Estimating the Empirical Lorenz Curve and Gini Coefficient in the Presence of Error

Chaya S. Moskowitz*   E. S. Venkatraman†

Elyn Riedel‡   Colin B. Begg**

*Memorial Sloan-Kettering Cancer Center, moskowc1@mskcc.org

†

‡

**

Estimating the Empirical Lorenz Curve and Gini Coefficient in the Presence of Error

Chaya S. Moskowitz, E. S. Venkatraman, Elyn Riedel, and Colin B. Begg

Abstract

The Lorenz curve is a graphical tool that is widely used to characterize the concentration of a measure in a population, such as wealth. It is frequently the case that the measure of interest used to rank experimental units when estimating the empirical Lorenz curve, and the corresponding Gini coefficient, is subject to random error. This error can result in an incorrect ranking of experimental units which inevitably leads to a curve that exaggerates the degree of concentration (variation) in the population. We explore this bias and discuss several widely available statistical methods that have the potential to reduce or remove the bias in the empirical Lorenz curve. The properties of these methods are examined and compared in a simulation study. This work is motivated by a health outcomes application which seeks to assess the concentration of black patient visits among primary care physicians. The methods are illustrated on data from this study.

Estimating the Empirical Lorenz Curve and Gini Coefficient in the Presence of Error

Chaya S. Moskowitz *

E. S. Venkatraman

Elyn R. Riedel

Colin B. Begg

Department of Epidemiology and Biostatistics

Memorial Sloan-Kettering Cancer Center

307 East 63rd Street, 3rd Floor, New York, NY 10021

*corresponding author, e-mail: moskowc1@mskcc.org

Abstract

The Lorenz curve is a graphical tool that is widely used to characterize the concentration of a measure in a population, such as wealth. It is frequently the case that the measure of interest used to rank experimental units when estimating the empirical Lorenz curve, and the corresponding Gini coefficient, is subject to random error. This error can result in an incorrect ranking of experimental units which inevitably leads to a curve that exaggerates the degree of concentration (variation) in the population. We explore this bias and discuss several widely available statistical methods that have the potential to reduce or remove the bias in the empirical Lorenz curve. The properties of these methods are examined and compared in a simulation study. This work is motivated by a health outcomes application which seeks to assess the concentration of black patient visits among primary care physicians. The methods are illustrated on data from this study.

1

1 INTRODUCTION

The Lorenz curve is a graphical statistic that was first introduced in 1905 as a tool for exhibiting the concentration of wealth in a population [1]. In this context members of the population are ranked in terms of their wealth and the cumulative wealth is plotted (on the y-axis) against the cumulative proportion of the population (on the x-axis). One can then select any quantile to characterize concentration using a statistic such as "Y% of the wealth is owned by X% of the population." Alternatively a summary index of concentration, the Gini coefficient [2], is frequently used.

Both the Lorenz curve and Gini coefficient have been primarily utilized in the economic and social sciences over the last century. In recent years, however, these methods have also seen application in other areas such as medical and health services research. For example, the Lorenz curve has been used to describe patterns of drug use. Hallas and Støvring [3] use the Lorenz curve to show that in 2003 in the County of Funen in Denmark 1% of users of opiod analgesics accounted for 19.3% of opioid consumption whereas 1% of users of insulin accounted for 4.7% of insulin consumption. They infer that there are not many heavy users of insulin, but, in contrast, there is a group of heavy users of opioid analgesics. The Lorenz curve and Gini coefficient have also been used to explore the distribution of health professionals in relation to the population distribution of patients. Chang and Halfon [4] examined the pediatrician-to-child ratios in the 50 states and showed a fourfold difference between the states with the highest (Maryland) and lowest (Idaho) ratios. The authors used Lorenz curve analyses to show that the concentration was greater among pediatricians than among all physicians, and that during the period 1982-1992, despite a 46% increase in the number of pediatricians nationwide, there was essentially no change in the national distribution, as evidenced by a minimal change in the Gini coefficient. Similar kinds of studies have been conducted by Brown [5] in Alberta who examined and compared the distributions of various kinds of health practitioners, and by Kobayashi and Takaki [6] in Japan who examined the

2

distribution of general physicians across 3,268 municipal entities.

As is described in further detail below, the estimation of both the Lorenz curve and the Gini coefficient involves ranking the units of observation on the basis of some quantity of interest and then estimating cumulative proportions. When there is error or variation in the measurement of the quantity of interest, an analysis that does not account for this error or variation may incorrectly rank the units and result in biased estimates for the Lorenz curve and Gini coefficient. This situation can arise in many different circumstances. For instance, the Lorenz curve has traditionally been used to study the distribution of income in a population. In this case, depending upon the study design, there may be variation in the measurement of income from a number of possible sources including error in the reported income (i.e. measurement error) and variation attached to estimating income if the income must be estimated for each member of the population.

In this manuscript we are particularly interested in the bias that may occur in a specific type of data configuration that can occur frequently in practice. This configuration is nested in that members of the population, or experimental units, are the primary units of analysis, but within each experimental unit there are multiple observations that are aggregated to form the outcome whose distribution is of interest. For example, using Lorenz curves and Gini coefficients Prakasam and Murthy [7] look at couples within states in India as the unit of analysis to explore the acceptance of family planning methods for different levels of literacy. Tickle [8] divides the North West Region of England into market areas and studies children in these areas who were included in the 1995/6 NHS epidemiology survey to examine the distribution of the frequency of dental caries using Lorenz curves and Gini coefficients. Elliot *et al.* [9] study geographical variations of sexually transmitted diseases across regions of Manitoba, Canada by looking at individuals within each region who have a reported infection. Nishiura *et al.* [10] use data on 76 provinces in Thailand to estimate the distribution of physicians, nurses, and hospital beds across the population.

3

Within this framework, our motivation for this work arises from a study of whether black patient visits to physicians are concentrated among a select group of physicians. Described in more detail below, this analysis involves ranking physicians (the primary units of analysis) on the basis of the proportion of patient visits to each physician that were made by black patients (that is, the patient visits constitute the multiple observations to be aggregated within physician). In this study one observes the number of black patient visits in a sample of patient visits for each physician.

To understand the problem, consider the fact that the goal of this research is to evaluate the extent to which care of black patients is concentrated within the population of physicians. Consider, for example, the extreme hypothesis that patients receive care from physicians randomly, i.e. without regard for race. Then, in the long run, the proportion of black patient visits in each physician's profile should converge to the same proportion, the population relative frequencies of black visits. However, since the Lorenz curve is constructed after ranking the physicians on the basis of the observed proportion of blacks, an observed "concentration" of blacks in a subset of the physicians is necessarily observed, with the degree of observed concentration increasing as the number of patients sampled per physicians decreases. In other words, the degree of concentration increases as the error variance increases. Thus, the "empirical" curve not only reflects the inherent degree of maldistribution of patients to doctors, it is also systematically influenced by the sample sizes one elects to use. That is, if we studied the doctors for two years rather than one year, we would inevitably observe less "concentration." Our goal in this article is to redefine the problem in such a way that the definition of "concentration" that we are endeavoring to estimate is invariant to the sampling used in the design of the study.

To our knowledge, this problem has been explicitly recognized by a limited number of investigators. Lee [11] compared the potential for "bias" in this setting with the inherent bias in estimating predictive accuracy for a model when the same dataset is used to both build and test the model, an

issue that has been studied widely by statistical theorists. Lee proposed the generation of bootstrap samples to reorder the experimental components while using the original data to estimate the Gini coefficient. Pham-Gia and Turkkan [12] proposed a parametric approach in the context of a study of inequality in income distributions when the incomes are measured with error, and in principle this has the potential to resolve the bias issue. In their formulation, they reason that *true income = observed income + error* and model the observed income with a parametric distribution and the error with a separate parametric distribution. They derive theoretical results for obtaining the distribution of the true income and the corresponding Lorenz curve assuming the observed income and error are independent and follow either beta or gamma distributions.

We are interested in "nonparametric" methods for estimating a Lorenz curve and Gini coefficient in this type of data configuration, where the methods are nonparametric in the sense that the ranking of the experimental units is unconstrained. The method by Pham-Gia and Turkkan is not directly applicable to the nested data configuration we consider and the assumption that the error and the observed concentration of observations are independent may not reasonable. While the bootstrap approach is a nonparametric method, we show that it has features that limit its applicability to data of this nature. In addition to the bootstrap, we consider three other approaches using random effects models for estimating the Lorenz curve and Gini coefficient.

In the next section, we describe the health services application that was our motivation for undertaking this work in more detail. In Section 3, we define the Lorenz Curve and Gini coefficient and discuss the potential bias in more detail. In Section 4, we present several different analytic strategies that can be used to estimate the Lorenz Curve and Gini coefficient. Section 5 contains the results of a simulation study evaluating these methods. In Section 6 we return to the health services application and use this dataset to illustrate the different methods. We make our concluding remarks in Section 7.

5

2   MOTIVATING APPLICATION

In a recently published article, Bach et al. [13] explore whether the quality of health care received by patients in the United States is associated with race. The authors study a sample of Medicare patients together with the patients' primary care physicians in order to assess if black patients receive a lower quality of care than do white patients. The analysis involved several different components, including a Lorenz curve type analysis studying whether patient visits made by black patients to their physicians were concentrated among a select group of physicians.

Specifically, they examine a sample of 5% of black and white Medicare beneficiaries who were treated during 2001 by 4,355 primary care physicians who participated in the 2000-2001 Community Tracking Study survey. The distribution of the black patients across these physicians was characterized by a Lorenz curve in which the cumulative proportion of black patient visits was plotted against the cumulative proportion of physicians in the population. A curve was constructed by ranking the physicians on the basis of the proportions of blacks in the physicians' profiles. In this case, the proportions of blacks patient visits in a physician's profile had to be estimated from the actual proportions of their patients who made office visits during 2001, submitted a suitable claim to Medicare, and were included in the 5% Medicare sample. Consequently, the observed proportion of blacks in an individual physician's profile varies considerably from the actual proportion that characterizes the physician's profile of patients.

3   LORENZ CURVE AND GINI COEFFICIENT

Suppose that N observations (patient visits) are dispersed among n experimental units (physicians). We represent the number of observations for each experimental unit as m_k, $k = 1, ..., n$. Interest lies in studying the concentration or distribution of a feature of each of the N observations across the

6

n members of the population. Let the individual observations be denoted by Y_{jk}, $j = 1, ..., m_k$, and let X_k denote the value of some summary for the kth experimental unit. In terms of our example, N is the number of patients in the sample seen by the n primary care physicians. Y is a binary indicator of whether or not a patient is black, and $X_k = \sum_{j=1}^{m_k} Y_{jk}$ is the number of black patients in the profile of the kth physician.

The Lorenz curve is constructed by ranking the units in terms of the measure of concentration, in our case the proportion of black patients in the physicians' profiles. Let $\tilde{t}_k = X_k/m_k$, $k = 1, ..., n$, be the observed proportions, and for notational convenience let the n units be ranked in ascending order on the basis of $\tilde{t}_k$. Suppose that $t_k = E(\tilde{t}_k)$ is the expected value of $\tilde{t}_k$. Let the n experimental units be conceptually re-ordered on the basis of these unknown proportions $\{t_k\}$, and let $\{k^*\}$ denote this re-ordering. Further, set

$$L(t) = \frac{\sum_{t_{k^*} \leq t} m_{k^*} t_{k^*}}{\sum_{k^*=1}^{n} m_{k^*} t_{k^*}} \quad \text{and} \quad G(t) = \frac{\sum_{t_{k^*} \leq t} m_{k^*}}{N}. \tag{1}$$

Then the underlying true Lorenz curve is a plot of $L(t)$ versus $G(t)$, i.e. it is a plot of the fraction of all black patients seen by the k physicians with the lowest proportions of black patients in their patient profiles against the fractions of all patients seen by the k physicians. (We note that this definition conditions, for convenience, on the observed sample sizes in each experimental unit.)

In practice the Lorenz curve is constructed empirically. That is, the experimental units are ranked on the basis of the observed proportions $\tilde{t}_k$, $k = 1, ..., n$. In this case, the axes of the "empirical" Lorenz curve are given by $\tilde{L}_k = (\sum_{\tilde{t}_k \leq t} m_k \tilde{t}_k)/(\sum_{k=1}^{n} m_k \tilde{t}_k)$ and $\tilde{G}_k = N^{-1} \sum_{\tilde{t}_k \leq t} m_k$. In this formulation the metric for plotting the curve is the patient, though only one point is plotted for each unit, i.e. physician. That is, the kth physician will be plotted a distance m_k/N to the right of the $(k-1)^{\text{th}}$ physician.

The definition of the Lorenz curve in (1) differs from the standard definition [14, 15, 16]. In the standard Lorenz curve analysis, there is a single quantity measured for each experimental unit, and

7

the metric for the horizontal axis is the proportion of experimental units. However in the context of our example, it is more relevant to quote concentration statistics in terms of patients rather than doctors.

The Gini coefficient is a commonly used numerical summary of the Lorenz curve and in applications frequently accompanies graphical presentation of the curve. A theoretical formula for the Gini coefficient is $(1/2) \int L(t) dG(t)$. We estimate it as

$$GC = \left| 1 - \sum_{k=1}^{n} (\tilde{G}_k - \tilde{G}_{k-1})(\tilde{L}_k + \tilde{L}_{k-1}) \right| \tag{2}$$

where $\tilde{G}_0 = \tilde{L}_0 = 0$. The coefficient is an estimate of the ratio of the area between the Lorenz curve and the $45°$ degree line to the area below the $45°$ degree line. The Gini coefficient ranges from zero for perfect equality in a distribution to one for maximum inequality.

In the hypothetical situation where there is perfect equality in the distribution so that the proportions of the features are evenly dispersed among the experimental units, $\tilde{t}_k = \tilde{t}_l \ \forall k, l$, and the Lorenz curve will fall on the $45°$ degree line connecting the origin (0,0) of the unit square to the top right corner (1,1). In this case the Gini coefficient will be zero. On the other end of the spectrum when all of the features are concentrated in a minimum number of experimental units (maximum inequality in the distribution), the Lorenz curve will lie along the horizontal axis before increasing linearly. The corresponding Gini coefficient in this case is one. In practice, the Lorenz curve lies somewhere between these two extremes.

The problem with which we are concerned arises because the quantities used to calculate the empirical Lorenz curve and Gini coefficient must be estimated from the data. Specifically, for each experimental unit we observe only an estimate, $\tilde{t}_k$, of the true unobserved measure t_k, where each of these estimates is based on a limited sample and/or limited time period of study, characterized by m_k. Thus, while the empirical Lorenz curve is estimated as described above using the $\{\tilde{t}_k\}$, the true curve involves a ranking of the data according to the $\{t_k\}$. The problem is that when

8

$\tilde{t}_k \neq t_k$ not only will the estimated relative frequencies $\{\tilde{t}_k\}$ be subject to random error, but also their ordering in the construction of $\{\tilde{L}_k\}$ will be redistributed in such a way as to maximize the apparent concentration. Members of the population with values of $\tilde{t}_k$ that are larger than their corresponding true t_k will be ranked higher than they should be while members of the population with values of $\tilde{t}_k$ that are smaller than their corresponding true t_k will be ranked lower than they should be. Consequently the empirically estimated Lorenz curve will be biased in the direction of increased apparent concentration, and the Gini coefficient will be overestimated.

Because of this systematic bias, we are interested in methods that will produce bias-corrected Lorenz curves. We next propose several methods that could be used for this purpose and describe them in detail. In these approaches we focus on adjusting the estimates of the proportions $\tilde{t}_k$ in recognition of the fact that the distribution of empirical proportions will always have greater variance that the distribution of true proportions from which they are generated.

4   ANALYTIC STRATEGIES FOR UNBIASED ESTIMATION

In this section we discuss a series of candidate analytic methods that have the potential to reduce or eliminate the bias. Three of these approaches involve random effects models while the fourth is a modification of the bootstrap approach discussed by Lee [11]. As mentioned in the Introduction, these are all nonparametric approaches in the sense that they do not impose constraints on how the experimental units are ranked.

4.1   Random Effects With Logistic Regression

The first method we consider is a random effects model. By using programs that are available in standard statistical software packages, we can estimate a random effect for each experimental unit. A random effects model of this nature induces "shrinkage" in the random effects estimators for the

9

empirical estimates. The effect is to reduce the spread of the proportions between physicians and consequently to reduce the Gini coefficient and raise the height of the Lorenz Curve toward the 45° line. The degree of shrinkage should be inversely related to the sample sizes of the individual experimental units. In this approach we consider the model:

$$E[Y_{jk}] = \text{logit}(t_k) = \theta + \gamma_k$$

for $j = 1, ..., m_k$ and $k = 1, ..., n$. Here the Y_{jk} are assumed to be independent Bernoulli variables conditional on the experimental unit, denoted by the subscript k. θ is a fixed effect common to all experimental units, and the $\{\gamma_k\}$ are the random effects. In order for the random effects to be identifiable, we need to postulate a model for their distribution. We follow convention and assume that γ_k, $k = 1, ..., n$, are independent observations from a Normal$(0, \sigma^2)$ distribution. t_k is then estimated as $\hat{t}_k = e^{\hat{\theta} + \hat{\gamma_k}}/(1 + e^{\hat{\theta} + \hat{\gamma_k}})$.

The parameters in this model can be estimated using previously developed methods for nonlinear mixed models. A number of such methods have been proposed and are available in statistical software packages. (Pinheiro and Bates [17] and Davidian and Giltinan [18] contain reviews of these methods.) In this article we use a numerical quadrature approach to integrating the likelihood of the data over the random effects. This approach is easily implemented for reasonably sized datasets in the software SAS through the PROC NLMIXED program.

4.2 Normal Random Effects Analysis

Although the preceding approach is conceptually appealing, it becomes computationally burdensome in large datasets. Historically, it has often been shown that analyses of binary data in large samples can be accomplished by simply using the more available and computationally less burdensome normal mixed models, in effect treating each binary outcome as if it were a continuous normal variate. We include this model for its accessibility and computational speed, and compare

10

its results with the preceding logistic regression model in our simulations. In this model

$$E[Y_{jk}] = \theta + \gamma_k$$

where γ_k represents the independent normal random effect for the kth physician. Since the variance of the Y_{jk} really reflects the variance of the binomial variate X_k, we must perform a weighted regression, with weights $1/m_k$ to more appropriately characterize the contributions of each patient. t_k is then estimated using $\hat{t}_k = \hat{\theta} + \hat{\gamma}_k$.

Multiple standard statistical software packages provide programs that can be used to estimate the parameters in this model. These packages use either maximum likelihood or restricted maximum likelihood to estimate the variance parameters and solve the mixed model equations to obtain estimates of the fixed and random effects.

4.3   Bayesian Random Effects Analysis

In this approach we assume that X_k follows a binomial distribution and consider the hierarchical model

$$
\begin{aligned}
X_k | t_k &\sim \text{Binomial}(m_k, t_k) \\
\text{logit}(t_k) &= r_k \\
r_k &\sim \text{Normal}(\mu, \sigma)
\end{aligned}
$$

where we place priors on the parameters of the normal distribution:

$$
\begin{aligned}
\mu &\sim \text{Normal}(\mu_b, \sigma_b^2) \\
\sigma &\sim \text{Gamma}(\delta_1, \delta_2).
\end{aligned}
$$

We have used the Gibbs sampler as implemented by WinBUGS to obtain posterior estimates of the $\{t_k\}$.

11

In our simulations we utilized relatively diffuse priors for μ and σ. In conducting our simulation studies we found that when using very diffuse hyperpriors for the Gamma distribution at times the Gibbs sampler had problems updating the model, particularly when there was a very high degree of inequality in the data. We found, however, that the choice of $\delta_1 = 0.01$ and $\delta_2 = 0.01$ worked in all scenarios. Further, in several scenarios we studied the difference between the simulation results presented below and other results (not shown) using a hyperprior of Gamma(0.0001, 0.0001) and found no substantial differences. We used $\mu_b = 0$ and $\sigma_b^2 = 10,000$ as our prior for μ.

4.4   Bootstrap

Lee [11] suggested using a bootstrap approach to estimate the Lorenz curve and Gini coefficient. The configuration examined by Lee was somewhat simpler than the one in which we are interested, in that there was only one observation per experimental unit (not multiple observations that need to be aggregated within unit), and the observation was assumed to be continuous or categorical. Lee applied the Lorenz curve and Gini coefficient to study how well a feature differentiates between two populations (for instance, using a biomarker to distinguish between diseased and non-diseased individuals). He proposed drawing bootstrap samples of experimental units within each population, then within each bootstrap sample reordering the data based on the bootstrap sample, using the original data to calculate the Gini coefficient in each bootstrapped sample, and then averaging the estimated Gini coefficients across the bootstrap samples to obtain a bias-corrected Gini coefficient.

To use a bootstrap approach to accommodate a two-stage design with multiple observations per experimental unit where the goal is to draw a Lorenz curve showing the concentration of a feature within a single population, we must apply the bootstrap differently from Lee's approach. We ideally want a resampling scheme which reflects the two-stage nature of our data. This can be achieved by first drawing bootstrap samples of experimental units (physicians) and then bootstrap

12

samples of observations (patients) within the bootstrapped units. Within each bootstrap sample we calculate the outcome measures $\{\tilde{t}_k\}$ and use these measures to estimate the cumulants of the Lorenz curve.

Specifically, for this approach we draw a bootstrap sample of experimental units. Within each experimental unit we resample observations. Let $X_k^{(b)}$ be the number of observations that have the binary feature Y for the kth bootstrapped experimental unit and $m_k^{(b)}$ be the total number of observations for the kth bootstrapped experimental unit. With the bootstrap sample we compute the proportions $t_k^{(b)} = X_k^{(b)}/m_k^{(b)}$. We rank the bootstrapped experimental units according to the $t_k^{(b)}$ and then calculate $L_k^{(b)} = \sum_{i=1}^{k} X_i^{(b)} / \sum_{i=1}^{n} X_i^{(b)}$ and $G_k^{(b)} = N^{(b)-1} \sum_{i=1}^{k} m_i^{(b)}$ where $N^{(b)} = \sum_{i=1}^{n} m_i^{(b)}$. We use these quantities to estimate the bootstrapped Gini coefficient, $GC^{(b)}$ by plugging the $L_k^{(b)}$ and $G_k^{(b)}$ into equation (2). We repeat this process B times to obtain B values of $L_k^{(b)}$, $G_k^{(b)}$, and $GC^{(b)}$. Let $GC^B = \frac{1}{B} \sum_{b=1}^{B} GC^{(b)}$. An estimate of the bias in the empirical Gini coefficient is then $\widehat{\text{bias}}_{GC} = GC - GC^B$. In other words, the bootstrap bias corrected Gini coefficient is equal to $GC^{B*} = GC - \widehat{\text{bias}}_{GC}$.

Constructing the bias corrected Lorenz curve is a little more complicated. The values of $G_k^{(b)}$ will be different across the bootstrap samples. That is, the points will be spaced at different intervals along the x-axis for different bootstrap samples. Therefore, simply averaging estimated cumulants across bootstrap samples is not ideal. Instead, we specify a set of Z points, $g_1, ...g_Z$, for the x-axis and use these points to interpolate the Lorenz curve in each bootstrap sample. In other words, for each specified point, g_z, we find the two values of $G_k^{(b)}$ such that $G_{k_1}^{(b)} < g_z \leq G_{k_2}^{(b)}$. Using linear interpolation, the point on the y-axis corresponding to g_z is $L_z^{(b)} = L_{k_1}^{(b)} + (g_z - G_{k_1}^{(b)})(L_{k_2}^{(b)} - L_{k_1}^{(b)})/(G_{k_2}^{(b)} - G_{k_1}^{(b)})$. Separately for each value of g_z, we take the average across bootstrap samples, $L_z^B = \frac{1}{B} \sum_{b=1}^{B} L_z^{(b)}$. In order to construct a bias corrected curve, we must also interpolate the empirically estimated curve. In the same manner as just described we obtain interpolated empirical

13

values $\{\hat{L}_z^E\}$. The biases in the estimates of the cumulants are $\widehat{\text{bias}}_{L_z} = L_z^B - \hat{L}_z^E$. The bias corrected estimates of the cumulants are then $L_z^{B*} = \hat{L}_z^E - \widehat{\text{bias}}_{L_z} = 2\hat{L}_z^E - L_z^B$. A bias corrected curve is obtained by plotting $\{g_z, L_z^{B*}\}$.

This approach is not necessarily an admissable approach for obtaining a bias corrected curve. By definition a Lorenz curve must be convex [14, 16]. By subtracting the bias from the empirical estimate on a cumulant basis, the resulting curve is not constrained to be convex. In the simulations and application we consider below, the curve still looks reasonable. However, it is technically not a Lorenz curve. Note that this is an issue only in our estimation of the Lorenz curve, not the Gini coefficient. Future work might consider how to obtain a bootstrap bias corrected curve that is restricted to be a convex function.

There are two potential problems that we foresee with this approach. The first is that it is unlikely to work well when there are a large number of experimental units in which none of the observations possess the binary feature $Y = 1$. In this case the bootstrap sample will continually estimate the probability as zero even though the true probability of occurrence is non-zero. Conversely, for experimental units in which all of the observations have the binary feature $Y = 1$, the bootstrap sample will continually estimate the probability as one. The second problem is that in order for the bootstrap estimate to converge to the "true" underlying parameter, we require that $m_k \to \infty$, not just that $n \to \infty$. In other words, we expect the bootstrap to be sub-optimal in the applications we consider because of the finite cluster sizes.

5   SIMULATION STUDY

We conducted a simulation study to (1) assess and characterize the magnitude of bias present in the empirically estimated Lorenz curve and Gini coefficient and (2) to evaluate the ability of the methods defined in Section 4 to produce bias-corrected estimates of the Lorenz curve and Gini

14

coefficient.

5.1 Simulation of Data

We generated data to have specific degrees of concentration as quantified by the Gini coefficient while also controlling the overall proportion of the observations with the feature $Y = 1$. The details of how we generated data are as follows:

We specified the total sample size, N, and the number of experimental units, n, and studied equal sample sizes per unit, $m_1 = m_k \ \forall k$, and unequal sample sizes generated from a multinomial distribution with probabilities α_k ranging from $1/n, ..., 1$, but scaled by the sum of these probabilities, $\sum_k \alpha_k$, so they add to one, in order to study potential bias when there is wide variation in sample sizes. Independently of N and the cluster size, we generated latent variables Z_k from a Normal(λ, 1) distribution where λ was prespecified, and then obtained probabilities $t_k^* = \Phi(z)$ where Φ is the cumulative distribution function for the standard normal distribution. In this scheme, the value of λ controls the degree of mal-distribution/concentration. When λ is large, the Z_k are large, and transforming them onto the Φ scale yields probabilities that are spread out across the range from zero to one corresponding to distributions with little concentration. As the value of λ decreases, the values of Z_k also decrease. When transformed onto the Φ scale, small Z_k values yield probabilities that are very small and restricted to a narrow range corresponding to distributions with higher degrees of concentration. In order to fix the value of $\pi = P(Y = 1)$, we first obtained $\{t_k^*\}$ as above, and then scaled the values iteratively until we obtained $\sum(t_k m_k)/N = \pi$. The observed $\{X_k\}$ were then randomly generated from Binomial(m_k, t_k) distributions.

In the simulations shown here, we fixed $N = 5000$. We studied situations where $\pi = 0.10$ and 0.25. We varied λ so as to obtain distributions with very little inequality, $GC = 0.05$, with moderate inequality, $GC = 0.25$ and $GC = 0.50$, and with high inequality, $GC = 0.75$. We also varied the

15

parameter n using values $n = 50, 100, 500$ which resulted in average cluster sizes of 100, 50, and 10. For each scenario we simulated 1000 datasets and report the mean values of the estimated Gini coefficients across the 1000 simulations.

In running these simulations, we found that in order to obtain convergence of the PROC NLMIXED program for the logistic regression, we needed to constrain the estimate of the variance of the random effects to be greater than zero.

5.2 Results

In Figure 1 we show examples of empirically estimated Lorenz curves from data simulated with equal cluster sizes. For this figure, data were generated with fixed $\pi = 0.10$, $N = 5000$, and then the cluster size and Gini coefficient were varied. Each estimated curve represents data from a single simulated dataset. Figure 1(a) depicts data with little inequality. The curve drawn using the "true" probabilities (which would be unobserved in a real application), lies very close to the $45°$ degree line (black line). The curve estimated from data with large cluster sizes of 100 (red line) lies close to this curve, but is clearly distinct from it suggesting a slightly higher degree of inequality in the data than is really present. The curve estimated from data with cluster sizes of 50 (green line) lies just below it. The Lorenz curve estimated from data with comparatively small cluster sizes of 10 (blue line) lies much further away and is very different from the true, unobserved curve. It incorrectly indicates a substantially larger degree of concentration in the data, and clearly represents a substantial degree of bias.

Figure 1(b) shows data with a slightly greater degree of inequality ($GC = 0.25$). Here all the estimated curves lie closer to the solid line depicting the true curve, but the overall trend is similar to that seen in Figure 1(a). Figures 1(c) and (d) show Lorenz curves drawn from data with increasing amounts of inequality. The estimated curves now lie increasingly closer to the true

16

curves. Figure 1(d) has all four curves relatively close to one another suggesting that in this scenario an empirically estimated curve may estimate the degree of inequality in the data with minimal bias regardless of cluster size.

There are two important general trends to observe from Figure 1. The first is that there is an inverse relationship between cluster size and the degree of bias in the estimated curve. As the cluster size decreases, the bias increases. The second point is that the degree of bias appears to decrease as the amount of concentration in the data increases. While there is substantial bias present in Figure 1(a) where $GC = 0.05$ (little concentration), the bias is small by Figure 1(d) where $GC = 0.75$ (high concentration).

We next explored the estimation of Lorenz curves using the analytical strategies presented in Section 4. In Figure 2, data were simulated with unequal cluster sizes with π, N, and n fixed at 0.10, 5000, and 500 respectively, to give an average cluster size of 10 for each curve. As before, the first panel, Figure 2(a), depicts data with little inequality. Here, the empirically estimated Lorenz curve (orange line) is the curve furthest away from the true, unobserved Lorenz curve. It incorrectly suggests a substantially larger degree of inequality than is actually present. The Bayesian and logistic random effects models produce curves (light blue and green lines, respectively) that lie on top of each other. These curves are the closest to the true curve and appear to do a reasonable job of correcting the bias in the empirically estimated curve. The curve produced by the normal random effects analysis (blue line) lies between the true curve and the empirically estimated curve. Out of the four analytic strategies, the bootstrap approach (magenta line) does the worst job of correcting the bias, yielding a curve that is relatively close to the empirically estimated curve.

Figure 2(b) shows data with a slightly greater degree of inequality. Here again the empirically estimated curve lies far away from the true curve. The Bayesian and logistic models now appear to over-correct for the bias, producing Lorenz curves that lie above the true curve. The normal

17

regression curve lies slightly below the true curve, and of the four analytic strategies gives the curve that is closest to the truth. In this situation, the bootstrapped curve only minimally corrects for the bias.

Figure 2(c) shows data with a moderate of inequality, and Figure 2(d) illustrates the situation when there is a high degree of inequality in the data. When the GC there appears to be relatively little bias.

Overall, these plots indicate that there is the potential for bias when estimating the Lorenz curve, that the bias decreases as the inequality in the data increases, and that a careful analysis of the data may help correct for the bias. It is difficult to draw general conclusions, however, based on the entire curve. To further characterize the bias and to simplify the table we utilize bias in the Gini coefficient as our measure of concentration.

Tables 1 through 4 contain these results. Table 1 and Table 2 summarize data that were simulated with $\pi = 0.10$. Table 3 and Table 4 summarize data that were simulated with $\pi = 0.25$. Table 1 contains results when the cluster sizes are equal, and Table 2 contains configurations where they vary. Looking first at the empirically estimated Gini coefficients in both tables, we see that the magnitude of bias depends at least partially on the cluster size. The bias can be determined by comparing the estimated GC with the known value in the first column. Thus, in the first row of Table 1, the logistic and normal GC estimates are 0.03 and 0.02, respectively, lower than the true value of 0.05. By contrast the empirical (0.18) and bootstrap (0.13) values clearly overestimate the true GC. Bias decreases as cluster size increases. For the case when GC=0.05, the difference in bias between the configurations where the average cluster size is 100 and where the average cluster size is 10 is considerable. Similar trends are evident in simulations with varying cluster sizes (Table 2).

The results in Table 3 and Table 4 are similar to the results in the first two tables. In these tables, we explore the effect of increasing the prevalence of the measure of interest to 25% ($\pi = 0.25$

18

in contrast to $\pi = 0.10$). The same pattern of increasing bias accompanying decreasing inequality and cluster size that was seen in Tables 1 and 2 is evident here as well. In contrast, however, in Tables 3 and 4 the bias in the empirically estimated curve is slightly less than we observed previously.

To understand why the bias decreases as the inequality increases, consider the following small, hypothetical example illustrated in Table 5. Suppose there are $n = 10$ experimental units and $N = 200$ equally distributed observations across the experimental units (ie. equal cluster sizes with 20 observations each) with $\pi = 0.1$. We consider first the situation where there is very little inequality in the distribution of the t_k's across the population. This scenario is represented in the first four columns of Table 5 where GC is approximately 0.05. Because the cluster sizes are equal and there is little concentration, all the t_ks have similar values in the region of about 10%. The order in which the units should be ranked based on those values is shown in the fourth column of the table. However, the estimates $\tilde{t}_k$ vary quite widely due to random variation, and the rankings based on these estimates (column 5) necessarily introduces a considerably increased correlation between the rankings and the $\tilde{t}_k$ values. This ranking is clearly very different from the ranking based on the t_k (column 4). The rank correlation coefficient comparing $\{t_k\}$ with $\{\tilde{t}_k\}$ is 0.30.

Consider now a similar data structure except with a moderate degree of concentration in the data. This scenario is represented in the right hand side of Table 5. Here the 10% of the observations with $Y = 1$ will tend to be concentrated in a limited number of experimental units. That is, roughly half the units have very small values of t_k while the remaining units have comparatively larger values of t_k. The observations for experimental units with small values of t_k will tend to have consistently small values of $\tilde{t}_k$. Conversely, the experimental units with higher values of t_k will generally tend to have higher rankings. In this case, the rank correlation coefficient comparing the two rankings is 0.92. Although this is a small and contrived example, it serves as a conceptual explanation of

19

why a low GC will lead to relatively high bias, and vice-versa. When there is little inequality in the population, there is great potential for variation in the estimates of the t_k and this necessarily results in a low rank correlation between the true rankings and the observed rankings. Conversely, when there is a large degree of inequality in the populations, there is less potential for disruptions of the appropriate ranking of the experimental units.

6 Health Care Data

Bach *et al.* [13] analyzed a 5% sample of black and white Medicare beneficiaries who were treated by 4,355 primary care physicians to assess whether visits made by black patients to physicians were concentrated in a relatively small proportion of physicians. Due to computational limitations, we illustrate the methods using only 500 of these physicians. There were 4892 total patient visits made to these 500 physicians. Three hundred and fifteen of the visits (6%) were made by black patients. The number of patient visits in the profile for a physician ranged from one to 58. The average number of patient visits to each physician was approximately 10 visits with a standard deviation of 8.8. Figure 3 shows the estimated Lorenz curves for these data. Of the 500 hundred physicians, 71% had no black patient visits in the 5% sample investigated. Because of this large number of zeros observed for the $\hat{X}_k$, as evidenced by the flat line in the empirically estimated curve along the x-axis, we suspect that the empirical estimate may be exaggerating the degree of concentration of the core of black patients in a small group of physicians. The empirically estimated curve suggests that 81% of black patient visits are concentrated in a relatively small group of physicians (94 out of the 500 in our sample), and these physicians account for only 20% of the total patient visits. The empirically estimated Gini coefficient corresponding to this Lorenz curve is 0.81.

All of the analytic strategies for reducing the potential bias that are considered here suggest that the degree of concentration of black patients within physicians is considerably less than suggested

by the empirical curve. The various methods suggest that between 64% to 78% of black patient visits are concentrated among physicians that have the highest concentrations of blacks and account for 20% of the total visits. These methods yield Gini coefficients ranging from 0.60 to 0.71. The Bayesian approach suggests the least amount of inequality. Based on our simulations it seems likely that the true Gini coefficient is in the middle of this range.

The interpretation of the Lorenz curve in the context of these nested data is not particularly intuitive. Ultimately, we would like to be able to make statements such as "$Y\%$ of black patients are seen by $X\%$ of physicians." This type of statement is simpler to understand and carries more meaning (and impact) than summarizing the Lorenz curve with the Gini coefficient. This information, however, cannot be directly obtained from Figure 3 where it is more difficult to articulate precisely what the individual points along the curve mean.

As an alternative, we suggest plotting the cumulative proportion of black patient visits (keeping the vertical axis the same) by the cumulative proportion of physicians. That is, we plot $\tilde{L}_k$ by Q_k where $Q_k = k/n$. In the situation where all physicians have the same patient volume (i.e. the m_k are equal across experimental units) this curve is equivalent to the Lorenz curve defined in equations (1). In contrast, when patient volumes differ across physicians the two curves are not equivalent. The same methods that we explore for estimating the Lorenz curve can be applied to this curve as well. The $\tilde{t}_k$ are estimated in the same way, used to rank the physicians, and then used to compute the $\tilde{L}_k$ just as with the curve in (1). The only difference is the quantity plotted along the horizontal axis. The bias we observed in the empirically estimated Lorenz curve together with the performance of the various methods we studied to correct this bias pertain to this curve as well.

It should be pointed out that the curve that we propose is technically not a Lorenz curve. Recall that a Lorenz curve must be convex. Our proposed plot can result in a curve that is not convex.

21

Consider the curves in Figure 4. The empirically estimated curve suggests that 84% of all black patient visits are concentrated within 20% of physicians. In contrast, the curves constructed using the other estimation methods we studied suggest that between 67% to 76% of all black patient visits are concentrated within 20% of physicians. The lack of convexity is especially noticeable for the bootstrap curve.

7 Discussion

In this manuscript we have studied the problem of estimating the Lorenz curve and Gini coefficient from a study in which the measure of interest is estimated with error for each experimental unit. We have demonstrated that ignoring variation in the outcome may lead to substantial bias in the empirically estimated Lorenz curve and Gini coefficient causing overestimation of the degree of concentration. This bias is most profound when there is relatively little concentration of the measure of interest in the population, and a relatively large error.

We have considered four analytic strategies to correct the bias. We are unable to conclude that a single method works best in all situations. For data with a relatively high degree of concentration (i.e. $GC \geq 0.50$), all of the methods possess reasonably good properties. The bootstrap does not work well, though, in data with little inequality. In this case, logistic regression with random effect provides better estimates, as does the closely related Bayesian hierarchical model, However, in data with a moderate degree of concentration ($GC = 0.25$ and $GC = 0.50$) and small cluster sizes, all of the methods are prone to noticeable bias.

In studying the Bayesian approach, we used diffuse hyperpriors for the distribution of the random effects centered at zero on the logit scale. However, in an application where there is prior information on the distribution of random effects, this information could be incorporated into the choice of parameters which may lead to an analytic approach for bias correction with better

22

properties than were seen in our simulation studies. Further, in the literature pertaining to the Bayesian random effects model, it has been pointed out that ranking experimental units based on the posterior estimates of the t_k may not work well (for example, Laird and Louis [19] and Louis and Shen [20]). Louis and Shen suggest using a loss function that is appropriate specifically for the ranks themselves and then using the posterior means of the ranks to rank the units. In this spirit we might order the experimental units based on the average of these posterior ranks and then use the estimated proportions, $\{\hat{t}_k\}$, to compute the Lorenz curve. This approach, however, could result in a curve that is at least partially concave. One avenue of future research might be to try to incorporate use of the posterior ranks while somehow restricting the curve to be convex.

In constructing the empirical Lorenz curve and Gini coefficient we have used maximum likelihood estimates (MLEs) of the probabilities $\{t_k\}$. As pointed out by Chew [21], in some applications point estimates other than the MLE may be preferred. Because the applications we consider have the potential for a relatively large number of zeros, the MLEs may not be the best point estimates to use. We explored constructing the empirical Lorenz curve and Gini coefficient using two other point estimates suggested by Chew. The first is based in game theory, assumes a squared error loss, and minimizes the expected loss. The second point estimate we studied was a Bayesian estimate which assumes that the prior probability density function for the probabilities is a Beta distribution with the first shape parameter equal to X_k and the second shape parameter equal to $m_k - X_k$. However, these methods do not perform as well as the bias correction methods we studied (data not shown).

Finally, in analyzing the 5% sample of Medicare benficiaries for this manuscript, we have ignored the fact that patients may have visited more than one physician. Patient level information that would have allowed us to track individual patients across potentially multiple visits to different physicians was not readily available to us. Future work might consider how to incorporate this

aspect of the data into the modelling process.

References

[1] Lorenz MO. Methods of measuring the concentration of wealth. *Publications of the American Statistical Association* 1905; **9**:209–219.

[2] Gini C. Sulla misura della concetrazione e della variability dei cartteri. *Atti del Reale Instituto Veneto di Scienze* 1914; **LXXII**:1203–1248.

[3] Hallas J, Støvring H. Templates for analysis of individual-level prescription data. *Basic & Clinical Pharmacology & Toxicology* 2006; **98**:260–265.

[4] Chang RKR, Halfon N. Geographic distribution of pediatricians in the United States: an analysis of the fifty states and Washington, DC. *Pediatrics* 1997; **100**:172–179.

[5] Brown MC. Using Gini-style indices to evaluate the spatial patterns of health practitioners: theoretical considerations and an application based on Alberta data. *Social Science & Medicine* 1994; **38**:1243–1256.

[6] Kobayashi Y, Takaki H. Geographic distribution of physicians in Japan. *Lancet* 1992; **340**:1391–1393.

[7] Prakasam CP, Murthy PK. Couple's literacy levels and acceptance of family planning methods: Lorenz curve analysis. *Journal of the Institute of Economic Research* 1992; **27**:1–11.

[8] Tickle M. The 80:20 phenomenon: help or hindrance to planning caries prevention programmes? *Community Dental Health* 2001; **19**:39–42.

[9] Elliot LJ, Blanchard JF, Beaudoin CM, Green CG, Nowicki DL, Matusko P, Moses S. Geographical variations in the epidemiology of bacterial transmitted infections in Manitoba, Canada. *Sexually Transmitted Infections* 2002; **71**:i139–i144.

[10] Nishiura H, Barua S, Lawpoolsri S, Kittitrakul C, Leman MM, Maha MS, Muangnoicharoen S. Health inequalities in Thailand: Geographic distribution of medical supplies in the provinces. *Southeast Asian Journal of Tropical Medicine and Public Health* 2004; **35**:735–740.

[11] Lee WC. Probabilistic analysis of global performances of diagnostic tests: interpreting the Lorenz curve-based summary measures. *Statistics in Medicine* 1999; **18**:455–471.

[12] Pham-Gia T, Turkkan N. Change in income distribution in the presence of reporting errors. *Mathematical and Computer Modelling* 1997; **25**:33–42.

[13] Bach PB, Pham HH, Schrag D, Tate RC, Hargaves JL. Primary care physicians who treat whites and blacks. *New England Journal of Medicine* 2004; **351**:575–584.

[14] Kendall MG, Stuart A. *The Advanced Theory of Statistics 1*. Charles Griffen and Company: London, 1963.

[15] Gastwirth JL. A general definition of the Lorenz curve. *Econometrica* 1971; **39**:1037–1039.

[16] Gastwirth JL. The estimation of the Lorenz curve and Gini index. *Review of Economics and Statistics* 1972; **54**:306–316.

[17] Pinheiro JC, Bates DM. *Mixed-Effects Models in S and S-Plus*. Springer: New York, 2000.

[18] Davidian M, Giltinan DM. *Nonlinear Models for Repeated Measurement Data*. Chapman and Hall: New York, 1995.

[19] Laird NM, Louis TA. Empirical Bayes ranking methods. *Journal of Educational Statistics* 1989; **14**:29–46.

[20] Louis TA, Shen W. Innovations in Bayes and Empirical Bayes methods: estimating parameters, populations and ranks. *Statistics in Medicine* 1999; **18**:2493–2505.

26

[21] Chew V. Point estimation of the parameter of the binomial distribution. *American Statistician* 1991; **25**:47–50.

Table 1: Equal cluster sizes: Average Gini coefficients across 1000 simulations when the m_k are equal. Data were generated with $N = 5000$ observations, $\pi = P(Y = 1) = 0.1$ and GC=0.05 for data with little inequality, $GC = 0.25$ and GC=0.50 for data with moderate inequality, and GC=0.90 for data with high inequality.

| GC	n	Cluster size	Estimated GC				
			Empirical	Logistic	Normal	Bayes	Bootstrap
0.05	50	100	0.18	0.03	0.02	0.03	0.12
	100	50	0.24	0.03	0.02	0.04	0.15
	500	10	0.50	0.02	0.02	0.03	0.35
0.25	50	100	0.31	0.22	0.21	0.20	0.26
	100	50	0.35	0.19	0.18	0.19	0.27
	500	10	0.55	0.11	0.11	0.10	0.41
0.50	50	100	0.52	0.48	0.47	0.49	0.50
	100	50	0.62	0.47	0.45	0.48	0.51
	500	10	0.64	0.45	0.43	0.36	0.58
0.75	50	100	0.75	0.74	0.74	0.74	0.75
	100	50	0.77	0.76	0.75	0.74	0.76
	500	10	0.81	0.73	0.71	0.71	0.78

28
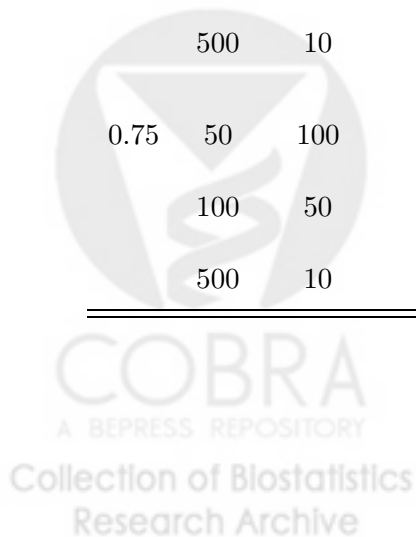
Table 2: Unequal cluster sizes: Average Gini coefficients across 1000 simulations when the m_k vary. Data were generated with $N = 5000$ observations, $\pi = P(Y = 1) = 0.1$ and GC=0.05 for data with little inequality, $GC = 0.25$ and GC=0.50 for data with moderate inequality, and GC=0.90 for data with high inequality. Also shown are average cluster size, range in cluster sizes, and average variation in cluster sizes across the simulations.

| GC	n	Cluster size			Estimated GC				
		Avg.	Range	Variation	Emp.	Logistic	Normal	Bayes	Boot
0.05	50	100	1-251	3298.4	0.17	0.03	0.09	0.04	0.11
	100	50	1-133	839.6	0.24	0.02	0.13	0.03	0.15
	500	10	1-38	35.9	0.50	0.03	0.26	0.03	0.38
0.25	50	100	1-237	3299.3	0.30	0.20	0.24	0.21	0.26
	100	50	1-133	839.6	0.34	0.19	0.24	0.21	0.27
	500	10	1-38	35.9	0.54	0.12	0.31	0.11	0.45
0.50	50	100	1-246	3304.2	0.51	0.48	0.47	0.45	0.50
	100	50	1-136	840.3	0.54	0.47	0.47	0.44	0.51
	500	10	1-37	35.9	0.65	0.40	0.45	0.36	0.57
0.75	50	100	1-237	3300.4	0.75	0.73	0.73	0.79	0.76
	100	50	1-133	841.2	0.77	0.75	0.74	0.74	0.76
	500	10	1-37	35.9	0.81	0.74	0.72	0.70	0.78

29
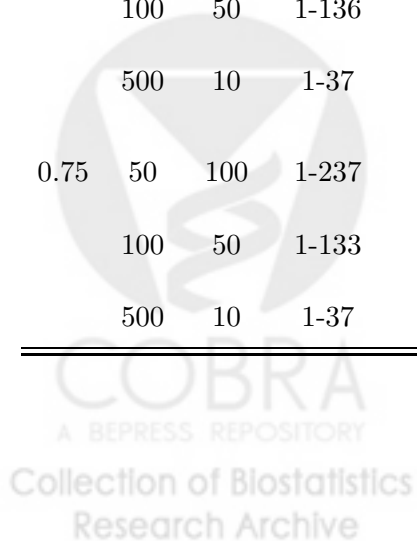
Table 3: Equal cluster sizes: Average Gini coefficients across 1000 simulations when the m_k are equal. Data were generated with $N = 5000$ observations, $\pi = P(Y = 1) = 0.25$ and GC=0.05 for data with little inequality, $GC = 0.25$ and GC=0.50 for data with moderate inequality, and GC=0.75 for data with high inequality.

| GC	n	Cluster size	Estimated GC				
			Empirical	Logistic	Normal	Bayes	Bootstrap
0.05	50	100	0.12	0.04	0.04	0.05	0.08
	100	50	0.15	0.03	0.03	0.03	0.10
	500	10	0.31	0.02	0.02	0.03	0.20
0.25	50	100	0.27	0.24	0.24	0.24	0.26
	100	50	0.29	0.23	0.23	0.23	0.26
	500	10	0.38	0.17	0.17	0.16	0.30
0.50	50	100	0.51	0.50	0.49	0.49	0.50
	100	50	0.52	0.50	0.49	0.49	0.50
	500	10	0.56	0.46	0.45	0.45	0.52
0.75	50	100	0.75	0.75	0.75	0.73	0.75
	100	50	0.76	0.75	0.75	0.74	0.75
	500	10	0.75	0.75	0.75	0.75	0.75
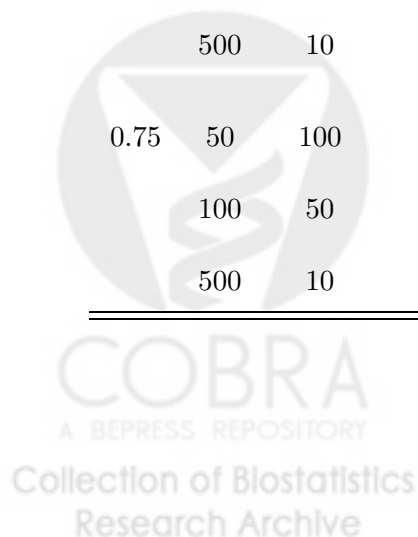
Table 4: Unequal cluster sizes: Average Gini coefficients across 1000 simulations when the m_k vary. Data were generated with $N = 5000$ observations, $\pi = P(Y = 1) = 0.25$ and GC=0.05 for data with little inequality, $GC = 0.25$ and GC=0.50 for data with moderate inequality, and GC=0.75 for data with high inequality. Also shown are average cluster size, range in cluster sizes, and average variation in cluster sizes across the simulations.

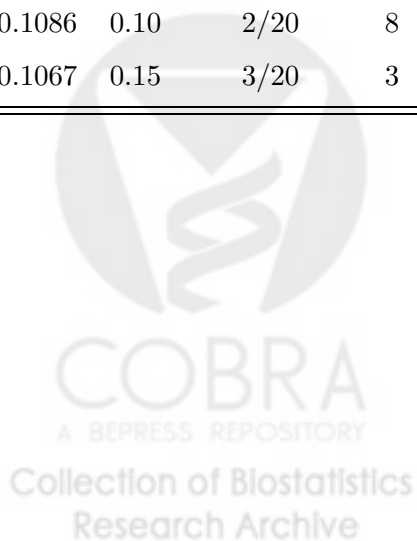| GC	n	Cluster size			Estimated GC				
		Avg.	Range	Variation	Emp.	Logistic	Normal	Bayes	Boot
0.05	50	100	1-251	3299.4	0.11	0.04	0.07	0.04	0.08
	100	50	1-133	842.2	0.15	0.03	0.09	0.03	0.10
	500	10	1-38	35.9	0.31	0.02	0.16	0.03	0.20
0.25	50	100	1-237	3299.1	0.27	0.24	0.24	0.25	0.26
	100	50	1-133	841.6	0.29	0.23	0.24	0.24	0.26
	500	10	1-38	35.9	0.39	0.18	0.25	0.17	0.31
0.50	50	100	1-246	3297.7	0.50	0.50	0.50	0.49	0.50
	100	50	1-136	840.2	0.51	0.49	0.49	0.48	0.50
	500	10	1-37	35.9	0.56	0.47	0.48	0.46	0.53
0.75	50	100	1-237	3302.1	0.75	0.76	0.75	0.73	0.76
	100	50	1-133	841.3	0.75	0.75	0.75	0.74	0.75
	500	10	1-37	35.9	0.75	0.75	0.75	0.75	0.75
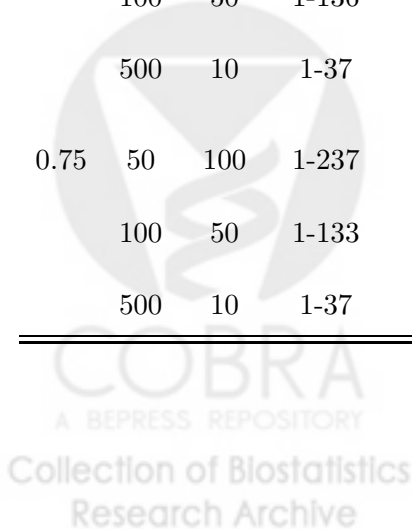
Table 5: Example of potential error in ranking experimental units. Data generated with $N = 200$, $n = 10$, and $\pi = 0.10$

| | | GC=0.05 | | | | | | GC=0.50 | | | |
| | | | Ranked by | | | | | | Ranked by | |
t_k	$\tilde{t}_k$	$\tilde{t}_k = X_k/m_k$	t_k	$\tilde{t}_k$		t_k	$\tilde{t}_k$	$\tilde{t}_k = X_k/m_k$	t_k	$\tilde{t}_k$
0.0466	0	0/20	1	2		0.2655	0.20	4/20	9	4
0.1084	0	0/20	6	3		0.1533	0.20	4/20	8	4
0.1069	0.15	3/20	4	1		0.3756	0.15	3/20	10	3
0.1084	0.20	4/20	6	2		0.0545	0.05	1/20	6	2
0.1079	0.05	1/20	5	4		0.0186	0	0/20	3	1
0.1086	0.15	3/20	8	5		0.0283	0.05	1/20	4	2
0.0895	0.15	3/20	2	1		0.0013	0	0/20	1	1
0.1085	0.30	6/20	7	3		0.0137	0	0/20	2	1
0.1086	0.10	2/20	8	1		0.0307	0.05	1/20	5	2
0.1067	0.15	3/20	3	1		0.0586	0.05	1/20	7	2
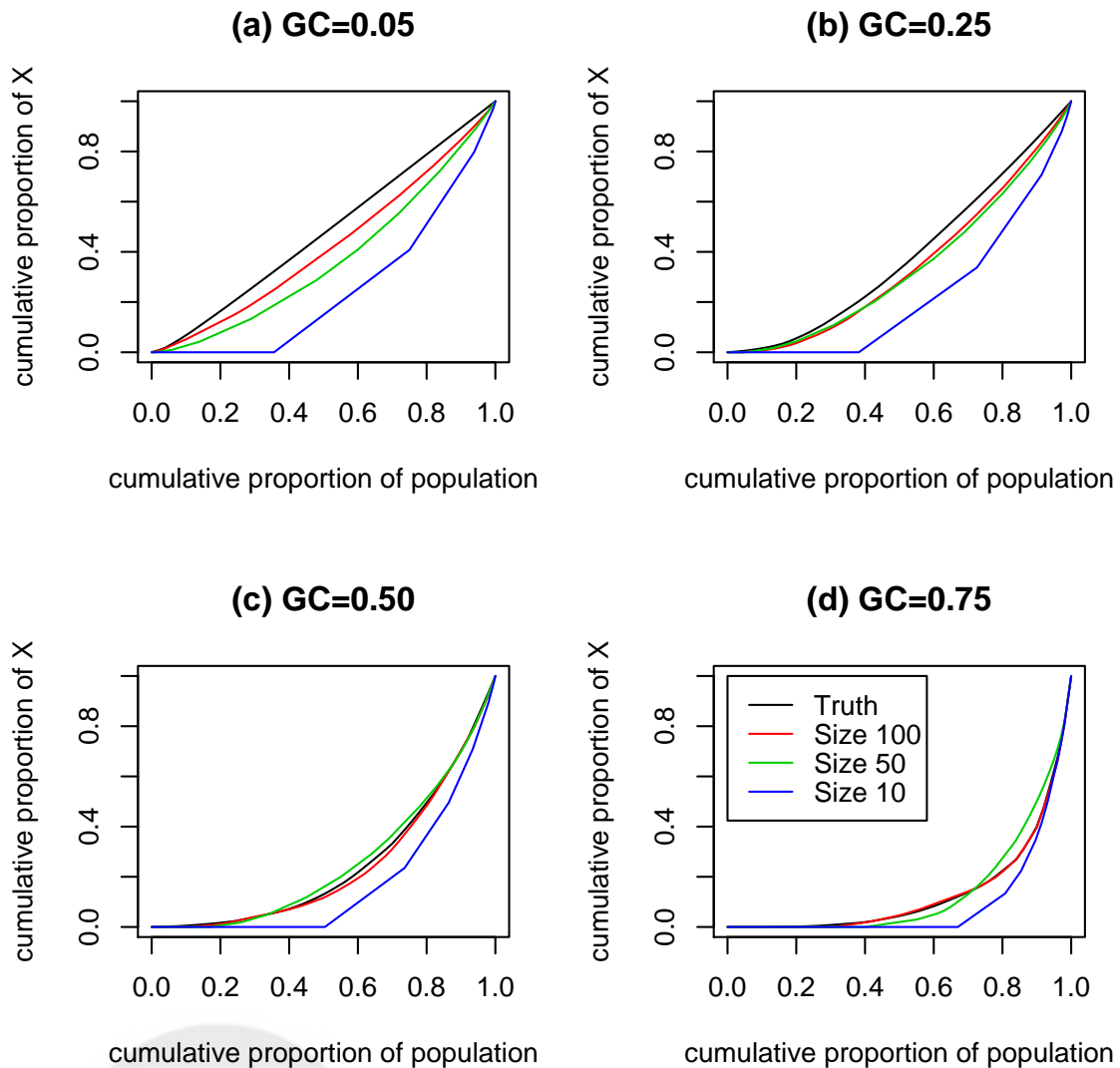
Figure 1: Bias in the empirically estimated Lorenz curve for data simulated to have equal cluster sizes of size 10, 50, and 100. Data were generated with $N = 5000$ observations and $\pi = P(Y = 1) = 0.1$.
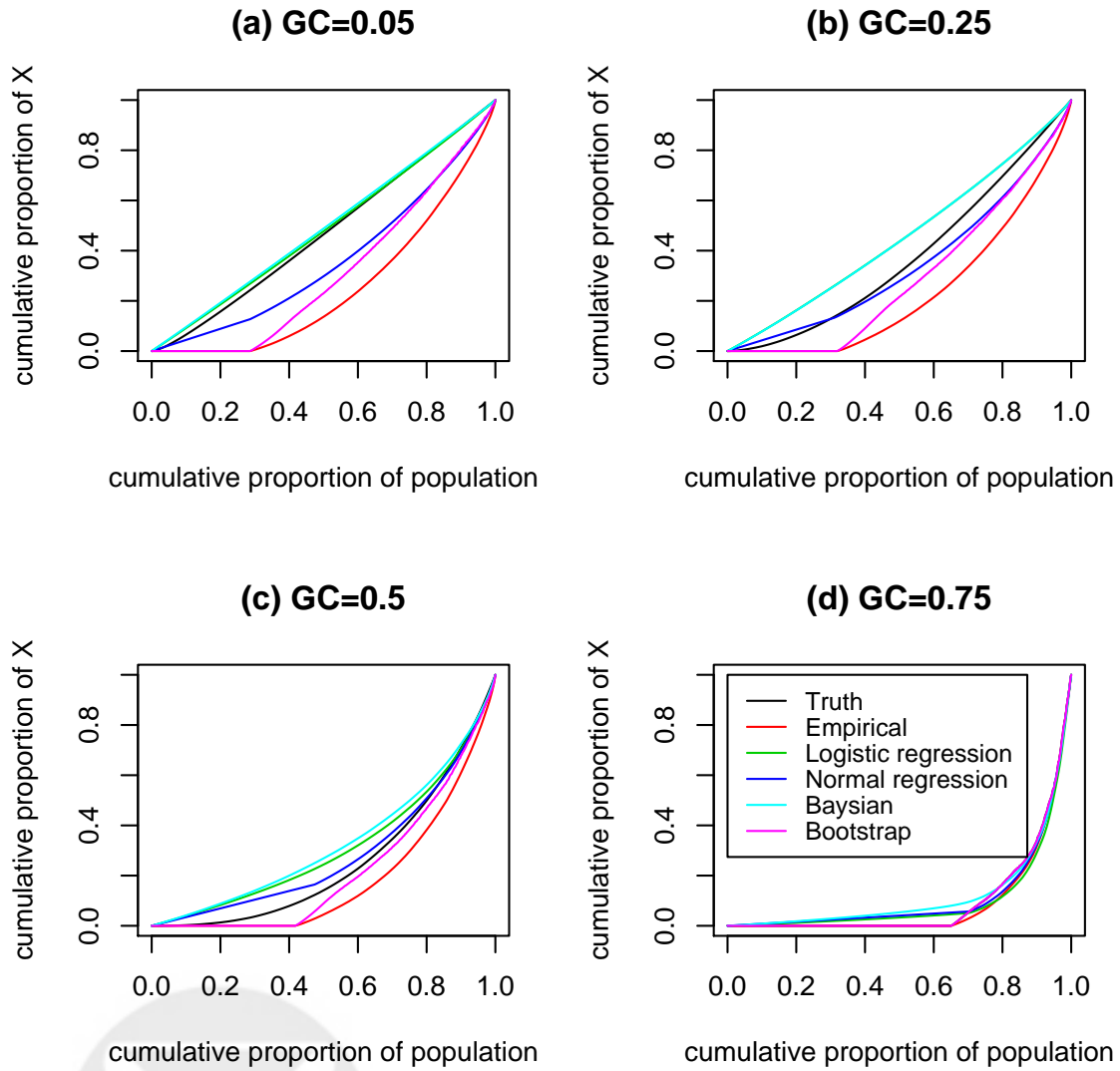
33
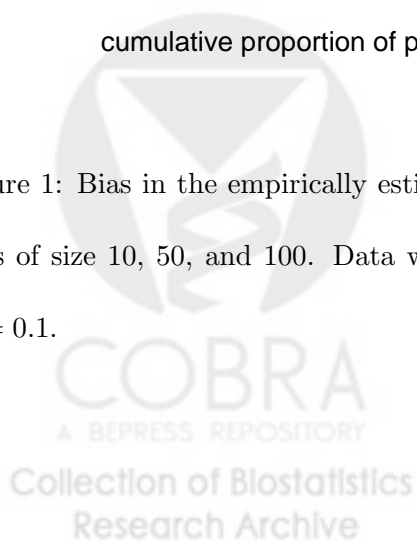
Figure 2: Analytic strategies for estimating the Lorenz curve from data simulated to have unequal cluster sizes of average size 10. Data were generated with $N = 5000$ and $\pi = 0.1$.
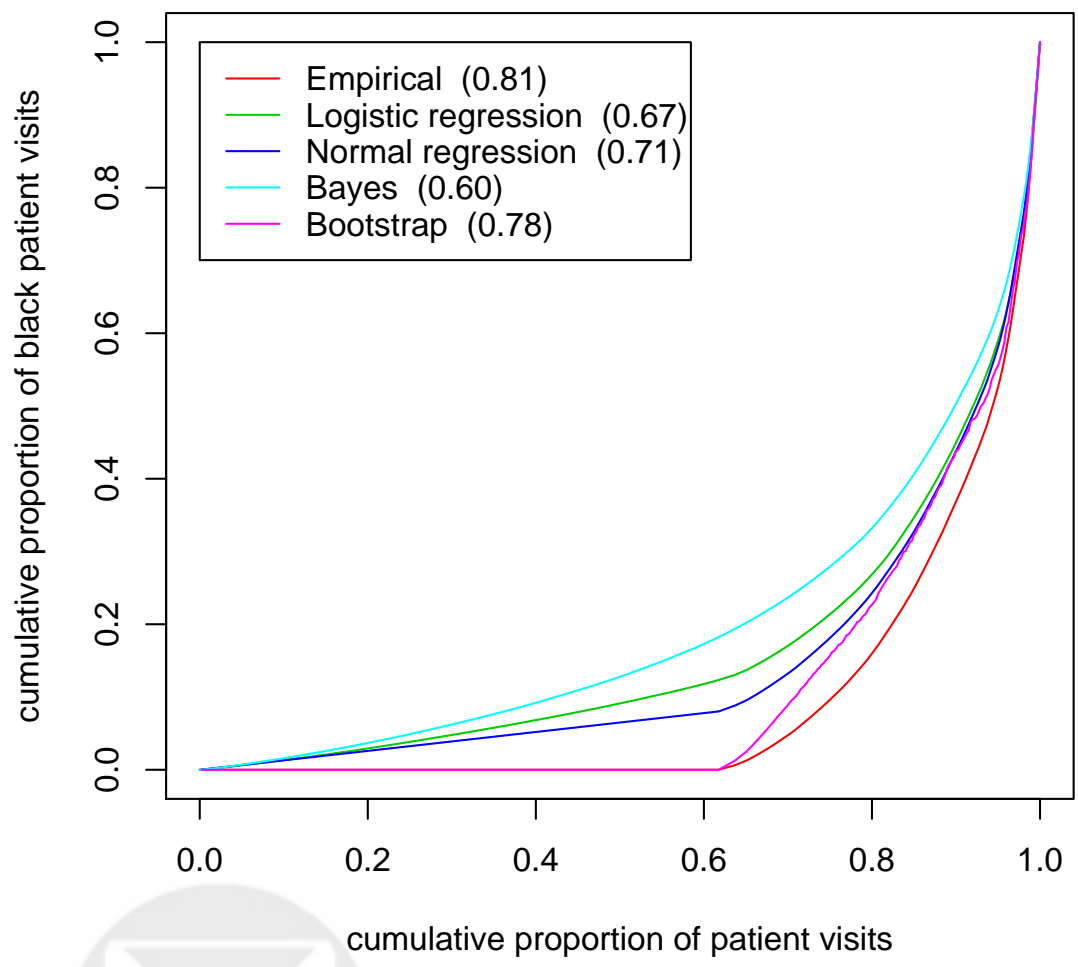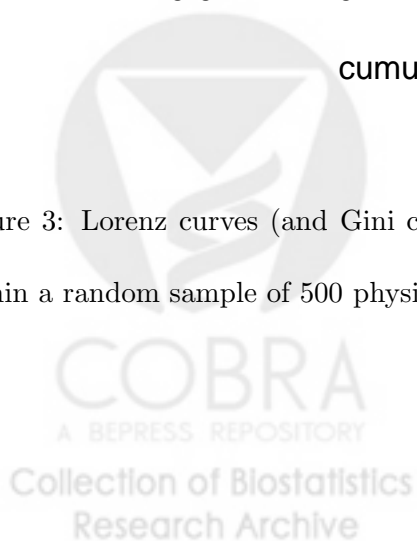
Figure 3: Lorenz curves (and Gini coefficients) showing the concentration of black patient visits within a random sample of 500 physicians.
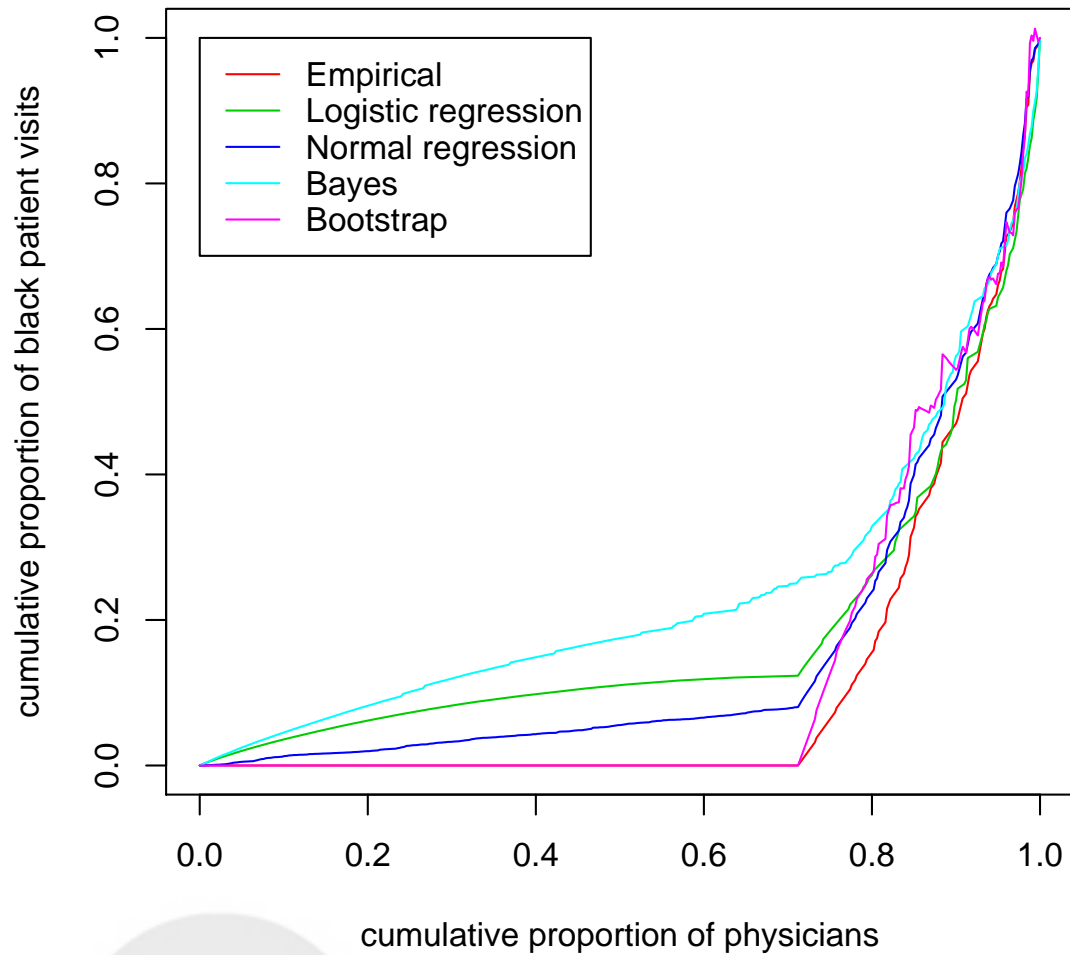
Figure 4: Alternate curve showing the concentration of black patient visits within physicians. In contrast to the Lorenz curve, here the cumulative proportion of physicians is plotted along the x-axis.