working paper 9612

Rounding in Earnings Data

by Mark E. Schweitzer and
Eric K. Severance–Lossin

FEDERAL RESERVE BANK OF CLEVELAND

Rounding in Earnings Data

by Mark E. Schweitzer
and Eric K. Severance-Lossin

Mark Schweitzer is an economist at the Federal Reserve Bank of Cleveland. Eric Severance-Lossin is a consultant at the Federal Reserve Bank of Cleveland.

December, 1996

Abstract

Earnings data are often reported in round numbers. In fact, in the March 1995 Current Population Survey (CPS), 71% of all full-time earnings responses are some multiple of $1,000. Rounding is typically ignored in analyses of earnings data, which effectively treats it as noise in the data. Our GMM estimates of a simple model of rounding indicate that this behavior is highly systematic and correlated with the respondents' earnings level. We find that the systematic nature of rounding can affect some commonly used statistics based on earnings data. The statistics we investigate in this analysis are inequality summary measures, earnings quantiles, kernel density estimates, and frequency plots of wage adjustments. We find that rounding alters most of these statistics substantially, that is, by more than the typical level of annual changes or reported standard errors.

1. Introduction

Rounding in reported earnings data may be an arcane issue; but it is one that has potentially large impacts on widely used income statistics. In particular, rounding can affect statistics that focus on specific points of the wage distribution or account for noncentral changes in the wage distribution, including most inequality measures. It also builds a nominal component into wages and wage changes, since rounding points are nominally oriented. This paper investigates the occurrence of gross rounding using a generalized method of moments (GMM) estimator. The estimated rounding model is used to back out the implied distribution of underlying unrounded wages. This distribution allows the impacts of rounding on other statistics of interest to be identified.

To the best of our knowledge, the only applicable work on rounding in economics literature is Hausman, Lo and MacKinlay's (1992) application of the ordered probit model to stock transaction prices. Both the model presented in this paper and theirs include a parametric rounding process. However, since we observe both rounded and unrounded observations in earnings data we are not forced to parameterize the underlying distribution. In transaction price data all observations are rounded and a parameterization of the underlying data generating process is needed to identify the model. We are able to allow the underlying data generation process to be arbitrary (up to smoothness restrictions) introducing greater flexibility into our model.

The approach employed in this analysis is general; however, our empirical work focuses on one of the primary sources of income data in the United States: the Annual

Demographic Supplements to the Current Population Survey (CPS). The qualitative results of our analysis probably apply to most data on incomes, because the CPS was the archetype for many later surveys, but the quantitative levels are likely to vary with specific questions and interview settings.

We find strong evidence of rounding that varies with income. At the level of rounding we consider, rounding is always a statistically significant phenomenon, although the amount of rounding at the nominal level we model has risen substantially over time. The statistics we focus on are also altered by the rounding process. Notably, earnings quantiles shift well beyond the conventional confidence boundaries. Three inequality measures are also tested, though these generally shift less in relation to their standard errors. Kernel density estimation is also affected by rounding, in that rounding points become modes at asymptotically optimal bandwidths. While we have not investigated all commonly used income statistics, our results suggest a general guideline that researchers should be concerned about possible effects of nominal rounding when working with statistics that involve minimal averaging.

2. Rounding in Earnings Data

All earnings data are rounded, at least at a low level – the data are only available to the dollar on an annual level. This is unlikely to be a problem for most analyses, because

this level of rounding is well within the desired level of precision for economic research. However, rounding also appears to occur over much broader ranges; i.e., individuals round up or down to thousands. Indeed, spikes at multiples of $1,000, $5,000 and $10,000 are quite large in this data set. This level of rounding, which is the focus of this paper, could be highly problematic.

The data set we use to explore characteristics of rounding in earnings data is one of the most heavily used research data sets on earnings in the United States, as well as the source of many government-provided earnings-base statistics. The March CPS earnings question is prototypical of earnings questions directed towards individuals. From1964 to 1979 the question was generally: "Last year (19XX) how much did {worker} receive in wages or salary before any deductions?" After 1980, the earnings questions attempt to include two sorts of information, first on primary and then on secondary jobs. The results of this two-stage procedure do not appear to be very different, but we have to focus on primary earning after 1986. The earnings section of the questionnaire (more formally known as the Annual Demographic Supplement) is run only once a year, not coincidentally near tax filing time, and is part of an extended interview process that concentrates on the employment situations of households that are surveyed eight times over a two-year period (if the residents do not move during that period). The survey is conducted either in person or by telephone (after an initial in-person survey) by a trained Census Bureau interviewer. The data are topcoded at nominal levels that are changed infrequently. Because neither topcoding nor high-income workers are our focus, we sidestep topcoding issues by considering only the central 96% of the distribution in all years.

In the most recent year (1994 earnings, surveyed in 1995), rounding to multiples $1,000 per year have become exceptionally common; 71% of observations are reported this way. Multiples of $100s of dollars per week are also reported frequently. Figure 1 shows a histogram of major spikes in the data set. The binwidth is $1, the highest level of precision available in the data set. Other than the fact that rounding is certainly present, few characteristics of the rounding process can be determined from this diagram.

Researchers have been concerned about measurement error in the CPS data set, although rounding was never the specific focus of this concern. Mellow and Sider (1983) matched CPS data with employer-reported records to measure the disagreement of the two data sources, without positing that one source is more accurate than the other. The general conclusion is that "there is no operational differences between the quality of earnings data obtained from workers in a household survey compared to what would be obtained directly from their employers." Like other research in this area, these analyses are typically treat error as non-systematic, while rounding is highly systematic.[1]

Lillard, Smith, and Welch (1986) were concerned about systematic errors implied by the Census Bureau's hot-deck imputation of earnings for nonrespondents. Hot decking recreates rounded observations at approximately the frequency with which they occur in the reported data, so this issue is largely orthogonal to rounding.

We focus on rounding in many years of data, and thus cannot work with a matched administrative data set. Finally, our focus is on nominal rounding, whether instigated by

[1] Two papers that used extensions to the Michigan Panel Study of Income Dynamics (Rodgers, Brown, and Duncan (1993) and Bound, Brown, Duncan, and Rodgers (1994)) are more detailed in their analysis, but also focus on unsystematic errors.

either the interviewee in error or by the employer; thus, a correctly reported yet rounded wage is still of interest for this analysis.

Not distinguishing between employer- and employee-based rounding may seem overly limiting and is, in fact, partially due to data limitation. However, most models of wages are based on real wage rates, which are more accurately represented by the latent unrounded wage. One who is truly interested in the rounded nominal wage would need additional data to distinguish between employer- and employee-based rounding.

3. A Statistical Model of Rounding

In order to model the type of rounding discussed in the previous section, we consider the following model. Let z be a random variable with density $\pi(z)$. Although we are interested in various characteristics of z, we are only able to observe a random variable x which is a rounded version of z. The observed random variable, x, is related to the latent variable, z, by

$$x = \left\{ \begin{array}{ll} k_r & with\ prob\ p_r(z, \Theta) \\ z & with\ prob\ 1 - \sum_{r=1}^{R} p_r(z, \Theta) \end{array} \right\}, \tag{1}$$

where $k_r, r = 1, ...R$ are rounding points and $\Theta \in I\!R^J$ is a vector of parameters. Let $\Psi(x)$ be the cumulative distribution function (cdf) for x, and $\psi(x)$ be the probability density

function (pdf) for x conditional on $x \notin \{k_1, ..., k_R\}$, so that

$$\pi(z) = \frac{\psi(x)}{1 - \sum_{r=1}^{R} p_r(x, \Theta)} \tag{2}$$

provided that $\sum_{r=1}^{R} p_r(z, \Theta) < 1 \; \forall z$. In order to get information about the random variable of interest, z, we first estimate the $p_r(z, \Theta)$'s and then use these estimates to adjust the distribution of the observed unrounded points $\Psi(x)$.

3.1 Moment Equations

The parameters Θ are estimated via a method of moments approach. Moment equations are generated by noting that

$$\begin{aligned} E[x = k_r] &= E[p_r(z, \Theta)] \\ &= E\left[\frac{p_r(x,\Theta)}{1 - \sum_{r=1}^{R} p_r(x,\Theta)} \mid x \notin \{k_i\}\right] \\ &= E\left[\frac{p_r(x,\Theta)(1 - \mathbf{1}[x \in \{k_i\}])}{1 - \sum_{r=1}^{R} p_r(x,\Theta)}\right]. \end{aligned} \tag{3}$$

Equating the first and last expressions in (3) allows us to write the moment equation in terms of the empirical distribution of the observed variable, x. With $g_r(x, \Theta)$ defined by

$$\begin{aligned} \int \mathbf{1}[x = k_r] - \frac{p_r(x, \Theta)(1 - \mathbf{1}[x \in \{k_i\}])}{1 - \sum_{r=1}^{R} p_r(x, \Theta)} \, d\Psi &\equiv \int g_r(x, \Theta) \, d\Psi = 0, \\ r &= 1, ..., R, \end{aligned} \tag{4}$$

Θ is estimated by

$$\widehat{\Theta} = \arg\min_{\Theta} T\left(\int g\left(x,\Theta\right) d\Psi_n\right), \tag{5}$$

where n is the sample size, Ψ_n is the empirical distribution of x, $g\left(x,\Theta\right) = [g_1\left(x,\Theta\right), ..., g_R\left(x,\Theta\right)]$ and $T : I\!R^R \rightarrow I\!R$ is twice continuously differentiable with $T\left(0\right) = 0, T\left(\Theta\right) \neq 0$ for $\Theta \neq 0$, and $\Delta \equiv \frac{\partial^2 T}{\partial\Theta\partial\Theta'}$ positive definite for all Θ. Under some regularity conditions (see Manski [1988]), the estimate given by (5) is asymptotically unbiased and normally distributed :

$$\sqrt{n}\left(\widehat{\Theta} - \Theta\right) \xrightarrow{D} N\left(0, \left(\Omega'\Delta\Omega\right)^{-1}\Omega'\Delta\Sigma\Delta\Omega\left(\Omega'\Delta\Omega\right)^{-1}\right), \tag{6}$$

where $\Sigma = E\left[g\left(x,\Theta\right)g\left(x,\Theta\right)'\right]$, and $\Omega = E\left[\frac{\partial g(x,\Theta)}{\partial\Theta}\right]$. Choosing $T\left(\cdot\right)$ such that $\Delta = \Sigma^{-1}$ minimizes $n\left(\widehat{\Theta} - \Theta\right)'\left(\widehat{\Theta} - \Theta\right)$ and simplifies (6) as

$$\sqrt{n}\left(\widehat{\Theta} - \Theta\right) \xrightarrow{D} N\left(0, \left(\Omega'\Sigma^{-1}\Omega\right)^{-1}\right). \tag{7}$$

Although Σ is not observed in practice, one can take

$$T\left(z\right) = z'\widehat{\Sigma}^{-1}z, \tag{8}$$

where $\widehat{\Sigma}$ is a consistent estimator of Σ, without changing the first order asymptotics. We use a two-step procedure that first estimates the parameters by taking $T\left(z\right) = z'Iz$, where I is the identity matrix, and then estimates Σ based on these parameters. Final estimates for Θ are obtained by estimating the model again using the $T\left(\cdot\right)$ given by (8).

3.2 Parameterization and Identification

In our empirical work we consider the following parameterization of $p_r(z, \Theta)$. Let a_q, $q = 1, ..., Q$ be disjoint intervals in $\mathbb{R}$ and k_r, $r = 1, ...R$ be rounding points, so that the x is related to the latent variable, z, by

$$x = \left\{ \begin{array}{ll} k_r & with\ prob\ p_{qr} \quad if\ z \in a_q \\ z & with\ prob\ 1 - \sum_{r=1}^{R} p_{qr} \quad if\ z \in a_q \end{array} \right\}, \tag{9}$$

so that

$$p_r(z, \Theta) = p_{qr}(\Theta)\, \mathbf{1}[z \in a_q] = p_{qr} \mathbf{1}[z \in a_q]. \tag{10}$$

To save on notation, the dependence of p_{qr} on Θ is suppressed. One could think of the model with $Q = R, k_i \in a_i$ and $p_{ij} = 0$ for $i \neq j$, so that each interval contains a rounding point and p_{ii} is the probability of a point in a_i being rounded to k_i. The model (9), however, is flexible enough to capture the more complicated, and hopefully more realistic, rounding structures considered in the next section. The probability density function, $\psi(x)$, of x conditional on $x \notin \{k_1, ..., k_R\}$, is given by

$$\pi(z) = \frac{\psi(z)}{1 - \sum_{r=1}^{R} p_{qr}} \qquad z \in a_q, \tag{11}$$

provided $\sum_{r=1}^{R} p_{qr} < 1$. The p_{qr}'s are functions of an unknown J dimensional vector of parameters, $\Theta = \{\theta_1, ..., \theta_J\}$. In addition, we assume that the linear map $\left[\frac{\partial p_{qr}}{\partial \theta_j}\right] : \mathbb{R}^J \rightarrow$

$\mathbb{R}^{QR}$ has rank J and is bounded. Then

$$g_r(x,\Theta) = \mathbf{1}[x = k_r] - \frac{\sum_{q=1}^{Q} p_{qr}\mathbf{1}[x \in a_q]}{1 - \sum_{r=1}^{R} p_{qr}}, \qquad r = 1, ..., R. \tag{12}$$

Under the assumption that $1 - \sum_{r=1}^{R} p_{qr} > 0, \forall q$ so that there is a positive probability of observing an unrounded point in each interval, and our assumptions on $\left[\frac{\partial p_{qr}}{\partial \theta_j}\right]$, most of the required regularity conditions given in Manski [1988] are trivially verified. The only conditions which require verification are the identification of Θ and the requirement that Ω has full rank. Both of these are consequences of the following proposition, whose proof is given in the appendix.

Proposition: *Under the above assumptions,* Ω *has rank* J.

The full rank of Ω is necessary for (6) to hold and is also sufficient for the identification of Θ. Identification follows directly from the implicit function theorem, by noting that $\frac{\partial E[g]}{\partial \Theta} = E\left[\frac{\partial g}{\partial \Theta}\right] = \Omega$ is full rank.

3.3 Statistics Based on the Latent Variable

A model of rounding is of little interest in itself, unless statistics based on the latent variable

can be estimated. Quantities of the form

$$F(\Pi) = \int f(z)\, d\Pi \tag{13}$$

are often of interest. In the context of this paper $F(\Pi)$ is a component of an inequality measure. For example the variance of log-wages can be written as $F_2(\Pi_w) - [F_1(\Pi_w)]^2$, where Π_w is the distribution of wages and

$$\begin{aligned} F_2(\Pi_w) &= \int \ln(w)^2\, d\Pi_w \\ F_1(\Pi_w) &= \int \ln(w)\, d\Pi_w. \end{aligned} \tag{14}$$

In order to estimate the integral (13), we consider the following additional moment equation

$$\int \mu - f(z)\, d\Pi = \int \frac{(\mu - f(z))\, \mathbf{1}[x \in a_q, x \notin \{k_r\}]\, d\Psi}{1 - \sum_{r=1}^{R} p_{qr}} = 0 \tag{15}$$

for some parameter μ. Note that μ need not be estimated simultaneously with Θ since, given any set of $\{\widehat{p}_{qr}\}$, the empirical version of (15) can be exactly satisfied by taking

$$\widehat{\mu} = \frac{\frac{1}{N}\sum_{i=1}^{N} \frac{f(x_i)\mathbf{1}[x_i \in a_q, x_i \notin \{k_r\}]}{1 - \sum_{r=1}^{R} \widehat{p}_{qr}}}{\frac{1}{N}\sum_{i=1}^{N} \frac{\mathbf{1}[x_i \in a_q, x_i \notin \{k_r\}]}{1 - \sum_{r=1}^{R} \widehat{p}_{qr}}}. \tag{16}$$

The estimates of μ and of Θ, $\left(\widehat{\mu}, \widehat{\Theta}\right)$, are jointly asymptotically normally distributed with mean (μ, Θ) and a variance-covariance matrix that may be calculated in a manner analogous to $\widehat{\Theta}$.

The Other important quantities of interest for determining inequality are the quantiles of Π_w. These are implicitly estimated by taking the appropriate quantile of the estimated cdf of z,

$$\widehat{q}_\alpha = \max\left\{q_\alpha \mid \widehat{\Pi}\left(q_\alpha\right) \leq \alpha\right\}. \tag{17}$$

In order to get standard errors for $\widehat{q}_\alpha$, the usual procedure of inverting (17) is applied, so that

$$\sqrt{N}\left(\widehat{q}_\alpha - q_\alpha\right) \xrightarrow{D} N\left(0, \sigma_\alpha\right), \tag{18}$$

where

$$\sigma_\alpha = \frac{\sqrt{\mathrm{Var}\left[\widehat{\Pi}\left(q_\alpha\right) - \Pi\left(q_\alpha\right)\right]}}{\pi\left(q_\alpha\right)}. \tag{19}$$

Since $\widehat{\Pi}\left(q_\alpha\right)$ can be written in the form of (13) its standard error can be estimated using the above method. The density, $\pi\left(q_\alpha\right)$, can be estimated via kernel density estimation, discussed below.

Another feature of the unobserved latent variable, z, that one might be interested in estimating is its density. An estimate of $\pi\left(z\right)$ is constructed by a slight modification of the usual kernel density estimator of Parzen (1962).

$$\widehat{\pi}\left(z\right) = \frac{\frac{1}{Mh}\sum\limits_{x_i \notin \{k_r\}} K\left(\frac{z - x_i}{h}\right)}{1 - \sum\limits_{r=1}^{R} \widehat{p}_{qr}} \qquad z \in a_q, \tag{20}$$

where M is the number of $x_i \notin \{k_r\}$ and h is a smoothing parameter chosen by the researcher. The estimate given in (20) is just the empirical version of (11) with $\psi\left(z\right)$ replaced by its kernel density estimator. It is an easy exercise to show that the asymptotic

properties of $\widehat{\pi}(z)$ are governed by the numerator and that the faster ($M^{-1/2}$) asymptotic rate of convergence of $\widehat{p}_{qr}$ does not affect the estimate. There is no first-order asymptotic cost to replacing p_{qr} with $\widehat{p}_{qr}$.

When h is selected in such a way that it tends to zero at a rate of $M^{-1/5}$, $\widehat{\pi}(z)$ is asymptotically normally distributed.

$$M^{2/5}\left(1-\sum_{r=1}^{R}\widehat{p}_{qr}\right)\left(\widehat{\pi}(z)-\pi(z)\right)\xrightarrow{D}N(b,v), \tag{21}$$

where,

$$\begin{aligned} b &= \mu_2(K)\,h_0^2\frac{\psi''(z)}{2} \\ \nu &= \|K\|_2^2\,\frac{\psi(z)}{h_0}, \end{aligned} \tag{22}$$

where $M^{1/5}h \to h_0$ as $M \to \infty$. An optimal value for h_0 in terms of minimizing expected mean squared error of $\widehat{\pi}(z)$ can be found by appropriately balancing the bias and variance given in (22). Since $\psi(z)$ and $\psi''(z)$ are unknown many methods have been proposed for selecting h in practice. For a comprehensive review of methods for selecting h in finite samples, see Jones, Marron and Sheather (1996). In our empirical work we use rule-of-thumb selection criteria based on a log-normal distribution.

4. Model Estimates

The model was estimated on full time/full year, CPS weekly earnings data from 1974 to

1994. We estimate probabilities of rounding to multiples of $1,000, $5,000 and $10,000 per year, along with multiples of $100 per week. The specific identification restrictions we employ were chosen to allow a broad variety of rounding phenomena: 1) rounding probabilities for multiples of $1,000 are a log linear function of underlying income; 2) constant rounding probabilities to multiples of $5,000 and $10,000 are estimated separately for rounded incomes of $30,000 or more; and 3) rounding to $10,000, $15,000, $20,000, $25,000, and weekly wage points are separately estimated. Rounding in the baseline year (1994) is estimated for all of these rounding points between $7,000 and $90,000. However, as the estimation goes back in time, rounding points must be dropped and specification estimation points adjusted, because the shrinking range of nominal incomes makes certain rounding points irrelevant.[2]

The results for 1994 earnings indicate that substantial rounding occurs for each of the rounding points specified in the analysis (see table 1). Furthermore, the positive association in rounding behavior and income levels is clear in the log linear 1,000's rounding term (the units are the log of the rounding point in thousands), along with most of the freely parameterized rounding points. Standard errors on the estimated rounding probabilities strongly reject zero probability of rounding for most rounding points. Only rounding to $15,000 and $600 per week are indistinguishable from zero in 1994 earnings at conventional hypothesis testing limits. Several rounding points are statistically significant, despite relative point estimates that indicate that only 1% or 2% of people who could round

[2] From 1977 to 1994 weekly wage points are estimated from $200 to $600 per week, anything less falls below minimum wage rates for full-time workers. Prior to 1977 weekly wage rounding points are for $100 to $500 per week. Similarly, the freely estimated points are for $5,000 to $20,000 prior to 1981.

to that value do. Low weekly wage rate rounding estimates probably reflect the fact that this particular survey encourages answers on an annual basis. These qualitative results persist even when the survey data go back 10 or 20 years, if at generally lower levels of rounding.

At the level of rounding modeled around the 1994 data, rounding declines in earlier years. The overall percentage rounding for five years is shown in lower portion of table 1, along with the minimum and maximum of the data set and the number of observations. The decline in rounding estimates is spread over most parameters, although specific points become more or less important in different years. The standard errors indicate that rounding to most of these levels continues to be statistically significant. In the earlier years (1979 and 1974 in table 1), the range of the data set requires that parameters refer to lower value; ultimately, some are no longer estimable. For example, there are no $5,000 or $10,000 rounding points in 1974 that are not estimated as part of the free rounding parameters for low multiples of $5,000. This certainly suggests the need for the models to be further adjusted, but we maintained the character of the specification for comparisons across years.

Given that the estimated parameters actually imply rounding from overlapping ranges of underlying earnings, it is useful to see the estimated rounding function relative to underlying earnings levels, as in figure 2. The fact that rounding probabilities rise with income is clearer in the presentation, because the combined levels of rounding (1,000's, 5,000's, etc.) indicate that at income levels above $2,500 approximately 80% of the sample will report a rounded wage. Even at the lowest 1994 wage levels substantial rounding occurs

(over 50%). Referring back to table 1, it is clear that the vast majority of rounding initially occurs towards $1,000's, as the only other relevant parameters ($10,000 and $200 per week) add up to only 3% of individuals at relevant income ranges rounding.

Our choices of the model parameterization were shaped around the obvious features of 1994 wage distribution. We view the model, applied back over 20 years, as an experiment on the effects of rounding at lower levels. Figure 2 also demonstrates that the previous years' estimates yield lower rounding levels for all real incomes. Another key difference in earlier years is that identified rounding ranges farther back in history imply larger real shifts in reported income from the underlying income distribution, because rounding is measured nominally. These facts alone make prior years a different experiment from 1994; however, there are also the subtle changes in survey questions and administration that were discussed above. Our strategy is to use these differences to explore the implications of rounding at different points of time and at different levels.

5. Implications of Rounding

The source of rounding effects on earnings statistics is the positioning of the mass points in the earnings density. While the probabilities can be quite large, if the distance between reported earnings and underlying earnings is small, then few statistics will be meaningfully altered. In addition, some statistics allow for errors to offset reducing the role of rounding

(for example, means). We focused on statistics that might be sensitive to subtle movements of mass in ranges of income: earnings inequality measures, quantiles and wage rigidity measures. In each area we cite one or two papers for more details on how these techniques are applied. These papers are cited as positive examples, rather than as a critique, in that rounding has not previously been identified in any of these literatures.

5.1 Inequality Measures

Inequality summary measures are potentially affected by rounding, because these measures weight portions of the distribution differently; thus, shifting weight around in the distribution might alter measured inequality. We choose three inequality measures to evaluate the effects of rounding: the Gini coefficient, variance of log earnings, and Theil's T. We include three measures because measures implicitly weight portions of the wage distribution unevenly. An interesting application of these measures using CPS data is Karoly (1992).

Given the models corrections and the data set, rounding can alter inequality measures by up to 3% of the measured inequality levels for each of these measures (see table 2). Errors are both positive and negative for all inequality measures, although on average the corrected distributions are less inequitable than the raw data. While none of the discrepancies were large enough to shift any trends, they are often as large as typical annual change in inequality or standard error estimates for these measures on a data set this size. The reduction in the amount of rounding corrected in earlier years does not lower the difference

between the corrected and uncorrected figures.

While the model appears to be tightly estimated, it is only useful if the quantities of interest are well estimated. Using techniques shown in equation (16), we estimate standard errors for two of the inequality measures as functionals of the wage distribution.[3] In each of these cases, the procedure yields estimates with tolerably tight confidence intervals. The standard errors are also presented in table 2. Year-to-year changes in quantities of interest are generally not significant, although longer trends can be. The standard errors are not suitable for comparing the corrected and uncorrected inequality measures because of the extremely high correlation between the measures.

Overall, rounding seems to distort neither the qualitative level of inequality nor the trend. However, researchers should be careful when making statements about the change in inequality between two adjoining years, because the effect of rounding differences could yield changes as large as the change seen in this period between adjacent years.

5.2 Quantiles

Quantiles are often used in place of inequality summary measures on the grounds that quantiles provide locational information on the distribution and are robust to aberrant data in the extremes of the distribution. These advantages have led to heavy use of quantiles in the earnings inequality literature (for example in Juhn, Murphy and Pierce [1993]). Quantile regression techniques have also gained more common application, but rely on

[3] The equation for the Gini coefficient is not of the same form, so the standard error cannot be calculated in the same manner.

accurate quantiles (see Buchinsky [1994]). Quantiles from any point in the distribution are potentially susceptible to variation due to rounding because they focus on points in the distribution, allowing small shifts of mass to alter their location substantially. In a data set with a large amount of rounding (over 50%) quantile estimates typically fall directly on rounding points, so subtle changes in the underlying distribution might cause the quantile estimate to shift from one rounding point to another.

Table 3 shows estimate differences due to rounding that are far larger than typical standard error estimates for all of the quantiles we estimated: the 10th percentile, the median, and the 90th percentile. Differences of over 4% were not uncommon, with positive and negative differences about equally frequent.. Again, the typical size of deviations does not decline with the level of estimated rounding declines. The earliest years actually stand out as having some of the largest deviations for each quantile.

Typical standard errors for quantile estimates in data sets of the size of the CPS are quite narrow. For example, using the method of Mood, Graybill and Boes (1963), STATA reports a 95% confidence interval for the 1994 median ($25,000) from $25,000 to $25,561 (a range of only 2.2%).[4] This means that the effects of rounding are potentially a substantial source of mistaken inference. The standard errors of our procedure (shown in table 3) indicate that even after the rounding correction, year-to-year changes are often statistically significant.

Another way to consider the scale of these changes is to note that real wages at the

[4] Mood, Graybill and Boes (1963) assume that the sampling distribution has a continuous cdf; since this is violated by rounding points, it is not appropriate for this data.

median generally change by less than 2% from year to year. Thus, a subtle change in the location of the underlying quantile with respect to rounding points could cause a far larger shift than is typically realized, or would be expected as the result of sampling errors. Likewise, smaller but substantial changes in the unrounded distribution may be absorbed into a single rounding point.

5.3 Density Estimation

Nonparametric density estimates have recently been applied to issues of income distribution by DiNardo, Fortin and Lemieux (1996). Nonparametric density estimation procedures rely on the result that as the sample size rises to infinity and the bandwidth goes to zero, the estimate converges to true density. With rounding, convergence to the underlying wage density will not occur, since the observed variable does not have a continuous pdf. Shrinking the bandwidth towards zero recreates the spikes associated with rounding. In fact, even at fairly large bandwidths, local modes will occur around any substantial rounding point.

The 1994 kernel density estimates for both the corrected and uncorrected samples are shown in figure 3. The uncorrected density has a large number of local modes. Bumps at most $5,000 multiples are particularly evident, but there is little sign of 1,000's rounding because it is largely suppressed by the bandwidth. Beyond their added features, rounding-induced bumps also hide features that may be locally significant, but are not larger than periodic local modes. The corrected density does indicate that distinct local modes may

exist around the $15,000, $20,000 and $35,000 levels of income that would be hard to distinguish from the periodic modes in the uncorrected case.

As with any nonparametric smoothing technique, the results are sensitive to the choice of smoothing parameter. In order to make a reasonable comparison between the uncorrected and rounding-corrected distributions an asymptotically optimal bandwidth was chosen for each, based on the assumption that the underlying distribution was log-normal. We chose the log-normal specification since it resembled the empirical distributions of the data more closely than the traditional normal approximation. Since, to the best of our knowledge, a rule of thumb based on a log-normal distribution has not been presented in the literature, we calculated a rule of thumb based on a log-normal distribution. The asymptotically optimal bandwidth is given by

$$h = \frac{\nu}{\delta}\left\{\frac{\|K\|_2^2}{N\mu_2^2(K)}\left(\frac{9\delta^4 + 20\delta^2 + 12}{32\sqrt{\pi}}\right)e^{(5\delta/2)^2}\right\}^{1/5}, \tag{23}$$

where ν is the geometric mean of the wages and δ is the standard deviation of log-wages. This rule of thumb was used with a Gaussian kernel for the graphs in figure 3.

Any analysis that is interested in frequency or size of modes should be sure to account for any rounding, because the rounding phenomenon is clearly capable of covering features of the density. In addition, researchers should be careful when comparing density estimates that do not have identical rounding patterns, as in comparisons across years or industries.

5.4 Wage Rigidity

Wage rigidity tests are quite different from the preceding statistics, in that they compare two year's worth of data. Several wage rigidity studies have focused on the prominence of a spike at zero wage change.[5] Rounding can directly affect these measures if underlying wage changes are small and the frequency of rounding high. because our data set is not matched, we cannot follow particular workers over time. Instead, we simulate the effects of rounding given our estimates, an assumed level of correlation between individuals' probabilities of rounding, and a wage growth assumption.

We consider the 1994 earnings data and construct an empirical distribution of the latent, unobserved, wage that is based on our estimates.

$$\Pi_n(z) = \frac{\Psi_n(z)}{1 - \sum_{r=1}^{R} \widehat{p}_{qr}} \qquad z \in a_q, \tag{24}$$

where $\Psi_n(z)$ is the empirical distribution of the unrounded observations. Based on this distribution, we construct the distribution of wages, $\Pi_{\gamma n}(z)$, subject to a fixed percentage increase in wages, γ.

$$\Pi_{\gamma n}(z) = \frac{\Psi_n(\gamma z)}{1 - \sum_{r=1}^{R} \widehat{p}_{qr}} \qquad z \in \gamma a_q, \tag{25}$$

In the simulation there is no rigidity in the latent wage. Based on these two distributions we calculate the expected percentage of zero-wage-change observations in the observed, rounded, data. We simulate for various levels of wage increases and under two extreme

[5] See McLaughlin (1994) for a detailed analysis, or Akerlof, Dickens and Perry (1996) for a summary of the literature and potential impacts of wage rigidities.

assumptions about rounding behavior. One assumption is that individuals round independently from year to year. The other is that rounding behavior is perfectly correlated from one year to the next.

The results of these simulations are presented in table 4. Except for large values of γ, there is a substantial amount of wage rigidity in the observed data, while there is no rigidity in the underlying latent wage. These simulation results seem to indicate that one should exercise a great deal of care in controlling for rounding effects when investigating wage rigidity. However, our results cannot be directly applied to any of the existing research because other studies use similar but distinct datasets, and because issues of whether the firm or the individual rounded the wage observation may be important for this topic.

6. Conclusions

Rounding is an extremely prominent phenomenon in Current Population Survey data. Although differences in questions and survey procedures cannot be ignored, we would be surprised if other household surveys did not exhibit similar rounding patterns. Even employer-based surveys may be affected, as some rounding may occur at the firm level, inducing a nominal pattern into the data. Many analyses of wage data could better describe the phenomenon of interest having abstracted from rounding patterns.

We estimated a simple model of rounding in order to construct the implied underlying

wage distributions. While these distributions are quite similar to the raw data in most respects, nonetheless certain statistics are substantially different when rounding points have been eliminated. In particular, quantiles, kernel density estimates, and measures of zero wage change are sometimes altered at levels comparable to annual changes and/or standard errors. While the set of measures we consider here is certainly not complete, our results seem to indicate that one should account for rounding when using statistics based on localized regions of the wage distribution.

References

Akerlof, George A., Dickens, William T., and Perry, George L. (1996), "The Macroeconomics of Low Inflation," *Brookings Papers on Economic Activity* 1, 1-76.

Bound, John, Brown Charles, Duncan, Greg J.,and Rodgers, Willard L.(1994), "Evidence on the Validity of Cross-sectional and Longitudinal Labor Market Data," *Journal of Labor Economics* 12, 345-368.

Buchinsky, Moshe (1994), "Changes in the U.S. Wage Structure 1963-1987: Application of Quantile Regression," *Econometrica* 62, 405-458.

DiNardo, John, Fortin, Nicole M., and Lemieux, Thomas (1996), "Labor Market Institutions and the Distribution of Wages, 1973-1992: A Semiparametric Approach," *Econometrica* 64, 1001-1044.

Hausman, J., Lo, A. and MacKinlay (1992), "An Ordered Probit Analysis of Transaction Stock Prices," *Journal of Financial Economics* 31, 319-379.

Jones, M. C., J. S. Marron and S. J. Sheather (1996), "A Brief Survey of Bandwidth Selection for Density Estimation," *Journal of the American Statistical Association* 91:433, 401-407.

Juhn, Chinhui, Murphy, and Brooks Pierce (1993), "Wage Inequality and the Rise in the Returns to Skill." *Journal of Political Economy*, 101, 410-442.

Karoly, Lynn (1992), "Changes in the Distribution of Individual Earnings in the United States: 1967-1986." *Review of Economics and Statistics*, 74, 107-114.

Lillard, Lee, Smith, James P., and Welch, Finis (1986), "What Do We Really Know about Wages? The Importance of Nonreporting and Census Imputation," *Journal of Political Economy* 94, 489-506.

Manski, C. (1988), *Analog Estimation Methods in Econometrics*, Chapman Hall, New York, NY.

McLaughlin, Kenneth J. (1994), "Rigid Wages?," *Journal of Monetary Economics* 34, 383-414.

Mellow, Wesley, and Sider, Hal (1983), "Accuracy of Response in Labor Market Surveys: Evidence and Implications," *Journal of Labor Economics* 1, 331-344.

Mood, Alexander M., Graybill, Franklin A., and Boes, Duane C. (1963), *Introduction to the Theory of Statistics*, McGraw-Hill, New York, NY.

Parzen, E. (1962), "On Estimation of a Probability Density and Mode," *Annals of Mathematical Statistics*, 35, 1065-76.

Rodgers, Willard L., Brown, C., and Duncan, G.(1993), "Errors in Survey Reports of Earnings, Hours Worked, and Hourly Wages," *Journal of the American Statistical Association*, 88, 1208-1218.

Silverman, B. (1986), *Density Estimation for Statistics and Data Analysis*, Chapman Hall, New York, NY.

Appendix: Proof of Proposition

Let $A_q = E\left[x \in a_q \; and \; x \notin \{k_r\}\right]$ and $S_q = 1 - \sum_{j=1}^{R} p_{jq}$. Consider the mappings $g_1 : I\!R^J \to I\!R^{QR}$ such that

$$g_1(\Theta) = P \equiv (p_{rq})$$

and $g_2 : I\!R^{QR} \to I\!R^R$ such that

$$g_2(P) = E\left[\mathbf{1}\left[x = k_r\right] - \frac{\sum_{q=1}^{Q} p_{qr}\mathbf{1}\left[x \in a_q\right]}{1 - \sum_{r=1}^{R} p_{qr}}\right] \quad r = 1, ..., R.$$

Then

$$\Omega = E\left[g(x, \Theta)\right] = g_2\left(g_1(\Theta)\right),$$

So that

$$\frac{\partial \Omega}{\partial \Theta} = \frac{\partial g_2}{\partial P} \cdot \frac{\partial g_1}{\partial \Theta}.$$

By assumption, $\frac{\partial g_1}{\partial \Theta}$ has rank J. Since $J \leq R$ it suffices to show that $\frac{\partial g_2}{\partial P}$ has rank R. Let $g_2(P) = \left[g_{21}(P), ..., g_{2R}(P)\right]$ so that

$$\frac{\partial g_{2r}}{\partial p_{ij}} = \frac{\delta_i^r \delta_j^q A_q}{S_q} + \frac{\delta_j^q A_q}{S_q^2} = \frac{\delta_j^q A_q}{S_q^2}\left(\delta_i^r S_q + 1\right).$$

Writing $\frac{\partial g_2}{\partial P}$ in matrix form we have

$$\frac{\partial g_2}{\partial P} = \left[\frac{A_1}{S_1^2}\left(S_1 I + E\right), ..., \frac{A_Q}{S_Q^2}\left(S_Q I + E\right)\right],$$

where I is the $R \times R$ identity matrix and E is an $R \times R$ matrix of ones. The result then follows from the fact that by assumption $A_q \neq 0, \mathrm{S}_q > 0 \; \forall q$ and $\gamma I + E$ is invertible for

$\gamma \neq 0, -R$. It can be verified that $(\gamma I + E)^{-1} = \frac{1}{\gamma} I - \frac{1}{\gamma(\gamma+R)} E$.

Table 1: Model Estimates and Data Characteristics

Parameters	1994	1989	1984	1979	1974
1000's constant	0.283	0.337	0.159	0.056	0.012
	(0.016)	(0.015)	(0.009)	(0.010)	(0.010)
1000's trend	0.097	0.082	0.128	0.151	0.147
	(0.006)	(0.005)	(0.004)	(0.004)	(0.005)
5000's	0.157	0.149	0.145	0.088	
	(0.007)	(0.007)	(0.009)	(0.008)	
10,000's	0.040	0.023	0.000	0.008	
	(0.006)	(0.005)	(0.006)	(0.006)	
$5,000				0.005	0.006
				(0.004)	(0.003)
$10,000	0.020	0.002	0.042	0.020	0.030
	(0.009)	(0.007)	(0.010)	(0.010)	(0.003)
$15,000	0.000	0.025	0.061	0.046	0.061
	(0.011)	(0.010)	(0.005)	(0.004)	(0.005)
$20,000	0.062	0.088	0.114	0.061	0.123
	(0.007)	(0.006)	(0.011)	(0.011)	(0.009)
$25,000	0.091	0.101	0.105		
	(0.011)	(0.009)	(0.007)		
$100/wk					0.022
					(0.001)
$200/wk	0.011	0.006	0.021	0.013	0.018
	(0.003)	(0.002)	(0.002)	(0.001)	(0.001)
$300/wk	0.036	0.017	0.019	0.012	0.011
	(0.003)	(0.002)	(0.001)	(0.001)	(0.001)
$400/wk	0.036	0.017	0.011	0.008	0.009
	(0.002)	(0.001)	(0.001)	(0.001)	(0.002)
$500/wk	0.023	0.013	0.034	0.006	
	(0.002)	(0.001)	(0.006)	(0.007)	
$600/wk	0.006	0.013	0.011	0.008	
	(0.005)	(0.005)	(0.002)	(0.003)	
Percentage Rounded	76.15	72.41	63.33	49.75	38.89
Total Observations	44128	47082	44404	49086	33950
Minimum	7000	6000	5000	3608	1880
Maximum	90000	70000	52000	37000	26400

Source: Authors' Calculations.

Table 2: Effects of Rounding on Inequality Measures

Year	Gini		Theil's T			Variance of Log Wages		
	Uncorrected	Corrected	Uncorrected	Corrected	Std. Err.	Uncorrected	Corrected	Std. Err.
1994	0.3003	0.2992	0.1425	0.1410	0.0256	0.3041	0.3040	0.0093
1993	0.2923	0.2868	0.1344	0.1292	0.0234	0.2923	0.2860	0.0102
1992	0.2863	0.2841	0.1287	0.1269	0.0215	0.2820	0.2791	0.0099
1991	0.2845	0.2831	0.1269	0.1256	0.0182	0.2777	0.2804	0.0101
1990	0.2843	0.2791	0.1267	0.1217	0.0165	0.2776	0.2724	0.0101
1989	0.2851	0.2810	0.1274	0.1237	0.0161	0.2811	0.2756	0.0104
1988	0.2823	0.2819	0.1247	0.1242	0.0173	0.2772	0.2796	0.0115
1987	0.2834	0.2815	0.1256	0.1236	0.0149	0.2770	0.2766	0.0114
1986	0.2840	0.2825	0.1263	0.1249	0.0125	0.2777	0.2777	0.0114
1985	0.2804	0.2802	0.1229	0.1227	0.0112	0.2692	0.2715	0.0114
1984	0.2786	0.2780	0.1212	0.1204	0.0107	0.2667	0.2667	0.0115
1983	0.2745	0.2733	0.1178	0.1164	0.0100	0.2554	0.2533	0.0122
1982	0.2719	0.2746	0.1156	0.1179	0.0097	0.2483	0.2518	0.0127
1981	0.2682	0.2700	0.1121	0.1134	0.0086	0.2435	0.2470	0.0128
1980	0.2663	0.2681	0.1105	0.1117	0.0073	0.2390	0.2428	0.0127
1979	0.2667	0.2675	0.1108	0.1113	0.0063	0.2424	0.2428	0.0127
1978	0.2660	0.2665	0.1102	0.1104	0.0061	0.2400	0.2399	0.0148
1977	0.2653	0.2664	0.1098	0.1104	0.0060	0.2410	0.2433	0.0154
1976	0.2606	0.2579	0.1058	0.1035	0.0048	0.2329	0.2265	0.0148
1975	0.2584	0.2554	0.1042	0.1015	0.0049	0.2306	0.2210	0.0174
1974	0.2728	0.2678	0.1175	0.1129	0.0052	0.2728	0.2586	0.0189

Source: Authors' Calculations.

Table 3: Effects of Rounding on Quantiles

Year	10 Percentile			Median			90 Percentile		
	Uncorrected	Corrected	Std. Err.	Uncorrected	Corrected	Std. Err.	Uncorrected	Corrected	Std. Err.
1994	11000	11240	130	25025	26100	310	55000	56400	900
1993	11000	10820	140	25000	25300	320	52000	51400	660
1992	10700	10600	120	25000	24840	260	50000	49800	690
1991	10150	10200	100	24000	24300	260	49000	48330	490
1990	10000	9800	100	23000	23050	240	47216	45760	390
1989	10000	9590	90	22000	22500	240	45000	44480	410
1988	9300	9150	90	21000	21520	270	43000	42890	390
1987	9000	8800	80	20000	20410	210	41600	41900	330
1986	8782	8500	60	20000	19500	150	40000	39700	270
1985	8500	8400	60	19000	18720	130	38812	38430	250
1984	8000	8070	60	18000	17920	120	36000	36420	260
1983	8000	7800	50	17000	17300	100	35000	34400	240
1982	7800	7700	50	16320	16500	100	32700	32940	240
1981	7280	7400	40	15442	15500	90	30000	30320	170
1980	6975	6800	40	14200	14400	80	28000	28020	160
1979	6035	6300	30	13000	13200	60	25116	25610	120
1978	5720	5720	30	12000	12200	60	23910	23650	120
1977	5200	5270	30	11200	11330	60	22000	22040	100
1976	5000	5100	30	10565	10500	50	20000	19900	80
1975	4680	4680	30	10000	9710	40	19000	18580	70
1974	4000	4100	30	9200	8940	40	18000	17500	80

Source: Authors' Calculations.

Table 4: Wage Rigidity Simulation

Wage Growth Rate	Perfect Correlation	Uncorrelated
1%	63.1%	33.5%
2%	49.8%	25.6%
3%	41.7%	21.1%
4%	36.2%	18.0%
5%	32.5%	15.9%
6%	28.2%	13.6%
7%	25.5%	12.1%
8%	23.5%	11.0%
9%	21.1%	9.7%
10%	18.8%	8.4%
11%	17.1%	7.4%
12%	15.1%	6.2%
13%	14.2%	5.8%
14%	13.2%	5.4%
15%	12.4%	5.1%
16%	11.3%	4.5%
17%	10.3%	3.9%
18%	8.9%	3.2%
19%	8.1%	2.8%

Source: Authors' Calculations.

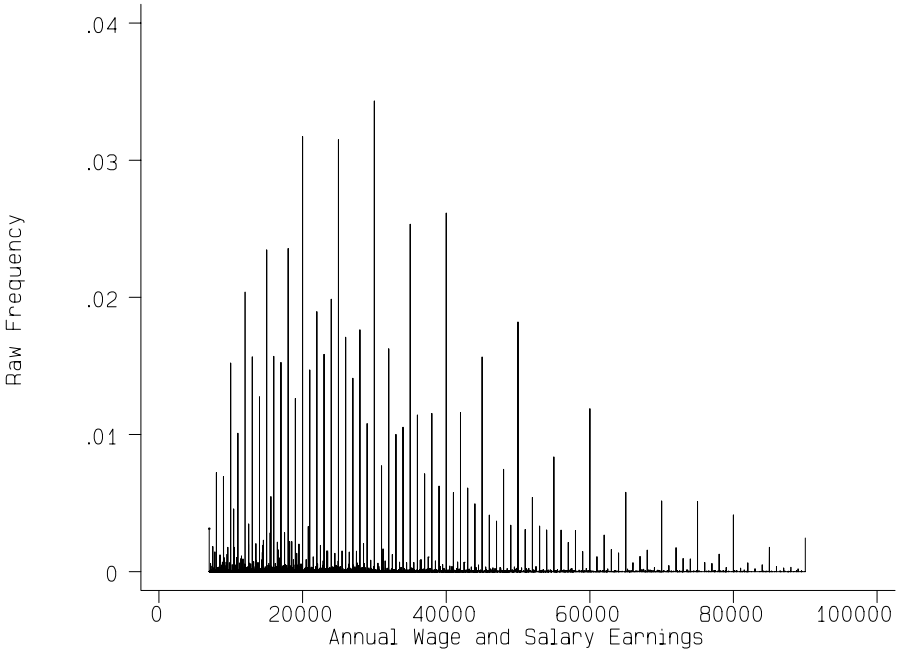

Figure 1: Histogram of 1994 Annual Earnings. Source: Authors' Calculations.

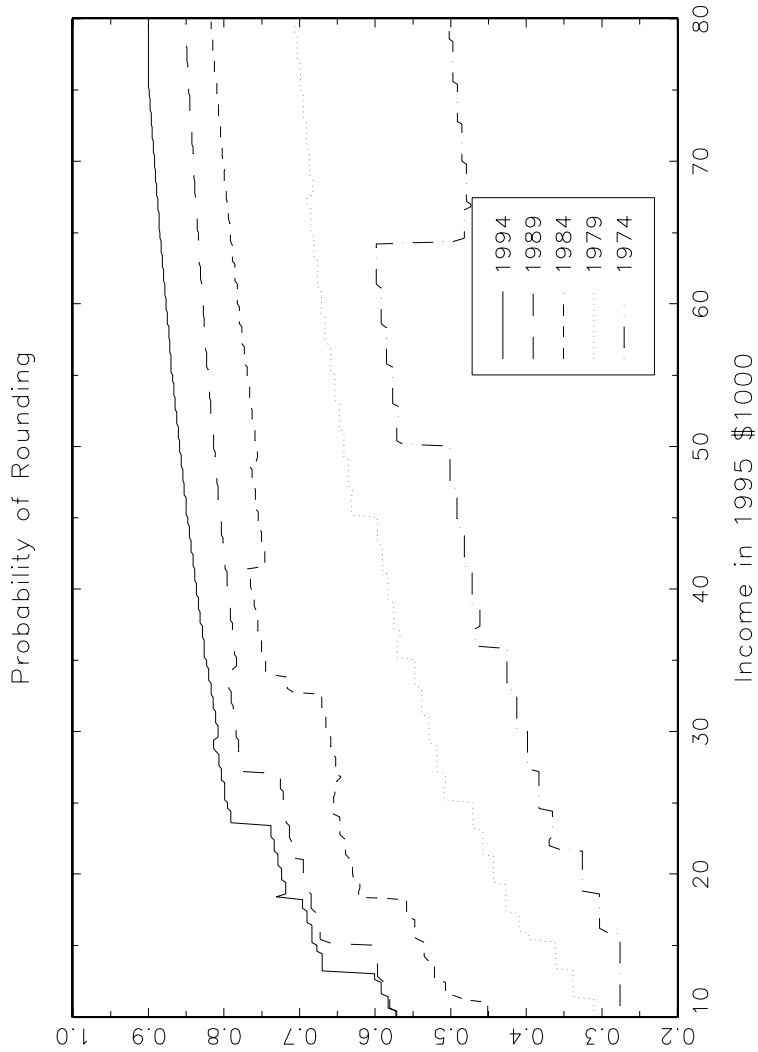

Figure 2: Rounding probabilities as a function of income. Source: Authors' Calculations.

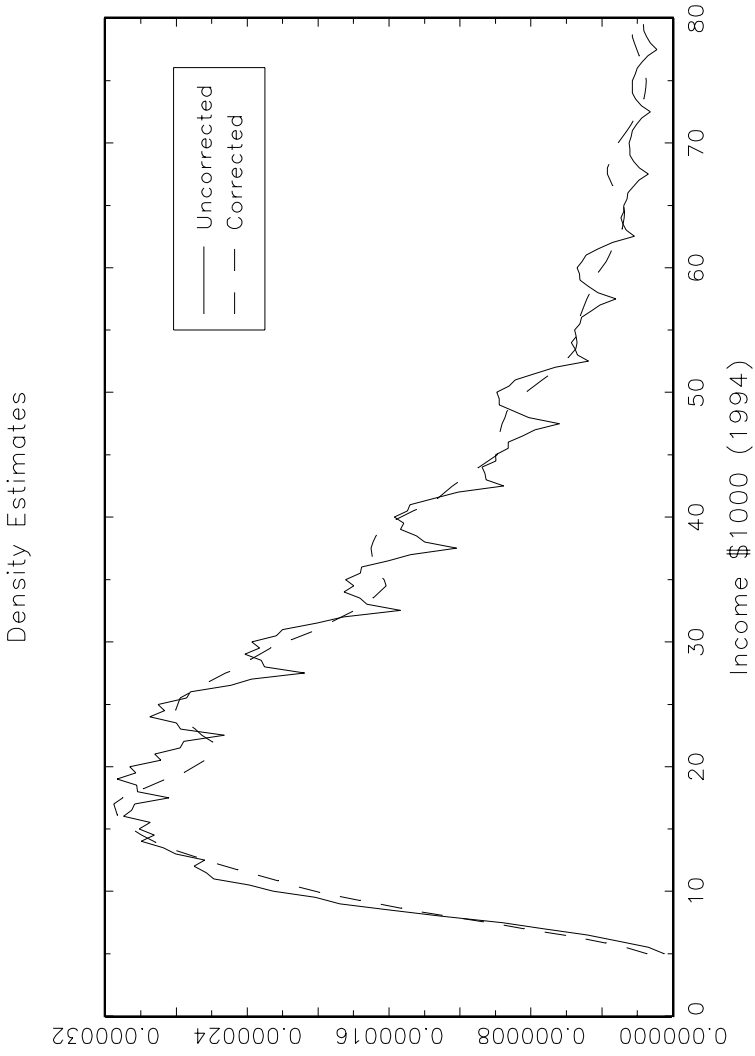

Figure 3: Nonparametric Density Estimates. Source: Authors' Calculations.