May 2020

Inequality in the United States 1946-2015

John Albert Schwendel
University of Wisconsin-Milwaukee

INEQUALITY IN THE UNITED STATES 1946-2015

by

John Albert Schwendel Jr.

A Dissertation Submitted in

Partial Fulfillment of the

Requirements for the Degree of

Doctor of Philosophy

in Economics

at

The University of Wisconsin-Milwaukee

May 2020

ABSTRACT

INEQUALITY IN THE UNITED STATES 1946-2015

by

John Albert Schwendel Jr.

The University of Wisconsin-Milwaukee, 2020
Under the Supervision of Professor Hamid Mohtadi

This dissertation consists of two essays which are concerned with the measurement and description of Income Inequality. Chapter 1 studies the measurement of inequality when data is presented in binned form. While various methods have been proposed for this purpose, I reveal an issue which has not yet been addressed in the literature. I demonstrate that differences in mean earnings can cause differences in the measures of inequality, even if underlying income shares remain constant. The consequence of this issue is that the usage of many of the available methods will create estimates of inequality not suitable for descriptive purposes and especially problematic for regression analysis. In this chapter I describe the nature of binned data, the methods available for estimation of inequality from this data, and a scaling issue detrimental to many estimation techniques. I then perform a series of simulations to demonstrate when certain methodologies perform better than others. I also establish if there are qualities of a specific dataset that would make a given method more appropriate in certain situations.

In chapter 2, using long run and geographically precise income data from the Internal Revenue Service, I estimate annual, state-level distributions of household income of the United States population from 1946 to 2015. Additionally and critically, by studying the central range of the U.S. distribution of income over this time period, I focus on an area of investigation that has received little or no attention in the literature which either focuses on a shorter time period or on top earners. I also use the findings from chapter 1 to select the most appropriate estimation technique for the data at hand. Various measures of inequality derived from the estimated distributional parameters are presented. National inequality represented by the Theil index shows the rise of inequality is concentrated in periods beginning in the late 1940s and the early 1960s. The movement of relative earnings among the 10th, 50th, and 90th percentile of earners supports the well-known hollowing out effect. Owing to the rich historical aspect of the data collected, I extend the understanding of the presence of this effect to the mid-20th century, i.e., the late 1940s and 1950s. State-level inequality follows a similar pattern to the national trend, but with some variation. While within-state inequality is the most substantial part of national inequality, an inverted W shaped pattern of between-state inequality is also found, with peaks in the early 1950s and early 2010s. The findings of this chapter both describe the history of United States inequality more fully and lay the foundation for better analysis of its causes and outcomes.

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGEMENTS

I would like to thank my advisor Hamid Mohtadi, who very early in my graduate studies expressed great confidence in my abilities. Throughout these studies, he pushed me and spent countless hours making sure I lived up to these abilities. I enjoyed our discussions on a variety of topics, and especially appreciate the economic expertise you so willingly shared to help me advance in my studies. Beyond your academic assistance I would like to thank you for always being a kind and understanding person and a good friend. I would also like to thank the other members of my committee. Scott Drewianka both encouraged me to pursue a Ph.D. and assisted me when I applied for the program. John Heywood always checked in both on my progress and to lend a friendly word. James Peoples happily took on the task of joining my committee and immediately began offering helpful insights. Thank you all for your advice in all areas, not just regarding my research. I deeply appreciate your willingness to listen to my endless questions and patiently offer intelligent, heartfelt, and frank advice.

To my friends, who saw me settling into a job that would not have fulfilled me and shook me out of complacency, thank you for making sure I pursued something more. To my friends within the program, thank you for making difficult tasks easier and long days shorter.

In addition I want to thank my family, who both encouraged me in this pursuit and who were always there when times were difficult. I would especially like to thank my mother who, as an educator herself, showed me how important and enjoyable the profession was. Finally I want to thank my father, who never failed to encourage me in all my endeavors, and who I wish could see me completing this one.

1 Mean Income, Scaling Issues, and Bias in Inequality Estimation from Binned Data

1.1 Introduction

Binned data is common in income surveys and in the tax data from a variety of nations. This type of data presents challenges to the researcher with regards to its usage for descriptive and analytical purposes. Various methods have been proposed for the purposes of estimating measures of inequality from this type of data. While these methods have been extensively analyzed and used in the literature, this paper will reveal a failing of many such methods which has yet to be addressed. While this problem, driven by the relationship between mean earnings across bins and the structure of the bins has been acknowledged in some cases, its effect has never been formally estimated. I show that this shortcoming produces systematic biases in various measures of inequality. The key insight regarding these biases is that income bins across regions or time will exhibit differential crowding based on the level of mean income for that level of observation. Many interpolation methods will be sensitive to such crowding, implying that inequality measures derived from these methods will be influenced not only by the underlying distribution, but also by the level of crowding in bins. This issue is especially important when comparing inequality over time or across regions. My formal demonstration of these issues allows me to compare how several candidate methodologies perform when this type of crowding is present.

This chapter will be organized as follows, section one will briefly describe binned income data and define its elements for usage in the rest of the chapter. Section two will outline a variety of methods used to estimate income inequality from binned data. It will also present examples of usage of these methods within the literature. Section three will present the results of a series of simulations that demonstrate the sensitivity of these methods to differential levels of mean income and will present results which formalize the systematic bias which is created. Section four will establish qualities of binned data that may limit or remove these biases and section five will conclude with some general guidelines for a researcher faced with data presented in this form.

1.2 Explanation of Binned Data

A reasonable place to begin is with a description of the type of data this paper will address. Binned data, sometimes referred to as bucketed or grouped, is distinct from individual level data. Whereas many surveys present observations on an individual and a weight assigning a representative value to those observations, binned data presents ranges of variables and the population within that range. When it comes to binned income data, this implies that information is given on how many individuals receive income between critical values. Often, the total income or mean income within each bin may also be provided. It should also be noted that customarily the last bin is only defined as containing all incomes above some value, with no upper bound.

Binned income data is common in tax data and small area data for the purposes of preserving anonymity. Even if individual level data is anonymized, if the population of the survey is specific enough, it is fairly straightforward to determine who an individual within a survey is and find specific information on that person. In the case of binned data, as no individual level observations are provided, it is only possible to find information on a specific person if a bin contains an extremely small number of individuals. As such, it makes it possible to provide data on small geographic areas and top earners without running afoul of privacy concerns.

For the purposes of this paper, I will define the elements of a binned dataset as follows:

$$[\alpha_\Theta, \alpha_{\Theta+1}) : The\, interval\, of\, a\, given\, bin.$$

$$\mu_\Theta : The\, mean\, value\, within\, a\, given\, bin.$$

$$n_\Theta = Number\, of\, individuals\, in\, a\, give\, bin$$

$$n = Number\, of\, individuals\, in\, a\, population$$

$$y_{\Theta} = Total\, income\, in\, a\, give\, bin$$

$$y = Total\, population\, income$$

1.3 Explanation of Candidate Methodologies

The first, and perhaps most obvious strategy for estimating inequality from this type of data arises from the fact that indices such as the Theil and the Gini can be calculated using only segments of the population and the corresponding percentage of total income which is held by that part of the population. This method requires information on the population within the bin and the average or total income of that bin. For example the Theil index, defined as:

$$T_t = \sum_{i=1}^{n} y_i ln \left(\frac{y_i}{x_i} \right)$$

$$y_i = percent\, of\, income\, held\, by\, individual\, or\, group\, i,$$

$$x_i = percent\, of\, population\, of\, individual\, or\, group\, i$$

can be calculated with n_{Θ}/n as the measure of x_i and y_{Θ}/y as the term y_i. This strategy will be referred to as the "Direct Method." As many income surveys have the top bin unbounded, this can require some assumptions on the nature of income within that bin. While data of this form does allow for direct calculation of inequality indices, constructing measures such as real earnings at the 10th, 50th, and 90th, percentiles are impossible unless some further interpolation is performed.[1] The motivation for the "Direct Method" is clear, most income surveys contain all the data necessary for implementation and as such, inequality indices can be directly calculated without any alteration of the data provided by the survey. While

[1]Unless population demographics allowed bins to somehow coincidentally fall exactly along these specific percentages, a very unlikely statistical fluke.

the above methodology is not common in the literature, its simplistic nature and immediate attractiveness offer a point of comparison to the methods which follow.

Cowell (1995) recommends a method which is more common in the literature and is based on the upper and lower bounds of inequality for a given presentation of binned data. These bounds are designed to be consistent with the binned data which are presented i.e. they maintain the reported bin populations, means, and boundary points. The lower bound is calculated by assuming every individual in a bin has income equal to the mean of that bin, μ_Θ. The upper bound is calculated by assuming $\frac{\alpha_{\Theta+1}-\mu_\Theta}{\alpha_{\Theta+1}-\alpha_\Theta}$ percent of a bin's population has income equal to the bottom value of that bin, α_Θ, and $\frac{\mu_\Theta-\alpha_\Theta}{\alpha_{\Theta+1}-\alpha_\Theta}$ percent of the bin's population has income equal to the top value of the bin, $\alpha_{\Theta+1}$. Cowell uses these bounds to derive what he refers to as a "2/3 1/3" rule. Specifically, he suggests a compromise measure of inequality indices which is simply a weighted average of these extremes. In the case of the Gini, the lower bound is weighted by one third and the upper bound by two thirds. For all other inequality indices, these are reversed. This method, like the prior method, is used only for the calculation of inequality indices. Other methods will be needed in order to provide measures like point estimates of income or income shares.

Each of the methods to be described below allow the researcher to create continuous distributions of a variety of types, and these distributions allow for effectively observing every individual in a population instead of observing groups defined by the survey itself. Two of these methods are extremely similar to the first strategy described above. Both assume uniform distributions within every bin. The first method, which is commonly used and generally simple, is referred to as the "Midpoint Approximation". This methodology is described by Heitjan (1989), and amounts to assuming that every individual within a bin has income equal to the midpoint of that bin. This is distinct from the direct method because the assumption of a uniform distribution within bins allows for the estimation of individual incomes and therefore, estimations of groups different than those present in the bin structure. This method creates a distribution which can be described with the probability mass function:

$$Pr(\frac{\alpha_\Theta + \alpha_{\Theta+1}}{2}) = \frac{n_\Theta}{n}$$

This same type of method can be performed using the mean of a bin as opposed to the midpoint, which would make it identical to the lower bound described by Cowell. It thus has the benefit of making both total income and mean bin incomes consistent with the underlying survey data. In this case the probability mass function is simply:

$$Pr(\mu_\Theta) = \frac{n_\Theta}{n}$$

This method will be referred to as the "Mean Approximation." It should be noted that the mean approximation, midpoint approximation, and direct method are all equivalent to linear interpolation of the Lorenz curve, due to the fact that they assume identical incomes and therefore, perfect equality within bins.

Cowell and Mehta (1982) offer several alternatives for estimations of inequality based on binned data. Among their proposed methods is the "Split-Histogram" method which assumes two uniform distributions within a bin that are split at an arbitrary value, b_Θ, within the bin. The share of individuals assigned to either side of the split is determined by the relative distance between either end point of the bin and the chosen point within the bin. For the general case, the density created by this method can be described as:

$$f(y) = \frac{n_\Theta}{n} \frac{\alpha_{\Theta+1} - 2\mu_\Theta + b_\Theta}{[\alpha_{\Theta+1} - \alpha_\Theta][b_\Theta - \alpha_\Theta]}, y\epsilon[\alpha_\Theta, b_\Theta)$$

or

$$f(y) = \frac{n_\Theta}{n} \frac{2\mu_\Theta - b_\Theta - \alpha_\Theta}{[\alpha_{\Theta+1} - \alpha_\Theta][\alpha_{\Theta+1} - b_\Theta]}, y\epsilon[b_\Theta, \alpha_{\Theta+1})$$

When this method uses the mean as the arbitrary point within a bin, which is the most common method in the literature, it is referred to as the "Mean Split-Histogram" method. In this case, the method simplifies to:

$$f(y) = \frac{n_\Theta}{n} \frac{\alpha_{\Theta+1} - \mu_\Theta}{[\alpha_{\Theta+1} - \alpha_\Theta][\mu_\Theta - \alpha_\Theta]}, y\epsilon[\alpha_\Theta, \mu_\Theta)$$

or

$$f(y) = \frac{n_\Theta}{n} \frac{\mu_\Theta - \alpha_\Theta}{[\alpha_{\Theta+1} - \alpha_\Theta][\alpha_{\Theta+1} - \mu_\Theta}, y\epsilon[\mu_\Theta, \alpha_{\Theta+1})$$

As is clear from these definitions, this method also requires one to make assumptions on open intervals in the highest income bin. The split-histogram method allows for construction of any measure of inequality, is straightforward to implement, and can be used with limited information presented in binned form, making it popular in the literature.

Another method proposed by Cowell and Mehta is polynomial interpolation of within bin distributions. In this case, the density can be described as:

$$f(y) = \sum_{k=0}^{K} \gamma_{\Theta k} y^k, y\epsilon[\alpha_\Theta, \alpha_{\Theta+1})$$

The authors discuss a variety of issues with this methodology, but they can be separated into two general categories, the order of the polynomial, K, being too small or too large. When the value chosen is too small, $K < 3$, the estimated distribution does not meet desirable qualities for an income distribution. If it is chosen to be too large, $K > 3$, the splines will have too many turning points and may display "implausible corkscrewing," which would not be consistent with empirically observed distributions.[2] Still, this method has not been abandoned in the literature. In fact, Blanchet Fournier, and Piketty (2017) describe a version of polynomial interpolation which they refer to as the "Generalized Pareto" method. Quintic splines are used to fit distributions of income within all but the top and bottom bins, which are fit using parametric assumptions.

The final method to be described and analyzed for this paper is a generalized least squares method. Binned income data effectively provides measures of empirical quantiles for

[2]The higher the order of the polynomial used, the more often its shape will exhibit changes in direction. Within-bin distributions do not change direction often, and thus this shape is unlikely to match with the underlying data.

the distribution of income within a population according to the bins provided.[3] For the least squares technique, using an assumed distribution, the parameters of that distribution are allowed to vary in order to minimize the sum of squared distances between the percentiles calculated from a theoretical distribution and the empirically observed population percentiles that fall at or below a given level of income. The parameters which minimize this measure are then selected to describe a continuous distribution. Once these parameters are estimated it paves the way for the estimation of any inequality measures of interest.

Many of these methods are found to be widely used for the estimation of inequality, with no consensus emerging on if any one is superior. What follows is several examples of the usage of these methodologies and the results derived from their implementation.

As mentioned the direct, midpoint, and mean methods are isomorphic with linear interpolation of the Lorenz curve, a method which was implemented by Panizza (2002). The resulting Lorenz curves were used to estimate a dataset of decennial, cross-state Gini indices. Panizza also used the split-histogram method to estimate quintiles of the distribution of income as an additional measure of inequality for his purposes. These indices were used to determine the effect of inequality on state-level per capita income growth from 1940-1980. Panizza finds a negative relationship between inequality and growth, though the magnitude of this effect is sensitive to the choice of the measure of inequality.

Frank (2009) also uses the split-histogram method in order to estimate the Theil index and two instances of the Atkinson index for the contiguous 48 states from 1945-2004. In addition, he implements Cowell's compromise Gini as an alternative measure of inequality. Similar to Panizza, these measures are used to estimate the effect of inequality on the growth rate of per capita income. Contrary to the results found by Panizza, Frank demonstrates a positive relationship between these measures, with the greatest effects observed for inequality among

[3]Presentations of binned income data provide information on upper and lower bounds and total numbers of filers within given bins. By dividing the cumulative sum of each bin population by the total population, it is straightforward to obtain the percent of individuals falling below a critical level of income.

the top portions of the income distribution. Both authors use the same United States tax data for their estimations, and both use some measure of the Gini index, so it is interesting that the signs of their effects differ. These differences may occur to do different panel regression methods, but may also arise from estimating inequality measures with different methodologies.

The split-histogram method is also used by Atkinson and Piketty (2010) to estimate the top income and population shares from binned income data for a variety of nations. This is done both for descriptive purposes and for subsequent explanation of the trends presented.

Polynomial interpolation of income distributions is also found within the inequality literature. Dikhanov and Ward (2001) use forth order polynomials to estimate distributions within bins of grouped data. The resulting distributions are used to calculate world incomes, with a specific focus on both within- and across-country inequality. Budd (1970) also uses this method with fourth order polynomials to estimate Lorenz curves for the purpose of describing post World War II income inequality.

The final method outlined above, the general least squares method, is found in Pinkovskiy and Sala-i-Martin (2009) who implement it in order to estimate distributions of incomes across 191 nations from 1970-2006. In their analysis they assume all distributions to be lognormal, though they do show their qualitative results are robust to a variety of distributional assumptions. They use the resulting distributions to calculate poverty rates as well as inequality indices. Chotikapanich and Griffiths (2008) use the same method in order to assess whether a mixture of Gamma densities can be used for accurate parametric fits of Canadian income distributions and resulting Gini indices.

1.4 Scale of Income, Bin Crowding, and Biases

While these methods have been analyzed for accuracy and in some cases, compared in terms of their ability to estimate a measure of inequality for a given population, the driving impetus of this section is to focus on a specific methodological issue not yet described in the literature. This issue arises from the fact that given a constant bin structure, as mean income changes across observations, whether temporal or geographic, bins will be crowded *differentially*. Even when bin structure is identical across regions or years, there will be variations in the share of income and population falling into specific bins due to several systematic factors. Specifically, crowding of populations inside upper and lower bins occurs respectively in wealthier or poorer geographic areas. In addition, as mean nominal incomes change over time due to inflation or deflation and growth or recession, similar crowding will occur.

While this issue of crowding has not been addressed, specific changes to bin structure can create the similar effects. The issue of bin structure has been discussed by Seiever (1979) who uses two alternate binning structures to analyze the effect of changing bin sizes and placement on the mean and midpoint approximations described above. He points out these methods are sensitive to such changes, and this sensitivity is problematic for cross-region comparison as well as the observation of time series trends. While Gatswirth (1975) does use several alternative binning structures, these are employed to analyze his proposed method against the midpoint approximation. Neither of these authors address the issue of crowding induced by the changing bin structure, nor do they discuss how even with a consistent bin structure, estimation bias and comparability issues will arise. This particular issue of bin structure will not be the focus here, where most analysis will be done holding the number and placement of bins constant, thus demonstrating that the underlying mechanism can be affected without changing bin structure. Even given this, the results to be described are consistent with these prior works.

Unfortunately, as will be discussed, many of the methods available for interpolation of distributions and construction of inequality measures are sensitive to the issue of crowding within bins, as crowding changes the underlying data used for inputs in these estimations. To my knowledge, this important issue has been overlooked in the literature. Below I present a discussion of intuitions on why several popular methods for dealing with binned data are sensitive to such crowding. For the issue at hand, the Theil and Gini indices are especially illustrative. A particular quality of these indices is their scale independence, as their underlying calculations are based on relative population and income shares. An estimate of either of these indices should then exhibit the same value regardless of the scale of income. This presents a new rubric by which to assess the methodologies described above.

Along with the intuitive explanations of the mechanisms which influence these estimations, a series of simulations will be performed. The goal of these simulations is to mirror the nature of the sampling and binning procedures that produce the data structure presented in binned income survey data.[4] To this end, an underlying distribution of incomes will be assumed.[5] From this distribution, random samples are drawn to represent the way in which a survey collects income data from the population. The income from these samples is then binned in order to conform to a given bin structure.[6] As an underlying distribution is assumed, the underlying inequality measures for that distribution are known prior to the sampling and binning. It is then possible to compare the true measures calculated from the assumed distributions to those that can be estimated by performing various interpolation strategies on the constructed binned data. This is the type of comparison that is common in the literature which looks at errors of specific methods in this somewhat static setting.

To show how mean income differences across areas or over time produce spurious measurements in the estimation, I perform a scaling exercise. Specifically, I rescale the sample to effectively create several simulated populations. This way, the true inequality indices are preserved and remain identical across the simulated populations as the relative

[4]The specifics of this exercise can be found in the Method Appendix to this chapter.

[5]Here this assumption will be based on estimated parameters for the 2015 distribution of adjusted gross income, found in the Internal Revenue Service's Statistics of Income, which closely matches a lognormal distribution with a location parameter of 3.5. To check for robustness, the scale parameter will be allowed to vary within a range of reasonable values, namely from 1 to 1.3 in .02 increments.

[6]The bins used for the presentation of 2015 Adjusted Gross Income Data from the Internal Revenue Service's Statistics of Income will be used. This structure is made up of coarsely defined bins and thus is ideal for illustrative purposes.

income shares income per individual are unchanged. The next steps are crucial: First, the scaled data are binned according to the same constant bin structure described above. Second, from this binned data, continuous distributions are interpolated using various prominent methods that have been proposed or used in the literature (see above). Finally, measures of inequality are calculated. This is meant to test whether crowding will create spurious results in measures of inequality.

Ideally, if population income shares do not change, but the mean income levels across simulated populations do, a proper estimation technique performed on binned data would produce the same inequality estimate for each simulated population. In this case, one would not have observed any pattern of misestimation caused by crowding induced in this way. Following this logic, if the estimated measures are observed to differ systematically from their true values, the underlying source of discrepancy must be the combined effects of the changed presentation of the binned income data and the method used to create a continuous distribution from that data. Given this characterization, I will establish the influence of bin crowding on various estimation methods. Specifically, for several of these methods, I will formally demonstrate the existence of a systematic bias in the measure of inequality that is correlated with the level of scaling and therefore the mean income of a given population. Due to this mechanism, differences in mean income will create differences in the estimated inequality, *even if the underlying level of inequality were to remain constant*. This implies that as intra-bin populations grow differently regions and time, the extent of crowding and the resulting bias of the indices will also vary, meaning that changes in the measure of inequality may be driven by systematic differences in mean income instead of solely being reflective of underlying changes in inequality.

The first method to undergo this diagnostic methodology is the direct method. While this method is not commonly used, it will be shown that its estimations do not differ greatly from the mean or midpoint approximations. In addition it establishes a clear intuition regarding the effects of bin crowding and as such, will serve as a reasonable comparison for all the methods which follow. Specifically, simulations show that as greater numbers of individuals are binned into a given income category, calculations will be performed as if the population within that category had identical incomes, understating inequality to an increasing degree in the case of the Theil, and creating a separate, but distinct pattern of bias in the Gini. The effects of this mechanism can be found in Figures 1.1 & 1.2 which summarize the results of the simulation outlined above and explained in detail in the Method Appendix. Figure 1.1 demonstrates the effects on estimations of the Theil index from such a treatment. As individuals from lower bins are pushed into higher income bins by my diagnostic methodology, each new individual in a bin represents a constant marginal increase in the population share of that bin, while their marginal contribution to the income share is decreasing. This makes the components y_i and x_i both grow in size, but because of the differential growth, these measures also become closer in value. The total effect is to make the term $ln(y_i/x_i)$ smaller, while also increasing the weight of this term, assigned by y_i. This decreases the weighted sum that defines the Theil, $\sum_{i=1}^{n} y_i ln\left(\frac{y_i}{x_i}\right)$.

When incomes are scaled downward, this effect is reversed. Interestingly though, when incomes are very small, it can be seen that the estimated Theil can be slightly higher than that calculated for unscaled income data. This is because underestimation occurs due to crowding within specific bins. At certain income scales, the bins are populated such that this crowding issue is even less prevalent than it is in the unscaled data.

While the trend in estimates for the Gini is similarly shaped, Figure 1.2 demonstrates that for the scales of income here, the Gini tends to be overestimated. This is due to the particular nature of the Gini, and its sensitivity to the number and placement of groups from which it is calculated. If the Gini for a given distribution of income is calculated across a fixed number of groups, the estimate of the Gini will be sensitive to the relative sizes of those groups. This is because linear interpolation of the Lorenz curve biases the measure of the Gini downward, as the linear representation will always fall inside of the actual Lorenz curve. Therefore if one is able to make this linear interpolation move closer to the actual

Lorenz curve (i.e. minimize the area between the linear interpolation and the curve), it will reduce the downward bias and increase the estimated Gini. This minimization occurs when the groups the Gini is calculated from represent identical population shares. For unscaled data, the binning structure used here tends to create a situation in which the lowest bin holds the greatest segment of the population. By scaling incomes up, more individuals are pushed into the bins that originally held smaller segments of the population. By shrinking the population share in the first bin, which originally held a plurality of individuals, and increasing the populations in the other bins, shares of the population become more similar across bins. As this occurs, a the linear representation of the Lorenz is moved closer to the actual Lorenz, thus increasing the estimated Gini.

The Cowell "2/3 1/3" rule also produces its own distinct pattern of bias as can be seen in Figures 1.3 & 1.4.[7] These trends are a weighted average of two separate measures, hence they exhibit fairly irregular shapes, but some broad conclusions can be made. First, while the trend of median Theil estimates exhibits a similar shape regardless of specified parameters, as can be seen in the panels of figure 1.3, its relationship to the true population Theil changes with the assumed scale parameter. Second, the downward bias of the mean approximation seems to overwhelm the upper bound for the estimation of the Gini, demonstrated by the consistent underestimation shown in figure 1.4. Finally, while neither index is estimated without bias, the estimates for the Gini adhere more closely to their true value. It may be due to this fact that the usage of this method in the literature is more common for estimation of the Gini than for other indices.

For the remaining methods, these analyses will be expanded to include several other common measures of inequality. As the prior methods are constrained to calculations involving observations of entire bins, the only measures of inequality that can be derived are those which can be calculated without further interpolation of binned data. For all

[7] As explained in the previous section, an assumption must be made on bounding for the top income bin in order to implement this method. As the income data here is simulated, I will simply use the actual highest sample income to bound this bin in my calculations. Unless otherwise noted, this will be used for all methods which require such a treatment

methods which follow, full distributions can be estimated, and thus incomes of every individual in a population can be derived. Of course this means that estimates of any arbitrary group can also be calculated. This affords me the opportunity to analyze estimations of income shares and compare them to their true population value.

In order to compare the efficacy of these methods in providing point estimates of real income, some adjustments will be needed. Given the particular diagnostic methodology here, the real incomes within the population from which samples are drawn do not change, but the samples themselves are scaled. In order to make these values comparable, I will take the real incomes estimated at the 10^{th}, 50^{th}, and 90^{th} percentile and "descale" them(i.e. I will divide these estimates by the value at which they were scaled).[8]

The first method for which these additional measures can be analyzed is the midpoint approximation Figures 1.5 & 1.6 summarize the effects of scaling and crowding on estimations of Theil and Gini indices calculated from this method. As can be seen, this estimation technique overestimates inequality. This result may be surprising, given that the mechanisms at play seem very similar to the previously described method. Specifically, they both assume large portions of the population have identical incomes. As noted by Gatswirth (1972), the tendency to overestimate is due to relative placements of midpoints and means within bins. Given the unimodality of most income distributions, the midpoint of a bin will usually be lower than the bin mean for incomes before the population mode, and higher than the bin mean thereafter. This will underestimate incomes for low income individuals, and overestimate them for high income individuals. While Gatswirth establishes that this will result in overestimation of inequality, I formally estimate this effect with relation to bin crowding as related to mean population incomes. In figures 1.5 & 1.6 as the bias reaches its peak, these two effects are at their greatest, with the population being split into bins before and after the population mean. After this point, progressively more incomes are pushed into bins that occur past the mode of the population, and thus

[8]This methodology is analogous to inflation adjusted measures in the literature. Instead of adjusting by the CPI, to normalize these incomes I adjust by the factor of scaling that is induced.

more bins have their incomes overestimated, making inequality look less extreme. This intuition is confirmed by figures 1.7 & 1.8 which show that the share of income held by the bottom 10% drops quickly initially and then levels out. Within the same range, the share held by the top 10% increases at first, when a large portion of the population has their income underestimated, then decreases due to overestimation of incomes for much of the population. Unsurprisingly, this method does not provide reasonable or consistent point estimates for real earnings as shown in figures 1.9 & 1.10 where the estimates of real income for the 10^{th} and 50^{th} percentile of earners oscillate wildly. This is due to the fact that the midpoints of bins remain constant and as populations are moved through these bins by scaling, the estimated income of an individual at a specific percentage of the population eventually jumps to the midpoint of the next bin. These effects are then tempered by the adjustment back to a normalized income, until the effect of scaling is severe enough to again shift the estimate to the midpoint of the subsequent bin. The 90^{th} percentile of earners have estimated incomes which increase and then flatten, as shown in figure 1.11. This is because the process of scaling incomes means that eventually the 90^{th} percentile of earners will always fall in the last bin and have their income estimated as the midpoint of that bin no matter how much higher incomes are scaled.

Given the bias created by the relationship between intra-bin means and midpoints, it may be attractive to implement the mean approximation described earlier. With this method though, the exclusion of any measure of within bin inequality means that inequality biases will always and everywhere be biased downward. As there is no longer an issue with the relative positions of midpoints and means, this implies that nothing will counteract this effect. Figures 1.12 & 1.13 illustrate these trends for the Theil and Gini respectively. At the lowest and highest mean incomes, the amount of individuals crowded into top and bottom bins will be at their most extreme. This will imply that the greatest portions of the population are assumed to exhibit perfect inequality at these extremes. Figures 1.14 & 1.15 reinforce this idea by showing a decrease in the share of income for bottom earners which levels out, and a more consistent decrease in the share for top earners. This method also does not improve on point estimates of real earnings, showing a similar oscillation to that previously described for the midpoint approximation, as in shown in Figures 1.16 & 1.17. The only striking difference can be found in Figure 1.18, which shows the estimate for the 90^{th} percentile of income decreases after this observation is pushed into the top bin. This is because unlike the midpoint approximation, which had a fixed value in the top bin, the adjusted mean of this top bin decreases as lower income individuals are crowded into that bin.

A method which does not assume this perfect within-bin equality is the mean split-histogram method. The version used here is proposed by Blanchet, Fournier, and Piketty and fits a Pareto distribution to the top, unbounded bin and uses other parametric methods in the bottom bin. All interior bins are still fit by the distributional function explained earlier. The results of income scaling on estimations of the Theil and Gini derived from this method can be found in Figures 1.19 & 1.20. As can be seen, this method provides the most accurate estimate of these indices so far. It should be noted that there is still a systematic downward bias as mean incomes increase. As found in Figure 1.22, this bias is due to a consistent decrease in the estimation of the share held by the top 10% of earners, which is recaptured in other portions of the earning distribution. What is interesting is this is likely driven by the split-histogram treatment in the inner bins because, as can be seen in Figure 1.21, the share of income estimated for the bottom 10% of earners is very accurate for most of the relevant range.

The trends in estimation of inequality indices by the generalized Pareto method are extraordinarily similar and can be observed in Figures 1.26 through 1.29. As this method makes the same assumptions on top and bottom bins as the split-histogram method implemented above, this reflects the marked impact of estimated values in these bins. This is unsurprising, given the fact that these are the bins that would be most sensitive to changes in mean income. The bottom bin is bounded at zero, so as mean incomes are decreased, crowding must occur there. Likewise as incomes increase, individuals crowd in the top, unbounded bin.

One aspect not captured in these figures is the nature of the two previous methods to overfit to sample data. The estimations derived for a given sample are very tightly clustered around the values presented in binned form for that sample. For the split-histogram method, this is due to the fact that, by construction, the distribution is fitted to the end points and the mean for each bin. In between these points, linear distributions are assumed. In the case of the generalized Pareto Methodology, these same points are fitted to the "corkscrew" polynomials described by Cowell and Mehta. In addition to overfitting, this creates problematic point estimates. If the points of interest fall along the endpoints or mean of a bin, these methods will fit to them very precisely. If they, on the other hand, fall between these values, these implausible shapes create the types of biases seen in Figures 1.23 through 1.25 for the split histogram method and 1.30 through 1.32 for the generalized Pareto method. What is interesting is that for the 90^{th} and 10^{th} levels of real earnings, the generalized Pareto does not seem to improve on the more simplistic split-histogram method, due to the irregular within bin distributions. By selecting polynomials flexible enough to fit well to observed values in binned data, this method creates implausible trends in levels of real earnings.

When these simulations are performed on the generalized least squares methodology, all measures described are found to be consistently estimated, which is demonstrated in Figures 1.33 through 1.39. It is not surprising that by assuming the correct underlying distribution, one can obtain accurate estimates of the parameters of the original data. This diagnostic methodology is able to reveal an additional benefit, this method of estimation is not influenced by scale driven differential crowding complications which affect the performance of the previously described methods. A final advantage, noted by Pinkovskiy and Sala-i-Martin (2009), is that this type of parametric estimation helps to correct for misreporting at the tails of the distribution. This is an advantage over other strategies which tend to overfit to potentially inaccurate values at the tails. Further research could alter the simulated data at these extremes in order to determine the robustness of this benefit.

These advantages should not lead the researcher to feel that this methodology is a silver bullet. If the underlying distributional assumption is incorrect, this method will perform worse than all those previously mentioned. This is clearly shown in Figures 1.40 & 1.41 which present estimates of the Theil derived from such misspecification. These show the effect of assuming an underlying lognormal distribution when the true population displays a Gamma or Weibull distribution respectively. All other measures show errors and biases which are at least as extreme, and thus will not be presented. The conclusion to be drawn is that despite the advantages of this method, using the least squares methodology improperly results in far worse estimates than nearly any of the alternative interpolation techniques which have been presented.

This necessitates a methodology to check for the proper distribution of a population given binned data. Conveniently, any goodness of fit test which is suitable for binned data will provide the criteria necessary for such a purpose. If underlying data follows any distribution from a variety of candidates, testing in this way will allow the researcher to avoid these pitfalls. In order to determine candidates, one will likely need to rely on existing literature, or test via contemporary micro datasets.

1.5 Correcting the Bias

All the simulations above were based on a set sample size and bin structure. Most of these methods demonstrated biases in estimations that resulted from bin crowding. Many biases of other statistical methods can be corrected or at least reduced by increasing sample size, so I will investigate whether this could also address the issues described here. Binned data does offer a complication to the strategy of increasing sample size. Increasing the sample from which binned data is calculated and presented does not actually increase the number of observations from which estimations are derived. For this to occur, the actual number of bins must be increased.

For these reasons I will perform two further analyses on the methods described above. First, the sample size will be increased by a factor of 10, effectively making the sample increase from 0.1% to 1% of the simulated population. The income from this sample will still be binned according to the same structure above and the resulting binned data will be used for estimation of inequality measures. It can then be determined whether increasing the sample size will remove the patterns of bias described earlier.

A second analysis will involve applying a much finer bin structure.[9] This new bin structure will be used to bin the exact same samples from the previous section, with the new bin values used in estimation. Again this is done to see if this type of change would reduce the biases explained for most of the methods above.

For the sake of brevity, I will only present the effects on the Theil and Gini, as effects of increasing sample size and changing bin structure are extremely similar for every other measure analyzed in the prior section. Additionally, because the simulations in the last section showed my results were robust to a wide variety of distributional parameters, for all results that follow I will only present the first iteration of these earlier results (i.e. the upper left hand panel in each of the prior treatments will be used as a point of reference.)[10]

These exercises will be performed on any of the methods from above which displayed a pattern of bias. Two general trends will hold for most of these methods. First, increasing the sample size does not remove the trend of the biases seen. It does reduce the variance in estimates from random sampling, but this tighter grouping still adheres to the trends in bias described above. Second, increasing the number and granularity of bins appears to reduce, but not completely remove biases. In addition, it nearly always flattens the shape of the bias, leaving a less systematic trend, while still having a directional effect on estimates.

When the sample size is increased for the direct method, the pattern of bias remains

[9]The above analysis uses a structure containing 9 bins, to show the other extreme the bin structure from the 1946 IRS SOI will be used which contains 31 bins and has greater granularity at high and low incomes.

[10]These panels represent a simulated population whose income structure follows a lognormal distribution with a location parameter of 3.5 and a scale parameter of 1.

identical. The only change is the size of band which represents 90% of the estimations from random sampling. Figures 1.42 & 1.43 demonstrate that this strategy only reduces noise due to sampling, but does not address the underlying issue related to bin crowding. When the number of bins and bin granularity is increased, though, the scaling has nearly no effect on the measure of the Theil. Figure 1.44 shows that measure stays nearly constant for the entire range. The Gini is still subject to the issue of placement of bins and resulting population segments that were discussed earlier, and figure 1.45 shows that this new binning structure lowers the magnitude, but does not remove the bias which can be created due to these mechanisms.

The Cowell "2/3 1/3" method presents similar results, with a smaller band of random sampling shown in Figures 1.46 & 1.47. The estimates of the Theil and Gini are now estimated with a fairly high accuracy, but still maintain a slightly downward bias, shown in Figures 1.48 & 1.49, demonstrating that the lower bound still seems to dominate the weighted average that comprises these measures.

The midpoint approximation is an interesting case, as changing the sample size does actually change the pattern of the bias. Specifically, it makes it more extreme. This is due to the fact that increasing the sample size allows for better observation of extreme values, creating many more samples in which the upper bin is included in calculations. This trend explains the similar shape, but higher level of estimated inequality indices seen in Figures 1.50 & 1.51. Another striking result is displayed in Figures 1.52 & 1.53. Prior results and analysis in other papers were suggestive that this method overestimated inequality indices. What is seen here is the fact that very fine bins both before and after the mean create a situation in which the bias caused by the unimodality issue explained above is now overwhelmed by the effect of assuming perfect within bin equality and as a result, for much of the range, inequality is underestimated.

The mean approximation returns to a pattern common to most of the methodologies. Specifically, increasing the sample size reduces the variance in estimates derived from random samples, but does not change the underlying bias, which is demonstrated in Figures 1.54 & 1.55. This is a somewhat unsurprising result. If one were able to increase bins to an arbitrarily high number, the mean approximation performed on binned data becomes isomorphic with direct estimation from weighted survey data. An extreme increase in bin number nearly eliminates the systematic bias, but still leaves a minor directional bias, as can be found in Figures 1.56 & 1.57.

The results for the split-histogram and generalized Pareto Methods are extremely similar. For both the increase in sample size reduces variation, but maintains the trend of bias, shown in Figures 1.58 & 1.59 for the split-histogram method and 1.62 & 1.63 for the generalized Pareto. Increasing the number of bins reduces the pattern and size of the bias to an extreme degree as seen in figures 1.60, 1.61, 1.64, and 1.65. This reinforces the idea that many of the problems with these methods arose from fitting unrealistic distributional shapes within bins. By greatly reducing the size of bins, the misspecification of these shapes becomes increasingly less important.

These results demonstrate that given a fine enough bin structure these methods come extremely close in performance to that of the least squares method. Unfortunately, as the researcher does not have the underlying individual level data for binned presentations available, a measure of what constitutes enough bins is fairly subjective. Future work may focus on defining spacing of bins relative to percentiles of the population and income within the bins, in order to build objective guidelines.

1.6 Results

Some general rules can be drawn for the researcher. First if it is reasonable to assume an underlying distribution for the data at hand, generalized least squares appears to perform best for a variety of purposes. If this is not the case, both the split-histogram and generalized Pareto Methodologies perform fairly well, though the bias present in estimations of indices and point estimates may be problematic for usage of these measures in any regression analysis. These problems are reduced as bin number and detail grow, but are unaffected by the sample size used to construct the binned data.

Given that many studies which employ this type of data are used to explain variables which can be highly correlated with mean income, this issue is by no means a minor one. Issues such as growth, the effect of education, or labor market institutions are likely correlated with both mean incomes and with the true level of inequality within a region or time period. This means that the methods described could create serious issues with endogeneity. Given the irregular, variable patterns of bias as well as the inability of the researcher faced with binned data to formally model this bias in a case by case basis, it is unlikely analytical exercises could be corrected in any consistent or effective manner.

Cross-regional and intertemporal comparisons may also be influenced by the issues outlined here. Given constant bin structures, it is not unusual to have extreme differences in mean incomes across geographic areas. In addition, whether bin structures remain constant or change across time, measures will be influenced by changes in nominal income, which can be driven by both inflationary forces and growth or recession within an economy. This implies that even for descriptive purposes, many of the methods described fall short, as spurious effects can create differences in the estimates of inequality that overwhelm those that may occur do to actual underlying changes in the distribution.

Even with these fairly serious concerns this chapter has demonstrated opportunities for researchers to benefit from binned income data. Given that this data can often be the best and only source of information on some segments of the population, historical time periods, or geographically specific areas, this may be especially useful in future work. By correctly estimating inequality from binned data, information on the trends, outcomes, and causes of this inequality can be better established.

2 The Distribution of Household Income in the United States, 1946-2015

2.1 Introduction

Earnings inequality in the United States has emerged as one of the single most important issues of public concern and policy debate. Although this surge of interest was preceded by considerable attention among economists, many of the well-known studies concerning inequality have focused on some specific aspects of inequality at the cost of neglecting others. Specifically, studies of inequality fall into three groups; those that focus on top earners at the expense of the remainder of the distribution, those that focus on the national level at the expense of intraregional and interregional trends, and those that focus on the short or intermediate run at the expense of the long run. To address these gaps, I implement a methodology from which I generate a new dataset, distinguished by its extension of historical data on the distribution of household incomes at both the national and state-level. The results reveal distinct new patterns of inequality that expand our understanding of historical inequality trends and are likely to prove pivotal to future inequality research.

For these purposes I use the binned Adjusted Gross Income (AGI) data present in the Internal Revenue Service's (IRS) Statistics of Income (SOI) reports. This allows me to create a state-level dataset of continuous earnings distributions for the main body of United States' earners by state and year from 1946-2015, thereby greatly extending the existing inequality data spatially and temporally.

This effort uncovers distinct changes in the U.S. inequality profile that had not been previously revealed. For example, I observe dramatic changes in the central segments of the household earnings distribution from the 1940s to the 1960s that entail large episodic increases in inequality and divergence in real earnings. Because these changes are described for annual household earnings from various sources, some of the changes may be distinct from those that involve individuals or hourly earnings. The trends here reveal many interesting and important dynamics concerning the extensive margin of earnings of both individuals and married couples which cannot be demonstrated using individual wage data.[11]

While studies on historical household earnings are in fact common in the literature, such studies are subject to the distributional limitations discussed earlier. For example, Piketty and Saez (2003) concentrate on the top 10% of household earners at a national level and while Frank (2009,2014) extended a similar analysis to state-level data, the focus for his description of inequality remained on the earners on the right tail of the distribution.[12] Sommeiller and Price (2018) provide a similar analysis where the upper percentiles are contrasted with trends from the remainder of the distribution and city and county level data are included in more recent years. These datasets do not provide an ability to investigate differential changes within the remaining part of the distribution, such as changes in real income at specific percentiles. As such, they may not be sufficient for analyzing issues related to wage inequality, recognized by Piketty (2014) as the primary driver in overall U.S. income inequality. Attention to the main body of earners across this particular time period is especially important to various labor market issues. Shedding light on the main body of earners is necessary for analysis of issues such as the prevalence of the middle class and the effect of labor market institutions including the minimum wage,[13] unionization,[14] and returns to skill.[15] Complete analysis of these types of issues requires data spanning back to key historical changes in many of these proposed drivers, including large educational changes in response to the G.I. Bill established in 1944, changes in the rate and type of unionization beginning in the 1940s, and legislative changes to the national minimum wage

[11] Examples include annual hours worked, non-wage sources of income, and dual income households.

[12] Frank (2009) does estimate state-level inequality indices of the entire population for his regression analysis though they are not directly presented. The underlying procedure used for some of his measures is subject to measurement bias related issues which are addressed in chapter 1.

[13] Autor, Manning, and Smith (2010) demonstrate that the real level of the minimum wage matters, not only for the left tail of earners, but may also have spillover effects into the main body of the earnings distribution.

[14] Union membership is shown by Dinardo et al. (1996) to drive inequality between the 50^{th} and 90^{th} percentiles of wage earners, and Card, Lemieux, and Riddell (2004) provide evidence that union membership also has differential effects on earnings based on skill level.

[15] Autor (2014) stresses that returns to skill are a major driver in U.S. inequality and apply mostly to "The Other 99%."

in 1961 and 1966. These specific events may also paint a clearer picture of the evolution of the middle class, especially given the popular notion of their rising affluence after World War II. Equally important is the fact that the falling of such affluence, the so-called hollowing out of the middle class, may be best understood in the context of the full post WWII history, rather than just a more recent slice of that history.[16]

An important feature of the data and thus of this paper, is its state-level coverage. One key benefit of such coverage is that it allows investigation of region-specific or state-specific trends, making it possible to examine whether the trend in inequality is homogenous across the nation. Additionally, many of the aforementioned drivers of inequality and their consequences can be examined by exploiting state and regional heterogeneities. The result will be a deeper understanding of the causes and the effects of inequality than has been possible. Finally, state-level data provides an opportunity to investigate whether changes in national inequality are driven by distributional changes *within* states or regions, or by changes in the relative mean earnings *between* them.

These considerations motivate the exercise at hand. By utilizing the Internal Revenue Service's Statistics of Income in conjunction with a parametric estimation approach, I construct estimations of the earnings distribution for the contiguous 48 states from 1946 until 2015. From these distributions I then construct state and national measures of inequality for the main body of earners, represented here by those between the 5$^{\text{th}}$ and 95$^{\text{th}}$ percent of the income distribution. This will also allow me to describe levels of real earnings at various percentiles of the earnings distribution.

I will also establish that the methodology used to obtain such estimates is the most appropriate choice for my efforts, given the importance of cross-state and cross-time comparability here, as well as the specific structure and qualities of IRS binned data. The consistency of this method is important for comparability of the measures themselves but

[16] Wolfson (1994) demonstrates that polarization cannot be accurately described by a single index. The use of Theil index, a measure used for reasons discussed later, does not address Wolfson's concern. But data on continuous distributions will.

could also be especially helpful when assessing causal relationships via regressions.

For the purpose of analyzing relative trends of inter- and intra-state inequality and their contribution to national inequality, the Theil is a natural choice. Its unique decomposition property allows me to estimate the relative intra-state and inter-state contributions to overall inequality as well as any divergence in earnings across states. Additionally, this decomposition will allow me to independently analyze the historical trends in each measure.

From this effort, a picture will emerge of a general increase in national inequality over the sample period with particularly acute rises in inequality observed from the mid 1940s to the mid 1950s, from the mid 1960s until the early 1970s, and across the 1980s. I also observe a relatively steady growth in inequality in the 1970s, and from the 1990s forward. Within state inequality measures for individual states generally follow a similar pattern. Also, while between-state inequality is not a large portion of the national measure, this component of inequality has been considerably more volatile. The time trend of this component changes direction much more often and severely than the within-state component. Many of these changes occur in time periods or segments of the earnings distribution that have not been explored in prior work, making it possible to not only describe the history of U.S. earnings inequality more fully and provide new insights into its evolution, but also to lay the basis for future analyses of its causes and its relevance to economic outcomes, thereby helping to inform policy decisions.

The rest of the paper will be organized as follows. Section two will discuss the features and advantages of the AGI data in the SOI over other income surveys, while section three will discuss the methodology necessitated by its particular features. Section four will discuss the results. Section five will compare these results to the existing literature and section six will conclude.

2.2 Data

Starting in 1916, the IRS has published annual data on individual earnings and taxes in the SOI. Of specific interest to this paper is the earnings data from 1946 onward, as prior to 1944 a majority of Americans were not required to file income taxes, and until 1946 the large number of earners deployed in World War II may drive the pattern of inequality. After a series of changes in the allowable deduction in the 1940s, the percentage of actual filers from the population of potential filers has consistently fluctuated between 85% and 95%, as demonstrated by Freeland and Hodge (2012). This makes the source from which the IRS draws its sample rather large.[17] While this population may exclude individuals with specific characteristics, Cilke (1998) shows that the majority of excluded individuals consists of dependents living at home and older individuals. These groups may be seen as outliers for the working population at large, and thus may represent an important, but distinct pattern from the ones drawn here.

The IRS SOI has several distinct advantages over the often-used Current Population Survey (CPS). First, the time period for which annual data are available offers two additional decades of information in the SOI as compared to many CPS surveys. The Annual Social and Economic Supplement does extend back this far, but data on early years are based on sample sizes a fraction of those found in the SOI and in later iterations of this CPS survey. Data from the decennial census is also available, and while this constitutes a similar sample size to that in the SOI, it does not offer yearly observations. In addition, the SOI performs sampling and weighting at both the national and state level, making it far more appropriate for the regional analysis that will be performed here.

Additionally, the SOI limits misreporting by the population from which samples are drawn: while other income surveys must rely on careful survey design and respondent training, information reported to the IRS carries with it a potential penalty for

[17] Additionally, as noted by Frank (2014) the pattern of the number of filers follows that of population changes in the United States after World War II.

misreporting. While there is still motivation to actively misreport, this incentive is greatest for those with the highest incomes, as the benefit of underreporting is the largest for those who pay the highest tax rates. This intuition is matched by evidence presented by Johns and Slemrod (2010) who show that as income increases, misreporting also monotonically increases until the top half percentile of the population. Because of this, this issue is far less important for the population of interest here than it is for studies on top earners. Other income surveys are not free of of this type of underestimation. Rector, Johnson, and Youssef (1999) demonstrate underreporting in the CPS exists at all levels of the income distribution, even in less affluent households. Also, as pointed out by Lemieux (2006), even if respondents honestly attempt to disclose earnings, some CPS measures are inherently subject to errors in estimation.[18] IRS penalties should discourage inaccurate data whether due to purposeful misreporting of income or to misestimation on the part of tax payers.

Finally, IRS data avoids top-coding and thus gives a much more precise measure of high earners and, in fact, tends to oversample these same individuals.[19] By contrast, the CPS top-codes above certain income thresholds. While top-coding is clearly problematic for observing the very highest earners, Schmitt (2003) shows how this censoring extends to lower earners in the distribution.[20]

With these advantages there are also specific challenges inherent to IRS data. First, there is no breakdown of the components of income at the state-level for all years in the sample period. To address this, I will rely on a measure of earnings that is consistently available: the pre-tax AGI.[21] While this may mask the effect of taxation and redistribution on inequality, pre-tax data will be more useful for explaining structural changes in the economy that influence inequality. Additionally, while the use of the AGI may limit precision in

[18]Lemieux shows that estimates of yearly earnings show greater variation than those of hourly earnings and asserts this is due to the high prevalence of hourly paid workers in the United States misestimating annual earnings.

[19]The sample drawn for the SOI represents a higher portion of the population for higher AGI brackets, the highest of which may include every individual in the actual population.

[20]Within the years studied, top-coding can be present for up to 10.3% of the workforce.

[21]State-level data on this measure were not published by the IRS for the years 1982-1988.

discussing specific types of earnings, it offers a more complete picture of inequality. Many labor studies actively exclude classes of earnings such as self-employment or focus on hourly wages. The former leaves out an important stratum of the population, while the later ignores the extensive margin of earnings (i.e., hours worked) and other sources of income.

As in many other income surveys, AGI is presented in binned form. Binned income data do not allow for observing each individual surveyed, but instead present data on the count of the population and the total amount of income falling within a given range of income values. Because IRS SOI binned data report this population in terms of "tax-units," which include both single and joint filers, my results can be interpreted as measures of household inequality, distinct from individual level data which are common in the literature. The difference between the two allows for a different set of questions to be asked. In the case of household level observations, the data allow for addressing issues such as the complementarity or substitutability of spouses' labor supply choices. For example, if a spouse experiences a decrease in their hourly wage rate, it may induce their partner to seek additional hours at their employment, or if they had not been employed, to enter the workforce. Effects of this type that occur on the extensive margin would be missed in literature concentrating on hourly wages and individual workers. Yet, understanding such effects is critical for many policy questions, especially in cases where households' consumption ability is of concern. In addition, household measures are arguably more directly related to certain inequality questions such as the strength or the hollowing of the middle class. Other studies dealing exclusively with individuals beg the question of what qualifies as middle class for households with two earners.

The usage of AGI as a measure of interest follows authors such as Frank (2009,2014) and Feenberg and Poterba (2000). All measures in these works report shares of income and measures of inequality that use AGI as the measure of group income and total AGI of taxpayers as the total income for the population of interest. This is a different treatment than those by Sommeiller and Price (2018) and Piketty and Saez (2003) who calculate total national and state income from Department of Commerce and Bureau of Economic Analysis data respectively. Both these studies make similar adjustments to AGI for their calculation of income shares in order to make them comparable with their definitions of total income. While these treatments differ slightly, the qualitative results in both cases remain extremely similar, suggesting that regardless of the precise definition of income the trends discussed would remain consistent.

2.3 Method

Because Adjusted Gross Income data are presented in "bins," using AGI for measuring inequality presents specific challenges. The binned income data in the SOI only provide information on the number of filing units and the corresponding total AGI for each given bin. Additionally, the number of bins and their thresholds (i.e., the values of AGI at which each bin begins and ends) change throughout the sample period.

Chapter 1 establishes an issue in the estimation of distributions and inequality measures from this type of data. Specifically, spurious results can be created because the estimated measure of inequality may change due to levels of mean income and not solely due to changes in the underlying distribution of income. This issue is of specific concern here for a variety of reasons.

First, mean AGI can vary by a fairly high degree across states. In the years studied, the highest mean income for a state in a given year ranges from 1.4 to 2.3 times the lowest mean income. These values mean that comparison of state inequality indices would likely be influenced by the issues explained in chapter 1, as these ranges are within the areas where large differences in estimations are seen.

In addition, the structure of bins used for AGI are often coarse, with the number of bins used for describing AGI ranging from as few as 5 to as many as 31. With this type of range, there are clearly many years which do not have fine enough bin structure to address the systematic bias described in chapter 1. In addition, this bin structure can remain constant across years. For example, the periods of 1970-1975 and 1979-1987 each have bin structures which remain constant. Given that in both these periods there were episodes of high inflation, this data source is also sensitive to the effects described in chapter 1 regarding constant bin structure and changing nominal incomes.

These qualities make several available methodologies used to map binned income data to inequality measures problematic for IRS AGI data. The method which was least sensitive to these type of issues was the least squares method. In order for the least squares technique to perform well though, the underlying data needs to follow a known distributional form. There is a large literature which studies the shape of the income distribution, and establishes the performance of several candidate distributions for this purpose.

An extremely popular assumption for estimation of income has been the lognormal distribution. Gibrat (1931) offered the theoretical basis for why a lognormal distribution approximates earnings for a majority of the population[22] and Atkinson (1975) and López and Servén (2006) demonstrate that the hypothesis of a lognormal distribution cannot be rejected. This distribution is also known to do well for the main body of the earners[23] which is the focus of this paper. However, despite the popularity of the lognormal distribution and its wide use, several other distributions have been found to be suitable, especially for households and for specific time periods. For example, Salem and Mount (1974) find that the gamma distribution provides a superior fit to lognormal for U.S. household income in the 1960s. Bandourian, McDonald, and Turley (2003), who also concentrate on household income, show that the Weibull distribution offers the best fit of a variety of two parameter distributions for several European countries as well as the U.S. for select years between 1967 and 1997.

The availability of public use micro-files from the IRS from 1962 onward allows me to select a method which choses the best distribution among the discussed candidates and

[22]Gibrat creates a theoretical model that assumes that current earnings result from past earnings multiplied by a random component. This results in a lognormal distribution.

[23]Estimates of percentage cutoffs for best performance vary slightly, but suggest that, when excluding the extreme tails, this is the most appropriate distributional form.

to check the validity of my general strategy.[24] This data has individual level observations with both AGI and relative weighting for that tax unit. As such it is possible to calculate inequality measures of interest for the sample population. It is also possible to bin the individual level data based on the structure present in the SOI for that year. Then that binned data can be used to estimate parameters for the three distributional candidates. These parameters are used to estimate the same measures of inequality that were calculated for the sample population. From this exercise I find that the best performing distribution, i.e. the distribution that provides the closest estimate of an inequality index as compared to one calculated for the underlying population, is also the distribution which demonstrates the lowest Jensen-Shannon divergence. This measure was developed by Lin (1991) and is defined as:

$$JSD(P|Q) = \frac{1}{2}K(P|M) + \frac{1}{2}K(Q|M), \;\; where\, M = \frac{1}{2}(P(x) + Q(x))$$

In the above formula, the K term is the Kullback Leibler devergence, defined as:

$$K(P|Q) = \sum P(x) log\left(\frac{P(x)}{Q(x)}\right)$$

The divergence here is calculated between the empirical quantiles, $P(x)$, calculated from the binned data and the theoretical quantiles, $Q(x)$, calculated from the parameters of each estimated distribution. When this criteria is used to select distributions, the estimates of inequality obtained from the least squares method performed on binned sample data are extremely similar to the true measures for the underlying sample. In addition, they closely track the time trends in the measures derived from individual level data. As such, the distribution for each state and year will be selected by this criterion.

Divergence measures are especially appropriate in this case, as they allow for a

[24]While these micro files provide an excellent check on methodology, they themselves are not appropriate for estimating the measures of interest here. They represent a shorter time period, are not weighted for state-level estimation, and are a subsample of the data in the SOI.

comparison of the best fitting distribution when there are several candidates to select from. In addition, other goodness of fit tests such as the Kolmogorov-Smirnov (KS), Anderson-Darling (AD), and Chi-square (CS) have been shown by Jäntschi and Bolboacă (2017) to reject the null hypothesis of zero divergence between two distributions when outliers or heavy tailed distributions are present. This could be especially problematic for my purposes as the Weibull and Lognormal, two of the candidate distributions, are themselves heavy tailed.

Using this strategy, distributional parameters will be recovered by year and by state that fully describe their respective earnings distributions. From these parameters annual, state-level inequality measures can be estimated in a manner which makes the calculated values comparable across states and time. Among these measures, the Theil index is of specific interest due to its decomposability property. It is possible to separate the measure of an entire population into mutually exclusive components which can be summed to the total value.[25] Here, this will allow a comparison of how much of the national level of the Theil is due to inequality inside of states, and how much is due to the divergence in the average income across states. In addition, the individual patterns of each of these measures can be observed for the sample period. The estimated parameters will also allow discussion on changes in the structure of real earnings.

2.4 Results

Using the above method, Figure I presents estimates of the Theil index at the national level, as well as its inter- and intra-state components. The first striking feature is a general increase in inequality, starting in the immediate aftermath of World War II. Even with this general trend, one observes episodic and relatively more severe increases in inequality for the

[25]The Theil is defined as: $T_t = \sum_{i=1}^{n} y_i ln\left(\frac{y_i}{x_i}\right)$, with y_i & x_i as the percentage of income and population comprised by individual i respectively. This can be decomposed into a within and between group component: $T_t = \sum_{k=1}^{n} y_k T_k + \sum_{k=1}^{m} y_k ln\left(\frac{y_k}{x_k}\right)$, with k groups or regions.

periods 1947-1956, 1963-1969, and to a lesser extent, across the 1980s.[26] This is especially notable as the first two periods pre-date (completely or partially) the coverage that has been available to most previous studies concerning the group of interest here. This result challenges the popular belief that the rise in inequality is a more recent phenomenon, at least among households. In fact in most other periods, the growth of inequality is surprisingly steady, with more extreme upward trends limited to the discussed time frames. Another result emerges from the decomposition of the National Theil index in Figure I. Here it is shown that the largest component and primary driver of national inequality is within-state inequality. While the importance of this component is at its lowest in the 1950s, it never represents less than 90% of the national measure. This result is not surprising, given that the between-state component is effectively calculated using the average income in each state. The relative size of these measures reflects that there is much greater disparity of income within states than there is between states' average income.

Nonetheless, the between-state component of the Theil is still informative for analyzing inequality across regions. Figure II summarizes changes in this measure. Divergence between state average incomes rises sharply for nearly a decade after the start of the sample period. It then declines until the late 1970s, albeit with greater variability. This is then followed by an increase in the inter-state divergence over the 1980s. Inter-state divergence then declines and is stable for the 1990s but oscillates after that point. To recap, the most important lesson from this aspect of the paper, afforded by my new data, is that inter-state divergence increased dramatically in the 1950s and was followed by an equally dramatic decline in mid 1960s to mid 1970s. Subsequent increases over the next five decades never reach the 1950s level of inter-state divergence.

One potential issue with using the Theil or any other single-valued measure of inequality is that a given change in an index could be the result of different underlying

[26] Many data points are missing for the 1980s as the IRS did not publish the necessary tables for these years. But because data are available for the beginning and the end of this decade, a general trend over this time period can still be discussed.

changes in the distribution of earnings. To address this issue, and to uncover the underlying changes in my case, the above estimation technique used for each state is repeated for the country as a whole. AGI groups were aggregated for the United States, and distributional parameters were estimated at the national level. The Theil calculated from this aggregate estimation is presented in Figure III, along with the Theil calculated by state-level estimates of distributions, which are subsequently used to estimate the within and between state components of a National Theil. As can be seen, the National Theil calculated by estimating a distribution for the nation as a whole shows an extremely similar pattern to the prior treatment, both qualitatively and quantitatively, implying that the prior results are not driven by varied state-level distributional assumptions.

Estimating the Theil index at the aggregate (national) level, one can then explore the underlying dynamics that influence the shifts in the overall index. To do this, I compare the share of income held by the top and bottom of the population over the entire period. Figure IV shows the relative share of total national earnings of the top and bottom 10% of the group of interest here (the 5^{th}-14^{th} percentile and the 86^{th}-95^{th} percentile of the total population). By comparing this trend to that in Figure III, I note that nearly all of the "dips" in the Theil index correspond to similar dips in the percent of earnings of the top 10%. Of course, the resulting decrease in inequality is only temporary as it bounces back along with the share of the top income group. On the whole, the share held by the relatively wealthy has risen over the entire post-WWII period. Moreover, the trend of the income share held by the bottom is largely downward sloping. This decline is seen as early as the mid-1940s to 1950s. These early post-war observations are documented in this paper for the first time.

But this does not tell the full story. So far, I have exclusively studied trends in relative shares, not in the levels of real earnings. This distinction is important to differentiate between possible causes for the same change in the discussed measures. For example, small "dips" in inequality, or decreases in the percentage of income held by top earners, could be caused by; decreased real income at the top with no changes at the bottom, income gains solely for lower earners, or simultaneous changes in both groups. The reverse could hold for an increase in inequality. To understand such changes better, the real earnings of the 10^{th}, 50^{th}, and 90^{th} percentile of earners are shown in Figure V. As can be seen, the lowest earning households show stagnant levels of real earnings for the entire sample period. The middle of the distribution shows steady, if modest, growth from the mid 1940s until the mid 1970s with a somewhat flat and even slightly declining profile thereafter. The highest earning households however, experience a rapid growth of earnings until 1968, followed by a smaller upward trend over the 1980s, and from the late 2000s until the end of the sample, interrupted by small short-run dips. The figure demonstrates decisively that the gap between the top and middle earners grew over the entire period. The dramatic jump in the earnings of the 90^{th} percentile begins in 2011, likely associated with the sharp increase in the equity market boom after the Great Recession, accompanied by the absence of any rise in real income of the mid-income group, dramatically increasing the extent of the real earnings gap.

It is again apparent that changes among high earners are the primary driver to changes in the Theil. As can be seen by comparing Figures III and V, dips in inequality represented by the Theil are principally linked with drops in real earnings in this top income group and less by gains among others. The fact that the episodic declines in inequality, when they occur, are due to the reduced real earnings of the top earners, rather than increased earnings of the bottom, is somewhat at odds with the popular notion of the source of increased equality.

Figure VI presents time trends in individual state Theil indices while Figures VII and VIII are the state-level equivalent of Figures IV and V respectively. While a full analysis of inequality across 48 states and 70 years is beyond the scope of this paper, some general trends can be observed. First, the patterns discussed seem largely mirrored within states, with general increases in inequality and stagnant levels of earnings for those on the low part of the distribution. Additionally, there is a great deal of heterogeneity among states. Given that many of the causes of inequality discussed earlier (e.g., minimum wage, unionization, returns to skills) exhibit state-level heterogeneity, this dataset may prove useful for using statistical inference to assess the contribution of these and other factors to overall inequality.

2.5 Comparison to Existing Literature

This dataset offers coverage which has not been available in the literature. Even given this, it will be informative to compare the results presented above to those that involve similar groups, time periods, and geographic focus. This will serve both to explain the novel nature of the results presented and show how they may be harmonized with existing data in order to give a wholistic view of the history of United States inequality.

One of the best known studies on the topic of historical United States inequality is that by Piketty and Saez (2003). As explained above, their coverage is different than that here, both because it focuses on top earners and because of its limitation to national level observations. In addition to these differences, the work here presents interesting contrasts to the conclusions made by these authors. Whereas these authors find the share of income held by the top decile of earners was extremely stable from the post war period until the 1970s, these are the same years during which some of the largest changes in the main body of the distribution took place. The work above demonstrates this to be the time period in which real incomes and income shares spread the fastest. This is also when these measures displayed their fastest increase for those at the top end of my population of study, showing that the trends among those with extremely high incomes differ greatly from those with relatively high ones. While these contrasts do not demonstrate a contradiction between the results, they do show the importance of looking at a greater segment of the population in order to describe inequality. If one were to only focus on the popularly used measure established by these authors, it would appear as though this time period was fairly uneventful.

Given this contrast, it may also be beneficial to see if the trends described here are consistent with those estimated for top earners. For this purpose, I compare my estimations with those of Frank (2009). Frank's series is ideal as he uses a definition of income consistent with the treatment here and offers a similar level of coverage, both in years and geographically. Frank's dataset establishes the trends of income shares among the top decile of the population by state and year. Using my distributional estimations, I am able to calculate the share of income held by the bottom 90% of the population, which directly implies what would be held by the remaining top decile. Figure 9 compares Frank's series to that implied by my work here. As can be seen these trends match extremely well. This is both suggestive that the results found here are consistent with prior work and demonstrates that in the future they may be harmonized with estimates of segments of the income distribution which are not the focus of this chapter. This provides the opportunity to describe the full income distribution of the United States. It also means this analysis can be performed for geographic areas and time periods which have not yet had this level of coverage.

There are some studies which deal with historic income within the main body of the distribution. Among these are the work of Budd (1970), who uses a variety of data sources to estimate annual, national Lorenz curves from the post World War II period through the 1960s. Given this type of analysis, his quantitative results are not directly comparable to mine, but the major qualitative results are consistent with the results here. Namely, he shows gains in income among individuals who fall between the 40^{th}-50^{th} and 94^{th}-96^{th} percentiles of the population. In addition, he shows decreases in income in his very top group, which is consistent with the trend of income among the top percentile of the national population demonstrated by Piketty and Saez.

As described earlier, the focus of this chapter also differs from other studies in its definition of the population and income of interest. Specifically the results here are all on the extensive margin of earnings, by which I mean annual income of tax units which are often composed of joint married filers as well as individuals. It may then also be interesting to contrast this with studies that deal with similar observations on the more intensive margin. For example, Goldin and Katz (1999) focus on male workers and use weekly earnings as their measure of income. At this unit of observation they still do observe increases in inequality, measured by the 90/10 log wage differential, from 1950-1970. Where my measures differ is the fact that the increases shown in this time period by Goldin and Katz are no more extreme than those observed in the rest of their sample period. This is suggestive that increased labor force participation by women may change the story of inequality significantly. In addition, they present results on returns to skill, showing large increases in this measure from 1950-1970, a decrease from 1970-1980, and then increases thereafter. Given the similarity in these directional results to those presented with regards to changes in Theil, it is very likely that returns to education are very important to the patterns described above.

Goldin and Margo (1992) use hourly earnings among males as their measure to establish trends in the wage structure over a similar time period. Their results on the log 90/10 wage differential are generally similar to those described above, but show a much flatter profile across the 1960s. This result differs from both my results earlier and those found by Goldin and Katz. The differences to Goldin and Katz demonstrate how results on inequality may change if one focuses on wages which may display trends that do not perfectly match with that of total income in a given time period.

While the results of this chapter do not contradict previous work in the area of historical inequality, they do demonstrate novel findings. Given that these findings are on a unit of observation that may be most comparable to the popular idea of inequality, they are certainly important to the population at large. In addition, as a measure of annual earnings for the main body of the population of households is important to consumption and welfare, this measure is also relevant to researchers and policy makers who are concerned with these outcomes. Finally, given that this measure presents trends that differ from those at other levels of observation, it is clear that the dataset created here could not be appropriately proxied by these prior findings.

2.6 Conclusion

The purpose of this paper has been to establish a dataset of state-level income distributions for the main body of earners (5^{th}-95^{th} percentile of the population), across a longer time period (extending back to WWII), and with a level of geographic precision (the 48 contiguous states) that had not been previously achieved. This is done for two reasons: to provide a more complete picture of the history of U.S. earnings inequality, and to help with future analysis of the causes and outcomes of inequality. These contributions are especially important given the intense debate among scholars and policy makers over the drivers and the consequences of inequality in the United States. The omission of the main body of earners in the early post-WWII period has deprived researchers from subjecting various hypotheses to the data from this *formative* and critical period. This work is now possible with this contribution.

The mapping from IRS binned data to continuous distributions for these purposes has called for selecting a method that outperforms alternatives in a non-trivial manner. The discussion on various methodologies in chapter 1 and on the qualities of the data here demonstrate the the methodology implemented creates comparable and accurate estimates of the measures of interest.

The main findings here are that while inequality has generally grown in the U.S. over this period, there are noticeable spikes, some of which occur earlier than prior literature has demonstrated. Additionally, when decreases in the measure of inequality are observed, this is generally due to income losses by high earners, rather than gains by low earners. A clearer picture of the history of the middle class also emerges. While the gap between the middle and top has consistently increased across the sample period, the middle class did exhibit gains in real earnings early on. However, this trend has flattened in the recent past. These results are especially important as much of the current literature seeks to explain changes that occur in the later part of the sample period here. Those explanations, while sufficient to explain changes in inequality in their respective time periods, are often specific to that time and therefore may not explain what factors influence inequality in the U.S. historically. Due to this, they do not establish the only avenue for changes in inequality, and do not provide a complete roadmap for avoiding or promoting certain changes in the income structure of the nation. Also, by following a strand of the literature which estimates inequality at the household level, the dataset created here is more suitable for discussing specific trends in inequality and policy implications. These qualities of the data make it especially well-suited for answering a host of economic questions, including those regarding labor market institutions, skill premia, and the middle class.

3 Figures

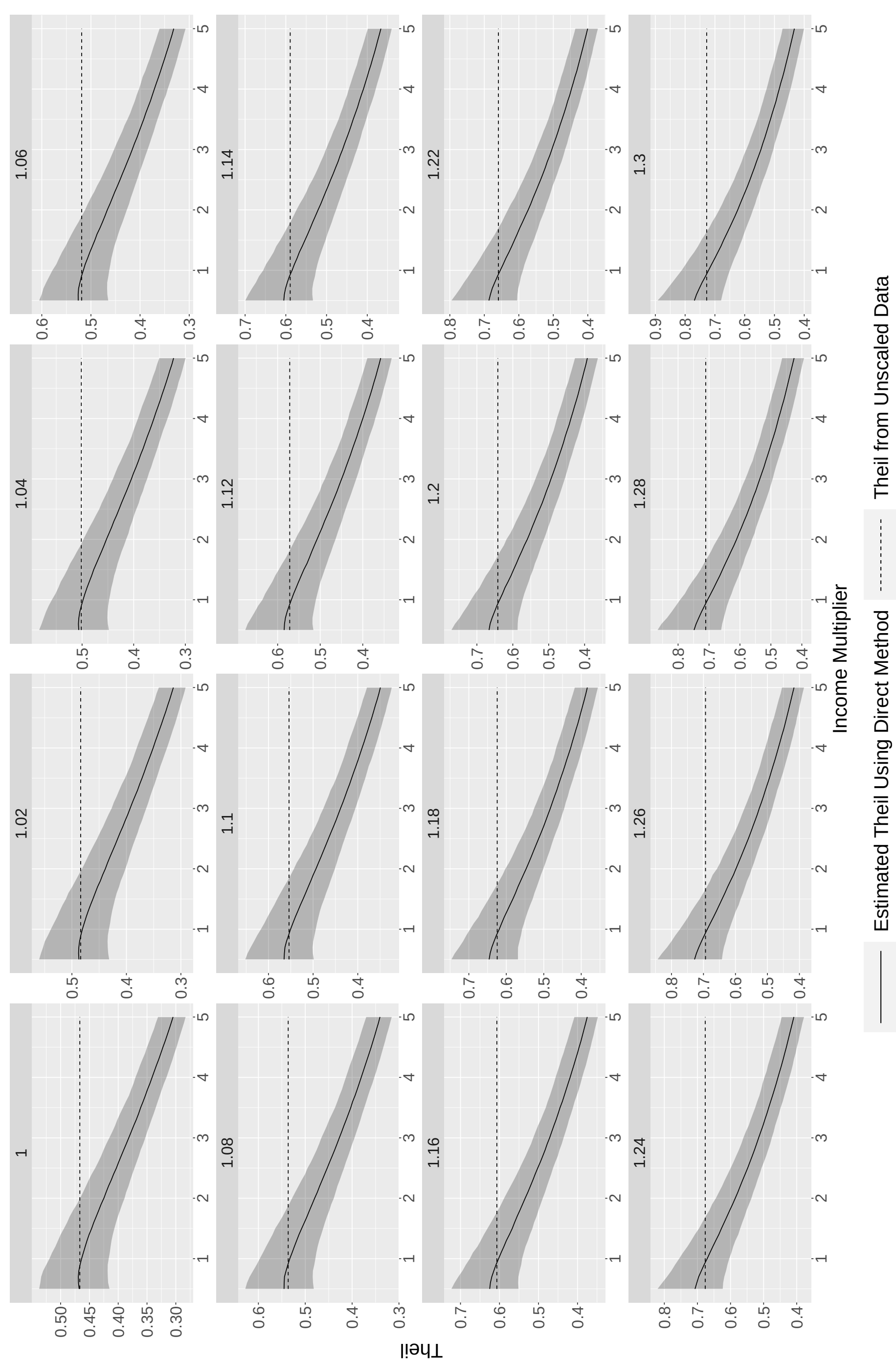


Figure 1.1: Performance of "Direct Method" for Estimation of the Theil Index
Theil estimated using the "Direct Method" on binned data created by random sampling from a population with a lognormal distribution of incomes with a location parameter of 3.5 and a scale parameter varying from 1-1.3. Samples then scaled from one half to five times their original value. Line is median estimate, bands represent the central 90% of estimates from random sampling from the distribution.

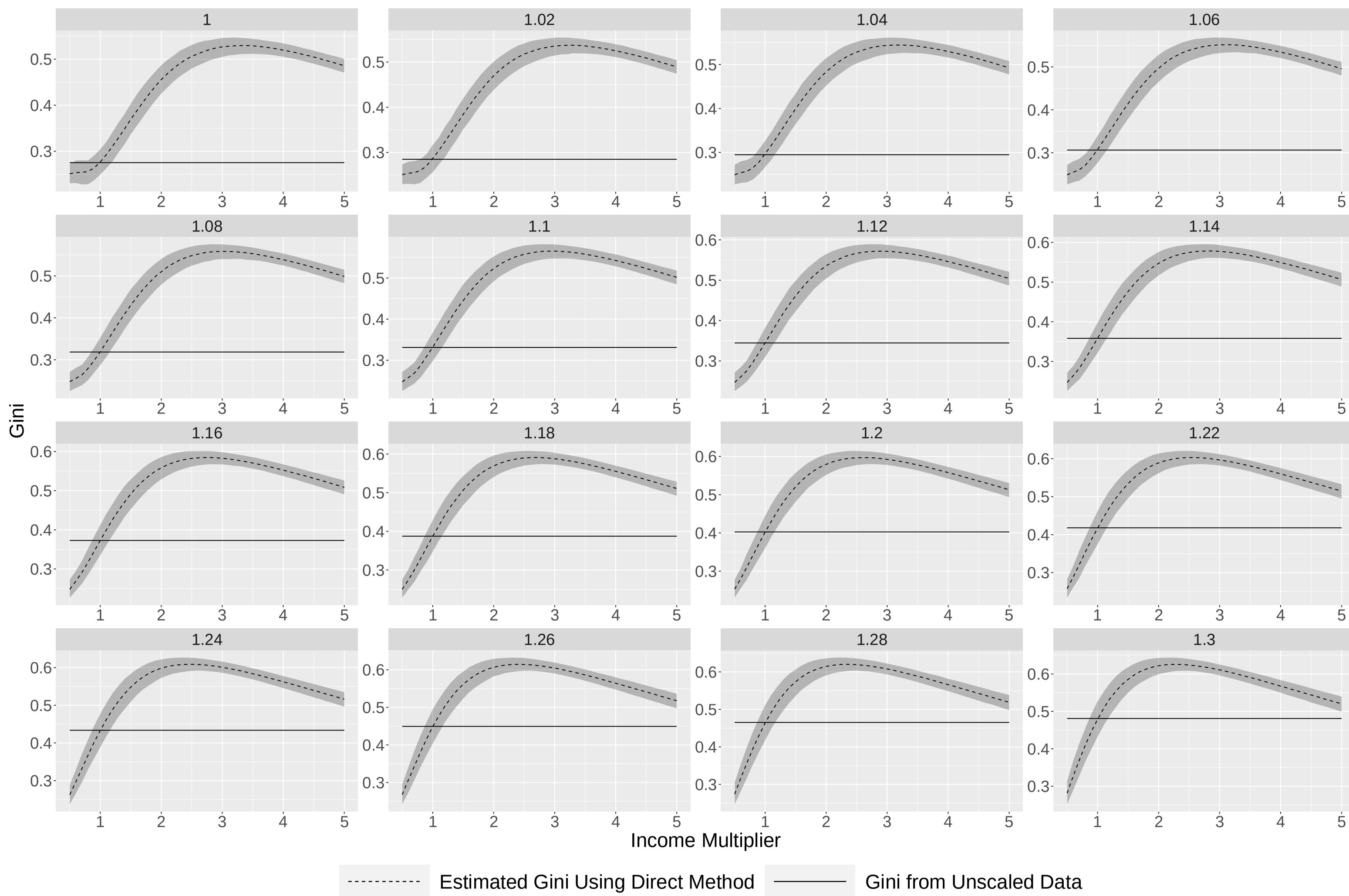


Figure 1.2: Performance of “Direct Method” for Estimation of the Gini Index
Gini estimated using the “Direct Method” on binned data created by random sampling from a population with a lognormal distribution of incomes with a location parameter of 3.5 and a scale parameter varying from 1-1.3. Samples then scaled from one half to five times their original value. Line is median estimate, bands represent the central 90% of estimates from random sampling from the distribution.

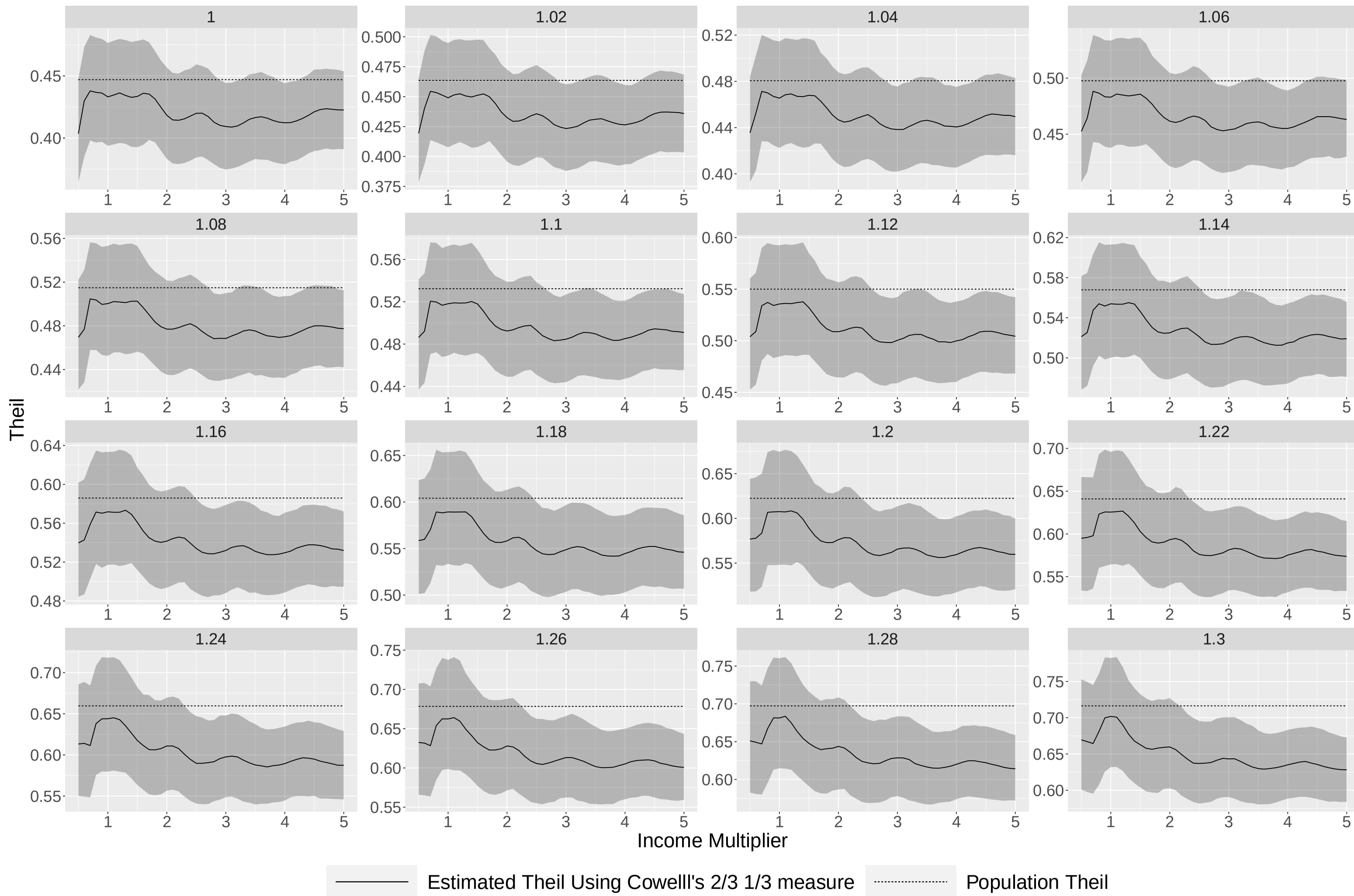


Figure 1.3: Performance of Cowell's compromise for Estimation of the Theil Index
Theil estimated using the Cowell's "2/3 1/3" compromise on binned data created by random sampling from a population with a lognormal distribution of incomes with a location parameter of 3.5 and a scale parameter varying from 1-1.3. Samples then scaled from one half to five times their original value. Line is median estimate, bands represent the central 90% of estimates from random sampling from the distribution.

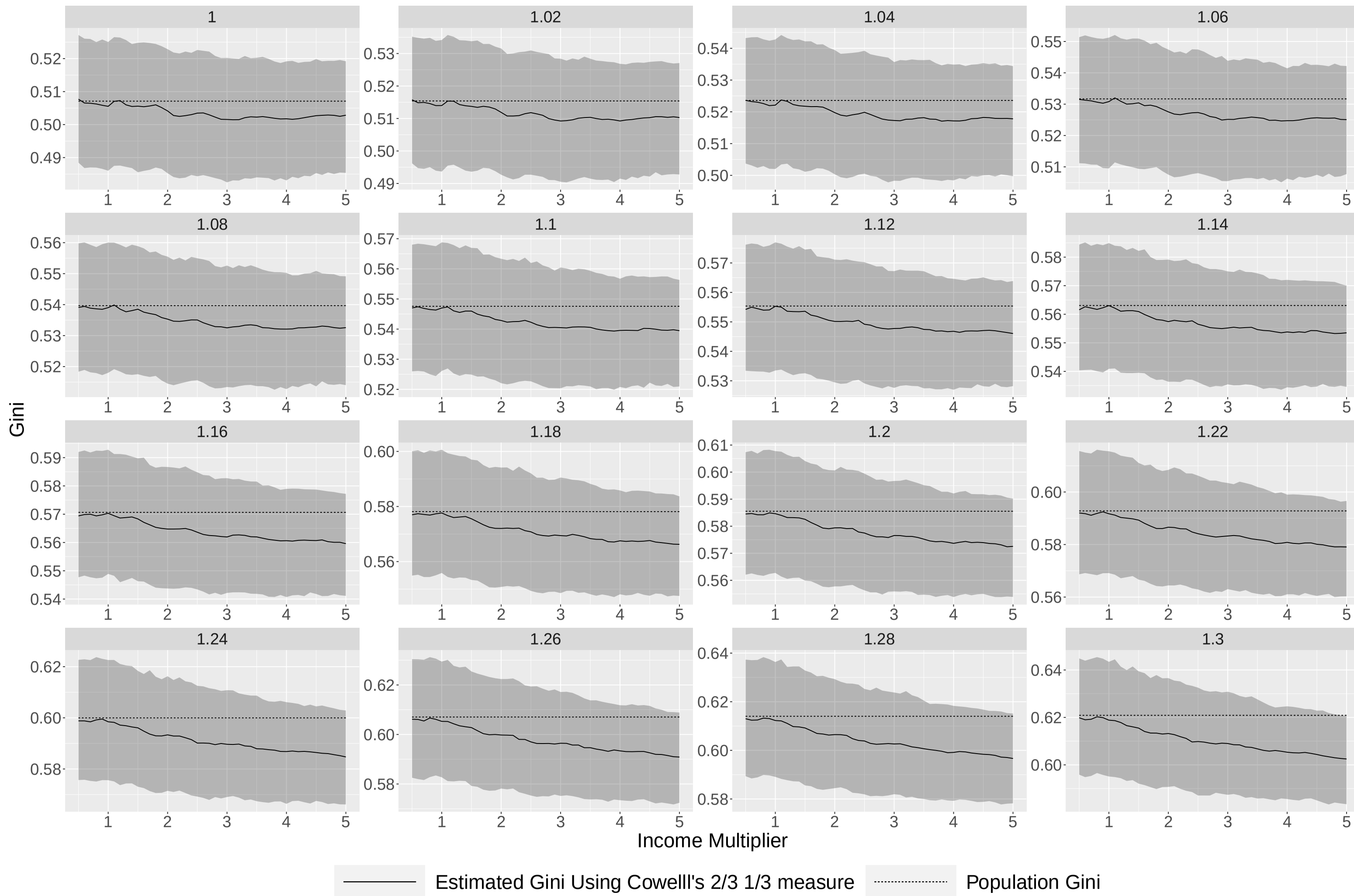


Figure 1.4

Gini estimated using the Cowell's "2/3 1/3" compromise on binned data created by random sampling from a population with a lognormal distribution of incomes with a location parameter of 3.5 and a scale parameter varying from 1-1.3. Samples then scaled from one half to five times their original value. Line is median estimate, bands represent the central 90% of estimates from random sampling from the distribution.

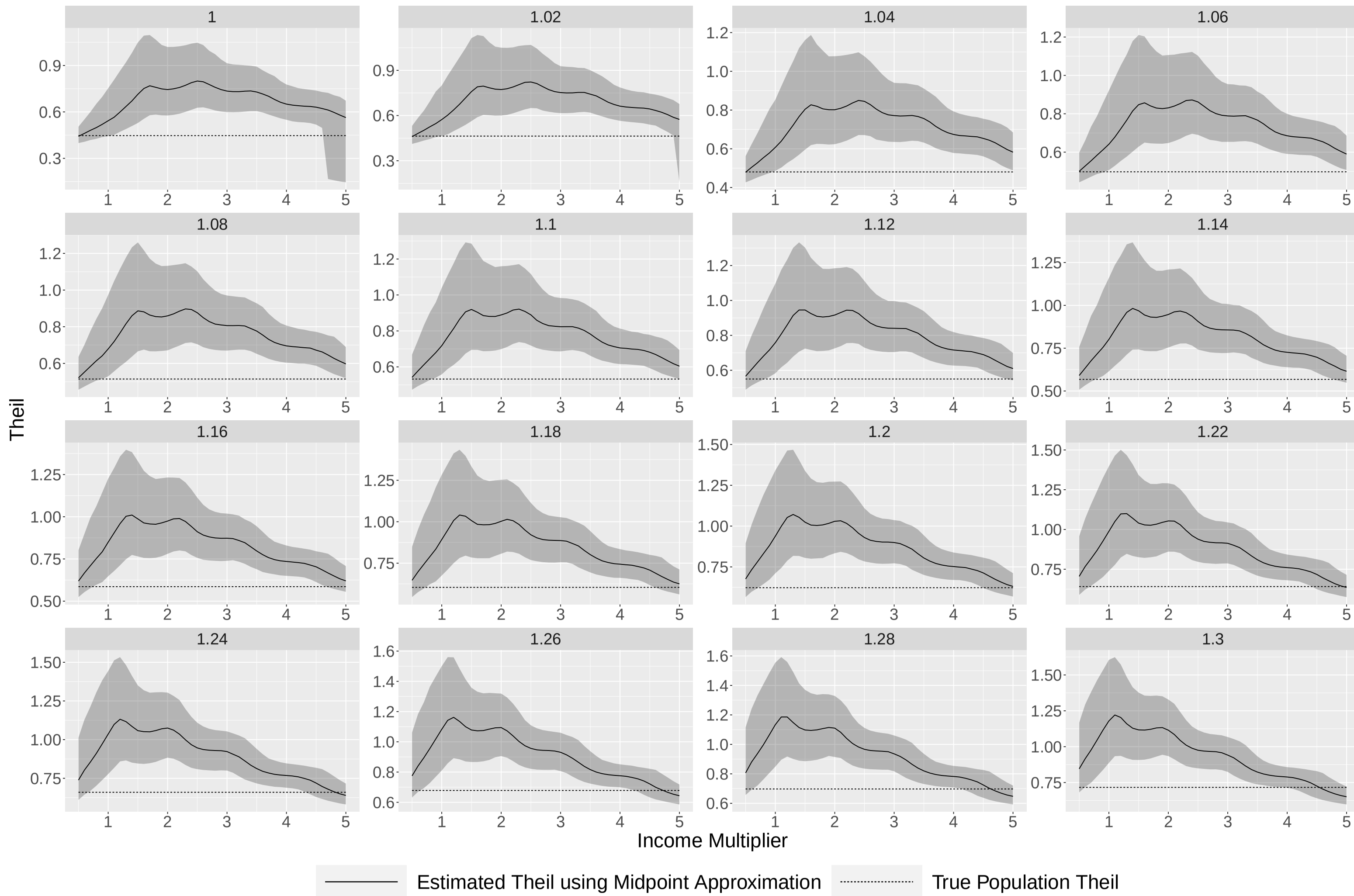


Figure 1.5: Performance of "Midpoint Approximation" for Estimation of the Theil Index
Theil estimated using the "Midpoint Approximation" on binned data created by random sampling from a population with a lognormal distribution of incomes with a location parameter of 3.5 and a scale parameter varying from 1-1.3. Samples then scaled from one half to five times their original value. Line is median estimate, bands represent the central 90% of estimates from random sampling from the distribution.

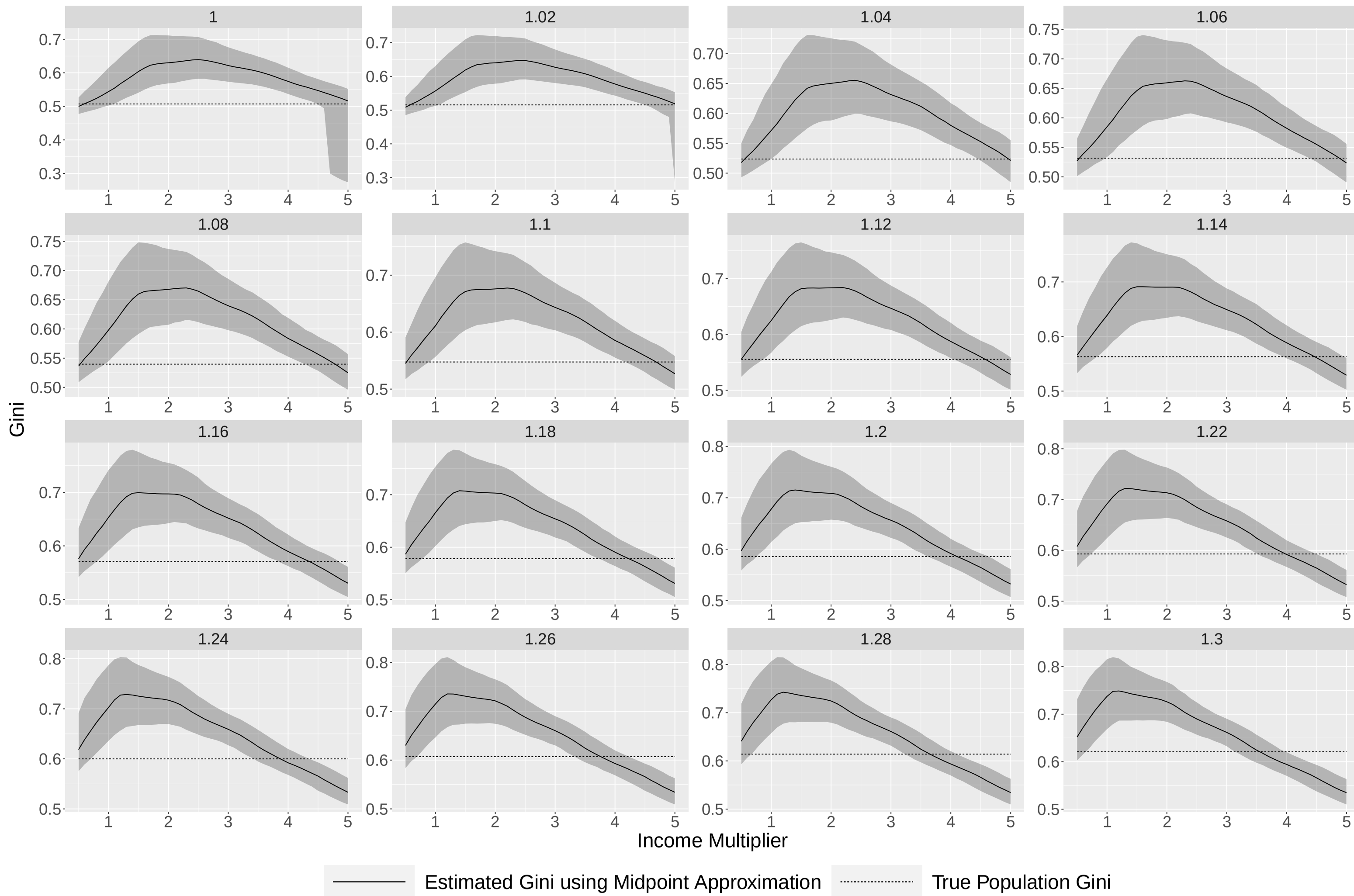


Figure 1.6: Performance of "Midpoint Approximation" for Estimation of the Gini Index
Gini estimated using the "Midpoint Approximation" on binned data created by random sampling from a population with a lognormal distribution of incomes with a location parameter of 3.5 and a scale parameter varying from 1-1.3. Samples then scaled from one half to five times their original value. Line is median estimate, bands represent the central 90% of estimates from random sampling from the distribution.

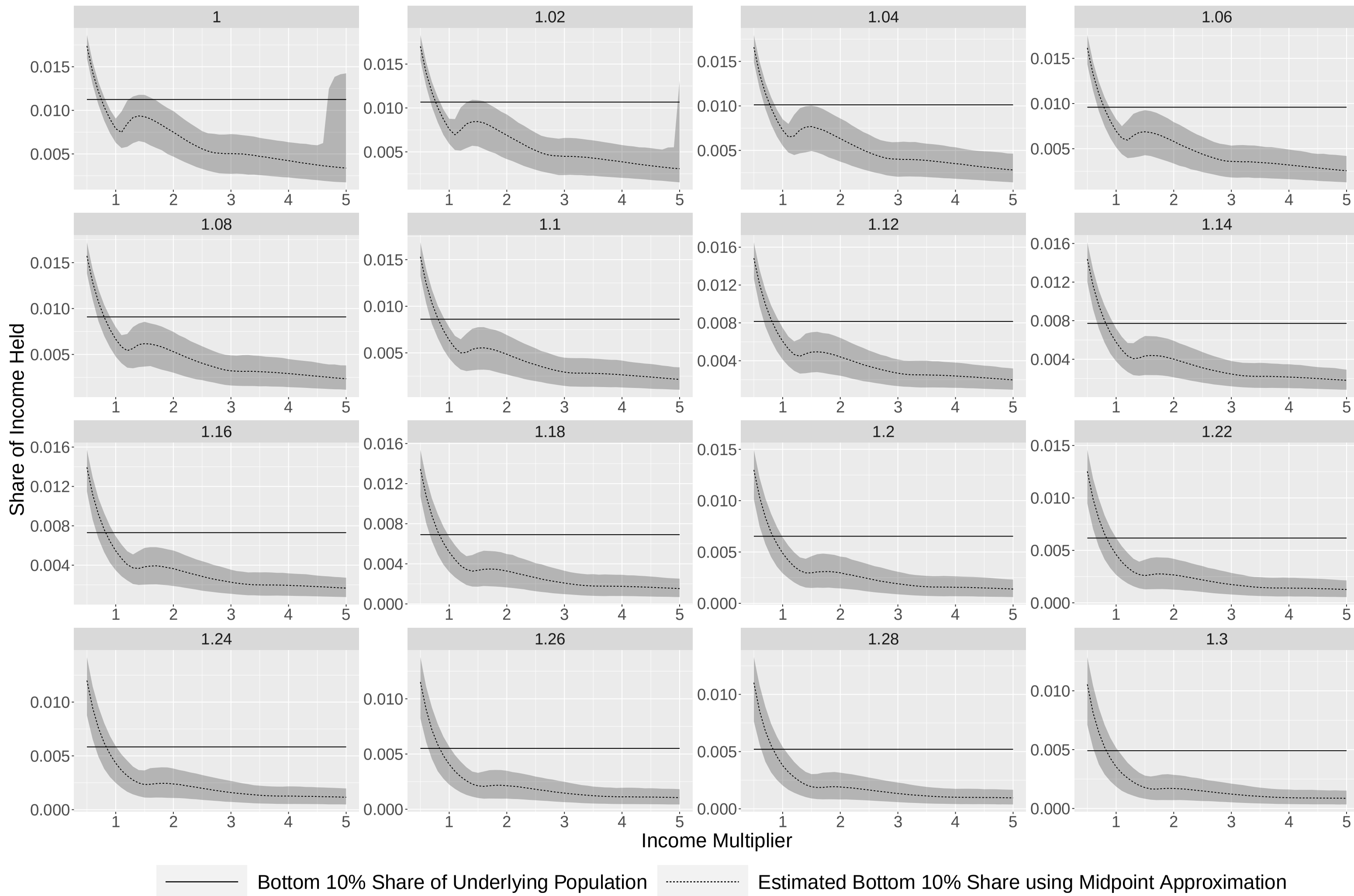


Figure 1.7: Performance of "Midpoint Approximation" for Estimation of Bottom Decile Population Share of Income
Bottom decile share of income estimated using the "Midpoint Approximation" on binned data created by random sampling from a population with a lognormal distribution of incomes with a location parameter of 3.5 and a scale parameter varying from 1-1.3. Samples then scaled from one half to five times their original value. Line is median estimate, bands represent the central 90% of estimates from random sampling from the distribution.

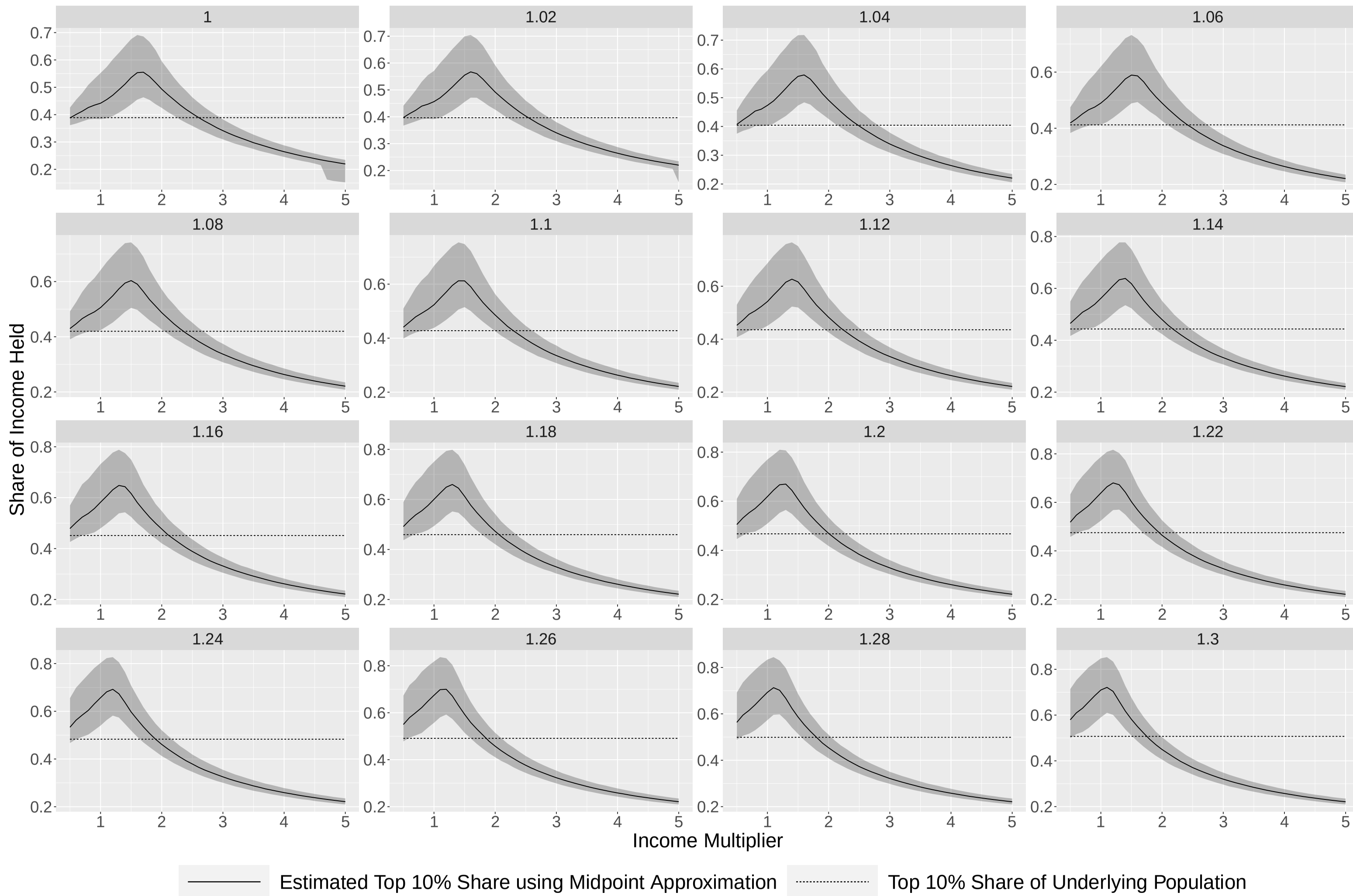


Figure 1.8: Performance of “Midpoint Approximation” for Estimation of Top Decile Population Share of Income
Top decile share of income estimated using the “Midpoint Approximation” on binned data created by random sampling from a population with a lognormal distribution of incomes with a location parameter of 3.5 and a scale parameter varying from 1-1.3. Samples then scaled from one half to five times their original value. Line is median estimate, bands represent the central 90% of estimates from random sampling from the distribution.

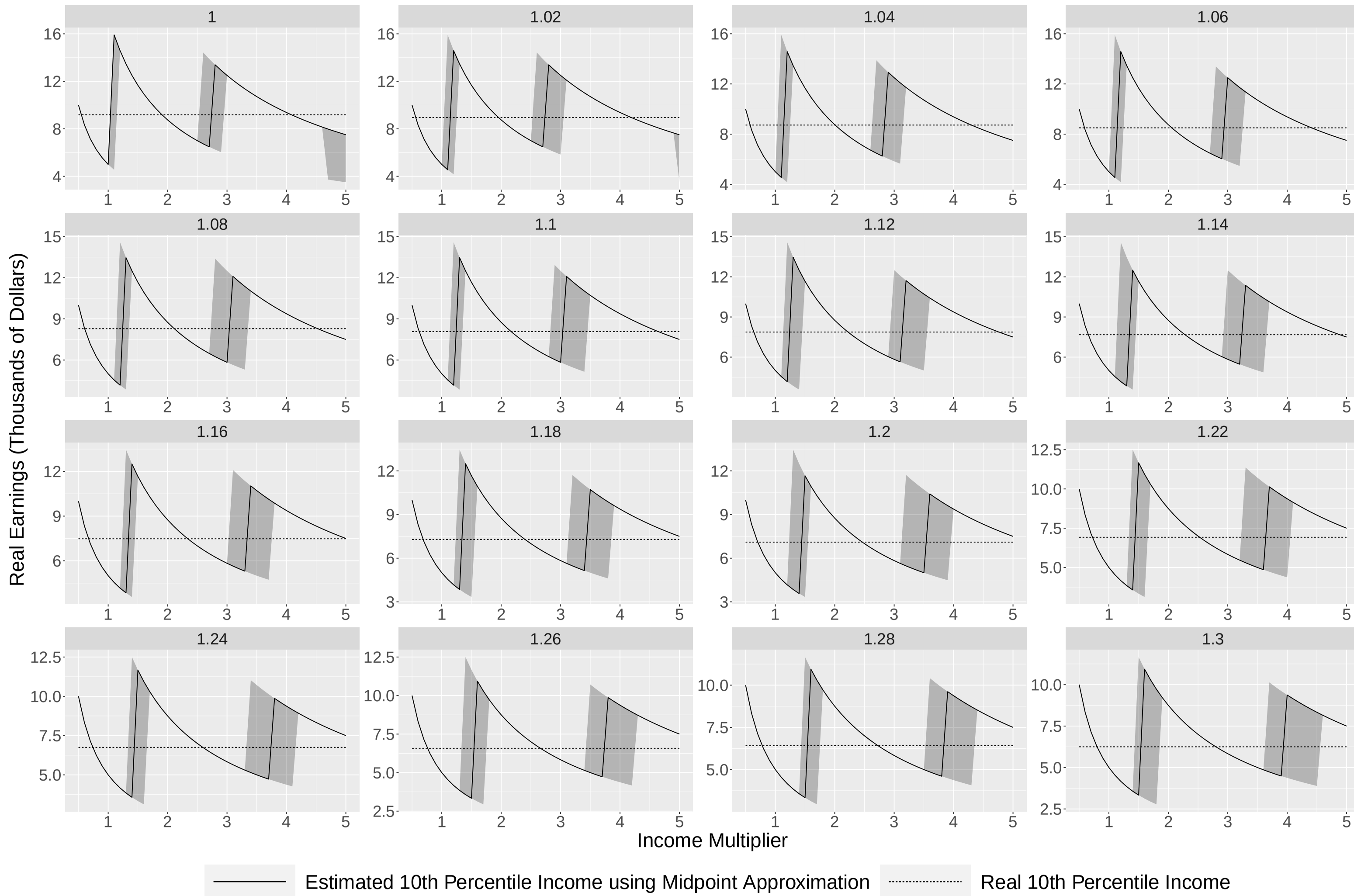


Figure 1.9: Performance of "Midpoint Approximation" for Estimation of 10^{th} Percentile of Income

Real income level for the 10^{th} percentile estimated using the "Midpoint Approximation" on binned data created by random sampling from a population with a lognormal distribution of incomes with a location parameter of 3.5 and a scale parameter varying from 1-1.3. Income normalized according to sample scale. Line is median estimate, bands represent the central 90% of estimates from random sampling from the distribution.

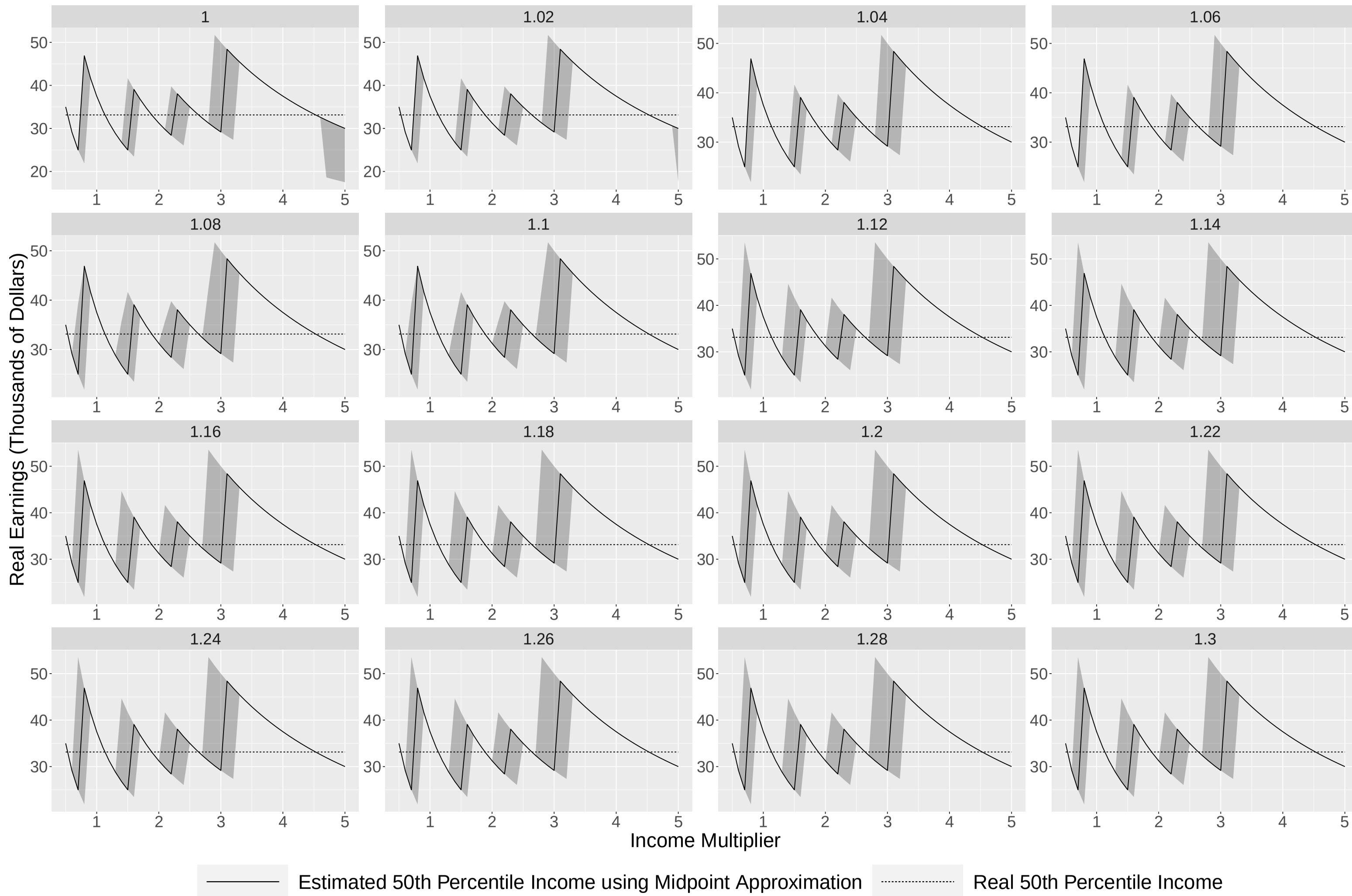


Figure 1.10: Performance of "Midpoint Approximation" for Estimation of 50^{th} Percentile of Income
Real income level for the 50^{th} percentile estimated using the "Midpoint Approximation" on binned data created by random sampling from a population with a lognormal distribution of incomes with a location parameter of 3.5 and a scale parameter varying from 1-1.3. Income normalized according to sample scale. Line is median estimate, bands represent the central 90% of estimates from random sampling from the distribution.

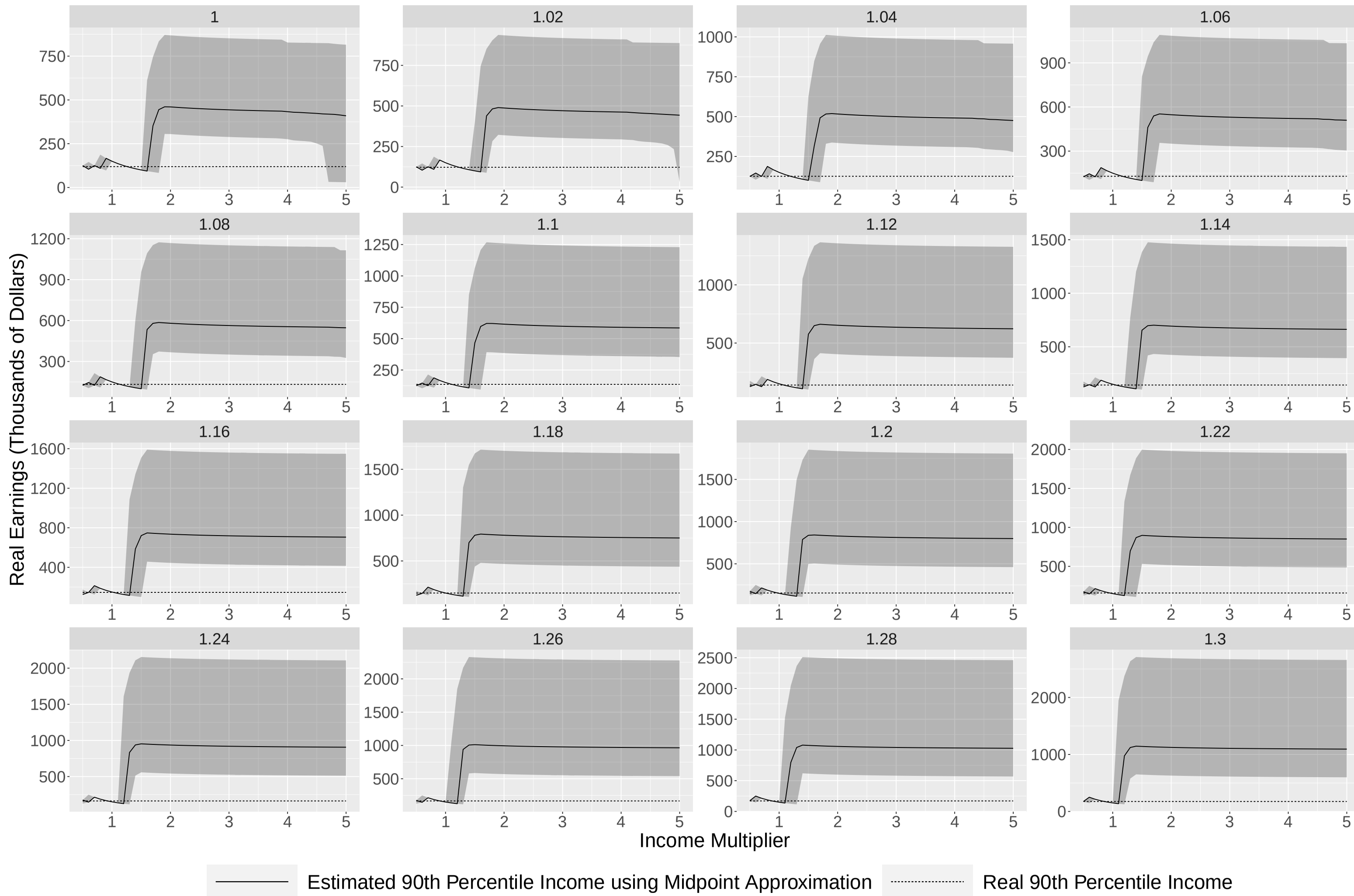


Figure 1.11: Performance of “Midpoint Approximation” for Estimation of 90$^{\text{th}}$ Percentile of Income
Real income level for the 90$^{\text{th}}$ percentile estimated using the “Midpoint Approximation” on binned data created by random sampling from a population with a lognormal distribution of incomes with a location parameter of 3.5 and a scale parameter varying from 1-1.3. Income normalized according to sample scale. Line is median estimate, bands represent the central 90% of estimates from random sampling from the distribution.

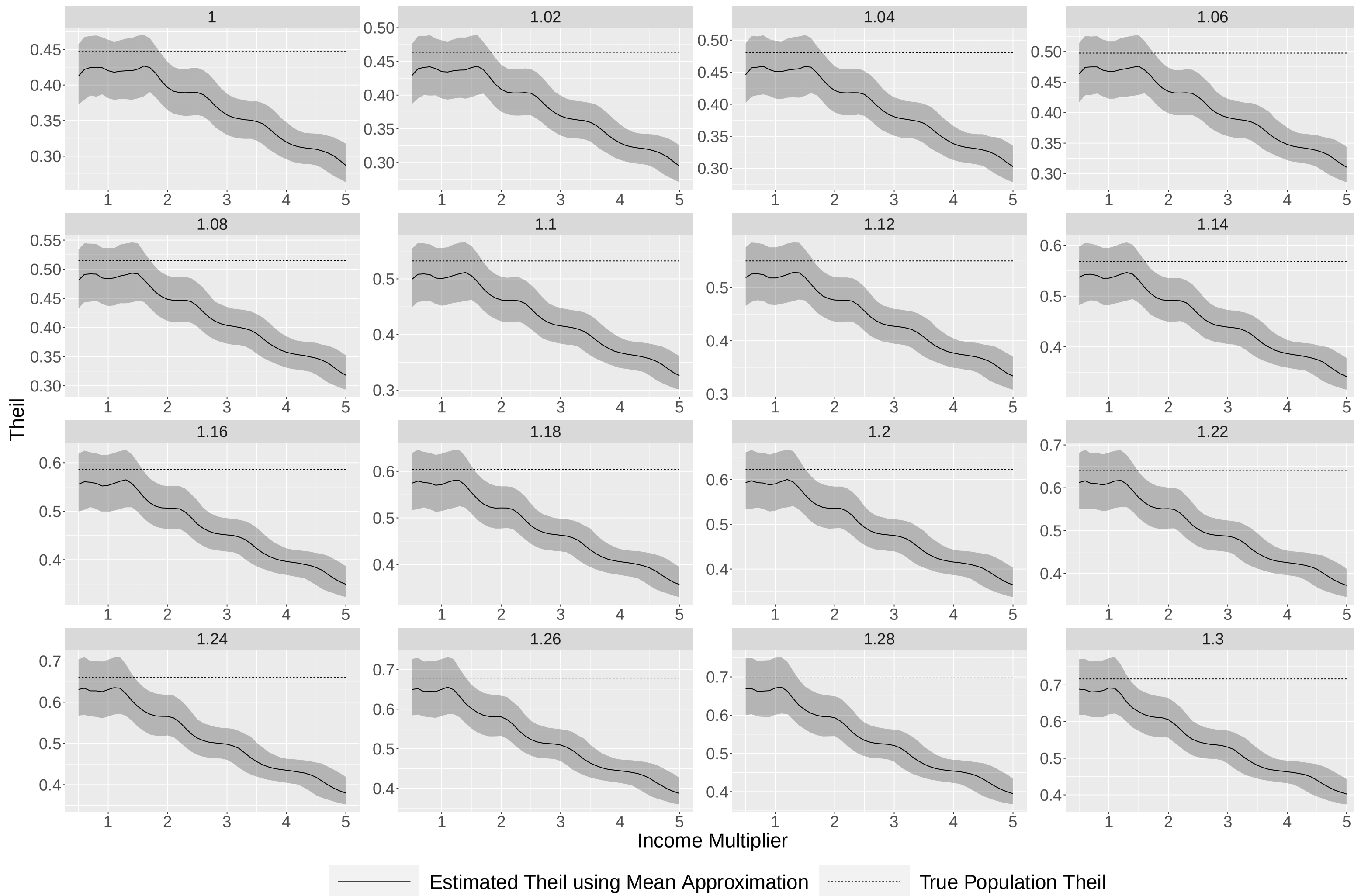


Figure 1.12: Performance of "Mean Approximation" for Estimation of the the Theil Index
Theil estimated using the "Mean Approximation" on binned data created by random sampling from a population with a lognormal distribution of incomes with a location parameter of 3.5 and a scale parameter varying from 1-1.3. Samples then scaled from one half to five times their original value. Line is median estimate, bands represent the central 90% of estimates from random sampling from the distribution.

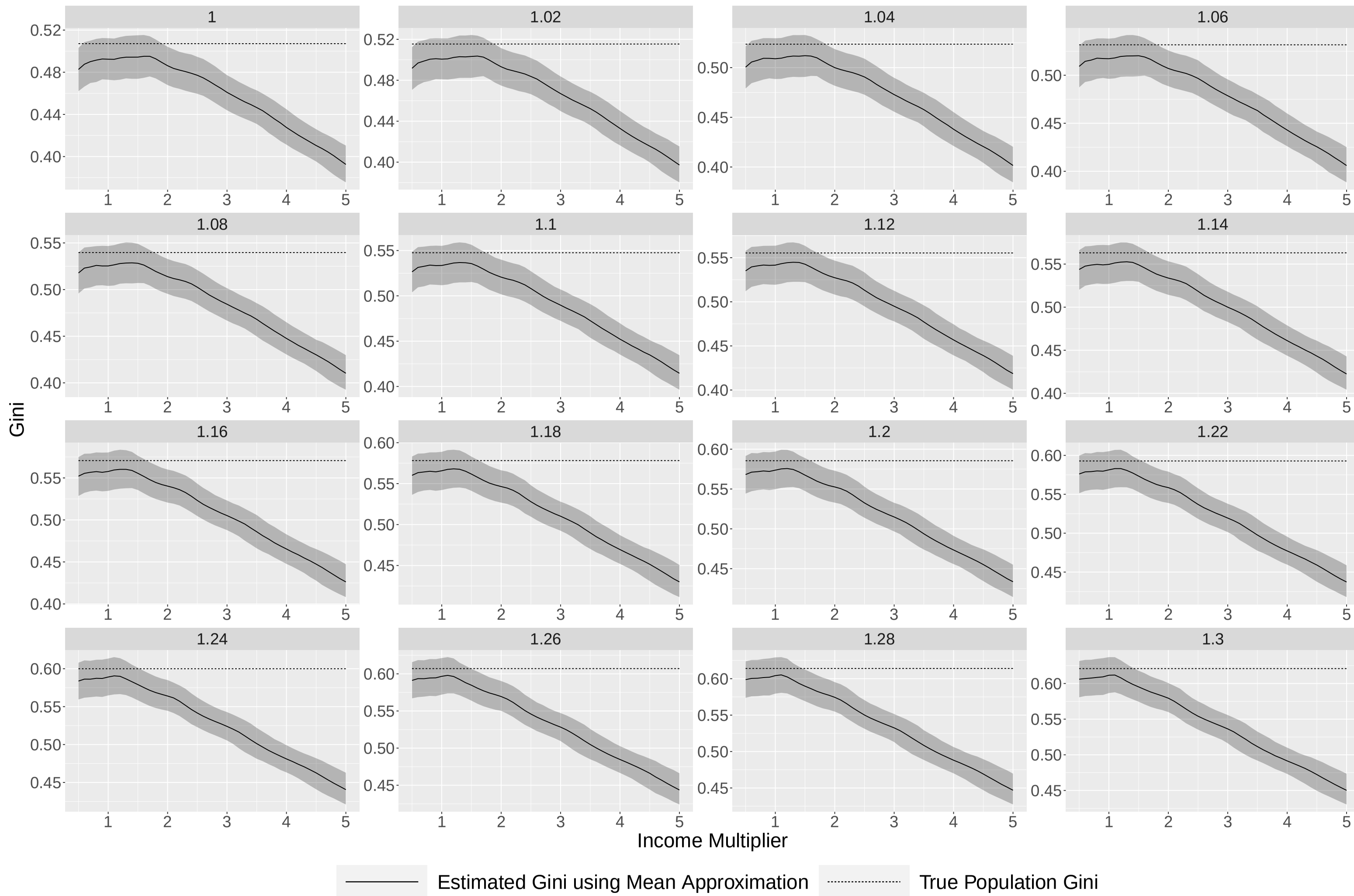


Figure 1.13: Performance of “Mean Approximation” for Estimation of the the Gini Index
Gini estimated using the “Mean Approximation” on binned data created by random sampling from a population with a lognormal distribution of incomes with a location parameter of 3.5 and a scale parameter varying from 1-1.3. Samples then scaled from one half to five times their original value. Line is median estimate, bands represent the central 90% of estimates from random sampling from the distribution.

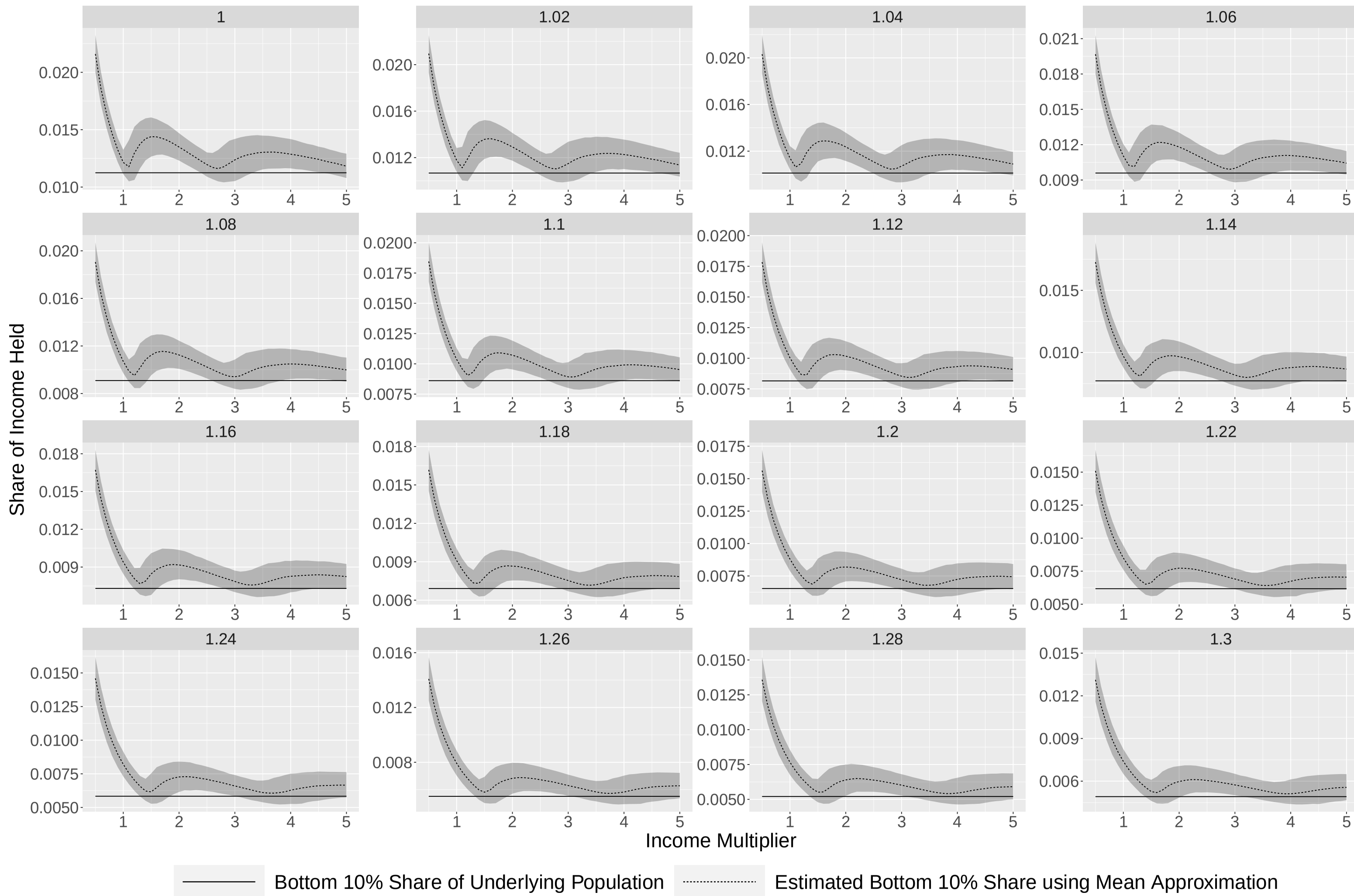


Figure 1.14: Performance of “Mean Approximation” for Estimation of Bottom Decile Population Share of Income
Bottom decile share of income estimated using the “Mean Approximation” on binned data created by random sampling from a population with a lognormal distribution of incomes with a location parameter of 3.5 and a scale parameter varying from 1-1.3. Samples then scaled from one half to five times their original value. Line is median estimate, bands represent the central 90% of estimates from random sampling from the distribution.

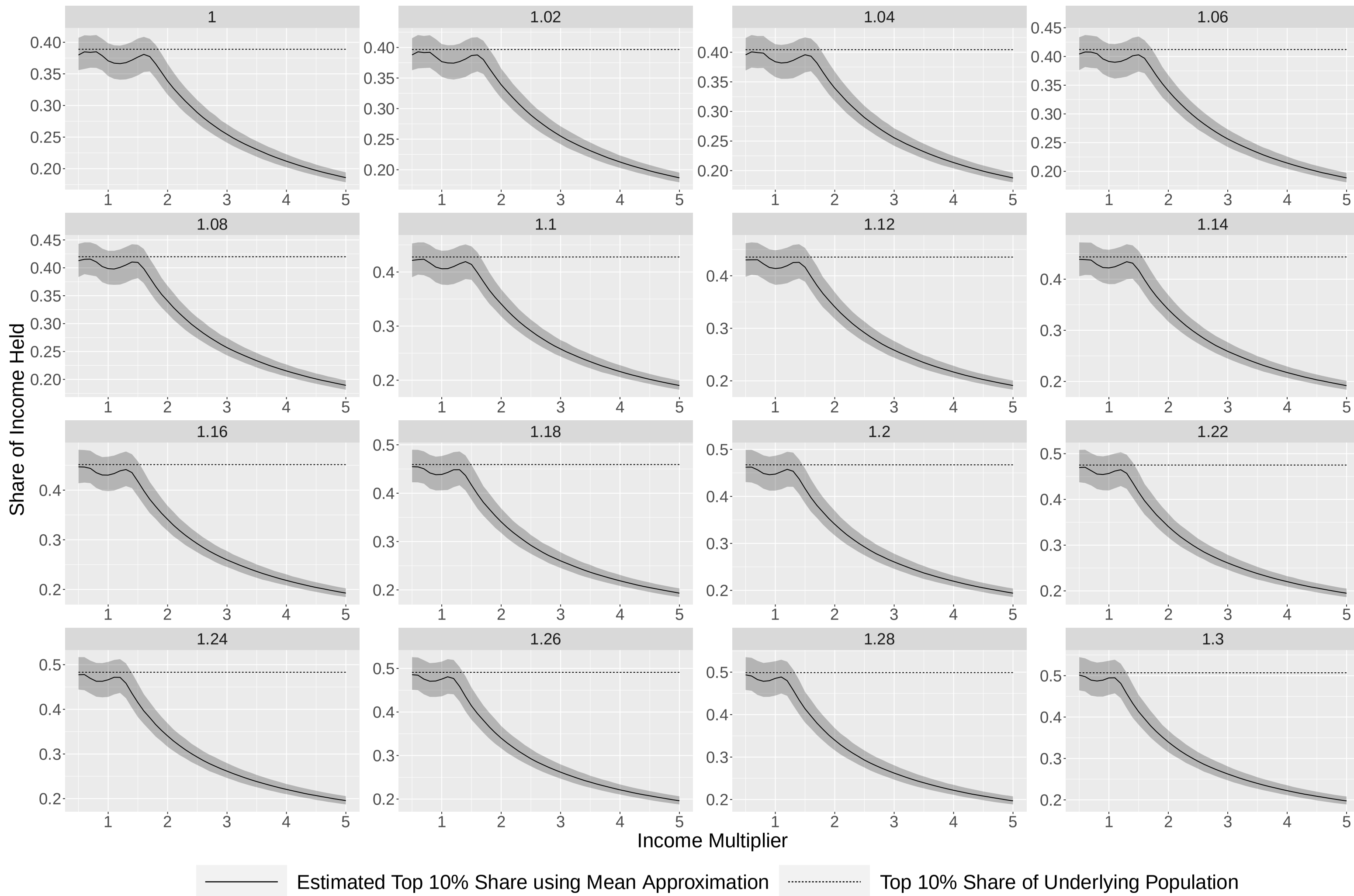


Figure 1.15: Performance of "Mean Approximation" for Estimation of Top Decile Population Share of Income
Top decile share of income estimated using the "MEan Approximation" on binned data created by random sampling from a population with a lognormal distribution of incomes with a location parameter of 3.5 and a scale parameter varying from 1-1.3. Samples then scaled from one half to five times their original value. Line is median estimate, bands represent the central 90% of estimates from random sampling from the distribution.

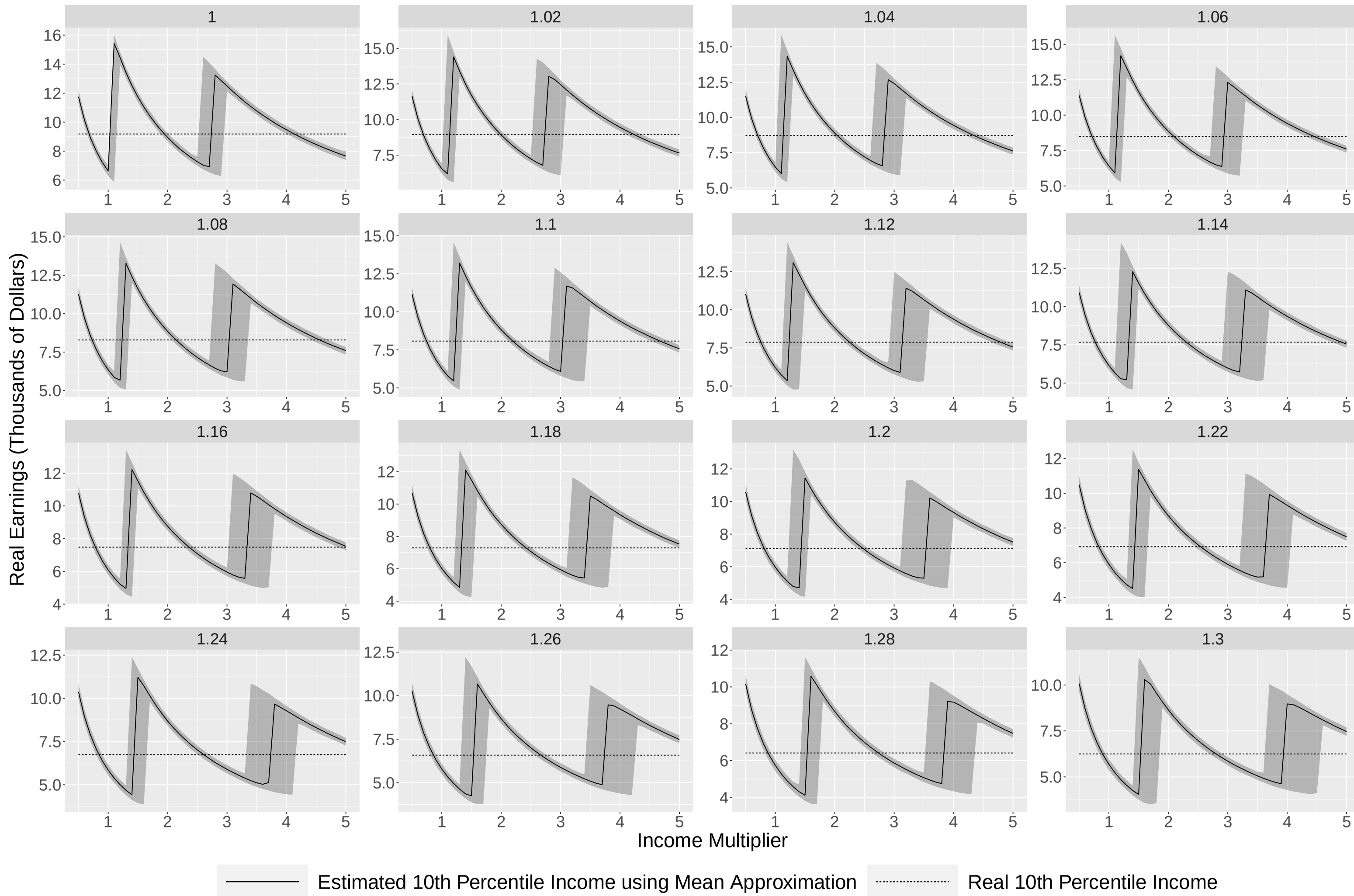


Figure 1.16: Performance of “Mean Approximation” for Estimation of 10^{th} Percentile of Income

Real income level for the 10^{th} percentile estimated using the “Mean Approximation” on binned data created by random sampling from a population with a lognormal distribution of incomes with a location parameter of 3.5 and a scale parameter varying from 1-1.3. Income normalized according to sample scale. Line is median estimate, bands represent the central 90% of estimates from random sampling from the distribution.

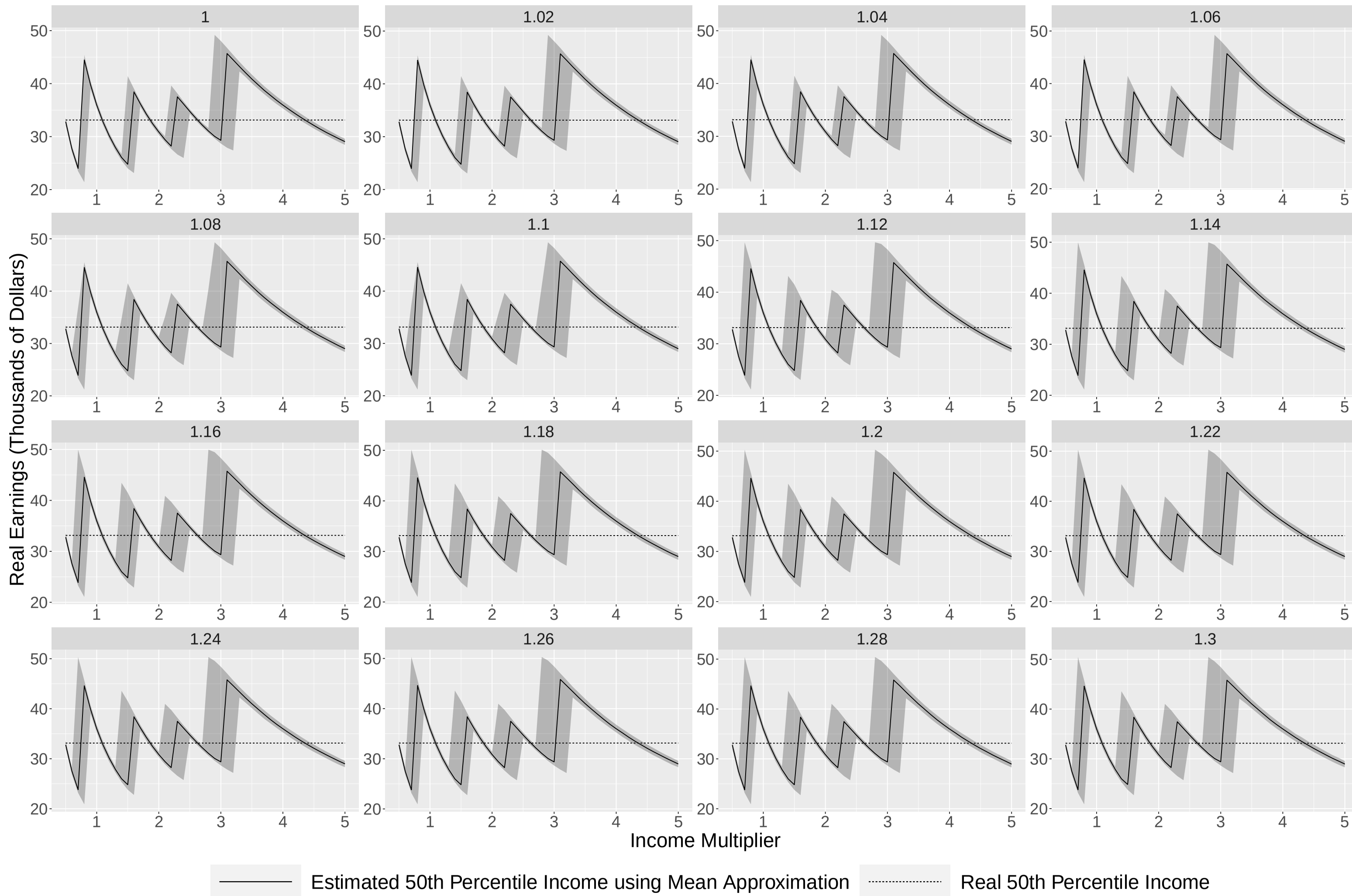
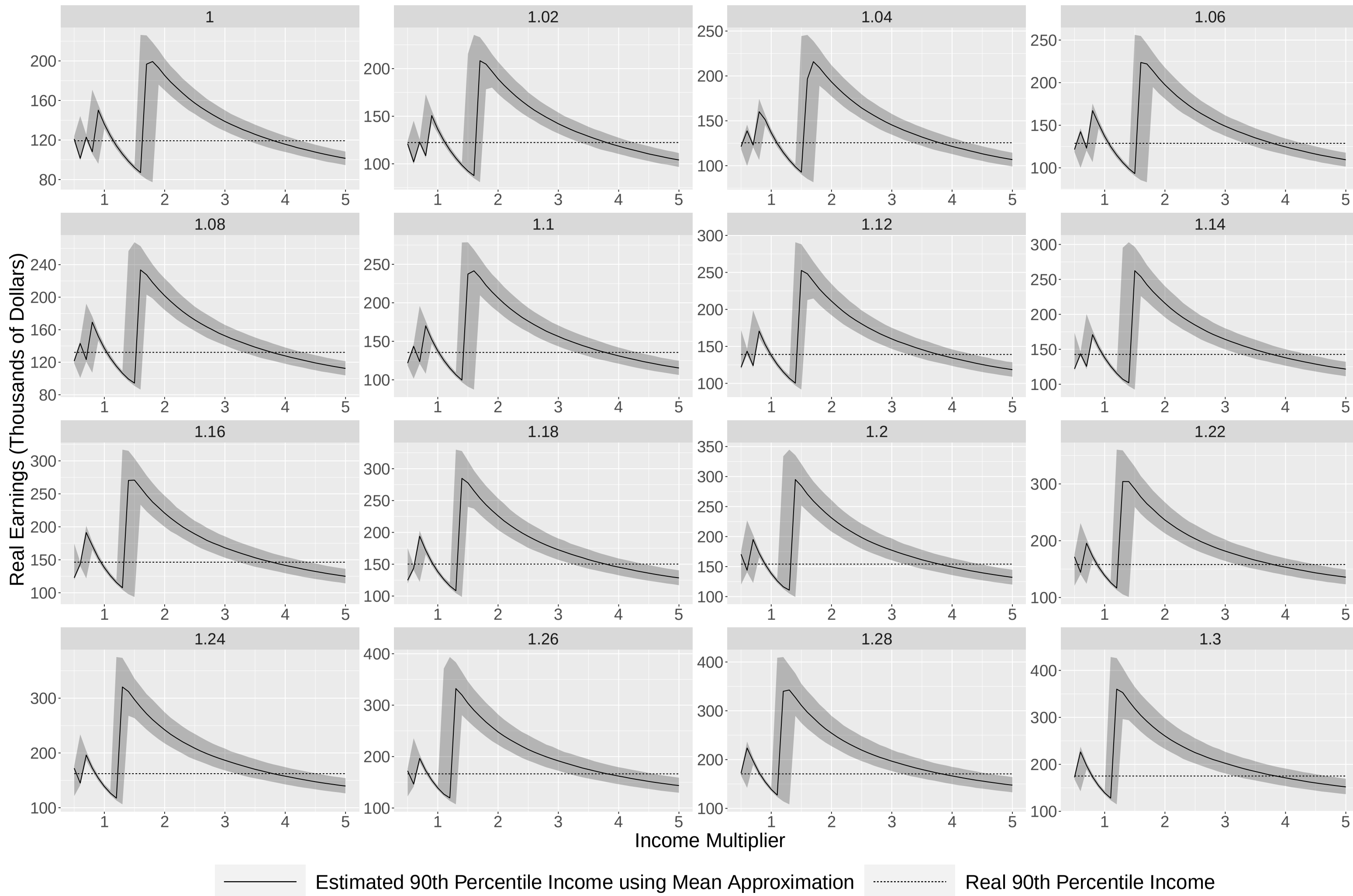


Figure 1.17: Performance of “Mean Approximation” for Estimation of 50^{th} Percentile of Income
Real income level for the 50^{th} percentile estimated using the “Mean Approximation” on binned data created by random sampling from a population with a lognormal distribution of incomes with a location parameter of 3.5 and a scale parameter varying from 1-1.3. Income normalized according to sample scale. Line is median estimate, bands represent the central 90% of estimates from random sampling from the distribution.



Performance of "Mean Approximation" for Estimation of 90^{th} Percentile of Income

Real income level for the 90^{th} percentile estimated using the "Mean Approximation" on binned data created by random sampling from a population with a lognormal distribution of incomes with a location parameter of 3.5 and a scale parameter varying from 1-1.3. Income normalized according to sample scale. Line is median estimate, bands represent the central 90% of estimates from random sampling from the distribution.

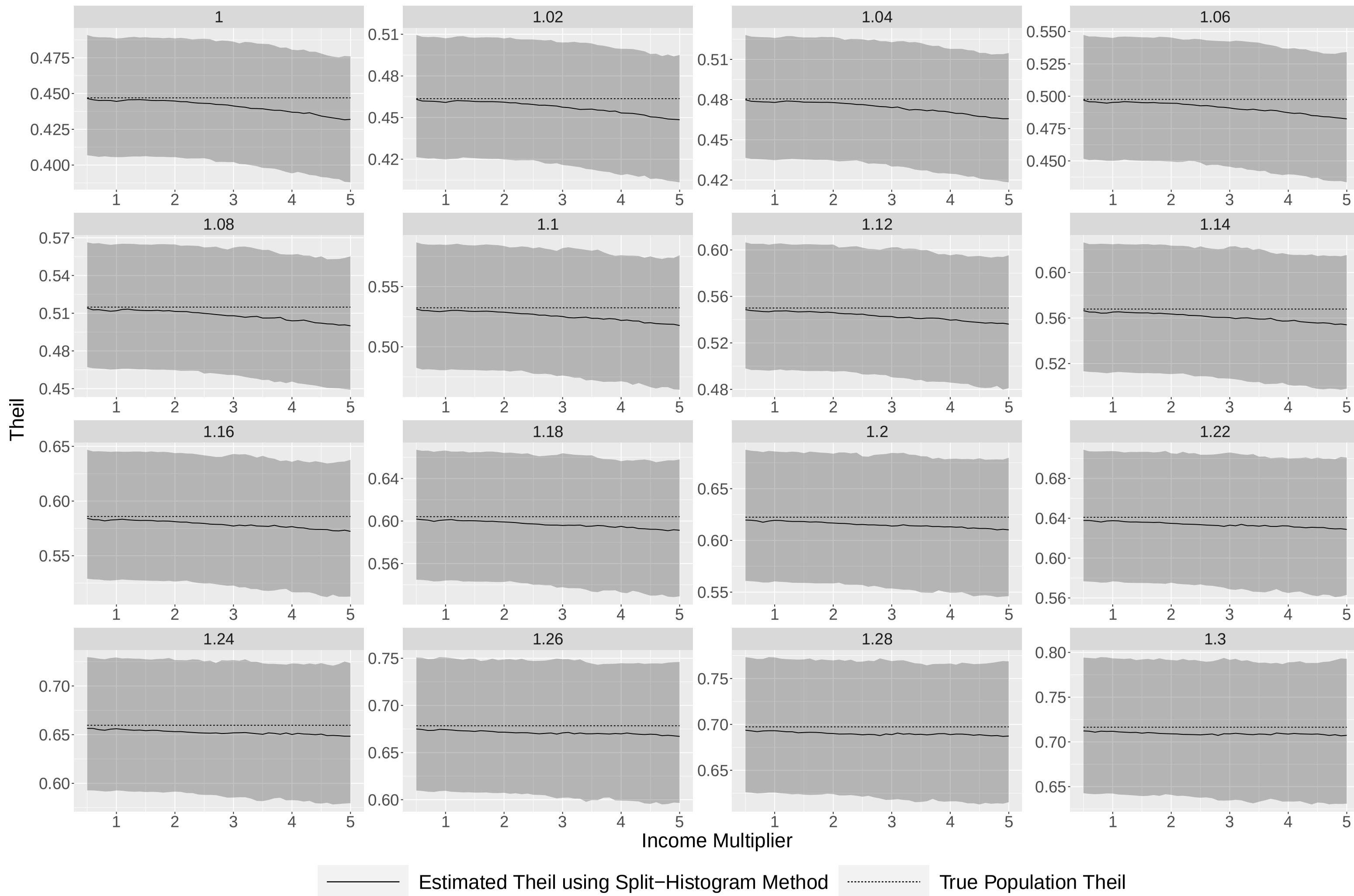


Figure 1.19: Performance of "Split-Histogram" for Estimation of the Theil Index
Theil estimated using the "Split-Histogram" method on binned data created by random sampling from a population with a lognormal distribution of incomes with a location parameter of 3.5 and a scale parameter varying from 1-1.3. Samples then scaled from one half to five times their original value. Line is median estimate, bands represent the central 90% of estimates from random sampling from the distribution.

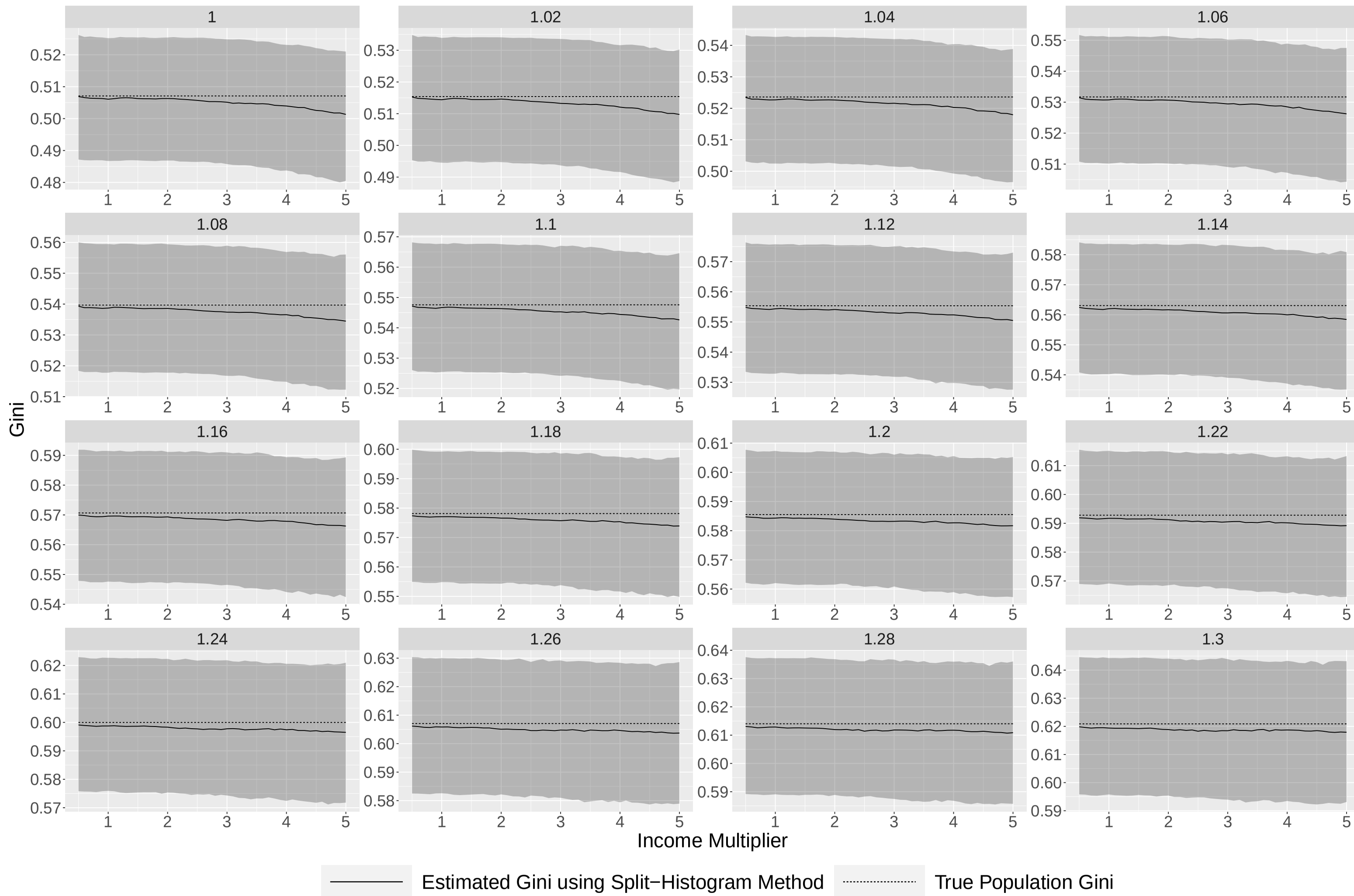


Figure 1.20: Performance of "Split-Histogram" for Estimation of the Gini Index
Gini estimated using the "Split-Histogram" method on binned data created by random sampling from a population with a lognormal distribution of incomes with a location parameter of 3.5 and a scale parameter varying from 1-1.3. Samples then scaled from one half to five times their original value. Line is median estimate, bands represent the central 90% of estimates from random sampling from the distribution.

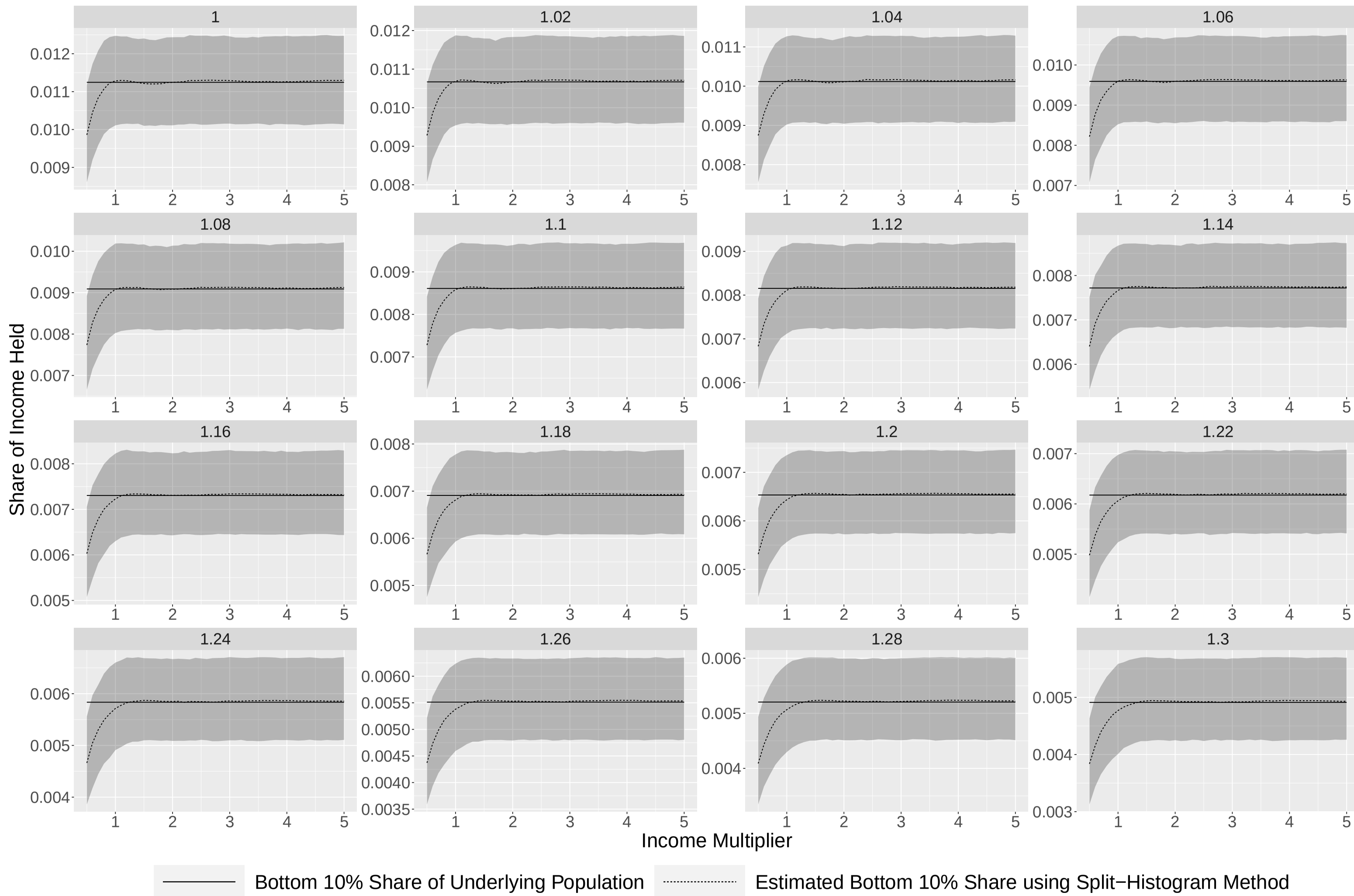


Figure 1.21: Performance of "Split-Histogram" for Estimation of Bottom Decile Population Share of Income
Bottom decile share of income estimated using the "Split-Histogram" method on binned data created by random sampling from a population with a lognormal distribution of incomes with a location parameter of 3.5 and a scale parameter varying from 1-1.3. Samples then scaled from one half to five times their original value. Line is median estimate, bands represent the central 90% of estimates from random sampling from the distribution.

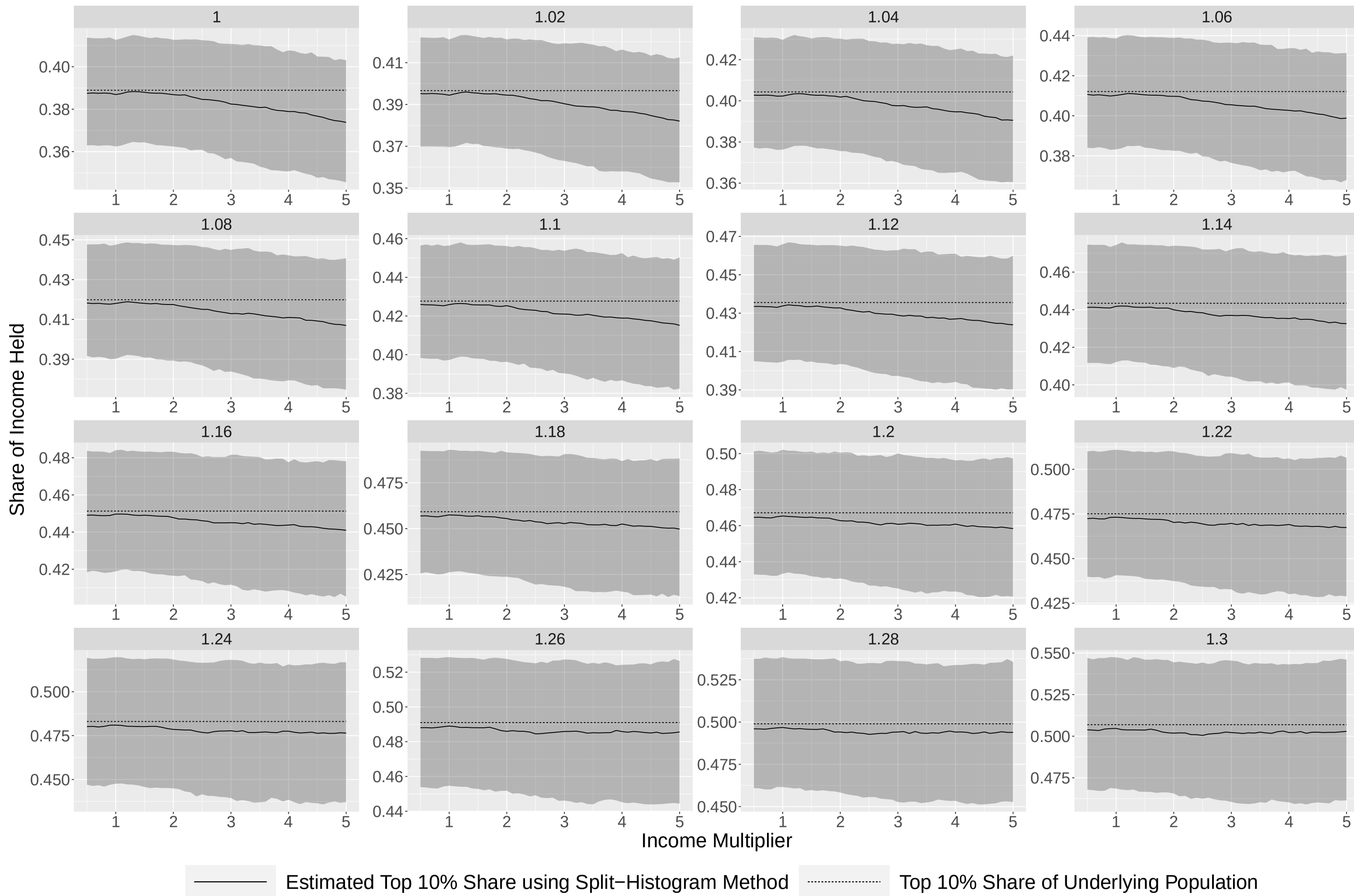


Figure 1.22
Top decile share of income estimated using the “Split-Histogram” method on binned data created by random sampling from a population with a lognormal distribution of incomes with a location parameter of 3.5 and a scale parameter varying from 1-1.3. Samples then scaled from one half to five times their original value. Line is median estimate, bands represent the central 90% of estimates from random sampling from the distribution.

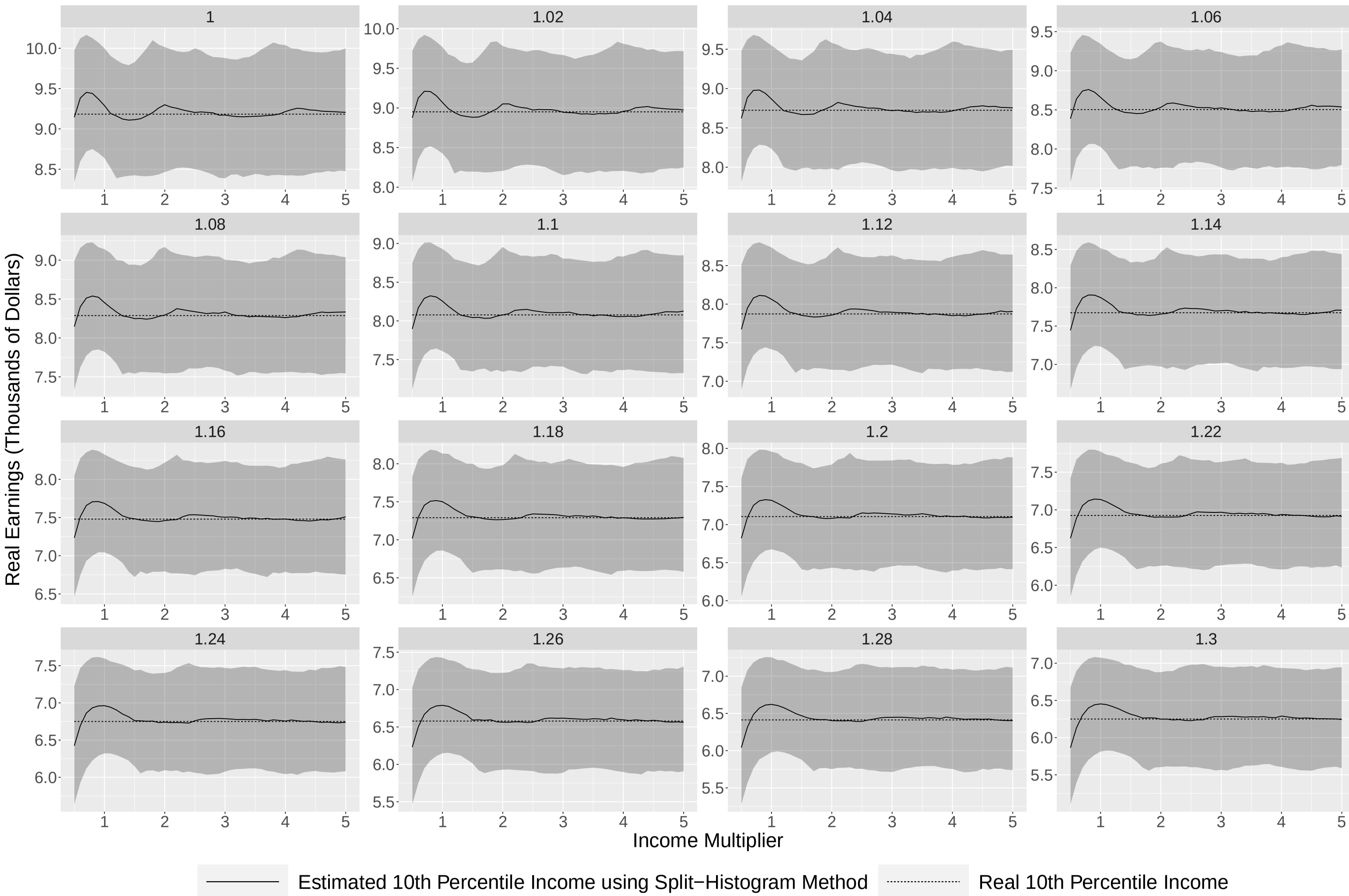


Figure 1.23: Performance of “Split-Histogram” for Estimation of 10^{th} Percentile of Income

Real income level for the 10^{th} percentile estimated using the “Split-Histogram” method on binned data created by random sampling from a population with a lognormal distribution of incomes with a location parameter of 3.5 and a scale parameter varying from 1-1.3. Income normalized according to sample scale. Line is median estimate, bands represent the central 90% of estimates from random sampling from the distribution.

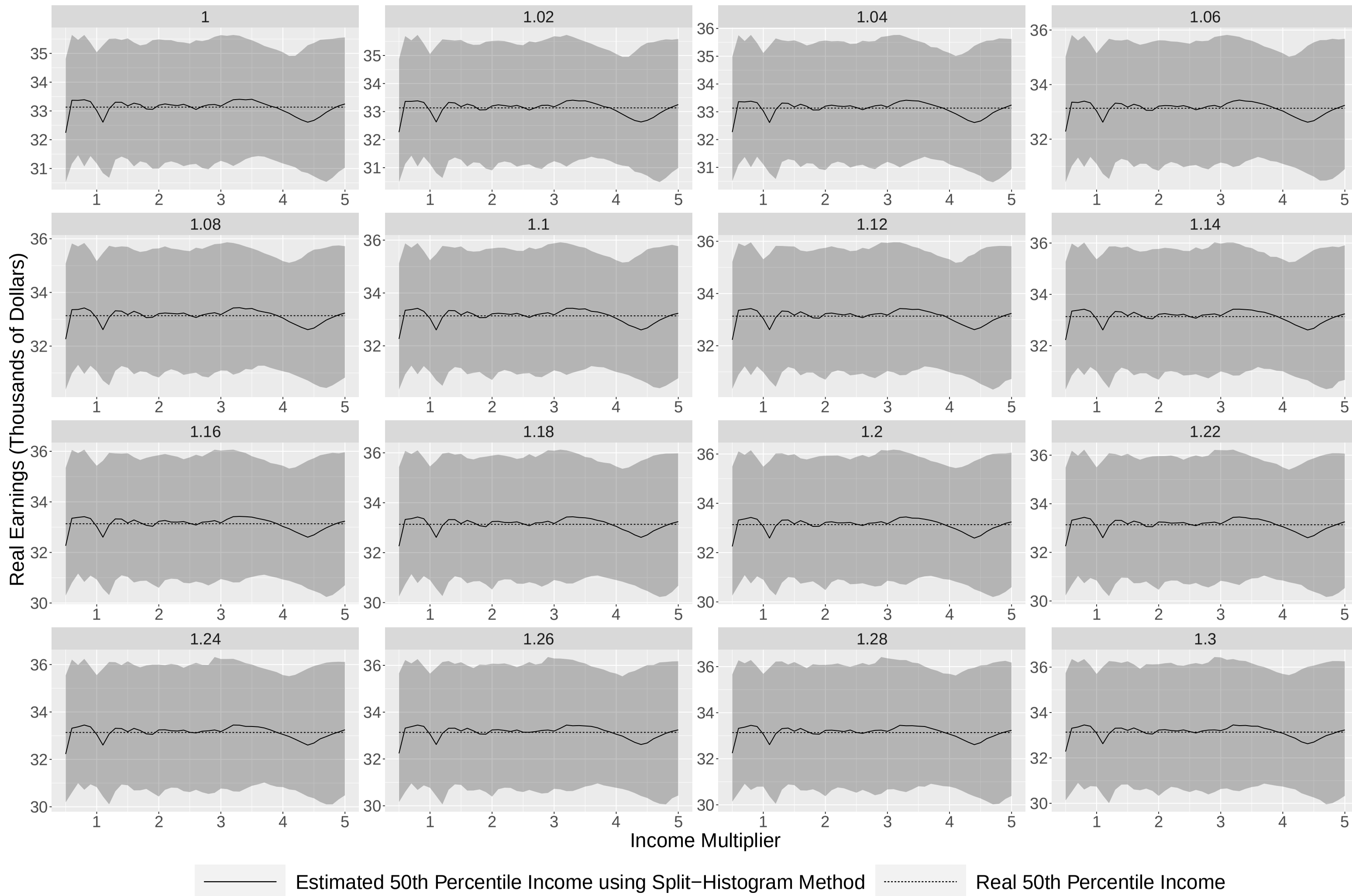


Figure 1.24: Performance of “Split-Histogram” for Estimation of 50$^{\text{th}}$ Percentile of Income

Real income level for the 50$^{\text{th}}$ percentile estimated using the “Split-Histogram” method on binned data created by random sampling from a population with a lognormal distribution of incomes with a location parameter of 3.5 and a scale parameter varying from 1-1.3. Income normalized according to sample scale. Line is median estimate, bands represent the central 90% of estimates from random sampling from the distribution.

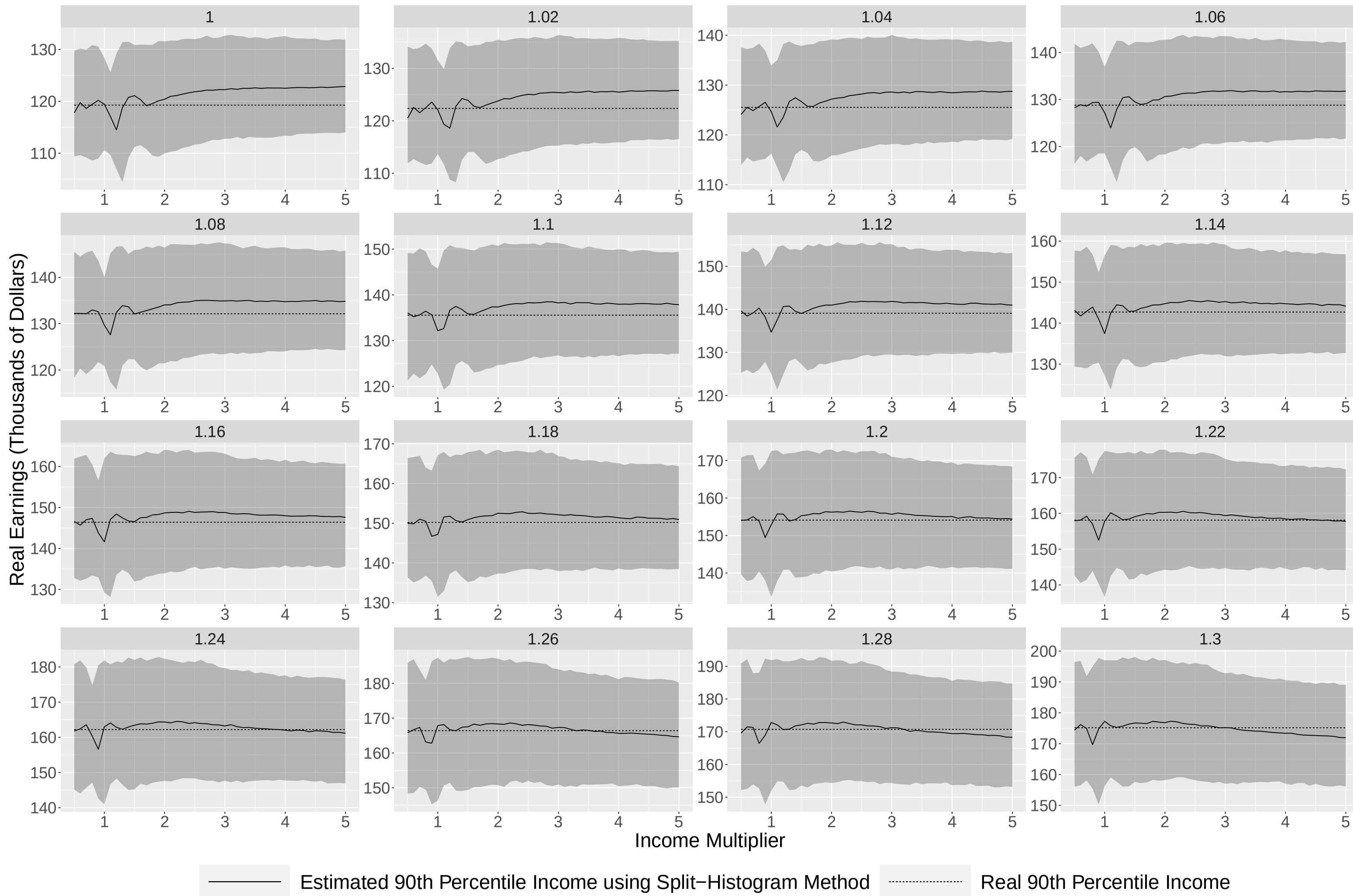


Figure 1.25: Performance of “Split-Histogram” for Estimation of 90$^{\text{th}}$ Percentile of Income

Real income level for the 90$^{\text{th}}$ percentile estimated using the “Split-Histogram” method on binned data created by random sampling from a population with a lognormal distribution of incomes with a location parameter of 3.5 and a scale parameter varying from 1-1.3. Income normalized according to sample scale. Line is median estimate, bands represent the central 90% of estimates from random sampling from the distribution.

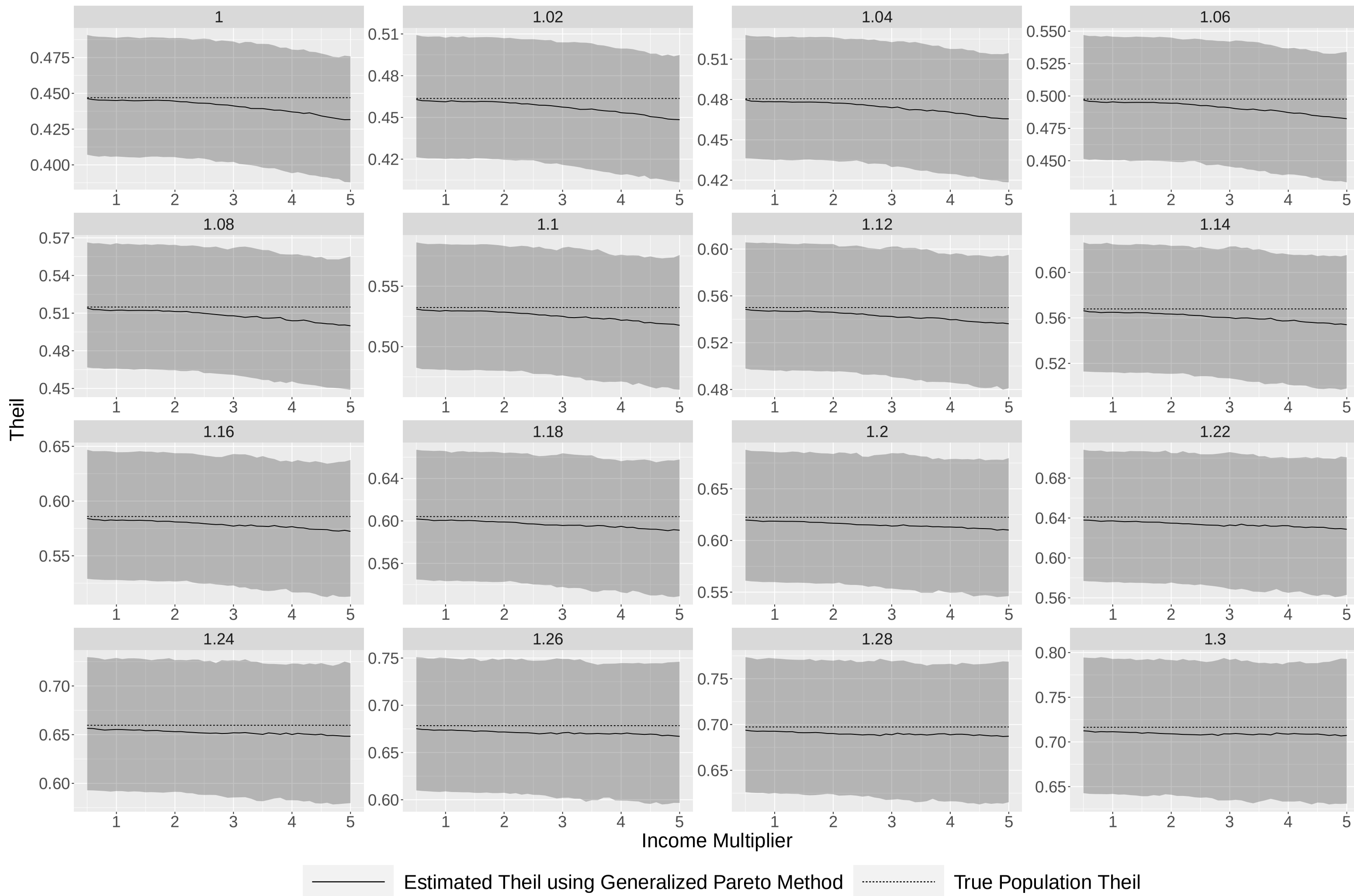


Figure 1.26: Performance of "Generalized Pareto" for Estimation of the Theil Index
Theil estimated using the "Generalized Pareto" method on binned data created by random sampling from a population with a lognormal distribution of incomes with a location parameter of 3.5 and a scale parameter varying from 1-1.3. Samples then scaled from one half to five times their original value. Line is median estimate, bands represent the central 90% of estimates from random sampling from the distribution.

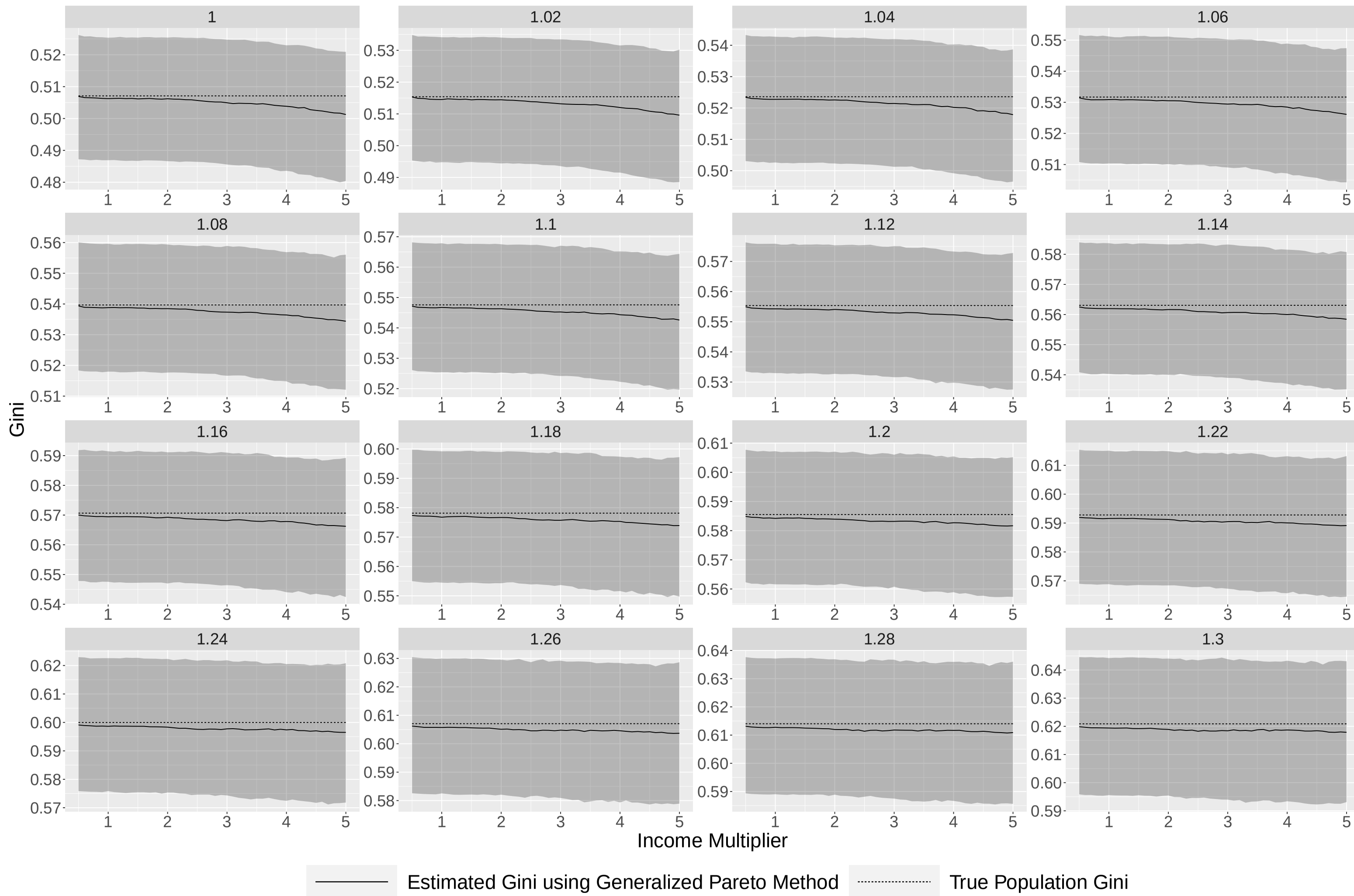


Figure 1.27: Performance of “Generalized Pareto” for Estimation of the Gini Index
Gini estimated using the “Generalized Pareto” method on binned data created by random sampling from a population with a lognormal distribution of incomes with a location parameter of 3.5 and a scale parameter varying from 1-1.3. Samples then scaled from one half to five times their original value. Line is median estimate, bands represent the central 90% of estimates from random sampling from the distribution.

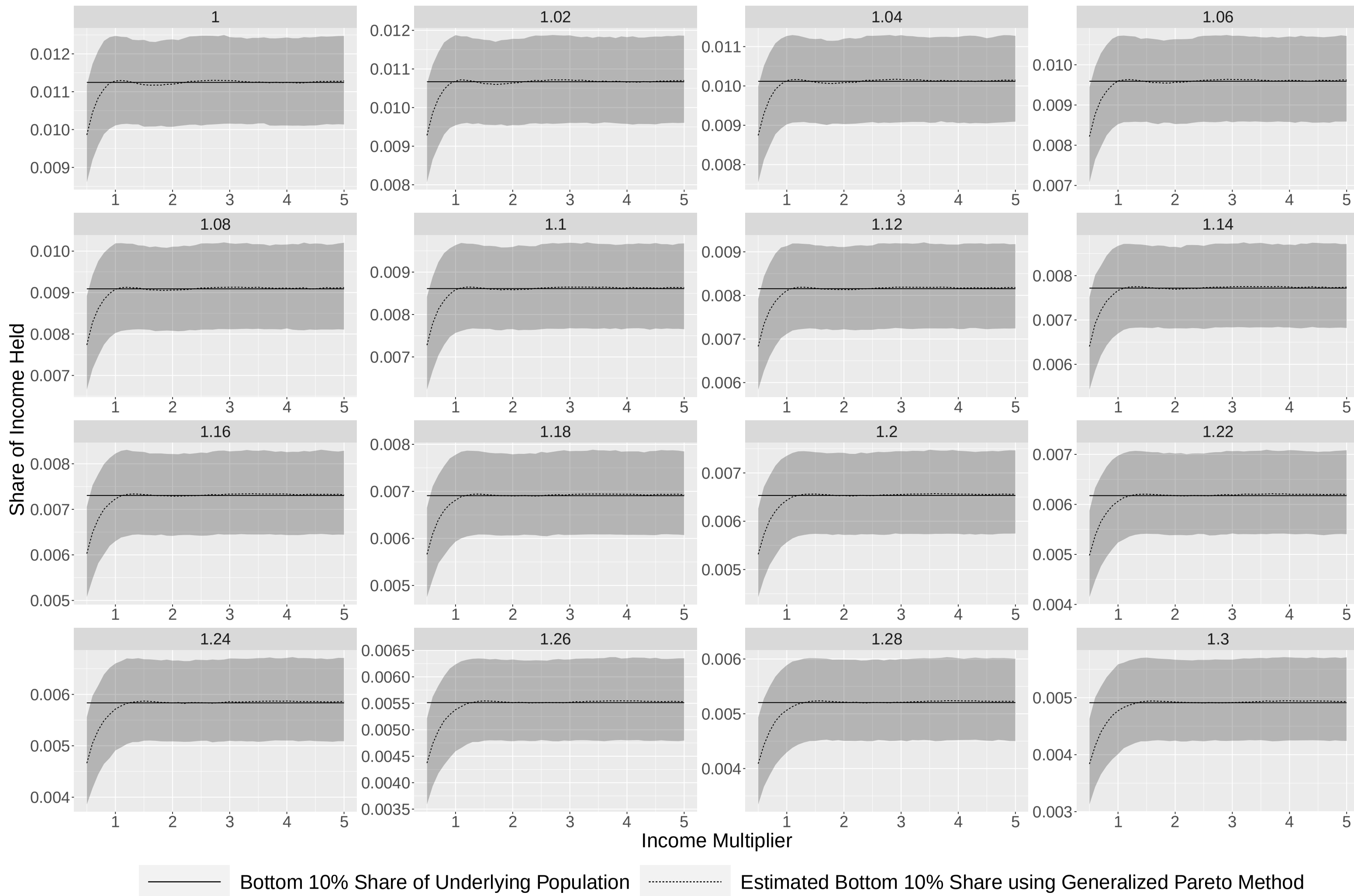


Figure 1.28: Performance of "Generalized Pareto" for Estimation of Bottom Decile Population Share of Income
Bottom decile share of income estimated using the "Generalized Pareto" method on binned data created by random sampling from a population with a lognormal distribution of incomes with a location parameter of 3.5 and a scale parameter varying from 1-1.3. Samples then scaled from one half to five times their original value. Line is median estimate, bands represent the central 90% of estimates from random sampling from the distribution.

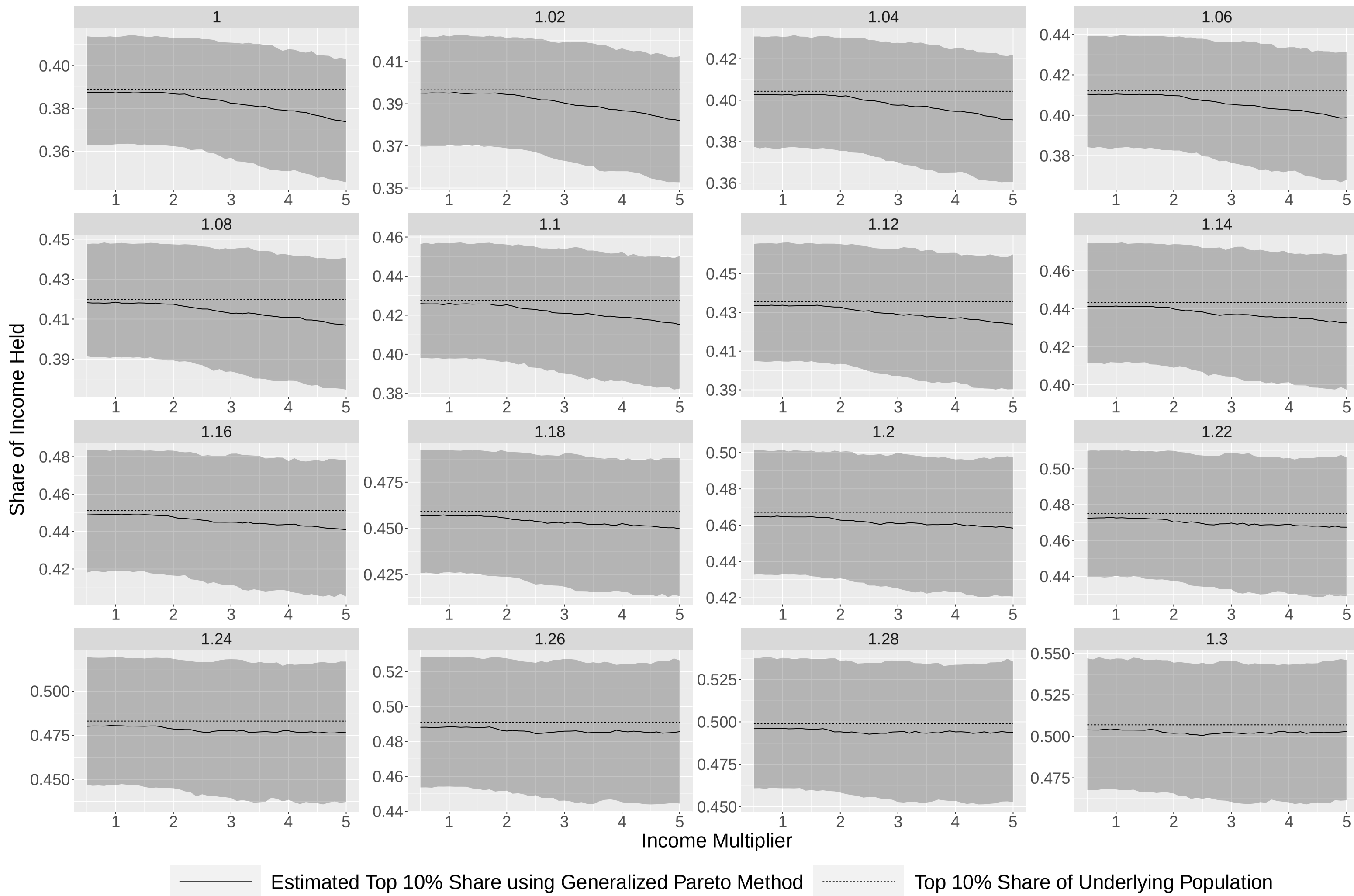


Figure 1.29: Performance of "Generalized Pareto" for Estimation of Top Decile Population Share of Income
Top decile share of income estimated using the "Generalized Pareto" method on binned data created by random sampling from a population with a lognormal distribution of incomes with a location parameter of 3.5 and a scale parameter varying from 1-1.3. Samples then scaled from one half to five times their original value. Line is median estimate, bands represent the central 90% of estimates from random sampling from the distribution.

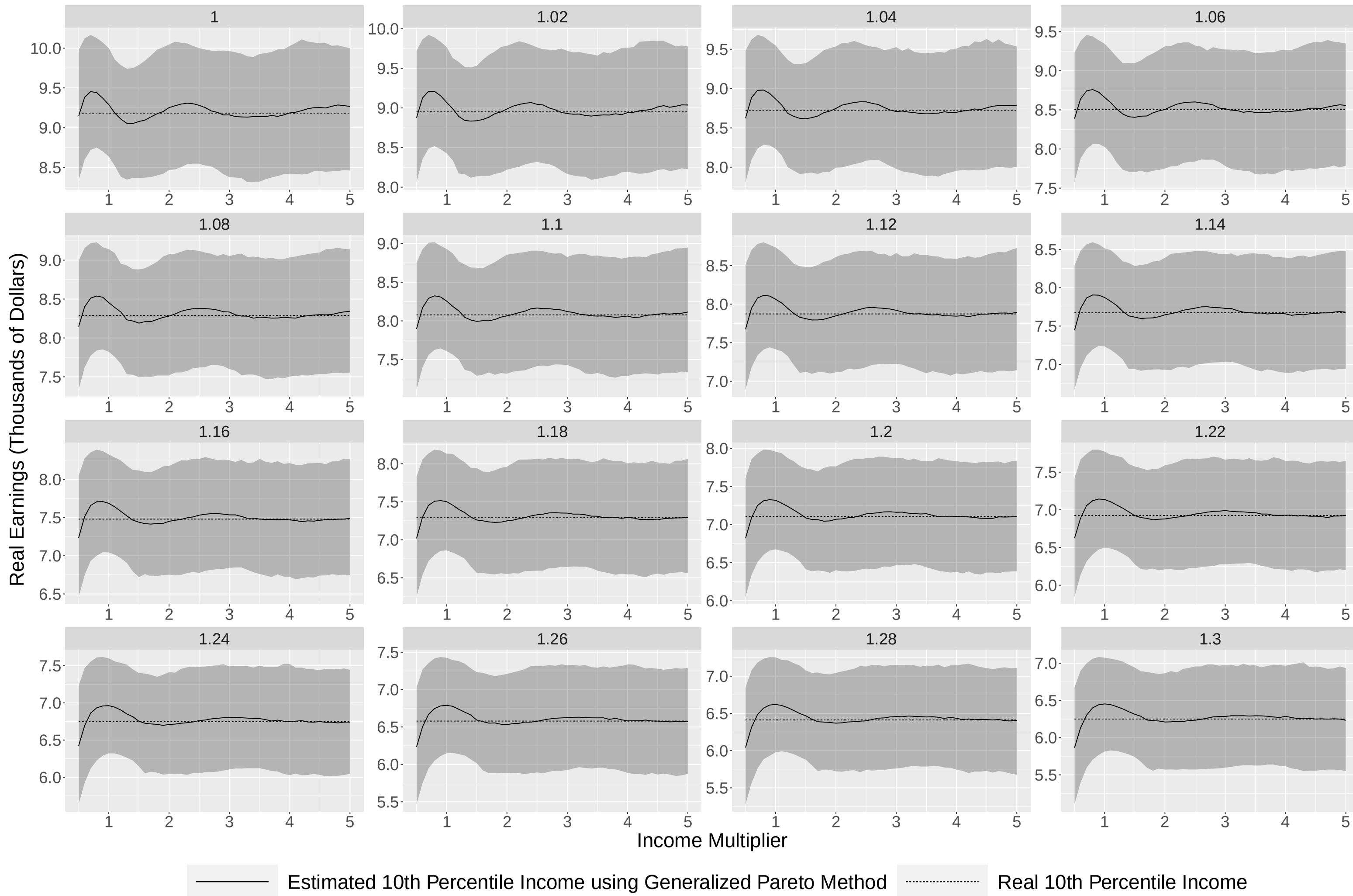


Figure 1.30: Performance of "Generalized Pareto" for Estimation of 10^{th} Percentile of Income

Real income level for the 10^{th} percentile estimated using the "Generalized Pareto" method on binned data created by random sampling from a population with a lognormal distribution of incomes with a location parameter of 3.5 and a scale parameter varying from 1-1.3. Income normalized according to sample scale. Line is median estimate, bands represent the central 90% of estimates from random sampling from the distribution.

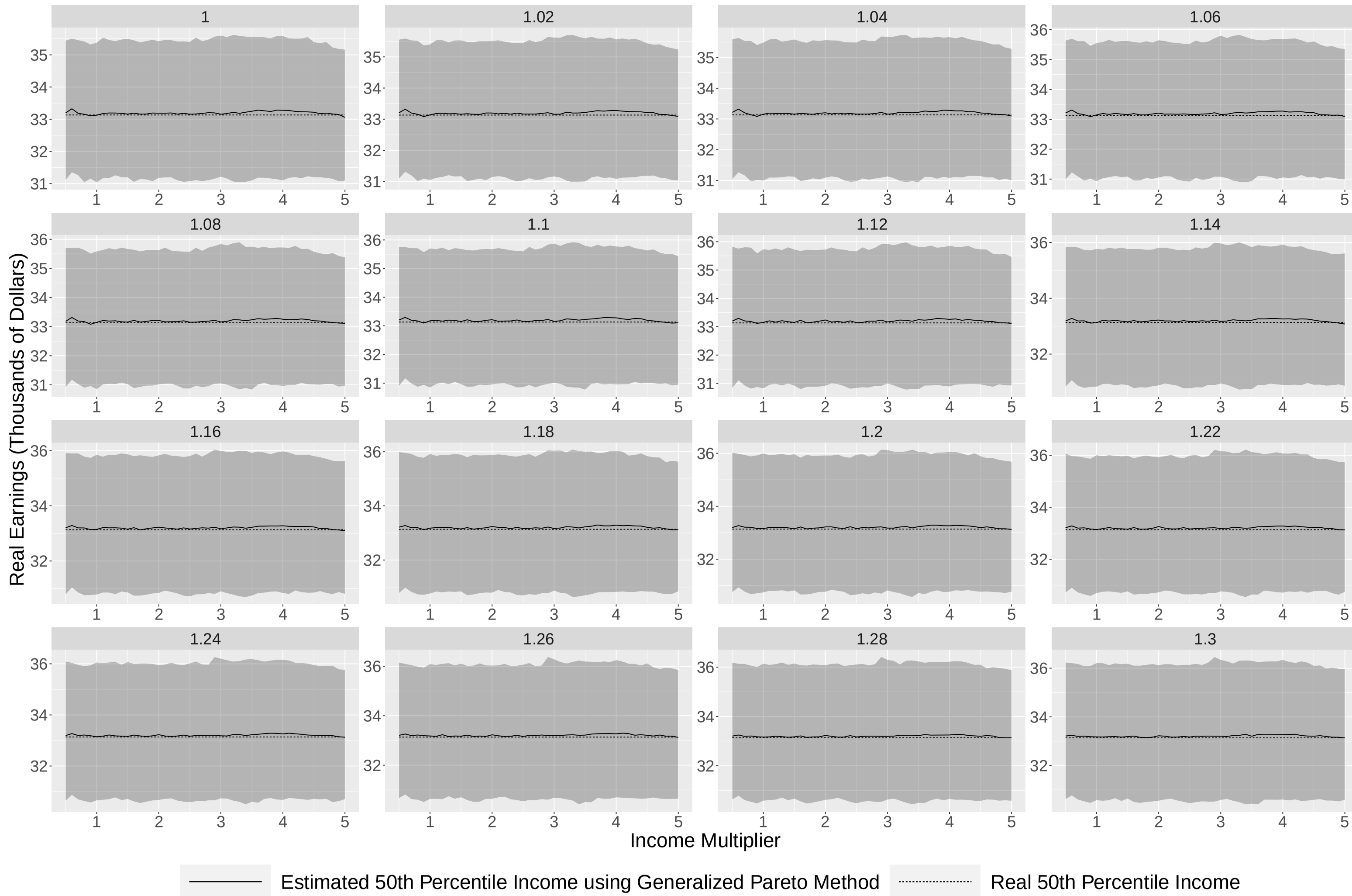


Figure 1.31: Performance of "Generalized Pareto" for Estimation of 50^{th} Percentile of Income
Real income level for the 50^{th} percentile estimated using the "Generalized Pareto" method on binned data created by random sampling from a population with a lognormal distribution of incomes with a location parameter of 3.5 and a scale parameter varying from 1-1.3. Income normalized according to sample scale. Line is median estimate, bands represent the central 90% of estimates from random sampling from the distribution.

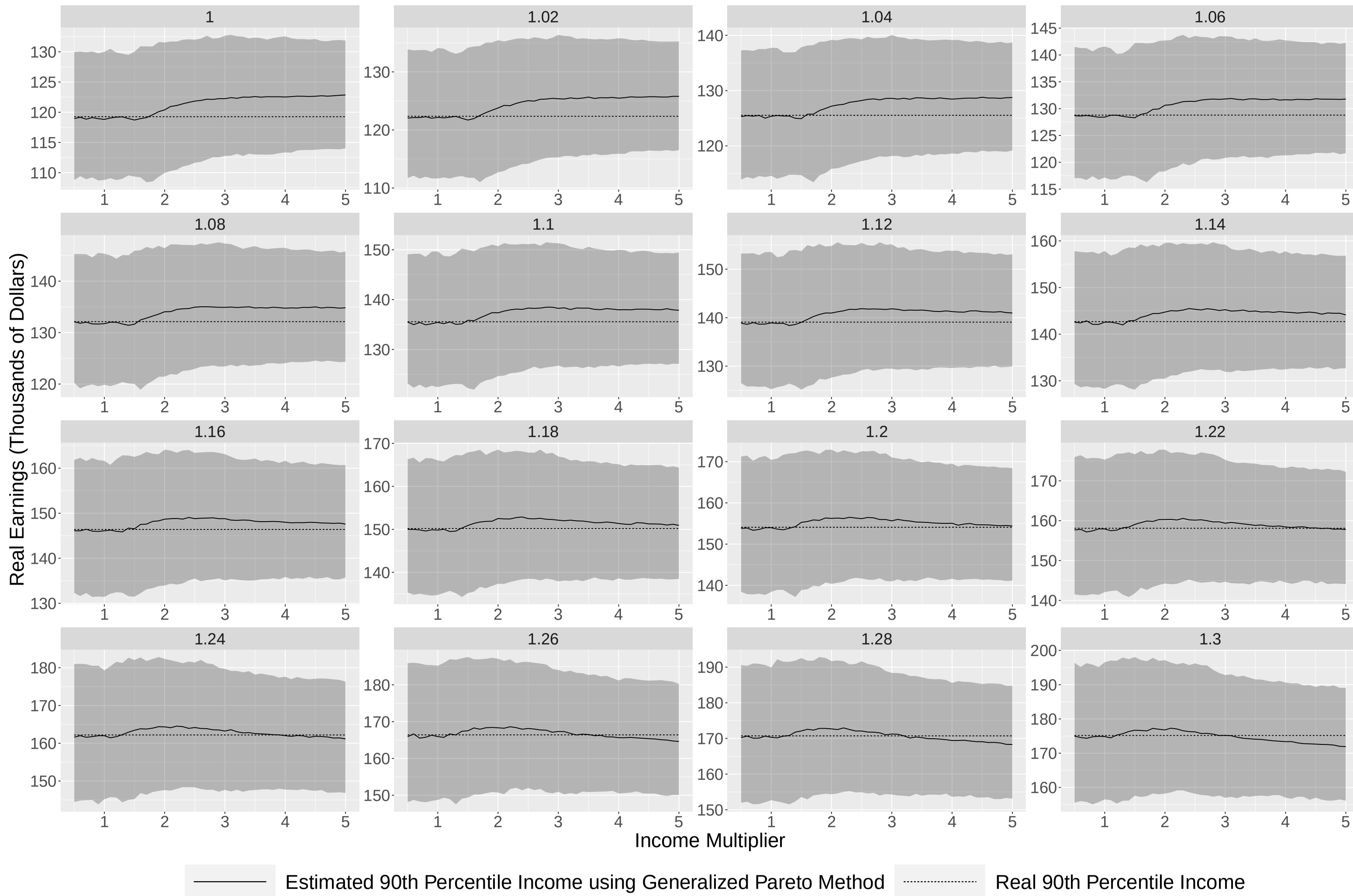


Figure 1.32: Performance of “Generalized Pareto” for Estimation of 90^{th} Percentile of Income
Real income level for the 90^{th} percentile estimated using the “Generalized Pareto” method on binned data created by random sampling from a population with a lognormal distribution of incomes with a location parameter of 3.5 and a scale parameter varying from 1-1.3. Income normalized according to sample scale. Line is median estimate, bands represent the central 90% of estimates from random sampling from the distribution.

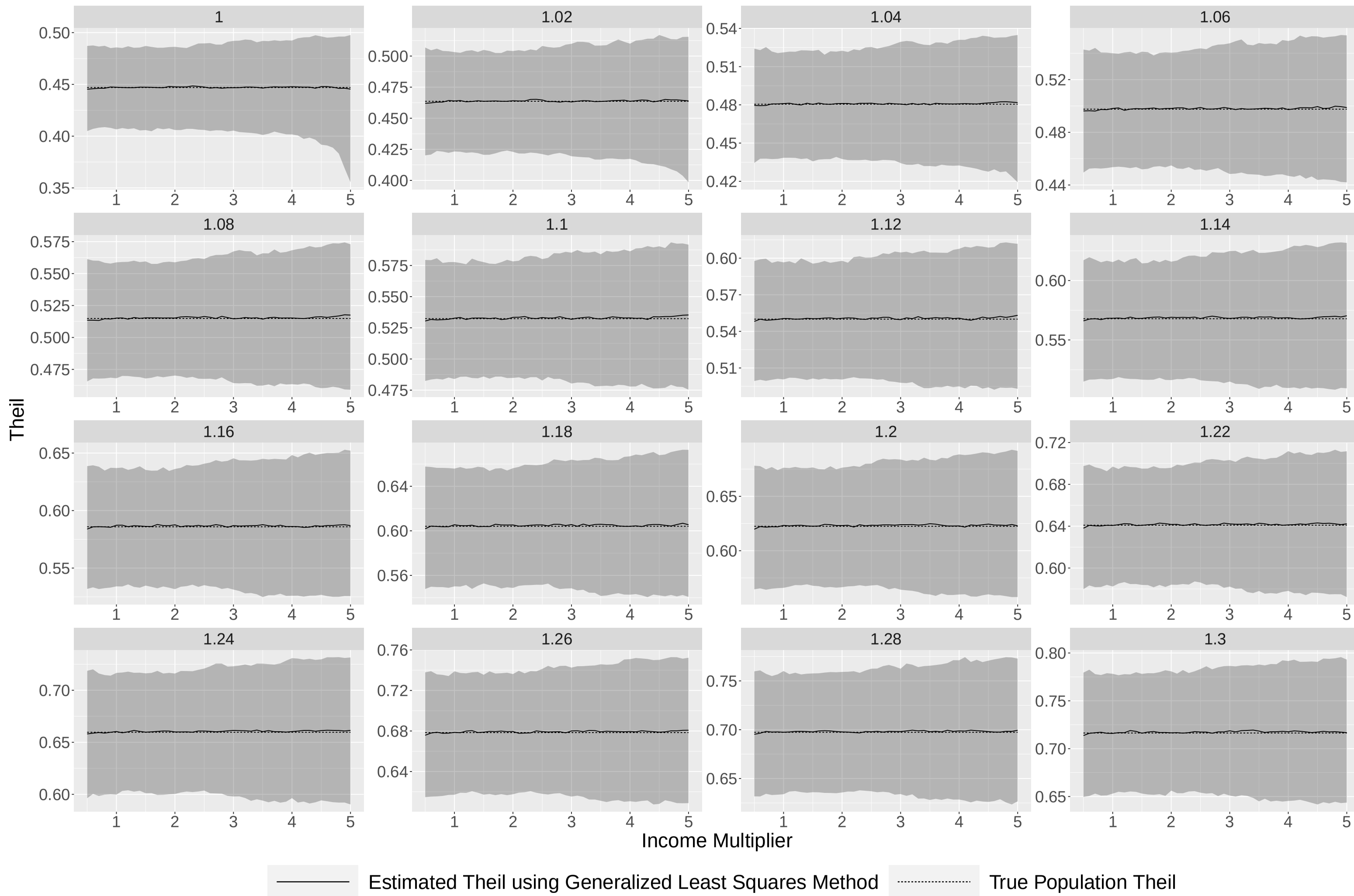


Figure 1.33: Performance of General Least Squares for Estimation of the Theil Index

Theil estimated using the Generalized Least Squares method on binned data created by random sampling from a population with a lognormal distribution of incomes with a location parameter of 3.5 and a scale parameter varying from 1-1.3. Samples then scaled from one half to five times their original value. Line is median estimate, bands represent the central 90% of estimates from random sampling from the distribution.

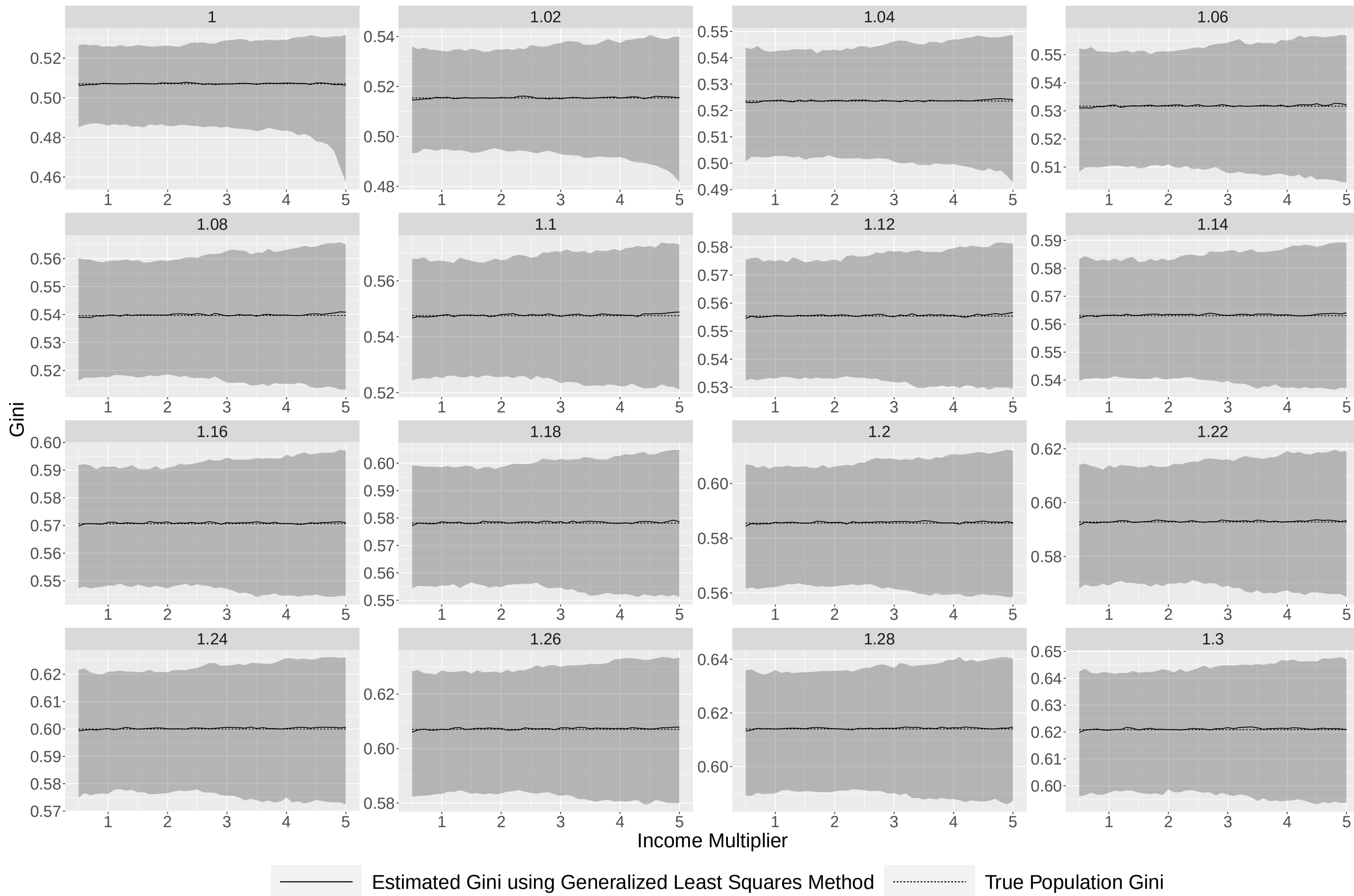


Figure 1.34: Performance of General Least Squares for Estimation of the Gini Index
Gini estimated using the Generalized Least Squares method on binned data created by random sampling from a population with a lognormal distribution of incomes with a location parameter of 3.5 and a scale parameter varying from 1-1.3. Samples then scaled from one half to five times their original value. Line is median estimate, bands represent the central 90% of estimates from random sampling from the distribution.

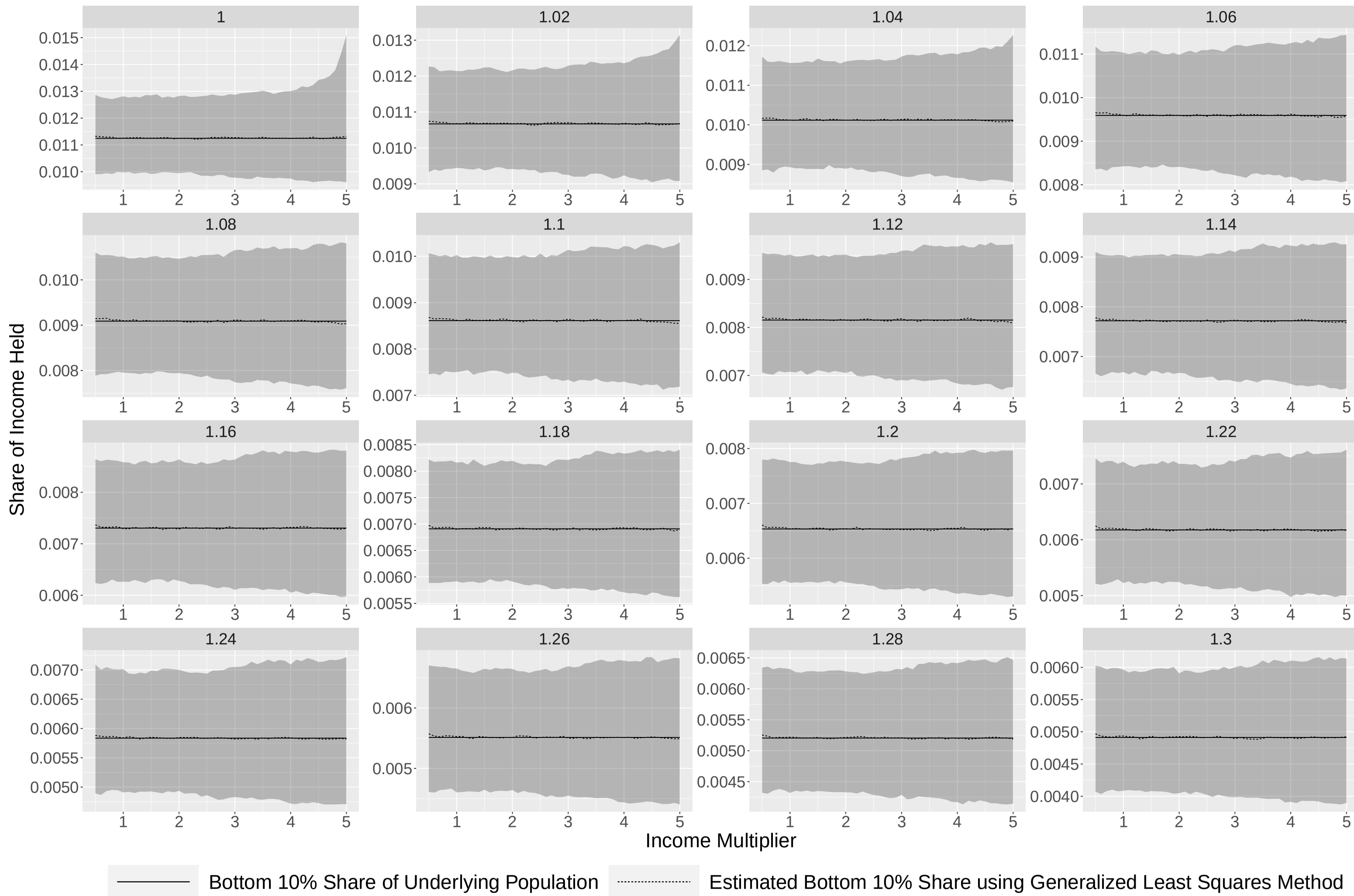


Figure 1.35: Performance of General Least Squares for Estimation of Bottom Decile Population Share of Income
Bottom decile share of income estimated using the Generalized Pareto method on binned data created by random sampling from a population with a lognormal distribution of incomes with a location parameter of 3.5 and a scale parameter varying from 1-1.3. Samples then scaled from one half to five times their original value. Line is median estimate, bands represent the central 90% of estimates from random sampling from the distribution.

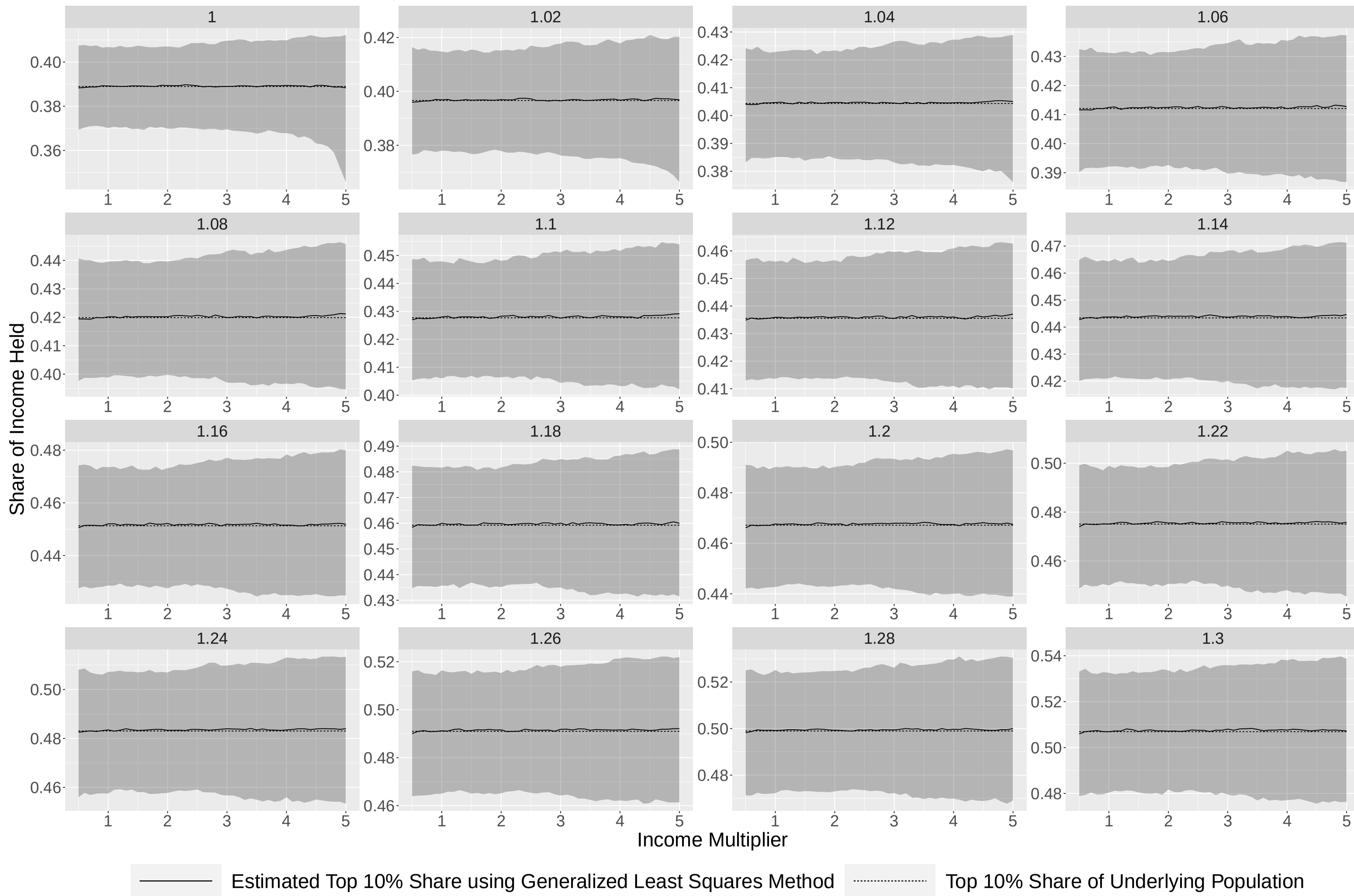


Figure 1.36: Performance of General Least Squares for Estimation of Top Decile Population Share of Income
Top decile share of income estimated using the Generalized Pareto method on binned data created by random sampling from a population with a lognormal distribution of incomes with a location parameter of 3.5 and a scale parameter varying from 1-1.3. Samples then scaled from one half to five times their original value. Line is median estimate, bands represent the central 90% of estimates from random sampling from the distribution.

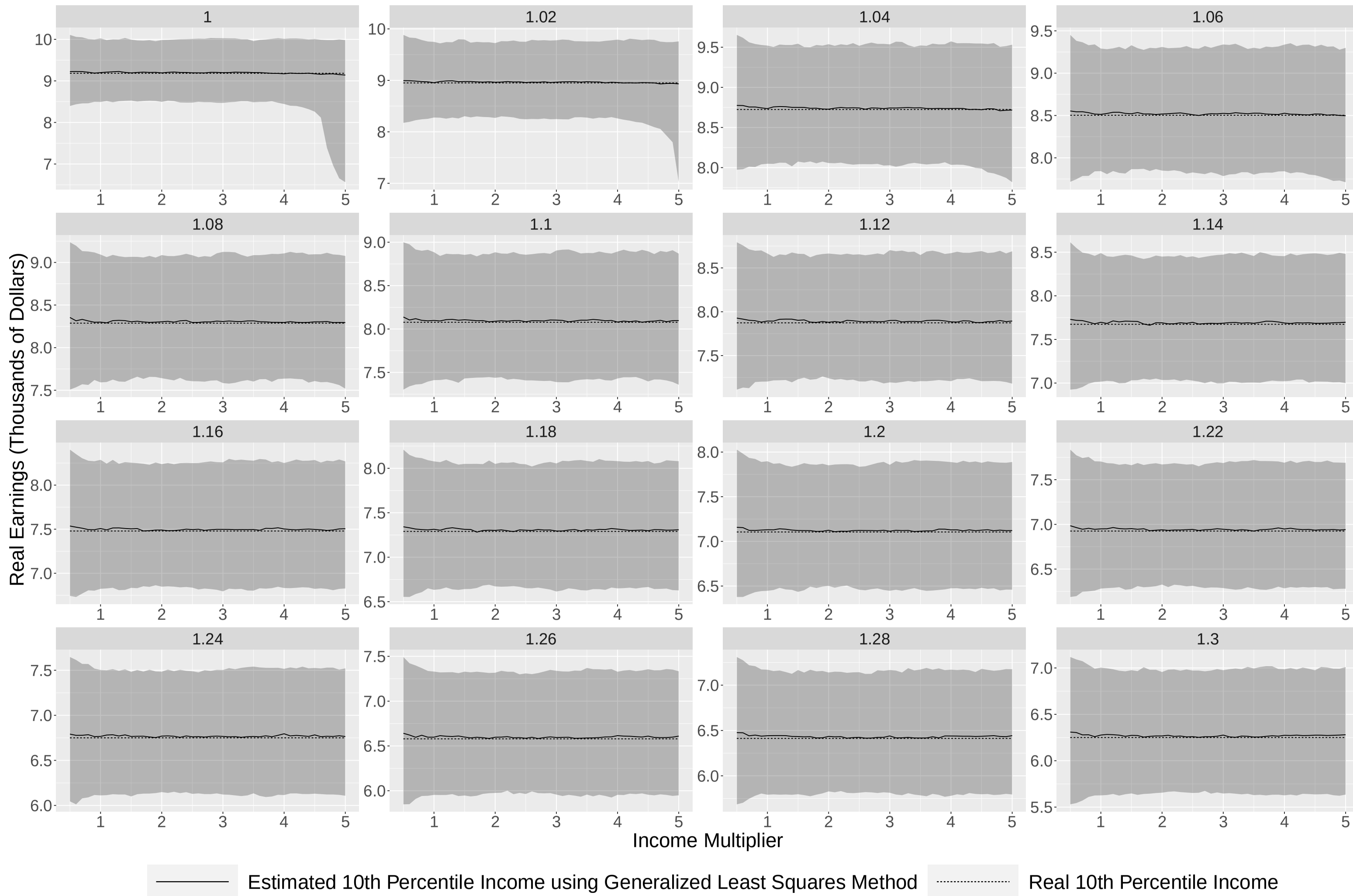


Figure 1.37: Performance of General Least Squares for Estimation of 10$^{\text{th}}$ Percentile of Income
Real income level for the 10$^{\text{th}}$ percentile estimated using the Generalized Least Squares method on binned data created by random sampling from a population with a lognormal distribution of incomes with a location parameter of 3.5 and a scale parameter varying from 1-1.3. Income normalized according to sample scale. Line is median estimate, bands represent the central 90% of estimates from random sampling from the distribution.

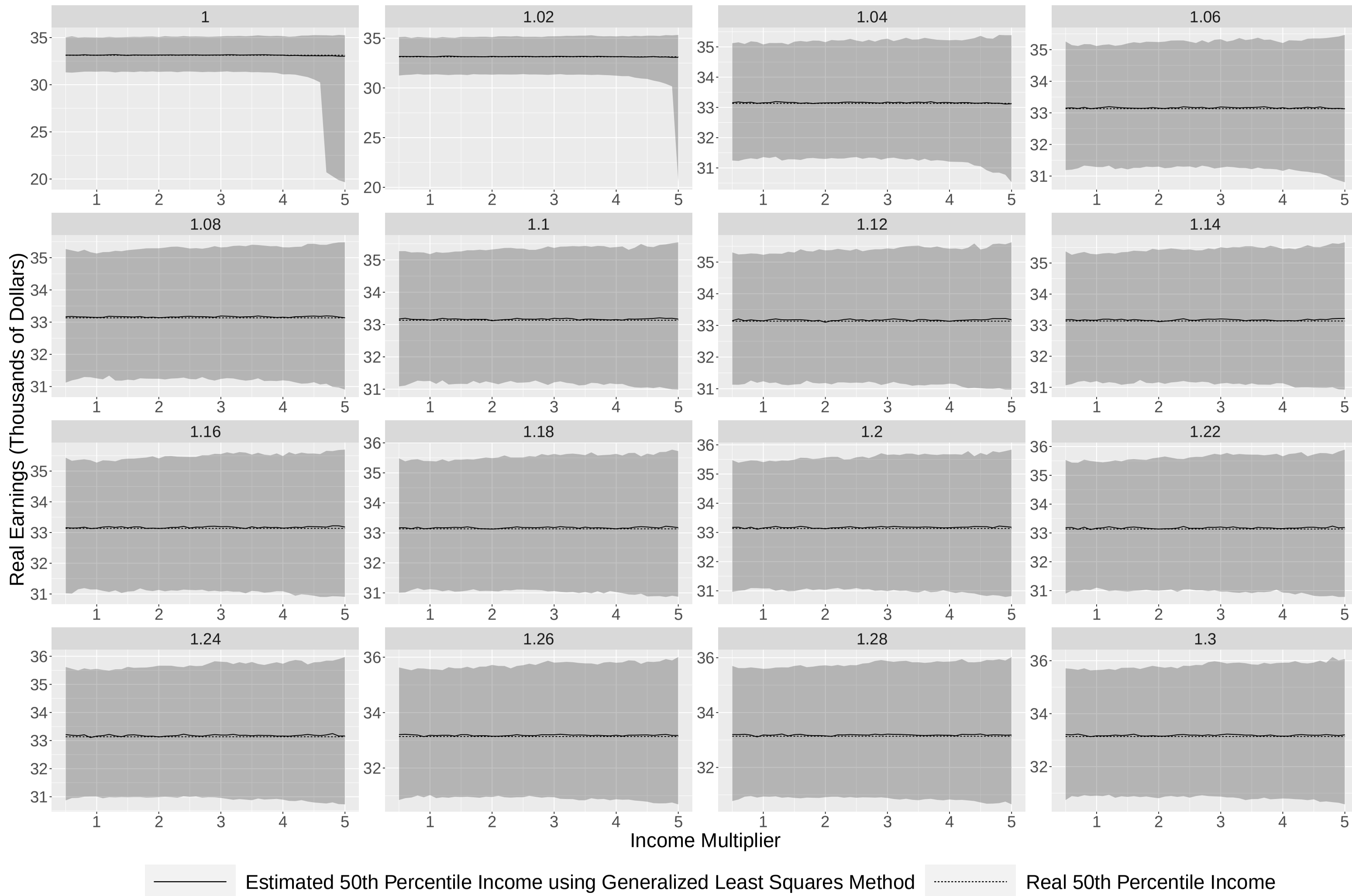


Figure 1.38: Performance of General Least Squares for Estimation of 50$^{\text{th}}$ Percentile of Income
Real income level for the 50$^{\text{th}}$ percentile estimated using the Generalized Least Squares method on binned data created by random sampling from a population with a lognormal distribution of incomes with a location parameter of 3.5 and a scale parameter varying from 1-1.3. Income normalized according to sample scale. Line is median estimate, bands represent the central 90% of estimates from random sampling from the distribution.

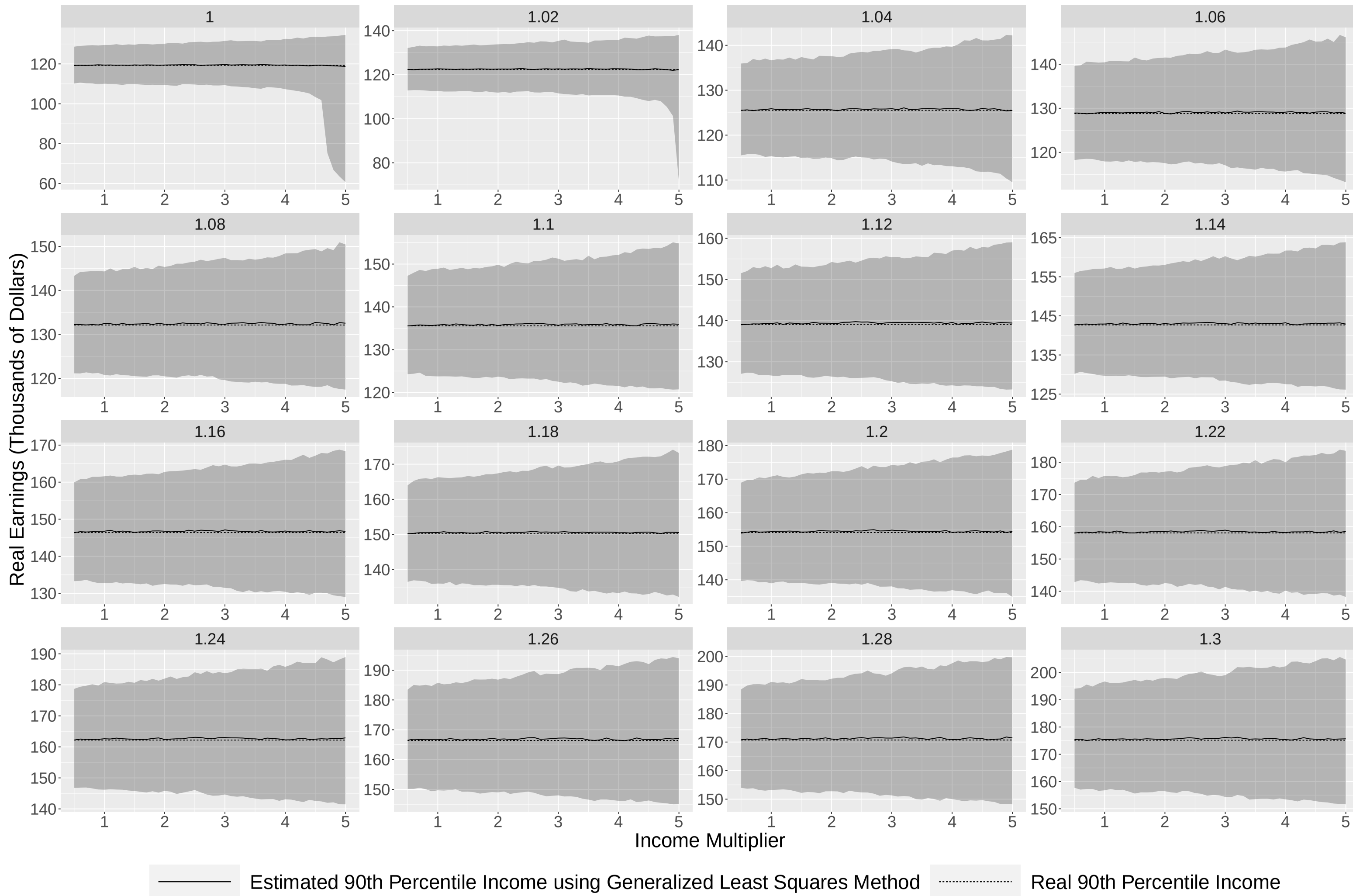


Figure 1.39: Performance of General Least Squares for Estimation of 90^{th} Percentile of Income
Real income level for the 90^{th} percentile estimated using the Generalized Least Squares method on binned data created by random sampling from a population with a lognormal distribution of incomes with a location parameter of 3.5 and a scale parameter varying from 1-1.3. Income normalized according to sample scale. Line is median estimate, bands represent the central 90% of estimates from random sampling from the distribution.

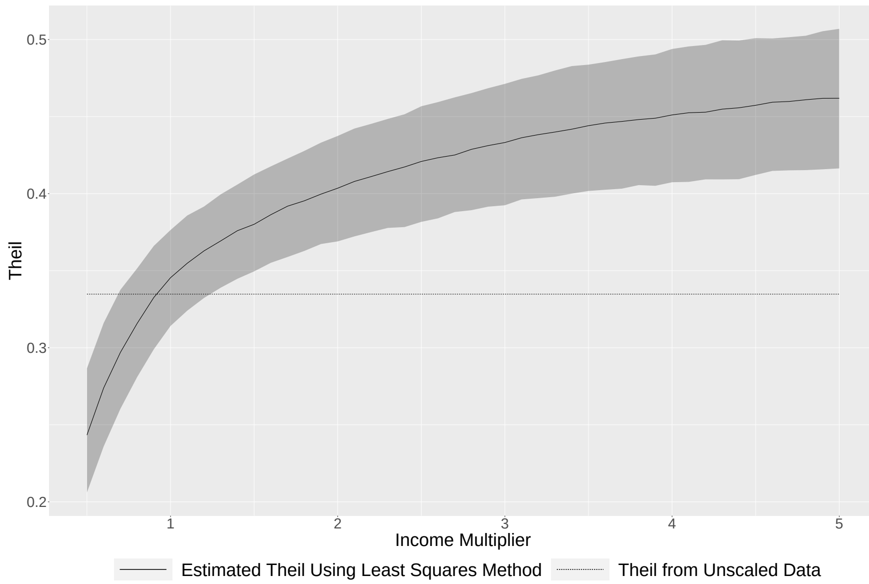


Figure 1.40: Performance of General Least Squares for Estimation of the Theil Index when Gamma Distribution is Misspecified
Theil estimated using the Generalized Least Squares method on binned data, assuming a lognormal distribution when the population actually follows a Gamma distribution. Line is median estimate, bands represent the central 90% of estimates from random sampling from the distribution.

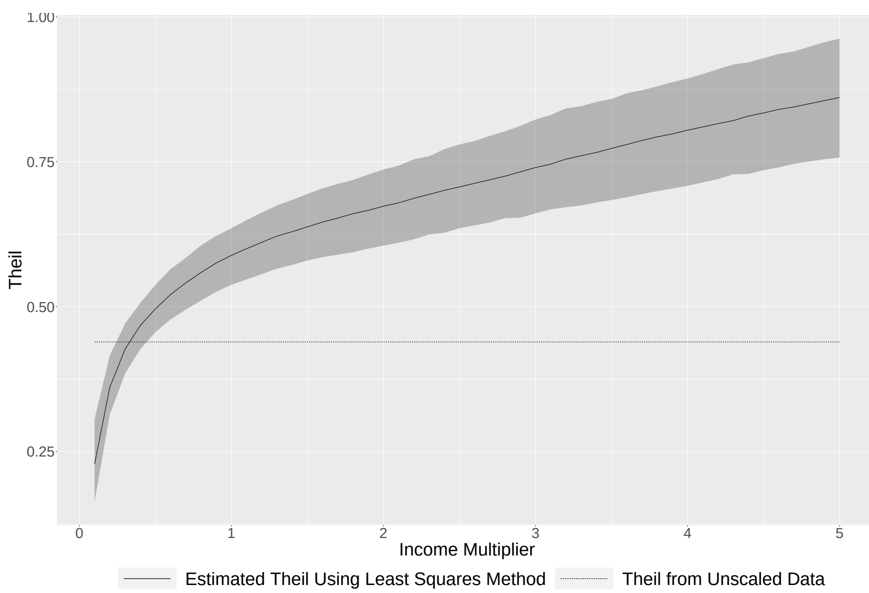


Figure 1.41: Performance of General Least Squares for Estimation of the Theil Index when Weibull Distribution is Misspecified
Theil estimated using the Generalized Least Squares method on binned data, assuming a lognormal distribution when the population actually follows a Weibull distribution. Line is median estimate, bands represent the central 90% of estimates from random sampling from the distribution.

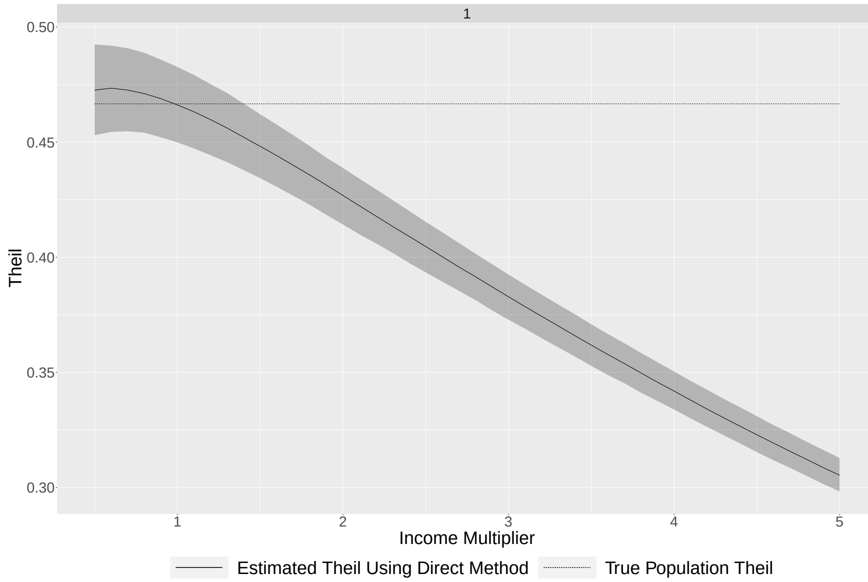


Figure 1.42: Performance of "Direct Method" for Estimation of the Theil Index, with Increased Sample Size Theil estimated using the "Direct Method" on binned data created by random sampling from a population with a lognormal distribution of incomes with a location parameter of 3.5 and a scale parameter varying from 1-1.3. Samples then scaled from one half to five times their original value. Line is median estimate, bands represent the central 90% of estimates from random sampling from the distribution. Using 10% Sample.

Figure 1.43: Performance of "Direct Method" for Estimation of the Gini Index, with Increased Sample Size Gini estimated using the "Direct Method" on binned data created by random sampling from a population with a lognormal distribution of incomes with a location parameter of 3.5 and a scale parameter varying from 1-1.3. Samples then scaled from one half to five times their original value. Line is median estimate, bands represent the central 90% of estimates from random sampling from the distribution. Using 10% Sample.

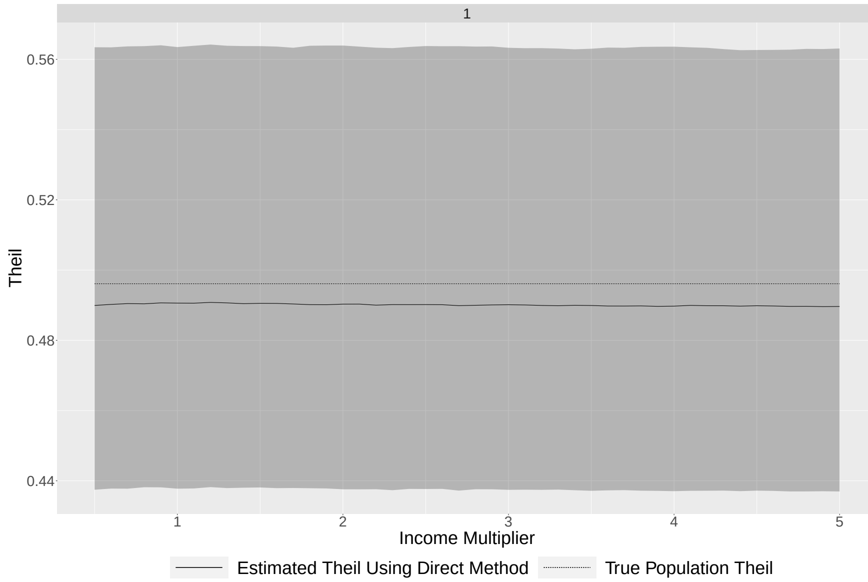


Figure 1.44: Performance of “Direct Method” for Estimation of the Theil Index, with Finer Bin Structure Theil estimated using the “Direct Method” on binned data created by random sampling from a population with a lognormal distribution of incomes with a location parameter of 3.5 and a scale parameter varying from 1-1.3. Samples then scaled from one half to five times their original value. Line is median estimate, bands represent the central 90% of estimates from random sampling from the distribution. Using finer bin structure.

Figure 1.45: Performance of “Direct Method” for Estimation of the Gini Index, with Finer Bin Structure Gini estimated using the “Direct Method” on binned data created by random sampling from a population with a lognormal distribution of incomes with a location parameter of 3.5 and a scale parameter varying from 1-1.3. Samples then scaled from one half to five times their original value. Line is median estimate, bands represent the central 90% of estimates from random sampling from the distribution. Using finer bin structure.

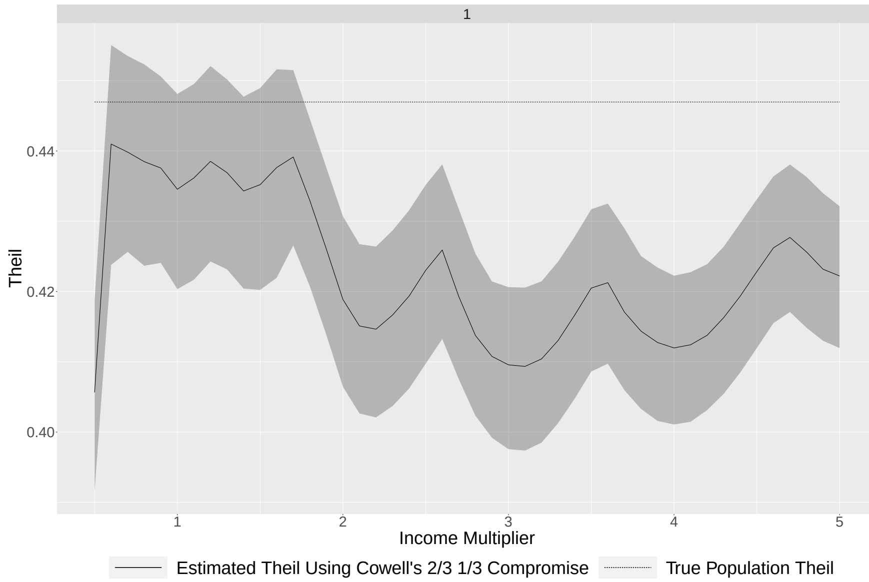


Figure 1.46: Performance of Cowell's compromise for Estimation of the Theil Index, with Increased Sample Size

Theil estimated using the Cowell's "2/3 1/3" compromise on binned data created by random sampling from a population with a lognormal distribution of incomes with a location parameter of 3.5 and a scale parameter varying from 1-1.3. Samples then scaled from one half to five times their original value. Line is median estimate, bands represent the central 90% of estimates from random sampling from the distribution. Using 10% sample.

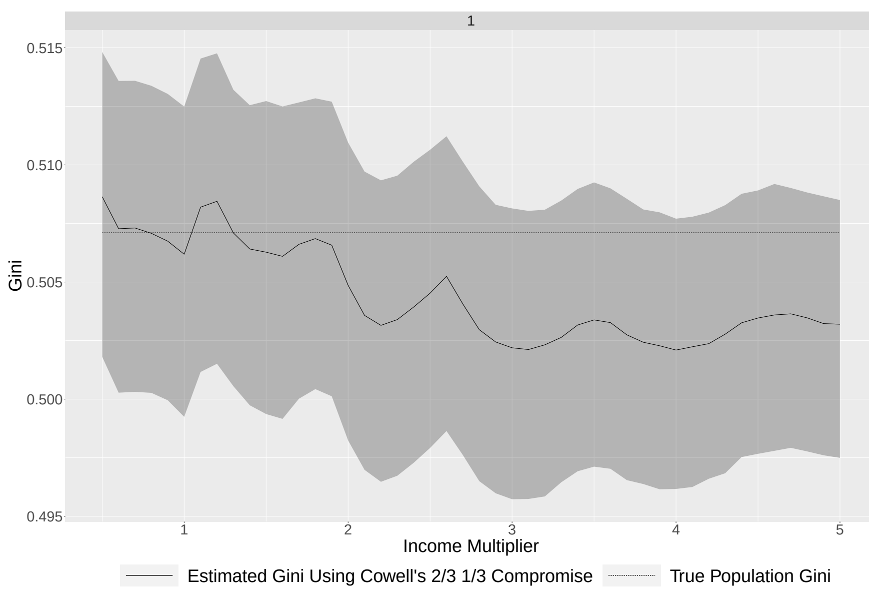


Figure 1.47: Performance of Cowell's compromise for Estimation of the Gini Index, with Increased Sample Size

Gini estimated using the Cowell's "2/3 1/3" compromise on binned data created by random sampling from a population with a lognormal distribution of incomes with a location parameter of 3.5 and a scale parameter varying from 1-1.3. Samples then scaled from one half to five times their original value. Line is median estimate, bands represent the central 90% of estimates from random sampling from the distribution. Using 10% sample.

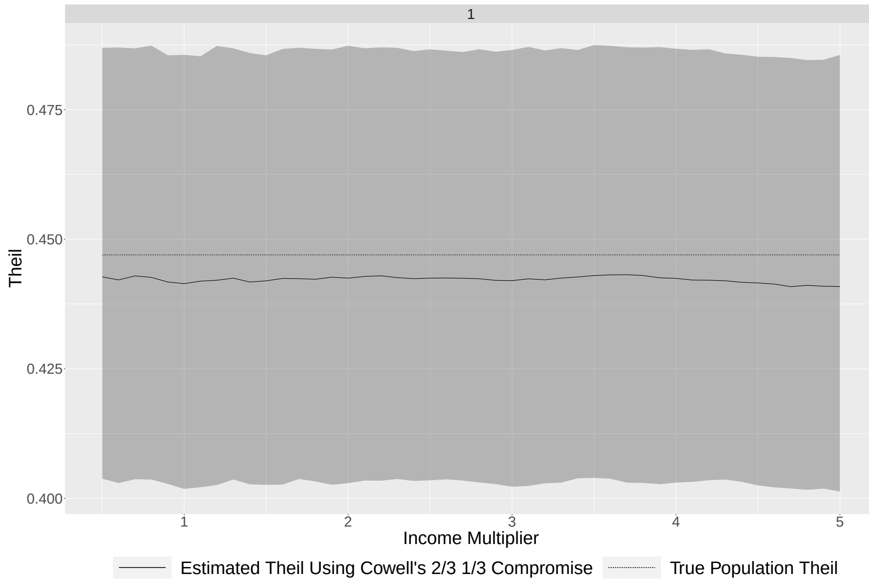


Figure 1.48: Performance of Cowell's compromise for Estimation of the Theil Index, with Increased Sample Size

Theil estimated using the Cowell's "2/3 1/3" compromise on binned data created by random sampling from a population with a lognormal distribution of incomes with a location parameter of 3.5 and a scale parameter varying from 1-1.3. Samples then scaled from one half to five times their original value. Line is median estimate, bands represent the central 90% of estimates from random sampling from the distribution. Using finer bin structure.

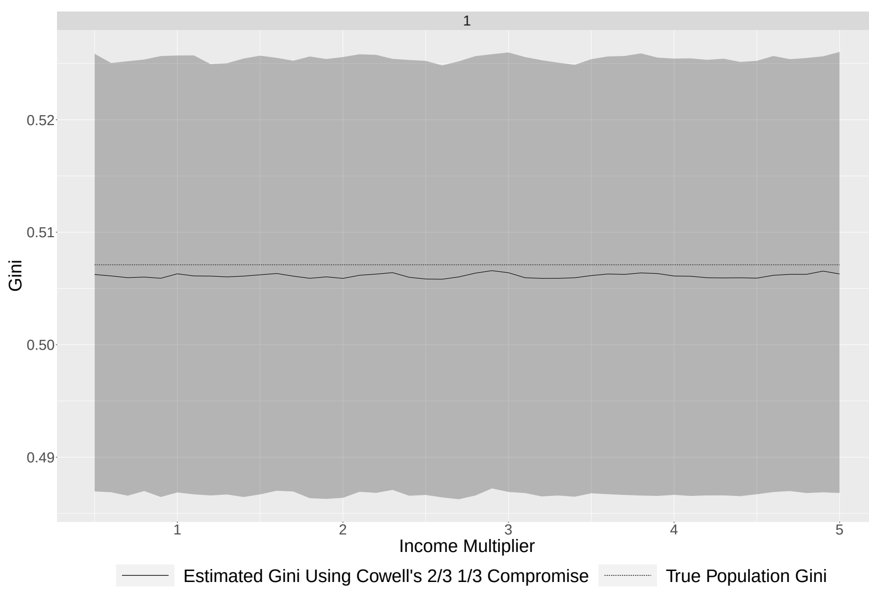


Figure 1.49: Performance of Cowell's compromise for Estimation of the Gini Index, with Increased Sample Size

Gini estimated using the Cowell's "2/3 1/3" compromise on binned data created by random sampling from a population with a lognormal distribution of incomes with a location parameter of 3.5 and a scale parameter varying from 1-1.3. Samples then scaled from one half to five times their original value. Line is median estimate, bands represent the central 90% of estimates from random sampling from the distribution. Using finer bin structure.

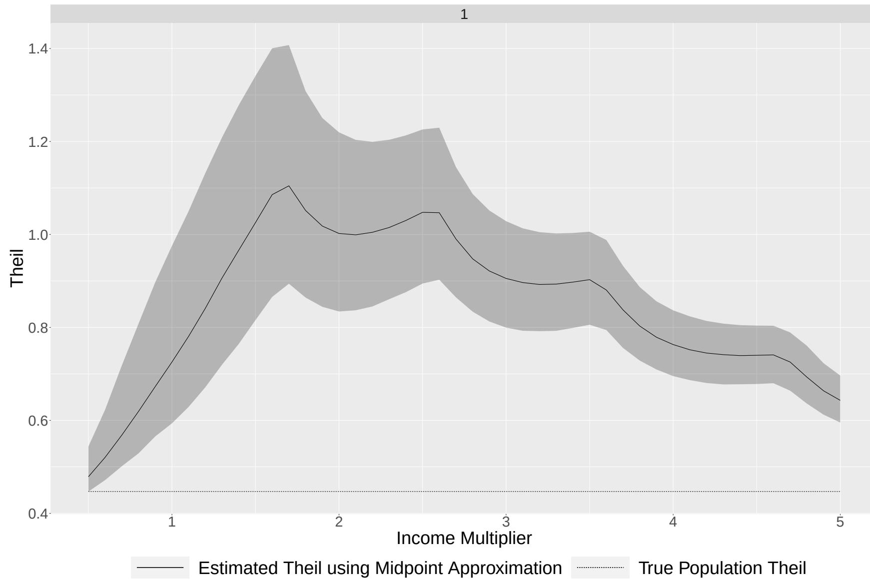


Figure 1.50: Performance of "Midpoint Approximation" for Estimation of the Theil Index, with Increased Sample Size
Theil estimated using the "Midpoint Approximation" on binned data created by random sampling from a population with a lognormal distribution of incomes with a location parameter of 3.5 and a scale parameter varying from 1-1.3. Samples then scaled from one half to five times their original value. Line is median estimate, bands represent the central 90% of estimates from random sampling from the distribution. Using 10% sample.

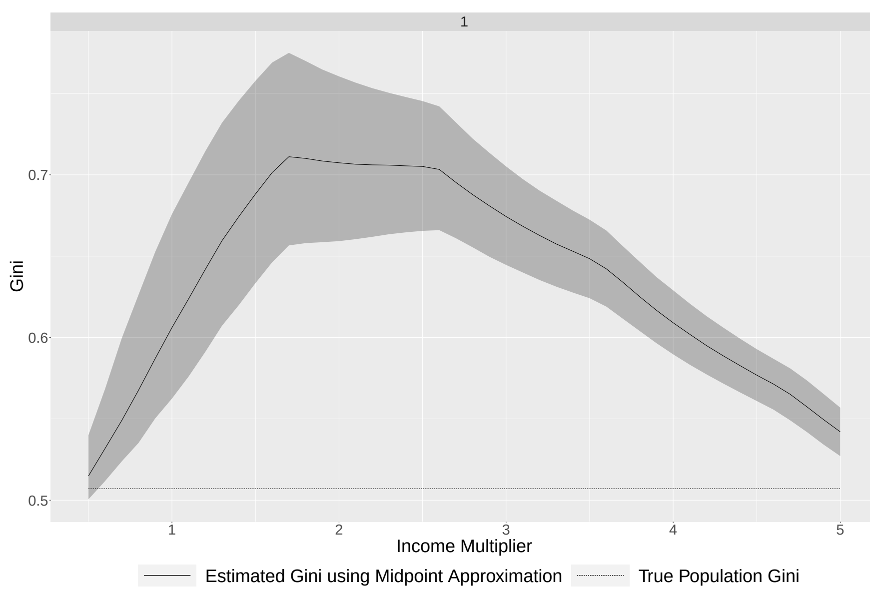


Figure 1.51: Performance of "Midpoint Approximation" for Estimation of the Gini Index, with Increased Sample Size
Gini estimated using the "Midpoint Approximation" on binned data created by random sampling from a population with a lognormal distribution of incomes with a location parameter of 3.5 and a scale parameter varying from 1-1.3. Samples then scaled from one half to five times their original value. Line is median estimate, bands represent the central 90% of estimates from random sampling from the distribution. Using 10% sample.

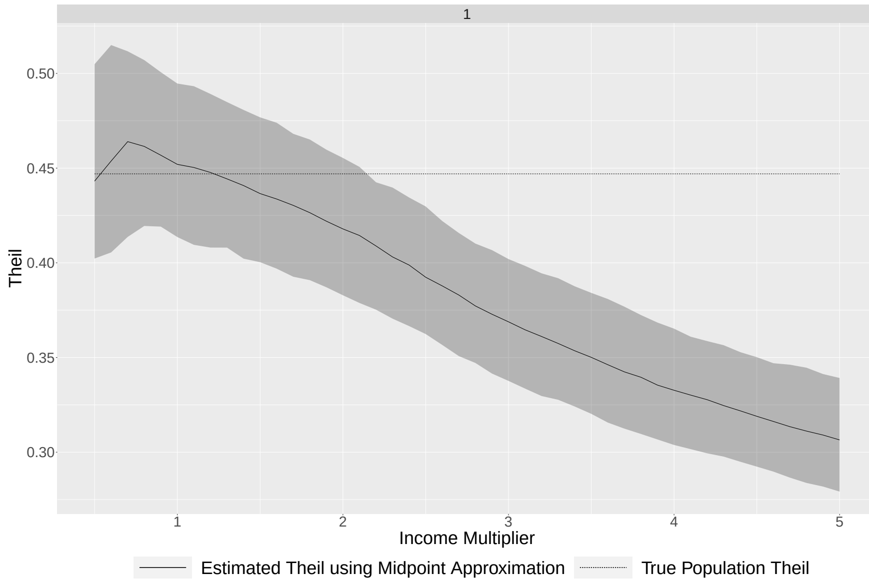


Figure 1.52: Performance of “Midpoint Approximation” for Estimation of the Theil Index, with Finer Bin Structure
Theil estimated using the “Midpoint Approximation” on binned data created by random sampling from a population with a lognormal distribution of incomes with a location parameter of 3.5 and a scale parameter varying from 1-1.3. Samples then scaled from one half to five times their original value. Line is median estimate, bands represent the central 90% of estimates from random sampling from the distribution. Using finer bin structure.

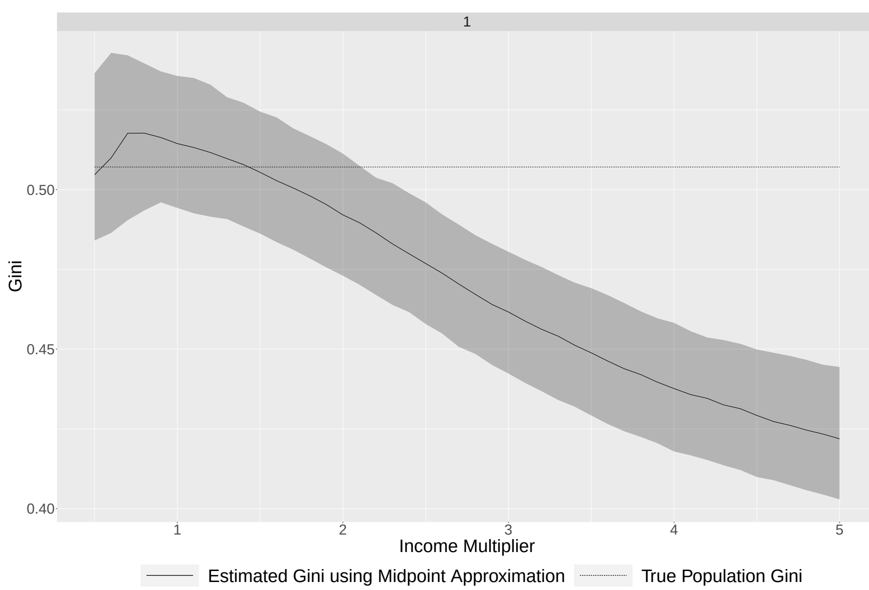


Figure 1.53: Performance of “Midpoint Approximation” for Estimation of the Gini Index, with Finer Bin Structure
Gini estimated using the “Midpoint Approximation” on binned data created by random sampling from a population with a lognormal distribution of incomes with a location parameter of 3.5 and a scale parameter varying from 1-1.3. Samples then scaled from one half to five times their original value. Line is median estimate, bands represent the central 90% of estimates from random sampling from the distribution. Using finer bin structure.

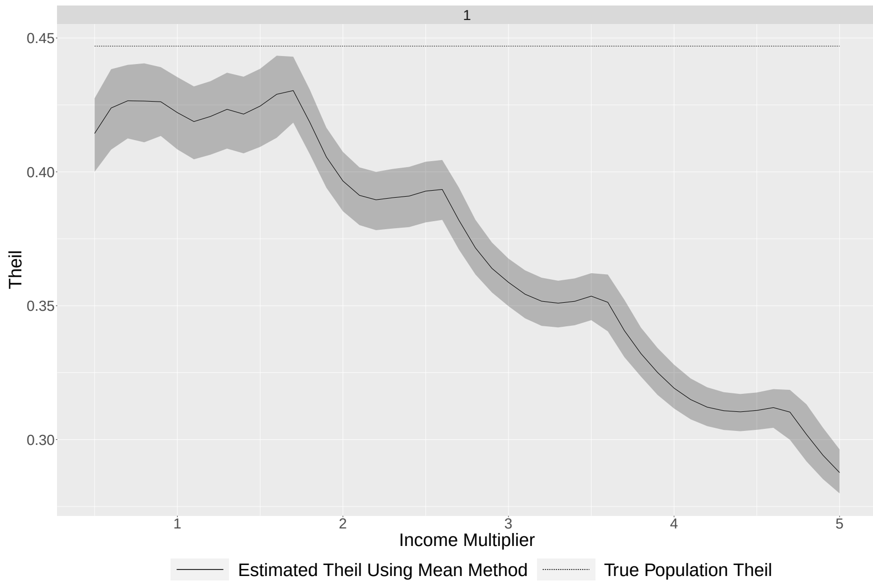


Figure 1.54: Performance of “Mean Approximation” for Estimation of the Theil Index, with Increased Sample Size
Theil estimated using the “Mean Approximation” on binned data created by random sampling from a population with a lognormal distribution of incomes with a location parameter of 3.5 and a scale parameter varying from 1-1.3. Samples then scaled from one half to five times their original value. Line is median estimate, bands represent the central 90% of estimates from random sampling from the distribution. Using 10% sample.

Figure 1.55
Gini estimated using the “Mean Approximation” on binned data created by random sampling from a population with a lognormal distribution of incomes with a location parameter of 3.5 and a scale parameter varying from 1-1.3. Samples then scaled from one half to five times their original value. Line is median estimate, bands represent the central 90% of estimates from random sampling from the distribution. Using 10% sample.

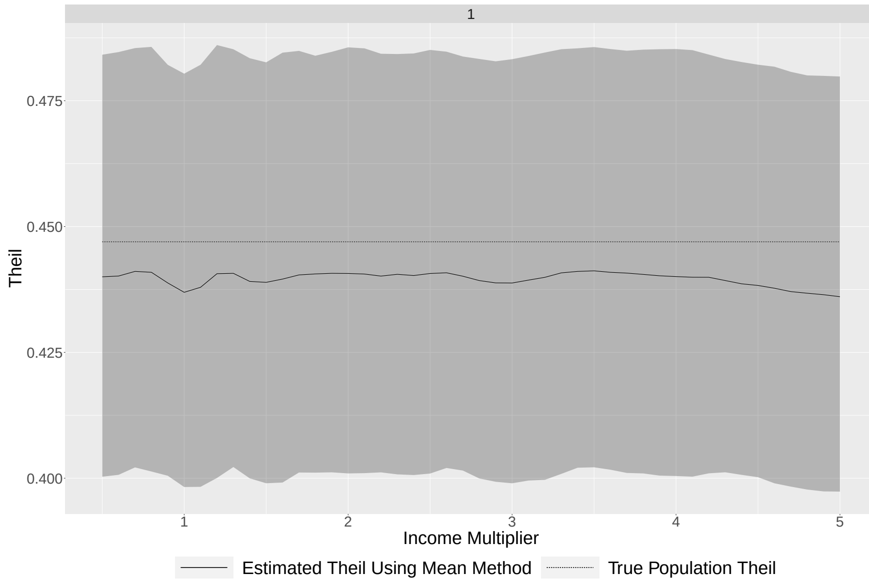


Figure 1.56: Performance of "Mean Approximation" for Estimation of the Theil Index, with Finer Bin Structure
Theil estimated using the "Mean Approximation" on binned data created by random sampling from a population with a lognormal distribution of incomes with a location parameter of 3.5 and a scale parameter varying from 1-1.3. Samples then scaled from one half to five times their original value. Line is median estimate, bands represent the central 90% of estimates from random sampling from the distribution. Using finer bin structure.

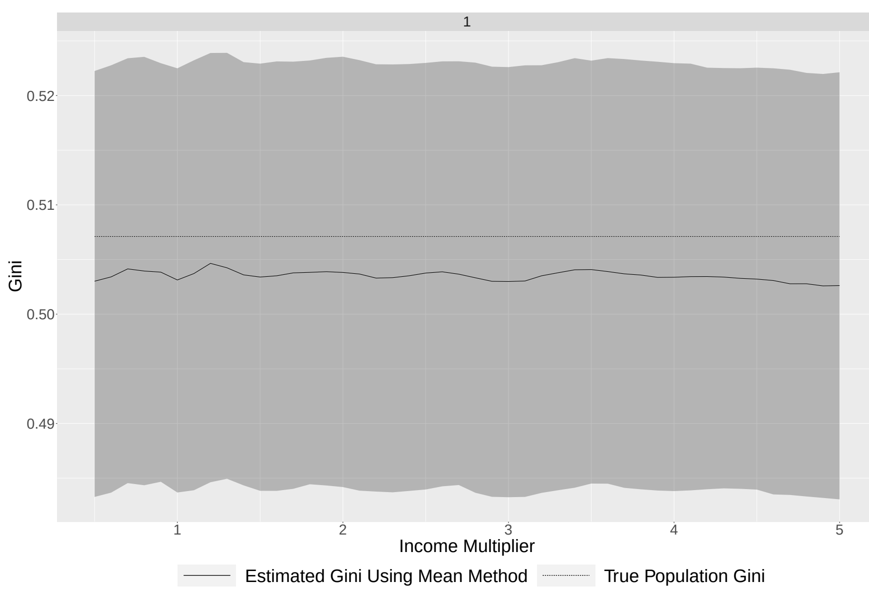


Figure 1.57: Performance of "Mean Approximation" for Estimation of the Gini Index, with Finer Bin Structure
Gini estimated using the "Mean Approximation" on binned data created by random sampling from a population with a lognormal distribution of incomes with a location parameter of 3.5 and a scale parameter varying from 1-1.3. Samples then scaled from one half to five times their original value. Line is median estimate, bands represent the central 90% of estimates from random sampling from the distribution. Using finer bin structure.

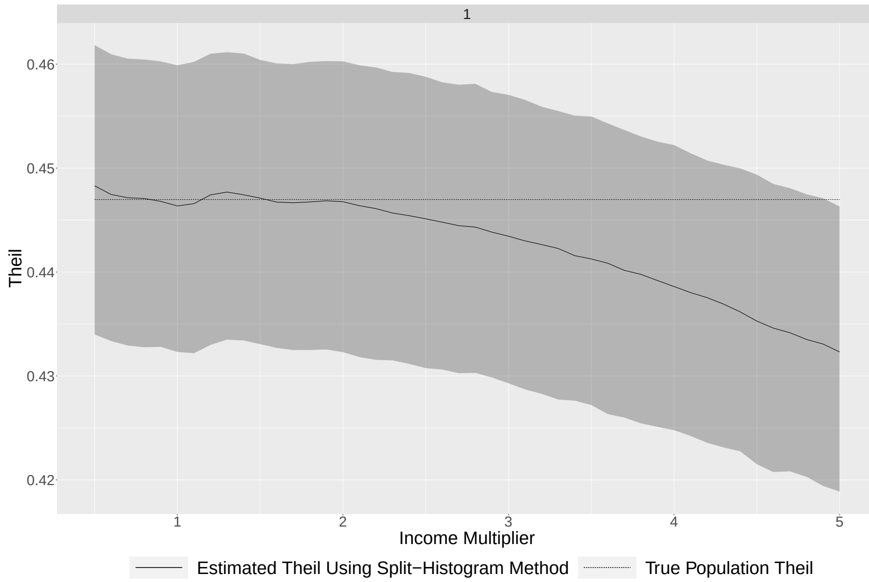


Figure 1.58: Performance of “Split-Histogram” Method for Estimation of the Theil Index, with Increased Sample Size
Theil estimated using the “Split-Histogram” method on binned data created by random sampling from a population with a lognormal distribution of incomes with a location parameter of 3.5 and a scale parameter varying from 1-1.3. Samples then scaled from one half to five times their original value. Line is median estimate, bands represent the central 90% of estimates from random sampling from the distribution. Using 10% sample.

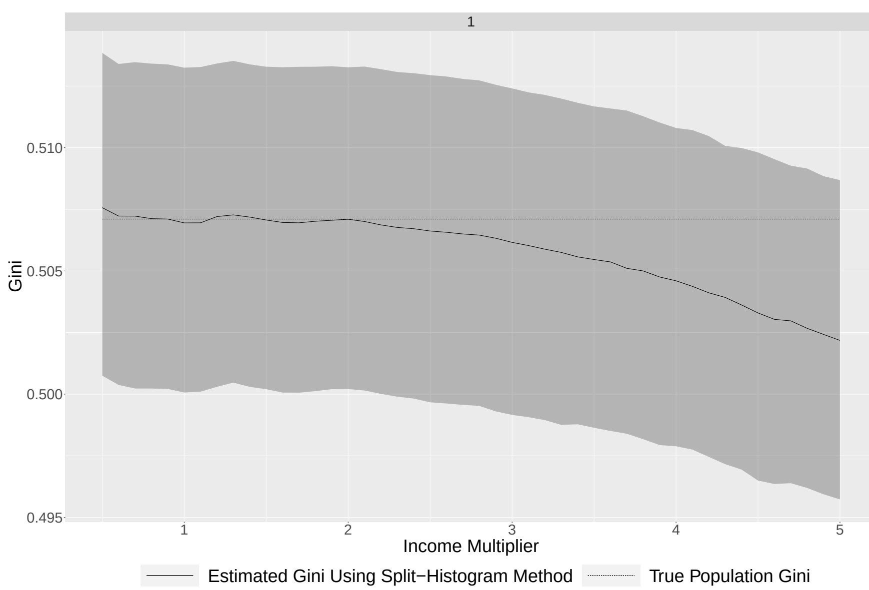


Figure 1.59: Performance of “Split-Histogram” Method for Estimation of the Gini Index, with Increased Sample Size
Gini estimated using the “Split-Histogram” method on binned data created by random sampling from a population with a lognormal distribution of incomes with a location parameter of 3.5 and a scale parameter varying from 1-1.3. Samples then scaled from one half to five times their original value. Line is median estimate, bands represent the central 90% of estimates from random sampling from the distribution. Using 10% sample.

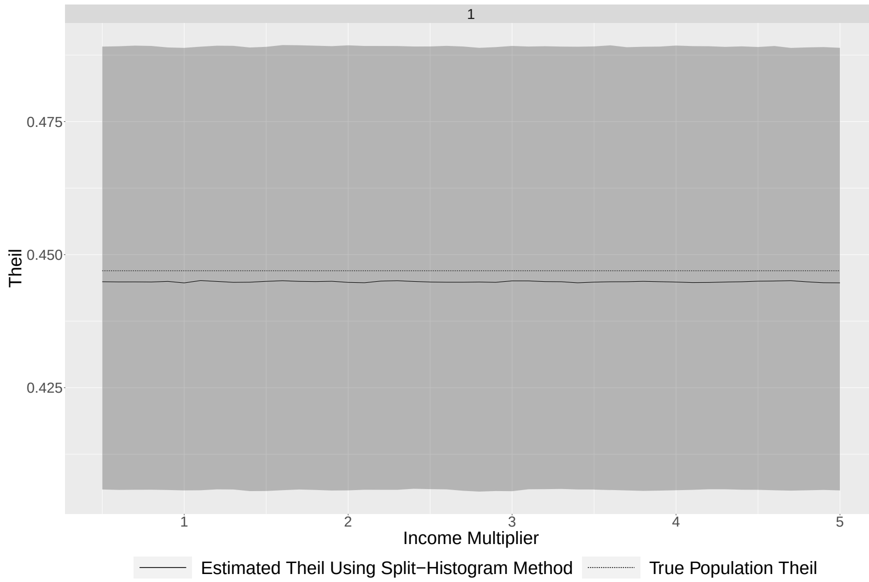


Figure 1.60: Performance of "Split-Histogram" Method for Estimation of the Theil Index, with Finer Bin Structure
Theil estimated using the "Split-Histogram" method on binned data created by random sampling from a population with a lognormal distribution of incomes with a location parameter of 3.5 and a scale parameter varying from 1-1.3. Samples then scaled from one half to five times their original value. Line is median estimate, bands represent the central 90% of estimates from random sampling from the distribution. Using finer bin structure.

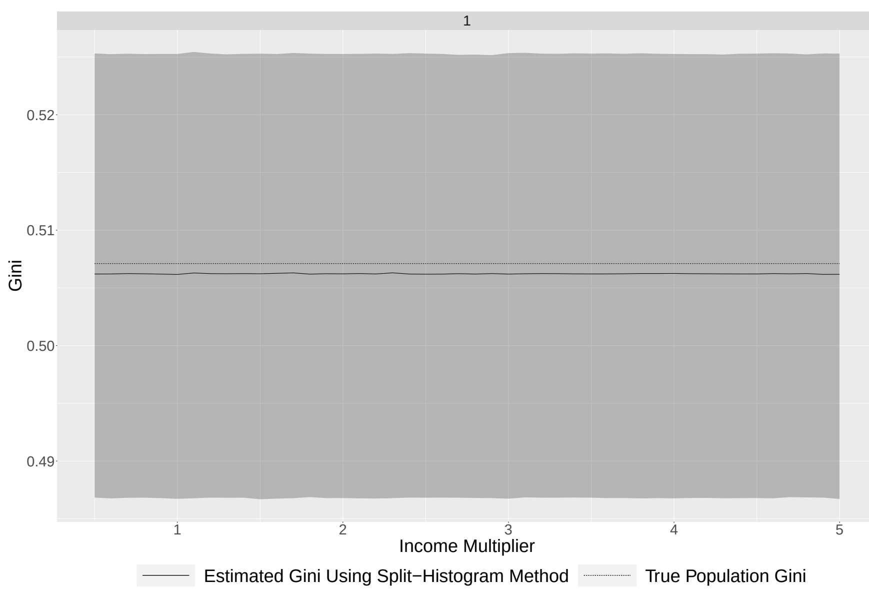


Figure 1.61: Performance of "Split-Histogram" Method for Estimation of the Gini Index, with Finer Bin Structure
Gini estimated using the "Split-Histogram" method on binned data created by random sampling from a population with a lognormal distribution of incomes with a location parameter of 3.5 and a scale parameter varying from 1-1.3. Samples then scaled from one half to five times their original value. Line is median estimate, bands represent the central 90% of estimates from random sampling from the distribution. Using finer bin structure.

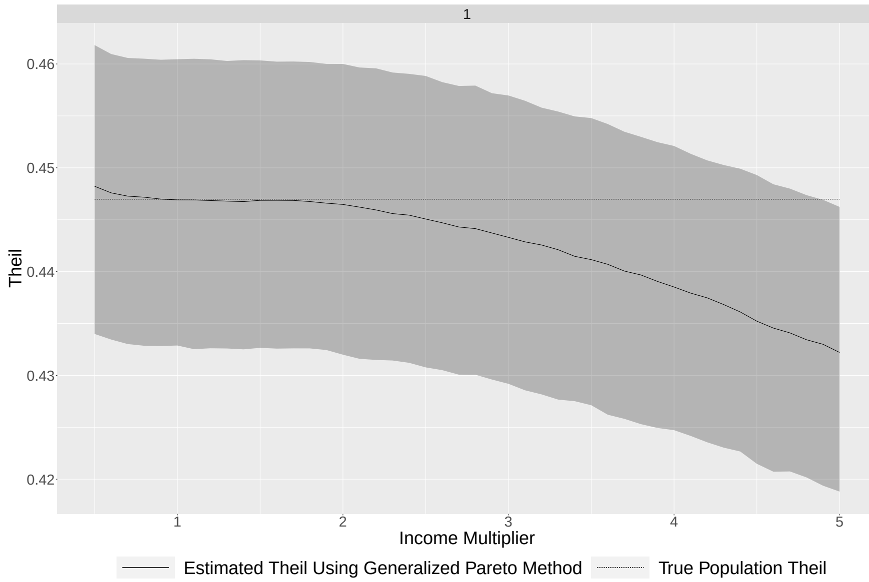


Figure 1.62: Performance of “Generalized Pareto” Method for Estimation of the Theil Index, with Increased Sample Size
Theil estimated using the “Generalized Pareto” method on binned data created by random sampling from a population with a lognormal distribution of incomes with a location parameter of 3.5 and a scale parameter varying from 1-1.3. Samples then scaled from one half to five times their original value. Line is median estimate, bands represent the central 90% of estimates from random sampling from the distribution. Using 10% sample.

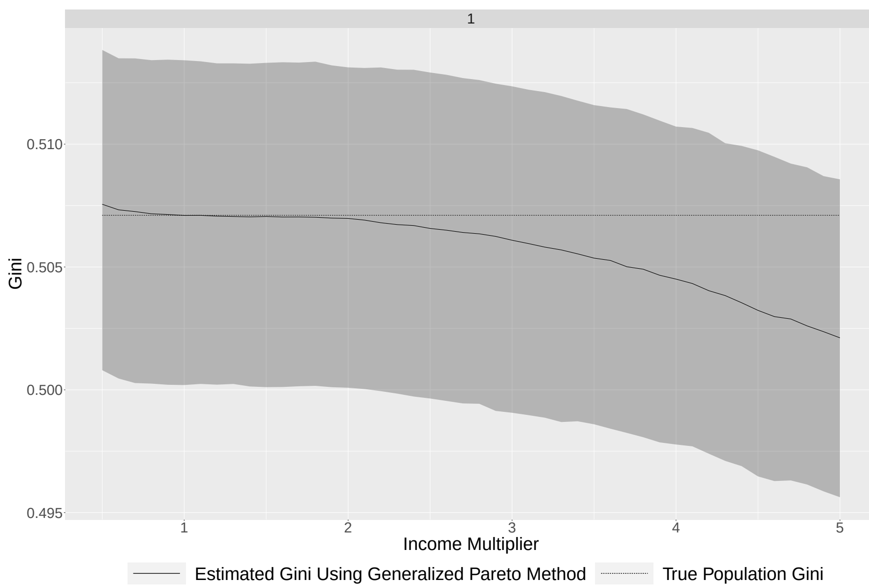


Figure 1.63: Performance of “Generalized Pareto” Method for Estimation of the Gini Index, with Increased Sample Size
Gini estimated using the “Generalized Pareto” method on binned data created by random sampling from a population with a lognormal distribution of incomes with a location parameter of 3.5 and a scale parameter varying from 1-1.3. Samples then scaled from one half to five times their original value. Line is median estimate, bands represent the central 90% of estimates from random sampling from the distribution. Using 10% sample.

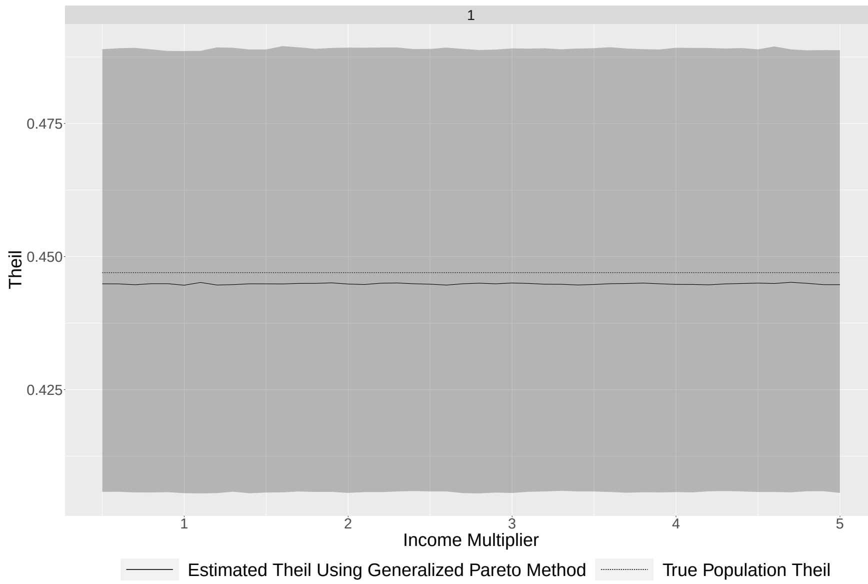
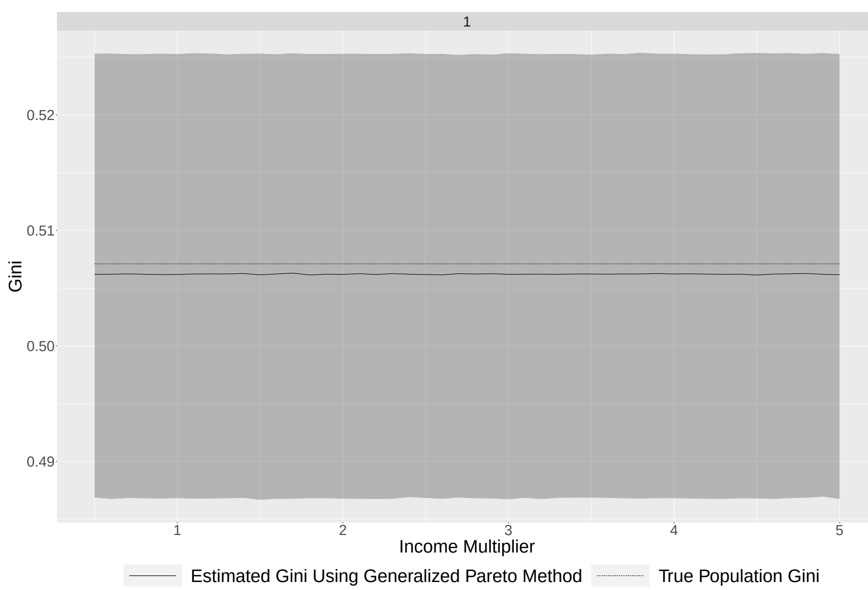


Figure 1.64: Performance of “Generalized Pareto” Method for Estimation of the Theil Index, with Finer Bin Structure
Theil estimated using the “Generalized Pareto” method on binned data created by random sampling from a population with a lognormal distribution of incomes with a location parameter of 3.5 and a scale parameter varying from 1-1.3. Samples then scaled from one half to five times their original value. Line is median estimate, bands represent the central 90% of estimates from random sampling from the distribution. Using finer bin structure.



Performance of “Generalized Pareto” Method for Estimation of the Gini Index, with Finer Bin Structure
Gini estimated using the “Generalized Pareto” method on binned data created by random sampling from a population with a lognormal distribution of incomes with a location parameter of 3.5 and a scale parameter varying from 1-1.3. Samples then scaled from one half to five times their original value. Line is median estimate, bands represent the central 90% of estimates from random sampling from the distribution. Using finer bin structure.

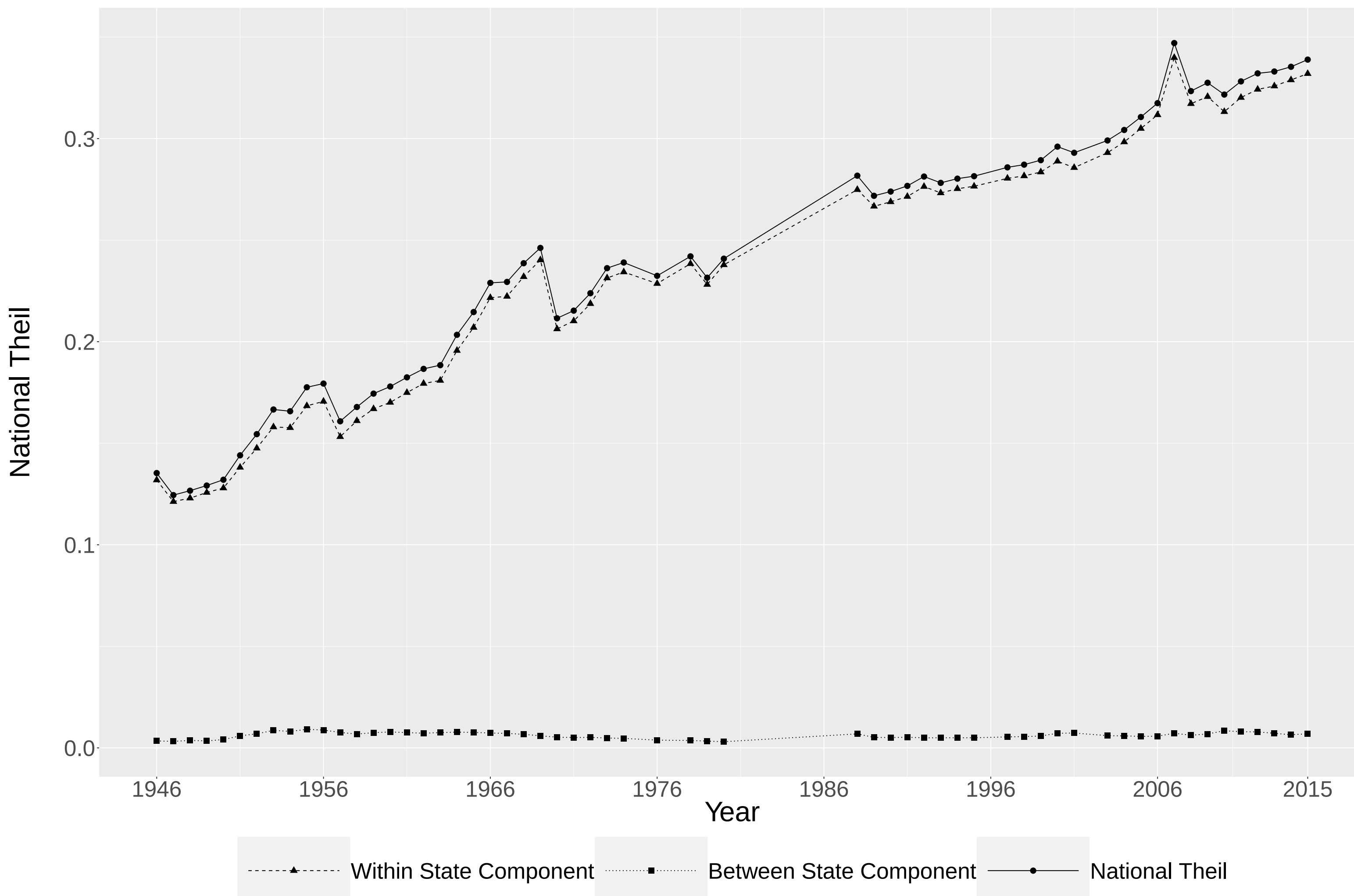


Figure 2.1: National Theil, Within- and Between-State Components, State-Level Estimation 1946-2015
National Theil, between-state, and within-state components. Calculated using state-level distributions, estimated with the least squares method, using Jensen-Shannon divergence to select distribution for each state-year observation.

Figure 2.2: National Theil Between-State Components 1946-2015
Between-state component of National Theil calculated using state level distributions, estimated with the least squares method, using Jensen-Shannon divergence to select distribution for each state-year observation.

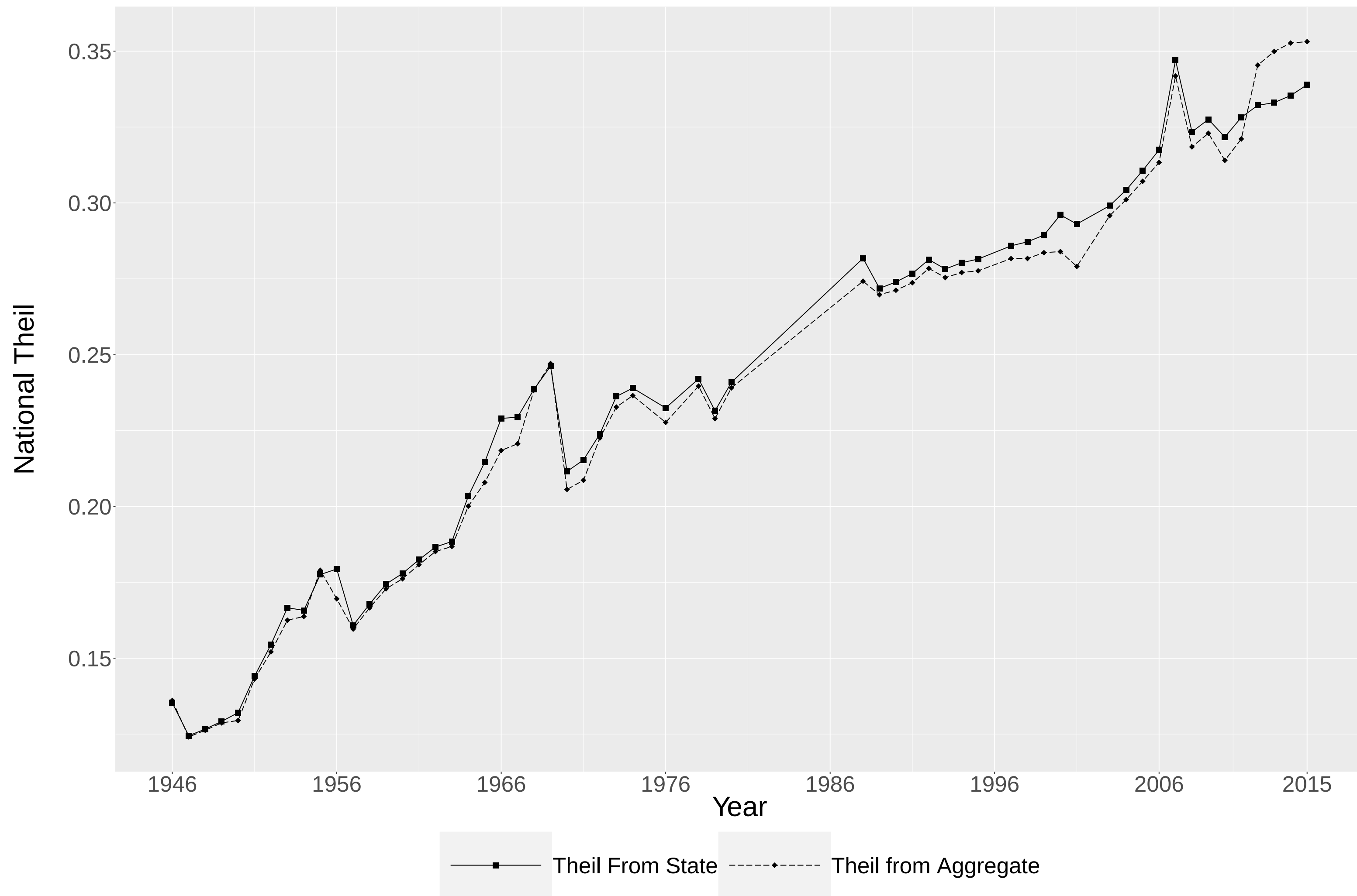


Figure 2.3: National Theil, Aggregate and State-Level Estimation 1946-2015
Comparison of National Theil calculated by state-level distributions versus national distribution estimated with least squares method, using Jensen-Shannon divergence to select distribution.

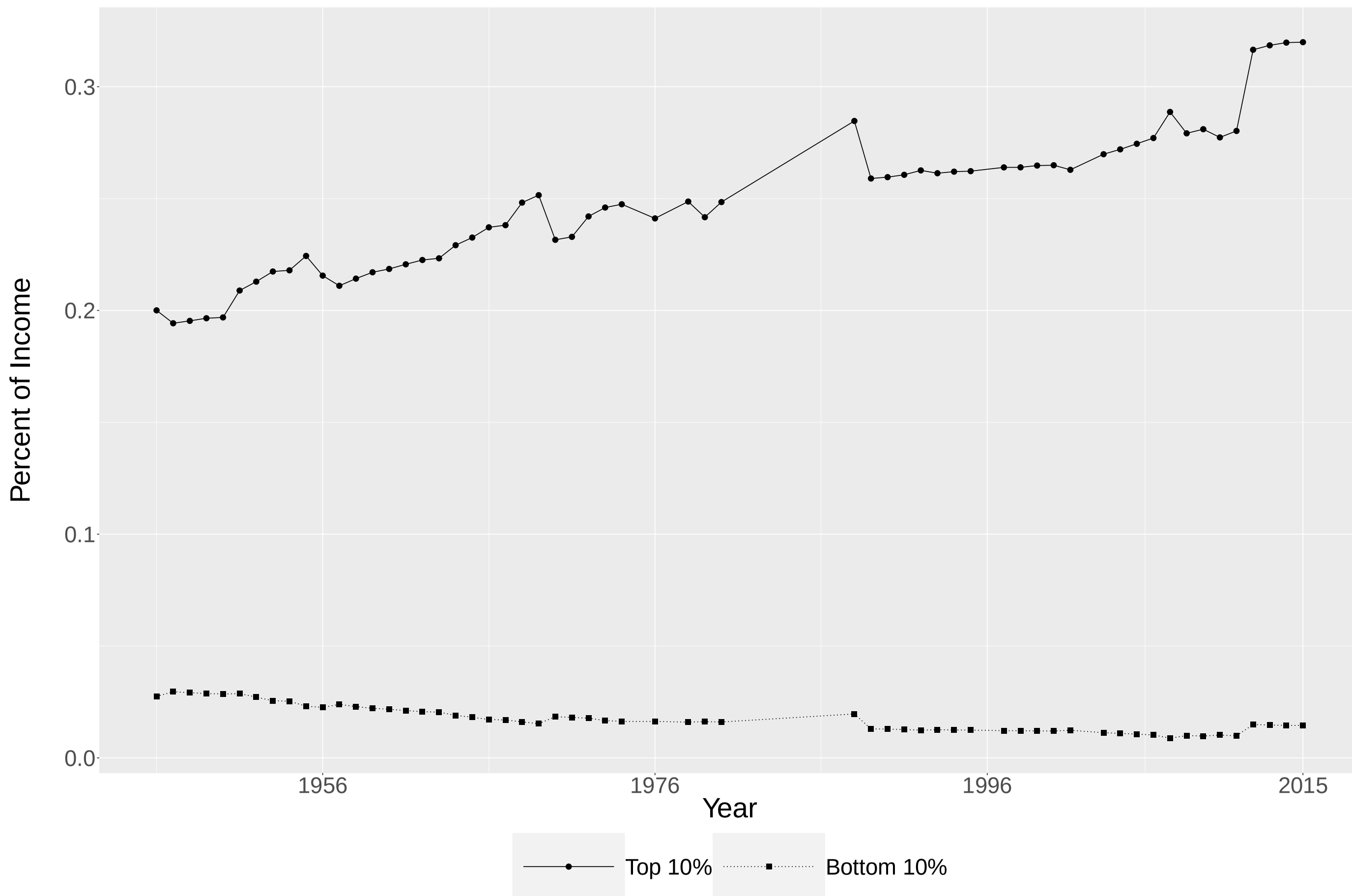


Figure 2.4: Income Share Top and Bottom Population Decile (National) 1946-2015
Percent of total national income held by the top and bottom 10% of earners of interest (5-95% of total population). Calculated using Jensen-Shannon divergence to select distribution.

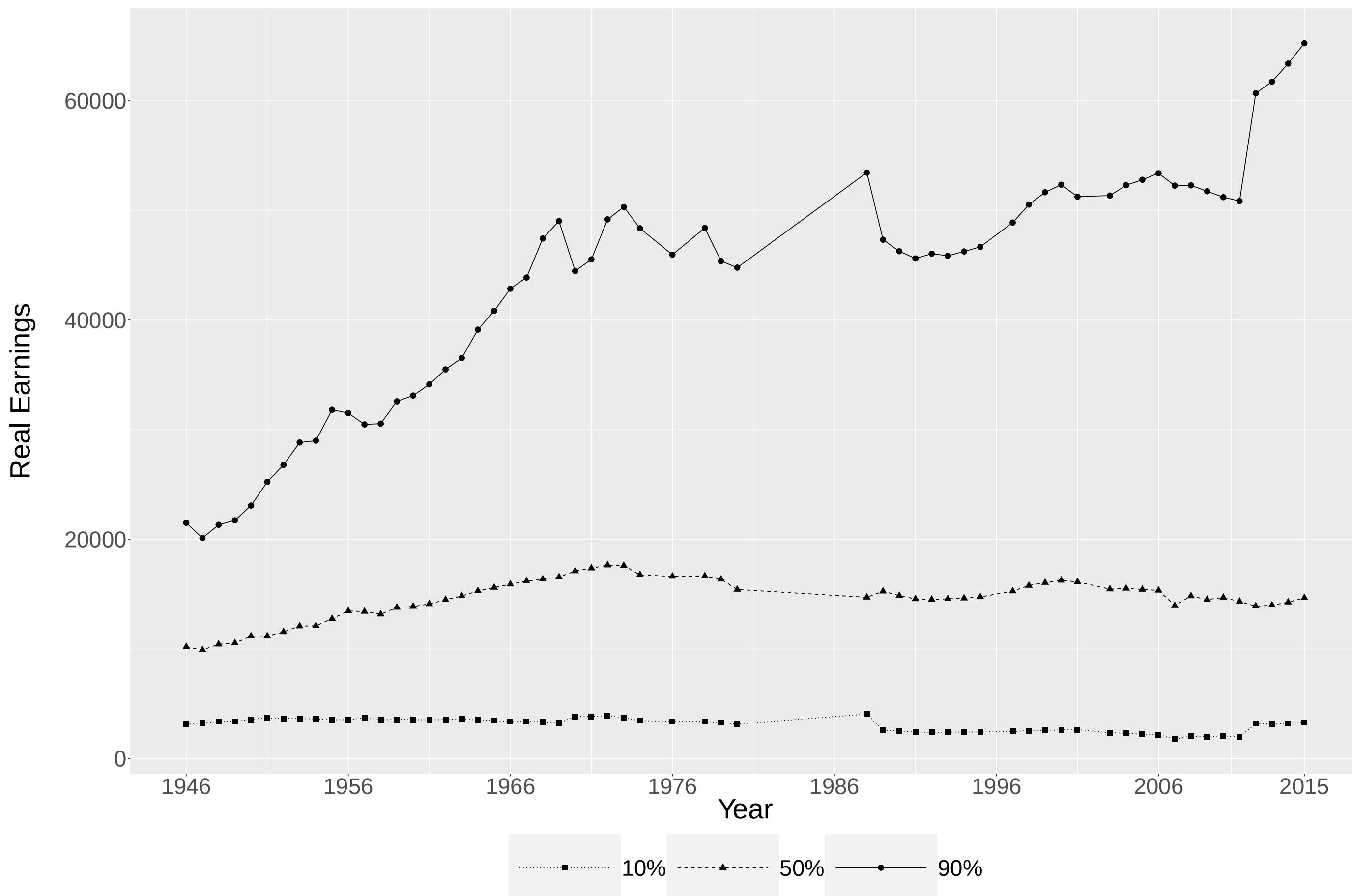


Figure 2.5: Real 10^{th}, 50^{th},90^{th} Percentile Income (National) 1946-2015
Real earnings at the 10^{th}, 50^{th}, and 90^{th} percentile of the national population in 1982-1984 adjusted dollars and using Jensen-Shannon divergence to select distribution.

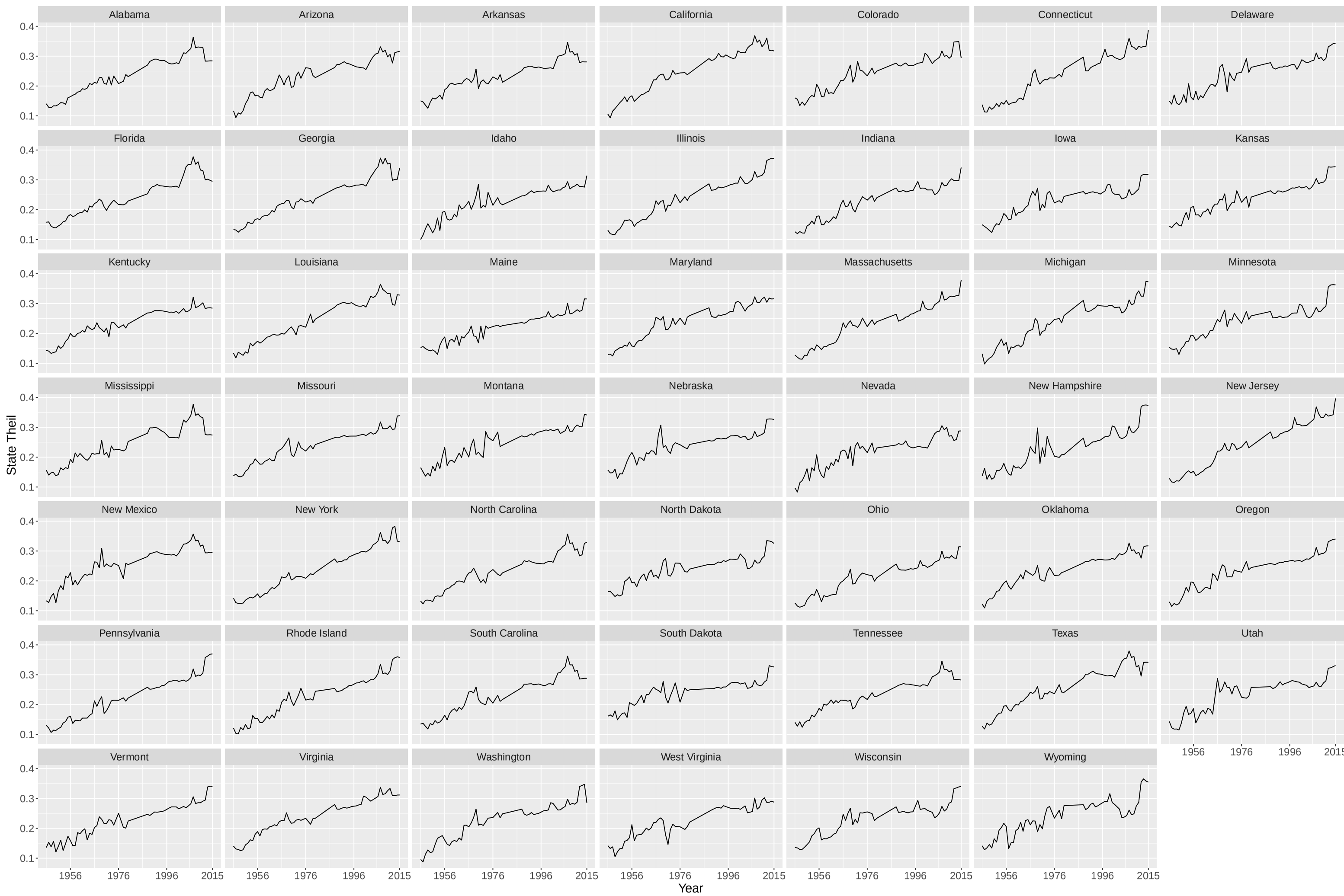


Figure 2.6: State Theils 1946-2015

Theil index for the contiguous 48 states. Calculated using least squares method, using Jensen-Shannon divergence to select distribution by state and year.

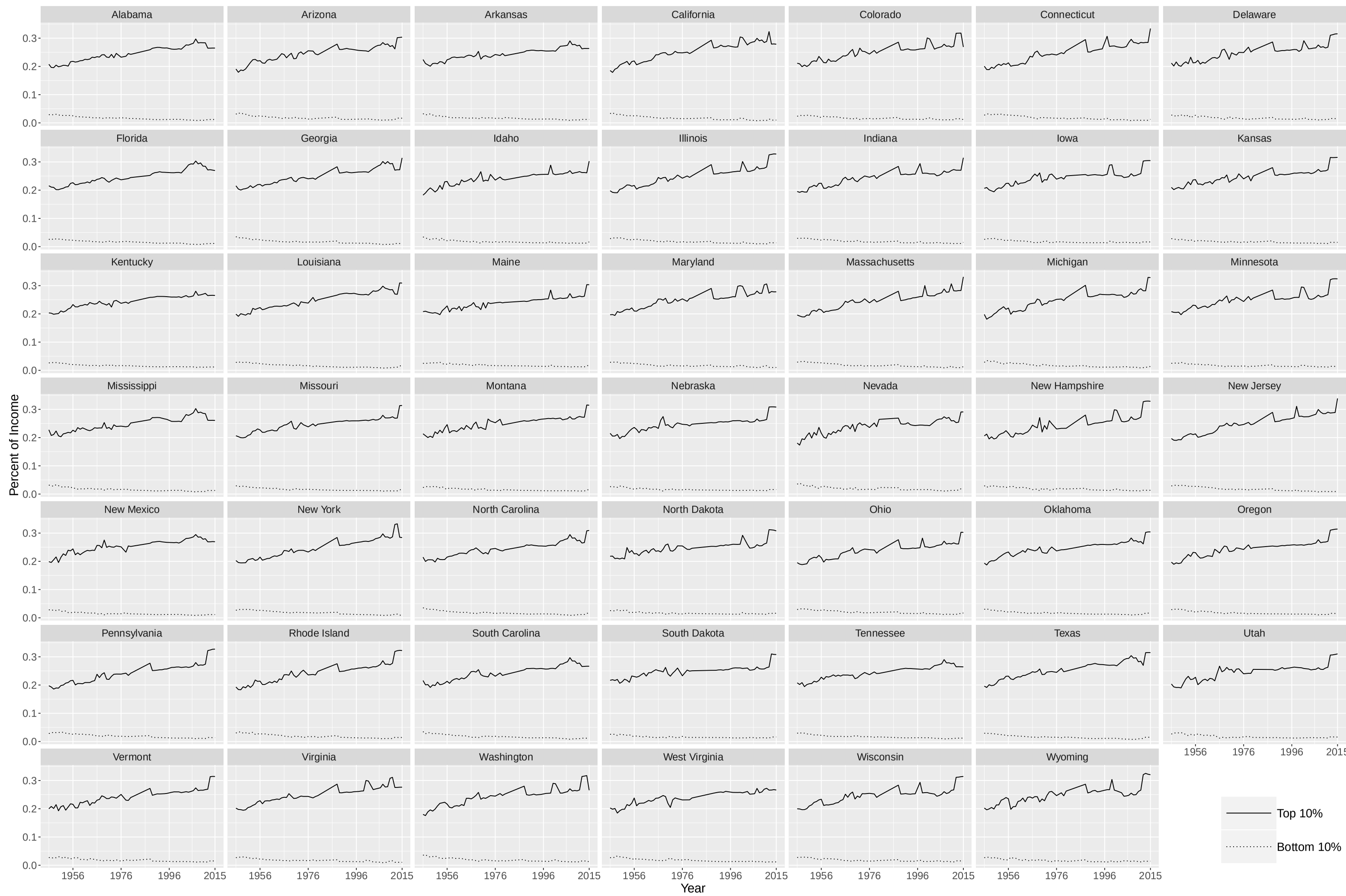


Figure 2.7: Income Share Top and Bottom Population Decile (State) 1946-2015
Percent of total income held by the top and bottom 10% of earners of interest (5-95% of total population) for the contiguous 48 states. Calculated using least squares method, using Jensen-Shannon divergence to select distribution by state and year.

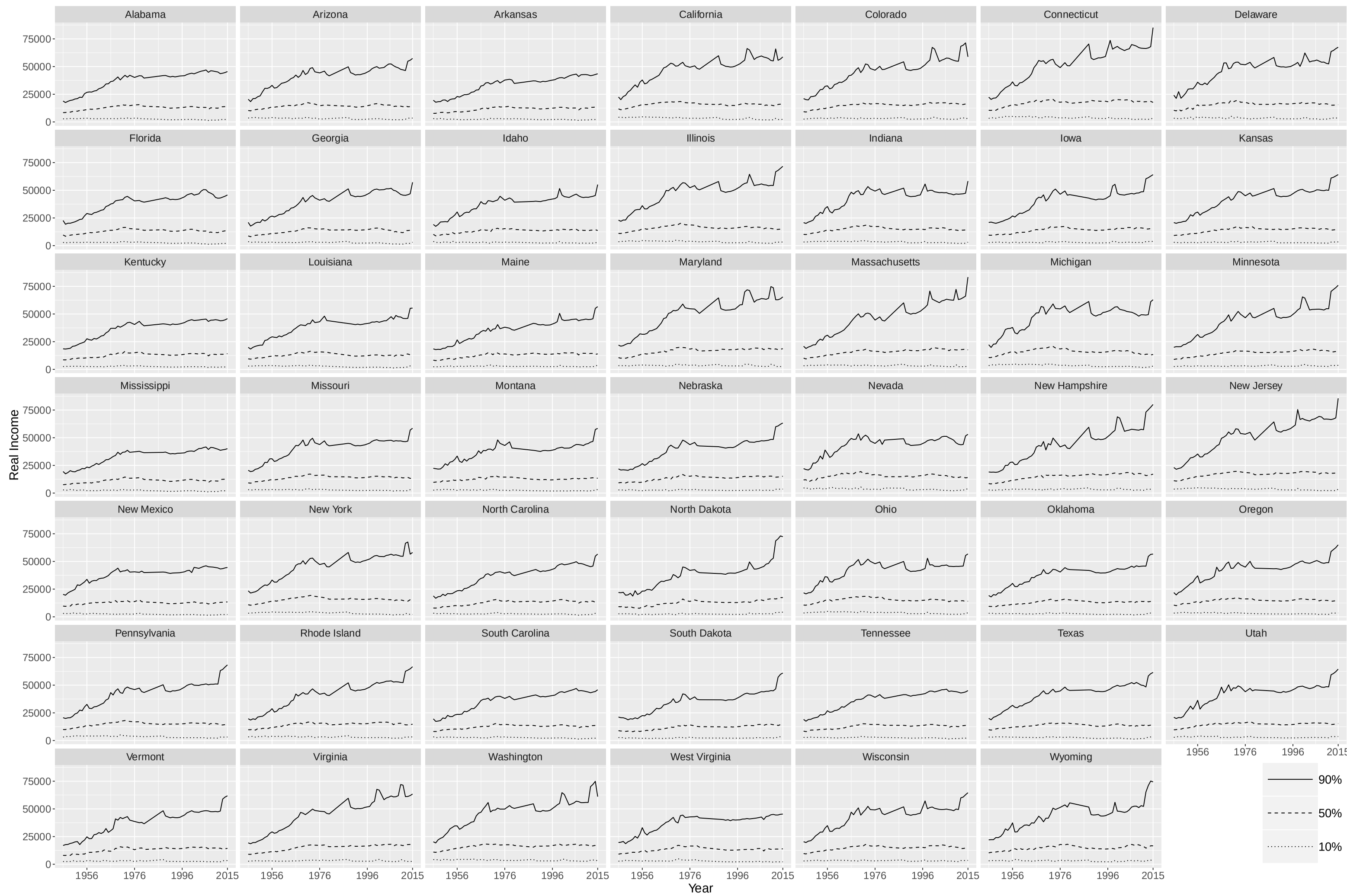


Figure 2.8: Real 10$^{\text{th}}$, 50$^{\text{th}}$,90$^{\text{th}}$ Percentile Income (State) 1946-2015
Real earnings at the 10$^{\text{th}}$, 50$^{\text{th}}$, 90$^{\text{th}}$ percentile of the population in 1982-1984 adjusted dollars, for the contiguous 48 states. Calculated using the least squares method, using Jensen-Shannon divergence to select distribution by state and year.

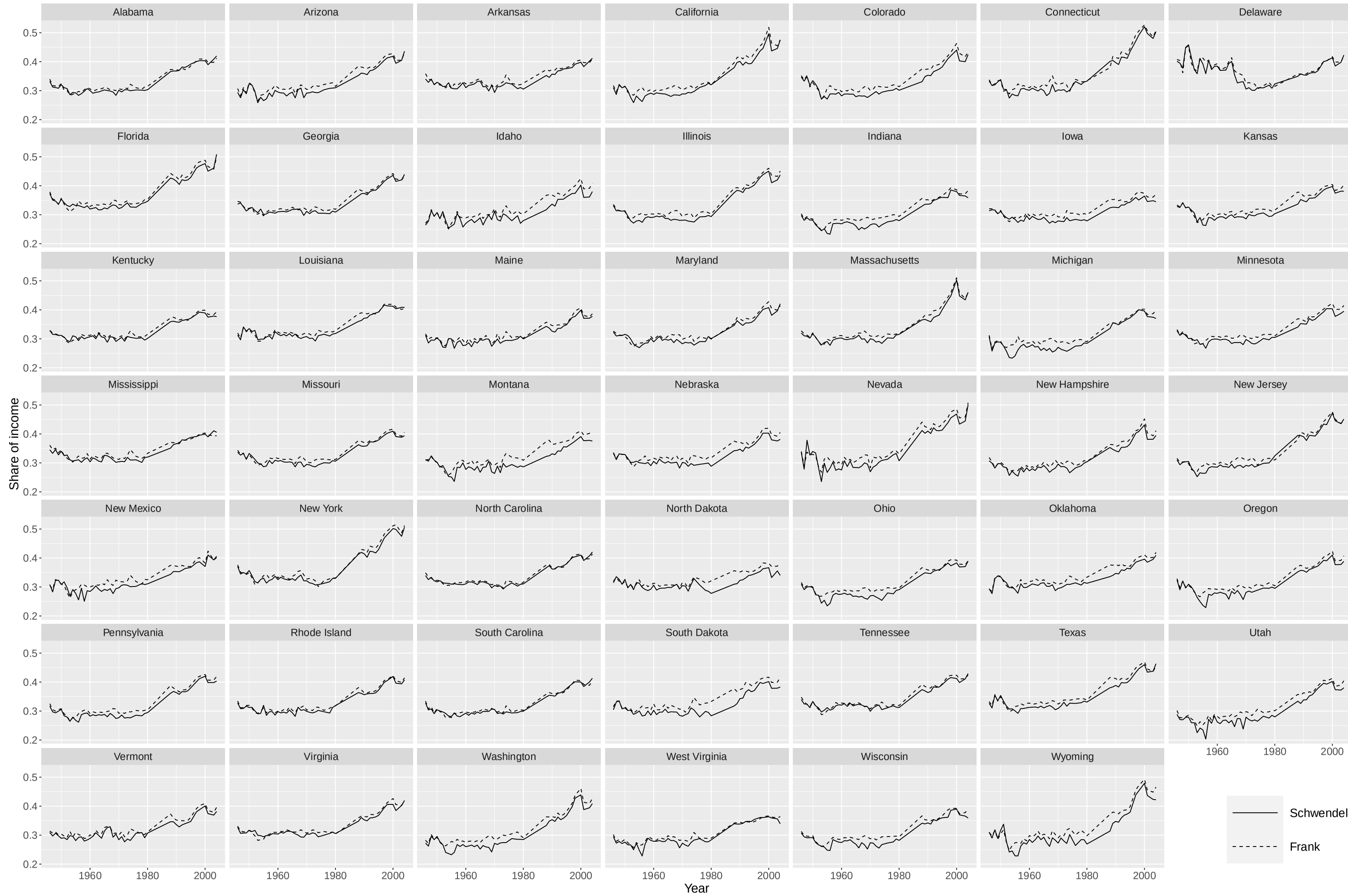


Figure 2.9: Comparison of Frank Series to Schwendel Series 1946-2004
Comparison of the share of income held by the top decile of the population estimated by Frank(2009) to that which results from the estimates within this chapter.

4 References

[1] Atkinson, A.B. (1975) *The Economics of Inequality.* Oxford: Oxford University Press.

[2] Atkinson, A.B. and T. Piketty (2010) *Top Incomes: A Global Perspective.* Oxford University Press.

[3] Autor, D.H. (2014) "Skills, education, and the rise of earnings inequality among 'the other 99 percent'." *Science*, 344: 843-851.

[4] Autor, D.H., A. Manning, and C.L. Smith (2010) "The Contribution of the Minimum Wage to U.S. Wage Inequality over Three Decades: A Reassessment." *NBER Working Papers*, 16533.

[5] Bandourian, R., J.B. McDonald, and R.S. Turley (2003) "Comparison of Parametric Models of Income Distribution across Countries and over Time." *LIS Working Papers*, 305.

[6] Blanchet, T., Fournier, J., and T. Piketty (2017) "General Pareto Curves: Theory and Applications." *WID Working Paper Series*: 201703.

[7] Budd, E. C. (1970) "Postwar Changes in the Size Distribution of Income in the U.S." *American Economic Review*, 60(2): 247-260.

[8] Card, D., T. Lemieux, and W.C. Riddell (2003) "Unionization and Wage Inequality: A Comparative Study of the U.S., the U.K., and Canada." *NBER Working Papers*, 9473.

[9] Chotikapanich D., and Griffiths W.E. (2008) "Estimating Income Distributions Using a Mixture of Gamma Densities." In: Chotikapanich D. (eds) *Modeling Income Distributions and Lorenz Curves. Economic Studies in Equality, Social Exclusion and Well-Being,* vol 5. Springer, New York, NY.

[10] Cilke, J. (1998) "A Profile of Non-filers." *U.S. Department of the Treasury, OTA paper*, 78.

[11] Cowell, F.A. (1995) *Measuring Inequality. London School of Economics Handooks Series.* 2nd ed. New York: Prentice Hall, Harvester Wheatsheaf

[12] Cowell, F.A., and F. Mehta (1982) "The Estimation and Interpolation of Inequality Measures." *The Review of Economic Studies*, 49(2): 273-290.

[13] Dikhanov, Y., and M. Ward (2001) Evolution of the Global Distribution of Income, 197099. *Session of International Statistical Institute*, 53.

[14] DiNardo, J., N.M. Fortin, and T. Lemieux (1996) "Labor Market Institutions and the Distribution of Wages, 1973-1992: A Semi-parametric Approach." *Econometrica*, 64(5): 1001-1044.

[15] Feenberg, D.R., and Poterba, J. M. (2000) "The Income and Tax Share of Very High-Income Households, 1960-1995." *American Economic Review*, 90(2): 264-270.

[16] Frank, M.W. (2009) "Inequality and Growth in the United States: Evidence from a New State-Level Panel of Income Inequality Measure." *Economic Inquiry*, 47(1): 55-68.

[17] Frank, M.W. (2014) "A New State-Level Panel of Annual Inequality Measures of the Period 1916-2005." *Journal of Business Strategies*, 31(1): 241-263.

[18] Freeland, W., and S. Hodge (2012) "Tax Equity and the Growth in Non-payers." *The Tax Foundation: Special Report*, 200.

[19] Gibrat, R. (1931) *Les inégalités économiques.* Paris: Sirey (1931).

[20] Goldin, C., and L.F. Katz (1999) "The Returns to Skill in the United States across the Twentieth Century." *NBER Working Paper*: 7126.

[21] Goldin, C., and R.A. Margo (1992) "The Great Compression: The Wage Structure in the United States at Mid-Century." *The Quarterly Journal of Economics*, 107(1): 1-34.

[22] Heitjan, D.F. (1989) "Inference from Grouped Continuous Data: A Review." *Statistical Science*, 4(2): 164-179.

[23] Jäntschi, L., and S. Bolboacă (2017) "Performances of Shannons Entropy Statistic in Assessment of Distribution of Data." *Ovidius University Annals of Chemistry*, 28(2): 30-42.

[24] Johns, A., and J. Slemrod (2010) "The Distribution of Income Tax Noncompliance." *National Tax Journal*, 63(3): 397-418.

[25] Lemieux, T. (2006) "Increasing Residual Wage Inequality: Composition Effects, Noisy Data, or Rising Demand for Skill?" *American Economic Review*, 96(3): 461-498.

[26] Lin, J. (1991) "Divergence Measures Based on the Shannon Entropy." *IEEE Transactions on Information Theory*, 37(1): 145-151.

[27] López J.H., and L. Servén (2006) "A Normal Relationship? Poverty, Growth and Inequality." *World Bank Policy Research Working Paper*: 3814.

[28] Panizza, U. (2002) "Income Inequality and Economic Growth:Evidence from American Data." *Journal of Economic Growth*, 7(1): 25-41

[29] Piketty, T., and S. Emmanuel (2003) "Income Inequality in the United States, 1913-1998." *Quarterly Journal of Economics*, 118(1): 1-39.

[30] Pinkovskiy, M. and X. Sala-i-Martin (2009) "Parametric Estimations of the World Distrubtion of Income." *NBER Working Paper*, 15433.

[31] Rector, R., K. Johnson, and Y. Sarah (1999) "The Extent of Material Hardship and Poverty in the United States." *Review of Social Economy*, 57(3): 351-387.

[32] Salem, A., and T. Mount (1974) "A Convenient Descriptive Model of Income Distribution: The Gamma Density." *Econometrica*, 42: 115-1128.

[33] Schmitt, J. (2003) "Creating a consistent hourly wage series from the Current Population Survey's Outgoing Rotation Group, 1979-2002." *Washington DC: Center for Economic and Policy Research*, Version 0.9.

[34] Seiver, D.A. (1979) "A note on the Measurement of Income Inequality with Interval Data." *The Review of Income and Wealth*, 25(2): 229-233.

[35] Sommeiller, E., and M. Price (2018) "The new gilded age: Income inequality in the U.S. by state, metropolitan area, and county." *Economic Policy Institute*.

[36] Wolfson, M.C. (1994) "When Inequalities Diverge." *American Economic Review*, 84(2): 353-358.

5 Appendices

5.1 Appendix A: Methodology Appendix to Chapter 1

A particular quality of the Theil index, or similar indices such as the Gini, is their scale independence, due to the fact that these indices are based on relative population and income shares. An estimate of the Theil should then exhibit the same ranking of inequality regardless of the scale of income. This poses a problem for any source of data which bins by values of income and not by percentiles of the population. The simulations in this chapter build upon prior work by showing that not only the placement of bins, but the relative crowding within bins can create changes in measures of inequality. Specifically, by increasing or decreasing mean earnings in a given population, concentration per bin is altered. In particular, as nominal income increases or decreases crowding will occur in high and low income bins respectively. As crowding becomes extreme, increasing portions of total population and of total income will be crowded into those bins, causing a variety of technical challenges for many estimation techniques.

The simulations at hand demonstrate conceptually what would happen to the estimates of various measures of inequality, if geographic or temporal observations had the same segments of the population receiving the same percentages of earnings, but *differed* in average earnings. The critical relevance of this exercise rests on the fact that individual incomes within a region or time period may be binned in a uniform structure, defined not by percentages of the population, but arbitrary levels of income. Thus any differences in subsequent estimates would be due solely to the differences in average earnings across observations. It should be also noted that the mechanism for these differences is that average earnings influence the number of individuals in the bins (e.g., more crowding at the top (bottom) bins for higher (lower) income observations) and this in turn influences the performance of various estimation techniques.

For this purpose, income bins from the Internal Revenue Services 2015 Statistics of income will be used. This year represents some of the coarsest bins in the sample period. It thus demonstrates the problem in its most severe form, making these bins ideal for illustrative purposes. Other years provide income data using more closely spaced bins at both the high and low end of the distribution. As shown in section 4 above, performing this analysis on these years will still result in a bias but it will be less severe. For 2015 the lognormal distribution is found to be the most appropriate distribution, both nationally and across states, based on various goodness of fit tests. It should be pointed out however, for other years and states other distributions prevail. When these same simulations are performed using the bin structure and estimated distributions for these years and areas, the trends remain extremely robust.

Given the lognormal distributional assumption, parameters are then chosen to allow bins to be populated in a manner similar to those observed across states in the actual IRS sample for 2015. From this distribution the incomes for a population of 1,000,000 individuals will be generated and from this population, 1,000 random draws are taken representing a sample of 1,000 incomes. Once drawn, each observation of individual incomes in that particular sample is then scaled by a factor of 0.5 to 5 in 0.1 increments.[27] To illustrate, each of the 1000 observed incomes comprising a given sample is first multiplied by 0.5 to create one theoretical population, next the original sample is multiplied by 0.6 to create another population and so on. This effectively creates 46 theoretical regions in which each of the 1000 individual incomes preserve their percentage of total income while average incomes differ. Of course, this also implies that for any given portion of the population chosen, the percentage of income held by that portion of the population will be the same in any of the constructed regions.

Each of these theoretical regions then have their 1,000 observed incomes binned according to the AGI classes discussed. This binned data can then be used to estimate inequality by

[27]Using the bounds 0.5 to 5 allows for a range of scaled incomes even greater than may normally be seen across regions or time. In addition, these bounds were selected as they allow for consistent comparison of all techniques which follow. Each technique requires some minimum number of bins to be populated in order for any estimation to be performed

several methods proposed in the literature for such a purpose. This approach addresses inter-regional and inter-temporal comparability of relative inequality measures created using these methods, with differences in estimated measures across the theoretical states arising from state-level average income differences, rather than from actual distributional differences. As such, each theoretical region will only exhibit different measures of inequality due to the influence that crowding has on a given estimation technique.

Before considering the results, I address possible sampling errors by repeating the above simulation 1,000 times, creating 1,000 samples of 1,000 draws which each in turn are transformed into the 46 theoretical regions as explained. If an estimation technique is consistent, as more samples are drawn, again *barring any effect from binning*, the median inequality index calculated at each multiple (each individual theoretical region) should approach the value calculated for the underlying population which was assumed. This value can be calculated directly from the density of the distribution from which the described samples are drawn.

The entire process described will then be repeated, allowing the scale parameter of the underlying population to vary from 1 to 1.3 in .02 increments. These will allow observed bins to vary in a manner that is similar to that observed across states in data presented in the IRS SOI. It is important to note that the step of multiplying random draws of income is equivalent to allowing the location parameter, assumed to be 3.5 here, to vary. That is to say, changing the location parameter of an underlying lognormal distribution should not change the value of the Theil for that distribution, as long the scale parameter remains constant, again barring any influence the binning procedure may have on the estimation technique. So, theoretical regions could also be created by changing the location parameter, but some intuition about the underlying relative levels of average income would be lost.

5.2 Appendix B: Data Appendix to Chapter 2

Slight variations in the way that AGI data are structured from year to year require some effort to make sure distributional estimates are based on as similar data as possible. Before 1955 individuals with AGI summing to zero or negative values were entirely excluded from the publications. For all the years after this date the IRS either includes these returns in the lowest AGI bracket or includes them as a separate category. It is impossible to perfectly disentangle the sub-categories within these bins, which include both positive and negative AGI. As such, for those years for which the number of tax units with negative AGI is available as a separate measure, those units will be included in the lowest AGI bracket. Because of this, the first ten years may not be entirely comparable to the rest of the sample, as they are the only years lacking data on negative AGI completely, but most of the years will be of similar construction for use in estimation. A separate issue is the presentation of taxable and non-taxable income. For the years 1959 until 1962, taxable and non-taxable returns are presented separately. For these years all AGI brackets of taxable and non-taxable income are combined where possible, excluding any category of non-taxable AGI which does not have an upper bound. Those tax units most likely constitute several taxable categories and require strong assumptions for combining them into overall totals. Due to these factors, income distribution measures for the specific years discussed may also not be fully comparable. In years when negative and positive, or taxable and non-taxable AGI are presented separately, estimations of inequality are made both including and excluding negative and non-taxable AGI for the purpose of comparison. This exercise shows that while some quantitative results slightly change, the qualitative trends in inequality are robust to both treatments.

Additionally, several years are absent from the sample completely, from 1982-1988 the necessary data is not made publicly available in any form.[28] Also, in 1952 the counts for Washington and Alaska are combined, in 1961 Maryland and the District of Columbia are

[28] Additionally, the years 1975, 1977, and 1981 are currently excluded due to low quality scans of the documents from which the data originates, meaning the underlying values cannot be currently verified. 2002 is also excluded, as it is not clear in what manner it deals with negative AGI.

combined, and in 1962 Delaware is not included in the reported IRS SOI data.

John Schwendel

Curriculum Vitae

Fields of Interest

Primary: Labor Economics, Applied Microeconomics, Economics of Inequality
Secondary: Machine Learning, Growth and Development

Education

Ph.D. Economics, University of Wisconsin-Milwaukee (expected May 2020) 2015-2020
M.A. Economics, University of Wisconsin-Milwaukee 2013-2014
B.S. Economics, University of Wisconsin-Madison 2004-2009
Certificate in Folklore, University of Wisconsin-Madison 2004-2009

Research

"The Distribution of Household Income in the United States, 1946-2015" (with Hamid Mohtadi)
"Residual Wage Inequality and Age Composition Effects"
"Skill-Biased Technical Change and Wage Inequality: Challenges in Theory and Data" (Master's Thesis)

Teaching Experience

Graduate Student Instructor, University of Wisconsin-Milwaukee 2016-2020
  Intermediate Microeconomics
  Economics of Personal Finance
  Principles of Microeconomics (online and in person)
  Introductory Economics
Teaching Assistant, University of Wisconsin-Milwaukee 2015-2018
  Principles of Macroeconomics (online and in person)
  International Trade
Guest Lecturer, University of Wisconsin-Milwaukee 2017-2020
  Economic Development (undergraduate and graduate)

Other Positions

Economics Department Graduate Assistant, University of Wisconsin-Milwaukee 2019
Provided outreach for integrated B.S./M.A. Program.
Research Assistant to Antu Murshid, University of Wisconsin-Milwaukee 2018
NSF Grant Study "Identifying Household Wealth Effects"
Economics Department Graduate Assistant, University of Wisconsin-Milwaukee 2016
Compiled teaching evaluation reports, managed class permissions and transfers.

Presentations

Econometric Society North American Summer Meeting, Seattle WA	2019
Midwest Economics Association Annual Conference, St. Louis MO	2019
University of Wisconsin-Milwaukee Microeconomics Workshop, Milwaukee WI	2019
Wisconsin Economics Association Annual Conference, Stevens Point WI	2018
University of Wisconsin-Milwaukee Labor Workshop, Milwaukee WI	2018

Honors & Awards

University of Wisconsin-Milwaukee

Richard Perlman Prize for Outstanding Paper in Labor Economics	2019
Chancellor's Graduate Student Award	2019
Chancellor's Graduate Student Award	2016

References

Hamid Mohtadi
University of Wisconsin-Milwaukee
Department of Economics
Email: Mohtadi@uwm.edu

Scott Drewianka
University of Wisconsin-Milwaukee
Department of Economics
Email: Sdrewian@uwm.edu

John S. Heywood
University of Wisconsin-Milwaukee
Department of Economics
Email: Heywood@uwm.edu

James H. Peoples
University of Wisconsin-Milwaukee
Department of Economics
Email: Peoples@uwm.edu