polynomial estimator of the copula and established the asymptotic normality of the Bernstein density copula. Under some regularity conditions, Sancetta and Satchell (2004) show that any copula can be approximated by a Bernstein copula. Asymptotic properties of the Bernstein density copula for α-mixing data are studied in Bouezmarni et al. (2010). Furthermore, Janssen et al. (2012) have shown that the Bernstein copula outperforms the classical empirical copula originally proposed by Deheuvels (1979).

Moreover, it is worth noting that an estimator of the partial derivative $C_u(u, v)$ can be derived using the Bernstein copula, but not the empirical copula. Using the Bernstein copula in (7), a nonparametric estimator of the first-order partial derivative of the copula function, $C_u(u, p)$, can be obtained as follows:

$$\widetilde{C}_u(u,p) = m \sum_{k_0=0}^{m-1} \sum_{k_1=0}^{m} \left[C_n\left(\frac{k_0+1}{m}, \frac{k_1}{m}\right) - C_n\left(\frac{k_0}{m}, \frac{k_1}{m}\right)\right] P_{m-1,k_0}(u) P_{m,k_1}(p). \tag{8}$$

The above estimator was introduced and its asymptotic properties were derived in Janssen et al. (2016). From (2) and (8), a nonparametric estimator of $CH(p)$ can be defined as follows:

$$\widehat{CH}_{m,n}(p) = \frac{\sum_{i=1}^{n} H_i \, \widetilde{C}_u(\widehat{F}_H(H_i), p)}{\sum_{i=1}^{n} H_i}. \tag{9}$$

We now use simulations and some examples of dependence between H and Y to illustrate the nonparametric estimator in (9). We generated $n = 100$ observations of (H_i, Y_i) using a Gaussian copula with correlation coefficients: (a) $\rho = 0.6$, (b) $\rho = -0.6$, (c) dependence between H and Y is quadratique, and (d) $\rho = 0$ (independence). The simulation results reported in Figure A2 of the appendix of the separate companion appendix show that the Bernstein copula-based estimator outperforms the empirical estimator as the former is much closer to the 45-degree line under independence and it is much smoother under other cases of dependence.

4.3 Gini health index estimation

We now use the previously developed estimators of CH to propose estimators of Gini health index. This index is a measure of dispersion intended to quantify the health inequality within a given socioeconomic group. Formally, the Gini health index is defined as follows:

$$G = 2\int_0^1 (p - CH(p))\, dp = 2\frac{Cov(H, F_Y(Y))}{\mathbb{E}(H)}. \tag{10}$$

Under independence between H and Y, the covariance term $Cov(H, F_Y(Y))$ is equal to zero, and therefore the Gini health index G is equal to zero. The index takes values between -1 and 1. On the one hand, a negative value of G indicates a pro-poor health, and in this case the health concentration curve will be located above the 45-degree line; see Subfigure 1(b). On the other hand, a positive value of G indicates a pro-rich health, and the health concentration curve in this case will be below the 45-degree line; see Subfigure 1(c). Furthermore, G is equal to zero when the concentration curve is close to the 45-degree line, see Subfigure 1(a). Observe, however, that the Gini health index G - which is equal to the integral of the difference between the 45-degree line and the curve $CH(p)$ - can be equal to zero even when the health concentration curve does not coincide with the 45-degree line as shown in Subfigure 2(c). Thus, $CH(p)$ curve should represent a better way of visualizing and detecting health inequalities.

Note that if H and Y are comonotone random variables - i.e., the dependence between the two variables can be modelled using the upper Fréchet–Hoeffding -, then $CH(p) = L_H(p)$ and G represents a Gini health coefficient that is constructed from Lorenz curve. In this case, we have $F_H(H) = F_Y(Y)$ and $G = \frac{2}{\mathbb{E}(H)} Cov(H, F_Y(Y)) = \frac{2}{\mathbb{E}(H)} Cov(H, F_H(H))$. Formally,

$$G = 2\int_0^1 (p - CH(p))\ dp = 2\int_0^1 (p - L_H(p))\, dp.$$

Now, if H and Y are countermonotonic random variables - i.e., the dependence between the two variables can be modelled using the lower Fréchet–Hoeffding -, then $CH(p) = 1 - L_H(1-p)$ and G

will be equal to minus Gini health index that is calculated based on Lorenz curve L_H. Formally,

$$G = 2\int_0^1 (p - CH(p))\, dp = 2\int_0^1 (p - 1 + L_H(1-p))\, dp = -\, 2\int_0^1 (p - L_H(p))\, dp.$$

We can show that G is equal to minus Gini health coefficient that is based on Lorenz curve by observing that $F_H(H) = 1 - F_Y(Y)$ and $Cov(H, F_Y(Y)) = -Cov(H, F_H(H))$.

Thus, using the expression of G in Equation (10) and the estimators of $CH(p)$ in Equations (5) and (9), we propose the following semiparametric

$$\widehat{G} = 2\int_0^1 \left(p - \widehat{CH}(p)\right)\, dp \tag{11}$$

and nonparametric

$$\widetilde{G} = 2\int_0^1 \left(p - \widehat{CH}_{m,n}(p)\right)\, dp \tag{12}$$

estimators of Gini health index, respectively.

We next establish the asymptotic properties of the semiparametric and nonparametric estimators of the health concentration curve CH. These properties can be used to establish the asymptotic properties of the estimators of the Gini health index.

5 Asymptotic properties of the estimators of CH

We investigate the consistency and asymptotic normality of the previously developed estimators of health concentration curve CH. The following subsections (5.1) and (5.2) state the asymptotic i.i.d. representations of the semiparametric and nonparametric estimators $\widehat{CH}(p)$ and $\widehat{CH}_{m,n}(p)$ in (5) and (9), respectively.

5.1 Asymptotic properties of semiparametric estimator of CH

In this subsection, we assume that the copula function C belongs to a known parametric family of copulas $\mathcal{C} = \{C(.;\theta), : \theta \in \Theta\}$, where Θ is a compact subset of $\mathbb{R}^q$. Let $\hat{\theta}_n$ be an estimator of the true parameter θ_0 that satisfies the following condition:

Assumption A: *The estimator $\hat{\theta}_n$ of θ_0 has the following asymptotic representation:*

$$\hat{\theta}_n - \theta_0 = n^{-1}\sum_{i=1}^{n} \xi_i + o_p(n^{-1/2}), \tag{13}$$

where $\xi_i = \xi(F_H(H_i), F_Y(Y_i); \theta_0)$ is a q-dimensional random vector such that $E(\xi) = 0$ and $E\left(||\xi||^2\right) < \infty$, with $||.||$ represents the Euclidean norm.

Note that the estimators of θ_0 that have been proposed in Shih and Louis (1995) and Genest, Ghoudi and Rivest (1995) satisfy Assumption **A**. Furthermore, Tsukuhara (2005) provides an asymptotic representation of $\hat{\theta}_n$ and establishes its asymptotic properties.

Now, before we state the asymptotic representation of the semiparametric estimator of CH, we need to introduce the following notations:

- $C_{uu} = \dfrac{\partial^2 C}{\partial u^2}$ and $C_{u,\theta} = \left(\frac{\partial C_u}{\partial \theta_1}, \dots, \frac{\partial C_u}{\partial \theta_q}\right)$.
- $r_\theta(p,\theta) = \left(\frac{\partial\, r(p,\theta)}{\partial \theta_1}, \dots, \frac{\partial\, r(p,\theta)}{\partial \theta_q}\right)^\top$, where $r(p,\theta) = \mathbb{E}\left[H C_u(F_H(H), p, \theta)\right]$.

Using the above notations, we consider the following additional assumptions that we need in order to establish the result of Theorem 1.

Assumption B: We assume that:

B1: **(i)** $\mathbb{E}|H| < \infty$ and **(ii)** $hF_H(h) \to 0$ as $h \to -\infty$**;**

B2: $C_{u,u}$ and $C_{u,\theta}$ are continuous on $(0,1)^2$ and $(0,1) \times \Theta$, respectively;

B3: $\mathbb{E}[H\, C_{u,u}(F_H(H), p; \theta_0)]^2 < \infty$;

B4: $\mathbb{E}[H\, \frac{\partial C_u}{\partial \theta_k}(F_H(H), p; \theta_0)]^2 < \infty$, for $k = 1, \dots, q$.

The following theorem states the asymptotic i.i.d. representation of the semiparametric estimator $\widehat{CH}(p)$ in (5) [see the proof of Theorem 1 in the Appendix of this paper].

Theorem 1 *Under Assumptions **A**, **B1-B4**, we have*

$$\widehat{CH}(p) - CH(p) = n^{-1} \sum_{i=1}^{n} \zeta_i + o_p(n^{-1/2}),$$

where,

$$\zeta_i = \mathbb{E}(H)^{-1} \left[H_i C_u(F_H(H_i), p, \theta_0) + \xi_i^\top r_\theta(p,\theta_0) + \eta(H_i, p, \theta_0) - H_i\, CH(p) \right],$$

with $\eta(H_i,p,\theta_0) = \mathbb{E}_H\left[H\left(\mathbb{I}(H_i \le H) - F_H(H)\right) C_{uu}(F_H(H), p, \theta)\right], r_\theta(p,\theta_0) = \mathbb{E}\left[H\, C_{u,\theta}(F_H(H), p, \theta_0)\right],$ ξ_i *is defined in Assumption **A**,* $\mathbb{E}_H$ *represents expectation with respect to* H, *and* $\mathbb{I}(.)$ *is an indicator function that takes value one if* $H_i \le H$ *and zero otherwise.*

Theorem 1 can be used to establish the asymptotic normality of the semiparametric estimator $\widehat{CH}$ with zero mean and asymptotic variance:

$$Var(\widehat{CH}(p)) = \mathbb{E}\left(\zeta_i^2\right) - \left(\mathbb{E}(\zeta_i)\right)^2,$$

where the expression of ζ_i is defined in Theorem 1. This variance is unknown, but it can be estimated by replacing the unknown quantities in the expressions of $\mathbb{E}(\zeta_i)$ and $\mathbb{E}\left(\zeta_i^2\right)$ by their empirical analogues. However, for testing and building confidence interval around $\widehat{CH}(p)$, we recommend to use bootstrap.

5.2 Asymptotic properties of nonparametric estimator of CH

We now establish the asymptotic i.i.d representation of the nonparametric estimator of CH. Let us first define some new terms. For independent random vectors $(U_1, V_1), \dots, (U_n, V_n)$, with joint distribution function C, we define

$$W_{i,m}(u,v) = m \sum_{k_0=0}^{m-1} \sum_{k_1=0}^{m} \left[A_i(k_0,k_1) - A_i(k_0,m)C_u(u,v)\right] P_{m-1,k_0}(u) P_{m,k_1}(v),$$

where

$$A_i(k_0,k_1) = \mathbb{I}\left(\frac{k_0}{m} < U_i \leq \frac{k_0+1}{m}, V_i \leq \frac{k_1}{m}\right) - P\left(\frac{k_0}{m} < U_i \leq \frac{k_0+1}{m}, V_i \leq \frac{k_1}{m}\right).$$

The following theorem states the asymptotic i.i.d. representation of the nonparametric estimator $\widehat{CH}_{m,n}(p)$ in (9) [see the proof of Theorem 2 in the Appendix of this paper].

Theorem 2 *Suppose that the second derivatives* $C_{u,u} = \frac{\partial^2 C(u,v)}{\partial^2 u}$ *and* $C_{u,v} = \frac{\partial^2 C(u,v)}{\partial u \partial v}$ *are Lipshitz continuous on* $[0,1]^2$. *If* $(m/n)\log\left(m^{1/2}n\right) \to 0$ *as* $n \to \infty$, *we have*

$$\widehat{CH}_{m,n}(p) - CH(p) = n^{-1}\sum_{i=1}^{n} \nu_i(p) + o_p(n^{-1/2}m^{1/4}),$$

where

$$\nu_i(p) = \mathbb{E}(H)^{-1}\left[\mathbb{E}_H\left[HW_{i,m}(F_H(H),p)\right] - H_i\, CH(p)\right].$$

Theorem 2 can be used to establish the asymptotic normality of the nonparametric estimator of CH with mean zero and - by adapting the proofs in Janssen et al. (2016) - asymptotic variance of order $O(n^{-1}m^{1/2})$. This asymptotic variance, however, is unknown as it depends on C_u. In practice we suggest to use bootstrap for constructing confidence intervals around $\widehat{CH}_{m,n}(p)$.

6 Monte Carlo simulations

In this section, we run Monte Carlo simulations to assess the performance of the estimators proposed previously. In particular, we calculate the Integrated Mean Square Errors (IMSE) of semiparametric and nonparametric estimators of health concentration curves, which we compare with the IMSE of the empirical estimator.

In our simulations, several data-generating processes (DGPs) are considered to model the dependence structure between health and socioeconomic variables. We use different copulas to generate data under different degrees of dependence measured by different values of Kendall's tau coefficient τ. The copulas under consideration are Gaussian copula, Student copula, Clayton copula, and Gumbel copula. The values of Kendall's tau coefficient under consideration are: (i) $\tau = -0.4$, 0.001, 0.01, 0.1, 0.2, 0.5, 0.7 for Gaussian and Student copulas and (ii) $\tau =$ 0.001, 0.01, 0.1, 0.2, 0.5, 0.7 for Clatyton and Gumbel copulas. To generate the data of health and socioeconomic variables, we also use an exponential distribution with the parameter $\lambda = 1$ and a Weibull distribution with the scale and shape parameters $\lambda = 2$ and $k = 10$, respectively.

The performances of the estimators $\widehat{CH}(p)$, $\widehat{CH}_{m,n}(p)$ and $\widehat{CH}_n(p)$ are assessed based on the IMSE that we calculate for different samples sizes: $n =$ 50, 100, 200, and 500. The calculation of IMSE is obtained using $B = 1,000$ replications. Formally, the IMSE is defined as:

$$\text{IMSE} \approx \frac{1}{I}\sum_{i=1}^{I}\left(\frac{1}{B}\sum_{j=1}^{B}\left(\widehat{CH}_j(p_i) - CH(p_i)\right)^2\right),$$

for $p_i = 0.01, 0.02, \ldots, 0.99$, and where $\widehat{CH}_j(p_i)$, for $j = 1, \ldots, B$, is the estimator of the concentration health $CH(p_i)$ that corresponds to the j-th replication. The integral $\int(.)$ of the MSE is approximated by replacing it with the sum $\sum^{I}(.)$ for $I = 99$.

For the semiparametric estimator $\widehat{CH}(p)$, we assume that the parametric copula is unknown and we use the **BiCopSelect** function implemented in **VineCopula R** package to select the appropriate copula from a set of copula families. This approach is expected to alleviate the negative effect that copula misspecification might have on the semiparametric estimator. We recall that copulas can be selected according to an Akaike or Bayesian Information Criteria [AIC and BIC, respectively]; see Akaike (1973), Schwarz (1978), and Manner (2007). To do so, all available copulas are first fitted using maximum pseudo-likelihood estimation. Then the information criteria of all fitted copulas are computed, and the copula that has the minimum AIC or BIC is selected. Formally, the AIC of a bivariate parametric copula density $c(F_H(H), F_Y(Y), \theta)$ is defined as

$$AIC = -2\sum_{i}^{n} ln[c(\hat{F}_H(H_i), \hat{F}_Y(Y_i), \hat{\theta}_n) + 2\kappa,$$

where κ represents the number of parameters: e.g.; $\kappa = 1$ for a copula that depends on one parameter, and $\kappa = 2$ for a copula with two parameters. Similarly, the BIC of bivariate copula density $c(F_H(H), F_Y(Y), \theta)$ is given by

$$BIC = -2\sum_{i}^{n} ln[c(\hat{F}_H(H_i), \hat{F}_Y(Y_i), \hat{\theta}_n) + ln(n)\kappa.$$

Furthermore, the bandwidth parameter m that we use to calculate $\widehat{CH}_{m,n}(p)$ is selected according to the rule of thumb $m = \left[a \times n^{2.5/5}\right]$, where $[.]$ is the integer part of $a \times n^{2.5/5}$. To assess the sensitivity of the estimation results with respect to m, we consider various values of $m \in \{1, 2, 3, 4, 5\}$, which is a common practice in nonparametric estimation when no optimal bandwidth is available. In the simulations, we find that the optimal value of a that works for all DGPs under consideration is equal to 4. Thus, to save space we only report results for $a = 4$; see Table 1. The other values of a also provide reasonable results and the latter are available upon request.

Table 1 shows that the IMSEs of all estimators are decreasing with the sample size. Interestingly, we find that the semiparametric estimator dominates both the empirical and the nonparametric estimators as it has the smallest IMSE under different copulas, degrees of dependence and sample sizes, except when the dependence between H and Y is weak or when τ takes values between 0.001 and 0.1. Thereafter, we find that the second best estimator is the nonparametric estimator that does better than the empirical estimator, except when the degree of dependence between H and Y takes values between $\tau = 0.5$ and $\tau = 0.7$ for Gaussian and Gumbel copulas.

The above results are obtained when the marginal distributions of H and Y are given by the exponential distribution with the parameter $\lambda = 1$. We have also got additional results after replacing the exponential distribution by the Weibull distribution with the parameters $\lambda = 2$ and $k = 10$. To save space, the latter results are not reported but they are available upon request. Using Weibull as a marginal distribution of H and Y, we find similar results to those discussed previously. The best estimator in terms of IMSE is the semiparametric estimator $\widehat{CH}(p)$, followed by the Bernstein estimator $\widehat{CH}_{m,n}(p)$ and then the empirical estimator $\widehat{CH}_n(p)$.

7 Inequality in COVID-19's infections and deaths

Since January 2020, COVID-19 pandemic led to millions of infections and deaths, and caused distressing economic worldwide. Acknowledging the importance of measures like lockdowns, testing, and face masks in reducing the transmission of COVID-19, concerns have arisen about the link between pre-existing social and economic inequalities and inequalities in COVID-19's infections and deaths, with the most-deprived classes are worst hit. In the United States, the spread of COVID-19 across the states have shown that not all Americans are equally at risk of infection and mortality from the virus. Furthermore, the World Health Organization (WHO) in the European Region pointed out that COVID-19's exposure risk and the severity of its health, social and economic impacts are not being felt equally in the European countries.

Moreover, a growing number of papers have investigated how social classes in societies are affected by COVID-19. They studied the effect of socioeconomic factors on COVID-19's infection and death rates, i.e.; the impact of factors that make for example low-income class and people of color more vulnerable than others, see Chen and Krieger (2021), Chin et al. (2020), McLaren (2021), and Brown and Ravallion (2020). None of the above-mentioned studies, however, use measures that are designed to detect health inequalities across socioeconomic factors. These studies are based on simple correlations and regressions and might not help detect these inequalities. In this section, we apply our semi- and nonparametric estimators of CH and Gini health index to quantify and examine inequalities in COVID-19's infections and deaths across and within U.S. states. The next subsection describes the U.S. data we use in our empirical analysis.

7.1 Data

We begin by describing our data and provide some descriptive statistics. Micro-data on COVID-19's infections and deaths that include socioeconomic factors at unit-record level (county level for the U.S. data) are less frequent. To obtain our data, we had to merge recorded counts of cases and deaths in the U.S. at county level with socioeconomic characteristics-average income, race, income inequality and poverty. Our dataset contains information about 2777 counties across 45 U.S. states. We excluded five U.S. states because of insufficient data, i.e.; in each of these states we have very few counties, which is not enough for the estimation of health concentration curves and Gini health index. The excluded states are: Kansas, Kentucky, Louisiana, Maine, Nevada. Our county-level variables fall into three general categories:

- Health variables on confirmed COVID-19's infections and deaths.
- Socioeconomic covariates on median income, total population, the percentage of population below the poverty line, percentages of residents that are African Americans, Whites, and Asians.
- Inequality measures such as Gini (economic) index and Gini health index that we estimated using our semi- and nonparametric estimators.

For data on confirmed COVID-19's infections and deaths, we draw on the U.S. Centers for Disease Control and Protection (CDC).[2] We use the most recent numbers available at the time of writing this paper (June 10, 2021). Data on county's population, population density, demographics and poverty rates were obtained from the US Census Bureau. Median income, and the poverty rate are estimated from survey data, but complemented by small-area estimation methods [see Rao and Molina (2015)]. Descriptive statistics of COVID-19's infection and death rates and socioeconomic variables in the 2777 counties can be found in Table 2 of the separate companion appendix. In this table, the share of African Americans refers to the proportion of the population that identifies as Black only, while the share "White" refers to the proportion of the population that identifies as White. The table shows that the average number of infections and deaths per county are 9.8% and 1.9%, while their standard deviation are 2.99% and 1%, respectively. Crowley county in Colorado recorded the highest number of infections (37%) and the second highest percentage of infections are in Chattahoochee in Georgia (36.8%). We also find that, overall, the average of poverty rate is 14.3%, ranging from 2.3% in Sterling County (Texas) to 54.7% in Todd County (South Dakota). Gini (economic) index is also estimated using small-area methods and varies widely across counties, from the lowest value of 0.302 in Skagway County (Alaska) to the highest value of 0.609 in Harding County (New Mexico). Furthermore, Table 3 of the separate companion appendix reports the correlation matrix of the socioeconomic variables. From this, we find a significant positive correlation between the share of African Americans and poverty rate and a significant but negative correlation between the share of White Americans and poverty rate. We next provide the results of the estimation of health concentration curves and Gini health index.

7.2 Estimation of health concentration curve

As we mentioned earlier, the health concentration curve CH allows one to obtain plots of cumulative percentage of the health variable - here COVID-19's infection and death rates - against the cumulative percentage of the population ranked by the socioeconomic variable - here poverty rate, population density, share of African Americans, share of White Americans, etc. These plots help visualize health inequalities by observing the position of the health concentration curve with respect

[2] The data are available through USA Facts : https://usafacts.org/visualizations/coronavirus-covid-19-spread-map/. An alternative source is the New York Times data site for COVID-19 obtainable from their Github repository. The NYT site, however, records infections and deaths according to the county in which they occurred, while CDC does so according to the person's place of residence.

to the line of equality (45-degree line) in a two-dimensional space. When the higher the socioeconomic variable is the richer the individual is, if CH lies above the 45-degree line, then health inequality is referred to as pro-rich - i.e., the rich have better health than the poor -, and in this case the associated health concentration index is negative. When CH lies under the 45-degree line, the health inequality is considered pro-poor - i.e., the poor have better health than the rich-, and the associated health concentration index is positive. If the health concentration curve coincides with the 45-degree line, then there is no socioeconomic health inequality, and the associated health concentration index is necessarily zero.

We now use the data described in the previous subsection to calculate the semiparametric and nonparametric estimators of health concentration curve for COVID-19's infection and death rates. For each state, we use three socioeconomic variables: income, poverty, and race (proportion of black/white people). The results are obtained for 45 U.S. states, but to save space and for a better presentation we only report the results for average, low and high income states, average, low and high poverty rate states, and average, low and high share of whites states.[3] The results are reported in Figures A3 to A11 of the separate companion appendix. Using income as a socioeconomic variable, Figures A3 and A4 show a clear pro-rich inequality for COVID-19's death rate both in low and high income states, except for West Virginia. Interestingly, for the average income states [see Figure A5], we find that the health concentration curve tend to match the 45-degree line, which indicates that individuals in these states had equal chance of dying from COVID-19 regardless of their socioeconomic position. We reach similar results when we replace income by poverty rate noting that low poverty rate is expected to be highly correlated with high income and vice-versa. Figures A6 and A7 confirm the pro-rich inequality for COVID-19's death rate for both low and high poverty rate states, with health concentration curve lies below the 45-degree line when the higher the poverty rate is the poorer the county is. The result in Figure A8 is also similar to the one we obtained in Figure A5: there is no inequality in COVID-19's death rate in the states with average poverty rate as the health concentration curve matches the 45-degree line. Moreover, Figures A9 and A10 indicate that the inequality in COVID-19's death rate is generally in favor of counties with low share of white people, except in state of California. However, the opposite happens in states with average share of white people, where we see that the inequality in COVID-19's death rate is in favor of counties with high share of white people.

To sum up, the socioeconomic variables income, poverty rate and ethnicity have an impact

[3] The rest of the results are available upon request.

on COVID-19's death rate, with the most-deprived classes are worst hit by the virus. In the next subsections, we provide additional analysis by estimating Gini health index and econometric models that link COVID-19's cases and deaths to Gini health index and other key socioeconomic factors.

7.3 Estimation of Gini health index

We now use the data to calculate the semiparametric and nonparametric estimators of Gini health index for COVID-19's death rates; see Equations (11) and (12), respectively. We focus on deaths because it should better reflects health inequalities across states with different levels of socioeconomic variables income, poverty, and race. Once the infection occurs, one might expect that COVID-19 deaths will be higher in the deprived states/areas due to the poor living conditions and underfunded local services such as health services that are essential for saving lives. To estimate the Gini health index, we excluded all the states with less than 10 counties, which affects the following states: New Hampshire, Conecticut, Hawaii, Kensas, Kentucky, Louisiana, Maine, Maryland, and RhodeIsland. The results for the remaining 40 U.S. states are reported in Figures A12 to A14 of the separate companion appendix.

Firstly, Figure A12 shows the results of estimating Gini health index of each U.S. state when the health variable H represents COVID-19's death rates and the socioeconomic variable Y is the median income. Since for this case the lowest socioeconomic position represents counties with lower median income, the negative Gini health index for the majority of U.S. states as shown in Figure A12 indicates a pro-rich health inequality. In other words, this result means that people in counties with high median income die less from COVID-19 compared to people who live in counties with low median income. Secondly, Figure A13 illustrates the results of semi- and non-parametrically estimating Gini health index when the socioeconomic variable Y is the poverty rate. Unlike in the previous case, the lowest socioeconomic position for the poverty rate represents rich counties, thus the positive Gini health index that we see in Figure A13 for the majority of U.S. states confirms again that there is a pro-rich health inequality as counties with high poverty rate have high COVID-19 mortality compared to counties with the low poverty rate. Finally, Figure A14 shows the results when the socioeconomic variable is the proportion of African Americans. Since for this case the lowest socioeconomic position represents counties with lowest proportion of African Americans, the result in Figure A14 indicates that there is a pro-non–African American health inequality across U.S. states as counties with high proportion of African Americans have high COVID-19 mortality compared to counties with the low proportion of African Americans.

This study offers some insights for policymakers to reduce health inequality, thus ensuring that everyone has the best chance of survival from COVID-19. The results indicate that U.S. states' authorities should allocate greater health resources to more deprived counties and counties with high proportion of African Americans to help reduce health inequalities in COVID-19 mortality.

7.4 Further analysis

In this subsection, we use U.S. state-level data for further analysis regarding the factors that affect COVID-19's infections and deaths. We regress state-level COVID-19's infections and deaths on measures of health and economic inequalities and other socioeconomic variables. In the regressions, we include all or some of the following covariates: state's population size, state's poverty rate, state's median income, state's Gini economic index, state's Gini health index we calculate using the estimators we developed in the previous sections, state's share of African Americans, and state's share of Asian Americans. The results are reported in Tables 4-5 of the separate companion appendix.

Tables 4-5 show that both economic (Gini index) and health (Gini health index) inequality measures have a positive impact on COVID-19's infections and deaths after controlling for other socioeconomic covariates. These effects are generally statistically significant. The effects of poverty rate on COVID-19's infections and deaths are statistically insignificant, which might be explained by the high correlations between poverty rate and other covariates. Furthermore, median income and population size have significant and positive impacts on both COVID-19's infections and deaths, whereas the impact of state's share of African Americans is positive and generally statistically insignificant and the impact of state's share of Asian Americans is negative and statistically significant. The results for COVID-19's infections and deaths are very similar.

The above results confirm once again that greater economic and health inequalities led to higher COVID-19's infection and death rates, with the most-deprived classes are worst hit. Policymakers are thus urged to set social and economic measures to reduce COVID-19's infections and deaths among most-deprived classes. At short-term, measures like free masks and sanitizers and targeted vaccination campaigns should help reduce COVID-19's infection and death rates in deprived states. At long-term, governments should set measures that can help reduce economic and social inequalities in society, which in turn will help reduce health inequality.

8 Conclusion

We derived semi- and non-parametric estimators of Health Concentration curve and Gini Health Coefficient that can quantify inequalities in COVID-19 infections and deaths and help identify the social classes that are most at risk of infection and dying from the virus. We first expressed CH in terms of copula function that we used to build our estimators of CH. For the semi-parametric estimator, a parametric copula was used to model the dependence between health and socioeconomic variables. The parameters of the copula were estimated using maximum pseudo-likelihood estimator after replacing the cumulative distribution of health variable by its empirical analogue. For the non-parametric estimator, we replaced the copula function by the Bernstein copula estimator. Furthermore, we used the estimators of CH to derive semi- and non-parametric estimators of Gini health coefficient. We establish the consistency and asymptotic normality of the estimators of CH. Using several data-generating processes and sample sizes, a Monte-Carlo simulation exercise showed that the semiparametric estimator outperforms the Bernstein-copula-based estimator, and that the latter does better than the empirical estimator in terms of Integrated Mean Squared Error**.** We also run an extensive empirical study to illustrate the importance of CH and Gini health index estimators for investigating inequality in COVID-19 infections and deaths in the U.S. The empirical results showed that socioeconomic variables like poverty, race, and economic prosperity of a state might explain the observed inequalities in COVID-19 infections and deaths.

References

[1] Akaike, H. (1998). “Information theory and an extension of the maximum likelihood principle,” In Selected papers of hirotugu akaike, Springer, New York, NY, pp. 199-213.

[2] Allison, R. A. and Foster, J. E. (2004). “Measuring health inequality using qualitative data,” *Journal of Health Economics,* vol. 23 (3), pp. 505-524.

[3] Bermudi, P. M. M., Lorenz, C., de Aguiar, B. S., Failla, M. A., Barrozo, L. V., & Chiaravalloti-Neto, F. (2021). “Spatiotemporal ecological study of COVID-19 mortality in the city of São Paulo, Brazil: shifting of the high mortality risk from areas with the best to those with the worst socioeconomic conditions,” *Travel medicine and infectious disease*, vol. 39, 101945.

[4] Bouezmarni, T., Rombouts, J., and Taamouti, A. (2010). "Asymptotic properties of the Bernstein density copula estimator for -mixing data," *Journal of Multivariate Analysis*, vol. 101, pp. 1-10.

[5] Chen, J. T., and Krieger, N. (2021). "Revealing the unequal burden of COVID-19 by income, race/ethnicity, and household crowding: US county versus zip code analyses," *Journal of Public Health Management and Practice*, vol. 27(1), pp. S43-S56.

[6] Chin, T., Kahn, R., Li, R., Chen, J.T., Krieger, N., Buckee, C.O., Balsari, S., and Kiang, M.V. (2020). "U.S. county-level characteristics to inform equitable COVID-19 response," medRxiv preprint.

[7] Deheuvels, P. (1979). "La fonction de dépendance empirique et ses propriétés. Un test non paramétrique d'indépendance," *Bulletins de l'Académie Royale de Belgique*, vol. 65(1), pp. 274-292.

[8] Ehlert, A. (2021). "The socio-economic determinants of COVID-19: a spatial analysis of German county level data," *Socio-Economic Planning Sciences*, 101083.

[9] Erreygers, G. and Van Ourti, T. (2011). "Measuring socioeconomic inequality in health, health care and health financing by means of rank-dependent indices: a recipe for good practice," *Journal of Health Economics*, vol. 30(4), pp. 685-694.

[10] Genest, C., Ghoudi, K., and Rivest, L. P. (1995). "A semiparametric estimation procedure of dependence parameters in multivariate families of distributions," *Biometrika,* vol. 82(3), pp. 543-552.

[11] Janssen, P., Swanepoel, J., and Veraverbeke, N. (2012). "Large sample behavior of the Bernstein copula estimator," *Journal of Statistical Planing and Inference.* vol. 142, pp. 1189-1197.

[12] Janssen, P., Swanepoel, J., and Veraverbeke, N. (2016). "Bernstein estimation for a copula derivative with application to conditional distribution and regression functionals," TEST vol. 25 (2), pp. 351-374.

[13] Knittel, C. R., and Ozaltun, B. (2020). "What does and does not correlate with COVID-19 death rates," (No. w27391). National Bureau of Economic Research.

[14] Krieger N, Chen, J.T, Waterman, P.D, Rehkopf, D.H, Subramanian, S.V. (2005). "Painting a truer picture of US socioeconomic and racial/ethnic health inequalities: The Public Health Disparities Geocoding Project," Am J Public Health, vol. 95: pp. 312-323.

[15] Krieger N, Chen, J.T, Waterman, P.D, Rehkopf, D.H, Subramanian, S.V. (2003). "Race/ethnicity, gender, and monitoring socioeconomic gradients in health: a comparison of area-based socioeconomic measures-the Public Health Disparities Geocoding Project," *American Journal of Public Health*, vol. 93, pp. 1655-1671.

[16] Krieger N, Chen J.T, Waterman P.D, Soobader M-J, Subramanian, S.V, Carson R. (2002). "Geocoding and Monitoring US Socioeconomic Inequalities in Mortality and Cancer Incidence: Does Choice of Area-Based Measure and Geographic Level Matter?-The Public Health Disparities Geocoding Project," *American Journal of Epidemiology*, vol. 156(5), pp. 471-82.

[17] Krieger N, Chen, J.T, Waterman, P.D, Soobader, M-J, Subramanian, S.V, Carson, R. (2003a). "Monitoring socioeconomic inequalities in sexually transmitted infections, tuberculosis, and violence: Geocoding and choice of area-based socioeconomic measures-The Public Health Disparities Geocoding Project (US)," *Public Health Reports*, vol. 118, pp. 240-260.

[18] Krieger N, Chen, J.T, Waterman, P.D, Soobader, M-J, Subramanian, S.V, Carson, R. (2003b). "Choosing area-based socioeconomic measures to monitor social inequalities in low birthweight and childhood lead poisoning –The Public Health Disparities Geocoding Project (US)," J*ournal of Epidemiology & Community Health*, vol. 57(3), pp. 186-199.

[19] Lassale, C., Gaye, B., Hamer, M., Gale, C. R., and Batty, G. D. (2020). "Ethnic disparities in hospitalisation for COVID-19 in England: The role of socioeconomic factors, mental health, and inflammatory and pro-inflammatory factors in a community-based cohort study," *Brain, behavior, and immunity*, vol. 88, pp. 44-49.

[20] Manner, H. (2007). "Estimation and model selection of copulas with an application to exchange rates," METEOR, Maastricht research school of Economics of Technology and Organizations.

[21] McLaren, J. (2021). "Racial disparity in COVID-19 deaths: Seeking economic roots with census data," *The BE Journal of Economic Analysis & Policy*, vol. 21(3), pp. 897-919

[22] Rao, J. N., & Molina, I. (2015). Small area estimation. John Wiley & Sons.

[23] Shih, J. H., and Louis, T. A. (1995). "Inferences on the association parameter in copula models for bivariate survival data," *Biometrics*, pp. 1384-1399.

[24] Sancetta, A. and Satchell, S. (2004). "The Bernstein copula and its applications to modeling and approximations of multivariate distributions," *Econometric Theory,* vol. 20, pp. 535-562.

[25] Schwarz, G. (1978). "Estimating the dimension of a model," *The Annals of Statistics*, PP. 461-464.

[26] Tsukahara, H. (2005). "Semiparametric estimation in copula models," *Canadian Journal of Statistics*, vol. 33(3), pp. 357-375.

[27] Wagstaff , A. (2002). "Inequality aversion, health inequalities and health achievement," *Journal of Health Economics,* vol. 21, pp. 627-641.

[28] Wagstaff, A. (2005). "The bounds of the concentration index when the variable of interest is binary, with an application to immunization inequality," *Health Economics*, vol. 14 (4), pp. 429-432.

[29] Wagstaff, A., Paci, P., and Van Doorslaer, E. (1991). "On the measurement of inequalities in health," *Social Science & Medicine* vol. 33 (5), pp. 545-557.

[30] Wagstaff, A., Van Doorslaer, E., and Paci, P. (1989). "Equity in the finance and delivery of health care: some tentative cross-country comparisons," *Oxford review of economic policy*, vol. 5(1), pp. 89-112.

[31] Wu, X., Nethery, R. C., Sabath, M. B., Braun, D., and Dominici, F. (2020). "Exposure to air pollution and COVID-19 mortality in the United States," medRxiv, 2020.04.05.20054502.

[32] Zheng, B. (2011). "A new approach to measure socioeconomic inequality in health," *The Journal of Economic Inequality,* vol. 9 (4), pp. 555-577.

9 Appendix: Proofs

This appendix contains the proofs of the main results in the text.

Poof of Proposition 1: Observe that

$$
\begin{aligned}
CH(p) &= \frac{\int_0^p \mathbb{E}\left[H \mid Y = F^{-1}(u)\right] du}{\int_0^1 \mathbb{E}\left[H \mid Y = F_Y^{-1}(u)\right] du} \\
&= \frac{\int_0^p \int_0^{+\infty} h f_H(h) c(F_H(h), u) dh du}{\int_0^1 \int_0^{+\infty} h f_H(h) c(F_H(h), u) dh du} \\
&= \frac{\int_0^{+\infty} h f_H(h) C_u(F_H(h), p) dh}{\int_0^{+\infty} h f_H(h) C_u(F_H(h), 1) dh} \\
&= \frac{\int_0^1 F_H^{-1}(u) C_u(u, p) du}{\mathbb{E}(H)},
\end{aligned}
$$

where the last equality is due to the fact that $\int_0^{+\infty} h f_H(h) C_u(F_H(h), 1) dh = \int_0^{+\infty} h f_H(h) dh = \mathbb{E}(H)$ since $C_u(u, 1) = 1$.

Poof of Theorem 1: We start with the following decomposition

$$
\widehat{CH}(p) - CH(p) = \frac{\hat{m}(p) - \bar{H} CH(p)}{\mathbb{E}(H)} + \frac{(\mathbb{E}(H) - \bar{H})(\hat{m}(p) - \bar{H}\, CH(p))}{\mathbb{E}(H)\bar{H}}, \tag{14}
$$

where $\bar{H}$ is the empirical mean of H and $\hat{m}(p) = n^{-1} \sum_{i=1}^n H_i C_u(\widehat{F}_H(H_i), p, \hat{\theta}_n)$. We study the first term in (14) since the second term is negligible with respect to the first term. Using Taylor expansion of C_u around $(F_H(H_i), p, \theta_0)$, we obtain

$$
\hat{m}(p) = n^{-1} \sum_{i=1}^n H_i\, C_u(F_H(H_i), p, \theta_0) + I_{n,1} + I_{n,2}, \tag{15}
$$

where

$$
I_{n,1} = n^{-1} \sum_{i=1}^n H_i\, C_{uu}(\tilde{F}_H(H_i), p, \tilde{\theta})(\widehat{F}_H(H_i) - F_H(H_i))
$$

and

$$
I_{n,2} = n^{-1} \sum_{i=1}^n H_i\, (\hat{\theta}_n - \theta_0)^\top C_{u,\theta}(\tilde{F}_H(H_i), p, \tilde{\theta}),
$$

with

$$
\tilde{F}_H(H_i) = F_H(H_i) + \alpha(\widehat{F}_H(H_i) - F_H(H_i)),\ \tilde{\theta} = \theta_0 + \alpha(\hat{\theta}_n - \theta_0),\ \alpha \in (0, 1).
$$

Under Assumptions B1 and B2, we have:

$$
I_{n,1} = n^{-1} \sum_{i=1}^n H_i\, C_{uu}(F_H(H_i), p, \theta_0)(\widehat{F}_H(H_i) - F_H(H_i)) + o_p(n^{-1/2}),
$$

and, from Assumption B4, we obtain

$$
\begin{aligned}
I_{n,2} &= n^{-1}\sum_{i=1}^{n} H_i\,(\hat{\theta}_n-\theta_0)^{\top} C_{u,\theta}(F_H(H_i),p,\theta_0)+o_p(n^{-1/2})\\
&= (\hat{\theta}_n-\theta_0)^{\top} n^{-1}\sum_{i=1}^{n} H_i\,C_{u,\theta}(F_H(H_i),p,\theta_0)+o_p(n^{-1/2})\\
&= (\hat{\theta}_n-\theta_0)^{\top}\mathbb{E}\left[H\,C_{u,\theta}(F_H(H),p,\theta_0)\right]+o_p(n^{-1/2})\\
&= \xi_i^{\top} r_\theta(p,\theta_0)+o_p(n^{-1/2}), \qquad (16)
\end{aligned}
$$

where, $r_\theta(p,\theta_0)=\mathbb{E}\left[H\,C_{u,\theta}(F_H(H),p,\theta_0)\right].$

Now, observe that $I_{1,n}$ is a V-statistic with the kernel $h_1(u,v,\theta_0)=\frac{1}{2}\left[h_2(u,v,\theta_0)+h_2(v,u,\theta_0)\right],$ where

$$
h_2(u,v,\theta_0)=uC_{uu}(F_H(u),p,\theta_0)(\mathbb{I}(v\le u)-F_H(u)).
$$

It can be shown that $\mathbb{E}(h_1(H_i,H_j,\theta_0))=0$, and from Assumption B3, we have $\mathbb{E}(h_1^2(H_i,H_j,\theta_0))<\infty$.

Hence, we obtain

$$
I_{n,1}=n^{-1}\sum_{i=1}^{n}\eta(H_i,p,\theta_0)+o_p(n^{-1/2}), \qquad (17)
$$

where

$$
\eta(H_i,p,\theta_0)=E_H\left[H\left(\mathbb{I}(H_i\le H)-F_H(H)\right)C_{uu}(F_H(H),p,\theta_0)\right].
$$

From (15), (16), and (17), we conclude the proof of Theorem 1.

Poof of Theorem 2: First, we recall the expression of the nonparametric estimator of $CH\,(p)$:

$$
\widehat{CH}_{m,n}(p)=\frac{\frac{1}{n}\sum_{i=1}^{n}H_i\,\widetilde{C}_u(\widehat{F}_H(H_i),p)}{\frac{1}{n}\sum_{i=1}^{n}H_i}=\frac{\tilde{m}(p)}{\bar{H}}.
$$

Note that from Janssen et al. (2016), and using the conditions of the theorem, we have

$$
\widetilde{C}_u(\widehat{F}_H(h),p)=\frac{1}{n}\sum_{i=1}^{n}W_{i,m}(F_H(h),p)+o_p(n^{-1/2}m^{1/4}).
$$

Using the above result, the numerator of $\widehat{CH}_{m,n}(p)$, can be written as:

$$
\begin{aligned}
\frac{1}{n}\sum_{j=1}^{n}H_j\widetilde{C}_u(\widehat{F}_H(H_j),p) &= \frac{1}{n}\sum_{j=1}^{n}\left[H_j\frac{1}{n}\sum_{i=1}^{n}W_{i,m}(F_H(H_j),p)\right]+o_p(n^{-1/2}m^{1/4})\\
&= \frac{1}{n}\sum_{i=1}^{n}\left[\frac{1}{n}\sum_{j=1}^{n}H_jW_{i,m}(F_H(H_j),p)\right]+o_p(n^{-1/2}m^{1/4})\\
&= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_H\left[HW_{i,m}(F_H(H),p)\right]+o_p(n^{-1/2}m^{1/4}). \qquad (18)
\end{aligned}
$$

The proof of Theorem 2 follows from Equation (18) and the following decomposition:

$$\widehat{CH}_{m,n}(p) - CH(p) = \frac{\tilde{m}(p) - \bar{H}CH(p)}{\mathbb{E}(H)} + \frac{(\mathbb{E}(H) - \bar{H})(\tilde{m}(p) - \bar{H}\,CH(p))}{\mathbb{E}(H)\bar{H}}.$$

Copula-based estimation of health inequality measures with an application to COVID-19: **Online Appendix**[*]

Taoufik Bouezmarni[†] Université de Sherbrooke

Mohamed Doukali[‡] University of East Anglia

Abderrahim Taamouti[§] University of Liverpool

January 26, 2023

[*] The authors would like to thank very much Serena Ng and the participants at the 2022 CISEM International Symposium on Statistics and Econometrics, 2022 Africa Meeting of the Econometric Society, and INSEA for their very useful comments.

[†] Département de Mathématiques, Université de Sherbrooke, CIREQ, CREAS. Email addresses: Taoufik.Bouezmarni@USherbrooke.ca

[‡] University of East Anglia, School of Economics, Norwich Research Park, NR4 7TJ, Norwich, United Kingdom email: m.doukali@uea.ac.uk

[§] *Corresponding author*. Department of Economics, University of Liverpool Management School. Address: Chatham St, Liverpool L69 7ZH. E-mail: Abderrahim.Taamouti@liverpool.ac.uk.

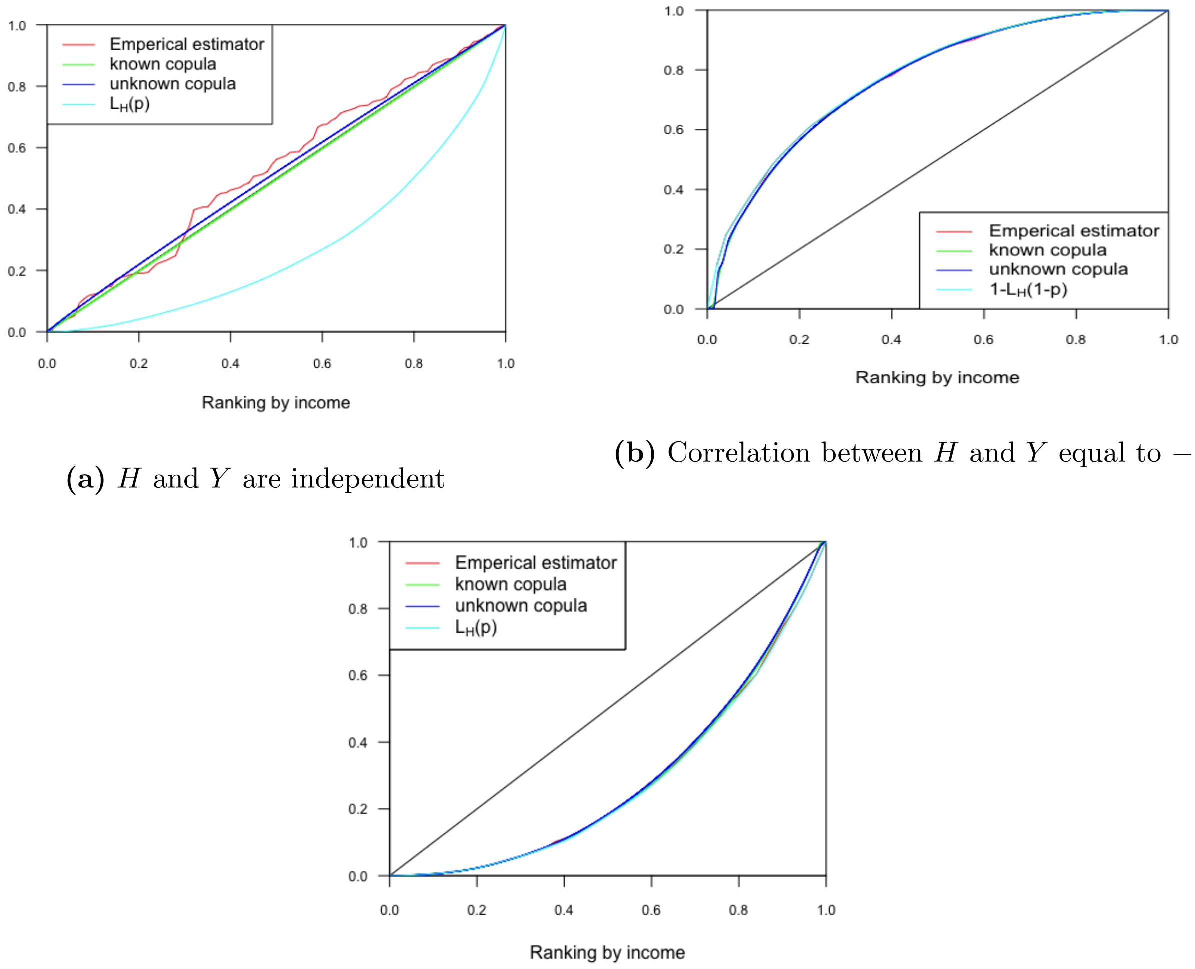
Copula-based estimation of health inequality measures with an application to COVID-19: Online Appendix

ABSTRACT

This online appendix contains all the simulation and empirical results stated in the main paper titled as "Copula-based estimation of health inequality measures with an application to COVID-19".

Keywords: Health concentration curve, Gini health coefficient, inequality, copula, semi/non-parametric estimators, COVID-19 infections and deaths.

Journal of Economic Literature classification: C13, C14, I14.



(a) H and Y are independent

(b) Correlation between H and Y equal to -0.99

(c) Correlation between H and Y equal to 0.99

Figure A1: *This figure illustrates the copula-based semiparametric estimator and the empirical estimator of the health concentration curve $CH(p)$ for different structures of dependence between H and Y.*

Table 1: Integrated Mean Squared Error of Semiparametric ($\widehat{CH}$), Empirical ($\widehat{CH}_n$) and Nonparametric ($\widehat{CH}_{m,n}$) estimators of CH. Simulation results based on 1000 replicates of $10^3 \times$ IMSE of the concentration health estimators.

Copula	τ	$n = 50$			$n = 100$			$n = 200$			$n = 500$		
		$\widehat{CH}$	$\widehat{CH}_n$	$\widehat{CH}_{m,n}$	$\widehat{CH}$	$\widehat{CH}_n$	$\widehat{CH}_{m,n}$	$\widehat{CH}$	$\widehat{CH}_n$	$\widehat{CH}_{m,n}$	$\widehat{CH}$	$\widehat{CH}_n$	$\widehat{CH}_{m,n}$
Gaussian	-0.4	1.52405	2.37677	2.14225	0.81371	1.15701	1.06885	0.38623	0.58638	0.53311	0.12977	0.22836	0.22454
	0.001	2.10629	3.44384	1.88307	1.07944	1.68181	1.00998	0.50564	0.85069	0.49509	0.19713	0.32445	0.19456
	0.01	2.13191	3.38774	1.85736	1.06366	1.73533	0.96506	0.51634	0.82562	0.51147	0.20159	0.32712	0.20469
	0.1	1.79796	3.20303	1.85608	0.87136	1.62803	1.04206	0.41574	0.81374	0.51939	0.17935	0.31983	0.22626
	0.2	1.50939	3.10968	1.86665	0.79499	1.50698	0.97972	0.46025	0.77342	0.51299	0.18221	0.30089	0.25418
	0.5	1.45913	1.98083	2.04345	0.84333	1.06459	1.17494	0.36143	0.51315	0.54520	0.10408	0.19161	0.16909
	0.7	1.08239	1.22984	1.44666	0.64150	0.64480	0.85689	0.22934	0.31696	0.27078	0.08308	0.12031	0.11829
Student	-0.4	1.50998	2.41594	2.09419	0.91477	1.32553	1.14913	0.45419	0.64521	0.57807	0.15306	0.25293	0.23254
	0.001	2.69858	3.84649	2.17324	1.34693	1.89724	1.13122	0.66351	0.98423	0.62257	0.22396	0.40470	0.24812
	0.01	2.79426	3.90271	2.18947	1.37388	1.85705	1.17449	0.68979	1.00119	0.62375	0.21079	0.38427	0.25512
	0.1	2.34374	3.67801	2.05408	1.15942	1.84015	1.14661	0.52027	0.95107	0.60335	0.20992	0.38355	0.26273
	0.2	2.00519	3.63507	2.12335	1.04043	1.70332	1.15646	0.53195	0.86679	0.61781	0.20103	0.34301	0.28498
	0.5	1.41260	2.04994	2.02757	0.74987	1.02980	1.01611	0.36943	0.48307	0.44746	0.15646	0.2214	0.24387
	0.7	1.05205	1.27782	1.44762	0.57896	0.62840	0.75690	0.34523	0.36431	0.48286	0.12641	0.14386	0.21322
Clayton	0.001	2.09081	3.42189	1.87387	1.00368	1.62497	0.91140	0.53735	0.83230	0.50405	0.21267	0.34863	0.20759
	0.01	2.04309	3.39695	1.87465	1.06874	1.62283	1.01943	0.51462	0.84717	0.51277	0.21135	0.34748	0.21627
	0.1	1.95105	3.44810	1.91312	0.91821	1.61878	1.02497	0.38758	0.82406	0.50239	0.13228	0.32879	0.20838
	0.2	1.55803	3.16271	1.76030	0.76749	1.56814	1.00488	0.34245	0.79809	0.49569	0.13989	0.31672	0.21340
	0.5	1.05728	2.33026	1.46983	0.46329	1.13629	0.88551	0.27416	0.55912	0.45783	0.09539	0.24638	0.20293
	0.7	0.84169	1.63300	1.34834	0.42886	0.76142	0.66166	0.20152	0.40266	0.35363	0.07494	0.16145	0.15101
Gumbel	0.001	2.23206	3.52253	1.96451	1.10604	1.72069	1.01641	0.51837	0.83419	0.50979	0.22490	0.36063	0.21245
	0.01	2.17886	3.53497	1.87014	1.02629	1.63997	0.98007	0.54707	0.83069	0.49453	0.21463	0.34587	0.20453
	0.1	2.08629	3.47387	1.99008	1.11354	1.71672	1.15681	0.54978	0.89121	0.60417	0.21383	0.35531	0.27247
	0.2	2.06448	3.18206	2.31921	1.01851	1.53778	1.27509	0.56014	0.81521	0.70188	0.20091	0.30514	0.39143
	0.5	1.35059	1.81019	1.91188	0.77963	0.94400	1.04940	0.37581	0.46746	0.50481	0.13160	0.18465	0.17968
	0.7	1.08024	1.15756	1.47898	0.62124	0.64797	0.85754	0.31388	0.32589	0.40235	0.09029	0.11647	0.11790

Table 2: Summary statistics

	Mean	**Std Dev.**	**Min**	**Max**
Cases	0.098	0.0299	0.0029	0.375
Deaths	0.019	0.010	0.000	0.102
Population	106542	341079	66	10081570
Population density	214.77	778.96	0.000	17179
Poverty rate	0.143	0.058	0.023	0.547
Median income	27776	5850	8641	70390
Gini Index	0.444	0.035	0.302	0.609
Share of African Americans	0.089	0.144	0	0.872
Share of White Americans	0.756	0.203	0.006	1
Share of Asian Americans	0.013	0.028	0	0.417

Note: This table provides the descriptive statistics of COVID-19's infection and death rates and socioeconomic variables for 2777 counties of 45 U.S. states. Data on COVID-19's cases and deaths rates are obtained from U.S. Centers for Disease Control and Protection (CC) and covers the period until June 10, 2021. Demographic variables are drawn primarily from the US Census and CDC.

Table 3: Correlation coefficients between socio-economic variables

	Density	**Median income**	**Population**	**Poverty rate**
Median income	0.296			
Population	0.366	0.227		
Poverty rate	-0.052	-0.742	-0.076	
Share of African Americans	0.128	-0.263	0.057	0.441
Share of White Americans	-0.185	0.194	-0.224	-0.475

Note: The table shows the correlation coefficients matrix of socioeconomic variables we use in our empirical analysis. These variables are described in Table 2.

Table 4: Regressions for COVID1-19 deaths

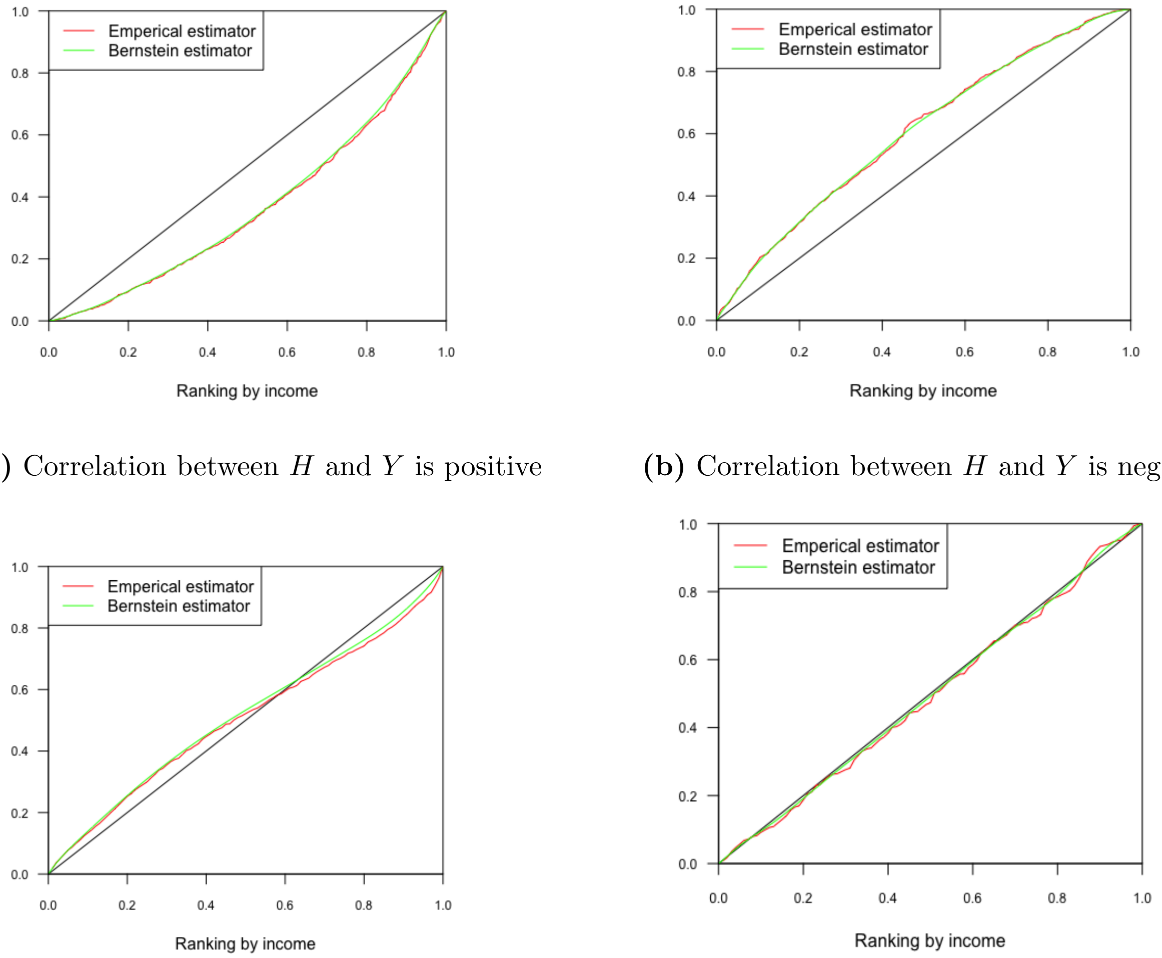
	Dependent variable: Log(COVID-19's deaths)						
	(1)	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**	**(7)**
Constant	7.87^{***} (0.17)	6.63^{***} (0.60)	-6.61^{**} (3.46)	-7.47^{**} (3.57)	-3.82 (3.60)	-6.64^{**} (3.61)	-7.52^{**} (3.70)
Population	0.13^{***} (0.01)	0.12^{***} (0.01)	0.08^{***} (0.017)	0.08^{***} (0.01)	0.09^{***} (0.01)	0.09^{***} (0.01)	0.09^{***} (0.015)
Poverty rate		10.19^{**} (4.79)	0.93 (4.79)	0.037 (4.87)	-0.872 (4.69)	-7.78 (4.89)	
Gini index			31.6^{**} (8.18)	33.5^{***} (8.39)	20.54^{***} (8.42)	30.36^{***} (8.53)	29.8^{***} (7.57)
Gini health index				1.68 (1.68)		4.00^{**} (1.92)	3.87^{**} (1.85)
Share of African Americans					2.76^{**} (1.34)	1.57 (1.34)	1.41 (1.33)
Share of Asian Americans						-6.91^{***} (2.31)	-7.24^{***} (2.31)
Median income							5.36^{**} (2.94)

Note: This table reports the estimation results of regressing state-level COVID-19's death on state's population size, state's poverty rate, state's median income, state's Gini (economic) index, state's Gini health index, state's share of African Americans, and state's share of Asian Americans,. The dependent variable is log (COVID-19's deaths). Standard errors in parentheses. *p<0.1; **p<0.05; ***p<0.01.

Table 5: Regressions for COVID1-19 infections

	Dependent variable: log(COVID-19's cases)						
	(1)	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**	**(7)**
Constant	12.1^{***} (0.14)	11.3^{***} (0.50)	4.62 (3.17)	3.69 (3.25)	7.09^{**} (3.30)	3.75 (3.08)	3.39 (3.22)
Population	0.11^{***} (0.01)	0.11^{***} (0.01)	0.09^{***} (0.01)	0.09^{***} (0.01)	0.09^{***} (0.015)	0.10^{***} (0.01)	0.10^{***} (0.01)
poverty rate		6.35 (3.95)	1.66 (4.38)	0.69 (4.44)	0.05 (4.31)	-7.84 (4.19)	
Gini index			16.0^{**} (7.48)	18.0^{**} (7.64)	10.5 (7.72)	20.1^{***} (7.30)	15.8^{**} (6.58)
Gini health index				1.82 (1.53)		4.73^{***} (1.64)	4.42^{***} (1.61)
Share of African Americans					2.45^{**} (1.23)	1.05 (1.14)	0.92 (1.15)
Share of Asian Americans						-7.84^{***} (1.97)	-7.79^{***} (2.01)
Median income							4.41^{**} (2.55)

Note: This table reports the estimation results of regressing state-level COVID-19's infections (cases) on state's population size, state's poverty rate, state's median income, state's Gini (economic) index, state's Gini health index, state's share of African Americans, and state's share of Asian Americans. The dependent variable is log (COVID-19's cases). Standard errors in parentheses. *p<0.1; **p<0.05; ***p<0.01.



(a) Correlation between H and Y is positive

(b) Correlation between H and Y is negative

(c) Correlation between H and Y changes sign

(d) H and Y are independent

Figure A2: *This figure illustrates the Bernstein copula-based estimator and the empirical estimator of the health concentration curve $CH(p)$ for different degrees of dependence between H and Y.*

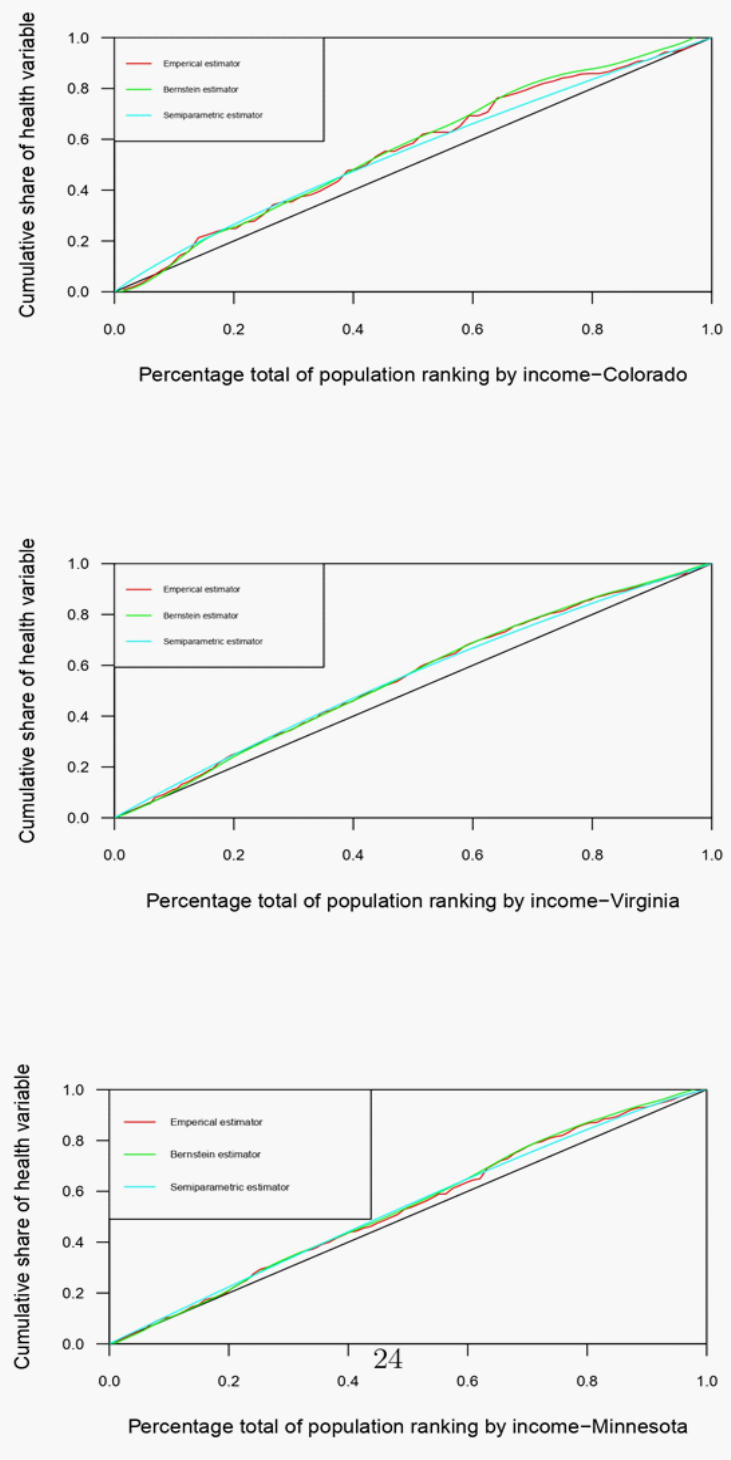


Figure A3: The empirical estimator, the semiparametric and the nonparametric estimator of the health concentration curve for COVID-19deaths. (high income states).

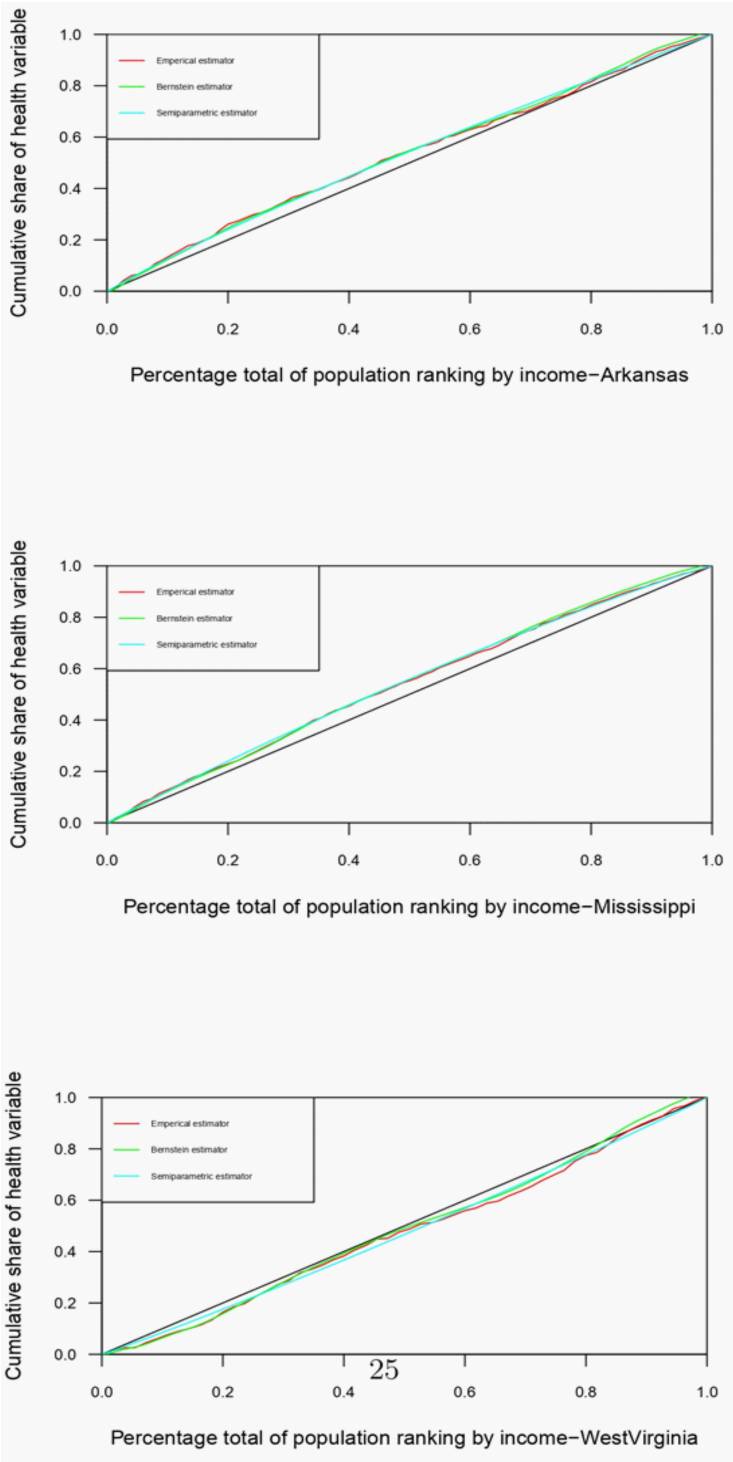


Figure A4: The empirical estimator, the semiparametric and the nonparametric estimator of the health concentration curve for COVID-19deaths. (low income states).

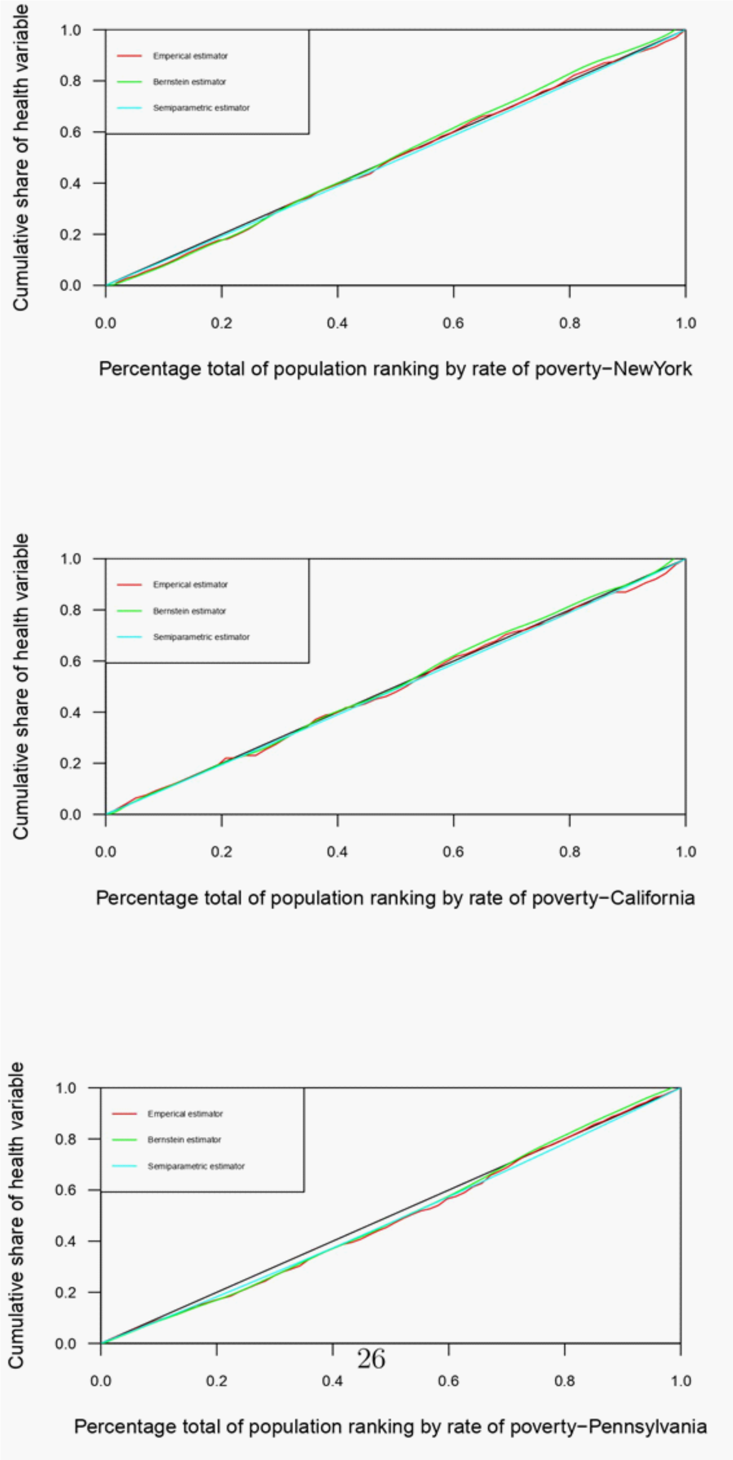


Figure A5: The empirical estimator, the semiparametric and the nonparametric estimator of the health concentration curve for COVID-19deaths. (average income states)

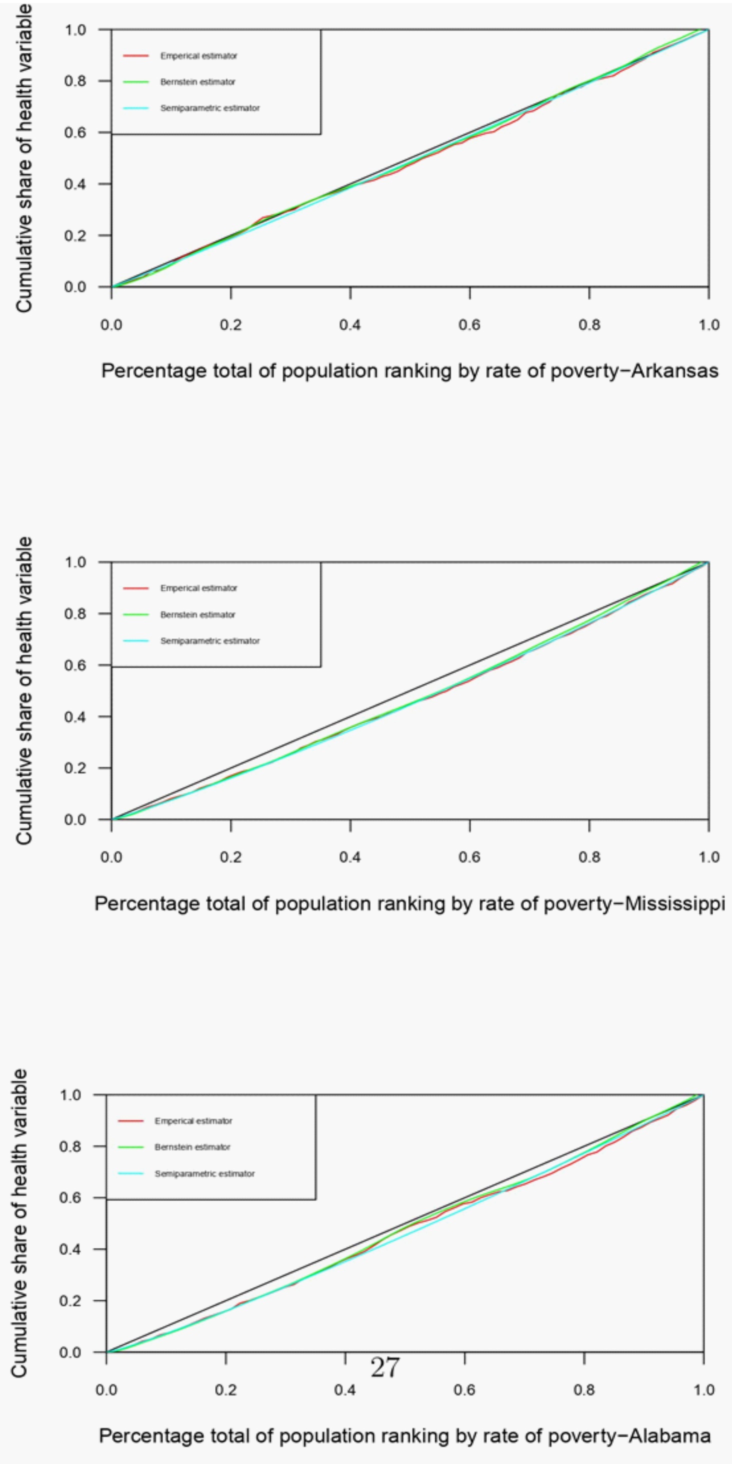


Figure A6: The empirical estimator, the semiparametric and the nonparametric estimator of the health concentration curve for COVID-19deaths. (high poverty rate states).

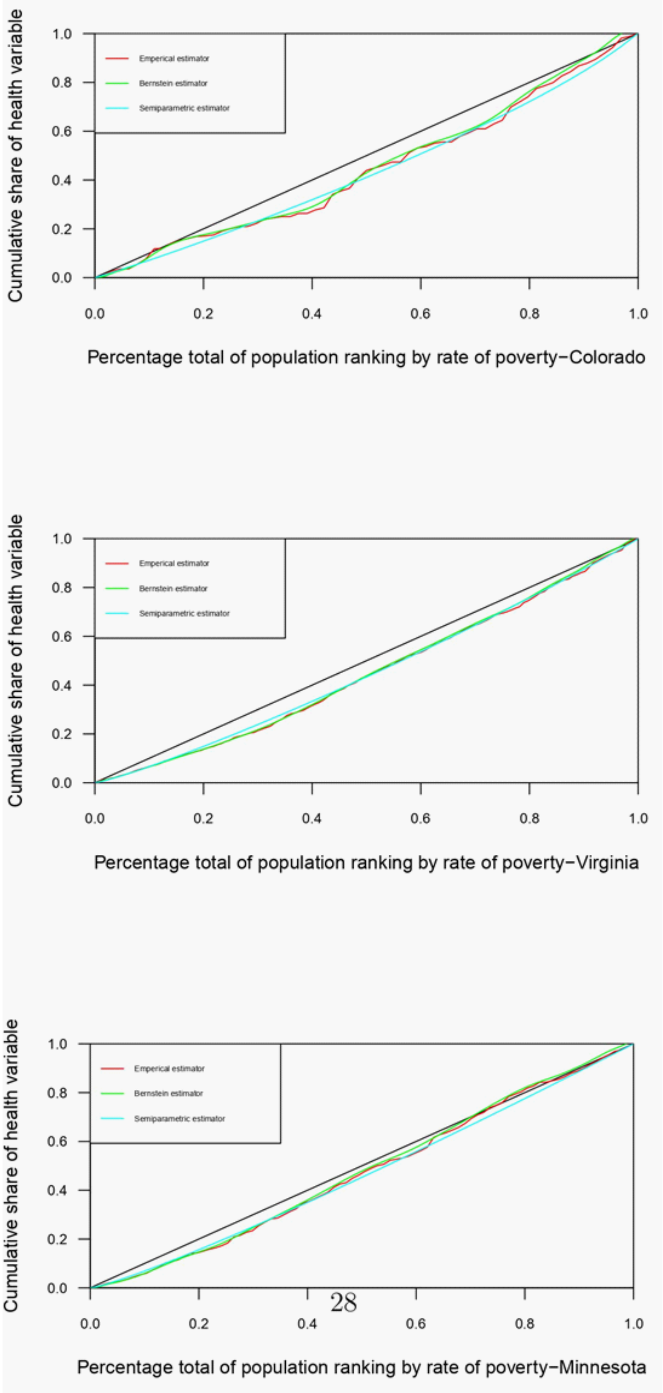


Figure A7: The empirical estimator, the semiparametric and the nonparametric estimator of the health concentration curve for COVID-19deaths. (low poverty rate states).

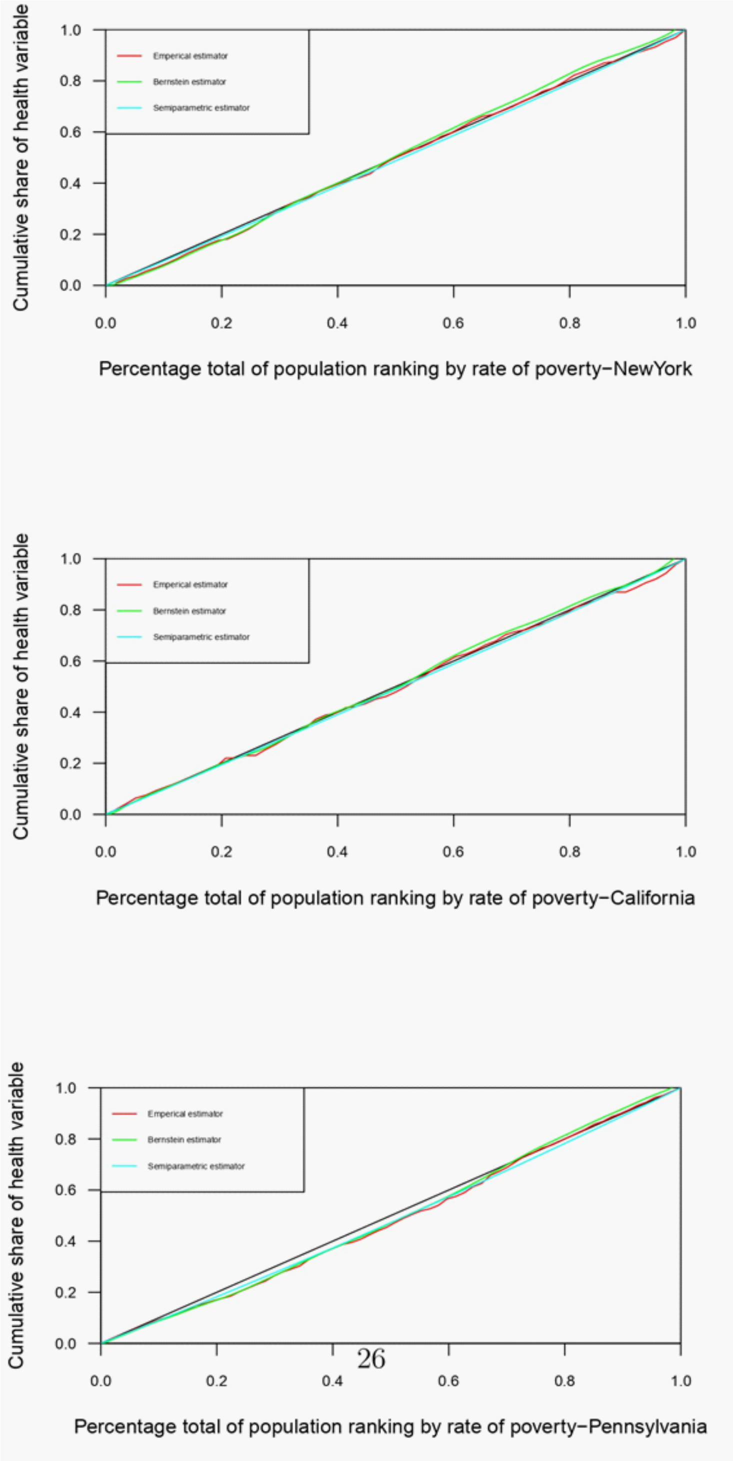


Figure A8: The empirical estimator, the semiparametric and the nonparametric estimator of the health concentration curve for COVID-19deaths. (average poverty rate states).

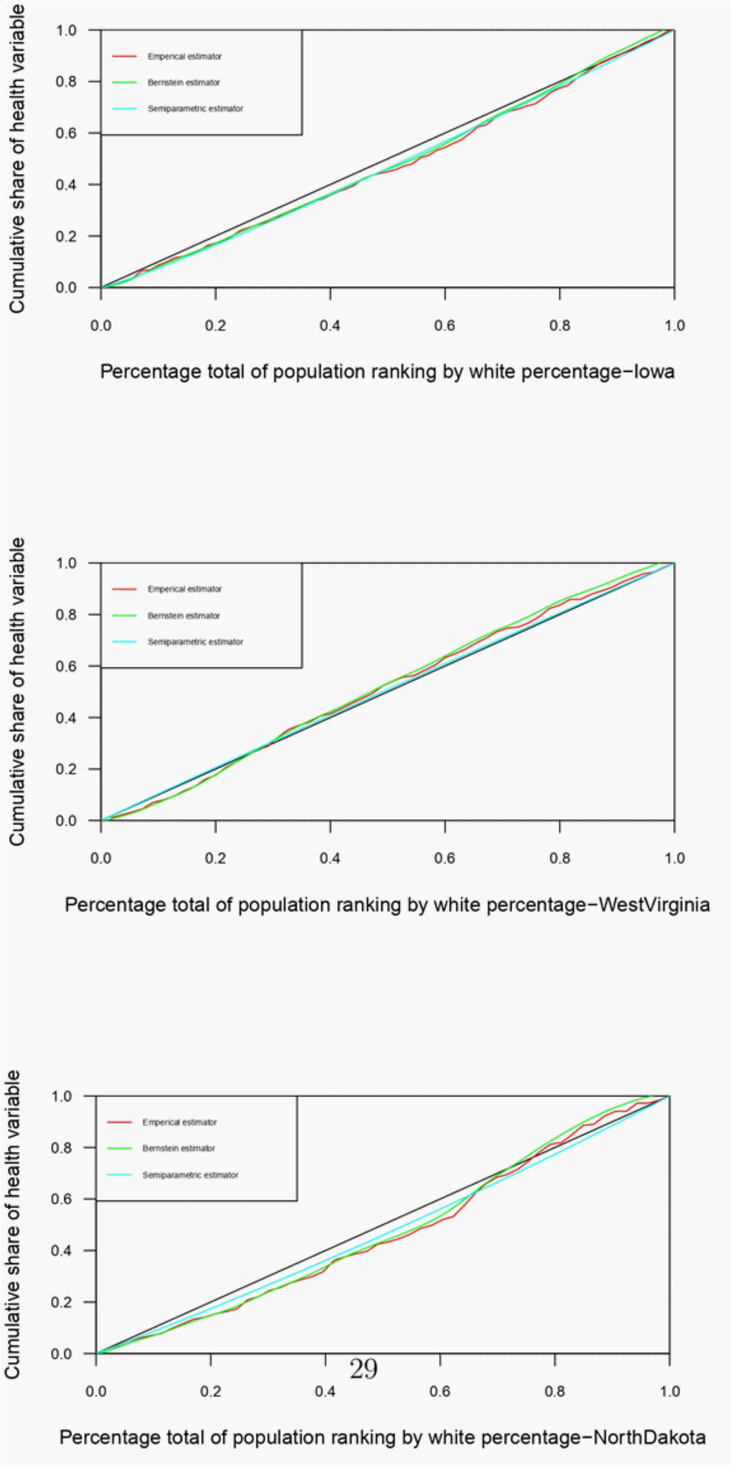


Figure A9: The empirical estimator, the semiparametric and the nonparametric estimator of the health concentration curve for COVID-19deaths. (high rate of white people states).

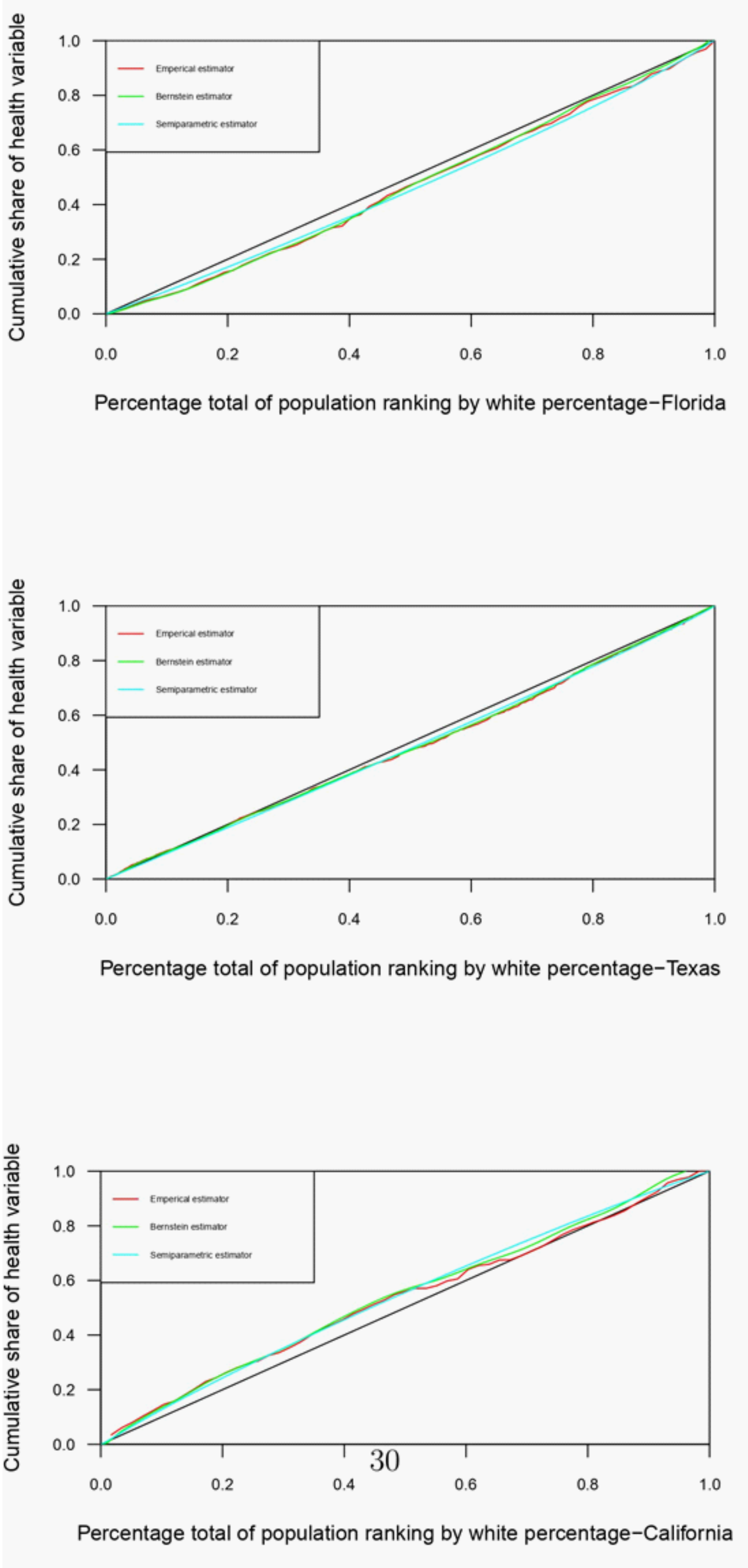


Figure A10: The empirical estimator, the semiparametric and the nonparametric estimator of the health concentration curve for COVID-19deaths. (low rate of white people states).

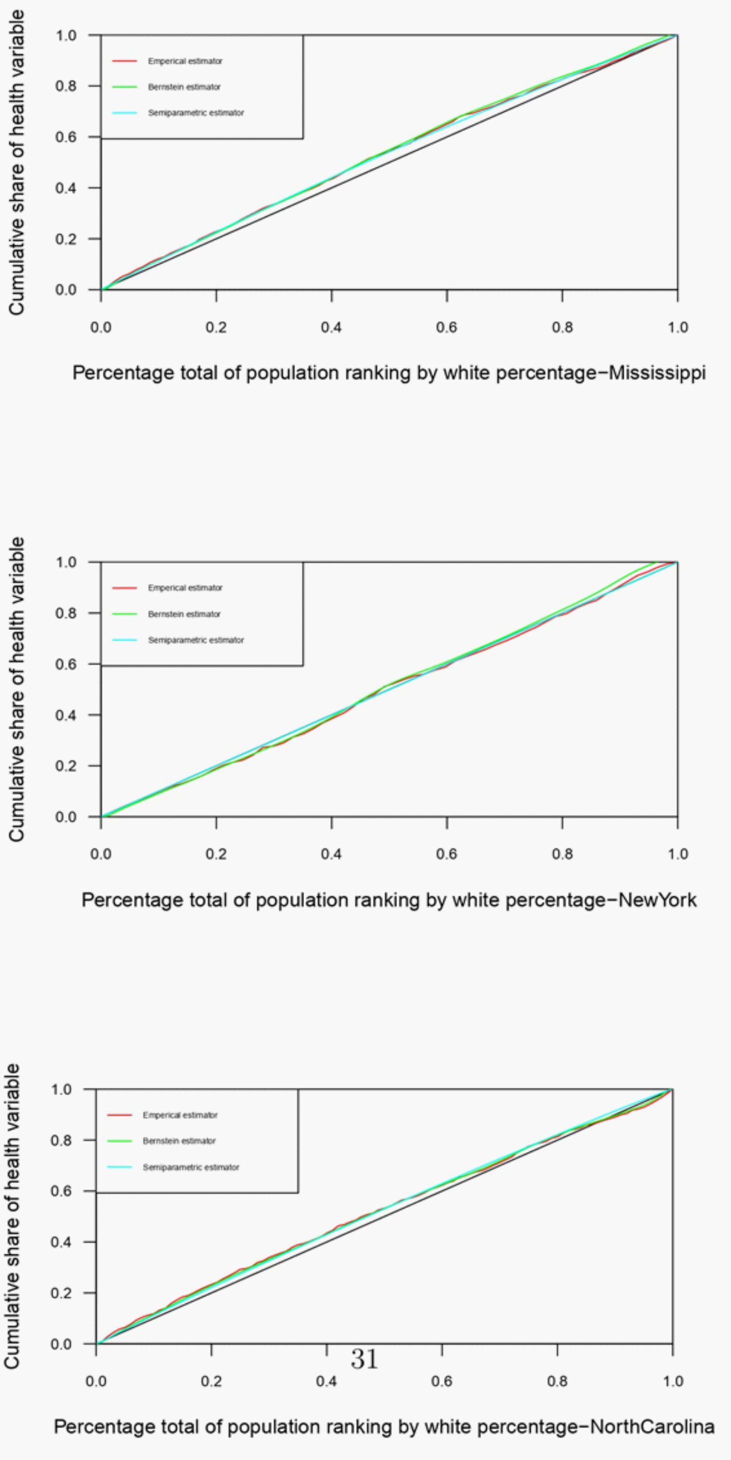
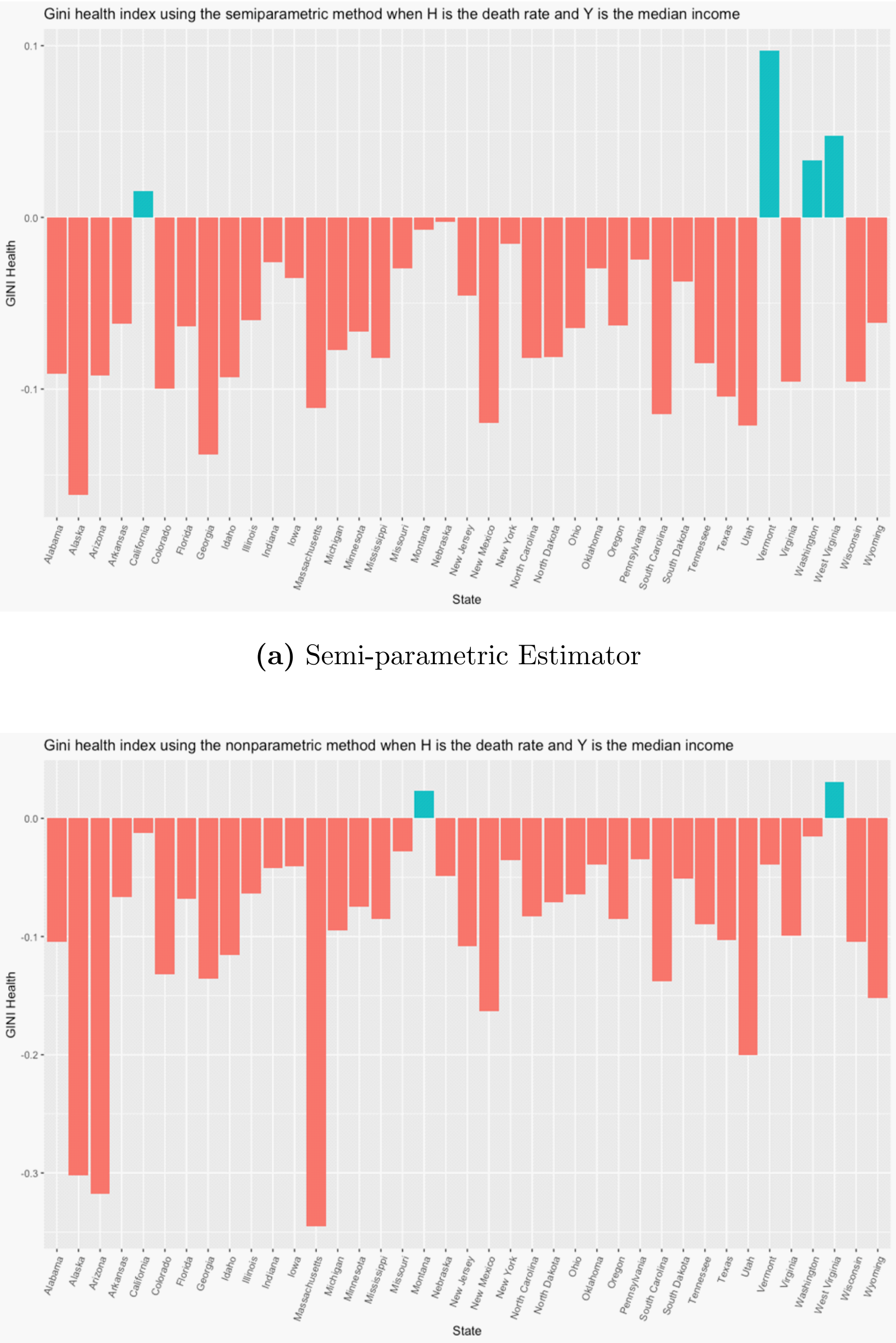


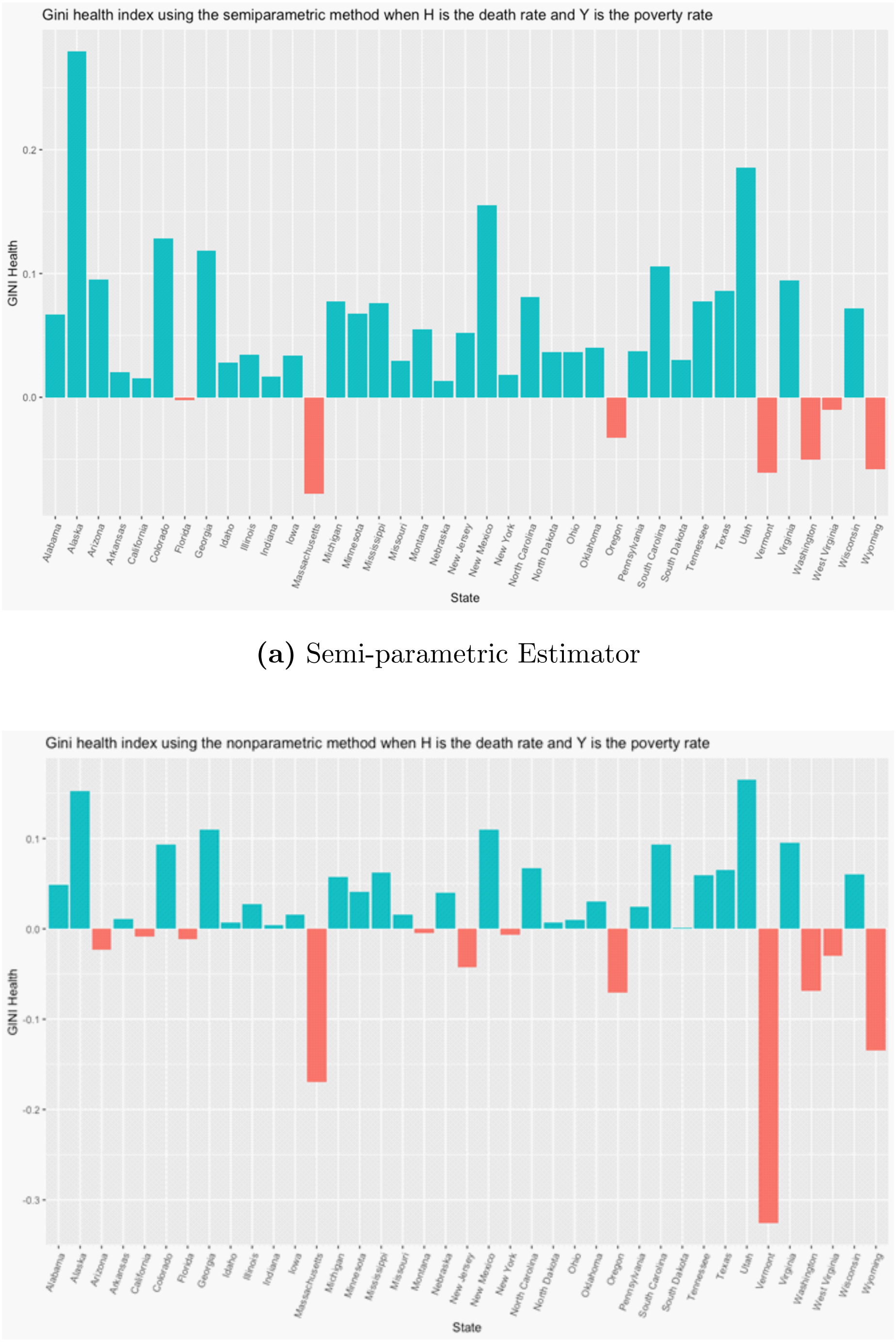
Figure A11: The empirical estimator, the semiparametric and the nonparametric estimator of the health concentration curve for COVID-19deaths. (average rate of white people states).



(a) Semi-parametric Estimator

(b) Non-parametric Estimator

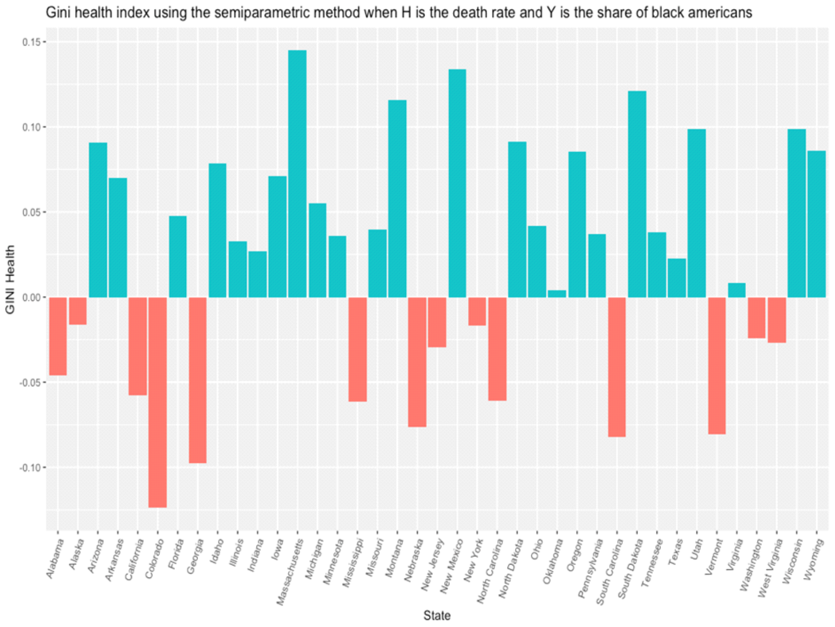
Figure A12: *The figure reports the semi/non-parametric estimators of Gini health index for 40 U.S. states when the health variable H represents COVID-19's death rates and the socioeconomic variable Y is the median income.*



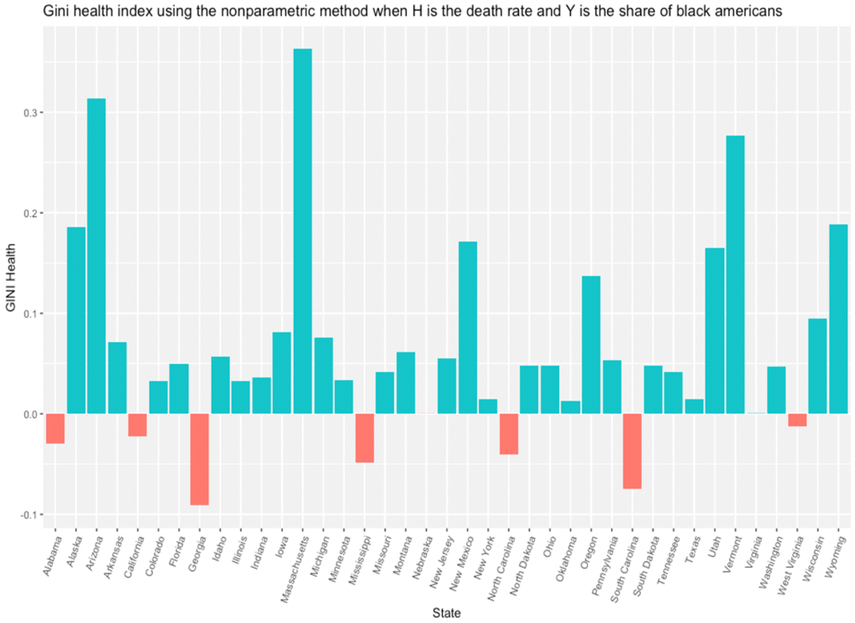
(a) Semi-parametric Estimator

(b) Non-parametric Estimator

Figure A13: *The figure reports the semi/non-parametric estimators of Gini health index for 40 U.S. states when the health variable H represents COVID-19's death rates and the socioeconomic variable Y is the poverty rate.*



(a) Semi-parametric Estimator



(b) Non-parametric Estimator

Figure A14: *The figure reports the semi/non-parametric estimators of Gini health index for 40 U.S. states when the health variable H represents COVID-19's death rates and the socioeconomic variable Y is the share of African Americans.*