Power Law Heteroskedasticity

David J. Price*

December 24, 2025

Abstract

Power laws are common in economics, as in city and firm sizes, and can cause extreme heteroskedasticity. I show that estimators based on observations exhibiting this extreme heteroskedasticity may not be consistent or asymptotically normal and may have unreliable confidence intervals. These problems can occur even without heteroskedasticity if weighted estimators are used. I construct a quasi-maximum likelihood estimator to form more accurate estimates and more reliable inference. This estimator is broadly useful when weighting is considered to improve estimators' precision. Simulations confirm it improves estimation precision and inference, while a replication shows it can lead to substantially different results.

Keywords: heteroskedasticity, power laws, asymptotics, quasi-maximum likelihood

JEL Classification: C12, C13, C18

---

*University of Toronto; david.price@utoronto.ca. Thanks to Bisma Khan for excellent research assistance. I am grateful for feedback from Isaiah Andrews, Han Hong, Davide Malacrino, Morten Støstad, Alonso Villacorta, and seminar participants at Stanford University and the University of Toronto.

1 Introduction

As documented by Gabaix (2009, 2016) and others, power laws are common in economic phenomena, from city and firm size to stock markets, income, and wealth. These power laws generally indicate that a "rank-size" rule approximately applies: that is, when observations are ranked from largest to smallest, the size A_t of observation t is given by

$$A_t \approx A_1 t^{-s} \tag{1}$$

for some parameter $s > 0$. This relationship approximately holds with $s = 1$ for the population of cities (see, for example, Eeckhout (2004) and Rozenfeld et al. (2011)); with $s = 1$ for the size of firms (Axtell, 2001); with $s = 1.5$ for the size of trades on various stock markets (Gabaix et al., 2006); and with $s = 3$ for short-term stock price changes (Gopikrishnan et al., 1999). Research has helped illuminate the cause of these power laws, and some econometric research has explored how to estimate the parameter s based on observational data. However, it has not previously been noted, as I show here, that these power laws can create an extreme form of heteroskedasticity that can lead to estimators for which the law of large numbers and the central limit theorem fail to apply, or for which usual robust standard errors are not consistent for the estimators' true standard deviation.

To understand how such behavior can come about, suppose a researcher wants to measure how some explanatory variable (such as a city-level policy) affects an individual level outcome (such as student test scores) in a setting with one observation per city in the United States. Note that the outcome variable may be poorly estimated for small cities because the data may come from surveys with only a few observations. This means that if a researcher using an unweighted technique like ordinary least squares (OLS) adds observations about progressively

smaller cities, they may essentially be adding as much noise as signal—or potentially more noise than signal, depending on the value of s. To correct for this heteroskedasticity, common practice is to use weighted estimators, such as weighted least squares (WLS), with a weight proportional to city size. However, if the measurement error in the outcome variable does not vary much with city size, this weighting can induce extreme heteroskedasticity in the transformed data: in particular, the largest cities have such a large weight that the estimator remains a nondegenerate random variable no matter how many more cities are added to the data set.

In this paper, I prove theorems about large-sample behavior using a simple example—estimating the population mean—to ensure that proofs are tractable. These theorems can serve as guides for more complicated and commonly-encountered scenarios, such as regressions and instrumental variables models—for example, estimating a causal effect of a treatment that is randomly assigned at the group level. I begin by showing that, under suitable regularity conditions and if there is any heteroskedasticity proportional to A_t, the OLS estimator for the mean converges to the population mean if and only if $s < 1$. When $s = 1$, the variance of the estimator converges to a constant; when $s > 1$, the variance diverges to infinity. Despite the failure of the law of large numbers, suitable regularity conditions ensure that the OLS estimator will be asymptotically normally distributed, and the usual standard error estimator converges to the true standard deviation of the OLS estimator.

Next, I show that large-sample problems can arise with the WLS estimator, particularly if the errors are homoskedastic. Essentially, the weighting involved in WLS induces extreme heteroskedasticity in the data. In the case of homoskedastic errors, a WLS estimator for the mean is consistent if and only if $s \leq 1$, it is generally asymptotically normal if and only if

$s \leq \frac{1}{2}$, and the usual standard error estimator is consistent if and only if $s \leq \frac{1}{2}$. In fact, when $s = 1$, the standard error estimator remains a random variable as the sample size increases to infinity and its variance is particularly high relative to other values of s, rendering it essentially uninformative about the estimator's standard deviation.

The failures of OLS and WLS are generally most severe when a researcher chooses the wrong estimator—OLS under heteroskedasticity, or WLS under homoskedasticity. A better estimator would be one that optimally chooses whether to weight, or even chooses some intermediate estimator between weighted and unweighted. I therefore propose a quasi-maximum likelihood (QML) estimator, which estimates the extent of heteroskedasticity present in the data and then estimates the equations of the model as efficiently as possible. This estimator is based on limited information maximum likelihood, so it can be used to estimate a single-equation model (such as estimating an average or a regression coefficient) or a multiple-equations model (such as a model based on instrumental variables). I implement this estimator using a new Stata command, "regoptwgt," which in many cases can easily replace often-used commands such as "regress" and "ivregress 2sls."

In simulations, I show that the QML estimator is similar to both OLS under homoskedasticity and WLS under heteroskedasticity, but in other cases it is much better: it has a lower root mean squared error, and confidence intervals have size closer to the nominal value than WLS. This comparison holds for estimating a mean, a simple regression, and an instrumental variables model. To explore how these issues can affect real-world research, I replicate key results in Autor et al. (2013, 2020), two highly-cited papers that study the effect of imports from China on labor market outcomes and political polarization in the United States. In both papers, an observation is a commuting zone (similar to a metropolitan area), and results

are weighted by population. Standard errors for almost all estimates are smaller using QML than with the authors' weighted estimators. Many of the results also differ in an economically significant way, and statistical significance is often reduced. For example, the effect of import exposure on manufacturing employment is only about half as large when estimated using QML, relative to Autor et al. (2013)'s weighted estimate. Autor et al. (2020) find that import exposure significantly ($p = .007$) increases voter turnout; using QML, I find results that are about a third as large, and statistically insignificant ($p = .280$).

This paper contributes to a long literature, as reviewed by MacKinnon (2012), that seeks to understand the ways heteroskedasticity can lead to incorrect inference, and that creates tools to mitigate its effects. Early papers, including Eicker (1963), Huber (1967), and White (1980), developed what is now called HC1, the most commonly-used estimator for heteroskedasticity-robust standard errors, which is used when invoking the "robust" option in many Stata commands. It is HC1 that I find can often lead to incorrect inference. Since those initial contributions, researchers have noted that HC1 can perform poorly in small samples and developed techniques to address this issue. For example, MacKinnon and White (1985) developed HC3, a jackknife standard error estimator, which improves inference in this setting (even though HC3 standard errors can also be uninformative about the estimator's standard deviation, and some of the estimators themselves remain inefficient, inconsistent and not asymptotically normal). Despite these advances, other literature, such as Young (2019, 2022), find that HC3 and similar corrections are rarely applied in practice, causing inference issues in many highly-cited papers.

The poor performance of HC1 is generally thought to be due to small sample issues. The proofs econometricians use to understand estimators' large-sample behavior rely on regularity

conditions that are often difficult or impossible to verify in practice, so the literature often attributes any problematic simulation results to small-sample behavior without rigorous proof of how the problems arise. I show that, in a common setting, heteroskedasticity can lead to problems even with an arbitrarily large sample size. For example, researchers using common estimators on a data set with a million firms may assume that heteroskedasticity will not affect their results if they use robust standard errors; I show that this confidence is misplaced. The results in this paper can also help place small-sample concerns onto more solid theoretical footing. Further, literature on heteroskedasticity generally focuses on problems with standard errors; I find that consistency and asymptotic normality are problems, too. I also add to this literature by proposing a QML estimator that improves efficiency while reducing problems with inference.

Additionally, most of this literature focuses on proving sufficient conditions for asymptotic properties—for example, conditions under which estimators are consistent. I add to this literature by specifying necessary conditions—for example, when an estimator will not be consistent. Such necessary conditions tell the researcher when a strategy is a bad choice; if only a sufficient condition is described, the researcher may hope a different sufficient condition may apply to their setting.

In this literature, heteroskedasticity is often thought to be problematic only when the variance of the error term is correlated with covariates in the model. For example, Greene (2018, p. 305) notes that "if the heteroscedasticity is not correlated with the variables in the model, then at least in large samples, the ordinary least squares computations, although not the optimal way to use the data, will not be misleading." My results contradict at least the common understanding of this assertion: heteroskedasticity (or even an adjustment for

heteroskedasticity when none exists) can cause misleading results in large samples even if it is uncorrelated with covariates. However, note that if weights are considered to be part of a model, the quotation covers most misleading results—except the nonconvergence of unweighted estimators under heteroskedasticity with $s \geq 1$. Further, note that (following conventional wisdom) heteroskedasticity is only problematic in this paper if it exists conditional on observables—that is, when group size is known. With heteroskedasticity, an unweighted estimator, and $s > 1$, it would be better to limit estimation to only observations of the largest groups, but that is only possible with group size information; a weighted estimator is not even possible without knowing the size.

This paper also relates to a growing literature on cluster-robust inference. In simulations, I show that the same problematic inference in the setting of one observation per group (for example, per city or per firm) can also be present when the data are not aggregated. This arises, for example, when survey data are used, but data are not collapsed to one observation per group, and errors are allowed to be clustered within groups using CV1, the most common clustering technique, which is used when invoking the "cluster" option in Stata. Such techniques were developed soon after heteroskedasticity-robust techniques, with early advances by Liang and Zeger (1986), Moulton (1986), and Arellano (1987). Similarly to small-sample behavior of HC1, literature such as Carter et al. (2017) has noted that CV1 can perform poorly when the size of clusters has a lot of variation. As with heteroskedasticity, resampling techniques like the jackknife (Bell and McCaffrey, 2002) and bootstrap (Cameron et al., 2008) have been found to perform well in simulations; they also improve inference in this setting. Literature reviews on this topic include Cameron and Miller (2011) and MacKinnon (2019), with practical guides by Cameron and Miller (2015) and MacKinnon

et al. (2023). One paper that is particularly closely related is that by Chiang et al. (2023), which was developed independently (and, to my knowledge, after most results in this paper); they prove necessary and sufficient conditions for cluster-robust inference to be valid and mention that the power-law size distribution of cities can cause estimators in the clustered setting to be non-Gaussian. I contribute to this literature by showing a new connection to the heteroskedasticity literature: CV1 in this setting gives almost exactly the same (problematic) results as HC1. I also contribute by proposing a QML technique that can improve efficiency and inference in simulations if data are collapsed to one observation per group. Additionally, as with heteroskedasticity, researchers may erroneously believe that they have a large enough sample size—for example, people in one million firms—to not worry about small sample problems. This paper shows that this intuition may be incorrect.

Many of the issues documented here arise from weighting, and the proposed QML solution is, essentially, a weighted estimator. To my knowledge, there is not a large literature on the use of weighted estimators. Dickins (1990) notes that weighting by A_t—often intended to increase the precision of estimators—can instead reduce it. He also appears to suggest the same quasi-maximum likelihood estimator as in this paper, though he does not give details about it, provide code to help researchers use it, or extend it to the instrumental variables setting, as in this paper. To my knowledge, this QML estimator has never been used since then.[1] Solon et al. (2015) also note the potential for weighting to make estimators less efficient, along with critiquing other motivations for the use of weights. I contribute to this literature by noting that weights can cause estimators to be not just inefficient but

[1]Based on a review of every available paper listed in `https://ideas.repec.org/` as citing Dickins (1990).

also inconsistent and lacking in asymptotic normality, with incorrect conventional robust standard errors.

The remainder of this paper proceeds as follows. Section 2 describes the theoretical setup used in the remainder of the paper. Section 3 discusses OLS estimation, while Section 4 discusses WLS estimation. Section 5 introduces the QML estimator. Section 6 presents simulations of OLS, WLS, and QML estimators, while Section 7 takes two papers that use weighted estimators and replicates them using QML. Section 8 concludes.

2 Theoretical setup

To understand how power laws can lead to large-sample problems, I use a simple model in which a researcher wants to estimate a mean. Results from this setup are intended to give intuition for more general regression or instrumental variables settings. For example, researchers may see one observation per city, and want to estimate the causal effect of a treatment that is randomly assigned at the city level. Consistency and asymptotic normality of such advanced estimators typically rely on properties of means, so results here should be applicable in that setting. I show that more advanced estimators face similar issues using simulations in Appendix Section D.

Consider an economic system where individual unit it appears within group t. Suppose there are T groups, and A_t individuals within any group t, where $A_t = A_1 t^{-s}$. Now, consider a model where $y_{it} = \theta + \epsilon_{it}$ for some parameter θ, with $\epsilon_{it} = \eta_{it} + \nu_t$. Suppose that the error terms $\{\eta_{it}\}$, $\{\nu_t\}$ are mean-zero and mutually independent, with identical variances for each error term (though possibly different variances between them): $\mathbb{V}[\eta_{it}] = \sigma_\eta^2$ for all it,

and $\mathbb{V}[\nu_t] = \sigma_\nu^2$ for all t. Now, suppose that we wish to estimate θ, but only the group-wide average y_t is observed, where

$$y_t \equiv \frac{1}{A_t} \sum_{i=1}^{A_t} y_{it}. \tag{2}$$

As an example, θ may be underlying ability on a test, which is equal for all individuals. However, when taking the test, individuals' scores may vary from θ for two reasons: first, due to random variation at the individual level (because the test is not perfect, so there is measurement error); and second, due to factors that vary at the city level and affect test scores, such as the education system. (Note that these sources of error mean that even getting data from all people in all cities would not give us the population value of θ.) Scores are only observed by the econometrician as an average at the city level.

We now define

$$\eta_t \equiv A_t^{\frac{1}{2}} \frac{1}{A_t} \sum_{i=1}^{A_t} \eta_{it}, \tag{3}$$

so that $\{\eta_t\}$ are homoskedastic with variance $\mathbb{V}[\eta_t] = \sigma_\eta^2$. Using this definition, we can rewrite

$$\begin{aligned} y_t &= \theta + A_t^{-1} \sum_{i=1}^{A_t} (\eta_{it} + \nu_t) \\ &= \theta + \epsilon_t \end{aligned} \tag{4}$$

where $\epsilon_t \equiv A_t^{-\frac{1}{2}} \eta_t + \nu_t$.

Setting $A_1 = 1$ (without loss of generality, as this simply involves a change in σ_η^2), we can now write

$$\epsilon_t = t^{\frac{s}{2}} \eta_t + \nu_t. \tag{5}$$

The first term in this sum will have a smaller variance for larger groups (that is, those with smaller values of t) relative to smaller groups.

In Sections 3 and 4, I examine the properties of the sample average (OLS in this setting) and the average weighted by A_t (WLS). I show these properties as conditions on the parameter s, where higher s indicates more extreme heteroskedasticity. Interestingly, the frequently-seen $s = 1$—the value for cities and firms, for example—is a particularly complicated value: for many of the theorems below, $s = 1$ is an edge case between useful asymptotic results and the lack thereof. In this case, even if these results hold for $s = 1$, estimates may converge very slowly to their limits, which can also cause problems in practice, where sample sizes are not infinite. Of course, real-world data will not perfectly obey a power law; in this case, the following theorems can be used as approximations to the true behavior of the estimators.

The asymptotic results presented in this paper assume that the variance of ϵ_1 is fixed, and as the sample grows, later observations have progressively larger variances. However, asymptotics could work differently; for example, the variance of the ϵ_T could be fixed, so that as the sample size grows, the variance of ϵ_1 would shrink. I discuss that possibility in Appendix Section B; essentially, the OLS estimator will be consistent under that asymptotic assumption, but WLS may still not be consistent, and conclusions in this paper about asymptotic normality or the consistency of estimators for the standard errors will remain unchanged.

3 Standard unweighted parameter estimates

Often, a researcher will attempt to estimate θ as an unweighted mean,

$$\hat{\theta}_{uw} \equiv \frac{1}{T}\sum_{t=1}^{T} y_t = \theta + \frac{1}{T}\sum_{t=1}^{T}\left(t^{\frac{s}{2}}\eta_t + \nu_t\right). \tag{6}$$

This unweighted estimate may be made, for example, if the functional form of the heteroskedasticity (the A_t in this example) is not known *a priori*. In this section I will show that, under general conditions, a necessary and sufficient condition for $\hat{\theta}_{uw} \xrightarrow{p} \theta$ is that $s < 1$. Regardless of consistency, though, $\hat{\theta}_{uw}$ is generally asymptotically normal, and usual standard error estimators are consistent.

Theorem 3.1. *Suppose $\hat{\theta}_{uw}$ is defined as in Equation 6, where η_t and ν_t are mean-zero, independently distributed, and homoskedastic with finite variances σ_η^2 and σ_ν^2, respectively. If $s < 1$, then $\hat{\theta}_{uw} \xrightarrow{p} \theta$. If additionally $\{\eta_t^2\}$ and $\{\nu_t^2\}$ are each uniformly integrable; $\sigma_\eta^2 > 0$; and $s \geq 1$; then $\hat{\theta}_{uw} \not\xrightarrow{p} \theta$.*

Proof. See Appendix Section A. □

Thus the unweighted estimator will not generally be consistent if $s \geq 1$. Figure 1 plots the variance of the unweighted estimator, $\mathbb{V}\left[\hat{\theta}_{uw}\right] = \frac{\sigma_\eta^2}{T^2}\sum_{t=1}^{T} t^s + \frac{\sigma_\nu^2}{T}$. As seen in that figure, for $s > 1$, each additional observation actually makes the estimator worse. Even with $s = 1$, having an infinite number of observations leads to an estimator with about the same variance as if there are two observations with $s = 0$ (that is, with homoskedastic errors).

[Figure 1 here]

However, if both η_t^2 and ν_t^2 are uniformly integrable, then $\hat{\theta}_{uw}$ will be asymptotically normal for any s. To prove this, I will use the following lemma, which will also be used later in this paper. It may also be useful in applying these results to populations that do not obey a power law.

Lemma 3.2. *Suppose $\{\epsilon_t\}$ are mean-zero, independently distributed, homoskedastic random variables with finite variance σ_ϵ^2. Further, suppose $\{\epsilon_t^2\}$ are uniformly integrable, and that*

there is some function $g(T)$ *such that, for all* T, $g(T) \neq 0$*; and a function* $f(t)$ *such that*

$$\lim_{T\to\infty} \sup_{t\leq T} \frac{f(t)^2}{\sum_{s=1}^{T} f(s)^2} = 0. \tag{7}$$

Define $X_{Tt} \equiv g(T)f(t)\epsilon_t$*;* $S_T \equiv \sum_{t=1}^{T} X_{Tt}$*; and* $s_T^2 \equiv \sum_{t=1}^{T} \mathbb{V}[X_{Tt}]$*. Then* $\frac{S_T}{s_T} \xrightarrow{d} \mathbb{N}(0,1)$*.*

Proof. See Appendix Section A. □

Using this lemma, we can now prove the following.

Theorem 3.3. *Suppose* $\hat{\theta}_{uw}$ *is defined as in Equation 6, where* η_t *and* ν_t *are mean-zero, independently distributed, and homoskedastic with finite variances* σ_η^2 *and* σ_ν^2*, respectively; and that* $\{\eta_t^2\}$ *and* $\{\nu_t^2\}$ *are uniformly integrable. Then* $g(T)\left(\hat{\theta}_{uw} - \theta\right) \xrightarrow{d} \mathbb{N}(0,1)$ *for some function* $g(T)$*.*

Proof. See Appendix Section A. □

Thus we know that $\left(\mathbb{V}\left[\hat{\theta}_{uw}\right]\right)^{-\frac{1}{2}}\left(\hat{\theta}_{uw} - \theta\right) \xrightarrow{d} \mathbb{N}(0,1)$. However, $\mathbb{V}\left[\hat{\theta}_{uw}\right]$ is usually not known *a priori*. Thus to perform inference, we must find some feasibly estimated $\hat{V}$ such that $\frac{\hat{V}}{\mathbb{V}\left[\hat{\theta}_{uw}\right]} \xrightarrow{p} 1$; using Slutsky, we can then show that $\hat{V}^{-\frac{1}{2}}\left(\hat{\theta}_{uw} - \theta\right) \xrightarrow{d} \mathbb{N}(0,1)$.

If $s \geq 1$, the estimates themselves will not converge, so the standard proof that heteroskedasticity-robust standard errors are consistent does not go through. However, if η_t and ν_t have finite kurtosis, then these estimated standard errors will indeed be consistent.

Theorem 3.4. *Define* $\hat{\epsilon}_t \equiv y_t - \hat{\theta}_{uw}$*, and* $\hat{V} \equiv \frac{T}{T-1}\frac{1}{T^2}\sum_{t=1}^{T}\hat{\epsilon}_t^2$*. As above, suppose that* $\hat{\theta}_{uw}$ *is defined as in Equation 6, where* η_t *and* ν_t *are mean-zero; independently distributed; homoskedastic with finite variances* σ_η^2 *and* σ_ν^2*, respectively; and have uniformly bounded kurtosis. Then* $\frac{\hat{V}}{\mathbb{V}\left[\hat{\theta}_{uw}\right]} \xrightarrow{p} 1$*.*

Proof. See Appendix Section A. □

4 Standard weighted parameter estimates

In empirical work, researchers often use population weights. That is, they estimate

$$\hat{\theta}_{pw} \equiv \frac{1}{\sum_{t=1}^{T} A_t} \sum_{t=1}^{T} A_t y_t. \tag{8}$$

Continuing to assume that $y_t = \theta + A_t^{-\frac{1}{2}} \eta_t + \nu_t$, and $A_t = A_1 t^{-s}$, this becomes

$$\hat{\theta}_{pw} = \theta + \frac{1}{\sum_{t=1}^{T} t^{-s}} \sum_{t=1}^{T} \left(t^{-\frac{s}{2}} \eta_t + t^{-s} \nu_t \right). \tag{9}$$

Under suitable regularity conditions, a necessary and sufficient condition for consistency is that $s \leq 1$.

Theorem 4.1. *Suppose $\hat{\theta}_{pw}$ is defined as in Equation 9, where η_t and ν_t are mean-zero, independently distributed, and homoskedastic with finite variances σ_η^2 and σ_ν^2, respectively, at least one of which is non-zero. Then $\hat{\theta}_{pw} \xrightarrow{p} \theta$ if and only if $s \leq 1$.*

Proof. See Appendix Section A. □

Some caution should be taken in interpreting the lack of convergence of the η_t term. This is because the motivation for the η_t terms is from measurement error, which might be quite accurate due to the law of large numbers. In fact, the WLS estimate is numerically identical to the mean of all observations in all cities; so if the only error is from idiosyncratic individual-level differences, rather than city level variation, then the law of large numbers may approximately apply to this error. Lack of convergence is therefore more interesting and useful for the ν_t term.

Despite the lack of convergence for the weighted estimator, it has some desirable properties. Figure 2 shows the variance of the WLS estimator, $\mathbb{V}\left[\hat{\theta}_{pw}\right] = \sigma_\eta^2 \left(\sum_{t=1}^{T} t^{-s}\right)^{-1} +$

$\sigma_\nu^2 \left(\sum_{t=1}^T t^{-s}\right)^{-2} \sum_{t=1}^T t^{-2s}$. For any s, adding more observations always improves the accuracy (as measured by the variance of the estimator). In fact, if $\sigma_\eta^2 = 1$ and $\sigma_\nu^2 = 0$, then this weighted estimator is generalized least squares, so $\hat{\theta}_{pw}$ is also the best linear unbiased estimator by the Gauss-Markov theorem.

[Figure 2 here]

In addition to the lack of convergence, the weighted estimator will often fail to be asymptotically normal. The following lemma, a partial converse of Lemma 3.2, will be useful in proving this.

Lemma 4.2. *Suppose $\{\epsilon_t\}$ are mean-zero, independently distributed, homoskedastic random variables with finite variance $\sigma_\epsilon^2 > 0$. Further, suppose there is some function $g(T)$ such that for all T, $g(T) \neq 0$; and a finite-valued function $f(t) > 0$ such that*

$$\lim_{T\to\infty} \frac{f(1)^2}{\sum_{t=1}^T f(t)^2} = C^2 \tag{10}$$

for some finite constant $C > 0$. Define $X_{Tt} \equiv g(T)f(t)\epsilon_t$; $S_T \equiv \sum_{t=1}^T X_{Tt}$; and $s_T^2 \equiv \sum_{t=1}^T \mathbb{V}[X_{Tt}]$. Then $\frac{S_T}{s_T}$ is not generally asymptotically normal, in the sense that, for all t, holding fixed the distribution of $\{\epsilon_k\}$ for $k \neq t$, there is at most one distribution of ϵ_t for which $\frac{S_T}{s_T} \xrightarrow{d} \mathbb{N}(0,1)$.

Proof. See Appendix Section A. □

Essentially, this lemma says that a sum satisfying certain conditions will converge to a normal distribution only if the summands have a particular relationship (for example, if all summands are already normal). The following theorem shows that this lemma applies to the weighted estimator for certain values of s.

Theorem 4.3. *Suppose $\hat{\theta}_{pw}$ is defined as in Equation 9, where $\{\eta_t\}$ and $\{\nu_t\}$ are mean-zero; independently distributed; and homoskedastic with finite variances σ_η^2 and σ_ν^2, respectively, with at least one strictly greater than zero; and $\{\eta_t^2\}$ and $\{\nu_t^2\}$ are uniformly integrable. If $s \leq \frac{1}{2}$, then $\hat{\theta}_{pw}$ is asymptotically normal (with suitable normalization). If $\frac{1}{2} < s \leq 1$, then $\hat{\theta}_{pw}$ is generally (in the sense of Lemma 4.2) asymptotically normal if and only if $\sigma_\eta^2 > 0$. If $s > 1$, then $\hat{\theta}_{pw}$ is not generally asymptotically normal in the sense of Lemma 4.2.*

Proof. See Appendix Section A. □

The extent to which the central limit theorem fails to apply can be inferred from excess kurtosis of the estimator, defined as $\mathbb{E}\left[\frac{\left(\hat{\theta}_{pw}-\theta\right)^4}{\mathbb{V}\left[\hat{\theta}_{pw}\right]^2}\right] - 3$. Excess kurtosis gives us a rough understanding of the weight on the tails of the distribution, which determines the extent to which p-values based on a normal approximation will be accurate (assuming accurate standard errors). When the central limit theorem applies, excess kurtosis will converge to 0, the value for a normal distribution. The following propositions calculate excess kurtosis in this setting, defining $H_\alpha(T) \equiv \sum_{t=1}^{T} t^{-\alpha}$ (and suppressing the argument for simplicity).

Proposition 4.4. *Define $\kappa_\eta^4 \equiv \frac{E(\eta_t^4)}{\sigma_\eta^4}$. If $\sigma_\nu^2 = 0$, $\mathbb{E}\left[\frac{\left(\hat{\theta}_{pw}-\theta\right)^4}{\mathbb{V}\left[\hat{\theta}_{pw}\right]^2}\right] - 3 = \left(\kappa_\eta^4 - 3\right)\frac{H_{2s}}{H_s^2}$.*

Proof. See Appendix Section A. □

Proposition 4.5. *Define $\kappa_\nu^4 \equiv \frac{E(\nu_t^4)}{\sigma_\nu^4}$. If $\sigma_\eta^2 = 0$, $\mathbb{E}\left[\frac{\left(\hat{\theta}_{pw}-\theta\right)^4}{\mathbb{V}\left[\hat{\theta}_{pw}\right]^2}\right] - 3 = \left(\kappa_\nu^4 - 3\right)\frac{H_{4s}}{(H_{2s})^2}$.*

Proof. See Appendix Section A. □

Excess kurtosis for various values of s is shown in Figure 3. For $s = 1$, excess kurtosis of the ν_t term with infinite observations is approximately equal to excess kurtosis where $s = 0$ (that is, homoskedasticity) and 3 data points. Thus if we do not think that the sum of 3

homoskedastic variables will be sufficiently normal, we should not think that the sum of an infinite number of $s = 1$ extreme heteroskedastic errors will lead to a sufficiently normal estimator.

[Figure 3 here]

One important note is that, although the distribution of the η_t term may fail to be generally normal, it may in fact be close to normal. This is because the η_t terms themselves often come from a measured variable, which might be approximately normal due to the central limit theorem. (This is related to the note, above, that lack of consistency for the η_t term is somewhat misleading.) As with consistency, then, the lack of asymptotic normality is more interesting and useful for the ν_t term.

Asymptotically, non-normality with $\frac{1}{2} < s \leq 1$ only occurs when $\sigma_\eta^2 = 0$, because the η_t term dominates the error otherwise. However, it is possible that σ_η^2 will be non-zero, but small enough that the $\{\nu_t\}$ terms will dominate the error for a finite number of observations. In this case, the asymptotics based on $\sigma_\eta^2 = 0$ may come closest to approximating the true distribution.

A final question is whether standard errors will be consistent—that is, whether $\frac{\hat{V}}{\mathbb{V}\left[\hat{\theta}_{pw}\right]} \xrightarrow{p} 1$, where $\hat{V}$ is the standard heteroskedasticity-robust standard errors. Where the estimator is asymptotically normal, this is clearly an important question. If the estimator is non-normal, even perfect standard errors should be interpreted cautiously, as the usual use of standard errors in testing is based on (asymptotic) normality. Still, correctly-estimated standard errors could at least give an idea of the dispersion of the distribution of the estimator. In fact, usual standard error estimators may not be consistent either. To explore the distribution of these standard errors, I begin with their expected value.

Proposition 4.6. *Define* $\hat{\epsilon}_t \equiv y_t - \hat{\theta}_{pw}$, *and* $\hat{V} \equiv \frac{T}{T-1}\left(\sum_{t=1}^{T} t^{-s}\right)^{-2} \sum_{t=1}^{T} t^{-2s}\hat{\epsilon}_t^2$. *As above, suppose that* $\hat{\theta}_{pw}$ *is defined as in Equation 9, where* η_t *and* ν_t *are mean-zero; independently distributed; and homoskedastic with finite variances* σ_η^2 *and* σ_ν^2, *respectively. Then* $\mathbb{E}\left[\frac{\hat{V}}{\mathbb{V}\left[\hat{\theta}_{pw}\right]}\right] = \frac{T}{T-1}\left(1 - 2H_s^{-1}\frac{\sigma_\eta^2 H_{2s} + \sigma_\nu^2 H_{3s}}{\sigma_\eta^2 H_s + \sigma_\nu^2 H_{2s}} + H_s^{-2}H_{2s}\right)$.

Proof. See Appendix Section A. □

The rate at which $\mathbb{E}\left[\frac{\hat{V}}{\mathbb{V}\left[\hat{\theta}_{pw}\right]}\right]$ converges to 1, for various sample sizes and with different values of s, is shown in Figure 4. Note that higher values indicate that $\hat{V}$ is more of an underestimate. Therefore, in that figure, we see that expected value of estimated standard errors are always smaller than the true standard deviations, and that the expected value only converges slowly to the truth, particularly for homoskedastic errors.

[Figure 4 here]

In fact, for the heteroskedastic η term, this expected value converges to 1 if and only if $s \leq 1$. For the homoskedastic ν term it converges to 1 if and only if $s \leq \frac{1}{2}$. (As with normality, in small sample, convergence of the estimated variance may closely approximate the worse homoskedastic case if the $\{\nu_t\}$ terms dominate even if $\sigma_\eta^2 > 0$.) As demonstrated in the following theorem, in cases of non-convergence, $\frac{\hat{V}}{\mathbb{V}\left[\hat{\theta}_{pw}\right]}$ is a nondegenerate random variable even in the limit, so in those cases standard errors provide little reliable guidance on the standard deviation of estimators, causing problems with inference. Simulations, shown in Section 6, demonstrate this issue in small sample.

Theorem 4.7. *Define* $\hat{\epsilon}_t \equiv y_t - \hat{\theta}_{pw}$, *and* $\hat{V} \equiv \frac{T}{T-1}\left(\sum_{t=1}^{T} t^{-s}\right)^{-2} \sum_{t=1}^{T} t^{-2s}\hat{\epsilon}_t^2$. *As above, suppose that* $\hat{\theta}_{pw}$ *is defined as in Equation 9, where* η_t *and* ν_t *are mean-zero; independently distributed; homoskedastic with finite variances* σ_η^2 *and* σ_ν^2, *respectively, at least one of which*

is non-zero; and have uniformly bounded kurtosis. If $s \leq \frac{1}{2}$*, then* $\frac{\hat{V}}{\mathbb{V}\left[\hat{\theta}_{pw}\right]} \xrightarrow{p} 1$*. If* $\frac{1}{2} < s \leq 1$*, and additionally* ν_t *takes on at least 3 values for some observation* t*, then* $\frac{\hat{V}}{\mathbb{V}\left[\hat{\theta}_{pw}\right]} \xrightarrow{p} 1$ *if and only if* $\sigma_\eta^2 > 0$*. If* $s > 1$*, then* $\frac{\hat{V}}{\mathbb{V}\left[\hat{\theta}_{pw}\right]}$ *does not converge.*

Proof. See Appendix Section A. □

To understand the extent to which $\hat{V}$ remains a random variable as the sample size increases, Figure 5 plots the variance of $\frac{\hat{V}}{\mathbb{V}\left[\hat{\theta}_{pw}\right]}$ for various values of s. Interestingly, for homoskedastic errors, among the values of s shown, $\mathbb{V}\left[\frac{\hat{V}}{\mathbb{V}\left[\hat{\theta}_{pw}\right]}\right]$ is largest for $s = 1$, suggesting that $s = 1$—a particularly common value of s—leads to standard errors that are particularly random. Note that for $s = 1$, not only is the variance of $\frac{\hat{V}}{\mathbb{V}\left[\hat{\theta}_{pw}\right]}$ quite large, it only gets larger as the number of observations increases. Higher moments (for example, skewness and kurtosis) of $\frac{\hat{V}}{\mathbb{V}\left[\hat{\theta}_{pw}\right]}$ are also quite extreme, so outliers are common. For example, in a simulation of WLS from Section 6, using 1,000 homoskedastic exponentially distributed ν terms, the 5th and 95th percentiles of $\frac{\hat{V}}{\mathbb{V}\left[\hat{\theta}_{pw}\right]}$ are .17 and 2.6, respectively. In other words, in about 5% of simulations, estimated variance is more than six times smaller than the true variance. In this setting, estimated standard errors may still be somewhat useful for inference, as discussed in Section 6, but they are essentially uninformative about the true standard deviation of the estimators. In Section 6, I also show that HC3 standard errors improve inference, particularly for a regression coefficient. However, even HC3 standard errors are similarly inaccurate—they are just larger on average. For example, in a simulation of WLS in a simple regression with HC3 standard errors, using 1,000 homoskedastic exponentially distributed ν terms, the 5th and 95th percentiles of $\frac{\hat{V}}{\mathbb{V}\left[\hat{\theta}_{pw}\right]}$ are .16 and 3.3, respectively.

Some researchers will estimate both OLS and WLS and choose to report the results from

the estimator with the smaller standard errors. In addition to the usual problems with pre-testing, the lack of consistency of standard errors means that there is no reason to believe that the researchers' choice will reflect the most accurate estimator.

[Figure 5 here]

5 Quasi-maximum likelihood

In Sections 3 and 4, we saw that in cases where data follow a power law, estimators face the most severe issues when the wrong one is used. For example, OLS is inconsistent only when errors are heteroskedastic, while WLS faces worse problems when errors are homoskedastic. A natural estimator, then, is one that defaults to the appropriate estimator in the homoskedastic and heteroskedastic cases while also optimally weighting for intermediate cases. Quasi-maximum likelihood (QML), in which both $\{\eta_t\}$ and $\{\nu_t\}$ are modeled as normal random variables, is a natural choice in this setting. I should note that, in cases when consistency and asymptotic normality fail with OLS and WLS estimators, I likewise cannot prove such large-sample properties with QML. However, in small-sample simulations, QML performs substantially better than either; see Section 6.

This QML estimator also has the advantage of being helpful even in settings where large-sample theory does apply. In many empirical settings, researchers are left wondering whether to use weights. This QML estimator can improve precision in such settings.

Some researchers may prefer weighted estimators to QML because they are concerned about heterogeneity and wish to estimate a population average partial effect. However, as noted by Solon et al. (2015), weighting by population also will not result in an estimator

for the population average partial effect, even without the heteroskedasticity discussed here. Given that neither estimator will result in the desired outcome, it is reasonable to use the estimator that can result in better inference and that is more precise (indeed, can be a maximum likelihood estimator) under homogeneity.

This QML estimator is based on limited information maximum likelihood (LIML), which can be used in both single-equation settings (analogous to OLS) or multiple-equation settings (analogous to two-stage least squares). Suppose we have a system of J equations and T observations, where observation t has weight A_t. We will write equation j for observation t as

$$y_{tj} = \mathbf{z}'_{tj}\boldsymbol{\delta}_j + \epsilon_{tj}, \tag{11}$$

where some of the variables $\mathbf{z}_{tj}$ may be endogenous. We will assume the error term for observation t in equation j is

$$\epsilon_{tj} \equiv A_t^{-\frac{1}{2}}\eta_{tj} + \nu_{tj}, \tag{12}$$

where all random variables $\{\eta_{tj}\}$ and $\{\nu_{tj}\}$ are independent from each other, except that within an observation, but across equations, the η terms may be correlated with each other, and the ν terms may be correlated with each other. Note that this specification allows for some equations to be homoskedastic and others to be heteroskedastic within the same model.

If we further assume that, for a given observation t, all η_{tj} are jointly normal, and all ν_{tj} are jointly normal, then the non-constant part of the objective function for LIML, following Hayashi (2002, p. 539), is

$$\mathcal{L} = -\frac{1}{2T}\sum_{t=1}^{T}\left[\ln\left(|\boldsymbol{\Sigma}_t|\right) + \boldsymbol{\epsilon}'_t\boldsymbol{\Sigma}_t^{-1}\boldsymbol{\epsilon}_t\right] \tag{13}$$

where $\boldsymbol{\epsilon}_t$ is a matrix of error terms ϵ_{tj}, and $\boldsymbol{\Sigma}_t$ is the covariance matrix of $\boldsymbol{\epsilon}_t$, with diagonal elements

$$\begin{aligned} \boldsymbol{\Sigma}_{t,jj} &= A_t^{-1}\sigma_{\eta,j}^2 + \sigma_{\nu,j}^2 && (14) \\ &= A_t^{-1}\exp(s_{\eta,j}) + \exp(s_{\nu,j}) && (15) \end{aligned}$$

where we define $s_{\eta,j} \equiv \ln(\sigma_{\eta,j}^2)$ and $s_{\nu,j} \equiv \ln(\sigma_{\nu,j}^2)$ for computational reasons. Off-diagonal elements are

$$\begin{aligned} \boldsymbol{\Sigma}_{t,j_1\neq j_2} &= A_t^{-1}\text{cov}\left(\eta_{j_1},\eta_{j_2}\right) + \text{cov}\left(\nu_{j_1},\nu_{j_2}\right) && (16) \\ &= A_t^{-1}\sigma_{\eta,j_1}\sigma_{\eta,j_2}\text{corr}\left(\eta_{j_1},\eta_{j_2}\right) + \sigma_{\nu,j_1}\sigma_{\nu,j_2}\text{corr}\left(\nu_{j_1},\nu_{j_2}\right) && (17) \\ &= A_t^{-1}\sqrt{\exp(s_{\eta,j_1})\exp(s_{\eta,j_2})}\left(2\ \text{invlogit}\left(c_{\eta,j_1j_2}\right)-1\right) \\ &\quad +\sqrt{\exp(s_{\nu,j_1})\exp(s_{\nu,j_2})}\left(2\ \text{invlogit}\left(c_{\nu,j_1j_2}\right)-1\right) && (18) \end{aligned}$$

where we define $c_{\eta,j_1j_2} \equiv \text{logit}\left(\left(\text{corr}\left(\eta_{j_1},\eta_{j_2}\right)+1\right)/2\right)$ and $c_{\nu,j_1j_2} \equiv \text{logit}\left(\left(\text{corr}\left(\nu_{j_1},\nu_{j_2}\right)+1\right)/2\right)$, again for computational reasons.

In usual use, LIML is estimated using a closed-form expression by concentrating out $\boldsymbol{\Sigma}$ (Davidson and MacKinnon, 1993, p.639–649). Such a trick is not possible in this setting because $\boldsymbol{\Sigma}_t$ does not consist of primitive parameters and also because it varies by observation t, so $\frac{\partial\mathcal{L}}{\partial\boldsymbol{\Sigma}}$ is not a meaningful concept. Parameters are therefore estimated by maximizing the full log-likelihood function.

This QML estimation technique is performed by the new Stata command "regoptwgt," which I detail in Appendix Section C.

6 Simulations

In this section, I perform Monte Carlo simulations of various estimators under differing levels of heteroskedasticity. I focus on the situation where $s = 1$, because that is a frequently occurring parameter (holding for cities and firms, for example); because it is a particularly difficult situation to understand theoretically, as it is an edge case between many asymptotic results holding and not; and because it is a situation in which the usual estimated standard errors have particularly poor behavior, as noted in Section 4. I simulate the following model with $T = 1{,}000$ observations and use various estimators for the expected value $\mathbb{E}[\epsilon_t]$:

$$\epsilon_t = k^{\frac{1}{2}} t^{\frac{s}{2}} \eta_t + \nu_t \tag{19}$$

$$\eta_t \sim \mathbb{N}(0, 1) \tag{20}$$

$$\nu_t \sim \text{Exp}(1) - 1 \tag{21}$$

$$k = \frac{\frac{H_{2s}(T)}{H_s(T)^2} - \frac{1}{T}}{\frac{H_{-s}(T)}{T^2} - H_s(T)^{-1}} e^{\Phi^{-1}(h)}, \tag{22}$$

where h is the level of heteroskedasticity on the x-axis and $\Phi^{-1}(\cdot)$ is the inverse of a standard normal CDF. Note that k is chosen such that heteroskedasticity is defined by $h = \Phi\left(\ln\left(\frac{\sigma_\eta/\sigma_\nu}{R}\right)\right)$, where R^2 is the variance ratio that would cause $\mathbb{V}\left[\hat{\theta}_{uw}\right] = \mathbb{V}\left[\hat{\theta}_{pw}\right]$. This means that the root mean squared (RMS) error of the unweighted and weighted estimates of an average should be exactly equal when $h = .5$.

Based on simulations, the QML estimator performs better than OLS or WLS under a range of levels of heteroskedasticity. Figure 6 shows the RMS error of OLS and WLS relative to QML, while Figure 7 shows the actual size of estimated 95% confidence intervals for each technique. The RMS error of WLS is over 5 times higher than OLS or QML

under homoskedasticity, while the RMS of OLS is almost 2 times higher than either WLS or QML under full heteroskedasticity. As expected, confidence intervals are correctly sized for OLS and close to correct for QML, while tests with nominal 5% size using WLS have actual size around 12%. Interestingly, confidence intervals for WLS are slightly incorrect even when errors are fully heteroskedastic—that is, when WLS is the best linear unbiased estimator—supporting the theoretical results above for the case where $\sigma_\nu^2 = 0$.

[Figure 6 here]

[Figure 7 here]

Given the failure of the central limit theorem and inconsistency of WLS standard errors, it is perhaps surprising that confidence interval problems are not more severe. From examining simulation results, it appears that when an estimator has an outlying value, the standard error often is larger than usual, which keeps rejection rates closer to nominal than might be expected. This does mean, however, that in this setting, standard errors should only be considered as building blocks of confidence intervals rather than taken at face value as the standard deviation of the estimator.

Appendix Section D presents the results of similar simulations for a regression coefficient and a coefficient estimated with instrumental variables (IV); the size of confidence intervals of different nominal sizes; the size of confidence intervals estimated using different techniques; and similar results when $s = 2$. In graphs showing size, the line marked "Disaggregated, CV1" simulates the setting described at the start of Section 2 without aggregating to the group level; instead, errors are clustered using CV1. Estimates will be the same as the weighted results; however, I also find that test size is nearly identical to the weighted case. (Interestingly, an exception is the size of a disaggregated test of an IV coefficient when

$\eta = 0$.) The similarity shows that the decision to aggregate to the group level changes little about the problems discussed here. It also indicates that, more generally, problems with HC1 heteroskedasticity-robust standard errors are closely related to problems with CV1 cluster-robust standard errors: in some settings, they produce nearly identical results if data are aggregated.

In general, more advanced inference methods improve inference even though, as noted in Section 4, standard errors from them are also inaccurate. HC3/jackknife estimators modestly improve inference on an average and greatly improve it for a regression or IV coefficient. Wild cluster bootstrap of the disaggregated data (using code from Roodman et al. (2019)) performs very well in almost all settings. Results for regression and IV coefficients are broadly similar to results for an average, justifying my use of the average in theoretical work. Interestingly, confidence intervals for the constant term in a simple regression display more severe issues; however, I do not include them in these simulations because they are rarely used in practice. Problems with accuracy and inference are substantially more severe when $s = 2$; for example, average and regression coefficients can be as much as 15 or 20 times less accurate, while IV coefficients can be substantially worse (likely because of weak instruments). Further, confidence intervals with nominal size of 5% can have actual size above 40%.

7 Replication

To examine how these issue might affect real-world research, in Table 1, I replicate key results from two highly-cited labor economics papers that examine the effect of increasing imports

from China on labor market outcomes (Autor et al., 2013) and political polarization (Autor et al., 2020) in the United States. Both papers use variation at the level of commuting zones and weight by some measure of the population in each zone. In urban areas, these zones roughly correspond to metropolitan areas, and they demarcate somewhat self-contained labor markets. The populations in these zones roughly follow a power law, so results from those papers are vulnerable to the critiques discussed here.

In the first two rows, we find the main results from Autor et al. (2013), who report that $1,000 of increased import exposure per worker causes manufacturing employment per working-age population to drop by 0.6 percentage points. Using QML, I find an estimate only half as large (0.3 percentage points), an economically significant difference. However—likely because the baseline statistic is so highly statistically significant (with a t-statistic of around -6), the QML estimate is also statistically significant at conventional levels (with a t-statistic of around -3), though it is less significant.

Changes in point estimates are similarly large for Autor et al. (2020), while changes in inference are more severe.[2] Their Table 3 examines effects on campaign contributions. Panel A of that paper reports that import exposure significantly increases donations; using QML, I find results that are only one-third as large, and not statistically significant. Panels B, C, and D report effects on contributions from left-wing, moderate, and right-wing donors, respectively. Autor et al. (2020) report large and significant effects for both non-moderate types of donors; the effects on these donors using QML are about half as large, and much less statistically significant. Table 4 of Autor et al. (2020) examines results of congressional

[2]Data from Autor et al. (2020)'s results using Pew and Nielsen data was not publicly released, so I do not replicate them here.

elections. Column 1 reports that import exposure significantly increases voter turnout in these elections; using QML, effects are only about one-third as large, and not statistically significant. Columns 2 through 5, on Republican vote share in various types of districts, show no significant effect in the original table. Column 6 shows that import exposure causes an increased probability of Republican victory; with QML, the effect is somewhat smaller and less significant. Finally, Table 5 of Autor et al. (2020) examines results of presidential elections, between 2000-2008 (Panel A) or 2000-2016 (Panel B). In both cases, with QML, the results are substantially smaller and less statistically significant.

Broadly, these results show that the type of extreme heteroskedasticity presented in this paper can have important effects on estimates and inference in real-world analysis. In this setting, standard errors are almost always smaller using QML than with the weighted estimators (which may themselves be incorrect anyway, as discussed above). When the absolute value of t-statistics from conventional WLS are very large (as with the value of around 6 for Autor et al. (2013)), the QML correction is unlikely to undo statistical significance. However, estimates may be substantially different, and even estimates with p-values below 1% may no longer be statistically significant with this correction.

[Table 1 here]

8 Concluding remarks

Power laws occur frequently in economic applications. The size of firms and cities approximately obey power laws, as do the size of stock market trades and short-term returns, as well as many other quantities. If accuracy of observations is related to size—or if the econometri-

cian estimates parameters assuming it is—then the variance of observations will also obey a power law. Traditional least squares tools can lead to inconsistent estimators and unreliable inference if this extreme heteroskedasticity is present. Problems are most severe when the wrong estimator is used—OLS under heteroskedasticity and WLS under homoskedasticity.

To fix this issue, I propose a QML estimator that can select between weighted and unweighted estimators, as well as intermediate estimators, and for multiple-equation models can even use weighting for some equations and no weighting for others. This estimator, which can be used as the "regoptwgt" command in Stata, is generally applicable to settings where a researcher is considering using weighted estimators to improve precision.

Many of the problems identified in this paper are most severe for high values of s—for example, in the size of stock trades ($s \approx 1.5$) or short-term stock price changes ($s \approx 3$). However, they may be a particularly complicated issue for studies that use observations at the level of cities or firms, including many staggered-rollout studies and those with Bartik instruments. In these cases, group size approximately follows a power law with $s = 1$, which is an edge case in many proofs. In this case, OLS estimates will not generally converge to the truth; instead, the variance of the estimator converges to a non-zero constant. On the other hand, the WLS estimator will converge to the parameter of interest, but slowly, because $s = 1$ is an edge case: if $s = 1 + \epsilon$ for any $\epsilon > 0$, then the WLS estimator does not converge, even for $\epsilon \to 0$. Additionally, WLS for $s = 1$ can cause particular problems with inference if the actual heteroskedasticity is small—for example, if we weight by city size, but estimates are just as accurate for small as large cities. In this case, asymptotic normality no longer holds in general and estimated standard errors are not consistent for the true standard deviation of the estimator—in fact, their variance relative to the truth is

particularly high with $s = 1$. Because of this, estimated confidence intervals may no longer have the correct size.

In practice, the impact of power laws on precision and inference may also cause important issues in an instrumental variables setting where the first stage is in danger of being weak. Incorrectly using weighted or unweighted estimators could lead to problems with weak instruments in cases where QML would not. (We avoid this issue in simulations by using only strong instruments when instrumental variables are used with $s = 1$, though the instrument is likely weak with $s = 2$, where weak instruments are harder to avoid.) Further, incorrect inference may mean that a weak instrument would be erroneously categorized as strong. That is particularly true because, as seen in Appendix Section D, the actual size of confidence intervals differs most from nominal when the nominal size is small. Testing for instrument strength requires tests with very small size: for example, an F-statistic of 10 (the usual rule-of-thumb) corresponds to a p-value of around 0.0016, and some recent literature suggests even larger F-statistics. (However, Andrews et al. (2019) caution that such pre-tests can be problematic for other reasons anyway.) In any case, formal proofs of the properties of regression and instrumental variables estimators under power law heteroskedasticity may be helpful future research.

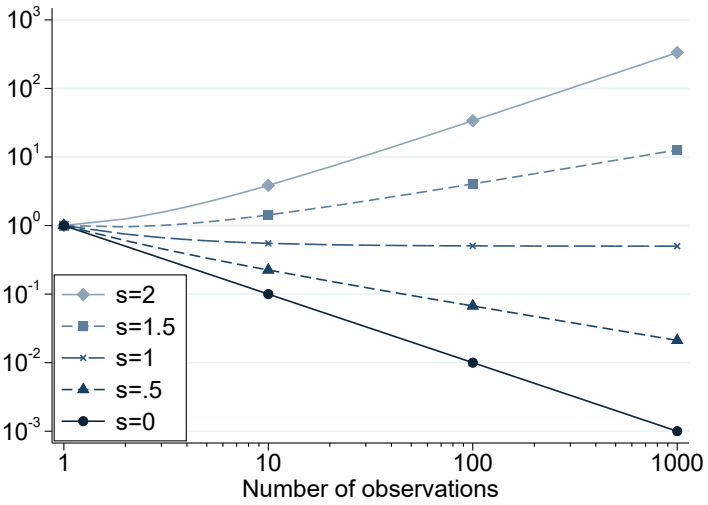
In this paper, I have assumed that the estimated model is homogeneous across observations; for example, that θ in Equation 4 is the same for all t. In many settings, this may be unrealistic, particularly when a more complicated model is estimated; causal effects are often found to be heterogeneous. As noted above, even weighting does not result in an estimator for the population average partial effect (Solon et al., 2015), so it is reasonable to instead choose QML, which has better inferential properties and is motivated by precision

under homogeneity, a useful working model. However, future work involves understanding the properties of the unweighted, weighted, and QML estimators in a setting with heterogeneous parameters and power law heteroskedasticity.

When encountering data that follows a power law, a few best practices will help researchers. If possible, using QML via regoptwgt or a similar methodology seems to solve most of the problems identified here. If that is not possible, trying both weighted and unweighted estimators may eliminate the worst-case scenario of using the wrong estimator—except in multi-equation models when some equations should be weighted and some should be unweighted. Researchers should also be careful not to simply choose the estimator with the smaller estimated standard error, because standard errors in this setting can be uninformative about the accuracy of the estimators. Using unweighted as a default may be preferable, as inference is never compromised and with $s = 1$ it has superior worst-case precision relative to QML (twice as bad rather than five times worse for weighted). Researchers would also be well-served by not simply relying on Stata's default heteroskedasticity-robust or cluster-robust standard errors, instead using more modern techniques like HC3/jackknife or bootstrap estimators.

Power laws are common in economics, and they can lead real-world data to be highly heteroskedastic. When it arises, researchers must take this heteroskedasticity into account if they hope to report precise estimates and accurate inference.

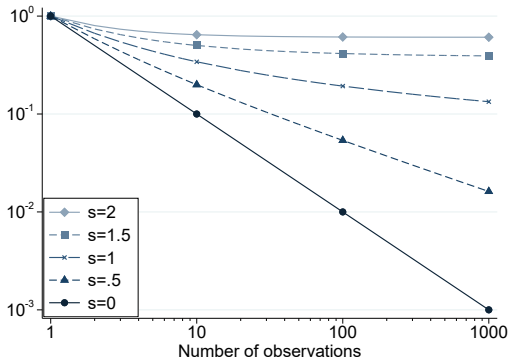
Figure 1: Variance of the OLS estimator



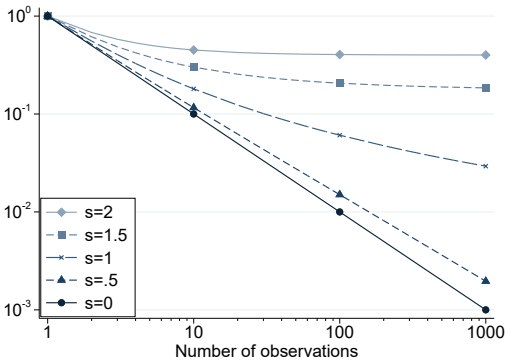
Notes: Variance is shown as sample size increases, with various values of s, setting $\sigma_\eta^2 = 1$ and $\sigma_\nu^2 = 0$. Additional data are assumed to be added from most accurate to least accurate.

Figure 2: Variance of the WLS estimator

(a) Setting $\sigma_\eta^2 = 1$ and $\sigma_\nu^2 = 0$



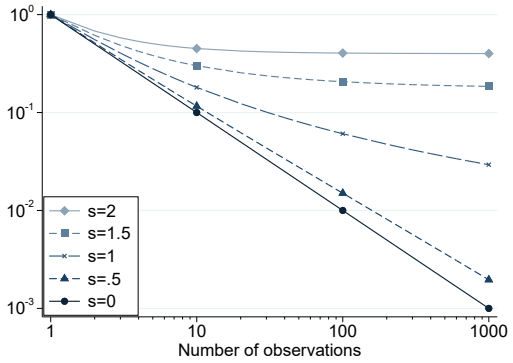
(b) Setting $\sigma_\eta^2 = 0$ and $\sigma_\nu^2 = 1$



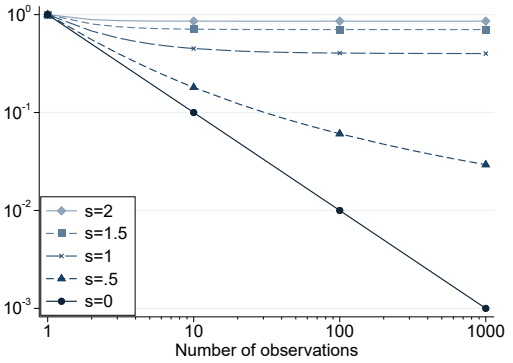
Notes: Variance is shown as sample size increases, with various values of s. Additional data are assumed to be added from most accurate to least accurate.

Figure 3: Excess kurtosis of the WLS estimator

(a) Setting $\sigma_\eta^2 = 1$ and $\sigma_\nu^2 = 0$



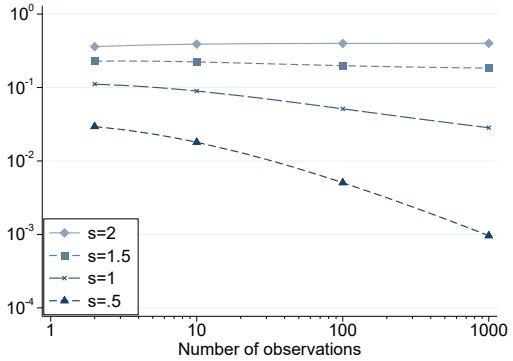
(b) Setting $\sigma_\eta^2 = 0$ and $\sigma_\nu^2 = 1$



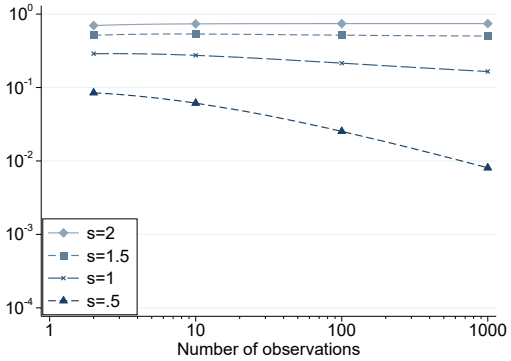
Notes: Excess kurtosis is shown as sample size increases, with various values of s. Additional data are assumed to be added from most accurate to least accurate. When the central limit theorem applies, excess kurtosis converges to 0, the value for a normal distribution. In both graphs, excess kurtosis of each η or ν is assumed to be 1.

Figure 4: Expected difference between estimated and true variance of the WLS estimator

(a) Setting $\sigma_\eta^2 = 1$ and $\sigma_\nu^2 = 0$.



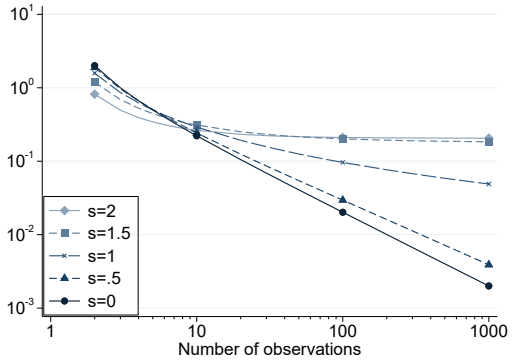
(b) Setting $\sigma_\eta^2 = 0$ and $\sigma_\nu^2 = 1$.



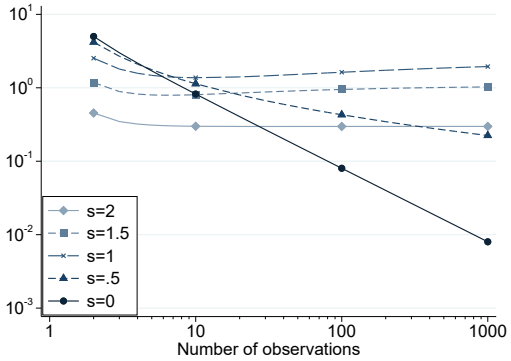
Notes: $-1 \times \left(\mathbb{E}\left[\frac{\hat{V}}{\mathbb{V}[\hat{\theta}_{pw}]} \right] - 1 \right)$. Expectation is shown as sample size increases, with various values of s. Additional data are assumed to be added from most accurate to least accurate. Note that high values in this graph mean the estimated variance is smaller than the true variance. $s = 0$ is not shown because for that value, $\mathbb{E}\left[\frac{\hat{V}}{\mathbb{V}[\hat{\theta}_{pw}]} \right] = 1$.

Figure 5: Variance of ratio between estimated and true variance of the WLS estimator

(a) Setting $\sigma_\eta^2 = 1$ and $\sigma_\nu^2 = 0$.

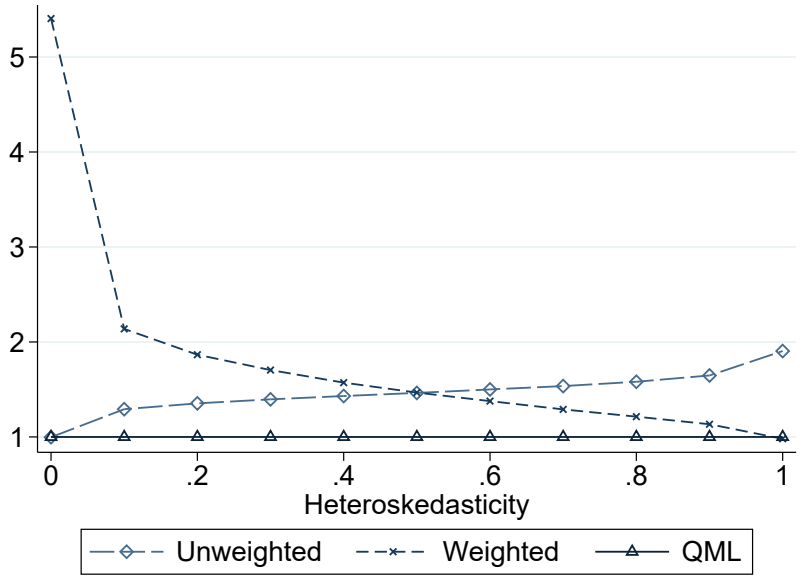


(b) Setting $\sigma_\eta^2 = 0$ and $\sigma_\nu^2 = 1$.



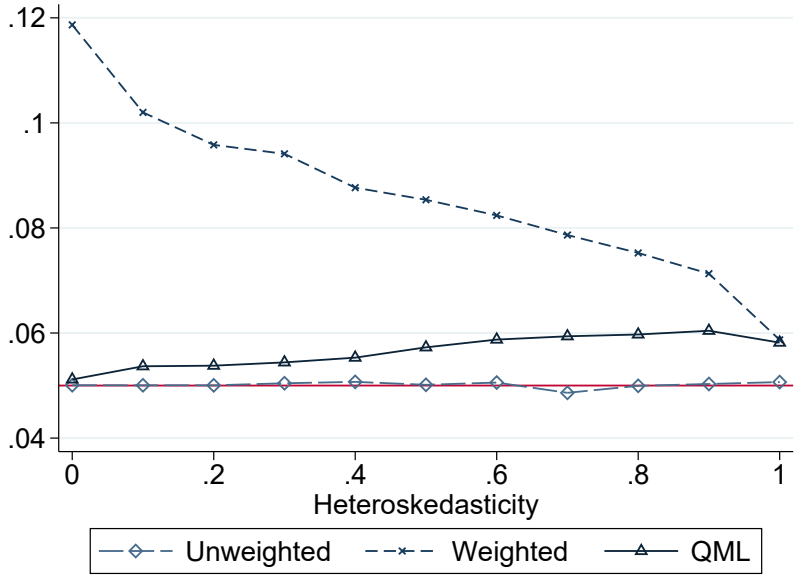
Notes: $\mathbb{V}\left[\frac{\hat{V}}{\mathbb{V}[\hat{\theta}_{pw}]}\right]$. Variance is shown as sample size increases, with various values of s. Additional data are assumed to be added from most accurate to least accurate. In Figure 5a, excess kurtosis is taken to be 0, as if η are normally distributed. In Figure 5b, excess kurtosis is taken to be 6, as if ν are exponentially distributed.

Figure 6: Relative root mean squared error in simulations estimating an average



Notes: Each graph shows the root mean squared error of estimators relative to QML for $s = 1$. Each point includes 100,000 simulations of an estimate. The error term is $\epsilon_t = k^{\frac{1}{2}} t^{\frac{s}{2}} \eta_t + \nu_t$, with $k = \frac{\frac{H_{2s}(T)}{H_s(T)^2} - \frac{1}{T}}{\frac{H_{-s}(T)}{T^2} - H_s(T)^{-1}} e^{\Phi^{-1}(h)}$, where h is heteroskedasticity and $\Phi(.)$ is the CDF of a normal distribution. For each simulation, there are 1,000 observations where $\eta_t \sim \mathbb{N}(0,1)$ and $\nu_t \sim \text{Exp}(1) - 1$.

Figure 7: Size of nominal 95% confidence intervals in simulations estimating an average



Notes: Each graph shows the size of nominally 95% confidence intervals for $s = 1$, using different estimation techniques. Each point includes 100,000 simulations of an estimate of the average. The error term is $\epsilon_t = k^{\frac{1}{2}} t^{\frac{s}{2}} \eta_t + \nu_t$, with $k = \frac{\frac{H_{2s}(T)}{H_s(T)^2} - \frac{1}{T}}{\frac{H_{-s}(T)}{T^2} - H_s(T)^{-1}} e^{\Phi^{-1}(h)}$, where h is heteroskedasticity and $\Phi(.)$ is the CDF of a normal distribution. For each simulation, there are 1,000 observations where $\eta_t \sim \mathbb{N}(0,1)$ and $\nu_t \sim \text{Exp}(1) - 1$.

Table 1: Replication of Autor et al. (2013, 2020)

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Paper	**Table**	**Color Panel**	**Use QML**	**Est.**	**S.E.**	**t-stat**	**p Value**
Labor	3	6	N	-0.60	0.10	-6.04	0.000
			Y	-0.30	0.10	-2.98	0.003
Polarization	3	A	N	37.23	21.05	1.77	0.077
			Y	12.70	15.91	0.80	0.425
Polarization	3	B	N	71.00	31.06	2.29	0.022
			Y	41.92	22.90	1.83	0.067
Polarization	3	C	N	23.60	19.63	1.20	0.229
			Y	4.42	14.03	0.32	0.753
Polarization	3	D	N	46.05	27.15	1.70	0.090
			Y	20.92	19.57	1.07	0.285
Polarization	4	1	N	5.27	1.94	2.72	0.007
			Y	1.87	1.73	1.08	0.280
Polarization	4	2	N	-1.08	5.98	-0.18	0.856
			Y	-3.75	5.45	-0.69	0.491
Polarization	4	3	N	-0.95	1.80	-0.53	0.599
			Y	-1.11	1.09	-1.02	0.307
Polarization	4	4	N	6.10	4.93	1.24	0.216
			Y	2.86	3.98	0.72	0.472
Polarization	4	5	N	-6.24	3.93	-1.59	0.112
			Y	-5.53	4.38	-1.26	0.207
Polarization	4	6	N	24.08	12.07	1.99	0.046
			Y	17.40	9.81	1.77	0.076
Polarization	5	A	N	1.59	0.85	1.86	0.062
			Y	1.02	0.66	1.55	0.121
Polarization	5	B	N	1.71	0.90	1.89	0.059
			Y	1.05	0.64	1.63	0.102

Notes: This table replicates results from Autor et al. (2013, 2020), using weighted instrumental variables (as in those papers) or QML (as suggested in this paper). Column 1 indicates whether results come from Autor et al. (2013) ("Labor") or Autor et al. (2020) ("Polarization"). Column 2 indicates the table of the result. Column 3 indicates the column or panel of the result. Column 4 indicates whether the result uses QML; "N" indicates the result is the same as the original, published version. Column 5 presents the point estimate. Column 6 presents the standard error. Column 7 presents the t-statistic (the estimate divided by the standard error). Column 8 presents the p-value of a test for the statistic being equal to zero.

References

Andrews, Isaiah, James H. Stock, and Liyang Sun. Weak instruments in instrumental variables regression: Theory and practice. *Annual Review of Economics*, 11(1):727–753, August 2019.

Arellano, Manuel. Computing robust standard errors for within-groups estimators. *Oxford Bulletin of Economics and Statistics*, 49(4):431–434, November 1987.

Autor, David, David Dorn, Gordon Hanson, and Kaveh Majlesi. Importing political polarization? the electoral consequences of rising trade exposure. *American Economic Review*, 110(10):3139–3183, October 2020.

Autor, David H., David Dorn, and Gordon H. Hanson. The China syndrome: Local labor market effects of import competition in the United States. *American Economic Review*, 103(6):2121–2168, October 2013.

Axtell, Robert L. Zipf distribution of U.S. firm sizes. *Science*, 293:1818–1820, September 2001.

Bell, Robert M. and Daniel F. McCaffrey. Bias reduction in standard errors for linear regression with multi-stage samples. *Survey Methodology*, 28(2):169–181, December 2002.

Cameron, A. Colin and Douglas L. Miller. Large sample estimation and hypothesis testing. In Ullah, Aman and David E. A. Giles, editors, *Handbook of Empirical Economics and Finance*, Handbook of Empirical Economics and Finance, chapter 1, pages 1–28. Chapman and Hall, 2011.

Cameron, A. Colin and Douglas L. Miller. A practitioner's guide to cluster-robust inference. *Journal of Human Resources*, 50(2):317–372, 2015.

Cameron, A. Colin, Jonah B. Gelbach, and Douglas L. Miller. Bootstrap-based improvements for inference with clustered errors. *The Review of Economics and Statistics*, 90(3):414–427, August 2008.

Carter, Andrew V., Kevin T. Schnepel, and Douglas G. Steigerwald. Asymptotic behavior of a t-test robust to cluster heterogeneity. *The Review of Economics and Statistics*, 99(4): 698–709, July 2017.

Chiang, Harold D., Yuya Sasaki, and Yulong Wang. Genuinely robust inference for clustered data. Papers 2308.10138, arXiv.org, Aug 2023.

Chung, Kai Lai. *A Course in Probability Theory*. Academic Press, second edition, 1974.

Davidson, Russell and James G. MacKinnon. *Estimation and Inference in Econometrics*. Oxford University Press, 1993.

Dickins, William T. Error components in grouped data: Is it ever worth weighting? *The Review of Economics and Statistics*, 72(2):328–333, May 1990.

Eeckhout, Jan. Gibrat's law for (all) cities. *American Economic Review*, 94(5):1429–1451, 2004.

Eicker, Friedhelm. Asymptotic normality and consistency of the least squares estimators for families of linear regressions. *The Annals of Mathematical Statistics*, 34(2):447–456, June 1963.

Gabaix, Xavier. Power laws in economics and finance. *Annual Review of Economics*, 1(1): 255–294, 2009.

Gabaix, Xavier. Power laws in economics: An introduction. *Journal of Economic Perspectives*, 30(1):185–206, 2016.

Gabaix, Xavier, Parameswaran Gopikrishnan, Vasiliki Plerou, and H. Eugene Stanley. Institutional investors and stock market volatility. *The Quarterly Journal of Economics*, 121 (2):461–504, None 2006.

Gopikrishnan, Parameswaran, Vasiliki Plerou, Luís A. Nunes Amaral, Martin Meyer, and H. Eugene Stanley. Scaling of the distribution of fluctuations of financial market indices. *Phys. Rev. E*, 60:5305–5316, Nov 1999.

Greene, William H. *Econometric Analysis*. Pearson, 8th edition, 2018.

Gu, Ariel and Hong Il Yoo. vcemway: A one-stop solution for robust inference with multiway clustering. *Stata Journal*, 19(4):900–912, December 2019.

Hayashi, Fumio. *Econometrics*. Princeton University Press, 2002.

Huber, Peter J. The behavior of maximum likelihood estimates under nonstandard conditions. In *Conference Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, pages 221–233. University of California Press, Berkeley, 1967.

Liang, Kung-Yee and Scott L. Zeger. Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1):13–22, 1986.

MacKinnon, James G. Thirty years of heteroskedasticity-robust inference. In Chen, Xiaohong and Norman R. Swanson, editors, *Recent Advances and Future Directions in Causality, Prediction, and Specification Analysis*, pages 437–461. Springer, New York, 2012.

MacKinnon, James G. How cluster-robust inference is changing applied econometrics. *Canadian Journal of Economics*, 52(3):851–881, August 2019.

MacKinnon, James G. and Halbert White. Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties. *Journal of Econometrics*, 29 (3):305–325, September 1985.

MacKinnon, James G., Morten Ørregaard Nielsen, and Matthew D. Webb. Cluster-robust inference: A guide to empirical practice. *Journal of Econometrics*, 232(2):272–299, None 2023.

Moulton, Brent R. Random group effects and the precision of regression estimates. *Journal of Econometrics*, 32(3):385–397, August 1986.

Roodman, David, James G. MacKinnon, Morten Ørregaard Nielsen, and Matthew D. Webb. Fast and wild: Bootstrap inference in stata using boottest. *Stata Journal*, 19(1):4–60, March 2019.

Rozenfeld, Hernan D., Diego Rybski, Xavier Gabaix, and Hernan A. Makse. The area and population of cities: New insights from a different perspective on cities. *American Economic Review*, 101(5):2205–2225, 2011.

Solon, Gary, Steven J. Haider, and Jeffrey Wooldridge. What are we weighting for? *Journal of Human Resources*, 50(2):301–316, 2015.

White, Halbert. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, 48(4):817–838, May 1980.

Young, Alwyn. Channeling fisher: Randomization tests and the statistical insignificance of seemingly significant experimental results. *The Quarterly Journal of Economics*, 134(2): 557–598, None 2019.

Young, Alwyn. Consistency without inference: Instrumental variables in practical application. *European Economic Review*, 147(C), 2022.

ONLINE APPENDIX

for "Power Law Heteroskedasticity"

by David J. Price

December 24, 2025

A Mathematical appendix

This section includes proofs for the theorems, propositions, and lemmas in the text.

Theorem 3.1.

Suppose $\hat{\theta}_{uw}$ is defined as in Equation 6, where η_t and ν_t are mean-zero, independently distributed, and homoskedastic with finite variances σ_η^2 and σ_ν^2, respectively. If $s < 1$, then $\hat{\theta}_{uw} \xrightarrow{p} \theta$. If additionally $\{\eta_t^2\}$ and $\{\nu_t^2\}$ are each uniformly integrable; $\sigma_\eta^2 > 0$; and $s \geq 1$; then $\hat{\theta}_{uw} \not\xrightarrow{p} \theta$.

Proof. (*Convergence if $s < 1$*) Using Chung (1974), Theorem 5.4.1, Corollary i (p. 125), a sufficient condition for $\hat{\theta}_{uw} \xrightarrow{p} \theta$ is that

$$\lim_{T\to\infty} \sum_{t=1}^{T} \frac{1}{t^2} \mathbb{E}\left[\left(t^{\frac{s}{2}}\eta_t + \nu_t\right)^2\right] < \infty. \tag{23}$$

(In fact, this is sufficient for almost sure convergence.) Because η_t and ν_t are independent and have mean zero, we have

$$\lim_{T\to\infty} \sum_{t=1}^{T} \frac{1}{t^2} \mathbb{E}\left[\left(t^{\frac{s}{2}}\eta_t + \nu_t\right)^2\right] = \lim_{T\to\infty} \left\{\sigma_\eta^2 \sum_{t=1}^{T} t^{s-2} + \sigma_\nu^2 \sum_{t=1}^{T} t^{-2}\right\}. \tag{24}$$

The sum on the left converges if $s < 1$, and the sum on the right always converges. To see this, note that by the definition of an integral as the sum of the area under a curve, for $p < -1$,

$$\sum_{t=1}^{T} t^p \leq 1 + \int_1^{T-1} t^p dt = \frac{1}{p+1}(T-1)^{p+1} + \frac{p}{p+1}. \tag{25}$$

□

Proof. (*Non-convergence if* $s \geq 1$) Using the methodology from the proof above, it is clear that the average of the ν_t terms converges. Thus a sufficient condition for non-convergence of $\hat{\theta}_{uw}$ is non-convergence of the average of the $t^{\frac{s}{2}}\eta_t$ terms.

Setting $\sigma_\nu^2 = 0$ for convenience, we have $\mathbb{V}\left[\hat{\theta}_{uw}\right] = \sigma_\eta^2 \frac{1}{T^2}\sum_{t=1}^{T} t^s$. Thus for $s = 1$, we have $\mathbb{V}\left[\hat{\theta}_{uw}\right] \to C$ for some constant C; and for $s > 1$, we have $\mathbb{V}\left[\hat{\theta}_{uw}\right] \to \infty$. In either case, if $\hat{\theta}_{uw} \xrightarrow{p} \theta$, then by Slutsky we would have $\left(\mathbb{V}\left[\hat{\theta}_{uw}\right]\right)^{-\frac{1}{2}}\left(\hat{\theta}_{uw} - \theta\right) \xrightarrow{p} 0$. But by Theorem 3.3, below (which does not use this result), $\left(\mathbb{V}\left[\hat{\theta}_{uw}\right]\right)^{-\frac{1}{2}}\left(\hat{\theta}_{uw} - \theta\right) \xrightarrow{d} \mathbb{N}(0,1)$, a contradiction. □

Lemma 3.2.

Suppose $\{\epsilon_t\}$ *are mean-zero, independently distributed, homoskedastic random variables with finite variance* σ_ϵ^2. *Further, suppose* $\{\epsilon_t^2\}$ *are uniformly integrable, and that there is some function* $g(T)$ *such that, for all* T, $g(T) \neq 0$; *and a function* $f(t)$ *such that*

$$\lim_{T\to\infty} \sup_{t\leq T} \frac{f(t)^2}{\sum_{s=1}^{T} f(s)^2} = 0. \tag{26}$$

Define $X_{Tt} \equiv g(T)f(t)\epsilon_t$; $S_T \equiv \sum_{t=1}^{T} X_{Tt}$; *and* $s_T^2 \equiv \sum_{t=1}^{T} \mathbb{V}\left[X_{Tt}\right]$. *Then* $\frac{S_T}{s_T} \xrightarrow{d} \mathbb{N}(0,1)$.

Proof. It is sufficient to prove that the Lindeberg condition applies with the assumptions above. That is, for all $c > 0$, we must prove that

$$\mathbb{L} \equiv \lim_{T\to\infty} \sum_{t=1}^{T} \int \frac{X_{Tt}^2}{s_T^2} \mathbb{I}\left[\frac{X_{Tt}^2}{s_T^2} \geq c\right] dP = 0. \tag{27}$$

Note that $s_T^2 = \sigma_\epsilon^2 \sum_{t=1}^{T} g(T)^2 f(t)^2$. All $g(T)$ functions thus cancel out from Equation 27.

Define $\alpha(t,T) \equiv \left(\frac{f(t)^2}{c\sigma_\epsilon^2 \sum_{s=1}^T f(s)^2}\right)^{-1}$, and $\alpha(T) = \inf_{t \leq T} \alpha(t,T)$. Then

$$\begin{aligned}
\mathbb{L} &= \frac{1}{\sigma_\epsilon^2} \lim_{T\to\infty} \frac{1}{\sum_{t=1}^T f(t)^2} \sum_{t=1}^T f(t)^2 \int \epsilon_t^2 \mathbb{I}\left[\epsilon_t^2 \geq \alpha(t,T)\right] dP \\
&\leq \frac{1}{\sigma_\epsilon^2} \lim_{T\to\infty} \frac{1}{\sum_{t=1}^T f(t)^2} \sum_{t=1}^T f(t)^2 \int \epsilon_t^2 \mathbb{I}\left[\epsilon_t^2 \geq \alpha(T)\right] dP \\
&\leq \frac{1}{\sigma_\epsilon^2} \lim_{T\to\infty} \sup_t \int \epsilon_t^2 \mathbb{I}\left[\epsilon_t^2 \geq \alpha(T)\right] dP \\
&= \frac{1}{\sigma_\epsilon^2} \lim_{\alpha\to\infty} \sup_t \int \epsilon_t^2 \mathbb{I}\left[\epsilon_t^2 \geq \alpha\right] dP = 0. \qquad (28)
\end{aligned}$$

The first inequality is by definition of $\alpha(T)$; the second inequality is from noting that the term is a convex combination of the positive-valued integrals; the final line's first equality is from noting that $\lim_{T\to\infty} \alpha(T) = \infty$ by the assumption in Equation 26 and using the fact that $f(t)^2 \geq 0$; and the final equality is from the assumption of uniform integrability. (Note that the first equality of the final line is only an equality on the condition that the second limit converges; but it so does by assumption.) □

Theorem 3.3

Suppose $\hat{\theta}_{uw}$ is defined as in Equation 6, where η_t and ν_t are mean-zero, independently distributed, and homoskedastic with finite variances σ_η^2 and σ_ν^2, respectively; and that $\{\eta_t^2\}$ and $\{\nu_t^2\}$ are uniformly integrable. Then $g(T)\left(\hat{\theta}_{uw} - \theta\right) \xrightarrow{d} \mathbb{N}(0,1)$ for some function $g(T)$.

Proof. First, note that $\hat{\theta}_{uw} - \theta = \sum_{t=1}^T \frac{1}{T} t^{\frac{s}{2}} \eta_t + \sum_{t=1}^T \frac{1}{T} \nu_t$. I will prove that each term converges to a normal distribution; the sum of independent normals is normal, so the sum also converges to a normal distribution. Using Lemma 3.2, I only need to prove that, for

$s \geq 0$,

$$\lim_{T\to\infty} \sup_{t\leq T} \frac{t^s}{\sum_{s=1}^{T} t^s} = 0. \tag{29}$$

Note that for $s \geq 0$, we have $\sum_{t=1}^{T} t^s \geq \int_0^T t^s ds = \frac{1}{s+1} T^{s+1}$, so the limit in Equation 29 is $\lim_{T\to\infty}(s+1)T^{-1} = 0$. □

Theorem 3.4

Define $\hat{\epsilon}_t \equiv y_t - \hat{\theta}_{uw}$, *and* $\hat{V} \equiv \frac{T}{T-1}\frac{1}{T^2}\sum_{t=1}^{T}\hat{\epsilon}_t^2$. *As above, suppose that* $\hat{\theta}_{uw}$ *is defined as in Equation 6, where* η_t *and* ν_t *are mean-zero; independently distributed; homoskedastic with finite variances* σ_η^2 *and* σ_ν^2, *respectively; and have uniformly bounded kurtosis. Then* $\frac{\hat{V}}{\mathbb{V}\left[\hat{\theta}_{uw}\right]} \xrightarrow{p} 1$.

Proof. If $\sigma_\eta^2 = 0$, then this is OLS in a homoskedastic setting, so clearly standard errors are consistent. If $\sigma_\eta^2 > 0$, then as $t \to \infty$, the η_t term dominates both the variance of the estimator, and estimated variance; for simplicity, I will therefore only consider this term.

$\frac{T}{T-1} \to 1$, so using Slutsky I will ignore this.

For estimated variance, we have

$$\hat{\epsilon}_t = t^{\frac{s}{2}}\eta_t - \frac{1}{T}\sum_{i=0}^{T} i^{\frac{s}{2}}\eta_i \tag{30}$$

$$\hat{\epsilon}_t^2 = t^s\eta_t^2 - 2t^{\frac{s}{2}}\eta_t\frac{1}{T}\sum_{i=1}^{T} i^{\frac{s}{2}}\eta_i + \left(\frac{1}{T}\sum_{i=1}^{T} i^{\frac{s}{2}}\eta_i\right)^2 \tag{31}$$

$$\hat{V} = \frac{1}{T^2}\sum_{t=1}^{T} t^s\eta_t^2 - \frac{1}{T}\left(\frac{1}{T}\sum_{t=1}^{T} t^{\frac{s}{2}}\eta_t\right)^2 \tag{32}$$

From Theorem 3.3, we know that

$$\left(\sigma_\eta^2\frac{1}{T^2}\sum_{t=0}^{T} t^s\right)^{-\frac{1}{2}}\left(\frac{1}{T}\sum_{t=1}^{T} t^{\frac{s}{2}}\eta_t\right) \xrightarrow{d} \mathbb{N}(0,1). \tag{33}$$

Note that the true variance is given by $\mathbb{V}\left[\hat{\theta}_{uw}\right] = \sigma_\eta^2 \frac{1}{T^2}\sum_{t=1}^{T} t^s$. Denote the first term of $\hat{V}$ in Equation 32 as $\hat{V}_1$ and the second term as $\hat{V}_2$, so $\hat{V} = \hat{V}_1 - \hat{V}_2$. Then squaring the formula in Equation 33,

$$\left(\sigma_\eta^2 \frac{1}{T^2}\sum_{t=0}^{T} t^s\right)^{-1}\left(\frac{1}{T}\sum_{t=1}^{T} t^{\frac{s}{2}}\eta_t\right)^2 = \left(\mathbb{V}\left[\hat{\theta}_{uw}\right]\right)^{-1} \times T \times \hat{V}_2 \xrightarrow{d} \chi^2, \tag{34}$$

and so $\left(\mathbb{V}\left[\hat{\theta}_{uw}\right]\right)^{-1}\hat{V}_2 \xrightarrow{p} 0$ by Slutsky. I therefore only need to show that

$$\left(\mathbb{V}\left[\hat{\theta}_{uw}\right]\right)^{-1}\hat{V}_1 = \left(\sigma_\eta^2 \frac{1}{T^2}\sum_{t=1}^{T} t^s\right)^{-1}\left(\frac{1}{T^2}\sum_{t=1}^{T} t^s\eta_t^2\right) \xrightarrow{p} 1, \tag{35}$$

which I will do by showing that it has expectation 1 and variance converging to 0. First,

$$\mathbb{E}\left[\left(\sigma_\eta^2 \frac{1}{T^2}\sum_{t=1}^{T} t^s\right)^{-1}\left(\frac{1}{T^2}\sum_{t=1}^{T} t^s\eta_t^2\right)\right] = \left(\sigma_\eta^2 \frac{1}{T^2}\sum_{t=1}^{T} t^s\right)^{-1}\left(\frac{1}{T^2}\sum_{t=1}^{T} t^s\sigma_\eta^2\right) = 1. \tag{36}$$

Second, defining $\kappa_\eta^4 \equiv \sup_t \mathbb{V}[\eta_t^2]$, which exists by assumption, we also have that

$$\mathbb{V}\left[\frac{1}{T^2}\sum_{t=1}^{T} t^s\eta_t^2\right] \leq \kappa_\eta^4 T^{-4}\sum_{t=1}^{T} t^{2s}, \tag{37}$$

and thus

$$\begin{aligned}
&\mathbb{V}\left[\left(\sigma_\eta^2 \frac{1}{T^2}\sum_{t=1}^{T} t^s\right)^{-1}\left(\frac{1}{T^2}\sum_{t=1}^{T} t^s\eta_t^2\right)\right] \leq \frac{\kappa_\eta^4}{\sigma_\eta^4}\left(\sum_{t=1}^{T} t^s\right)^{-2}\left(\sum_{t=1}^{T} t^{2s}\right) \\
&\leq \frac{\kappa_\eta^4}{\sigma_\eta^4}\left(\int_0^T t^s ds\right)^{-2}\left(\int_0^{T+1} t^{2s} ds\right) = \frac{\kappa_\eta^4}{\sigma_\eta^4}\frac{(s+1)^2}{2s+1}(T^{s+1})^{-2}(T+1)^{2s+1} \\
&= \frac{\kappa_\eta^4}{\sigma_\eta^4}\frac{(s+1)^2}{2s+1}\left(\frac{T+1}{T}\right)^{2s+1} T^{-1} \to 0,
\end{aligned} \tag{38}$$

where the inequality between summation and integration is because t^s is increasing in t. □

Theorem 4.1

Suppose $\hat{\theta}_{pw}$ is defined as in Equation 9, where η_t and ν_t are mean-zero, independently distributed, and homoskedastic with finite variances σ_η^2 and σ_ν^2, respectively, at least one of which is non-zero. Then $\hat{\theta}_{pw} \xrightarrow{p} \theta$ if and only if $s \leq 1$.

Proof. (Convergence if $s \leq 1$) Clearly, $\mathbb{E}\left[\hat{\theta}_{pw}\right] = \theta$. I will prove that $\mathbb{V}\left[\hat{\theta}_{pw}\right] \to 0$, which is sufficient. We then have

$$\begin{aligned}
\mathbb{V}\left[\hat{\theta}_{pw}\right] &= \sigma_\eta^2 \left(\sum_{t=1}^{T} t^{-s}\right)^{-1} + \sigma_\nu^2 \left(\sum_{t=1}^{T} t^{-s}\right)^{-2} \sum_{t=1}^{T} t^{-2s} \\
&\leq \sigma_\eta^2 \left(\int_{t=1}^{T} t^{-s} dt\right)^{-1} + \sigma_\nu^2 \left(\int_{t=1}^{T} t^{-s} dt\right)^{-2} \left(1 + \int_{t=1}^{T} t^{-2s} dt\right) \\
&= \sigma_\eta^2 (1-s) \left(T^{1-s} - 1\right)^{-1} \\
&\quad + \sigma_\nu^2 (1-s)^2 \left(T^{1-s} - 1\right)^{-2} \left(1 + \frac{1}{1-2s}(T^{1-2s} - 1)\right) \\
&\to 0, \qquad (39)
\end{aligned}$$

with a similar expression if $s = 1$ (in which case the integral for t^{-s} becomes a logarithm); or if $s = \frac{1}{2}$ (in which case the integral for t^{-2s} becomes a logarithm). □

Proof. (Non-convergence if $s > 1$) Note that $\sum_{t=1}^{T} t^{-s} \to C$ for some constant C. Now, define $X_{lim} \equiv C^{-1} \lim_{T\to\infty} \sum_{t=2}^{T} \left(t^{-\frac{s}{2}} \eta_t + t^{-s} \nu_t\right)$. (In fact, X_{lim} will be a non-degenerate random variable with finite variance, but I need not prove that here.) We now have that $\hat{\theta}_{pw} \xrightarrow{p} \theta + C^{-1}(\eta_1 + \nu_1) + X_{lim}$. Note that $C^{-1}(\eta_1 + \nu_1)$ is independent of X_{lim}. A non-degenerate random variable, plus an independent random variable (degenerate or not) cannot equal a constant. Thus $\hat{\theta}_{pw}$ does not converge to a constant. □

Lemma 4.2

Suppose $\{\epsilon_t\}$ *are mean-zero, independently distributed, homoskedastic random variables with finite variance* $\sigma_\epsilon^2 > 0$. *Further, suppose there is some function* $g(T)$ *such that for all* T, $g(T) \neq 0$; *and a finite-valued function* $f(t) > 0$ *such that*

$$\lim_{T\to\infty} \frac{f(1)^2}{\sum_{t=1}^{T} f(t)^2} = C^2 \tag{40}$$

for some finite constant $C > 0$. *Define* $X_{Tt} \equiv g(T)f(t)\epsilon_t$; $S_T \equiv \sum_{t=1}^{T} X_{Tt}$; *and* $s_T^2 \equiv \sum_{t=1}^{T} \mathbb{V}[X_{Tt}]$. *Then* $\frac{S_T}{s_T}$ *is not generally asymptotically normal, in the sense that, for all* t, *holding fixed the distribution of* $\{\epsilon_k\}$ *for* $k \neq t$, *there is at most one distribution of* ϵ_t *for which* $\frac{S_T}{s_T} \xrightarrow{d} \mathbb{N}(0,1)$.

Proof. First, note that all $g(T)$ expressions cancel out from $\frac{S_T}{s_T}$, so they will be ignored. Next, note that $s_\infty \equiv lim_{T\to\infty} s_T$ is finite, or else Equation 40 would not hold. Because of this, $\left(\frac{S_T}{s_T} - \frac{S_T}{s_\infty}\right) \xrightarrow{p} 0$ by Slutsky; so I will prove the above theorem about $\frac{S_T}{s_\infty}$. We now have that

$$\begin{aligned}\frac{S_T}{s_\infty} &= (s_\infty)^{-1} f(t)\epsilon_t + (s_\infty)^{-1} \sum_{k\neq t}^{T} f(k)\epsilon_k \\ &= C\frac{f(t)}{f(1)}\epsilon_t + (s_\infty)^{-1} \sum_{k\neq t}^{T} f(k)\epsilon_k. \end{aligned} \tag{41}$$

Define ϕ_A as the characteristic function of $(s_\infty)^{-1}\sum_{k\neq t}^{T} f(k)\epsilon_k$; ϕ_t as the characteristic function of ϵ_t; and ϕ_N as the characteristic function of a standard normal random variable. If $\frac{S_T}{s_\infty} \xrightarrow{d} \mathbb{N}(0,1)$, then ϕ_t is uniquely defined by

$$\phi_t(x) = \frac{\phi_N\left(\frac{f(1)}{Cf(t)}x\right)}{\phi_A\left(\frac{f(1)}{Cf(t)}x\right)}. \tag{42}$$

This may not be a valid characteristic function; if it is not, then there is no distribution of ϵ_t that would lead to asymptotic normality. □

Theorem 4.3

Suppose $\hat{\theta}_{pw}$ is defined as in Equation 9, where $\{\eta_t\}$ and $\{\nu_t\}$ are mean-zero; independently distributed; and homoskedastic with finite variances σ_η^2 and σ_ν^2, respectively, with at least one strictly greater than zero; and $\{\eta_t^2\}$ and $\{\nu_t^2\}$ are uniformly integrable. If $s \leq \frac{1}{2}$, then $\hat{\theta}_{pw}$ is asymptotically normal (with suitable normalization). If $\frac{1}{2} < s \leq 1$, then $\hat{\theta}_{pw}$ is generally (in the sense of Lemma 4.2) asymptotically normal if and only if $\sigma_\eta^2 > 0$. If $s > 1$, then $\hat{\theta}_{pw}$ is not generally asymptotically normal in the sense of Lemma 4.2.

Proof. Asymptotic normality of $\hat{\theta}_{pw}$ will depend on the normality of the sum $\sum_{t=1}^{T} \left(t^{-\frac{s}{2}}\eta_t + t^{-s}\nu_t\right)$, where the normalization term can be dropped because it is nonrandom, and multiplication by a nonrandom variable does not impact normality. Combining Lemma 3.2 and Lemma 4.2, and because $sup_{t\geq 1} t^{-s} = 1$, a necessary and sufficient condition for (general) normality of the η_t term is that $\sum_{t=1}^{T} t^{-s} \to \infty$, i.e. that $s \leq 1$; and for (general) normality of the ν_t term is that $\sum_{t=1}^{T} t^{-2s} \to \infty$, i.e. that $s \leq \frac{1}{2}$. Furthermore, if $\frac{1}{2} < s \leq 1$ and $\sigma_\eta^2 > 0$, then the variance of the η_t term goes to ∞, while the variance of the ν_t term goes to a constant; thus, after normalization, the ν_t term disappears, and we are only left with the η_t term, which is asymptotically normal. □

Proposition 4.4

Define $\kappa_\eta^4 \equiv \frac{E(\eta_t^4)}{\sigma_\eta^4}$. If $\sigma_\nu^2 = 0$, $\mathbb{E}\left[\frac{\left(\hat{\theta}_{pw}-\theta\right)^4}{\mathbb{V}\left[\hat{\theta}_{pw}\right]^2}\right] - 3 = \left(\kappa_\eta^4 - 3\right)\frac{H_{2s}}{H_s^2}$.

Proof. Excess kurtosis is calculated with the binomial theorem. We have that

$$\mathbb{E}\left[\frac{\left(\hat{\theta}_{pw}-\theta\right)^4}{\mathbb{V}\left[\hat{\theta}_{pw}\right]^2}\right] = \frac{1}{\left(\sigma_\eta^2\left(\sum_{t=1}^T t^{-s}\right)^{-1}\right)^2}\mathbb{E}\left[\left(\frac{\sum_{t=1}^T t^{-\frac{s}{2}}\eta_t}{\sum_{t=1}^T t^{-s}}\right)^4\right] \tag{43}$$

$$= \frac{\left(\sum_{t=1}^T t^{-2s}\sigma_\eta^4\kappa_\eta^4 + 6\frac{1}{2}\left(\sum_{t=1}^T t^{-s}\sigma_\eta^2\right)^2 - 6\frac{1}{2}\sum_{t=1}^T t^{-2s}\sigma_\eta^4\right)}{\sigma_\eta^4\left(\sum_{t=1}^T t^{-s}\right)^2} \tag{44}$$

$$= \frac{\kappa_\eta^4 H_{2s} + 3H_s^2 - 3H_{2s}}{H_s^2} = \left(\kappa_\eta^4 - 3\right)\frac{H_{2s}}{H_s^2} + 3 \tag{45}$$

So excess kurtosis is $\left(\kappa_\eta^4 - 3\right)\frac{H_{2s}}{H_s^2}$. □

Proposition 4.5

Define $\kappa_\nu^4 \equiv \frac{E(\nu_t^4)}{\sigma_\nu^4}$. *If* $\sigma_\eta^2 = 0$, $\mathbb{E}\left[\frac{\left(\hat{\theta}_{pw}-\theta\right)^4}{\mathbb{V}\left[\hat{\theta}_{pw}\right]^2}\right] - 3 = \left(\kappa_\nu^4 - 3\right)\frac{H_{4s}}{\left(H_{2s}\right)^2}$.

Proof. We now have

$$\mathbb{E}\left[\frac{\left(\hat{\theta}_{pw}-\theta\right)^4}{\mathbb{V}\left[\hat{\theta}_{pw}\right]^2}\right] = \frac{1}{\left(\sigma_\nu^2\left(\sum_{t=1}^T t^{-s}\right)^{-2}\sum_{t=1}^T t^{-2s}\right)^2}\mathbb{E}\left[\left(\frac{\sum_{t=1}^T t^{-s}\nu_t}{\sum_{t=1}^T t^{-s}}\right)^4\right] \tag{46}$$

$$= \frac{\left(\sum_{t=1}^T t^{-4s}\sigma_\nu^4\kappa_\nu^4 + 6\frac{1}{2}\left(\sum_{t=1}^T t^{-2s}\sigma_\nu^2\right)^2 - 6\frac{1}{2}\sum_{t=1}^T t^{-4s}\sigma_\nu^4\right)}{\sigma_\nu^4\left(\sum_{t=1}^T t^{-2s}\right)^2} \tag{47}$$

$$= \frac{\kappa_\nu^4 H_{4s} + 3H_{2s}^2 - 3H_{4s}}{\left(H_{2s}\right)^2} = \left(\kappa_\nu^4 - 3\right)\frac{H_{4s}}{\left(H_{2s}\right)^2} + 3 \tag{48}$$

So excess kurtosis is $\left(\kappa_\nu^4 - 3\right)\frac{H_{4s}}{\left(H_{2s}\right)^2}$. □

Proposition 4.6

Define $\hat{\epsilon}_t \equiv y_t - \hat{\theta}_{pw}$, *and* $\hat{V} \equiv \frac{T}{T-1}\left(\sum_{t=1}^T t^{-s}\right)^{-2}\sum_{t=1}^T t^{-2s}\hat{\epsilon}_t^2$. *As above, suppose that* $\hat{\theta}_{pw}$ *is defined as in Equation 9, where* η_t *and* ν_t *are mean-zero; independently distributed; and ho-*

moskedastic with finite variances σ_η^2 *and* σ_ν^2*, respectively. Then* $\mathbb{E}\left[\frac{\hat{V}}{\mathbb{V}\left[\hat{\theta}_{pw}\right]}\right] = \frac{T}{T-1}\left(1 - 2H_s^{-1}\frac{\sigma_\eta^2 H_{2s}+\sigma_\nu^2 H_{3s}}{\sigma_\eta^2 H_s+\sigma_\nu^2 H_{2s}} + H$

Proof. The true variance of the weighted estimator is shown in the first line of Equation 39.

To estimate the variance, however, the standard equations lead to

$$\hat{\epsilon}_t \equiv y_t - \hat{\theta}_{pw} = t^{\frac{s}{2}}\eta_t + \nu_t - \left(\sum_{k=1}^{T} k^{-s}\right)^{-1}\sum_{k=1}^{T}\left(k^{-\frac{s}{2}}\eta_k + k^{-s}\nu_k\right) \tag{49}$$

$$\begin{aligned}\hat{\epsilon}_t^2 = &\left(t^{\frac{s}{2}}\eta_t + \nu_t\right)^2 - 2\left(t^{\frac{s}{2}}\eta_t + \nu_t\right)\left(\sum_{k=1}^{T} k^{-s}\right)^{-1}\sum_{k=1}^{T}\left(k^{-\frac{s}{2}}\eta_k + k^{-s}\nu_k\right)\\ &+ \left(\sum_{k=1}^{T} k^{-s}\right)^{-2}\left(\sum_{k=1}^{T}\left(k^{-\frac{s}{2}}\eta_k + k^{-s}\nu_k\right)\right)^2\end{aligned} \tag{50}$$

$$\begin{aligned}\hat{V} &\equiv \frac{T}{T-1}\left(\sum_{t=1}^{T} t^{-s}\right)^{-2}\sum_{t=1}^{T} t^{-2s}\hat{\epsilon}_t^2\\ &= \frac{T}{T-1}\left(\sum_{t=1}^{T} t^{-s}\right)^{-2}\left(\sum_{t=1}^{T}\left(t^{-\frac{s}{2}}\eta_t + t^{-s}\nu_t\right)^2\right)\\ &\quad - 2\frac{T}{T-1}\left(\sum_{t=1}^{T} t^{-s}\right)^{-3}\left(\sum_{t=1}^{T}\left(t^{-\frac{s}{2}}\eta_t + t^{-s}\nu_t\right)\right)\left(\sum_{t=1}^{T}\left(t^{-\frac{3s}{2}}\eta_t + t^{-2s}\nu_t\right)\right)\\ &\quad + \frac{T}{T-1}\left(\sum_{t=1}^{T} t^{-s}\right)^{-4}\left(\sum_{t=1}^{T} t^{-2s}\right)\left(\sum_{t=1}^{T} t^{-\frac{s}{2}}\eta_t + t^{-s}\nu_t\right)^2\end{aligned} \tag{51}$$

We now have that

$$\mathbb{E}\left[\frac{\hat{V}}{\mathbb{V}\left[\hat{\theta}_{pw}\right]}\right] = \frac{T}{T-1} - 2\frac{T}{T-1}H_s^{-1}\frac{\sigma_\eta^2 H_{2s} + \sigma_\nu^2 H_{3s}}{\sigma_\eta^2 H_s + \sigma_\nu^2 H_{2s}} + \frac{T}{T-1}H_s^{-2}H_{2s}. \tag{52}$$

□

Theorem 4.7

Define $\hat{\epsilon}_t \equiv y_t - \hat{\theta}_{pw}$*, and* $\hat{V} \equiv \frac{T}{T-1}\left(\sum_{t=1}^{T} t^{-s}\right)^{-2}\sum_{t=1}^{T} t^{-2s}\hat{\epsilon}_t^2$*. As above, suppose that* $\hat{\theta}_{pw}$ *is defined as in Equation 9, where* η_t *and* ν_t *are mean-zero; independently distributed; ho-moskedastic with finite variances* σ_η^2 *and* σ_ν^2*, respectively, at least one of which is non-zero;*

and have uniformly bounded kurtosis. If $s \leq \frac{1}{2}$*, then* $\frac{\hat{V}}{\mathbb{V}\left[\hat{\theta}_{pw}\right]} \xrightarrow{p} 1$*. If* $\frac{1}{2} < s \leq 1$*, and additionally* ν_t *takes on at least 3 values for some observation* t*, then* $\frac{\hat{V}}{\mathbb{V}\left[\hat{\theta}_{pw}\right]} \xrightarrow{p} 1$ *if and only if* $\sigma_\eta^2 > 0$*. If* $s > 1$*, then* $\frac{\hat{V}}{\mathbb{V}\left[\hat{\theta}_{pw}\right]}$ *does not converge.*

Proof. (*Convergence*) First, note that if $s \leq 1$, $H_s \to \infty$, $\frac{H_{2s}}{H_s} \to 0$, $\frac{H_{3s}}{H_s} \to 0$, and $\frac{H_{4s}}{H_s} \to 0$ as $T \to \infty$. From this, we can see from Equation 52 that $\mathbb{E}\left[\frac{\hat{V}}{\mathbb{V}\left[\hat{\theta}_{pw}\right]}\right] \to 1$ as $T \to \infty$. I therefore only must prove that $\mathbb{V}\left[\frac{\hat{V}}{\mathbb{V}\left[\hat{\theta}_{pw}\right]}\right] \to 0$ as $T \to \infty$. I begin by proving this for the case where $\nu_t = 0$, and then the case where $\eta_t = 0$. I then note that if neither term equals zero, the η_t terms will dominate, so convergence follows from convergence of that term.

$\frac{T}{T-1} \to 1$, so using Slutsky I will ignore this.

First, assume $\nu_t = 0$ and $s \leq 1$. Further, assume that η_t has a variance of 1 (since $\mathbb{V}\left[\hat{\theta}_{pw}\right] = \sigma_\eta^2 H_s^{-1}$, we are dividing by the variance of η, so this is without loss of generality) and define $\kappa_\eta^4 \equiv \sup_t \mathbb{E}[\eta_t^4]$, which exists by assumption. Defining $\hat{V}_1$ as the first term of Equation 51, $\hat{V}_2$ as the second term, and $\hat{V}_3$ as the third term (that is, $\hat{V} = \hat{V}_1 - \hat{V}_2 + \hat{V}_3$),

we have

$$
\begin{aligned}
& \mathbb{V}\left[\frac{\hat{V}}{\mathbb{V}\left[\hat{\theta}_{pw}\right]}\right] \\
= \; & H_s^2\left(\mathbb{V}[\hat{V}_1]+\mathbb{V}[\hat{V}_2]+\mathbb{V}[\hat{V}_3]-2\mathbb{C}[\hat{V}_1,\hat{V}_2]+2\mathbb{C}[\hat{V}_1,\hat{V}_3]-2\mathbb{C}[\hat{V}_2,\hat{V}_3]\right) \qquad (53) \\
= \; & H_s^{-2}\mathbb{E}\left[\sum_{t=1}^{T} t^{-s}\eta_t^2 \sum_{k=1}^{T} k^{-s}\eta_k^2\right]-1 \\
& +4H_s^{-4}\mathbb{E}\left[\sum_{t=1}^{T} t^{-\frac{s}{2}}\eta_t \sum_{k=1}^{T} k^{-\frac{3s}{2}}\eta_k \sum_{l=1}^{T} l^{-\frac{s}{2}}\eta_l \sum_{m=1}^{T} m^{-\frac{3s}{2}}\eta_m\right]-\left(2H_s^{-1}\frac{H_{2s}}{H_s}\right)^2 \\
& +H_s^{-6}H_{2s}^2\mathbb{E}\left[\sum_{t=1}^{T} t^{-\frac{s}{2}}\eta_t \sum_{k=1}^{T} k^{-\frac{s}{2}}\eta_k \sum_{l=1}^{T} l^{-\frac{s}{2}}\eta_l \sum_{m=1}^{T} m^{-\frac{s}{2}}\eta_m\right]-(H_s^{-2}H_{2s})^2 \\
& -2\times 2H_s^{-3}\mathbb{E}\left[\sum_{t=1}^{T} t^{-s}\eta_t^2 \sum_{k=1}^{T} k^{-\frac{s}{2}}\eta_k \sum_{l=1}^{T} l^{-\frac{3s}{2}}\eta_l\right]+2\times 1\times 2H_s^{-2}H_{2s} \\
& +2\times H_s^{-4}H_{2s}\mathbb{E}\left[\sum_{t=1}^{T} t^{-s}\eta_t^2 \sum_{k=1}^{T} k^{-\frac{s}{2}}\eta_k \sum_{l=1}^{T} l^{-\frac{s}{2}}\eta_l\right]-2\times 1\times H_s^{-2}H_{2s} \\
& -2\times 2H_s^{-5}H_{2s}\mathbb{E}\left[\sum_{t=1}^{T} t^{-\frac{s}{2}}\eta_t \sum_{k=1}^{T} k^{-\frac{3s}{2}}\eta_k \sum_{l=1}^{T} l^{-\frac{s}{2}}\eta_l \sum_{m=1}^{T} m^{-\frac{s}{2}}\eta_m\right]+2\times 2H_s^{-2}H_{2s}\times H_s^{-2}H_{2s} \qquad (54)
\end{aligned}
$$

$$
\begin{aligned}
&= H_s^{-2}\mathbb{E}\left[\sum_{t=1}^{T}\sum_{k=1}^{T} t^{-s}k^{-s}\eta_t^2\eta_k^2\right] - 1 \\
&\quad +4H_s^{-4}\mathbb{E}\left[\sum_{t=1}^{T}\sum_{k=1}^{T}\sum_{l=1}^{T}\sum_{m=1}^{T} t^{-\frac{s}{2}}k^{-\frac{3s}{2}}l^{-\frac{s}{2}}m^{-\frac{3s}{2}}\eta_t\eta_k\eta_l\eta_m\right] - 4H_s^{-4}H_{2s}^2 \\
&\quad +H_s^{-6}H_{2s}^2\mathbb{E}\left[\sum_{t=1}^{T}\sum_{k=1}^{T}\sum_{l=1}^{T}\sum_{m=1}^{T} t^{-\frac{s}{2}}k^{-\frac{s}{2}}l^{-\frac{s}{2}}m^{-\frac{s}{2}}\eta_t\eta_k\eta_l\eta_m\right] - H_s^{-4}H_{2s}^2 \\
&\quad -4H_s^{-3}\mathbb{E}\left[\sum_{t=1}^{T}\sum_{k=1}^{T}\sum_{l=1}^{T} t^{-s}k^{-\frac{s}{2}}l^{-\frac{3s}{2}}\eta_t^2\eta_k\eta_l\right] + 4H_s^{-2}H_{2s} \\
&\quad +2H_s^{-4}H_{2s}\mathbb{E}\left[\sum_{t=1}^{T}\sum_{k=1}^{T}\sum_{l=1}^{T} t^{-s}k^{-\frac{s}{2}}l^{-\frac{s}{2}}\eta_t^2\eta_k\eta_l\right] - 2H_s^{-2}H_{2s} \\
&\quad -4H_s^{-4}H_{2s}\mathbb{E}\left[\sum_{t=1}^{T}\sum_{k=1}^{T}\sum_{l=1}^{T}\sum_{m=1}^{T} t^{-\frac{s}{2}}k^{-\frac{3s}{2}}l^{-\frac{s}{2}}m^{-\frac{s}{2}}\eta_t\eta_k\eta_l\eta_m\right] + 4H_s^{-4}H_{2s}^2 \qquad (55) \\
&\leq H_s^{-2}\left[H_s^2 + H_{2s}(\kappa_\eta^4 - 1)\right] - 1 \\
&\quad +4H_s^{-4}\left[H_sH_{3s} + 2H_{2s}^2 + H_{4s}(\kappa_\eta^4 - 3)\right] - 4H_s^{-4}H_{2s}^2 \\
&\quad +H_s^{-6}H_{2s}^2\left[3H_s^2 + H_{2s}(\kappa_\eta^4 - 3)\right] - H_s^{-4}H_{2s}^2 \\
&\quad -4H_s^{-3}\left[H_sH_{2s} + H_{3s}(\kappa_\eta^4 - 1)\right] + 4H_s^{-2}H_{2s} \\
&\quad +2H_s^{-4}H_{2s}\left[H_s^2 + H_{2s}(\kappa_\eta^4 - 1)\right] - 2H_s^{-2}H_{2s} \\
&\quad -4H_s^{-5}H_{2s}\left[3H_{2s}H_s + H_{3s}(\kappa_\eta^4 - 3)\right] + 4H_s^{-4}H_{2s}^2 \qquad (56) \\
&\to 0 \qquad (57)
\end{aligned}
$$

where the limit uses the facts, as noted previously, that $H_s \to \infty$, $\frac{H_{2s}}{H_s} \to 0$, $\frac{H_{3s}}{H_s} \to 0$, and $\frac{H_{4s}}{H_s} \to 0$ as $T \to \infty$.

Next, assume $\eta_t = 0$ and $s \leq \frac{1}{2}$. Note that now, $H_{2s} \to \infty$. Further, assume that ν_t has a variance of 1 (since $\mathbb{V}\left[\hat{\theta}_{pw}\right] = \sigma_\nu^2 H_s^{-2} H_{2s}$, we are dividing by the variance of ν, so this is without loss of generality) and define $\kappa_\nu^4 \equiv \sup_t \mathbb{E}[\nu_t^4]$, which exists by assumption.

$$\mathbb{V}\left[\frac{\hat{V}}{\mathbb{V}\left[\hat{\theta}_{pw}\right]}\right] \tag{58}$$

$$= H_s^4 H_{2s}^{-2}\left(\mathbb{V}[\hat{V}_1] + \mathbb{V}[\hat{V}_2] + \mathbb{V}[\hat{V}_3] - 2\mathbb{C}[\hat{V}_1, \hat{V}_2] + 2\mathbb{C}[\hat{V}_1, \hat{V}_3] - 2\mathbb{C}[\hat{V}_2, \hat{V}_3]\right) \tag{59}$$

$$\begin{aligned}
= \; & H_{2s}^{-2}\mathbb{E}\left[\sum_{t=1}^{T} t^{-2s}\nu_t^2 \sum_{k=1}^{T} k^{-2s}\nu_k^2\right] - 1 \\
& + 4H_s^{-2}H_{2s}^{-2}\mathbb{E}\left[\sum_{t=1}^{T} t^{-s}\nu_t \sum_{k=1}^{T} k^{-2s}\nu_k \sum_{l=1}^{T} l^{-s}\nu_l \sum_{m=1}^{T} m^{-2s}\nu_m\right] - \left(2H_s^{-1}\frac{H_{3s}}{H_{2s}}\right)^2 \\
& + H_s^{-4}\mathbb{E}\left[\sum_{t=1}^{T} t^{-s}\nu_t \sum_{k=1}^{T} k^{-s}\nu_k \sum_{l=1}^{T} l^{-s}\nu_l \sum_{m=1}^{T} m^{-s}\nu_m\right] - (H_s^{-2}H_{2s})^2 \\
& - 2 \times 2H_s^{-1}H_{2s}^{-2}\mathbb{E}\left[\sum_{t=1}^{T} t^{-2s}\nu_t^2 \sum_{k=1}^{T} k^{-s}\nu_k \sum_{l=1}^{T} l^{-2s}\nu_l\right] + 2 \times 1 \times 2H_s^{-1}\frac{H_{3s}}{H_{2s}} \\
& + 2 \times H_s^{-2}H_{2s}^{-1}\mathbb{E}\left[\sum_{t=1}^{T} t^{-2s}\nu_t^2 \sum_{k=1}^{T} k^{-s}\nu_k \sum_{l=1}^{T} l^{-s}\nu_l\right] - 2 \times 1 \times H_s^{-2}H_{2s} \\
& - 2 \times 2H_s^{-3}H_{2s}^{-1}\mathbb{E}\left[\sum_{t=1}^{T} t^{-s}\nu_t \sum_{k=1}^{T} k^{-2s}\nu_k \sum_{l=1}^{T} l^{-s}\nu_l \sum_{m=1}^{T} m^{-s}\nu_m\right] + 2 \times 2H_s^{-1}\frac{H_{3s}}{H_{2s}} \times H_s^{-2}H_{2s}
\end{aligned} \tag{60}$$

$$
\begin{aligned}
&= H_{2s}^{-2}\mathbb{E}\left[\sum_{t=1}^{T}\sum_{k=1}^{T} t^{-2s}k^{-2s}\nu_t^2\nu_k^2\right] - 1 \\
&\quad +4H_s^{-2}H_{2s}^{-2}\mathbb{E}\left[\sum_{t=1}^{T}\sum_{k=1}^{T}\sum_{l=1}^{T}\sum_{m=1}^{T} t^{-s}k^{-2s}l^{-s}m^{-2s}\nu_t\nu_k\nu_l\nu_m\right] - 4H_s^{-2}H_{2s}^{-2}H_{3s}^2 \\
&\quad +H_s^{-4}\mathbb{E}\left[\sum_{t=1}^{T}\sum_{k=1}^{T}\sum_{l=1}^{T}\sum_{m=1}^{T} t^{-s}k^{-s}l^{-s}m^{-s}\nu_t\nu_k\nu_l\nu_m\right] - H_s^{-4}H_{2s}^2 \\
&\quad -4H_s^{-1}H_{2s}^{-2}\mathbb{E}\left[\sum_{t=1}^{T}\sum_{k=1}^{T}\sum_{l=1}^{T} t^{-2s}k^{-s}l^{-2s}\nu_t^2\nu_k\nu_l\right] + 4H_s^{-1}H_{3s}H_{2s}^{-1} \\
&\quad +2H_s^{-2}H_{2s}^{-1}\mathbb{E}\left[\sum_{t=1}^{T}\sum_{k=1}^{T}\sum_{l=1}^{T} t^{-2s}k^{-s}l^{-s}\nu_t^2\nu_k\nu_l\right] - 2H_s^{-2}H_{2s} \\
&\quad -4H_s^{-3}H_{2s}^{-1}\mathbb{E}\left[\sum_{t=1}^{T}\sum_{k=1}^{T}\sum_{l=1}^{T}\sum_{m=1}^{T} t^{-s}k^{-2s}l^{-s}m^{-s}\nu_t\nu_k\nu_l\nu_m\right] + 4H_s^{-3}H_{3s} \qquad (61) \\
&\leq H_{2s}^{-2}\left[H_{2s}^2 + H_{4s}(\kappa_\nu^4 - 1)\right] - 1 \\
&\quad +4H_s^{-2}H_{2s}^{-2}\left[H_{2s}H_{4s} + 2H_{3s}^2 + H_{6s}(\kappa_\nu^4 - 3)\right] - 4H_s^{-2}H_{2s}^{-2}H_{3s}^2 \\
&\quad +H_s^{-4}\left[3H_{2s}^2 + H_{4s}(\kappa_\nu^4 - 3)\right] - H_s^{-4}H_{2s}^2 \\
&\quad -4H_s^{-1}H_{2s}^{-2}\left[H_{2s}H_{3s} + H_{5s}(\kappa_\nu^4 - 1)\right] + 4H_s^{-1}H_{3s}H_{2s}^{-1} \\
&\quad +2H_s^{-2}H_{2s}^{-1}\left[H_{2s}^2 + H_{4s}(\kappa_\nu^4 - 1)\right] - 2H_s^{-2}H_{2s} \\
&\quad -4H_s^{-3}H_{2s}^{-1}\left[3H_{3s}H_{2s} + H_{5s}(\kappa_\eta^4 - 3)\right] + 4H_s^{-3}H_{3s} \\
&\to 0 \qquad (62)
\end{aligned}
$$

Note that if $\sigma_\eta^2 > 0$ and $\sigma_\nu^2 > 0$, the η term will dominate as $T \to \infty$. So the fact that the variance goes to zero as $T \to \infty$ when $\nu_t = 0$ means that $\mathbb{V}\left[\frac{\hat{V}}{\mathbb{V}\left[\hat{\theta}_{pw}\right]}\right] \to 0$ even when $\sigma_\eta^2 > 0$ and $\sigma_\nu^2 > 0$.

□

Proof. (*Non-convergence if* $s > 1$) Note that $H_s \to C_1$ for some constant C_1, while $H_{2s} \to C_2$

for some other constant C_2. Now, using the first line of Equation 39,

$$\mathbb{V}\left[\hat{\theta}_{pw}\right] \to \sigma_\eta^2 C_1^{-1} + \sigma_\nu^2 C_1^{-2} C_2, \tag{63}$$

a constant. We also have

$$\hat{\theta} = \frac{1}{\sum_{t=1}^T t^{-s}} \sum_{t=1}^T t^{-s}(\theta + \epsilon) = \theta + H_s^{-1} \sum_{t=1}^T t^{-s}\epsilon_t \tag{64}$$

$$\hat{\epsilon}_t = y_t - \hat{\theta} = \theta + \epsilon - \hat{\theta} = \epsilon_t - H_s^{-1} \sum_{k=1}^T k^{-s}\epsilon_k \tag{65}$$

$$\hat{\epsilon}_t^2 = \epsilon_t^2 - 2\epsilon_t H_s^{-1} \sum_{k=1}^T k^{-s}\epsilon_k + H_s^{-2} \left(\sum_{k=1}^T k^{-s}\epsilon_k\right)^2 \tag{66}$$

$$\hat{V} \equiv \frac{T}{T-1}\left(\sum_{t=1}^T t^{-s}\right)^{-2} \sum_{t=1}^T t^{-2s}\hat{\epsilon}_t^2 = \frac{T}{T-1} H_s^{-2} \sum_{t=1}^T t^{-2s}\hat{\epsilon}_t^2 \tag{67}$$

To simplify, I will show that $\tilde{V} \equiv \sum_{t=1}^T t^{-2s}\hat{\epsilon}_t^2$ converges to a nondegenerate random variable. By Slutsky,

$$\text{plim}\frac{\hat{V}}{\mathbb{V}\left[\hat{\theta}_{pw}\right]} = \text{plim}\left(\sigma_\eta^2 C_1^{-1} + \sigma_\nu^2 C_1^{-2} C_2\right)^{-1} C_1^{-2}\tilde{V} \tag{68}$$

$$= \left[(\sigma_\eta^2)^{-1} C_1^{-1} + (\sigma_\nu^2)^{-1} C_2^{-1}\right] \text{plim}\tilde{V} \tag{69}$$

so if $\tilde{V}$ converges to nondegenerate random variable, $\frac{\hat{V}}{\mathbb{V}\left[\hat{\theta}_{pw}\right]}$ does as well.

Now,

$$\tilde{V} = \sum_{t=1}^{T} t^{-2s}\epsilon_t^2 - 2H_s^{-1}\sum_{t=1}^{T} t^{-2s}\epsilon_t \sum_{k=1}^{T} k^{-s}\epsilon_k + H_s^{-2}\left(\sum_{t=1}^{T} t^{-2s}\right)\left(\sum_{k=1}^{T} k^{-s}\epsilon_k\right)^2 \tag{70}$$

$$\begin{aligned}
=&\epsilon_1\epsilon_2\left(-2H_s^{-1}2^{-s} - 2H_s^{-1}2^{-2s} + 2H_s^{-2}H_{2s}2^{-s}\right) \\
&+ \epsilon_1^2\left(1 - 2H_s^{-1} + H_s^{-2}H_{2s}\right) \\
&+ \epsilon_2^2\left(2^{-2s} - 2H_s^{-1}2^{-3s} + H_s^{-2}H_{2s}2^{-2s}\right) \\
&+ \epsilon_1\left(-2H_s^{-1}\sum_{k=3}^{T} k^{-s}\epsilon_k - 2H_s^{-1}\sum_{t=3}^{T} t^{-2s}\epsilon_t + 2H_s^{-2}H_{2s}\sum_{t=3}^{T} t^{-s}\epsilon_t\right) \\
&+ \epsilon_2\left(-2H_s^{-1}2^{-2s}\sum_{k=3}^{T} k^{-s}\epsilon_k - 2H_s^{-1}2^{-s}\sum_{t=3}^{T} t^{-2s}\epsilon_t + 2H_s^{-2}H_{2s}2^{-s}\sum_{t=3}^{T} t^{-s}\epsilon_t\right) \\
&+ \sum_{t=3}^{T} t^{-2s}\epsilon_t^2 - 2H_s^{-1}\sum_{t=3}^{T} t^{-2s}\epsilon_t \sum_{k=3}^{T} k^{-s}\epsilon_k + H_s^{-2}H_{2s}\left(\sum_{t=3}^{T} t^{-s}\epsilon_t\right)^2
\end{aligned} \tag{71}$$

Crucially, the only term in Equation 71 that includes both ϵ_1 and ϵ_2 is the first term.

Now, note that by assumption, at least one of σ_η^2 and σ_ν^2 is non-zero, so both ϵ_1 and ϵ_2 each take on at least two values each with nonzero probability. Denote these $\epsilon_{1,1}$ and $\epsilon_{1,1}$ for ϵ_1; and $\epsilon_{2,2}$ and $\epsilon_{2,2}$ for ϵ_2. For some fixed values of $\{\epsilon_3 ... \epsilon_T\}$, denote by $\tilde{V}_{i,j}$ the value of $\tilde{V}$ when $\epsilon_1 = \epsilon_{1,i}$ and $\epsilon_2 = \epsilon_{2,j}$. Now, we calculate

$$\begin{aligned}
\left(\tilde{V}_{1,1} - \tilde{V}_{1,2}\right) - \left(\tilde{V}_{2,1} - \tilde{V}_{2,2}\right) =& \left((\epsilon_{1,1}\epsilon_{2,1} - \epsilon_{1,1}\epsilon_{2,2}) - (\epsilon_{1,2}\epsilon_{2,1} - \epsilon_{1,2}\epsilon_{2,2})\right) \\
&\times \left(-2H_s^{-1}2^{-s} - 2H_s^{-1}2^{-2s} + 2H_s^{-2}H_{2s}2^{-s}\right) \\
=& \left(\epsilon_{1,1} - \epsilon_{1,2}\right)\left(\epsilon_{2,1} - \epsilon_{2,2}\right) \\
&\times \left(-2H_s^{-1}2^{-s} - 2H_s^{-1}2^{-2s} + 2H_s^{-2}H_{2s}2^{-s}\right) \\
\xrightarrow{p}& \left(\epsilon_{1,1} - \epsilon_{1,2}\right)\left(\epsilon_{2,1} - \epsilon_{2,2}\right) \qquad (72) \\
&\times \left(-2C_1^{-1}2^{-s} - 2C_1^{-1}2^{-2s} + 2C_1^{-2}C_2 2^{-s}\right) \qquad (73)
\end{aligned}$$

where all other terms in $\tilde{V}$ are dropped because they do not contain both ϵ_1 and ϵ_2. This limit can only equal 0 if $\epsilon_{1,1} = \epsilon_{1,2}$ or $\epsilon_{2,1} = \epsilon_{2,2}$, which is not true by assumption. Therefore plim $\left[\left(\tilde{V}_{1,1} - \tilde{V}_{1,2}\right) - \left(\tilde{V}_{2,1} - \tilde{V}_{2,2}\right)\right]$ takes on at least two values with nonzero probability, so plim$\tilde{V}$ also takes on at least two values with nonzero probability, so it is a nondegenerate random variable.

□

Proof. (*Non-convergence if* $\frac{1}{2} < s \leq 1$, ν_t *takes on at least 3 values for some* t, *and* $\sigma_\eta^2 = 0$) We now have

$$\mathbb{V}\left[\hat{\theta}_{pw}\right] = \sigma_\nu^2 H_s^{-2} H_{2s}, \tag{74}$$

so

$$\frac{\hat{V}}{\mathbb{V}\left[\hat{\theta}_{pw}\right]} = \frac{T}{T-1} H_{2s}^{-1} \sum_{t=1}^{T} t^{-2s} \hat{\epsilon}_t^2 \tag{75}$$

Because $\frac{T}{T-1} H_{2s}^{-1} \to C_2$, and continuing to define $\tilde{V} \equiv \sum_{t=1}^{T} t^{-2s} \hat{\epsilon}_t^2$, convergence of $\frac{\hat{V}}{\mathbb{V}\left[\hat{\theta}_{pw}\right]}$ again depends on convergence of $\tilde{V}$. However, because now $H_s \to \infty$, $(-2H_s^{-1}2^{-s} - 2H_s^{-1}2^{-2s} + 2H_s^{-2}H_{2s}2^{-}$ 0, so the previous proof does not go through.

Instead, consider that t for which $\epsilon_t = \nu_t$ takes on at least 3 values. We have

$$\tilde{V} = \nu_t^2 \left(t^{-2s} - 2H_s^{-1} t^{-3s} + H_s^{-2} H_{2s} t^{-2s}\right) \tag{76}$$

$$+ \nu_t \left(-2H_s^{-1} t^{-2s} \sum_{k \neq t}^{T} k^{-s} \nu_k - 2H_s^{-1} t^{-s} \sum_{k \neq t}^{T} k^{-2s} \nu_k + 2H_s^{-2} H_{2s} t^{-s} \sum_{k \neq t}^{T} k^{-s} \nu_k\right) \tag{77}$$

$$+ \sum_{k \neq t}^{T} k^{-2s} \nu_k^2 - 2H_s^{-1} \sum_{k \neq t}^{T} k^{-2s} \nu_k \sum_{k \neq t}^{T} k^{-s} \nu_k + H_s^{-2} H_{2s} \left(\sum_{k \neq t}^{T} k^{-s} \nu_k\right)^2 \tag{78}$$

Because $H_s \to \infty$, as $T \to \infty$,

$$\tilde{V} - \left(t^{-2s} \nu_t^2 + \sum_{k \neq t}^{T} k^{-2s} \nu_k^2\right) \xrightarrow{p} 0 \tag{79}$$

(the variance of all other terms goes to 0 because the kurtosis of ν_k are uniformly bounded). Thus the convergence of $\tilde{V}$ depends on the convergence of $t^{-2s}\nu_t^2 + \sum_{k\neq t}^{T} k^{-2s}\nu_k^2$. Because ν_t takes on at least 3 values, ν_t^2 must take on at least two values, so it is a nondegenerate random variable. A non-degenerate random variable, plus an independent random variable (degenerate or not) cannot equal a constant. Thus $\tilde{V}$, and therefore $\frac{\hat{V}}{\mathbb{V}\left[\hat{\theta}_{pw}\right]}$, does not converge to a constant.

□

B Alternative asymptotics

Suppose we now assume that the least accurate observation's variance is held fixed. In other words, the error term of Equation 5 will be rewritten as

$$\epsilon_t = \left(\frac{A_t}{A_T}\right)^{-\frac{1}{2}} \eta_t + \nu_t = \left(\frac{t}{T}\right)^{\frac{s}{2}} \eta_t + \nu_t, \tag{80}$$

This formulation is a valid description of the data; for any given sample size, it is equivalent to Equation 5, but with a different value for σ_η^2.

In this formulation, the unweighted estimator will be consistent. The unweighted estimator is now

$$\hat{\theta}_{uw} \equiv \frac{1}{T}\sum_{t=1}^{T} y_t = \theta + \frac{1}{T}\sum_{t=1}^{T}\left(\left(\frac{t}{T}\right)^{\frac{s}{2}} \eta_t + \nu_t\right) \tag{81}$$

To see that this is consistent, note that the variance of the (unbiased) estimator is given by

$$\begin{aligned}\mathbb{V}\left[\hat{\theta}_{uw}\right] &= \mathbb{E}\left[\left(\frac{1}{T}\sum_{t=1}^{T}\left(\frac{t}{T}\right)^{\frac{s}{2}} \eta_t + \nu_t\right)^2\right] = \sigma_\eta^2 \frac{1}{T^2}\sum_{t=1}^{T}\left(\frac{t}{T}\right)^{\frac{s}{2}} + \frac{1}{T}\sigma_\nu^2 \\ &\leq \sigma_\eta^2 \frac{1}{T^2}\sum_{t=1}^{T} 1 + \frac{1}{T}\sigma_\nu^2 = \frac{1}{T}\left(\sigma_\eta^2 + \sigma_\nu^2\right) \to 0.\end{aligned} \tag{82}$$

However, it remains true that, for $s > 1$, the unweighted estimator would be more efficient if only the most accurate observation is used. It is simply that each observation becomes arbitrarily accurate as sample size increases.

The weighted estimator is now given by

$$\begin{aligned}\hat{\theta}_{pw} &\equiv \frac{1}{\sum_{t=1}^{T}\left(\frac{t}{T}\right)^{-s}}\sum_{t=1}^{T}\left(\frac{t}{T}\right)^{-s} y_t = \frac{1}{\sum_{t=1}^{T} t^{-s}}\sum_{t=1}^{T} t^{-s} y_t \\ &= \theta + \frac{1}{\sum_{t=1}^{T} t^{-s}}\sum_{t=1}^{T}\left(\frac{1}{T^{\frac{s}{2}}} t^{-\frac{s}{2}} \eta_t + t^{-s}\nu_t\right)\end{aligned} \tag{83}$$

Note that the ν_t term in this expression is the same as in Equation 9; thus, all results from above for the ν_t term in the weighted estimator will remain valid. Focusing instead on the

η_t term, we see that it converges; if $\sigma_\nu^2 = 0$, then

$$\mathbb{V}\left[\hat{\theta}_{pw}\right] = T^{-s}\sigma_\eta^2 \left(\sum_{t=1}^{T} t^{-s}\right)^{-1} \to 0 \tag{84}$$

$$\tag{85}$$

because $T^{-s} \to 0$ for all $s > 0$, and $\left(\sum_{t=1}^{T} t^{-s}\right)^{-1}$ is bounded (and goes to 0 if $s = 0$).

Results for asymptotic normality of the η_t term will not change. The extra $T^{\frac{s}{2}}$ will function as the $g(T)$ term in Lemma 3.2 and Lemma 4.2, and therefore will not affect the conclusions.

Results on standard errors for the η_t term will also not be affected, as the extra $T^{\frac{s}{2}}$ will factor out of $\frac{\hat{V}}{\mathbb{V}\left[\hat{\theta}_{pw}\right]}$ if $\nu_t = 0$.

There are, of course, infinite other asymptotic assumptions that agree with any finite amount of data, each of which has different asymptotic implications. For example, if we assume that the errors in all remaining observations will be homoskedastic, then of course the usual assumptions apply and estimators are consistent, asymptotically normal, and have consistent standard errors. However, in any asymptotic specification in which all data obey a power law, much in this paper will remain unchanged. For example, if $s > 1$, the OLS estimator will always be less precise as more observations are added. Further, conclusions about asymptotic normality and consistency of the estimator for standard errors will remain unchanged. Finally, simulation results, because they are based on a small sample, are unaffected by such assumptions.

C Implementation of quasi-maximum likelihood in Stata

This section details the new Stata command "regoptwgt," which implements the QML procedure described in Section 5. It is available at https://www.davidjonathanprice.com and will soon be available through the Statistical Software Components (SSC) archive.

In many cases, regoptwgt can simply replace other commands to replace WLS with QML. For example, the following command:

- `reg y x1 x2 x3 [w=wgt]`

can be replaced with:

- `regoptwgt y x1 x2 x3 [w=wgt]`

Likewise, the following command:

- `ivregress 2sls y x1 (x2 x3 = x4 x5 x6) [w=wgt]`

can be replaced with:

- `regoptwgt y x1 (x2 x3 = x4 x5 x6) [w=wgt]`

Some details about the command that are important to note:

- By default, regoptwgt initializes the search for parameters based on results from an unweighted analysis (OLS for the one-equation case, two-stage least squares in the multiple-equation case). In the simulations in Section 6, if QML fails to converge, the process is repeated without this initialization. In a set of 6.6 million simulation estimations, both of these processes failed to converge for the same estimation only 99 times. (Of these, 98 were cases with $s = 2$ and $\nu = 0$—that is, the highest heteroskedasticity I tested. Extra care may be needed in that extreme setting.)

- The log likelihood function is maximized using the lf1 method. This means that the first derivatives of the log likelihood function are hard-coded into the command and are used to find the maximum.

- By default, regoptwgt cycles through four maximization techniques: Newton-Raphson (nr), Berndt-Hall-Hall-Hausman (bhhh), Davidon-Fletcher-Powell (dfp), and Broyden-Fletcher-Goldfarb-Shanno (bfgs).

- Although the creation of regoptwgt is motivated by power laws, as described in this paper, this command can improve precision any time a researcher is considering using weights.

- Standard errors may be clustered according to one or more variables. Note that clustering does not change estimator; it only adjusts standard errors. Future work could allow for covariance within clusters to affect the optimal weight, which could results in a more precise estimator.

- If standard errors are clustered along more than one dimension, the command uses vcemway, as described by Gu and Yoo (2019).

- The command regoptwgt can be quite fast; in simulations, estimation was typically complete within five seconds. However, although regoptwgt works with any number of endogenous variables, it is substantially slower with two or more endogenous right-hand-side variables. This is because it uses mata to calculate some parameters, and overhead from the use of mata appears to slow down the command. Future work involves speeding up this calculation.

- At present, factor variables are not supported by the command. Future work involves incorporating them.

D Additional figures

All figures in this section are results of simulations. They are based on the following model:

$$\epsilon_t = k^{\frac{1}{2}} t^{\frac{s}{2}} \eta_t + \nu_t \tag{86}$$

$$\eta_t \sim \mathbb{N}(0,1) \tag{87}$$

$$\nu_t \sim \text{Exp}(1) - 1 \tag{88}$$

$$k = \frac{\frac{H_{2s}(T)}{H_s(T)^2} - \frac{1}{T}}{\frac{H_{-s}(T)}{T^2} - H_s(T)^{-1}} e^{\Phi^{-1}(h)} \tag{89}$$

$$z_t \sim \mathbb{N}(0,1) \tag{90}$$

$$w_t \sim \mathbb{N}(0,1) \tag{91}$$

$$x_t = 2z_t + w_t + \xi_t \tag{92}$$

$$\xi_t \sim \mathbb{N}(0,1) \tag{93}$$

$$y_t = w_t + \epsilon_t \tag{94}$$

where h is the level of heteroskedasticity on the x-axis and $\Phi^{-1}(\cdot)$ is the inverse of a standard normal CDF.

Estimators for an average are estimating the mean of ϵ_t. Estimators of a regression coefficient have ϵ_t as a dependent variable and z_t as an independent variable. Estimators of an instrumental variables coefficient have y_t as a dependent variable, x_t as an endogenous regressor, and z_t as an excluded instrument.

In addition to this model, some estimators are calculated with a "disaggregated" technique, similar to the setup at the start of Section 2. For these techniques, each of the 1,000 initial observations is expanded so the total number of observations in the group formed by

each initial observation t is approximately equal to $4000 \times t^{-s}$. I then set

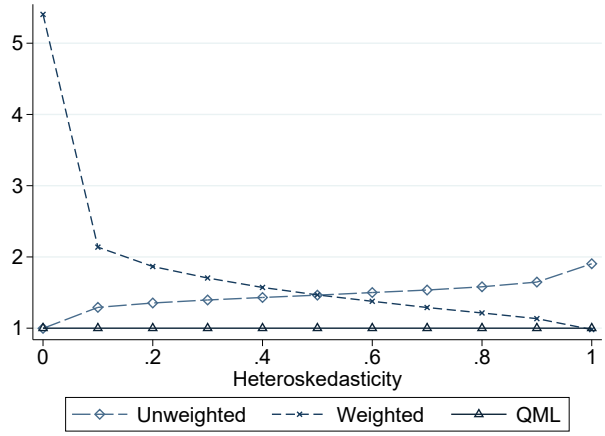
$$\epsilon_{it} = (4000 \times k)^{\frac{1}{2}} \eta_{it} + \nu_t \tag{95}$$

$$\eta_{it} \sim \mathbb{N}(0, 1) \tag{96}$$

Figure 8 shows the root mean squared (RMS) error on an estimated average for different methods, relative to the RMS of QML, with $s = 1$. Figure 9 shows the same for a regression coefficient; Figure 10 shows the same for a coefficient from an instrumental variables regression. Figures 11, 12, and 13 are equivalent graphs for $s = 2$.

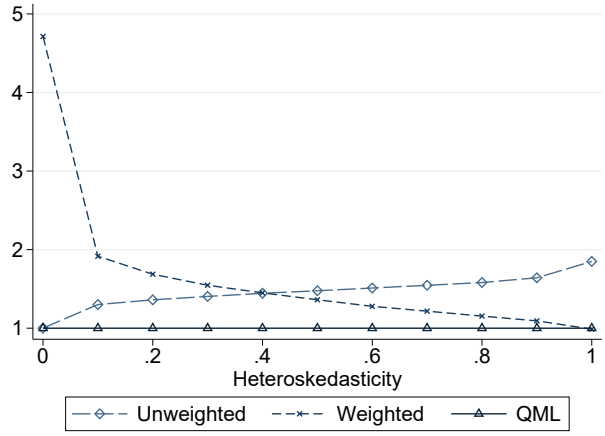
Figure 14 shows the size of confidence intervals with various nominal sizes when testing a true condition on an average. Figure 15 shows the same for a regression coefficient. Figure 16 shows the same for an instrumental variables coefficient. Figures 17, 18, and 19 are equivalent graphs for $s = 2$. "Unweighted" and "Weighted, HC1" use Stata's "robust" option. "Weighted, HC3" uses Stata's "vce(hc3)" option. "Disaggregated, CV1" uses Stata's "cluster(t)" option. "Disaggregated, Jackknife" uses Stata's "cluster(t) vce(jackknife, mse)" options. "Disaggregated, Bootstrap" uses a wild cluster bootstrap with the user-written Stata command "boottest," as described by Roodman et al. (2019).

Figure 8: Relative root mean squared error in simulations estimating an average, $s = 1$



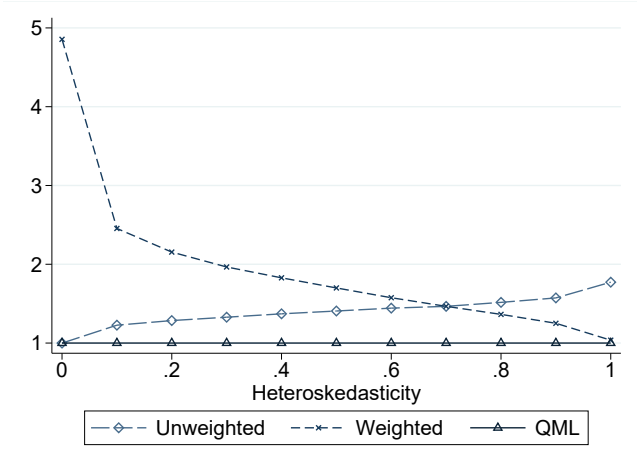
Notes: Each graph shows the root mean squared error of estimators relative to QML for $s = 1$. Each point includes 100,000 simulations of an estimate. The error term is $\epsilon_t = k^{\frac{1}{2}} t^{\frac{s}{2}} \eta_t + \nu_t$, with $k = \frac{\frac{H_{2s}(T)}{H_s(T)^2} - \frac{1}{T}}{\frac{H_{-s}(T)}{T^2} - H_s(T)^{-1}} e^{\Phi^{-1}(h)}$, where h is heteroskedasticity and $\Phi(.)$ is the CDF of a normal distribution. For each simulation, there are 1,000 observations where $\eta_t \sim \mathbb{N}(0, 1)$ and $\nu_t \sim \text{Exp}(1) - 1$.

Figure 9: Relative root mean squared error in simulations estimating a regression coefficient, $s = 1$



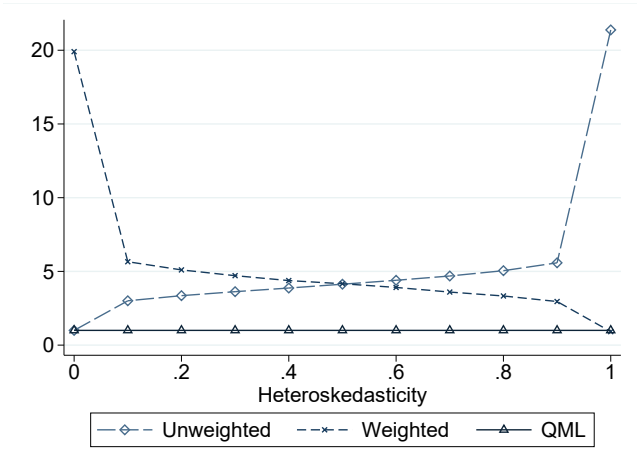
Notes: Each graph shows the root mean squared error of estimators relative to QML for $s = 1$. Each point includes 100,000 simulations of an estimate. The error term is $\epsilon_t = k^{\frac{1}{2}} t^{\frac{s}{2}} \eta_t + \nu_t$, with $k = \frac{\frac{H_{2s}(T)}{H_s(T)^2} - \frac{1}{T}}{\frac{H_{-s}(T)}{T^2} - H_s(T)^{-1}} e^{\Phi^{-1}(h)}$, where h is heteroskedasticity and $\Phi(.)$ is the CDF of a normal distribution. For each simulation, there are 1,000 observations where $\eta_t \sim \mathbb{N}(0, 1)$ and $\nu_t \sim \text{Exp}(1) - 1$.

Figure 10: Relative root mean squared error in simulations estimating an instrumental variables coefficient, $s = 1$



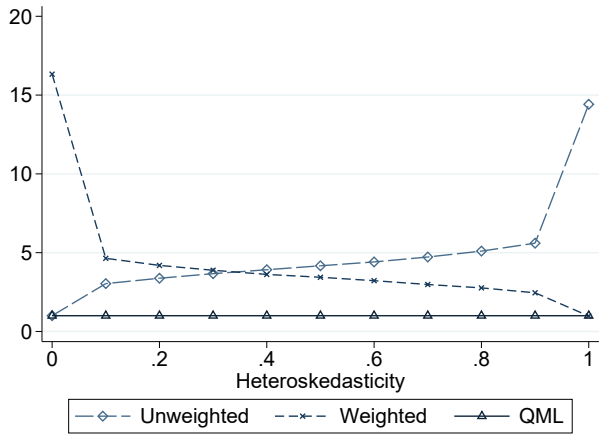
Notes: Each graph shows the root mean squared error of estimators relative to QML for $s = 1$. Each point includes 100,000 simulations of an estimate. The error term is $\epsilon_t = k^{\frac{1}{2}} t^{\frac{s}{2}} \eta_t + \nu_t$, with $k = \frac{\frac{H_{2s}(T)}{H_s(T)^2} - \frac{1}{T}}{\frac{H_{-s}(T)}{T^2} - H_s(T)^{-1}} e^{\Phi^{-1}(h)}$, where h is heteroskedasticity and $\Phi(.)$ is the CDF of a normal distribution. For each simulation, there are 1,000 observations where $\eta_t \sim \mathbb{N}(0, 1)$ and $\nu_t \sim \text{Exp}(1) - 1$.

Figure 11: Relative root mean squared error in simulations estimating an average, $s = 2$



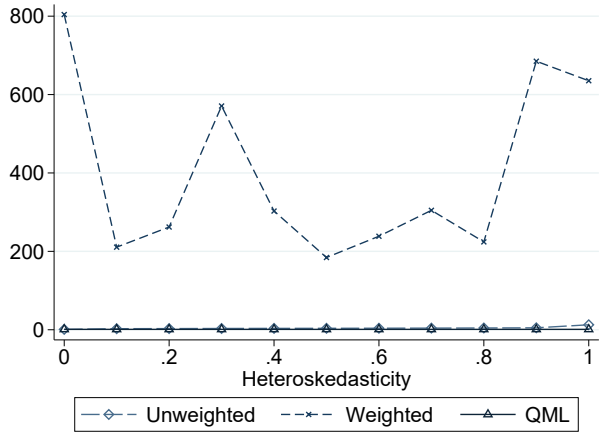
Notes: Each graph shows the root mean squared error of estimators relative to QML for $s = 2$. Each point includes 100,000 simulations of an estimate. The error term is $\epsilon_t = k^{\frac{1}{2}} t^{\frac{s}{2}} \eta_t + \nu_t$, with $k = \frac{\frac{H_{2s}(T)}{H_s(T)^2} - \frac{1}{T}}{\frac{H_{-s}(T)}{T^2} - H_s(T)^{-1}} e^{\Phi^{-1}(h)}$, where h is heteroskedasticity and $\Phi(.)$ is the CDF of a normal distribution. For each simulation, there are 1,000 observations where $\eta_t \sim \mathbb{N}(0, 1)$ and $\nu_t \sim \text{Exp}(1) - 1$.

Figure 12: Relative root mean squared error in simulations estimating a regression coefficient, $s = 2$



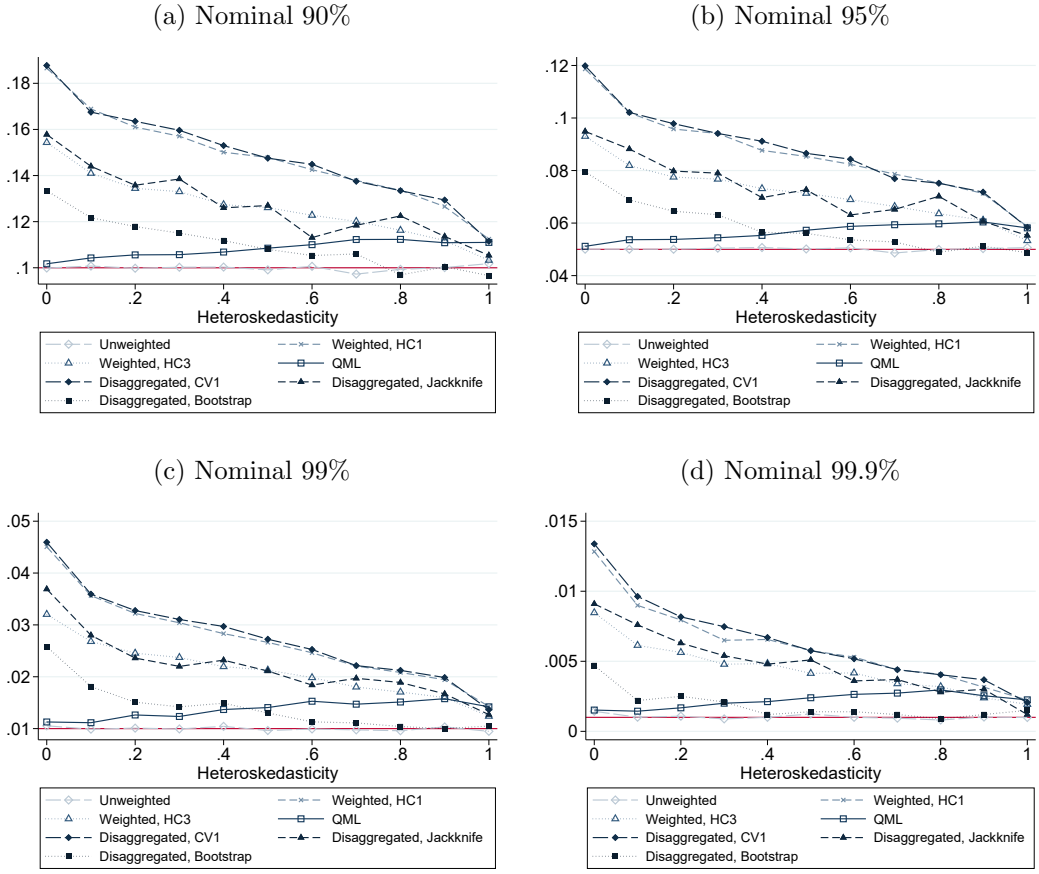
Notes: Each graph shows the root mean squared error of estimators relative to QML for $s = 2$. Each point includes 100,000 simulations of an estimate. The error term is $\epsilon_t = k^{\frac{1}{2}} t^{\frac{s}{2}} \eta_t + \nu_t$, with $k = \frac{\frac{H_{2s}(T)}{H_s(T)^2} - \frac{1}{T}}{\frac{H_{-s}(T)}{T^2} - H_s(T)^{-1}} e^{\Phi^{-1}(h)}$, where h is heteroskedasticity and $\Phi(.)$ is the CDF of a normal distribution. For each simulation, there are 1,000 observations where $\eta_t \sim \mathbb{N}(0, 1)$ and $\nu_t \sim \text{Exp}(1) - 1$.

Figure 13: Relative root mean squared error in simulations estimating an instrumental variables coefficient, $s = 2$



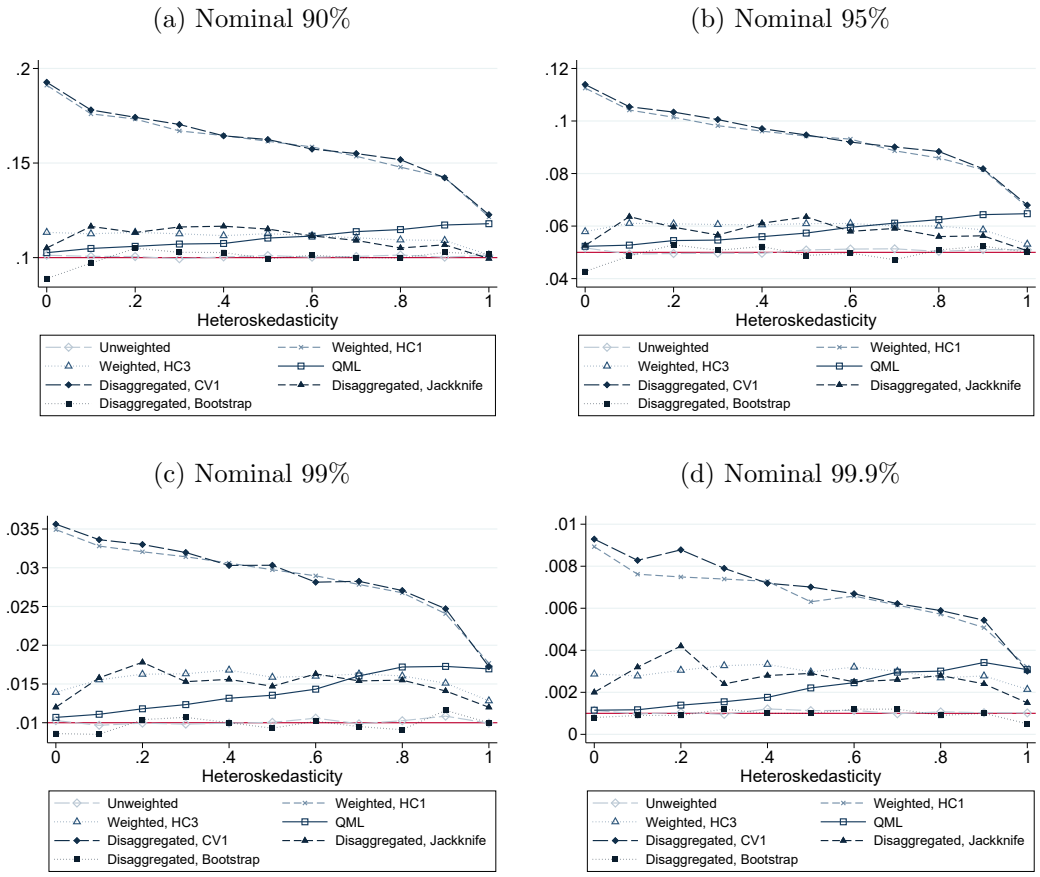
Notes: Each graph shows the root mean squared error of estimators relative to QML for $s = 2$. Each point includes 100,000 simulations of an estimate. The error term is $\epsilon_t = k^{\frac{1}{2}} t^{\frac{s}{2}} \eta_t + \nu_t$, with $k = \frac{\frac{H_{2s}(T)}{H_s(T)^2} - \frac{1}{T}}{\frac{H_{-s}(T)}{T^2} - H_s(T)^{-1}} e^{\Phi^{-1}(h)}$, where h is heteroskedasticity and $\Phi(.)$ is the CDF of a normal distribution. For each simulation, there are 1,000 observations where $\eta_t \sim \mathbb{N}(0, 1)$ and $\nu_t \sim \text{Exp}(1) - 1$.

Figure 14: Size of confidence intervals in simulations estimating an average, $s = 1$



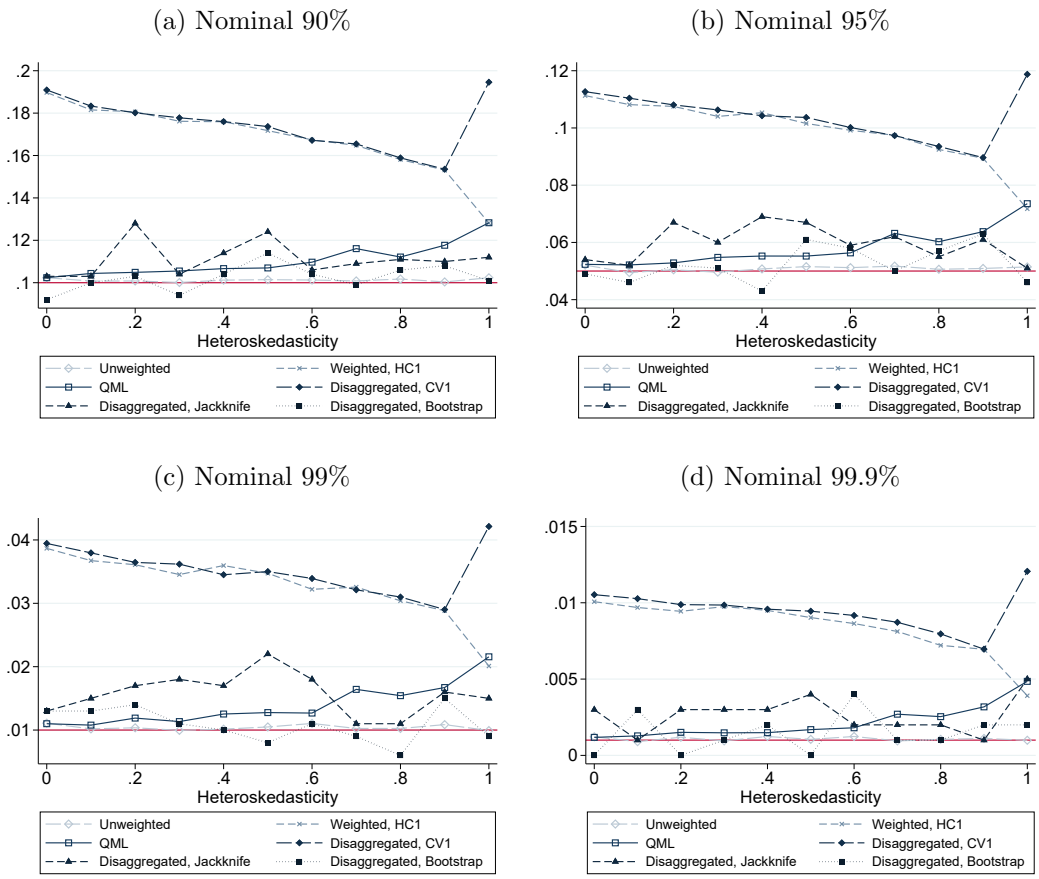
Notes: Each graph shows the size of nominally 95% confidence intervals for $s = 1$, using different estimation techniques. Each point includes 10,000 simulations for jackknife and bootstrap estimators and 100,000 simulations for all other estimators. The error term is $\epsilon_t = k^{\frac{1}{2}} t^{\frac{s}{2}} \eta_t + \nu_t$, with $k = \frac{\frac{H_{2s}(T)}{H_s(T)^2} - \frac{1}{T}}{\frac{H_{-s}(T)}{T^2} - H_s(T)^{-1}} e^{\Phi^{-1}(h)}$, where h is heteroskedasticity and $\Phi(.)$ is the CDF of a normal distribution. For each simulation, there are 1,000 observations where $\eta_t \sim \mathbb{N}(0,1)$ and $\nu_t \sim \text{Exp}(1) - 1$. However, "disaggregated" techniques expand each of the 1,000 initial observations so the size of the group formed by each initial observation t is approximately equal to $4000 \times t^{-s}$.

Figure 15: Size of confidence intervals in simulations estimating a regression coefficient, $s = 1$



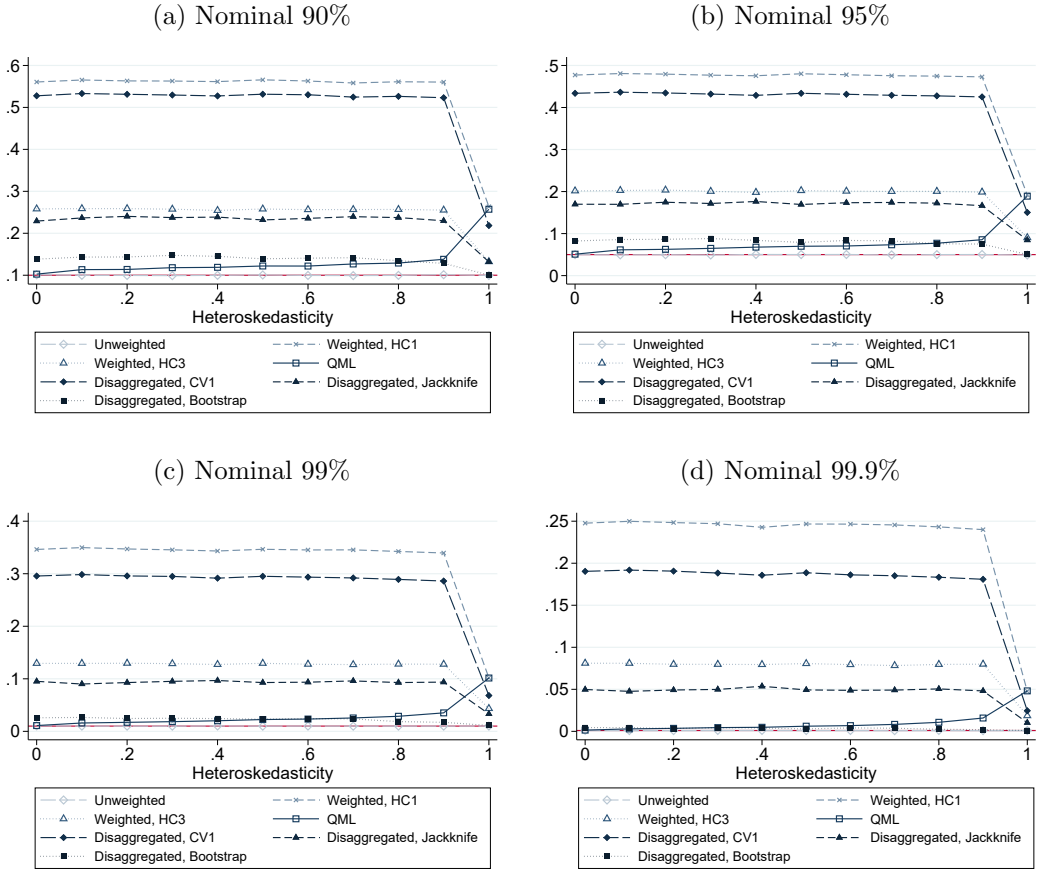
Notes: Each graph shows the size of nominally 95% confidence intervals for $s = 1$, using different estimation techniques. Each point includes 10,000 simulations for jackknife and bootstrap estimators and 100,000 simulations for all other estimators. The error term is $\epsilon_t = k^{\frac{1}{2}} t^{\frac{s}{2}} \eta_t + \nu_t$, with $k = \frac{\frac{H_{2s}(T)}{H_s(T)^2} - \frac{1}{T}}{\frac{H_{-s}(T)}{T^2} - H_s(T)^{-1}} e^{\Phi^{-1}(h)}$, where h is heteroskedasticity and $\Phi(.)$ is the CDF of a normal distribution. For each simulation, there are 1,000 observations where $\eta_t \sim \mathbb{N}(0,1)$ and $\nu_t \sim \text{Exp}(1) - 1$. However, "disaggregated" techniques expand each of the 1,000 initial observations so the size of the group formed by each initial observation t is approximately equal to $4000 \times t^{-s}$.

Figure 16: Size of confidence intervals in simulations estimating an instrumental variables coefficient, $s = 1$



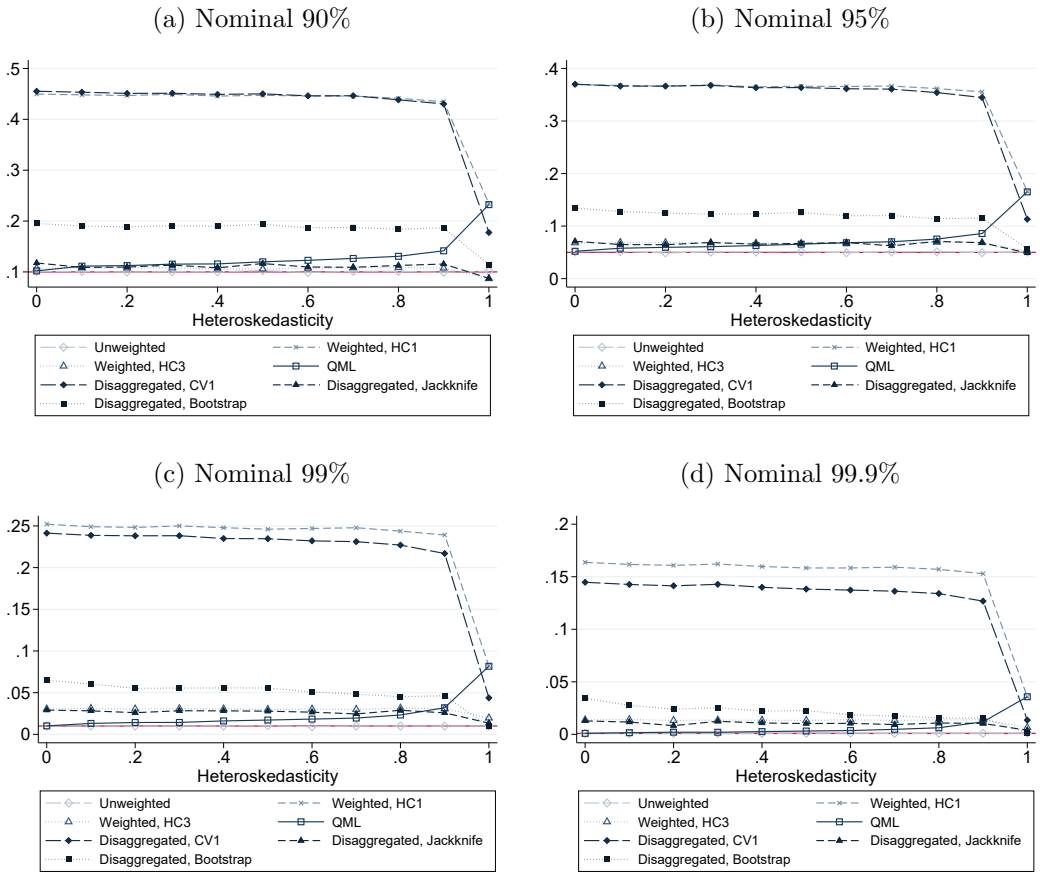
Notes: Each graph shows the size of nominally 95% confidence intervals for $s = 1$, using different estimation techniques. Each point includes 1,000 simulations for jackknife and bootstrap estimators and 100,000 simulations for all other estimators. The error term is $\epsilon_t = k^{\frac{1}{2}} t^{\frac{s}{2}} \eta_t + \nu_t$, with $k = \frac{\frac{H_{2s}(T)}{H_s(T)^2} - \frac{1}{T}}{\frac{H_{-s}(T)}{T^2} - H_s(T)^{-1}} e^{\Phi^{-1}(h)}$, where h is heteroskedasticity and $\Phi(.)$ is the CDF of a normal distribution. For each simulation, there are 1,000 observations where $\eta_t \sim \mathbb{N}(0,1)$ and $\nu_t \sim \text{Exp}(1) - 1$. However, "disaggregated" techniques expand each of the 1,000 initial observations so the size of the group formed by each initial observation t is approximately equal to $4000 \times t^{-s}$.

Figure 17: Size of confidence intervals in simulations estimating an average, $s = 2$



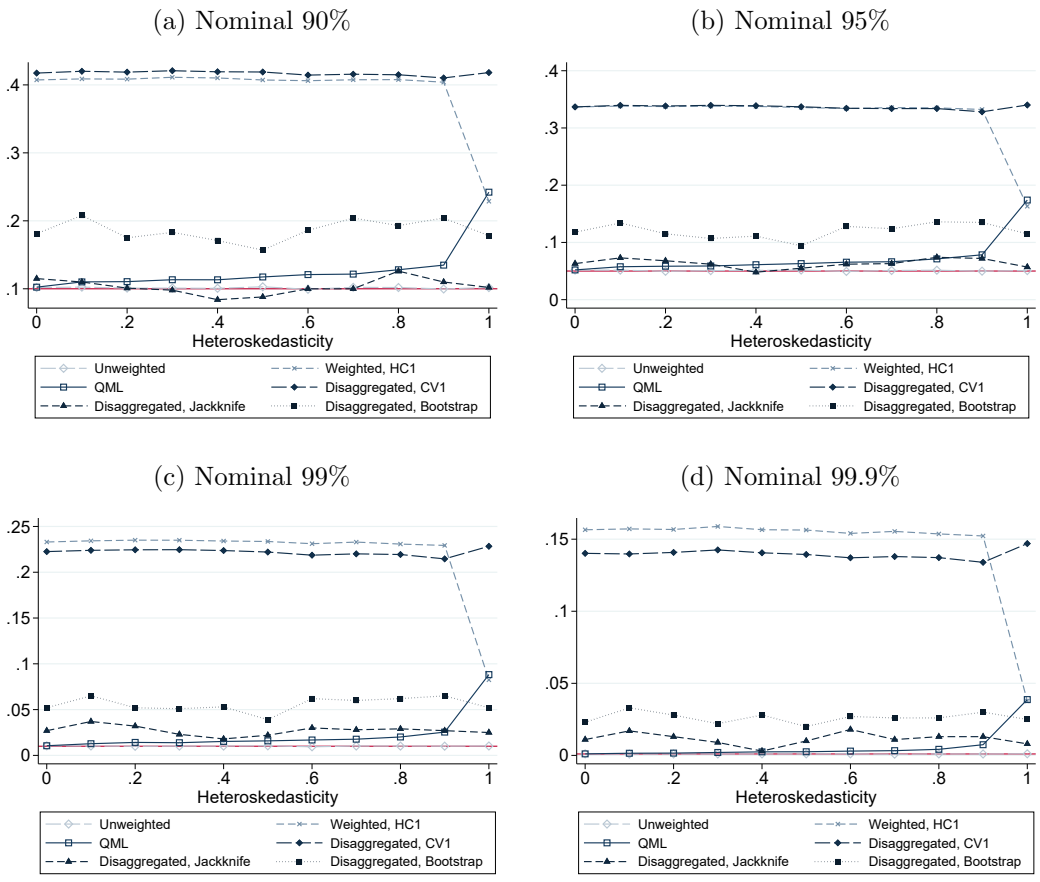
Notes: Each graph shows the size of nominally 95% confidence intervals for $s = 2$, using different estimation techniques. Each point includes 10,000 simulations for jackknife and bootstrap estimators and 100,000 simulations for all other estimators. The error term is $\epsilon_t = k^{\frac{1}{2}} t^{\frac{s}{2}} \eta_t + \nu_t$, with $k = \frac{\frac{H_{2s}(T)}{H_s(T)^2} - \frac{1}{T}}{\frac{H_{-s}(T)}{T^2} - H_s(T)^{-1}} e^{\Phi^{-1}(h)}$, where h is heteroskedasticity and $\Phi(.)$ is the CDF of a normal distribution. For each simulation, there are 1,000 observations where $\eta_t \sim \mathbb{N}(0,1)$ and $\nu_t \sim \text{Exp}(1) - 1$. However, "disaggregated" techniques expand each of the 1,000 initial observations so the size of the group formed by each initial observation t is approximately equal to $4000 \times t^{-s}$.

Figure 18: Size of confidence intervals in simulations estimating a regression coefficient, $s = 2$



Notes: Each graph shows the size of nominally 95% confidence intervals for $s = 2$, using different estimation techniques. Each point includes 10,000 simulations for jackknife and bootstrap estimators and 100,000 simulations for all other estimators. The error term is $\epsilon_t = k^{\frac{1}{2}} t^{\frac{s}{2}} \eta_t + \nu_t$, with $k = \frac{\frac{H_{2s}(T)}{H_s(T)^2} - \frac{1}{T}}{\frac{H_{-s}(T)}{T^2} - H_s(T)^{-1}} e^{\Phi^{-1}(h)}$, where h is heteroskedasticity and $\Phi(.)$ is the CDF of a normal distribution. For each simulation, there are 1,000 observations where $\eta_t \sim \mathbb{N}(0,1)$ and $\nu_t \sim \text{Exp}(1) - 1$. However, "disaggregated" techniques expand each of the 1,000 initial observations so the size of the group formed by each initial observation t is approximately equal to $4000 \times t^{-s}$.

Figure 19: Size of confidence intervals in simulations estimating an instrumental variables coefficient, $s = 2$



Notes: Each graph shows the size of nominally 95% confidence intervals for $s = 2$, using different estimation techniques. Each point includes 1,000 simulations for jackknife and bootstrap estimators and 100,000 simulations for all other estimators. The error term is $\epsilon_t = k^{\frac{1}{2}} t^{\frac{s}{2}} \eta_t + \nu_t$, with $k = \frac{\frac{H_{2s}(T)}{H_s(T)^2} - \frac{1}{T}}{\frac{H_{-s}(T)}{T^2} - H_s(T)^{-1}} e^{\Phi^{-1}(h)}$, where h is heteroskedasticity and $\Phi(.)$ is the CDF of a normal distribution. For each simulation, there are 1,000 observations where $\eta_t \sim \mathbb{N}(0,1)$ and $\nu_t \sim \text{Exp}(1) - 1$. However, "disaggregated" techniques expand each of the 1,000 initial observations so the size of the group formed by each initial observation t is approximately equal to $4000 \times t^{-s}$.