An Empiricist's Guide to Nonparametric Analysis in Accounting

This study provides an overview of several nonparametric statistical techniques and how they can be applied to accounting research. Nonparametric estimation allows researchers flexibility in their analysis by removing restrictive assumptions on the relation between variables of interest. In recent years, academic accounting researchers have increasingly begun to use statistical techniques other than traditional ordinary least square (OLS) regressions. This is, in large part, due to pressure by academic researchers to document not only average associations, but also to make causal claims that are valid over the entire sample (Gow, Larcker, and Reiss 2016). In this paper, we suggest that two statistical tools commonly employed in economics and finance research, nonparametric estimation of a kernel density function and locally weighted regressions, may be useful in accounting research as well. The goal of our study is not to derive the statistical properties of these techniques.[1] Rather, we provide insight into the properties, applications, and limitations of these techniques to aid in their application to empirical accounting research.

OLS is appealing on a number of dimensions and appropriate for many accounting-related studies. It produces consistent and unbiased estimates under well-known conditions, requires minimal computing power, and its underlying mechanisms are understood within the academic community. Despite these positive properties, a major drawback of OLS is that it only reports the conditional average linear relation between the independent and dependent variables. Nonparametric analysis provides a range of statistical techniques that allows the researcher to document relations between variables, both linear and nonlinear, across the entire distribution of

[1] These can be obtained in econometric textbooks such as Hansen (2014) and Cameron and Trivedi (2005).

the dependent variable. Furthermore, OLS is generally not robust to the inclusion of a small number of outliers, which may be problematic in the large panel data studies often conducted in accounting (e.g., Kraft, Leone, and Wasley (2006)). Several recent papers use more advanced tools such as iterative reweighted least squares, median regression estimates, and matching techniques to assure readers that results are not due to a handful of influential observations (e.g., Dyreng and Lindsey 2009; De Simone 2016; DeFond, Erkens, and Zhang 2016). However, these techniques still only generate a measure of central tendency, which represent important but single key statistics. Nonparametric analysis can be useful in providing the researcher and reader with a much richer set of information by examining associations along the entire distribution of the data.

In this paper, we examine two distinct settings in accounting research that are particularly well suited for the application of nonparametric analysis: 1) effective tax rates in the financial services sector and 2) the relation between firm size and audit fees. We choose these diverse settings to demonstrate the flexibility of nonparametric analysis across broad areas of accounting research. Furthermore, prior research has examined these variables at length. We contribute directly to these streams of literature by extending prior findings using univariate analysis and providing new insights through nonparametric techniques. The methods presented in this paper can be directly applied to a host of other settings. For example, behavioral accounting researchers commonly use data that are identically and independently distributed (IID). With IID data, nonparametric analysis can be used to obtain statistical inferences in univariate analysis. Gow et al. (2016) call for more descriptive studies in accounting research; the nonparametric techniques we present can be adapted to meet the needs of most descriptive studies, not just

those using IID data, to provide details on the underlying data that may be interesting in their own right or necessary to aid in research design choices.

We begin our analysis by examining financial service firms and effective tax rates (ETRs). The financial industry is a major sector of the U.S. economy, yet many accounting studies do not include these firms because they have structurally different financial statements and regulatory environments than other types of firms. This exclusion is often warranted when running a pooled regression, particularly when examining items such as leverage or cash holdings that differ widely across financial and non-financial firms. Little accounting research examines financial firms specifically in their tax planning; therefore, our analysis directly contributes to the growing body of tax research that focuses on in-depth descriptive analysis using summary tax measures (e.g., Dyreng et al. (2017); Chen, Koester, and Shevlin (2017); Drake, Lusch, and Hamilton (2017)).[2] Nonparametric analysis is well suited for univariate analysis, allowing for a natural extension of these studies. While a number of other factors may influence ETRs, it is crucial to first understand the distribution and time series properties of ETRs. Trends may emerge in the data that influence future research design choices and spur further research. Without a simple description of the relation between the two variables of interest, it is difficult to disentangle if more complex relations are simply a byproduct of research design choices.

We begin our nonparametric analysis by using kernel density estimates for a sample of financial service firms (SIC 6000-6499) between 1992 and 2016 to examine the distribution of effective tax rates. We impose similar data restrictions as Dyreng et al. (2017) to ensure

[2] A notable exception is Hodder, McNally, and Weaver (2003).

comparability. While we find a similar downward slope in ETRs over time on average, we supplement OLS by plotting the entire ETR distribution over time. We perform several time-series partitions. Importantly, we use this setting to document how research design choices ultimately influence the nonparametric analysis by isolating the impact of these choices. This approach allows us to demonstrate that the choice of bandwidth (analogous to bin width in a histogram) is the most important researcher-imposed design choice.[3] Our results document that the distribution of ETRs among financial services firms is becoming less focused around the mean, with more probability mass occurring for below average ETRs. This result provides a deeper understanding of the time-trend in Dyreng et al. (2017) by suggesting an actual shift has occurred in the distribution of ETRs, which is more prevalent after the 2007-2009 recession.

Next, we examine the relation between audit fees and firm size. A univariate regression of audit fees on size, commonly measured as the natural logarithm of total assets, consistently explains over 70% audit fees. In one seminal paper on audit fees, Simunic (1980) provides early evidence that the relation between audit fees and firm size is nonlinear. More recently, Cullinan, Du, and Zheng (2016) examine various mathematical transformation of audit fees. However, many papers that utilize audit fees as a measure of audit quality do not include higher-order terms for size to account for this non-linearity, such as squares or indicators (e.g., Francis, Reichelet, and Wang 2005; Beck and Mauldin 2014). One means by which researchers can reduce the influence of nonlinearity is through different sample matching procedures (see Shipman, Swanquist, and White 2017 for a discussion of propensity score matching in auditing

[3] Compared to other research design choices such as selecting a kernel weighting function.

research).[4] Given the overall importance of audit fees as a proxy for audit quality, we contribute to the audit literature by examining the post-SOX relation between total assets and audit fees.

In order to establish a baseline result, we use OLS to regress both logged and raw audit fees on a firm's total assets and find a positive and statistically significant coefficient. We then use the setting of audit fees to document how several research design choices may influence results. First, we note that discarding small firms through sample attrition may add nonlinearity. It is fairly common to require that each observation report all necessary control variables to be included in the sample. Thus, many smaller firms that may not be covered by analysts may be removed from a sample. Next, we show how both winzorizing and logging all variables leads to a much more pronounced association. These techniques help ensure that a small number of influential observations do not drive results, but they also remove much of the variation in the measures, which may be of interest to researchers. Lastly, we document that in a sample of very large firms, the relation between audit fees and size systematically diminishes.

The idea that any given statistic cannot convey the entire message of a dataset is not new. Anscombe (1973) creates four datasets with nearly identical common statistical properties, but with very different underlying distributions. He urges authors to plot the data as a best practice. The most basic nonparametric analysis is a histogram, with more sophisticated techniques including the estimation of density functions and weighted local linear regression (e.g., Fan (1992)). Use of histograms as a "bunching analysis" has grown in popularity among economists (e.g., Saez (2010)). These techniques largely rely on a visual depiction with confidence intervals.

[4] In a concurrent working paper, Beardsley, Imdieke, and Omer (2018) examine the possibility that non-audit services have a nonlinear relation with audit quality. We view this paper as a complement to our study.

As with any statistical method, understanding the key assumptions and mechanisms will help determine the most appropriate tool for the research question and data.

It is important to point out the limitations of the methods presented in this paper as well. There are statistical costs associated with using nonparametric estimation. First, these techniques require significantly larger data samples than OLS, and the researcher must trade-off between variance and bias in estimates. However, much accounting research uses samples well in excess of 1,000 observations, reducing concerns over the need for significant data. Second, reliability of estimation is lower in the tails of the distribution, where data points cannot be weighted symmetrically. Next, the approaches discussed in this paper suffer from the "curse of dimensionality," as the techniques are not designed to handle a large number of regressors. While some of these techniques may help accounting researchers provide more unique insight, they should generally be considered complements to more structured regression analysis. Lastly, nonparametric analysis does not help to solve endogeneity problems, such as reverse causality, which are typically a concern of much accounting literature.

New methods have the potential to make a significant impact on accounting research. Some methodological papers are introduced after a technique has been misused in practice (e.g., Larcker and Rusticus (2010); Lennox, Francis, and Wang (2012); Shipman, Swanquist, and Whited (2017)), while others discuss new tools (e.g., Feng, Kim, and Kimbrough (2018)). The goal of our paper is to expand upon the relatively new stream of research that uses nonparametric analysis to provide empiricists with an introductory guide to successfully implementing common nonparametric techniques. The use of statistical packages (Stata, SAS, etc.) is fairly pervasive in accounting research, however, few receive formal training on many of the advanced methods these packages can perform. This can be problematic if researchers do not understand the basic

assumptions and limitations of econometric techniques. In the below analysis we present an overview of nonparametric estimation. Then, we discuss estimating density functions and performing nonparametric regression. Lastly, we show how the techniques discussed can be used in the tax and audit literature.

1. Background on Regression Analysis

Before discussing the nuances of nonparametric analysis, we briefly summarize some of the statistical properties of OLS estimation for comparative purposes. First, OLS estimation yields the best linear unbiased estimator (frequently shortened to "BLUE") based on the Gauss-Markov Theorem. OLS is considered "best" in that it yields an estimate with the lowest variance. It is "linear" in that it fits a linear regression line to the data and "unbiased" in that the estimate eventually converges to the true parameter. For OLS to be considered BLUE, several assumptions must hold. These assumptions include: the error term and the explanatory variables are uncorrelated, the data are IID, and the explanatory variables have finite variance.[5] A common misconception is that the error term must be normally distributed. This is not a requirement for consistency; however, a normal error term does allow for more efficiency.

There are two commonly cited issue with OLS in accounting research that uses panel data. Specifically, data are rarely IID and often include outliers. Because a given firm's actions today are almost always serially correlated with its actions yesterday, and because there is commonly unobserved heterogeneity, the data underlying most archival accounting research are not IID. Most regressions include some form of correction for serial dependence in the error

[5] Additionally, the $\mathrm{E}(X^4) < m \; and \; \mathrm{E}(Y^4) < m$; where $m < \infty$, although this condition is rarely necessary.

terms and often include industry, year, or firm fixed effects to control for heterogeneity.[6] Nonparametric analysis is not designed to overcome the limitations from non-IID data, but it can help alleviate the influence of outliers. OLS is generally not considered robust to outliers in data. This issue has been circumvented in some research by "trimming" data or using regressions more robust to the influence of outliers (Kraft, Leone, and Wasley 2006; Dyreng and Lindsey 2009; Leone, Minutti-Meza, and Wasley 2017). The effect of outliers merits analysis in many situations. Cohen and Lys (2003) states, in the context of analyst forecast errors, "It seems advisable for researchers to conduct diagnostics including the impact of influential observations and to examine the sensitivity of their conclusions depending on whether parametric or non-parametric tests are used." (page 151).

2. Nonparametric Analysis

In this section we provide a brief overview of the importance of visualizing data in accounting research and the statistical properties of nonparametric estimation techniques. It is neither feasible nor practical to cover every nonparametric technique. Therefore, we focus on two popular nonparametric techniques: kernel density estimation and locally weighted regression analysis.

2.1. Visualizing data

Nonparametric techniques were pioneered by Rosenblatt (1956), Nadaraya (1964), and Watson (1964), among others, and have been widely deployed in many areas of the social sciences. As statistical software now offers numerous "canned" visual analyses these techniques are becoming more accessible for accounting researchers. Still, the most common form of visual

[6] Gow, Oremazabal, and Taylor (2010) provide an overview of the appropriate ways of clustering standard errors in OLS regressions using accounting data.

analysis is the histogram. We examine the first issue of 2018 for the top three accounting journals (*The Accounting Review*, *Journal of Accounting Research*, and *Journal of Accounting and Economics*) to examine how prevalent visualization is. Among the 28 empirical papers in these issues, only 54% present any visual evidence of a treatment effect or a descriptive discussion of their sample. Nearly all these figures are time-series trends or histograms. In this paper, we cover different nonparametric estimation techniques that visually depict data or analysis that have garnered less attention in the accounting literature.

We note two important caveats. First, despite limited evidence of the use of nonparametric estimation in recent accounting literature, we cannot observe if the techniques we discuss are being used to properly guide research design choices, as we suggest they should be. The methods section of many papers may not include the minutia of each research design choice, especially those related to identifying the research question. Furthermore, some papers may have had visual aids cut due to publishing space limits, and we did not examine supplemental online materials. Second, of the papers that *do* incorporate visual depictions, most use rudimentary plots of data. We suggest that nonparametric analysis can be used to present broader depictions of data and univariate regression analysis. In this paper, we introduce accounting researchers to the uses and implementation of such tools.

2.2 Description of the kernel density function and estimation

The most basic form of nonparametric estimation is a histogram. In a histogram, the researcher sets up bins of a certain width and plots the number of observations that falls into each bin. The canonical examples of histograms in empirical accounting research are Figures 1 through 3 of Burgstahler and Dichev (1997) (BD97). In this analysis, the authors plot the distribution of earnings and note a significant discontinuity around zero. They attribute this result

to earnings management, with firms manipulating accruals to move from a small loss to a small profit. We use the variables in BD97 to help describe a histogram more formally. We then extend this definition to kernel density estimates.

We follow Cameron and Trivedi (2005) Chapter 9 to obtain the estimator of a histogram through the statistical properties of the estimate (Section 9.3).[7] For a sample of N observations $\{x_i, i = 1,\ldots,N\}$ consider the continuous variable x_0 which is evaluated at point x and has a distribution function $f(x)$. Furthermore, the variable h, the bin width, will help determine the range of points that will be included in a bin. Going back to the BD97 example, x represents earnings and h represents the distance to the midpoint of the earning bin (.0025/2). The histogram is estimated with the following equation:

$$\hat{f}_{\text{Hist}}(x_0) = \frac{1}{N}\sum_{i=1}^{N}\frac{1(x_0 - h < x_i < x_0 + h)}{2h} \qquad (1)$$

The indicator function is one if the observation occurs within the range $x_0 \pm h$. This equation leads to a step function (based on each bin) that weights all observations within a given bin equally. We can now extend equation (1) to a case where we weight each point within a bandwidth. The theory behind weighting (with a kernel weighting function) is that points closer to x are likely more representative and should be considered more than points further from x. In terms of BD97, this would be equivalent to suggesting that firms with earnings closest to zero are more similar than the firms toward either tail of this bin. It is important to note that the final kernel destiny function fits a smooth estimate based on the estimation at each point, x_i.[8] This analysis also relies on several assumptions about the kernel weighting function, $K(\cdot)$. These

[7] We refer to statistical properties generally as asymptotic properties unless otherwise noted that we are discussing finite sample attributes.
[8] While most statistical packages will use the same evaluation criteria for bandwidths across a specific number of points along the entire distribution, the analysis can theoretically be performed pointwise.

assumptions include symmetry around zero, continuity, and integration to one. We discuss some commonly used kernel density functions in Section 2.3.4. The kernel density estimate can then be written as:

$$\hat{f}(x_0) = \frac{1}{Nh} \sum_{i=1}^{N} K\left(\frac{1(\ x_i - x_0)}{h}\right) \qquad (2)$$

While this function looks similar to that of a histogram, equation (2) uses a weighting function instead, $K(\cdot)$, of summing a series of indicators over a particular range. Given the function for kernel density estimates, there are two primary research design choices: bandwidth, h, and the weighting function, $K(\cdot)$. The bandwidth is the distance around data points that will be evaluated by the weighting function (either point-wise or at discrete points along a distribution). The weighting function specifies how much weight to assign to particular observations within the bandwidth. In general the choice of bandwidth is more important than the choice of weighting function, because many weighting functions have similar properties. We illustrate this characteristic in our example on bank ETRs (Section 3).

2.3.1 Kernel Density Regression

The above discussion largely focuses on understanding the distribution of a single variable. While this is useful in many settings, we next turn our attention to the case of nonparametric regression. A broad literature in statistics and econometrics discuses many types of nonparametric and semiparametric techniques; for tractability, we limit discussion to constant and local linear regression. In this section we continue to assume that both x and y have positive econometric properties (that is, that they are IID, continuous, twice differentiable, and have finite variances). Assume we want to estimate the following regression equation:

$$y_i = m(x_i) + \varepsilon_i \qquad (3)$$

We follow Cameron and Trivedi and start by showing a general case where we can estimate the function $m(\cdot)$ by taking the average values of y_i for points that are within h of each observation x:[9]

$$\widehat{m}(x_0) \equiv \frac{\sum_{i=1}^{N} 1\left(\left|\frac{x_i - x_0}{h}\right| < 1\right) y_i}{\sum_{i=1}^{N} 1\left(\left|\frac{x_i - x_0}{h}\right| < 1\right)} \qquad (4)$$

Similar to a histogram, this assigns equal weight to all observations that fall within the bandwidth h. If we wish to vary the weights placed on observations that are closer to x_0, we need a weighting function. In this case we use a kernel weighting function, as above, and obtain the following estimate of $m(\cdot)$:

$$\widehat{m}(x_0) \equiv \frac{\frac{1}{Nh}\sum_{i=1}^{N} K\left(\frac{x_i - x_0}{h}\right) y_i}{\frac{1}{Nh}\sum_{i=1}^{N} K\left(\frac{x_i - x_0}{h}\right)} \qquad (5)$$

This regression technique estimates a constant for each value of x and then plots a curve via interpolation (Opsomer and Breidt (2011)). We can extended this technique beyond a constant to estimate a weighted local linear model (Fan (1992)).[10] Said differently, equation (5) is simply a generalized version of the weighted local polynomial function to the power zero. To show this extension, we turn our attention to the estimation of the following functional form of $m(\cdot)$:

$$m(x) = \alpha_0 + \beta_0 (x - x_0) \qquad (6)$$

To generate a local linear regression we set α and β to minimize:

$$\sum_{i=1}^{N} K\left(\frac{x_i - x_0}{h}\right) (y_i - \alpha_0 - \beta_0 (x_i - x_0))^2 \qquad (7)$$

[9] This estimator can be derived by integrating the joint density function of x and y. See Hardle and Linton (1994).
[10] Higher order polynomials can be estimated in a similar fashion. See Hardle and Linton equation (21) or Cameron and Trivedi equation (9.31). See Fan and Gijbels (1996) for properties of these higher order polynomial estimators.

As in the previous analysis, the researcher chooses the bandwidth, h, and the weighting function $K(\cdot)$. Equation (7) can be estimated using weighted least squares at each point x_0. This leads to a smooth linear estimate around each observation x. Equation (7) produces the locally weighted linear estimate of $\widehat{m}(x)$ (Cameron and Trivedi 2005).

2.3.2 *Properties of the kernel density estimates*

Unlike OLS, where the estimate of $\hat{\beta}$ converges to an unbiased estimate of β, the expectation of the kernel function yields a biased estimate, with this bias increasing in h.[11] However, this does not imply that the researcher should set h as small as possible to reduce the bias. The bandwidth h also appears in the denominator when computing the variance, leading to the classic tradeoff between reducing bias and reducing variance. See Cameron and Treveti (2005), Hardel and Linton (1994), or Hansen (2014) for derivations of the trade-off between variance and bias.

It is also worth noting that the kernel density (KD) estimator converges more slowly than the OLS rate of $\sqrt{N}$.[12] In essence, KD estimates the function along the entire data distribution, requiring significantly more data than OLS. This distinction is analogous to the difference between running a pooled regression versus several industry-level regressions. In order to obtain valid inferences at the industry level, the econometrician would need a sufficiently large number of observations per industry. It is important to note that many statistical packages do not generate the KD estimate based on every data point due to the processing power required. Instead they select a number of points to calculate the KD estimate, typically 50 (Stata 15). Assuming we

[11] Cameron and Treveti note "the kernel estimator is biased in the size $O(h^2)$, where we use the order of magnitude notation that a function $a(h)$ is $O(h^k)$ if $a(h)/h^k$."

[12] Per the delta method: $\sqrt{N}(\hat{\beta} - \beta) \rightarrow N(\mu, \sigma^2)$

need 30 data points at each of these 50 windows to generate valid inferences, this would require at least 1,500 observations – more if the data are not uniformly distributed.

2.3.3 Choice of bandwidth

Setting the bandwidth, h, may be the most important research design choice in nonparametric analysis. In Section 3 we show, using accounting data, that the choice of bandwidth is more important than the choice of kernel function. Jones, Marron, and Sheather (1996) describe the importance of bandwidth choice as follows: "when insufficient smoothing is done, the resulting density or regression estimate is too rough and contains spurious features that are artifacts of the sampling process. When excessive smoothing is done, important features of the underlying structure are smoothed away" (pg 401). Thus, the researcher must select an appropriate bandwidth so as to not under- or over- smooth the data.

Statistical packages often use some derivation of Silverman's (1986) plug-in estimate to generate a reasonable estimate, h^*, of the optimal bandwidth.[13] For example, Stata generates the optimal bandwidth as though the underlying data were Guassian and a Guassian kernel were used.[14] It is important to note that the choice of bandwidth should change based on the chosen kernel weighting function (discussed in Section 2.3.4).[15] Hardle and Linton (1994) note that the plug-in estimates yields an h^* that works well for most symmetric distributions. Cameron and Trevedi (2005) suggest checking the robustness of inferences by setting the bandwidth to h^*, $2h^*$, and $0.5h^*$. Another approach to choosing the bandwidth is to use what is known as the

[13] Note this does not necessarily generate the most efficient or optimal bandwidth that would minimize the mean integrated squared error (i.e., $\int E\,[(\hat{f}(x_0) - f(x_0))^2]dx_0$).
[14] This may oversmooth the density function based on statements from the Stata 15 guide.
[15] Similarly, Hansen (2014) states: "the bandwidth should take the form $h = cn^{-1/5}$. The optimal constant c depends on the kernel k, the bias function $B(x)$ and the marginal density $f_x(x)$." (pg 254).

"leave-out" cross-validation technique. However, they note that this technique is more "computationally burdensome."[16]

2.3.4 Choice of Weighting Function

In a simple histogram, points within the bin are given equal weight. In more advanced nonparametric analysis, there are numerous options for kernel weighting functions. For brevity, we discuss two particularly relevant functions: Epanechnikov and Gaussian (normal). If an optimal bandwidth is selected, the Epanechnikov kernel is the most efficient in that it minimizes the mean integrated squared error. For this reason, Stata uses the Epanechnikov kernel as the default setting. The Gaussian kernel is also important because, unlike other kernels (e.g., Epanechnikov), it gives some weight to observations that fall outside of the bandwidth. This feature reduces estimation concerns where there may be gaps in the data, but as we move outside of the bandwidth, these points receive very little weight.[17] As illustrated in the examples below comparing the Epanechnikov, Gaussian, and other kernel weighting functions, the choice of kernel is not nearly as important as the choice of bandwidth.

2.4 Obtaining Statistical Inferences

The tradeoff between bias and variance creates a unique challenge for empiricists. When performing statistical tests, once must decide if it is more important to determine the correct estimate or the correct significance level. We propose that it is more important to obtain an accurate estimate (one with small bias) than a narrow confidence interval. If the researcher

[16] Stata 15 command npregress uses this method for determining bandwidth. Other cross-validation techniques such as k-fold are less computationally burdensome, but their discussion is beyond the scope of this paper.
[17] Consider the example of a random variable centered on zero and normally distributed. The probability density function is defined from - ∞ to ∞, however as we move into the tails of the distribution, the probability approaches zero.

knows the confidence interval has an inflated variance, this approach reports a more conservative estimate of the confidence interval around the estimates (i.e., compensates for a small bias with a smaller p-value). To implement this approach, we propose the researcher should obtain an estimate with a small bias by under-smoothing, thus setting *h* smaller than optimal. Next, plot a 95% confidence interval around the under-smoothed density estimate.

2.5 Additional Econometric Discussion

Several other econometric items merit discussion. The first of these is a technique called k-nearest neighbor estimation (often abbreviated as k-NN or KNN). Instead of focusing on the area around a point, X_0, based on the bandwidth, *h*, this technique uses a weighted average of the values of the *k* closest observations to the point. Another nonparametric technique is the use of splines. Opsomer and Breidt (2011) describe splines as "an alternative approach… to represent the fit as a piecewise polynomial, with the pieces connecting at points called knots." (page 976). This methodology has recently been used in the capital markets literature to examine cross-sectional attributes that explain equity returns in Freyberger, Neuhierl, and Weber (2017).[18] A full description of these estimators is beyond the scope of this paper. Lastly, kernel density regression does not handle end points as well as locally weighted linear regressions (Fan 1992).

3. Kernel Density Estimates Using Accounting Data

3.1 Effective Tax Rates and the Financial Services Industry

The financial service industry is often excluded from analysis examining the impact of taxation on various economic outcomes. This omission is likely due to two key factors. First, banks and insurance companies are heavily regulated and subject to different tax rules than non-

[18]This paper uses splines in conjunction with a group LASSO procedure. See Section 3 of their paper for a formal discussion of these techniques.

financial firms. From this perspective, their inclusion in a pooled analysis would be inappropriate to the extent that it induces noise in regression estimates. Second, the financial services industry has a structurally different balance sheet than industrial firms (e.g., banks have extremely high amounts of leverage). Many empirical models in accounting are designed with industrial or technology firms in mind. In these settings, banks are often lost to sample attrition due to data requirements. Despite these limitations, a nascent stream of literature has begun to reexamine tax outcomes in the financial services sector (e.g., Schandlbauer (2017); Donohoe, Lisowsky, and Mayberry (2016); Gallemore, Mayberry, and Wilde (2017); Hall and Lusch (Forthcoming).

We fill one void in the literature by performing a descriptive analysis of effective tax rates (ETRs) in the financial services industry. Dyreng, Hanlon, Maydew, and Thornock (2017) document that average ETRs have declined over time; however, they provide limited insight into changes across the ETR distribution. This work has spurred a growing interest in examining macroeconomic trends in tax attributes (e.g., Gaertner, Laplante, and Lynch (2016); Chen, Koester, and Shevlin (2017); Drake, Lusch, and Hamilton (2017), among others). This analysis is particularly well suited for nonparametric techniques because ETRs are a summary measure, and nonparametric analysis is best suited for a small number of variables (similar to using a portfolio analysis). Additionally, nonparametric analysis can provide robust insight into the entire distribution of ETRs.

3.2 Sample selection, variable construction, and descriptive statistics

We begin our empirical analysis by obtaining a sample of firms with historical standard industrial classification (SICH) codes between 6000-6499 for the period 1993 to 2016 from Compustat. Prior to 2003, a significant number of financial services firms are missing SICH codes; thus we backfill these with SIC codes. Our sample is comprised of non-real estate

financial services firms after the implementation of FAS 109, which standardized GAAP ETR reporting. We also require firms to be domiciled in the US. We follow Dyreng et al. (2017) and eliminate very small firms (less than $10m in assets), firms with fewer than 5 observations in our sample, and unprofitable firm-year observations. Collectively, these screens are designed to ensure we capture time-trends rather than changes in sample composition as small firms enter and exit the sample. Though these criteria will create a survivorship bias, we view them as appropriate given the severity of the recession in 2007-2009 that affected much of the financial services industry. Because the data necessary to compute cash ETRs is poorly populated in the earlier portion of our sample, we use GAAP ETR as our primary summary measure. We report a final GAAP ETR sample of 20,957 observations and a cash ETR sample of 13,063 observations. We present the sample selection criteria in Table 1.

[Insert Table 1 Here]

Similar to prior research, we define cash ETR as cash taxes paid (TXPD) over pretax income (PI) and GAAP ETR as tax expense (TXT) over pretax income (PI). We follow Dyreng et al. (2017) and do not adjust the denominator by special items. We manually winsorize ETRs by setting negative values to zero and values greater than 100% to one. This winsorization procedure is not problematic using nonparametric techniques because we do not focus on the tails of the distribution. In addition to these variables, we create several control variables for our OLS analysis consistent with prior literature. These control variables include an indicator for multinational firms, which is set to one if the firm reports a non-zero absolute value of pre-tax foreign income (PIFO) or foreign tax expense (TXFO) and zero otherwise; the log of total assets; advertising expense; return on assets; and an indicator for tax loss carryforwards. Lastly, we create a measure for time-trends by subtracting 1993 from the fiscal year of the observation.

We present the descriptive statistics for the pooled sample in Table 2 Panel A. Based on the difference in the sample composition, we split the data into the GAAP and cash ETR sub-samples. The mean ETRs of 30.2% (GAAP) and 28.7% (cash) are in line with prior literature.[19] The majority of our sample is larger domestic-only firms, firms on average generate a 2% return on assets and spend less than 1% of total assets on advertising expense, and fewer than 10% of firms have tax loss carryforwards. In Figure 1 we plot the time series analysis of ETRs and show a downward time-trend, although the trend is less pronounced than that of Dyreng et al (2017). There is an increase in the number of firm-years in the cash ETR sample after 2003 due to changes in the reporting of firms in SICH 6020. We attribute the spike in cash ETRs in 2008 to the recession that eliminated many unprofitable financial service firms from the sample.

[Insert Table 2 and Figure 1 Here]

3.3 Pooled nonparametric analysis

We begin our nonparametric analysis by pooling all observations among the financial services industry and plotting a histogram of the GAAP ETRs. For comparison, we overlay a normal distribution with the same mean and standard deviation as our dataset. From visual inspection, the data do not appear to be normally distributed.[20] The winsorization technique applied places a substantial number of observations at zero. This "left winsorization" appears to be more prevalent than the "right winsorization" placing observations at one. Empirical researchers working in the tax avoidance field are familiar with the issues associated with manually winsorizing the data, however the technique is still common (e.g., Henry and Sansing (2014); De Simone, Nickerson, Seidman, and Stomberg (2017)). If we limit our attention to the

[19] Drake, Hamilton, and Lusch (2017) report a mean GAAP ETR of 32.2% and Dyreng et al. (2017) report a mean cash ETR of 29.1%.
[20] Tests of skewness and kurtosis confirm the data are not normally distributed.

interquartile range (ETRs between 26-37%), consistent with expectations, we observe a relatively symmetric distribution around the mean. However, the full distribution has significantly more probability mass around the central tendencies, very little mass in the right tail of the distribution, and a left tail that spikes at zero. Next, for comparison purposes, we overlay the default kernel density function over the histogram.

[Insert Figure 2 Here]

Having described our data using histograms, we next demonstrate the importance of weighting function and bandwidth choices using kernel density estimation. In Figure 3, we illustrate several different weighting function choices. First, we allow the program (Stata 15) to set the optimal bandwidth and use the default Epanechnikov kernel weighting function (Panel A).[21] This function is identical to the one we overlay on the histogram in the Figure 2. Stata sets the default bandwidth to 0.0091, which we maintain in Panels B and C, for comparative purposes. Next, we use a Guassian kernel, which weights points based on a normal distribution (Panel B), and a triangular kernel, which weights observations based on the absolute value of their distance from x (Panel C). In the last figure, Panel D, we compare the three different weighting functions for ETRs between 15% and 45%, which includes the entire interquartile range, using the default bandwidth (which changes based on the sample). Because there is minimal probability mass outside of the ETR range of 15-45% (except for 0%) we will continue to use this range for comparison purposes.

[Insert Figure 3 Here]

[21] Stata has two different Epanechnikov functions: Epanechnikov(defualt) and Epan2. The software discusses the slight distinction between these two weighting functions.

Upon comparison of the three different weighting functions, Panels A and B (Guassian and Epanechnikov kernels) appear remarkably similar. When comparing these to the triangular kernel, there are slight differences in the tails, where these graphs are least well defined. In Panel D, differences between the three kernels emerge (for example, the triangular kernel is more jagged), but these differences are minor. Based on this comparison, we echo the discussion from prior statistical guides that suggests differences in weighting function do impact the estimation process but that these differences are generally small unless there are holes in the data. To the extent that there are holes in the data, a Gaussian kernel is best suited to estimate the function because it gives weights to points outside the bandwidth.

For the next analysis, we hold constant the Epanechnikov kernel weighting function and systematically vary the bandwidth. There do not appear to be any holes in the data along this range, minimizing the need to use the Guassian kernel. We select three different bandwidths, one that undersmooths (Panel A), the default setting (Panel B), and one that oversmooths (Panel C). We use bandwidths of 0.0005 to undersmooth, 0.0074 based on the default setting, and 0.1 to oversmooth.[22] These density functions are graphed in Figure 4, along with the comparison in Panel D.

[Insert Figure 4 Here]

When we examine Panels A-C, the differences between them are striking compared to the differences observed in Figure 3. In Panel A, when we undersmooth by setting the bandwidth smaller than optimal, the figure appears jagged. This image is consistent with the observation from Jones et al. (1996) that an undersmoothed estimate "is too rough." In Panel B, the default

[22] Note that the default bandwidth changes when we limit the sample to ETRs between 15 and 45%.

setting is significantly smoother, yet still shows the level of detail needed to identify the greater probability mass to the left of the mean. In Panel C, the oversmoothing procedure has essentially "smoothed away" all details. What remains is a relatively symmetric curve with very little additional insight into the distribution. The comparison among the different bandwidths (Panel D) leads to an interesting quandary: should a researcher undersmooth in order to gain better insight into the data? We reiterate the advice of Cameron and Trevedi (2005), which suggests halving and doubling the default (or optimal) bandwidth and reporting all three for robustness. As with most empirical questions, however, the best answer is often to examine multiple design choices and clearly explain why a given choice is preferable. This paper offers some initial guidance on these choices.

3.4 Time-Series Analysis

Dyreng et al. (2017), among others, note a significant time-series decline in cash ETRs for a sample of non-financial firms. Our sample selection criteria are similar to Dyreng et al. (2017) allowing for comparability between the samples. Note that this is not a replication of their study. We simply attempt to use the same sampling criteria to compare the financial services industry to the non-financial services industry over roughly the same time period (their sample is 1988-2012; ours is 1993-2016). We deviate from Dyreng et al. (2017) and focus on GAAP ETRs due to data availability but present regression results for cash ETRs for robustness. Additionally, we note a spike in cash ETRs in 2008. This is likely due to removing loss observations around the financial crisis. We perform various permutations of the following equation:

$$GAAP\_ETR_{it} = \alpha + \beta_1 Time_{it} + \varepsilon_{it} \qquad (8)$$

Table 3 presents time-series analysis of GAAP (Panel A) and cash (Panel B) ETRs. Column (1) shows the results of a pooled OLS regression with a single regressor, a count variable for time. Columns (2) and (3) present the results of industry and firm fixed effect regressions. Columns (4) and (5) limit the sample to large firms (those in the top quartile of assets) and multinational firms (those reporting non-missing values of foreign income or taxes (pifo or txtfo) in Compustat). Lastly, Column (6) includes controls for multinationals, the log of assets, advertising expense, return on assets, and tax loss carryforwards. Across all six specifications, the coefficient on time is negative and significant. The magnitude of the time-trend is not as large as that reported in Dyreng et al., which could be due to the different sample periods. Our Figure 1 notes a rise in ETRs after 2012, a time period outside the Dyreng et al. sample. Additionally, our sample begins in 1993, not 1988, though Dyreng et al. Figure 3 does not visually support the suggestion that ETR's were higher in 1988 than in 1993. These results make a novel contribution to the literature by extending prior studies to a major sector of the U.S. economy (financial service industry).

[Insert Table 3 Here]

We now show how plotting kernel density functions is complimentary to more structured OLS regression analysis. We use the Epanechnikov kernel and default bandwidth.[23] In Figure 5 we plot ETRs across three different time periods. Observations prior to 2002 are labeled "Early," observations from 2002-2008 are labeled "Middle," and observations after 2008 are labeled "Late." Panel A reports ETRs over the range [0, 1], and Panel B reports over the range (.15, .45).

[Insert Figure 5 Here]

[23] In untabled analysis we undersmooth and report no noteworthy differences in the figures.

In Panel A, we note that there are very few observations with ETRs over 50% in any of the three periods. We conclude that GAAP ETR spikes are not common among our sample of firms in the financial services industry. Consistent with our OLS analysis, we identify a clear negative time-trend in ETRs. In both Panels A and B, we see a left shift in the distribution. This shift does not appear to be a tail effect, but instead firms around the mean lower their ETRs. Additionally, the probability mass has become significantly flatter over time.[24] One can interpret this result as a reduced occurrence of firms reporting an average ETR but an increased occurrence of firms reporting ETRs slightly below average. Overall, this increased dispersion seems to be localized to the observations below the median of the distribution.

We also perform our analysis using cash ETRs (untabulated) but find the results less systematic than those for GAAP ETRs due to the spike in cash ETRs around 2008. To ensure that our inferences are not systematically affected by the financial crisis, we plot the densities for the 1990s and 2010s and remove all observations between 2000 and 2010 (untabulated) and find a similar shift in GAAP ETRs. We also examine cash ETRs in the 1990s vs. 2010s, which show the same left shift in the distribution as GAAP ETRs. However, as we undersmooth, the pattern becomes less distinct. Collectively, the nonparametric analyses supplement the OLS results and point to a systematic shift in the distributions of ETRs over time.

4. Nonparametric Regression Estimates Using Accounting Data

4.1 Audit Fees Analysis

A key area of research in accounting examines the determinants of audit quality. While researchers have studied many different measures meant to capture the supply of and demand for

[24] Measures of kurtosis suggest the tails of the ETR distribution are drawing inward over time.

high-quality audits, one commonly used measure is audit fees. For a recent survey of the auditing literature please see DeFond and Zhang (2014). In a typical audit fees study, the econometrician will regress the natural log of audit fees on a host of independent variables. Swanquist and Whited (2018) document that there has been a dramatic increase in the number of "control" variables used in accounting studies published in top journals over time. The inclusion of such a multitude of variables in the regression can introduce both measurement error and endogeneity. They suggest applying a more parsimonious model and documenting that the results are robust to the inclusion of additional controls. This approach is similar to the suggestion from Gow, Larcker and Reiss (2016) that in order to eventually document causal inferences, accounting research must first provide more high-quality descriptive studies that offer insight on institutional details. In this spirit, we provide a detailed descriptive analysis on the relation between audit fees and firm size that can be used for future audit research.

4.2 Sample Selection, Variables, and OLS Regression Analysis

We begin with the total population of U.S. domiciled firms that report both audit fees (*audit_fees)* in Audit Analytics and total assets (*AT*) in Compustat after 2003. We begin our analysis in 2004 because it allows us to analyze firm-years entirely after the passage of Sarbanes-Oxley (SOX), which drastically changed the nature of audits. After eliminating very small firms (those with less than $1m in assets), we generate our final sample of 79,279 firm-year observations[25]. The sample selection criteria are listed in Table 1. We examine both the raw values (*AT* and *Audit_Fees*) and the natural log (*Size* and *Log_Audit_Fees*) of both total assets and audit fees. We then create squared and cubic terms of *AT* and *Size* to address nonlinearity

[25] Note the two analyses use different cut-offs for "very small" firms based on norms in the literature.

(for a discussion of nonlinearity using mutual fund data see Cullinan, Du, and Zheng 2016). To ensure the raw variables are on a similar scale (millions), we divide *audit_fees* by one million and rename the variable A*udit_Fees_mill.* All variables are winsorized at the 1 and 99 percentiles. While it is a common practice in accounting research to both log and winsorize the same variable, it is worth noting that this technique essentially smooths the data twice.

We perform three sets of baseline OLS regressions, each examining both the untransformed variables and the log transformed variables. These include only a linear term; linear and quadratic terms; and lastly linear, quadratic, and cubic terms. For brevity we only report the full log-transformed regression equation:

$$Log\_Audit\_Fees_{it} = \alpha + \beta_1 Size_{it} + \beta_2 Size^2_{it} + \beta_3 Size^3_{it} + \varepsilon_{it} \quad (9)$$

The results of these three regressions are reported in Table 4. For the purposes of our analysis, we focus on nonlinearity in the independent variable. We could also examine non-linearity in the dependent variable by using maximum likelihood estimation; however the functional form of the variables need not be estimated or assumed in the nonparametric analysis.[26] Across all of the specifications presented in Table 4, size appears to be a key determinant of audit fees. The univariate regressions, which do not include year or industry fixed effects, have R^2 of .531 and .656. It is important to note that the fit is better for logged values than raw values, which is likely due to smoothing away a significant amount of variation. With this smoothness, we also find the squared and cubed terms are slightly less significant than in the regression using raw values.

[26] For example, the command boxcox in STATA performs maximum likelihood estimation to obtain the Box-Cox parameters for both independent and dependent variables that fit a model with better econometric properties. Discussion of these techniques are outside the scope of this paper.

[Insert Table 4 Here]

Next, we demonstrate how sample selection criteria can affect results. We do this in two ways: first, by limiting our sample based on an ad hoc cut off of $1M in audit fees and second, by running the regression by size quartiles. In Panel A of Table 5, we show that when we restrict our sample, non-linearity appears to be less of an issue. In this specification, when we include squared and cubic terms, the squared term is only marginally significant, and the cubic term is insignificant. The R^2 is lower for the restricted sample than the full sample. If the researcher has a preconceived notion that the R^2 should be higher, they may be deceived by the addition of industry and year fixed effects. If we include year and industry fixed effects (Column 4), it would appear as though there is no loss of explanatory power, demonstrating the potential pitfalls of statistical heuristics. Comparing Tables 4 and 5 reveals a perplexing result. When examining the model with both a squared and cubed term (Column (6) of Table 4 and Column (3) of Table 5), it appears that nonlinearity is reduced for the restricted sample. However, when we compare the models without the cubic term (Column (5) of Table 4 and Column (2) of Table 5), the point estimate and t-statistic on the squared term are larger in Table 5 than Table 4, suggesting greater nonlinearity in the restricted sample. These conflicting results are an inherent problem with OLS and functional form misspecification. Without any theoretical grounds, we cannot determine which is the better model or the "true" parameter of interest. In the following section, we discuss how nonparametric analysis can help alleviate this problem.

In Panel B of Table 5, we partition the sample based on firm size. This partition allows the estimated relation between size and audit fees to change based on where along the distribution of the x variable an observation falls. Panel B reveals that both the point estimate on size and the explanatory power differ from the previous regressions. It is important to note that

R^2 are not directly comparable between different samples and should be viewed as a single measure of "goodness of fit." These results underscore the importance of understanding how sample selection choices may influence results.

[Insert Table 5 Here]

4.3 Nonparametric Regression

In this section, we perform nonparametric analysis on the relation between audit fees and size to complement the OLS results above. We first perform a kernel density regression and then a locally weighted linear regression. Additionally, we show how several research design choices may affect the nonparametric results. These procedures use the Stata command lpoly with either a degree of 0 or 1. Higher degrees yield higher-order polynomial smoothing. This command uses a plug-in estimate for bandwidth and defaults to an Epanechnikov kernel. In Figure 6 we show both the kernel and local linear estimates for raw and logged values.

[Insert Figure 6 Here]

In Panel A of Figure 6, the nonparametric regression with log-transformed variables shows almost a linear relation. While there is some evidence of nonlinearity, we feel comfortable concluding a regression with only a linear term for size is sufficient. Additionally, because the log transformation has already smoothed away some variation, there is little difference between a constant weighted and a local weighted linear regression. In comparison, the results presented in Table 4, Column (6) show that both the squared and cubed terms are statistically significant, though the magnitude is small and these variables add little explanatory power to the regression. Figure 6, Panel B reveals much more variation in the raw values. Accordingly, we also report a higher order polynomial for comparison. Based on the visual depiction from Panel B, we focus

on firms with over $100B in assets in Panel C. Untabulated regression analysis shows that for these firms (N=1,495), the R-squared is now below 10% and the point estimate is significantly smaller. For the largest firms in the sample, it appears the relation between audit fees and total assets is tenuous. For the largest firms, characteristics other than size may be significantly more important determinants of audit fees.

In Figure 7 we explore several additional features of the relation between audit fees and size. First, in Panel A, we examine how the relation changes when we impose sample selection restrictions. Parallel to our limited-sample OLS specification above (results presented in Table 5), we exclude observations with less than $1M in audit fees. Recall that above, our OLS results yielded conflicting conclusions regarding whether there was more or less nonlinearity in the relation in the restricted sample as compared to the full sample. Nonparametric analysis helps us overcome this limitation of OLS by visually depicting the relation between audit fees and size. The relation appears to have greater nonlinearity in the restricted sample. Research design choices may create or eliminate the need for non-linear terms using OLS, and we caution researchers against including these terms by default as they may be unnecessary or misspecified. We suggest utilizing nonparametric analysis to help determine the most appropriate specification.

[Insert Figure 7 Here]

Next, we demonstrate how nonparametric analysis can be used to detect differences in the relation between two distinct groups: Big 4 and Non-Big 4 auditors. In Panel B, we examine the relation between *Size* and *Log Audit Fees* for Big 4 and non-Big 4 auditors. Because the right tail of the distribution for firms with non-Big 4 auditors is not well defined, we use logs. If we use raw data, the regression lines become very jagged, making it difficult to interpret them and to

draw inferences. A comparison of the relation across the two types of firms shows that the slopes of the two lines diverge at a certain point. In Panel C, we use raw values for firms below median size to obtain better insight. We see that this divergence in slopes occurs relatively quickly. For these observations, fees appear to be more strongly related to size among Big 4 clients than non-Big 4 clients. Overall, these results present a descriptive analysis on the relation between audit fees and size. This analysis highlights how nonparametric analysis may aid researchers in designing more structured analysis, for example by providing insight on sample selection criteria and the need for non-linear terms.

5. Discussion on the Benefits and Limitations

Nonparametric analysis has many benefits, and we suggest that it can and should play a larger role in accounting research. In this section, we summarize the key benefits and applications of nonparametric analysis, as well as some limitations of these techniques.

By facilitating visual inspection of data, nonparametric analysis can be particularly helpful in guiding research design choices. For example, as illustrated above, it can provide insight on sample selection criteria and the need for non-linear terms in a more structured regression analysis. It can also be used to identify anomalies or outliers in datasets. Influential observations are not uncommon in several accounting variables (for example, analyst forecast errors, stock pricing errors, share volume, and market share) and require careful treatment. Researchers can also use nonparametric analysis to examine the distribution of residuals. Because many econometric techniques rely on properties of these distributions, this type of analysis can have significant implications for methodological choices.

One drawback of the nonparametric methods we discuss is that they are generally constrained to the examination of two variables.[27] We suggest, however, that this drawback is actually also a strength. It is often useful to show the relation between two variables of interest alone, even if the analysis will eventually include control variables. Without the most simplistic understanding of the relation, it is difficult to justify modeling a more complex one. In this sense, nonparametric tools are complementary to more structured regression techniques commonly used in accounting research, like OLS. We encourage researchers to remember that the average treatment effects identified by OLS regressions are just that – they represent average associations and are only valid for the mean observation. Nonparametric analysis can lend further insight by describing relations at other points along the distribution.

A key benefit of nonparametric analysis is that it allows for statistical analysis under very weak assumptions. It is important to note that the tools presented in this paper can be applied to all types of data, but statistical inferences are based on IID data. This makes nonparametric analysis well suited to experimental studies with sufficiently large participant pools, since data are generally IID when treatment and control groups are randomly assigned. A strength of nonparametric analysis for accounting research in general is that empirical accounting studies often analyze large data sets. Nonparametric estimators such as the kernel density estimator discussed above converge to their true values more slowly than the OLS estimator and as such require more data.

6. Conclusion

[27] It is possible to create three dimensional images examining the relation between three variables. We forgo discussion of these graphs because they are difficult to interpret, detracting much of the benefit of nonparametric analysis that comes from straightforward visualization. Additionally, there are statistical nonparametric techniques that allow for numerous regressors (e.g., splines, generalized additive models); however, their discussion is outside the scope of this paper.

In this paper, we offer an overview of several nonparametric techniques that may be useful in accounting research. We illustrate two of these techniques, kernel density estimation and locally weighted regression, by analyzing time-series properties of effective tax rates in the financial service industry and the relation between audit fees and firm size. We discuss the statistical properties, limitations, and key inputs of these techniques in the hope that accounting researchers will use them to complement more common econometric techniques like OLS. Nonparametric analysis has a multitude of applications in accounting research, such as identifying outliers, examining the distribution of residuals, and guiding research design choices.

The techniques we present in this paper can aid researchers in presenting descriptive analysis. Gow, Larcker, and Reiss (2016) call for more research in accounting that presents causal links. Specifically they state: "We believe that accounting research can benefit substantially from more in-depth descriptive research…. this type of research is essential to improve our understanding of causal mechanisms" (pg 499). We heed this call and extend the current tax and audit literature. We present an in-depth descriptive analysis on the time-series properties of effective tax rates in the financial service industry and nonlinearities in the relation between audit fees and firm size. Taken together, our results provide both methodological and non-methodological contributions.

REFERENCES

Anscombe, F.J., 1973. Graphs in Statistical Analysis. Am. Stat. 27, 17–21.

Beardsley, E.L., Imdieke, A., Omer, T.C., 2018. Evidence of a Nonlinear Association between Auditor-Provided Non-Audit Services and Audit Quality.

Beck, M.J., Mauldin, E.G., 2014. Who's really in charge? Audit committee versus CFO power and audit fees. Account. Rev. 89, 2057–2085. doi:10.2308/accr-50834

Burgstahler, D., Dichev, I., 1997. Accounting. J. Account. Econ. 24, 99–126.

Cameron, A., Trivedi, P., 2005. *Microeconometrics: Methods and Applications*. New York, Cambridge University Press.

Chen, N., Koester, A., 2017. On the Divergence between Corporate Tax Expense and Tax Paid Working paper

Cohen, D.A., Lys, T.Z., 2003. A note on analysts' earnings forecast errors distribution. J. Account. Econ. 36, 147–164. doi:10.1016/j.jacceco.2003.11.002

DeFond, M., Erkens, D.H., Zhang, J., 2016. Do Client Characteristics Really Drive the Big N Audit Quality Effect? New Evidence from Propensity Score Matching. Manage. Sci. mnsc.2016.2528. doi:10.1287/mnsc.2016.2528

DeFond, M., Zhang, J., 2014. A review of archival auditing research. J. Account. Econ. 58, 275–326. doi:10.1016/j.jacceco.2014.09.002

De Simone, L., 2016. Does a common set of accounting standards affect tax-motivated income shifting for multinational firms? J. Account. Econ. 61, 145–165. doi:10.1016/j.jacceco.2015.06.002

De Simone, L., Nickerson, J., Seidman, J.K., Stomberg, B., 2016. How Reliably Do Empirical Tests Identify Tax Avoidance? Work. Pap. doi:10.2139/ssrn.2534058

Donohoe, M.P., Lisowsky, P., Mayberry, M.A., 2016. Taxes, Competition, and Organizational Form, University of Illinois.

Drake, K., Hamilton, R., Lusch, S.J., 2017. The sources of declining effective tax rates for multinational and domestic firms : Insight from effective tax rate reconciliations.

Dyreng, S.D., Hanlon, M., Maydew, E.L., Thornock, J.R., 2017. Changes in corporate effective tax rates over the past 25 years. J. financ. econ. 124, 441–463. doi:10.1016/j.jfineco.2017.04.001

Dyreng, S.D., Lindsey, B.P., 2009. Using financial accounting data to examine the effect of foreign operations located in tax havens and other countries on U.S. multinational firms' Tax rates. J. Account. Res. 47, 1283–1316. doi:10.1111/j.1475-679X.2009.00346.x

Fan, J., 1992. Design-adaptive Nonparametric Regression. J. Am. Stat. Assoc. 87, 998–1004. doi:10.1080/01621459.1992.10476255

Fan, J., Gijbels, I., 1996. *Local Polynomial Modelling and Its Applications*. London: Chapman and Hall.

Feng, I.R., Kim, O., Kimbrough, M., 2017. The marginal coefficient: a new approach for identifying observation level sensitives. Work. Pap.

Francis, J.R., Reichelt, K., Wang, D., 2005. The Pricing of National and City-Specific Reputations for Industry Expertise in the U.S. Audit Market. Account. Rev. 80, 113–136. doi:10.1007/s11113-013-9277-6

Freyberger, J., Neuhierl, A., Weber, M., 2017. Dissecting Characteristics Nonparametrically. Stock. Sch. Econ. TAU Financ. Conf. doi:10.3386/w23227

Gaertner, F.B., Laplante, S.K., Lynch, D.P., 2016. Trends in the Sources of Permanent and Temporary Book-Tax Differences During the Schedule M-3 Era. Natl. Tax J. 69, 785–808. doi:10.17310/ntj.2016.4.03

Gallemore, J., Mayberry, M., Wilde, J., 2017. Corporate Taxation and Bank Outcomes: Evidence from U.S. State Taxes.

Gow, I.D., Larcker, D.F., Reiss, P.C., 2016. Causal Inference in Accounting Research. J. Account. Res. 54, 477–523. doi:10.1111/1475-679X.12116

Gow, I.D., Ormazabal, G., Taylor, D.J., 2010. Correcting for Cross-Sectional and Time-Series Dependence in Accounting Research. Account. Rev. 85, 483–512.

Hall, C., Lusch, S., Forthcoming. Strategic Cost Shifting and State Tax Minimization. J. Man. Acct. Res.

Hansen, B. 2014. *Econometrics*. Current edition Online at https://www.ssc.wisc.edu/~bhansen/econometrics/

Härdle, W., & Linton, O. (1994). Chapter 38 Applied nonparametric methods. *Handbook of Econometrics*, *4*, 2295–2339. https://doi.org/10.1016/S1573-4412(05)80007-8

Henry, E., Sansing, R., 2014. Data Truncation Bias and the Mismeasurement of Corporate Tax Avoidance D. Work. Pap.

Hodder, L., McAnally, M.L., Weaver, C.D., 2003. The influence of tax and nontax factors on banks' choice of organizational form. Account. Rev. 78, 297–325. doi:10.2308/accr.2003.78.1.297

Jones, M.C., Marron, J.S.S., Sheather, S.J.J., 1996. A Brief Survey of Bandwidth Selection for Density Estimation. J. Am. Stat. Assoc. 91, 401–407. doi:10.1080/01621459.1996.10476701

Kraft, A., Leone, A.J., Wasley, C., 2006. An analysis of the theories and explanations offered for the mispricing of accruals and accrual components. J. Account. Res. 44, 297–339. doi:10.1111/j.1475-679X.2006.00202.x

Larcker, D.F., Richardson, S.A., 2004. Fees paid to audit firms, accrual choices, and corporate governance. J. Account. Res. 42, 625–658. doi:10.1111/j.1475-679X.2004.t01-1-00143.x

Larcker, D.F., Rusticus, T.O., 2010. On the use of instrumental variables in accounting research. J. Account. Econ. 49, 186–205. doi:10.1016/j.jacceco.2009.11.004

Lennox, C.S., Francis, J.R., Wang, Z., 2012. Selection models in accounting research. Account. Rev. 87, 589–616. doi:10.2308/accr-10195

Leone, A.J., Minutti-Meza, M., Wasley, C., 2017. Influential Observations and Inference in Accounting Research. Work. Pap.

Nadaraya, E.A., 1964. On Estimating Regression. Theory Probab. its Applications 186–190.

Opsomer J.D., Breidt F.J. (2011) Nonparametric Regression Using Kernel and Spline Methods. In: Lovric M. (eds) *International Encyclopedia of Statistical Science*. Springer, Berlin, Heidelberg

Rosenblatt, M., 1956. Remarks on Some Nonparametric Estimates of a Density Function. Ann. Math. Stat. 27, 832–837.

Saez, E., 2010. Do Taxpayers Bunch at Kink Points ? Am. Econ. J. Econ. Policy 2, 180–212.

Schandlbauer, A., 2017. How do financial institutions react to a tax increase? J. Financ. Intermediation 30, 86–106. doi:10.1016/j.jfi.2016.08.002

Schepens, G., 2016. Taxes and bank capital structure. J. financ. econ. 120, 585–600. doi:10.1016/j.jfineco.2016.01.015

Shipman, J.E., Swanquist, Q.T., Whited, R.L., 2017. Propensity score matching in accounting research. Account. Rev. 92, 213–244. doi:10.2308/accr-51449

Skinner, D.J., 2008. The rise of deferred tax assets in Japan: The role of deferred tax accounting in the Japanese banking crisis. J. Account. Econ. 46, 218–239. doi:10.1016/j.jacceco.2008.07.003

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. London: Chapman and Hall.

Simunic, D.A., 1980. The Pricing of Audit Services : Theory and Evidence. J. Account. Res. 18, 161–190.

Swanquist, Q.T., Whited, R.L., 2018. Out of Control: The Use and Misuse of Controls in Accounting Research. Work. Pap.

Watson, G.S., 1964. Smooth regression analysis. Indian J. Stat. 26, 359–372. doi:10.2307/25049340

Appendix A

Panel A: Variable Names

Variable Name	Definition
Financial Statement Data	
GAAP_ETR	Tax expense (TXT) over pretax income (PI)
Cash_ETR	Taxes paid (TXPD) over pretax income (PI)
Time	The current fiscal year minus 1993
MNC	An indicator set to one if the firm reports a non-zero absolute value of pre-tax foreign income (PIFO) or foreign tax expense (TXFO) and zero otherwise
Size	the log of total assets (AT)
$Size^2$	Size squared
$Size^3$	Size cubed
AD	Advertising expense (XAD) scaled by total assets (AT)
ROA	Net income (PI-TXT) over total assets (AT)
NOL	An indicator set to one if a firm has a positive value of tax loss carryforwards (TLCF)
Audit_Fees_mill	Audit fees (audit_fees) divided by 1,000,000
Log_Audit_Fees	The log of audit fees (audit_fees)
AT	Total assets (AT)
AT^2	Total assets squared
AT^3	Total assets cubed

TABLE 1

Sample Selection Criteria

U.S Domiciled Firms in SIC 6000-64999 after 1992	28,471
Eliminating firms with losses, less than $10m in assets or missing tax expense	(4,748)
Eliminating firms with less than 5 years of data	(2,766)
GAAP ETR Sample	20,957
Eliminating firms without cash ETR	(7,894)
Cash ETR Sample	13,063
U.S Domiciled Firms After 2003	83,849
Eliminating firms with less than $1m in assets, missing total assets, missing audit fees	(4,570)
Audit Fees Sample	79,279

TABLE 2

Descriptive Statistics

Variable	N	Mean	Std Dev	25th Pctl	50th Pctl	75th Pctl
Panel A: GAAP ETR Sample						
GAAP_ETR	20,957	0.302	0.118	0.264	0.326	0.364
MNC	20,957	0.091	0.287	0.000	0.000	0.000
Size	20,957	7.288	1.929	5.994	6.902	8.323
AD	20,957	0.001	0.009	0.000	0.000	0.001
ROA	20,957	0.022	0.051	0.007	0.010	0.015
NOL	20,957	0.063	0.244	0.000	0.000	0.000
Cash ETR Sample						
Cash_ETR	13,063	0.287	0.209	0.151	0.276	0.375
MNC	13,063	0.144	0.351	0.000	0.000	0.000
Size	13,063	7.567	2.029	6.224	7.180	8.711
AD	13,063	0.002	0.011	0.000	0.000	0.001
ROA	13,063	0.029	0.061	0.007	0.010	0.025
NOL	13,063	0.099	0.299	0.000	0.000	0.000
Panel B: Audit Fees Sample						
Audit_Fees_mill	79,279	1.86	0.68	3.68	0.21	1.72
Log_Audit_Fees	79,279	13.35	13.43	1.51	12.25	14.36
Size	79,279	6.13	6.24	2.54	4.37	7.85
$Size^2$	79,279	43.98	39.00	31.91	19.09	61.66
$Size^3$	79,279	347.75	243.56	359.72	83.44	484.21
AT	79,279	7,066.05	515.43	26,687.02	79.02	2,572.35
AT^2	79,279	762,117,234	265,667	5,098,161,801	6,245	6,616,964
AT^3	79,279	1.34E+14	136,932,504	1.05.E+15	493,470	1.70E+10

Panel C: Audit Fees and Size by Year

		Average		Median	
Year	N	Audit Fees ($Mill)	AT ($Mill)	Audit Fees ($Mill)	AT ($Mill)
2004	7,082	1.410	5,012	0.444	312
2005	6,893	1.551	5,324	0.535	342
2006	6,648	1.749	5,747	0.608	387
2007	6,358	1.802	6,288	0.674	424
2008	6,036	1.873	6,516	0.697	477
2009	5,905	1.789	6,798	0.670	498
2010	5,824	1.796	7,160	0.662	539
2011	5,759	1.863	7,510	0.700	594
2012	5,802	1.927	7,884	0.709	619
2013	5,894	2.006	8,094	0.761	638
2014	5,869	2.098	8,481	0.814	704
2015	5,666	2.187	8,814	0.885	784
2016	5,282	2.321	9,594	0.933	913
2017	261	1.884	5,913	0.523	406

TABLE 3
Financial Firms Tax Rate Time-Series

Panel A: GAAP ETRs

D.V.= GAAP ETR	(1)	(2)	(3)	(4)	(5)	(6)
VARIABLES	Pool Sample	Industry FE	Firm FE	Large Firms	MNCs	Controls
Constant	0.3303***	0.3289***	0.3259***	0.3392***	0.3562***	0.3232***
	(127.42)	(133.21)	(125.24)	(68.65)	(32.53)	(41.07)
Time	-0.0026***	-0.0025***	-0.0022***	-0.0032***	-0.0035***	-0.0026***
	(-12.47)	(-11.91)	(-9.14)	(-8.93)	(-4.62)	(-11.75)
Controls	No	No	No	No	No	Yes
Observations	20,957	20,957	20,957	5,239	1,900	20,957
R-squared	0.022	0.080	0.398	0.038	0.024	0.092
Fixed Effects	None	Industry	Firm	None	None	Industry
Cluster	Firm	Firm	Firm	Firm	Firm	Firm

Panel B: Cash ETR

D.V.= Cash ETR	(1)	(2)	(3)	(4)	(5)	(6)
VARIABLES	Pool Sample	Industry FE	Firm FE	Large Firms	MNCs	Controls
Constant	0.3172***	0.3494***	0.3573***	0.3007***	0.2998***	0.3759***
	(46.87)	(51.61)	(48.47)	(25.38)	(19.72)	(27.51)
Time	-0.0022***	-0.0045***	-0.0051***	-0.0024***	-0.0025**	-0.0039***
	(-5.15)	(-9.37)	(-9.48)	(-3.47)	(-2.53)	(-7.48)
Observations	13,063	13,063	13,063	3,924	1,882	13,063
R-squared	0.004	0.038	0.269	0.006	0.006	0.048
Fixed Effects	None	Industry	Firm	None	None	Industry
Cluster	Firm	Firm	Firm	Firm	Firm	Firm

Note: We present the results of regression Equation (8). Panel A (B) uses GAAP ETR (Cash ETR) as the dependent variable. All variables are defined in Appendix A. We define industry fixed effects based on 4-digit SIC codes. Large firms are those in the highest quartile of assets and MNCs are those where *MNC*=1. Other controls are *Size, AD, ROA* and *NOL*. *, **, and *** signify statistical significance at the 10%, 5%, and 1% significance level, respectively.

TABLE 4

Association between Audit Fees on Size

VARIABLES	(1) Raw Linear	(2) Raw Quadratic	(3) Raw Cubic	(4) Log Linear	(5) Log Quadratic	(6) Log Cubic
AT	0.0001***	0.0002***	0.0003***			
	(31.62)	(23.21)	(23.81)			
AT^2		-0.0000***	-0.0000***			
		(-12.93)	(-10.53)			
AT^3			0.0000***			
			(7.79)			
Size				0.4823***	0.3984***	0.5179***
				(161.54)	(38.22)	(21.97)
$Size^2$					0.0069***	-0.0156***
					(7.80)	(-3.50)
$Size^3$						0.0012***
						(4.97)
Constant	1.1469***	0.8308***	0.6553***	10.3945***	10.6043***	10.4427***
	(49.98)	(39.51)	(34.45)	(589.74)	(400.86)	(300.21)
Observations	79,279	79,279	79,279	79,279	79,279	79,279
R-squared	0.531	0.596	0.618	0.656	0.658	0.658
Fixed Effects	None	None	None	None	None	None
Cluster	Firm	Firm	Firm	Firm	Firm	Firm

Note: We present the results of regression Equation (9). In Columns 1-3 the dependent variable is raw audit fees, and in Columns 4-6 it is the natural log of audit fees. All variables are defined in Appendix A. *, **, and *** signify statistical significance at the 10%, 5%, and 1% significance level, respectively.

TABLE 5
Association between Audit Fees on Size: Cross-sectional

Panel A: Audit Fees in Excess of $1M

VARIABLES	(1) Log Linear audit fees > $1m	(2) Log quadratic audit fees > $1m	(3) Log cubic audit fees > $1m	(4) Log Linear w FE audit fees > $1m
Size	0.3254***	-0.3405***	-0.2947**	0.4032***
	(55.70)	(-11.30)	(-2.11)	(72.29)
$Size^2$		0.0397***	0.0340*	
		(20.73)	(1.84)	
$Size^3$			0.0002	
			(0.30)	
Observations	31,084	31,084	31,084	31,084
R-squared	0.496	0.535	0.535	0.660
Fixed Effects	None	None	None	SIC4 & Year
Cluster	Firm	Firm	Firm	Firm

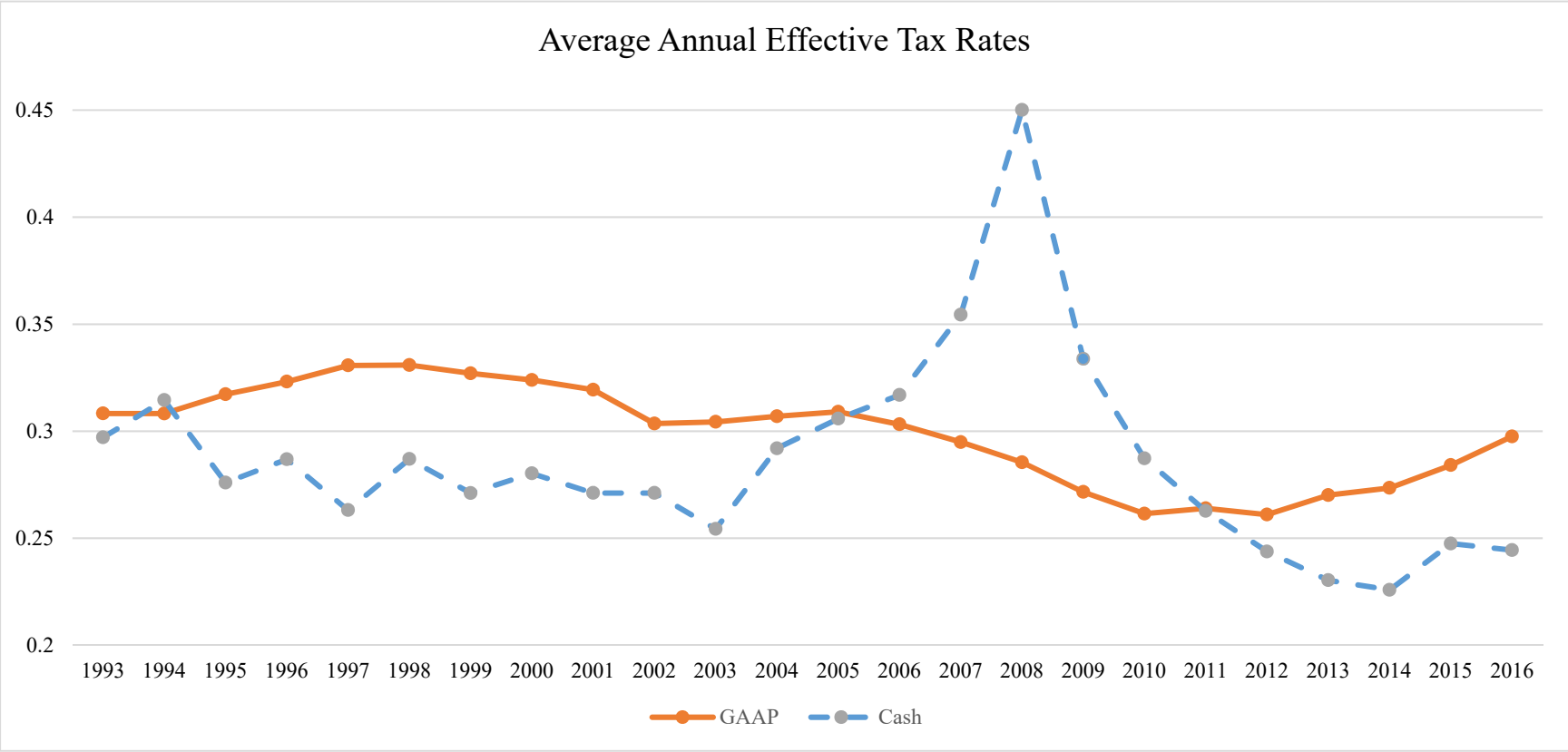
Panel B: Audit Fees on Size by Size Quartile

VARIABLES	(1) Log Linear Smallest Quartile	(2) Log Linear Second Quartile	(3) Log Linear Third Quartile	(4) Log Linear Largest Quartile
Size	0.4593***	0.3587***	0.5767***	0.5619***
	(53.74)	(15.05)	(17.83)	(43.55)
Observations	19,819	19,820	19,820	19,820
R-squared	0.306	0.041	0.071	0.384
Fixed Effects	None	None	None	None
Cluster	Firm	Firm	Firm	Firm

Note: We present the results of regression Equation (9). In both Panels the dependent variable is the natural log of audit fees. All variables are defined in Appendix A. *, **, and *** signify statistical significance at the 10%, 5%, and 1% significance level, respectively.

FIGURE 1

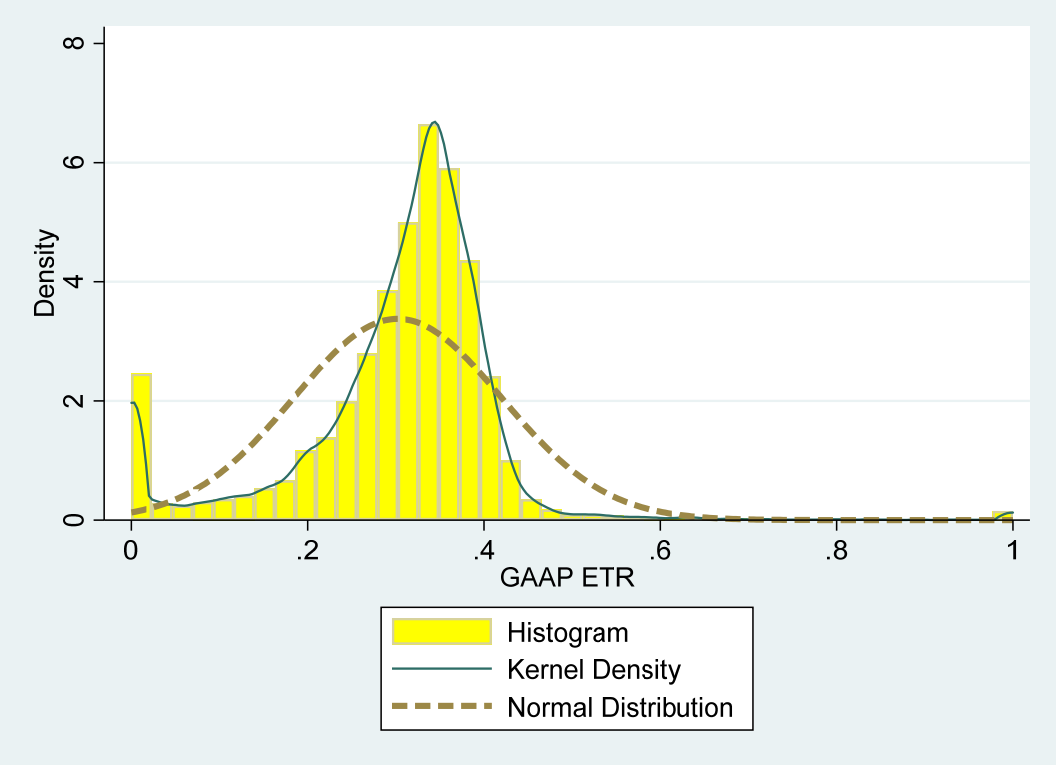
Time-trends in ETRs



Note: We present the annual average GAAP and Cash Effective Tax Rates in this figure for the financial services industry.

FIGURE 2

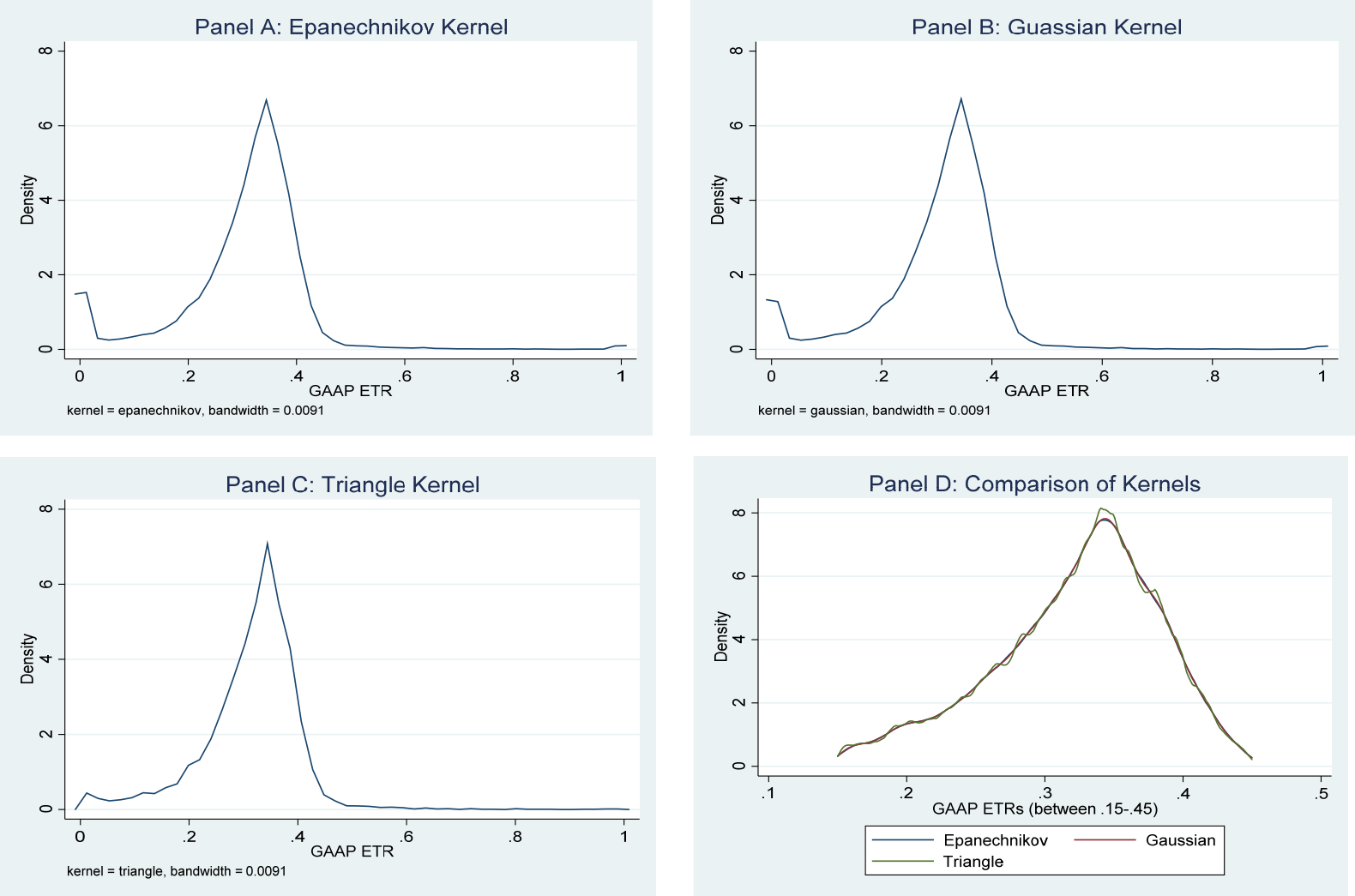
Histograms of GAAP ETRs in the Financial Service Industry



Note: This figure plots the distribution of GAAP ETRs for financial services firms from 1993 to 2016. The thick yellow bars represent the histogram estimates. A normal distribution is overlaid in the dotted brown line. The solid green line is the kernel density estimates.

FIGURE 3

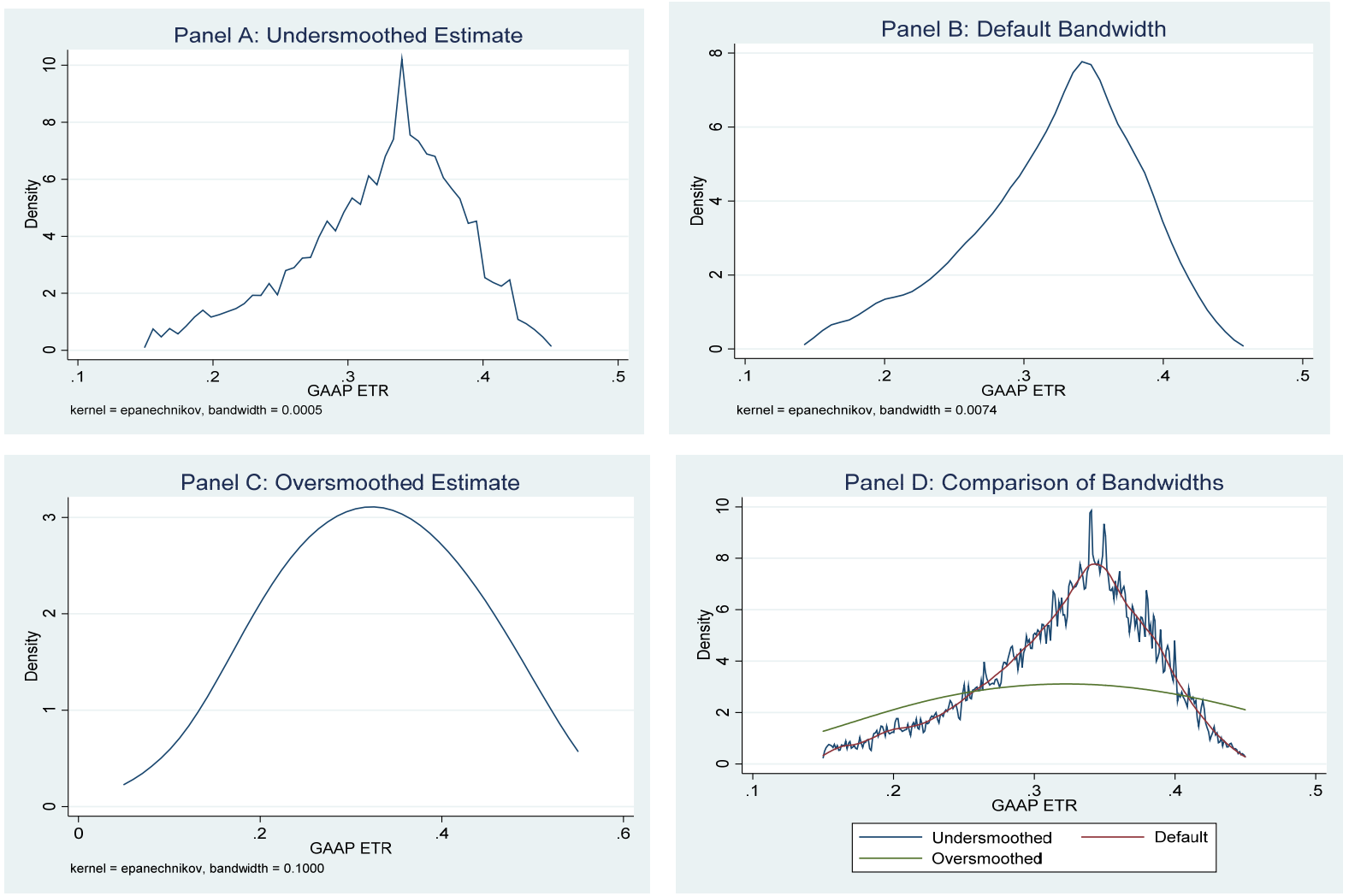
GAAP ETRs: Examining Choice of Weighting Function



Note: These figures plot the distribution of GAAP ETRs by financial services firms from 1993 to 2016. In Panels A-C we use the entire distribution of data. In Panel D, we examine ETRs between 15-45%. We allow STATA 15 to set the default bandwidth and compare three different weighting functions.

FIGURE 4

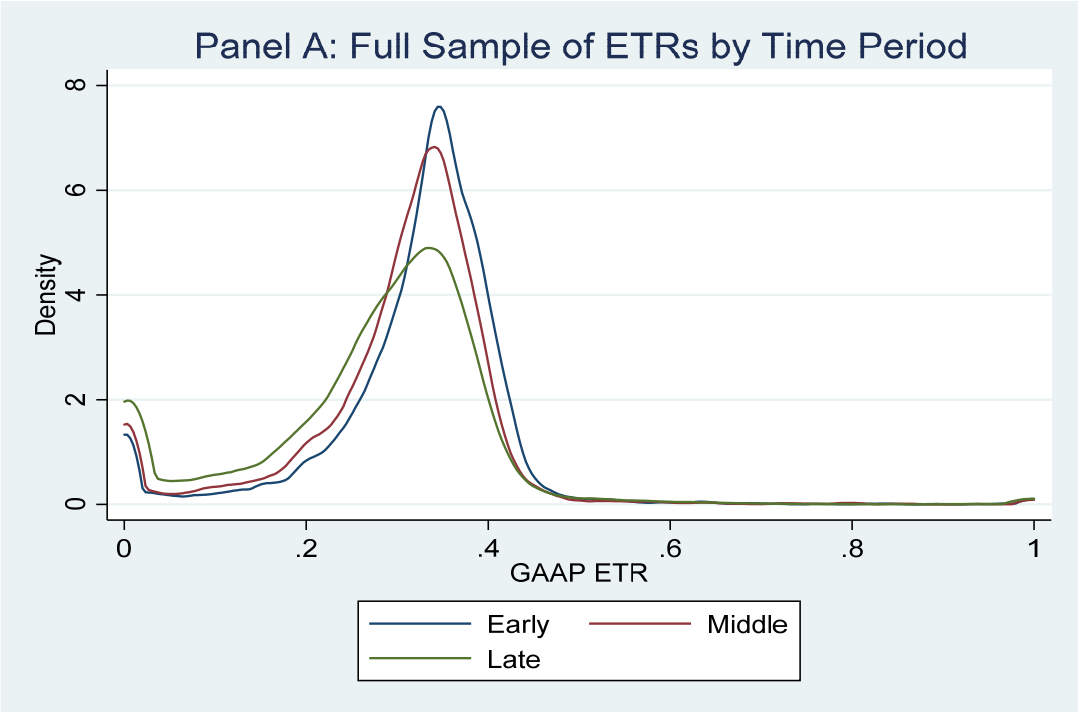
GAAP ETRs: Examining Choice of Bandwidth

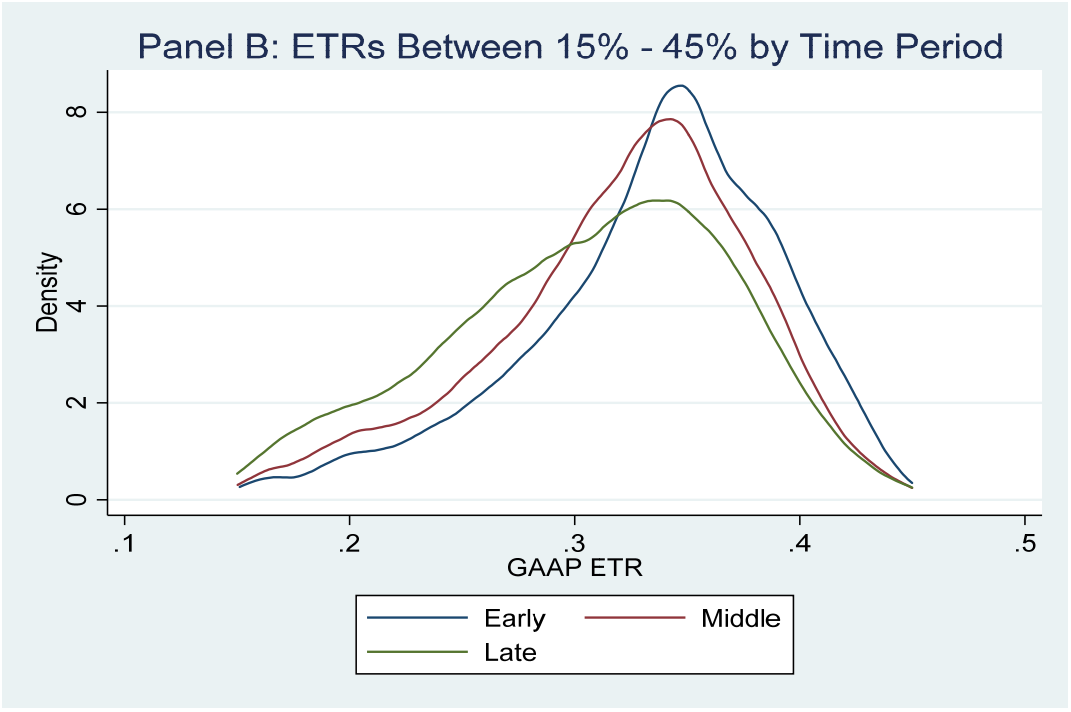


Note: These figures plot the distribution of GAAP ETRs by financial services firms from 1993 to 2016. In Panels A-C we use the entire distribution of data. In Panel D, we examine ETRs between 15-45%. We allow STATA 15 to set the default weighting function and compare different bandwidths.

FIGURE 5

Time-series Analysis

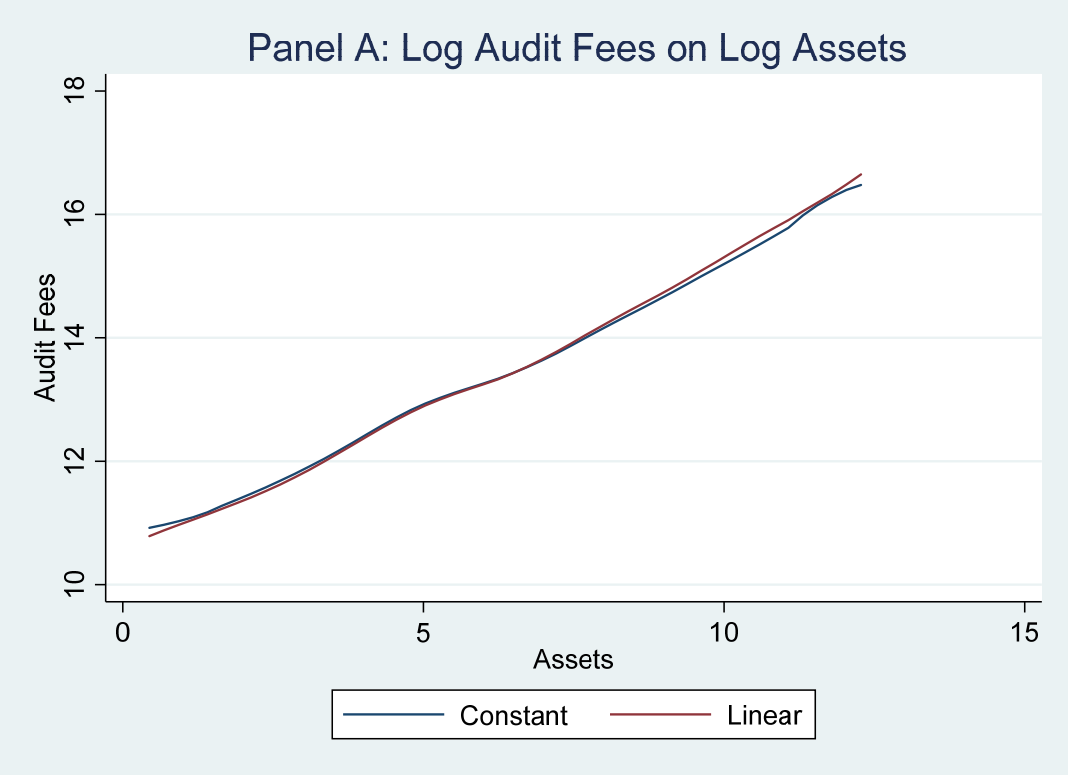


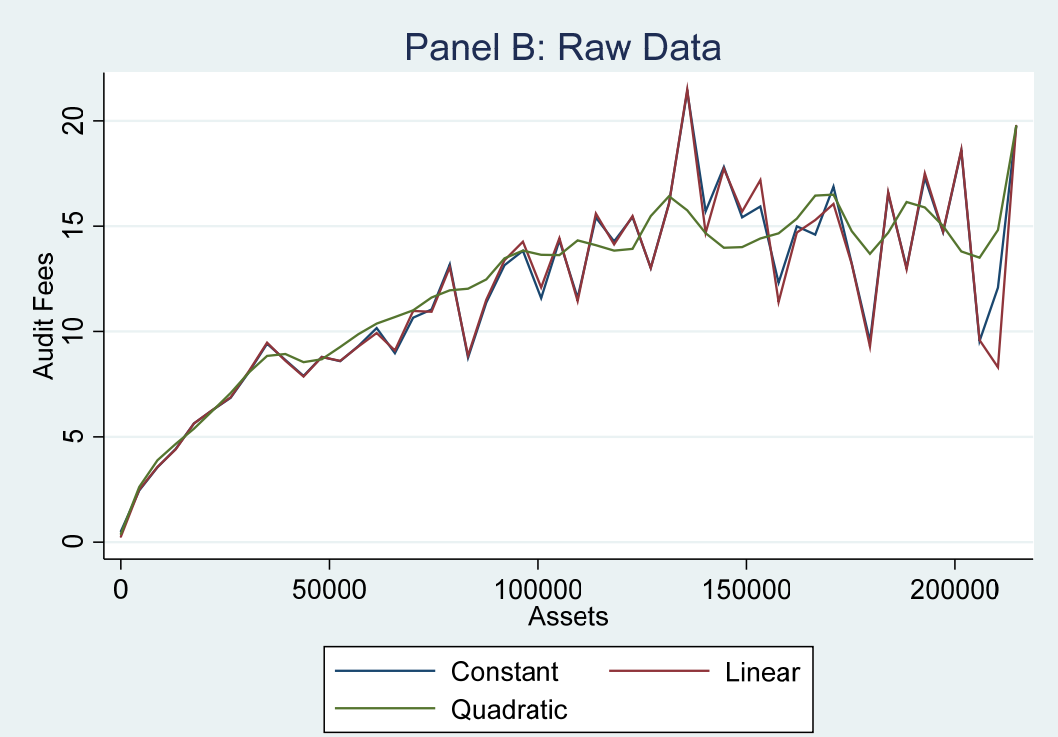


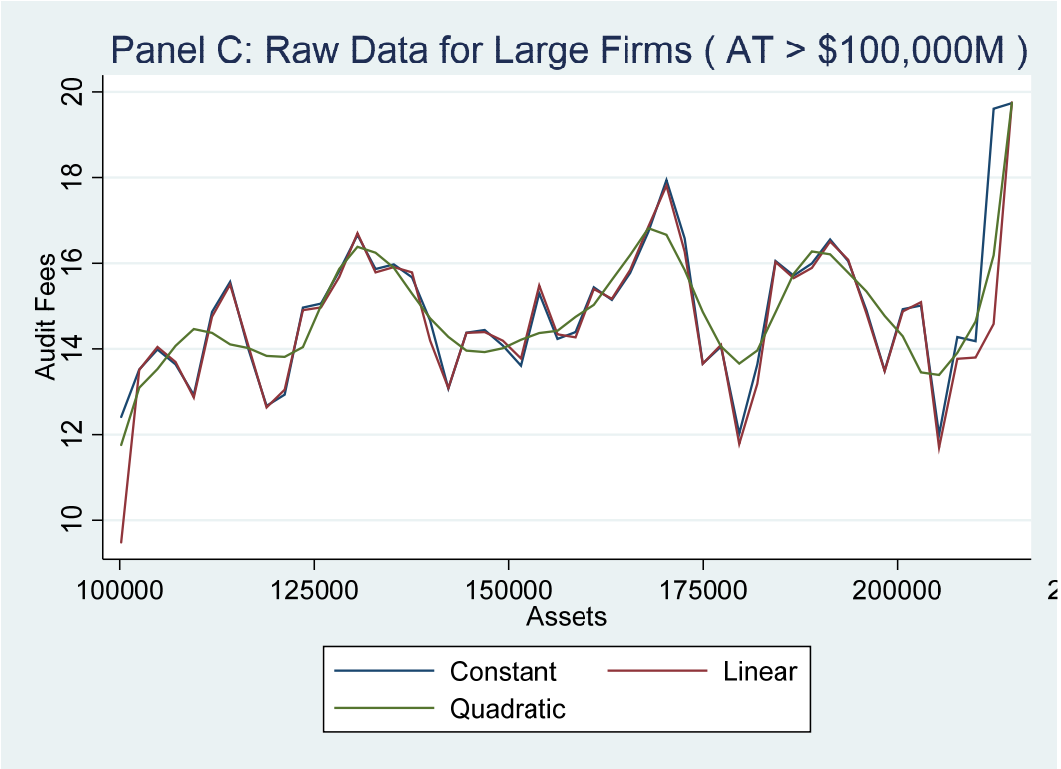
Note: These figures plot the distribution of GAAP ETRs by financial services firms from 1993 to 2016. In Panel A, we use the entire distribution of data, and in Panel B, we examine ETRs between 15-45%. We allow STATA 15 to set the default kernel and weighting function. Early years are pre-2002, middle years are 2002-2008, and ate years are post-2008.

FIGURE 6

Audit Fees and Size



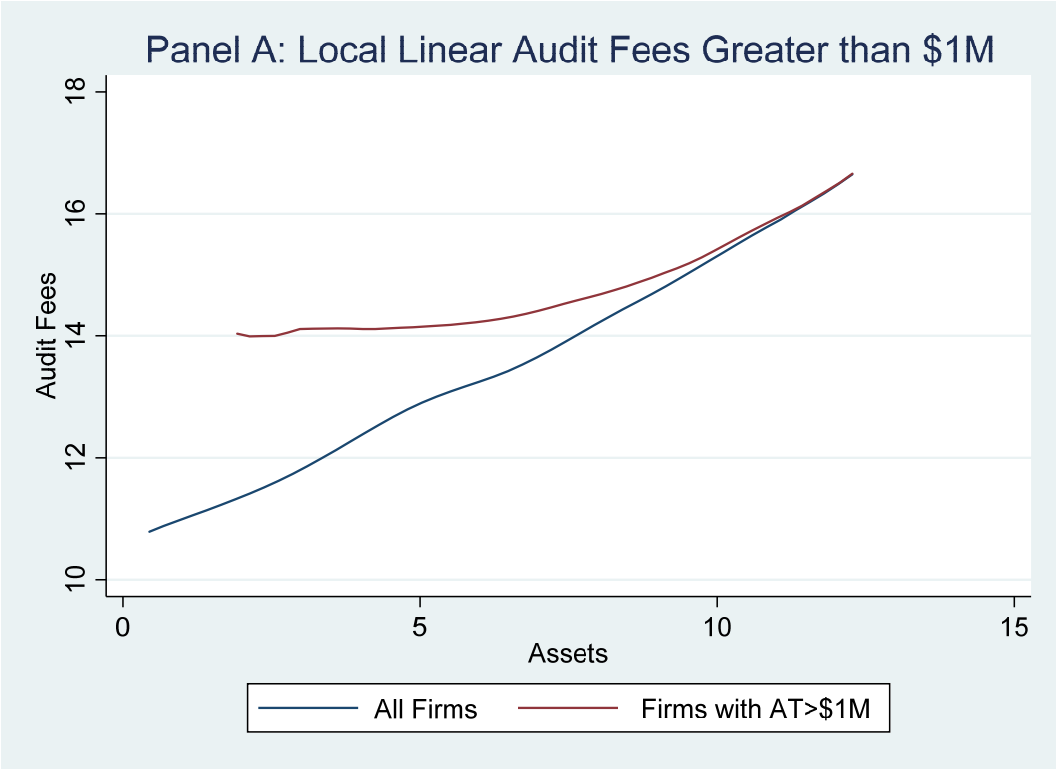


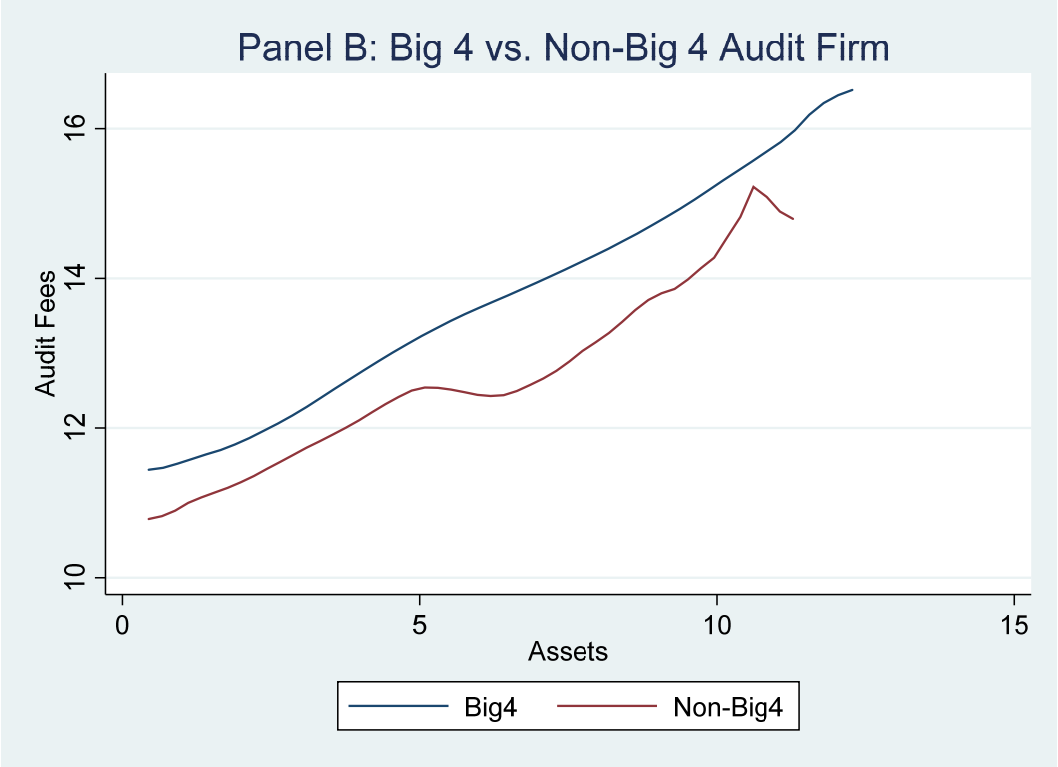


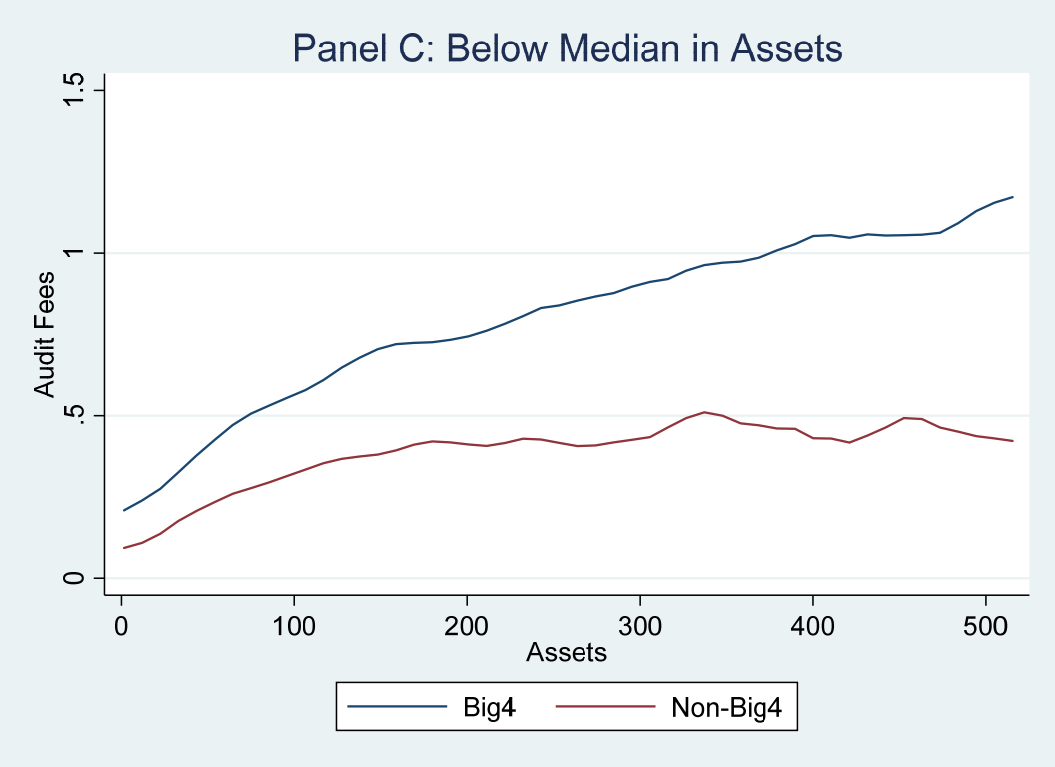
Note: These figures represent the nonparametric regression of audit fees on firm total assets. Audit fees and total assets are logged in Panel A but untransformed in Panels B and C. In Panels A through C, we report the kernel density (constant) and locally weighted linear regression. In Panels B and C we also report locally weighted quadratic regression. Panels A and B use the full sample and Panel C reports only firms with total assets over $100 Billion. Bandwidths are set to the default by Stata 15.

FIGURE 7

Audit Fees and Size: Additional Analysis







Note: These figures represent the nonparametric regression of audit fees on firm total assets. In Panel A we perform locally weighted linear regression with default bandwidth and weighting function. We report both the full sample and the sample of firm-years with at least $1M in audit fees. In Panels B and C we report the kernel density (constant) regression for Big 4 vs. Non-Big 4 auditors. In Panel C we limit the sample to firm-years with below median assets. In all three panels we use log transformed audit fees and total assets.