Diseconomies of Scale in Active Management: Robust Evidence

Ľuboš Pástor

Robert F. Stambaugh

Lucian A. Taylor

Min Zhu*

July 13, 2021

Abstract

We take a deeper look at the robustness of evidence presented by Pástor, Stambaugh, and Taylor (2015) and Zhu (2018), who find that an actively managed mutual fund's returns relate negatively to both fund size and the size of the active mutual fund industry. When we apply robust regression methods, we confirm both studies' inferences about scale diseconomies at the fund and industry levels. Moreover, data errors play no role, as both studies' results are insensitive to applying various error screens and using alternative return benchmarks. We reject constant returns to scale even after dropping 25% of the most extreme return observations. Finally, we caution that asymmetric removal of influential observations delivers biased conclusions about diseconomies of scale.

JEL classifications: G11, G23, J24
Keywords: returns to scale, active management, mutual funds

*Pástor is at the University of Chicago Booth School of Business. Stambaugh and Taylor are at the Wharton School of the University of Pennsylvania. Zhu is at the Business School, University of Queensland. Pástor and Stambaugh are also at the NBER. Pástor is additionally at the National Bank of Slovakia and the CEPR. The views in this paper are the responsibility of the authors, not the institutions they are affiliated with. This research was funded in part by the Fama-Miller Center for Research in Finance at the University of Chicago Booth School of Business.

1. Introduction

Do active investment managers face diseconomies of scale? This question is central to understanding the characteristics and performance of individual active managers as well as the active management industry. The existence of scale diseconomies is supported by empirical evidence in a variety of forms. For example, Pástor, Stambaugh, and Taylor (2020) document numerous relations among the characteristics of active mutual funds that are consistent with fund managers facing decreasing returns to scale (DRS). As that study explains, the characteristic-based evidence of DRS is especially strong and avoids a potential challenge to finding evidence of DRS by examining fund returns. Specifically, the model of Berk and Green (2004) predicts that, in equilibrium, money flows in and out of funds each period such that investors always expect each fund to have zero net return in excess of a passive benchmark. Those fund flows thus reduce variation in fund returns, thereby limiting the ability to detect scale effects in returns.

Despite that challenge, fund returns do provide evidence of DRS—evidence we find to be robust. We focus on the studies by Pástor, Stambaugh, and Taylor (2015) and Zhu (2018). The first of these studies, hereafter denoted as PST, finds industry-level DRS, reporting a significant negative relation between a fund's return and the size of the active mutual fund industry. That study also estimates a negative relation between a fund's return and the fund's size, consistent with fund-level DRS, but the relation is not statistically significant. As those authors explain, however, estimating the fund-level relation risks a finite-sample bias. They employ a recursive-demeaning (RD) procedure that avoids the bias but has less power. Zhu (2018) refines the RD procedure, adding an intercept in its first stage and using a slightly different instrumental variable. With that refinement, Zhu finds a significant negative relation between fund return and fund size, especially when fund size is represented by its logarithm. Together, these two studies provide significant evidence of DRS at both the fund and industry levels. We examine the robustness of their inferences and find them to be supported consistently.

Our focus on the above two studies is motivated by the recent critique from Adams, Hayunga, and Mansi (2021), hereafter AHM. They argue that data errors produce a small number of influential observations that account for the significant industry-level DRS found by PST as well as the significant fund-level DRS found by Zhu (2018). We thus begin our analysis by performing additional data-cleaning steps to eliminate the various errors that AHM allege. We find that results from the PST and Zhu procedures are unaffected, even when we apply a full battery of data-cleaning steps that jointly remove 3% of the original observations. We also find that the PST and Zhu inferences are unaffected by using alternative benchmarks based on the three-factor model of Fama and French (1993), as opposed to Morningstar benchmarks. This latter evidence contradicts the claim in AHM's abstract that a "major source of these errors is the incorrect use of Morningstar's current performance benchmarks to measure historical return performance." Overall, in contrast to AHM's allegations, we show that data errors play no role in the conclusions of PST or Zhu.

As for results being driven by a small set of influential observations, we turn to standard robust-regression methods that mitigate such effects. We apply both M-estimation and MM-estimation to a variety of regression specifications, including the RD procedures of both PST and Zhu. The results consistently confirm the original findings of significant industry-level and fund-level DRS.

The critique of AHM essentially rests on removing influential observations. There are two basic problems with their analysis. First, instead of using data errors to identify influential observations, AHM identify influential observations and then inspect them for data errors. In actuality, data errors are irrelevant to the PST and Zhu results, as noted above. Second, AHM identify influential observations asymmetrically, treating large positive residuals differently from large negative residuals. We show that eliminating observations in such a manner produces results very different from eliminating observations symmetrically. AHM's asymmetric approach has less ex ante appeal than the symmetric approach, and both outlier-removal approaches are unappealing compared to robust regression. Moreover, we explain why asymmetric outlier removal imparts a positive bias to the estimated relation between a fund's benchmark-adjusted return and the size of the fund or the fund industry. We link the bias to heteroskedasticity that is implied by various equilibrium models and also supported by the data. Specifically, the volatility of a fund's return is negatively related to the size of both the fund and the industry. This bias explains AHM's puzzling finding that the estimated direction of returns to scale flips as they remove progressively more outliers.

The ultimate message advanced by AHM is that one cannot reject the null hypothesis of constant returns to scale, at both the fund and industry levels, against a DRS alternative. As we explain later, that null hypothesis can be tested with an OLS regression of fund returns on industry size, including fund fixed effects. This regression produces a significantly negative coefficient on industry size, thereby rejecting the null and thus contradicting AHM's conclusion. Moreover, it continues to do so even after removing up to 25% of the sample by eliminating the observations with the most extreme returns.

The remainder of our study is organized as follows. Section 2 establishes the irrelevance of data errors. Section 3 confirms the original PST and Zhu inferences by using robust regression methods. Section 4 illustrates the pitfalls of removing influential observations asymmetrically. Section 5 demonstrates the robustness of rejecting the null of constant returns to scale. Section 6 concludes.

2. Data errors play no role

The principal message advanced by AHM is that data errors largely account for PST's and Zhu's inferences supporting DRS at both the industry and fund levels. We disagree.

Consider the approach AHM follow in drawing their conclusion. After identifying 168 observations deemed influential for PST's inference about industry-level DRS, AHM "next investigate the origins of these influential observations and the reasons for their outlyingness" (subsection 3.1.2.1). In fact, however, AHM just investigate whether those 168 observations contain errors, as opposed to showing that errors thereby discovered caused the observations to be identified as influential. This is problematic because the "errors" among the 168 observations could be smaller than the errors in the rest of the sample.

To illustrate this point, consider a hypothetical example in which 100% of all sample observations contain an error component (e.g., deviations of the Morningstar-assigned benchmarks from some unobservable ideal benchmarks), but these errors are small, random, and inconsequential. Following AHM's procedure would reveal, correctly, that 100% of the influential observations contain errors. However, that revelation would be meaningless because 100% of the non-influential observations also contain small random errors, and because all of these errors are irrelevant anyway.

2.1. Alleged Data Errors

Moreover, the observations that AHM allege to be erroneous often are not. Fully half (84) of the 168 influential observations pertain to the Growth Fund of America, the largest actively managed U.S. fund. AHM argue that Morningstar is mistaken in classifying this fund as U.S. large growth and in assigning the Russell 1000 Growth index as its benchmark. This is a provocative allegation—AHM are effectively accusing Morningstar, the largest provider of mutual fund ratings, of misclassifying the largest U.S. active fund, which manages over $270 billion as of this writing in mid-2021. Morningstar continues classifying the Growth Fund of

America in the same way, with the same benchmark, to this day.[1] The fund also continues to present itself as large-cap growth and compare its performance to Russell 1000 Growth.[2] AHM even "question whether the fund should be included in the sample as the construction process of PST explicitly calls for the removal of non-domestic and non-equity funds." Yet as of this writing, the fund's asset mix is 86% U.S. equities, 11% non-U.S equities, and 3% cash. While these numbers vary somewhat over time, we are perfectly happy having this fund in our sample. We disagree with AHM's judgment that the largest active U.S. fund is a data error (or, more precisely, 84 data errors).

AHM also note that a substantial number of the 168 observations correspond to bear-market funds, or long-short funds, to which Morningstar assigns the S&P 500 index as the benchmark. AHM do not explain why such funds should be any less subject to DRS than are other active funds. Nor do AHM explain why such funds should not be benchmarked against a buy-and-hold market index. AHM instead simply declare that the inclusion of such funds in the sample is erroneous. We think it is fine to keep these funds.

To be fair, among the 168 observations, AHM do identify some funds that PST intended to exclude, such as some international funds that escaped PST's keyword search of the most recent Morningstar (MS) category. Of course, such funds account for other observations as well, not just a few of the 168. Also, PST use the benchmark for a fund's most recently assigned MS category rather than the historical MS category assigned at each point in time. Therefore, any difference between the current and historical category can be viewed as a data error. Again, though, such errors potentially afflict many observations, not just the observations that AHM identify as influential. Moreover, such errors are likely to be random with respect to the assessment of DRS. Finally, such benchmark errors matter only for funds whose styles change substantially over time. AHM do not address the magnitudes of these errors, nor do they show that any identification of influential observations is caused by such errors. Identifying influential observations and then inspecting them for data errors does not really address the question of whether data errors affect results. The latter question is worthy of consideration, however, and we address it head on.

2.2. Additional Data Cleaning

To investigate whether data errors affect PST's and Zhu's inferences about DRS, we check whether the inferences change after applying three additional data-cleaning steps to the

[1] See `https://www.morningstar.com/funds/xnas/agthx/portfolio`.

[2] See `https://www.capitalgroup.com/individual/investments/fund/agthx`.

original PST sample. First, we obtain data on Morningstar's historical category designations, and we use them in place of the most recent MS category. This change affects the dependent variable (*GrossR*, fund return minus the return on the benchmark for the fund's MS category) as well as clustering (by month × MS category). The resulting dataset is slightly smaller than the original PST dataset because returns for the historical benchmarks are sometimes missing. Second, we additionally drop observations from seven MS categories that do not apply to active U.S. equity funds: "Convertibles," "High Yield Bond," "Multisector Bond," "Target-Date 2000-2010," "Target-Date 2011-2015," "World Allocation," and "World Stock." The analysis in PST screens out those MS categories using each fund's most recent MS category, but using historical MS categories screens out additional observations. Third, we also drop funds whose names contain at least one of 45 keyword strings, such as " muni ," "foreign," "oversea," "austral," "lifecycle," "property," and "gold."[3] The MS category variable already helps exclude bond funds, international funds, target-date funds, real estate funds, commodity funds, and others, but this screen based on keywords in fund names excludes additional observations. This third sample, arguably the cleanest and most conservative, eliminates about 3% of the observations in the original sample.

To investigate further the effects of benchmarks, we use the Fama and French (1993) three-factor model to construct a benchmark for each fund, replacing the fund's MS benchmark. Specifically, we compute Fama-French-adjusted return as the fund's excess gross return minus the Fama-French factors times their respective betas. We compute each fund's Fama-French betas by estimating a time series regression of the fund's excess gross return on the three Fama-French factors, using all available data for the fund.[4] We apply the same three cleaning steps to this sample as well.

2.3. Evidence

For each of the above datasets, Table 1 displays results for the two key regressions estimated originally by PST and reported in columns 5 and 9 of that study's Table 3.[5] The first four columns estimate the OLS regression of *GrossR* on just industry size, with fund fixed

[3] The full list of strings includes: treasury, govt, government, barclay, municipal, muni , convertible, investment-grade, fixed-income, fxd-inc, mortgage, consumer price index, t-bill, acwi, world, global, developing market, emerging market, latin america, international, oversea, intl , foreign, canad, asia, japan, brazil, mexico, russia, euro , austral, china, india, europe, eafe, msci em, target date, trgt date, lifecycle, property, real estate, reit, commodity, gold.

[4] In the rare cases when a fund has fewer than 24 monthly observations, we set the fund's betas equal to their average values across funds in the same sector. We compute those average betas from sector-level panel regressions of excess gross fund returns on the Fama-French factors.

[5] In their Table 4, AHM also analyze the robustness of a fixed-effect regression when fund size is included, but that regression suffers from the finite-sample bias pointed out by PST.

effects included. Column (a) gives results without the three additional data-cleaning steps described above, and columns (b) through (d) report results with those steps successively applied. Panel A reports results using MS benchmarks, while Panel B uses Fama-French benchmarks. The last four columns correspond to the first four except that they use two independent variables, industry size and fund size, along with PST's RD procedure.

Table 1 dispels AHM's data-error narrative for PST. The data-cleaning steps that cumulatively eliminate 3% of the sample have virtually no effect on any of the results; the differences between columns (a) and (d) are negligible. The differences between Panels A and B are somewhat larger, indicating that using alternative return benchmarks does have some effect, not surprisingly, but inferences are unaffected. The coefficient on industry size is consistently negative and significant, confirming the original PST results. The coefficient on fund size is consistently negative but insignificant, again as in PST.

Table 2 presents corresponding results for the two key regressions estimated by Zhu (2018) and reported in the fourth column of Panels A and B in Table 3 of her study. Otherwise the organization of Table 2 parallels that of Table 1. The first four columns estimate the regression of $GrossR$ on fund size, while the last four columns estimate the regression on the log of fund size. Zhu's RD procedure is used to estimate both regressions.

Table 2 reveals the irrelevance of data errors to Zhu's results as well. The differences between columns (a) and (d) are small. Again, using alternative benchmarks has a somewhat larger effect, but inferences are unaffected. The coefficients in Table 2 are consistently negative and significant, on both fund size and log fund size, with the latter producing a stronger rejection of constant returns to scale. Zhu (2018) obtains the same inferences.

3. Robust regressions confirm our results

While OLS has long been the workhorse of econometrics, OLS estimates are well known to be sensitive to outliers. Robust regression analysis provides an alternative to OLS that downweights outliers in an effort to obtain more robust estimates. Although robust regression is not widely used in the social sciences, statisticians have been using it for decades.

The reason why OLS estimates are sensitive to outliers is that the quadratic loss function inherent to OLS increases sharply with the magnitude of the residuals. To moderate this sensitivity, robust regression techniques use different loss functions that are less responsive to outlying residuals. For example, the M-estimator of Huber (1964) minimizes a loss function that is quadratic for small residuals but linear for large residuals. For another example, the MM-estimator of Yohai (1987) is typically used with Tukey's bisquare weighting function, which downweights all nonzero residuals relative to OLS and completely disregards large residuals.[6] Both M-estimation and MM-estimation effectively downweight observations with large residuals relative to OLS. Both are also among the most popular methods of robust estimation nowadays (Maronna, Martin and Yohai, 2006).

In this section, we demonstrate the robustness of the results of PST and Zhu (2018) by applying M-estimation as well as MM-estimation. Following PST and Zhu (2018), we consider the regressions of $GrossR$, a benchmark-adjusted gross fund return computed using historical MS benchmarks, on various combinations of industry size, fund size, and log fund size. We use the cleanest version of our sample (i.e., the sample used in columns (d) of Tables 1 and 2) for the period March 1993 through December 2011, which matches PST's main sample. For the fixed-effect estimation, we first apply a within-fund demeaning of variables and then run robust regression on those demeaned variables. For the RD estimation, we run OLS in the first-stage regression to extract predicted fund size and then run robust regression at the second stage, in which the predicted fund size is used.[7] To properly compute heteroskedasticity-robust standard errors, we use residuals weighted by the robust loss function rather than the usual residuals calculated as the difference between the response and the predicted response (Heritier et al., 2009).

Table 3 presents the results from both M-estimation (Panel A) and MM-estimation (Panel B). Both panels tell the same story. In the fixed-effect robust regression, the slope on industry size is negative and highly significant, with the t-statistic of about -5. This slope is negative and significant also when fund size is included the regression, regardless of whether we use PST's RD or Zhu's RD and whether or not we take the log of fund size. The slope on fund size is also negative and significant, especially when we take its log and when we use Zhu's RD. In the latter case, the t-statistics on log fund size are between -5 and -6, whether or not industry size is included. Overall, robust regressions provide strong evidence of DRS at

[6] For both loss functions, large residuals are defined as having scaled absolute values in excess of k, where k is the so-called tuning constant. Smaller values of k produce more resistance to outliers, but at the expense of lower efficiency when the true residual distribution is normal. For the Huber loss function, the standard value of k is 1.345, whereas for Tukey's function, the standard value is $k = 4.685$. In both cases, these values of k provide 95% asymptotic efficiency relative to OLS if the residuals are normal (e.g., Maronna, Martin and Yohai, 2006). We use these standard values of k to obtain our estimates in Table 3.

[7] We run OLS at the first stage of RD because doing so produces a better instrument than running robust regression. Recall that the sole purpose of the first-stage regression is to obtain a variable (predicted fund size) that is highly correlated with the original fund size but uncorrelated with the return innovation. We find that applying robust analysis to the first-stage regression reduces the correlation between predicted fund size and the actual fund size, which compromises the effectiveness of the first-stage regression. Therefore, our robust version of the RD estimator involves robust regression only at the second stage.

both the fund level and the industry level.

4. Pitfalls in removing influential observations

Instead of applying robust regression to the full sample, AHM shrink the sample by removing some influential observations and reestimate the model in the resulting subsample. To identify influential observations, AHM use the multivariate outlier detection approach of Rousseeuw and van Zomeren (1990). While identifying outliers in this manner makes sense, removing them seems problematic because outliers can contain valuable information. In general, it is bad practice to remove legitimate datapoints so as to produce a better-fitting model. If the outlier is a natural element of the population being studied, dropping it from the data set leads to biased inference. Indeed, Rousseeuw and van Zomeren (1990) themselves caution against removing outliers, writing on page 637 that "We would like to stress that we are *not* advocating that one simply remove the outliers." (The emphasis on "not" is added by Rousseeuw and van Zomeren.) Despite these warnings, AHM use Rousseeuw and van Zomeren's method to not only identify but also remove outliers.

To motivate this removal, AHM assert at the beginning of their Section 2.3 that the outliers identified by their procedure are "errant." We are puzzled by this characterization. The fact that an observation is identified as influential ex post does not make it errant. This point is well understood by Rousseeuw and van Zomeren (1990), who argue on page 634 that "we do not necessarily want to *delete* the outliers; it is only our purpose to *find* them, after which the user may decide whether they are to be kept, deleted, or corrected, depending on the situation." (Again, the emphasis is added by Rousseeuw and van Zomeren.) While AHM subsequently examine these outliers for possible data errors, they do not use the data error criterion to exclude observations. Instead, AHM first use a statistical procedure to identify outliers, then they exclude those outliers, and finally they analyze whether any of the excluded outliers are data errors. We find it more reasonable to first identify data errors and then exclude them, as we note in Section 2.

4.1. Symmetric versus asymmetric outlier removal

Another problem with AHM's approach is that they remove outliers asymmetrically. They do mention symmetric removal in passing, but the results they emphasize and tabulate remove *positive* vertical outliers (i.e., observations with large positive vertical distances) while retaining *negative* outliers. Such an asymmetric treatment seems difficult to justify ex ante. Moreover, it biases the estimates, as we show below.

To analyze the effects of AHM's outlier-removal procedure, we first apply it to the main model estimated by PST: the regression of $GrossR$ on both industry size and fund size. We implement the procedure in the cleanest version of our dataset, as before, and use the same sample period of March 1993 through December 2011. We identify and remove outliers in the same way as AHM, eliminating up to 10% of our sample.

Table 4 is the counterpart of Panel C of AHM's Table 4, except that our sample is somewhat different. Panels A and B of Table 4 report the results based on asymmetric outlier removal. In Panel A, we remove only *positive* vertical outliers, mimicking AHM's approach. Our results are similar to those reported by AHM. When no observations are excluded, the coefficient on industry size is negative and significant, with a t-statistic of about -2.5. The coefficient remains negative and significant until 0.1% observations are excluded, but then it loses significance, and its sign eventually flips to positive and significant.

In Panel B of Table 4, we remove only *negative* vertical outliers. Unlike in Panel A, the coefficient on industry size is negative and significant throughout. Moreover, the coefficient becomes increasingly negative: by the time 10% observations are removed, its magnitude increases more than sixfold and its t-statistic drops below -10. Had AHM chosen to emphasize negative instead of positive outliers, they would have reached the opposite conclusion. The stark difference between Panels A and B shows that AHM's outlier removal procedure introduces large bias. We explain the source of this bias in Section 4.2.

In Panel C of Table 4, we remove influential observations symmetrically, eliminating both positive and negative outliers to the same degree. The coefficient on industry size then varies only modestly as a growing fraction of the sample is removed. The coefficient remains negative throughout, and it remains significant even when 5% of the observations are removed. Therefore, PST's evidence of industry-level DRS is robust if influential observations are removed symmetrically.

Next, we apply AHM's outlier-removal procedure to the main regression estimated by Zhu (2018): that of $GrossR$ on log fund size. We use the same sample as in Table 4. Table 5 reports the results in the same format as Table 4. The patterns in both tables are remarkably similar. When no observations are removed, the coefficient on log fund size is negative and significant ($t = -5.53$). When we gradually remove positive vertical outliers, the coefficient eventually turns positive and significant (Panel A). When we gradually remove negative vertical outliers, the coefficient grows even more negative and more significant, with a t-statistic dropping below -10 (Panel B). When we remove outliers symmetrically, the coefficient remains negative and highly significant throughout (Panel C).

Overall, the conclusions of PST about industry size, as well as those of Zhu (2018) about fund size, are robust to removing influential observations in a symmetric fashion. Compared to AHM's asymmetric approach, the symmetric approach is easier to justify ex ante. Moreover, AHM's approach produces biased estimates, as we explain next.

4.2. Bias induced by asymmetric outlier removal

The patterns reported in Tables 4 and 5 are remarkably similar. As we move from left to right, the slopes on both industry size (Table 4) and log fund size (Table 5) flip from negative to positive in Panel A and from negative to more negative in Panel B. These patterns are caused by the bias induced by an asymmetric removal of outliers.

To explain the source of the bias, we proceed in two steps. First, we show that funds tend to have more volatile benchmark-adjusted returns when they are smaller, and also when industry size is smaller. We also explain that this heteroskedasticity is implied by several equilibrium models. Second, we explain how this heteroskedasticity induces the patterns observed in Tables 4 and 5.

Starting with the first step, we regress $GrossR$ on both fund size and industry size, with fund fixed effects. We then regress squared residuals from that regression on both fund size and industry size. We find negative and highly significant slopes on both size variables. To the extent that squared residuals proxy for residual volatility, this result indicates negative relations between that volatility and both fund and industry sizes.

These negative relations make sense from the perspective of equilibrium models that include decreasing returns to scale. In a number of such models, the larger the fund, the less its optimal portfolio tends to deviate from the benchmark, and thus the lower is the volatility of the fund's benchmark-adjusted return. In the model of Berk and Green (2004), for example, once a fund invests the optimal dollar amount in active positions, any additional capital is indexed. Therefore, the more additional capital the fund receives, the lower is the volatility of the fund's benchmark-adjusted return. In general equilibrium, Stambaugh (2014) shows that a larger fund's dollar active positions are optimally less aggressive, and he derives a negative relation between a fund's size and the volatility of its benchmark-adjusted return. A similar result follows from the model of Pástor, Stambaugh, and Taylor (2020), which implies that larger funds hold portfolios that are more similar to the benchmark index. The same logic naturally extends from fund size to industry size, because when funds become bigger, so does their industry. As a result, benchmark-adjusted fund returns are less volatile when industry size is larger.

The second step is to explain why AHM's asymmetric outlier-removal procedure induces a bias, given the negative volatility-size relation. We illustrate this bias in Figure 1, which plots a hypothetical sample exhibiting two properties: (i) a negative relation between return and size and (ii) a negative relation between volatility and size. Consistent with the first condition, the fitted OLS relation based on all observations, plotted as the solid line, has a negative slope. We then remove influential observations using the procedure of AHM. We delete 5% of the observations with the largest positive residuals (vertical distance), and we delete 5% of the observations for which size has the largest Mahalanobis distance (thus omitting observations with the largest and smallest sizes). We plot the removed observations in red and the retained ones in green. The fitted OLS relation based on just the retained observations is plotted as the dashed line.

AHM's method of removing observations biases against finding DRS, as illustrated in Figure 1 by the dashed slope having a higher slope than the solid line. Because the observations at lower size are more volatile, the observations that AHM remove based on large, positive vertical distances tend to lie in the upper left quadrant. Removing those observations therefore increases the slope of the fitted line, potentially even flipping it from negative to positive, as illustrated in Figure 1. This effect can explain why the slopes in Panels A of Tables 4 and 5 eventually flip from negative to positive when enough positive outliers are removed. AHM's puzzling "findings of non-negative returns to scale," which they mention at the beginning of their conclusion, seem to be an artifact of their biased estimation approach.

The same logic also explains why the slopes in Panels B of Tables 4 and 5 become more negative as we exclude a growing number of observations. The removal of negative vertical outliers removes influential observations from the bottom left quadrant of Figure 1, thereby mechanically tilting the OLS line downward. When vertical outliers are removed symmetrically, there is no obvious bias, which is why the slopes in Panels C of Tables 4 and 5 change so little as we move from left to right across the table.

5. Robustness of DRS to dropping extreme returns

The main message that AHM advance with respect to scale diseconomies is that there are no decreasing returns to scale of any kind. (The title of AHM's study is *"Scale and Performance in Active Management are Not Negatively Related."*) We reject AHM's message even when a large fraction of observations, those with the most extreme benchmark-adjusted fund returns, are removed from the sample. Given the large variability in performance across funds and time, this approach to removing outliers seems natural, in addition to being simple and transparent.

The AHM null hypothesis of no DRS can be tested via an OLS panel regression of $GrossR$ on industry size, with fund fixed effects included. A significantly negative coefficient on industry size is then sufficient to reject the AHM null hypothesis. Under that null, neither fund size nor industry size enters, so fund size can be excluded from the regression without inducing an omitted-variable bias. Recursive demeaning and the associated loss of power are thereby avoided. As before, we use the cleanest version of our dataset and the sample period March 1993 through December 2011. We progressively remove up to a quarter of our sample observations, which represent the biggest outliers in terms of $GrossR$.

Table 6 reports results, with column 1 of this table matching column 4 of Table 1. In the following columns, we reestimate the regression after dropping the $(x/2)\%$ lowest and $(x/2)\%$ highest values of $GrossR$, for $x \in \{5\%, 10\%, 15\%, 20\%, 25\%\}$. Our main finding is that the coefficient on industry size is always negative and statistically significant. For example, the corresponding t-statistic is -2.67 when we exclude 15% of the most extreme observations, and it is -2.11 when we exclude a quarter of the sample. The negative relation between industry size and fund performance is remarkably robust to excluding return outliers. By finding a robustly negative slope coefficient on industry size, we reject AHM's main conclusion of no DRS.

6. Conclusions

We find robust evidence of DRS at both the fund and industry levels. Our evidence reaffirms the conclusions of PST and Zhu (2018) but contradicts those of AHM. AHM's conclusions are based on an asymmetric outlier-removal procedure that is biased and hard to justify ex ante. In contrast, our conclusions are based on a variety of evidence, including robust regression and two symmetric ways of removing outliers. We obtain that evidence after applying additional data-cleaning steps to a sample already carefully constructed by matching CRSP and Morningstar sources. Neither those data-cleaning steps nor the use of alternative benchmarks make a difference, dispelling AHM's data-error narrative for the PST and Zhu results.

Table 1
Robustness of PST to additional data cleaning

Column 1 (5) of Panel A repeats the analysis in column 5 (9) in Panel A of Table 3 of PST. The remaining columns show how those results change after we perform additional data-cleaning steps. The table header indicates the data-cleaning steps used to create each sample: (a) original data-cleaning steps from PST; (b) use historical MS categories; (c) also drop seven additional MS categories; (d) also drop funds with flagged names. Columns (1)–(4) contain OLS FE estimates, and columns (5)–(8) contain IV estimates using the PST RD procedure. Panel B is the same as Panel A but uses Fama-French adjusted returns in place of Morningstar-adjusted returns as the dependent variable. We estimate funds' Fama-French loadings in fund-by-fund full-sample time-series regressions; if the fund has fewer than 24 monthly observations, then we set its factor loadings to their sector averages. In Panel B, the only difference between columns 1 and 2 is that we cluster by month × {most-recent MS category} in column 1, and by month × {historical MS category} in column 2. Both panels use data from March 1993 through December 2011. The remaining details are the same as in Table 3 in PST.

	Data-cleaning steps							
	(a)	(b)	(c)	(d)	(a)	(b)	(c)	(d)
				Panel A: Morningstar benchmarks				
IndustrySize	-0.0326	-0.0304	-0.0308	-0.0310	-0.0275	-0.0263	-0.0267	-0.0267
	(-3.60)	(-3.76)	(-3.79)	(-3.80)	(-2.12)	(-2.46)	(-2.48)	(-2.47)
FundSize					-0.430	-0.313	-0.316	-0.320
					(-1.27)	(-1.06)	(-1.07)	(-1.10)
Observations	283,046	282,833	281,563	274,729	270,625	270,415	269,231	262,622
% obs. dropped	0.00	0.08	0.52	2.94	0.00	0.08	0.52	2.96
				Panel B: Fama-French benchmarks				
IndustrySize	-0.0264	-0.0264	-0.0266	-0.0268	-0.0210	-0.0210	-0.0212	-0.0214
	(-3.02)	(-2.74)	(-2.74)	(-2.75)	(-2.09)	(-1.96)	(-1.97)	(-1.97)
FundSize					-0.482	-0.482	-0.485	-0.471
					(-1.64)	(-1.55)	(-1.56)	(-1.54)
Observations	283,046	283,046	281,776	274,942	270,625	270,625	269,441	262,832
% obs. dropped	0.00	0.00	0.45	2.86	0.00	0.00	0.44	3.05

Table 2
Robustness of Zhu (2018) to additional data cleaning

Column 1 (5) of Panel A repeats the estimation reported in the fourth column of Panel A (B) of Table 3 of Zhu (2018), but using the original PST dataset. The other columns show how those results change after we perform additional data-cleaning steps. The remaining details are the same as in Table 1.

	Data-cleaning steps							
	(a)	(b)	(c)	(d)	(a)	(b)	(c)	(d)
	Panel A: Morningstar benchmarks							
FundSize	-0.8471	-0.7791	-0.7789	-0.7626				
	(-2.20)	(-2.06)	(-2.06)	(-2.05)				
log(FundSize)					-0.0028	-0.0025	-0.0025	-0.0024
					(-6.31)	(-5.59)	(-5.54)	(-5.53)
Observations	271,819	271,607	270,416	263,785	271,819	271,607	270,416	263,785
% obs. dropped	0.00	0.08	0.52	2.96	0.00	0.08	0.52	2.96
	Panel B: Fama-French benchmarks							
FundSize	-0.6077	-0.6077	-0.6035	-0.5783				
	(-2.47)	(-2.42)	(-2.41)	(-2.39)				
log(FundSize)					-0.0015	-0.0015	-0.0015	-0.0014
					(-4.17)	(-4.06)	(-3.99)	(-3.88)
Observations	271,819	271,819	270,628	263,997	271,819	271,819	270,628	263,997
% obs. dropped	0.00	0.00	0.44	2.88	0.00	0.00	0.44	2.88

Table 3
Robustness to using robust-regression techniques

This table presents results from robust regressions of *GrossR*, a benchmark-adjusted gross fund return computed using historical Morningstar benchmarks, on various combinations of industry size, fund size, and log fund size. Panel A reports results from M-estimation with the Huber loss function. Panel B reports results from MM-estimation with Tukey's bisquare loss function. We use standard tuning constant values that deliver 95% relative efficiency for the normal distribution. We use the cleanest version of our sample (i.e., the sample used in columns (d) from Table 1) for the period March 1993 through December 2011. The *t*-statistics in Panels A and B are based on M-residuals or MM-residuals (Heritier et al., 2009), respectively, clustered by style×month (OLS FE) and further by fund (RD specifications).

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Panel A: M-estimation									
IndustrySize	-0.0219					-0.0135	-0.0139	-0.0127	-0.0136
	(-4.87)					(-2.41)	(-2.52)	(-2.28)	(-2.47)
FundSize		-0.2507	-0.3850			-0.3391	-0.3879		
		(-1.64)	(-3.64)			(-2.20)	(-4.02)		
log(FundSize)				-0.0025	-0.0012			-0.0014	-0.0012
				(-1.96)	(-5.85)			(-1.13)	(-5.86)
Panel B: MM-estimation									
IndustrySize	-0.0192					-0.0088	-0.0092	-0.0081	-0.0090
	(-5.02)					(-2.05)	(-2.18)	(-1.90)	(-2.14)
FundSize		-0.2685	-0.3076			-0.3224	-0.3127		
		(-2.00)	(-3.78)			(-2.40)	(-4.07)		
log(FundSize)				-0.0018	-0.0009			-0.0011	-0.0009
				(-1.66)	(-5.41)			(-1.05)	(-5.48)
Estimator	FE	PST RD	Zhu RD	PST RD	Zhu RD	PST RD	Zhu RD	PST RD	Zhu RD

Table 4
Symmetric versus asymmetric outlier removal: PST specification

This table presents results obtained from regressing *GrossR*, a benchmark-adjusted gross fund return computed using historical Morningstar benchmarks, on industry size and fund size, computed as in PST. The estimation is conducted by PST's RD procedure while removing a progressively growing fraction of multivariate outliers, which we identify by following AHM's procedure. Panel A implements AHM's asymmetric approach, which removes positive vertical outliers only. Panel B removes negative vertical outliers only. Panel C follows a symmetric approach that removes both positive and negative vertical outliers. We use the cleanest version of our sample for the period March 1993 through December 2011. The *t*-statistics are clustered by style×month and fund.

	Percent of observations dropped									
	0.00%	0.025%	0.05%	0.10%	0.25%	0.50%	1.00%	2.50%	5.00%	10.00%
Panel A: Remove positive vertical outliers										
IndustrySize	-0.0267	-0.0251	-0.0240	-0.0214	-0.0158	-0.0068	0.0081	0.0225	0.0626	0.1226
	(-2.47)	(-2.34)	(-2.23)	(-2.02)	(-1.52)	(-0.67)	(0.75)	(1.67)	(5.15)	(9.25)
FundSize	-0.320	-0.328	-0.322	-0.342	-0.318	-0.512	-1.053	1.226	0.393	0.738
	(-1.10)	(-1.08)	(-1.01)	(-0.94)	(-0.87)	(-1.04)	(-1.02)	(0.74)	(0.36)	(0.68)
Panel B: Remove negative vertical outliers										
IndustrySize	-0.0267	-0.0277	-0.0289	-0.0310	-0.0357	-0.0417	-0.0521	-0.0737	-0.1057	-0.1653
	(-2.47)	(-2.56)	(-2.67)	(-2.86)	(-3.30)	(-3.83)	(-4.39)	(-5.43)	(-7.37)	(-10.22)
FundSize	-0.320	-0.333	-0.314	-0.301	-0.267	-0.262	-0.182	-0.408	-0.466	0.114
	(-1.10)	(-1.07)	(-0.98)	(-0.84)	(-0.74)	(-0.55)	(-0.17)	(-0.33)	(-0.41)	(0.10)
Panel C: Remove vertical outliers symmetrically										
IndustrySize	-0.0267	-0.0265	-0.0261	-0.0262	-0.0257	-0.0246	-0.0214	-0.0220	-0.0198	-0.0149
	(-2.47)	(-2.46)	(-2.42)	(-2.43)	(-2.46)	(-2.41)	(-2.07)	(-2.06)	(-2.01)	(-1.68)
FundSize	-0.320	-0.338	-0.354	-0.312	-0.318	-0.275	-0.468	0.187	0.225	0.367
	(-1.10)	(-1.11)	(-1.11)	(-0.88)	(-0.89)	(-0.63)	(-0.51)	(0.19)	(0.26)	(0.49)

Table 5
Symmetric versus asymmetric outlier removal: Zhu (2018) specification

This table is the counterpart of Table 4, except that the independent variable is the logarithm of fund size, following Zhu (2018), and we use Zhu's RD.

	Percent of observations dropped									
	0.00%	0.025%	0.05%	0.10%	0.25%	0.50%	1.00%	2.50%	5.00%	10.00%
Panel A: Remove positive vertical outliers										
log(FundSize)	-0.0024	-0.0023	-0.0022	-0.0020	-0.0017	-0.0014	-0.0009	0.0001	0.0013	0.0029
	(-5.53)	(-5.22)	(-5.00)	(-4.72)	(-4.12)	(-3.52)	(-2.48)	(0.36)	(3.10)	(5.80)
Panel B: Remove negative vertical outliers										
log(FundSize)	-0.0024	-0.0025	-0.0025	-0.0026	-0.0028	-0.0031	-0.0035	-0.0044	-0.0057	-0.0077
	(-5.53)	(-5.70)	(-5.85)	(-6.00)	(-6.30)	(-6.75)	(-7.20)	(-8.20)	(-9.53)	(-10.70)
Panel C: Remove vertical outliers symmetrically										
log(FundSize)	-0.0024	-0.0024	-0.0023	-0.0023	-0.0022	-0.0021	-0.0020	-0.0019	-0.0018	-0.0017
	(-5.53)	(-5.48)	(-5.41)	(-5.35)	(-5.38)	(-5.31)	(-5.43)	(-5.46)	(-5.55)	(-5.62)

Table 6
Robustness to dropping extreme values of *GrossR*

This table presents results obtained from regressing *GrossR*, a benchmark-adjusted gross fund return computed using historical Morningstar benchmarks, on industry size, computed as in PST. Column 1 matches column 4 in Table 1. In the following columns, we reestimate the model after dropping the $(x/2)\%$ lowest and $(x/2)\%$ highest values of *GrossR*; the table's bottom row shows the values of x. We estimate the model by OLS with fund FEs, using the cleanest version of our sample for the period March 1993 through December 2011.

	(1)	(2)	(3)	(4)	(5)	(6)
IndustrySize	-0.0310	-0.0168	-0.0119	-0.0101	-0.00745	-0.00584
	(-3.80)	(-3.05)	(-2.65)	(-2.67)	(-2.31)	(-2.11)
Observations	274,729	260,992	247,256	233,520	219,784	206,046
% obs. dropped	0%	5%	10%	15%	20%	25%

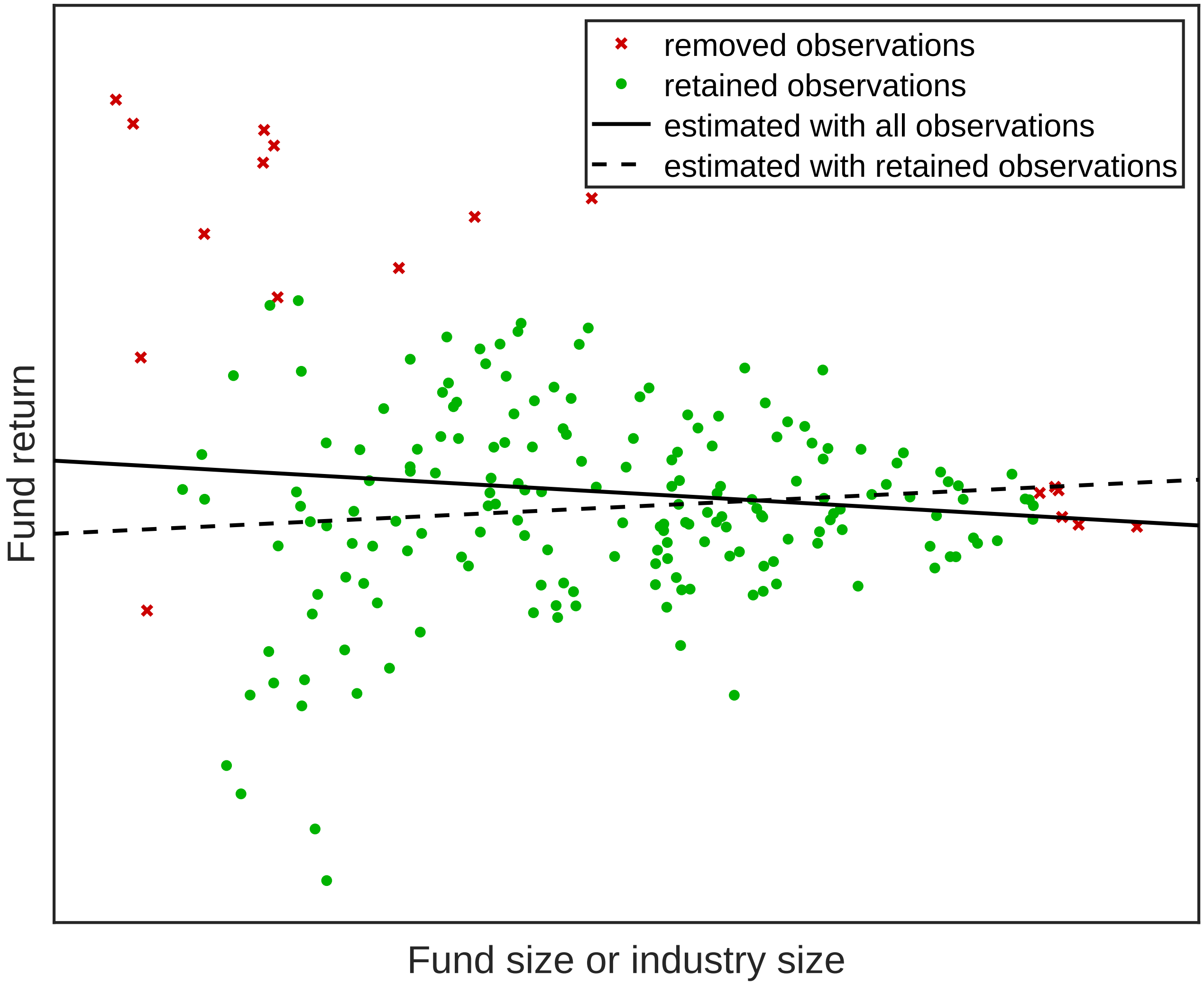

Figure 1. Bias induced by AHM's outlier removal. The figure illustrates how AHM's procedure for removing influential observations imparts an upward bias to the return-size relation estimated with the remaining observations.

REFERENCES

Adams, John, Darren Hayunga, and Sattar Mansi, 2021, Scale and performance in active management are not negatively related, forthcoming in *Critical Finance Review.*

Berk, Jonathan B., and Richard C. Green, 2004, Mutual Fund Flows and Performance in Rational Markets, *Journal of Political Economy* 112, 1269–1295.

Fama, Eugene F. and Kenneth R. French, 1993, Common risk factors in the returns on stocks and bonds, *Journal of Financial Economics* 33, 3–56.

Heritier, Stephane, Eva Cantoni, Samuel Copt, and Maria-Pia Victoria-Feser, 2009, *Robust Methods in Biostatistics*, John Wiley & Sons, Ltd.

Huber, Peter J., 1964, Robust estimation of a location parameter, *Annals of Statistics* 53, 73–101.

Maronna, Ricardo A., Douglas R. Martin, and Victor J. Yohai, 2006, *Robust Statistics: Theory and Methods*, Wiley Series in Probability and Statistics, Chichester: John Wiley & Sons, Ltd.

Pástor, Ľuboš, Robert F. Stambaugh, and Lucian A. Taylor, 2015, Scale and skill in active management, *Journal of Financial Economics* 116, 23–45.

Pástor, Ľuboš, Robert F. Stambaugh, and Lucian A. Taylor, 2020, Fund tradeoffs, *Journal of Financial Economics* 138, 614–634.

Rousseeuw, Peter J., and Bert C. van Zomeren, 1990, Unmasking multivariate outliers and leverage points, *Journal of the American Statistical Association* 85, 633–639.

Stambaugh, Robert F., 2014, Presidential address: Investment noise and trends, *Journal of Finance* 69, 1415–1453.

Yohai, V.J., 1987, High breakdown point and high breakdown-point and high efficiency robust estimates for regression, *Annals of Statistics* 15, 642–656.

Zhu, Min, 2018, Informative fund size, managerial skill, and investor rationality, 2018, *Journal of Financial Economics* 130, 114–134.