Lucky Factors

Campbell R. Harvey

Duke University, Durham, NC 27708 USA
National Bureau of Economic Research, Cambridge, MA 02138 USA

Yan Liu[*]

Texas A&M University, College Station, TX 77843 USA

Current version: December 21, 2015

Abstract

We propose a new method to select amongst a large group of candidate factors — many of which might arise as a result of data mining — that purport to explain the cross-section of expected returns. The method is robust to general distributional characteristics of both factor and asset returns. We allow for the possibility of time-series as well as cross-sectional dependence. The technique accommodates a wide range of test statistics. Our method can be applied to both asset pricing tests based on portfolio sorts as well as tests using individual asset returns. In contrast to recent asset pricing research, our study of individual stocks finds that the original market factor is by far the most important factor in explaining the cross-section of expected returns.

Keywords: Factors, Factor selection, Variable selection, Bootstrap, Data mining, Orthogonalization, Multiple testing, Predictive regressions, Fama-MacBeth, GRS.

[*] Current Version: December 21, 2015. First posted on SSRN: November 20, 2014. Previously circulated under the title "How Many Factors?" and "Incremental Factors". Send correspondence to: Campbell R. Harvey, Fuqua School of Business, Duke University, Durham, NC 27708. Phone: +1 919.660.7768, E-mail: cam.harvey@duke.edu . We appreciate the comments of Tony Cookson, Gene Fama, Thomas Flury, Ken French, Bryan Kelly, Hagen Kim, Marco Rossi, and Michael Stutzer as well as seminar participants at Texas A&M University, Baylor University, UC Boulder, Wharton Jacobs Levy Conference, New York, Society for Quantitative Analysis, New York, and Man Quant Conference, New York. We thank Yong Chen for supplying us with the mutual fund data. We thank Gene Fama, Ken French, Bryan Kelly and Lu Zhang for sharing their factor returns data.

1 Introduction

There is a common thread connecting some of the most economically important problems in finance. For example, how do we determine that a fund manager has "outperformed" given that there are thousands of managers and even those following random strategies might outperform? How do we assess whether a variable such as a dividend yield predicts stock returns given that so many other variables have been tried? Should we use a three-factor model for asset pricing or a new five factor model given that recent research documents that over 300 variables have been published as candidate factors? The common thread is multiple testing or data mining.

Our paper proposes a new method that enables us to better identify the flukes. The method is based on a bootstrap that allows for general distributional characteristics of the observables, a range of test statistics (e.g., R^2, t-ratios, etc.), and, importantly, preserves both the cross-sectional and time-series dependence in the data. Our method delivers specific recommendations. For example, for a p-value of 5%, our method delivers a marginal test statistic. In performance evaluation, this marginal test statistic identifies the funds that outperform or underperform. In our main application which is asset pricing, it will allow us to choose a specific group of factors, i.e., we answer the question: How many factors?

Consider the following example in predictive regressions to illustrate the problems we face. Suppose we have 100 candidate X variables to predict a variable Y. Our first question is whether any of the 100 X variables appear to be individually significant. This is not as straightforward as one thinks because what comes out as significant at the conventional level may be "significant" by luck. We also need to take the dependence among the X variables into account since large t-statistics may come in bundles if the X variables are highly correlated. Suppose these concerns have been addressed and we find a significant predictor, how do we proceed to find the next one? Presumably, the second one needs to predict Y in addition to what the first variable can predict. This additional predictability again needs to be put under scrutiny given that 99 variables can be tried. Suppose we establish the second variable is a significant predictor. When should we stop? Finally, suppose instead of predictive regressions, we are trying to determine how many factors are important in a cross-sectional regression. How should our method change in order to answer the same set of questions but accommodate the potentially time-varying risk loadings in a Fama-MacBeth type of regression?

We provide a new framework that answers the above questions. Several features distinguish our approach from existing studies.

First, we take data mining into account.[1] This is important given the collective effort in mining new factors by both academia and the finance industry. Data mining has a large impact on hypothesis testing. In a single test where a single predetermined variable X is used to explain the left-hand side variable Y, a t-statistic of 2.0 suffices to overcome the 5% p-value hurdle. When there are 100 candidate X variables and assuming independence, the 2.0 threshold for the maximal t-statistic corresponds to a p-value of 99% — not even near the acceptable 5%.[2] Our paper proposes appropriate statistical cutoffs that control for the search among the candidate variables.

While cross-sectional independence is a convenient assumption to illustrate the point of data snooping bias, it turns out to be a big assumption. First, it is unrealistic for most of our applications since almost all economic and financial variables are intrinsically linked in complicated ways. Second, a departure from independence may have a large impact on the results. For instance, in our previous example, if all 100 X variables are perfectly correlated, then there is no need for a multiple testing adjustment and the 99% p-value incorrectly inflates the original p-value. Recent work on mutual fund performance shows that taking cross-sectional dependence into account can materially change inference.[3]

Our paper provides a framework that is robust to the form and amount of cross-sectional dependence among the variables. In particular, our method maintains the dependence information in the data matrix, including higher moment and nonlinear dependence. Additionally, to the extent that higher moment dependence is difficult to measure in finite samples and this may bias standard inference, our method automatically takes sampling uncertainty (i.e., the observed sample may underrepresent the population from which it is drawn from) into account and provides inference that does not rely on asymptotic approximations.

Our method uses a bootstrap method. When the data are independent through time, we randomly sample the time periods with replacement. Importantly, when we bootstrap a particular time period, we draw the entire cross-section at that point in time. This allows us to preserve the contemporaneous cross-sectional dependence structure of the data. Additionally, by matching the size of the resampled data with the original data, we are able to capture the sampling uncertainty of the original sample. When the data are dependent through time, we sample with blocks to capture time-series dependence, similar in spirit to White (2000) and Politis and Romano (1994). In essence, our method reframes the multiple hypothesis testing problem in

[1] Different literature uses different terminologies. In physics, multiple testing is dubbed "looking elsewhere" effect. In medical science, "multiple comparison" is often used for simultaneous tests, particularly in genetic association studies. In finance, "data mining", "data snooping" and "multiple testing" are often used interchangeably. We also use these terms interchangeably and do not distinguish them in this paper.

[2] Suppose we have 100 tests and each test has a t-statistic of 2.0. Under independence, the chance to make at least one false discovery is $1 - 0.95^{100} = 1 - 0.006 = 0.994$.

[3] See Fama and French (2010) and Ferson and Chen (2015).

regression models in a way that permits the use of bootstrapping to make inferences that are both intuitive and distribution free.

Empirically, we show how to apply our method to both predictive regression and cross-sectional regression models — the two areas of research for which data snooping bias is likely to be the most severe. However, our method applies to other types of regression models as well. Essentially, what we are providing is a general approach to specifying a regression model when a researcher is faced with multiple variables to choose from.

Our paper adds to the recent literature on the multidimensionality of the cross-section of expected returns. Harvey, Liu and Zhu (2016) document 316 factors discovered by academia and provide a multiple testing framework to adjust for data mining. Green, Hand and Zhang (2013) study more than 330 return predictive signals that are mainly accounting based and show the large diversification benefits by suitably combining these signals. McLean and Pontiff (2015) use an out-of-sample approach to study the post-publication bias of discovered anomalies. The overall finding of this literature is that many discovered factors are likely false. But how many factors are true factors? We provide a new testing framework that simultaneously addresses multiple testing, variable selection, and test dependence in the context of regression models.

Our method is inspired by and related to a number of influential papers, in particular, Foster, Smith and Whaley (FSW, 1997) and Fama and French (FF, 2010). In the application of time-series prediction, FSW simulate data under the null hypothesis of no predictability to help identify true predictors. Our method bootstraps the actual data, can be applied to a number of test statistics, and does not need to appeal to asymptotic approximations. More importantly, our method can be adapted to study cross-sectional regressions where the risk loadings can potentially be time-varying. In the application of manager evaluation, FF (2010) (see also, Kosowski et al., 2006, Barras et al., 2010, and Ferson and Chen, 2015) employ a bootstrap method that preserves cross-section dependence. Our method departs from theirs in that we are able to determine a specific cut-off whereby we can declare that a manager has significantly outperformed or that a factor is significant in the cross-section of expected returns.[4]

Despite the discovery of hundreds of factors to compete with the original factor proposed by Sharpe (1964), our analysis of value weighted individual stocks identifies one dominant factor – the one proposed by Sharpe. We also find a role for a profitability factor but its contribution is economically small. Our analysis of equal weighting provides some evidence of value and size factors, yet consistent with the value-weighted analysis these additional factors provide a modest contribution compared to the market factor. It is striking that the market factor is the dominant

[4] See Harvey and Liu (2015) for the application of our method to investment fund performance evaluation.

factor used in the practice of corporate finance (see, Graham and Harvey, 2001) yet this factor has long been out of favor in asset pricing research.

Our paper is organized as follows. In the second section, we present our testing framework. In the third section, we apply our method to the selection of risk factors. We offer insights on both tests based on traditional portfolio sorts as well as raw tests based on individual assets. Some concluding remarks are offered in the final section.

2 Method

Our framework is best illustrated in the context of predictive regressions. We highlight the difference between our method and the current practice and relate to existing research. We then extend our method to accommodate cross-sectional regressions.

2.1 Predictive Regressions

Suppose we have a $T \times 1$ vector Y of returns that we want to predict and a $T \times M$ matrix X that includes the time-series of M right-hand side variables, i.e., column i of matrix X (X_i) gives the time-series of variable i. Our goal is to select a subset of the M regressors to form the "best" predictive regression model. Suppose we measure the goodness-of-fit of a regression model by the summary statistic Ψ. Our framework permits the use of an arbitrary performance measure Ψ, e.g., R^2, t-statistic or F-statistic. This feature stems from our use of the bootstrap method, which does not require any distributional assumptions on the summary statistics to construct the test. In contrast, Foster, Smith and Whaley (FSW, 1997) need the finite-sample distribution on R^2 to construct their test. To ease the presentation, we describe our approach with the usual regression R^2 in mind but will point out the difference when necessary.

Our bootstrap-based multiple testing adjusted incremental factor selection procedure consists of three major steps:

Step I. Orthogonalization Under the Null

Suppose we already selected k $(0 \leq k < M)$ variables and want to test if there exists another significant predictor and, if there is, what it is. Without loss of generality, suppose the first k variables are the pre-selected ones and we are testing among the rest $M - k$ candidate variables, i.e., $\{X_{k+j}, j = 1, \ldots, M - k\}$. Our null hypothesis is that none of these candidate variables provides additional explanatory power of Y, following White (2000) and FSW (1997). The goal of this step is to modify the data matrix X such that this null hypothesis appears to be true in-sample.

To achieve this, we first project Y onto the group of pre-selected variables and obtain the projection residual vector $Y^{e,k}$. This residual vector contains information that cannot be explained by pre-selected variables. We then orthogonalize the $M-k$ candidate variables with respect to $Y^{e,k}$ such that the orthogonalized variables are uncorrelated with $Y^{e,k}$ for the entire sample. In particular, we individually project $X_{k+1}, X_{k+2}, \ldots, X_M$ onto $Y^{e,k}$ and obtain the projection residuals $X^e_{k+1}, X^e_{k+2}, \ldots, X^e_M$, i.e.,

$$X_{k+j} = c_j + d_j Y^{e,k} + X^e_{k+j}, \quad j = 1, \ldots, M-k, \tag{1}$$

where c_j is the intercept, d_j is the slope and X^e_{k+j} is the residual vector. By construction, these residuals have an in-sample correlation of zero with $Y^{e,k}$. Therefore, they appear to be independent of $Y^{e,k}$ if joint normality is assumed between X and $Y^{e,k}$.

This is similar to the simulation approach in FSW (1997), in which artificially generated independent regressors are used to quantify the effect of the multiple testing. Our approach is different from FSW because we use real data. In addition, we use bootstrap or block bootstrap to approximate the empirical distribution of test statistics.

We achieve the same goal as FSW while losing as little information as possible for the dependence structure among the regressors. In particular, our orthogonalization guarantees that the $M-k$ orthogonalized candidate variables are uncorrelated with $Y^{e,k}$ in-sample.[5] This resembles the independence requirement between the simulated regressors and the left-hand side variables in FSW (1997). Our approach is distributional free and maintains as much information as possible among the regressors. We simply purge $Y^{e,k}$ out of each of the candidate variables and therefore keep all the distributional information among the variables that is not linearly related to $Y^{e,k}$ intact. For instance, the tail dependency among all the variables — both pre-selected and candidate — is preserved. This is important because higher moment dependence may have a dramatic impact on the test statistics in finite samples.[6]

A similar idea has been applied to the recent literature on mutual fund performance. In particular, Kosowski et al. (2006) and Fama and French (2010) subtract the in-sample fitted alphas from fund returns, thereby creating "pseudo" funds that

[5] In fact, the zero correlation between the candidate variables and $Y^{e,k}$ not only holds in-sample, but also in the bootstrapped population provided that each sample period has an equal chance of being sampled in the bootstrapping, which is true in an independent bootstrap. When we use a stationary bootstrap to take time dependency into account, this is no longer true as samples on the boundary time periods are sampled less frequently. But we should expect this correlation to be small for a long enough sample as the boundary periods are a small fraction of the total time periods.

[6] See Adler, Feldman and Taqqu (1998) for how distributions with heavy tails affect standard statistical inference.

exactly generate a mean return of zero in-sample. Analogously, we orthogonalize candidate regressors such that they exactly have a correlation of zero with what is left to explain in the left-hand side variable, i.e., $Y^{e,k}$.

Step II. Bootstrap

Let us arrange the pre-selected variables into $X^s = [X_1, X_2, \ldots, X_k]$ and the orthogonalized candidate variables into $X^e = [X^e_{k+1}, X^e_{k+2}, \ldots, X^e_M]$. Notice that for both the residual response vector $Y^{e,k}$ and the two regressor matrices X^s and X^e, rows denote time periods and columns denote variables. We bootstrap the time periods (i.e., rows) to generate the empirical distributions of the summary statistics for different regression models. In particular, for each draw of the time index $t^b = [t^b_1, t^b_2, \ldots, t^b_T]'$, let the corresponding left-hand side and right variables be Y^{eb}, X^{sb}, and X^{eb}.

The diagram below illustrates how we bootstrap. Suppose we have five periods, one pre-selected variable X^s, and one candidate variable X^e. The original time index is given by $[t_1 = 1, t_2 = 2, t_3 = 3, t_4 = 4, t_5 = 5]'$. By sampling with replacement, one possible realization of the time index for the bootstrapped sample is $t^b = [t^b_1 = 3, t^b_2 = 2, t^b_3 = 4, t^b_4 = 3, t^b_5 = 1]'$. The diagram shows how we transform the original data matrix into the bootstrapped data matrix based on the new time index.

$$[Y^{e,k}, X^s, X^e] = \underbrace{\begin{bmatrix} y^e_1 & x^s_1 & x^e_1 \\ y^e_2 & x^s_2 & x^e_2 \\ y^e_3 & x^s_3 & x^e_3 \\ y^e_4 & x^s_4 & x^e_4 \\ y^e_5 & x^s_5 & x^e_5 \end{bmatrix}}_{\text{Original data matrix}} \begin{pmatrix} t_1 = 1 \\ t_2 = 2 \\ t_3 = 3 \\ t_4 = 4 \\ t_5 = 5 \end{pmatrix} \Rightarrow \begin{pmatrix} t^b_1 = 3 \\ t^b_2 = 2 \\ t^b_3 = 4 \\ t^b_4 = 3 \\ t^b_5 = 1 \end{pmatrix} \underbrace{\begin{bmatrix} y^e_3 & x^s_3 & x^e_3 \\ y^e_2 & x^s_2 & x^e_2 \\ y^e_4 & x^s_4 & x^e_4 \\ y^e_3 & x^s_3 & x^e_3 \\ y^e_1 & x^s_1 & x^e_1 \end{bmatrix}}_{\text{Bootstrapped data matrix}} = [Y^{eb}, X^{sb}, X^{eb}]$$

Returning to the general case with k pre-selected variables and $M - k$ candidate variables, we bootstrap and then run $M - k$ regressions. Each of these regressions involves the projection of Y^{eb} onto a candidate variable from the data matrix X^{eb}. Let the associated summary statistics be $\Psi^{k+1,b}$, $\Psi^{k+2,b}$, $\ldots$, $\Psi^{M,b}$, and let the maximum among these summary statistics be Ψ^b_I, i.e.,

$$\Psi^b_I = \max_{j \in \{1,2,\ldots,M-k\}} \{\Psi^{k+j,b}\}. \tag{2}$$

Intuitively, Ψ^b_I measures the performance of the best fitting model that augments the pre-selected regression model with one variable from the list of orthogonalized candidate variables.

The max statistic controls for data snooping bias. With $M-k$ factors to choose from, the factor that is selected may appear to be significant through random chance. We adopt the max statistic as our test statistic to control for multiple hypothesis testing, similar to White (2000), Sullivan, Timmermann and White (1999) and FSW (1997). Our bootstrap approach allows us to obtain the empirical distribution of the max statistic under the joint null hypothesis that none of the $M-k$ variables is true. Due to multiple testing, this distribution is very different from the null distribution of the test statistic in a single test. By comparing the realized (in the data) max statistic to this distribution, our test takes multiple testing into account.

Which statistic should we use to summarize the additional contribution of a variable in the candidate list? Depending on the regression model, the choice varies. For instance, in predictive regressions, we typically use the R^2 or the adjusted R^2 as the summary statistic. In cross-sectional regressions, we use the t-statistic to test whether the average slope is significant.[7] One appealing feature of our method is that it does not require an explicit expression for the null distribution of the test statistic. It therefore can easily accommodate different types of summary statistics. In contrast, FSW (1997) only works with the R^2.

For the rest of the description of our method, we assume that the statistic that measures the incremental contribution of a variable from the candidate list is given and generically denote it as Ψ_I or Ψ_I^b for the b-th bootstrapped sample.

We bootstrap $B = 10,000$ times to obtain the collection $\{\Psi_I^b, b = 1, 2, \ldots, B\}$, denoted as $(\Psi_I)^B$, i.e.,

$$(\Psi_I)^B = \{\Psi_I^b, b = 1, 2, \ldots, B\}. \tag{3}$$

This is the empirical distribution of Ψ_I, which measures the maximal additional contribution to the regression model when one of the orthogonalized regressors is considered. Given that none of these orthogonalized regressors is a true predictor in population, $(\Psi_I)^B$ gives the distribution for this maximal additional contribution when the null hypothesis is true, i.e., null of the $M-k$ candidate variables is true. $(\Psi_I)^B$ is the bootstrapped analogue of the distribution for maximal R^2's in FSW (1997). Similar to White (2000) and advantageous over FSW (1997), our bootstrap method is essentially distribution-free and allows us to obtain the exact distribution of the test statistic through sample perturbations.[8]

Our bootstrapped sample has the same number of time periods as the original data. This allows us to match the sampling uncertainty of the original data with the

[7] In cross-sectional regressions, sometimes we use the average pricing errors (e.g., mean absolute pricing error) as the summary statistics. In this case, Ψ^{eb} should be understood as the minimum among the average pricing errors for the candidate variables.

[8] We are able to generalize FSW (1997) in two significant ways. First, our approach allows us to maintain the distributional information among the regressors, helping us avoid the Bonferroni type of approximation in equation (3) of FSW (1997). Second, even in the case of independence, our use of bootstrap takes the sampling uncertainty into account, providing a finite sample version of what is given in equation (2) of FSW (1997).

bootstrapped sample. When there is little time dependence in the data, we simply treat each time period as the sampling unit and sample with replacement. When time dependence is an issue, we use a block bootstrap, as explained in detail in the appendix. In either case, we only resample the time periods. We keep the cross-section intact to preserve the contemporaneous dependence among the variables.

Step III: Hypothesis Testing and Variable Selection

Working on the original data matrix X, we can obtain a Ψ_I statistic that measures the maximal additional contribution of a candidate variable. We denote this statistic as Ψ_I^d. Hypothesis testing for the existence of the $(k+1)$-th significant predictor amounts to comparing Ψ_I^d with the distribution of Ψ_I under the null hypothesis, i.e., $(\Psi_I)^B$. With a pre-specified significance level of α, say 5%, we reject the null if Ψ_I^d exceeds the $(1-\alpha)$-th percentile of $(\Psi_I)^B$, that is,

$$\Psi_I^d > (\Psi_I)^B_{1-\alpha}, \tag{4}$$

where $(\Psi_I)^B_{1-\alpha}$ is the $(1-\alpha)$-th percentile of $(\Psi_I)^B$.

The result of the hypothesis test tells us whether there exists a significant predictor among the remaining $M-k$ candidate variables, after taking multiple testing into account. Had the decision been positive, we declare the variable with the largest test statistic (i.e., Ψ_I^d) as significant and include it in the list of pre-selected variables. We then start over from Step I to test for the next predictor, if not all predictors have been selected. Otherwise, we terminate the algorithm and arrive at the final conclusion that the pre-selected k variables are the only ones that are significant.

2.2 Panel Regression Models

Our method can be applied to panel regression models commonly used in asset pricing tests, where asset returns are regressed on a set of common factors. We demean factor returns such that the demeaned factors have zero impact in explaining the cross-section of expected returns. However, their ability to explain variation in asset returns in time-series regressions is preserved. This way, we are able to disentangle the time-series vs. cross-sectional contribution of a candidate factor.

We start by writing down a time-series regression model,

$$R_{it} - R_{ft} = a_i + \sum_{j=1}^{K} b_{ij} f_{jt} + \epsilon_{it}, i = 1, \ldots, N, \tag{5}$$

in which the time-series of excess returns $R_{it} - R_{ft}$ are projected onto K contemporaneous factor returns f_{it}. Factor returns are the long-short strategy returns corresponding to zero cost investment strategies. If the set of factors are mean-variance efficient (or, equivalently, if the corresponding beta pricing model is true), the cross-section of regression intercepts should be indistinguishable from zero. This constitutes the testable hypothesis for the classic Gibbons, Ross and Shanken (GRS, 1989) test.

The GRS test is widely applied in empirical asset pricing. However, several issues hinder further applications of the test, or time-series tests in general. First, the GRS test almost always rejects. This means that almost no model can adequately explain the cross-section of expected returns. As a result, most researchers use the GRS test statistic as a heuristic measure for model performance (see, e.g., Fama and French, 2015a). For instance, if Model A generates a smaller GRS statistic than Model B, we would take Model A as the "better" Model, although neither model survives the GRS test. But does Model A "significantly" outperform B? The original GRS test cannot answer answer this question because the overall null of the test is that all intercepts are strictly at zero. When two competing models both generate intercepts that are not at zero, the GRS test is not designed to measure the relative performance of the two models. Our method provides a solution to this problem. In particular, for two models that are nested, it allows us to tell the incremental contribution of the bigger model relative to the smaller one, even if both models fail to meet the GRS null hypothesis.

Second, compared to cross-sectional regressions (e.g., the Fama-MacBeth regression), time-series regressions tend to generate a large time-series R^2. This makes them appear more attractive than cross-sectional regressions because the cross-sectional R^2 is usually much lower.[9] However, why would it be the case that a few factors that explain more than 90% of the time-series variation in returns are often not even significant in cross-sectional tests? Why would the market return explain a significant fraction of variation in individual stock and portfolio returns in time-series regressions but offer little help in explaining the cross-section? These questions point to a general inquiry into asset pricing tests: is there a way to disentangle the time-series vs. cross-sectional contribution of a candidate factor? Our method achieves this by demeaning factor returns. By construction, the demeaned factors have zero impact on the cross-section while having the same explanatory power in time-series regressions as the original factors. Through this, we test a factor's significance in explaining the cross-section of expected returns, holding its time-series predictability constant.

Third, the inference for the GRS test which is based on asymptotic approximations can be problematic. For instance, MacKinlay (1987) shows that the test tends to have low power when the sample size is small. Affleck-Graves and McDonald (1989) show that nonnormalities in asset returns can severely distort its size and/or power. Our method relies on bootstrapped simulations and is thus robust to small-sample or

[9] See Lewellen, Nagel and Shanken (2010).

nonnormality distortions. In fact, bootstrap based resampling techniques are often recommended to mitigate these sources of bias.

Our method tries to overcome the aforementioned shortcomings in the GRS test by resorting to our bootstrap framework. The intuition behind our method is already given in our previous discussion on predictive regressions. In particular, we orthogonalize (or more precisely, demean) factor returns such that the orthogonalized factors do not impact the cross-section of expected returns.[10] This absence of impact on the cross-section constitutes our null hypothesis. Under this null, we bootstrap to obtain the empirical distribution of the cross-section of pricing errors. We then compare the realized (i.e., based on the real data) cross-section of pricing errors generated under the original factor to this empirical distribution to provide inference on the factor's significance. We describe our panel regression method as follows.

Without loss of generality, suppose we only have one factor (e.g., the excess return on the market $f_{1t} = R_{mt} - R_{ft}$) on the right-hand side of (5). Taking unconditional expectations on both sides of (5), we have

$$E(R_{it} - R_{ft}) = a_i + b_{i1}E(f_{1t}). \tag{6}$$

The mean excess return of the asset can be decomposed into two parts. The first part is the time-series regression intercept (i.e., a_i), and the second part is the product of the time-series regression slope and the average factor return (i.e., $b_{i1}E(f_{1t})$).

In order for the one-factor model to work, we need $a_i = 0$ across all assets. Imposing this condition in (6), we have $b_{i1}E(f_{1t}) = E(R_{it} - R_{ft})$. Intuitively, the cross-section of $b_{i1}E(f_{1t})$'s need to line up with the cross-section of expected asset returns (i.e., $E(R_{it} - R_{ft})$) in order to fully absorb the intercepts in time-series regressions. This condition is not easy to satisfy in time-series regressions because the cross-section of risk loadings (i.e., the b_i) are determined by individual time-series regressions. The risk loadings may happen to line up with the cross-section of asset returns and thereby making the one-factor model work or they may not. This suggests the possibility that some factors (e.g., the market factor) may generate large time-series regression R^2's but contribute little to explaining the cross-section of asset returns.

Another important observation from (6) is that by setting $E(f_{1t}) = 0$, factor f_{1t} exactly has zero impact on the cross-section of expected asset returns. Indeed, if $E(f_{1t}) = 0$, the cross-section of intercepts from time-series regressions (i.e., the a_i) exactly equal the cross-section of average asset returns (i.e., $E(R_{it} - R_{ft})$) that the factor model is supposed to help explain in the first place. On the other hand, whether or not the factor mean is zero does not matter for time-series regressions. In

[10] More precisely, our method makes sure that the orthogonalized factors have a zero impact on the cross-section of expected returns *unconditionally*. This is because panel regression models with constant risk loadings focus on unconditional asset returns.

particular, both the regression R^2 and the slope coefficient (i.e., b_{i1}) are kept intact when we alter the factor mean.

The above discussion motivates our test design. For the one-factor model, we define a "pseudo" factor $\tilde{f}_{1t}$ by subtracting the in-sample mean of f_{1t} from its time-series. This demeaned factor maintains all the time-series predictability of f_{1t} but has no role in explaining the cross-section of expected returns. With this pseudo factor, we bootstrap to obtain the distribution of a statistic that summarizes the cross-section of pricing errors (i.e., regression intercepts). Candidate statistics include mean/median absolute pricing errors, mean squared pricing errors, and absolute t-statistics. We then compare the realized statistic for the original factor (i.e., f_{1t}) to this bootstrapped distribution.

Our method generalizes straightforwardly to the situation when we have multiple factors. Suppose we have K pre-selected factors and we want to test the $(K+1)$-th factor. We first project the $(K+1)$-th factor onto the pre-selected factors through a time-series regression. We then define the new pseudo factor by subtracting the regression intercept from the $(K+1)$-th factor. This is analogous to the previous one-factor model example. In the one-factor model, demeaning is equivalent to projecting the factor onto a constant.

We use an example to illustrate how our method works when there are multiple factors. Suppose we have one pre-selected factor (f_{1t}) in the baseline model. The regression equation for asset i is:

$$R_{it} - R_{ft} = a_i + b_{i1} f_{1t} + e_{it}. \tag{7}$$

Now suppose we add another factor f_{2t} to the baseline model and denote the augmented model as:

$$R_{it} - R_{ft} = a_i^* + b_{i1}^* f_{1t} + b_{i2}^* f_{2t} + e_{it}^*. \tag{8}$$

If f_{2t} were a true factor, compared to a_i, a_i^* should be closer to zero. We therefore want to compare a_i^* with a_i. In general, $a_i \neq a_i^*$. Our goal is to adjust f_{2t} such that the adjusted f_{2t} (denoted as f_{2t}^*) guarantees $a_i = a_i^*$, that is, the regression intercept under the augmented model is the same as the intercept under the baseline model. Our description in the previous paragraph achieves this. In particular, let the regression equation that projects f_{2t} onto f_{1t} be:

$$f_{2t} = \alpha + \beta f_{1t} + \varepsilon_t, \tag{9}$$

and define f_{2t}^* as

$$f_{2t}^* \equiv f_{2t} - \alpha = \beta f_{1t} + \varepsilon_t. \tag{10}$$

Thus defined f_{2t}^*, when substituting f_{2t} in (8), makes sure that $a_i^* = a_i$. To see this, we replace f_{2t}^* with f_{2t} in (8) and rewrite the regression equation as:

$$\begin{aligned} R_{it} - R_{ft} &= a_i^* + b_{i1}^* f_{1t} + b_{i2}^*(\beta f_{1t} + \varepsilon_t) + e_{it}^* && (11) \\ &= a_i^* + (b_{i1}^* + b_{i2}^*\beta) f_{1t} + \underbrace{(b_{i2}^*\varepsilon_t + e_{it}^*)}_{u_{it}^*}. && (12) \end{aligned}$$

By construction, both ε_t and e_{it}^* are orthogonal to f_{1t} and a vector of ones. Hence, by treating $u_{it}^* = b_{i2}^*\varepsilon_t + e_{it}^*$ as the new regression residual and by comparing (12) with (7), we must have:

$$a_i^* = a_i, \quad b_{i1}^* + b_{i2}^*\beta = b_{i1}. \tag{13}$$

Our adjustment makes economic sense. Taking unconditional expectations on both sides of (10), we have

$$E(f_{2t}^*) = \beta E(f_{1t}). \tag{14}$$

Therefore, the adjusted factor f_{2t}^* is absorbed by the pre-selected factor f_{1t} in the sense that its premium is completely explained by its exposure to the pre-selected factor. When this happens, the adjusted factor has zero incremental impact on the cross-section of expected returns. In the meantime, it has perfect time-series correlation with the original factor in-sample and has the same time-series correlation with the pre-selected variable as the original factor. Hence, the adjusted factor preserves the time-series properties of the original factor aside from the mean.

With this pseudo factor, we bootstrap to generate the distribution of pricing errors. In this step, the difference from the one-factor case is that, for both the original regression and the bootstrapped regressions based on the pseudo factor, we always keep the original K factors in the model. This way, our test captures the incremental contribution of the candidate factor.

2.3 Fama-MacBeth Regressions

Our method can also be adapted to test factor models in cross-sectional regressions. In particular, we show how an adjustment of our method applies to Fama-MacBeth type of regressions (FM, Fama and MacBeth, 1973) — a framework that allows for time-varying risk loadings.

One hurdle in applying our method to FM regressions is the time-varying slopes in cross-sectional regressions. In particular, separate cross-sectional regressions are performed for each time period to obtain a collection of cross-sectional regression slopes. We test the significance of a factor by looking at the time averaged cross-sectional slope coefficient. Therefore, in the FM framework, the null hypothesis is that the slope is zero in population. We adjust our method such that this condition exactly holds in-sample for the adjusted regressors.

First, we need to orthogonalize. Suppose we run a FM regression on a baseline model and obtain the panel of residual excess returns. In particular, at time t, let the vector of residual excess returns be Y_t. We are testing the incremental contribution of a candidate factor in explaining the cross-section of expected returns. Let the vector of risk loadings (i.e., β's) for the candidate factor be X_t. Suppose there are n_t assets in the cross-section at time t so the dimension of both Y_t and X_t is $n_t \times 1$. Notice that n_t can be time-dependent as it is straightforward for our method to handle unbalanced panels. In a typical FM regression, we would project Y_t onto X_t. For our orthogonalization to work, we reverse the process, similar to what we do in predictive regressions. More specifically, we stack the collection of Y_t's and X_t's into two column vectors that have a dimension of $\sum_{t=1}^{T} n_t \times 1$, and run the following constrained regression model:

$$\begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_T \end{bmatrix}_{\sum_{t=1}^{T} n_t \times 1} = \begin{bmatrix} \phi_1 \\ \phi_2 \\ \vdots \\ \phi_T \end{bmatrix}_{\sum_{t=1}^{T} n_t \times 1} + \xi_{1\times 1} \cdot \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_T \end{bmatrix}_{\sum_{t=1}^{T} n_t \times 1} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_T \end{bmatrix}_{\sum_{t=1}^{T} n_t \times 1}, \quad (15)$$

where ϕ_t is the constant vector of intercepts for time t, $\xi_{1\times 1}$ is a scalar, and $[\varepsilon_1', \varepsilon_2', \ldots, \varepsilon_T']'$ is the vector of projected regressors that will be used in the follow-up bootstrap analysis. This is a constrained regression as we have a single regression slope (i.e., ξ) throughout the sample. Had we allowed different slopes across time, we would have the usual unconstrained regression model where X_t is projected onto Y_t period-by-period. Having a single slope coefficient is key for us to achieve the null hypothesis in-sample for the FM model.

Alternatively, we can view the above regression model as an adaptation of the orthogonalization procedure that we use in predictive regressions. It pools returns and factor loadings together to estimate a single slope coefficient. What is different, however, is the use of separate intercepts for different time periods. This is natural since the FM procedure allows time-varying intercepts and slopes. To purge the variation in Y_t's out of X_t's, we need to allow for time-varying intercepts as well. Mathematically, the time-dependent intercepts allow the regression residuals to sum up to zero within each period. This property proves very important in that it allows us to form the FM null hypothesis in-sample, as we shall see later.

Next, we scale each residual vector ε by its sum of squares $\varepsilon'\varepsilon$ and generate the orthogonalized regressor vectors:

$$X_t^e = \varepsilon_t/(\varepsilon_t'\varepsilon_t), \ t = 1, 2, \ldots, T. \quad (16)$$

These orthogonalized regressors are the FM counterparts of the orthogonalized regressors in predictive regressions. They satisfy the FM null hypothesis in cross-sectional regressions. In particular, suppose we run cross-sectional OLS with these orthogonalized regressor vectors for each period:

$$Y_t = \mu_t + \gamma_t X_t^e + \eta_t, \ t = 1, 2, \ldots, T, \tag{17}$$

where μ_t is the $n_t \times 1$ vector of intercepts, γ_t is the scalar slope for the t-th period, and η_t is the $n_t \times 1$ vector of residuals. We show in Appendix A that the following FM null hypothesis holds in-sample:

$$\sum_{t=1}^{T} \gamma_t = 0. \tag{18}$$

The above orthogonalization is the only step that we need to adapt to apply our method to the FM procedure. The rest of our method follows for factor selection in FM regressions. In particular, with a pre-selected set of right-hand side variables, we orthogonalize the rest of the right-hand side variables to form the joint null hypothesis that none of them is a true factor. We then bootstrap to test this null hypothesis. If we reject, we add the most significant one to the list of pre-selected variables and start over to test the next variable. Otherwise, we stop and end up with the set of pre-selected variables.

2.4 Discussion

Across the three different scenarios, our orthogonalization works by adjusting the right-hand side or forecasting variables so they appear irrelevant in-sample. That is, they achieve what are perceived as the null hypotheses in-sample. However, the null varies across the regression models. As a result, a particular orthogonalization method that works in one model may not work in another model. For instance, in the panel regression model the null is that a factor does not help reduce the cross-section of pricing errors. In contrast, in Fama-MacBeth type of cross-sectional regressions, the null is that the time averaged slope coefficients is zero. Following the same procedure as what we do in panel regressions will not achieve the desired null in the FM regressions.

Our method builds on the statistics literature on bootstrap. Jeong and Maddala (1993) suggest that there are two uses of bootstrap that can be justified both theoretically and empirically. First, bootstrap provides a tractable way to conduct statistical analysis (e.g., hypothesis tests, confidence intervals, etc.) when asymptotic theory is not tractable for certain models. Second, even when asymptotic theory is available, it may not be accurate in smaller samples.[11]

Our approach solves at least two problems. First, it is a daunting task to derive asymptotic distributions given the complicated structure of the cross-section of equity returns, e.g., unbalanced panel, cross-sectional dependency, number of firms (N) is large relative to the number of time periods (T), etc. Second, as shown in Affleck-Graves and McDonald (1989), the GRS test is distorted when the returns for test portfolios are non-normally distributed. The problem is likely to be even worse given our use of individual stocks as test assets. Our bootstrap method allows us to overcome these difficulties and conduct robust statistical inference.

More specially, our method falls into the category of nonparametric bootstrap that is routinely used for hypothesis testing. Hall and Wilson (1991) provide two valuable guidelines. The first, which can have a large impact on test power, is that bootstrap resampling should be done in a way that reflects the null hypothesis, even if the true hypothesis is distant from the null.[12] The second is to use pivotal statistics, that is, statistics whose distributions do not depend on unknown parameters.[13]

The design of our tests closely follows these principles. Take our panel regression model as an example. The first step orthogonalization, which is core to our method, ensures that the null hypothesis that a factor has no explanatory power for the cross-section of expected returns is exactly achieved in-sample. Our method therefore abides by the first principle and can potentially have a higher test power compared to alternative designs of the hypothesis tests. In addition, when constructing the test statistics corresponding to the panel regression model, we make sure that pivotal statistics (e.g., t-statistics of the regression intercepts) are considered along with other test statistics.

[11] For other references on bootstrap and its applications to financial time series, see Li and Maddala (1996), Veall (1992, 1998), Efron and Tibshirani (1993), and MacKinnon (2006).

[12] Young (1986), Beran (1988) and Hinkley (1989) discuss the first guideline in more detail.

[13] To give an example of the use of pivotal statistics in bootstrap hypothesis testing, suppose our sample is $\{x_1, x_2, \ldots, x_n\}$ and the hypothesis under test is that the population mean equals θ_0, i.e., $H_0 : \theta = \theta_0$. A test statistic one may want to use is $\hat{\theta}^* - \theta_0$, where $\hat{\theta}^* = \sum_{i=1}^n x_i/n$ is the sample mean. However, this statistic is not pivotal in that its distribution depends on the population standard deviation σ, which is an unknown parameter. According to Hall and Wilson (1991), a better statistic is to divide $\hat{\theta}^* - \theta_0$ by $\hat{\sigma}^*$, where $\hat{\sigma}^*$ is the standard deviation estimate. The new test statistic $(\hat{\theta}^* - \theta_0)/\hat{\sigma}^*$ is an example of a pivotal test statistic.

3 Identifying Factors

3.1 Candidate Risk Factors

In principle, we can apply our method to the grand task of sorting out all the risk factors that have been proposed. One attractive feature of our method is that it allows the number of risk factors to be larger than the number of test portfolios, which is infeasible in conventional multiple regression models. However, we do not pursue this in the current paper but instead focus on a select group of prominent risk factors. The choice of the test portfolios is a major confounding issue. Different test portfolios lead to different results. In contrast, individual stocks avoid the arbitrary portfolio construction. We apply our method to both popular test portfolios and individual stocks.

In particular, we apply our panel regression method to 14 risk factors that are proposed by Fama and French (2015a), Frazzini and Pedersen (2014), Novy-Marx (2013), Pastor and Stambaugh (2003), Carhart (1997), Asness, Frazzini and Pedersen (2013), Hou, Xue and Zhang (2015), Harvey and Siddique (2000), and Herskovic, Kelly, Lustig and Van Nieuwerburgh (2014).[14]

We first provide acronyms for factors. Fama and French (2015a) add profitability (*rmw*) and investment (*cma*) to the three-factor model of Fama and French (1993), which has market (*mkt*), size (*smb*) and book-to-market (*hml*) as the pricing factors. Hou, Xue and Zhang (2015) propose similar profitability (*roe*) and investment (*ia*) factors. Other factors include betting againist beta (*bab*) in Frazzini and Pedersen (2014), gross profitability (*gp*) in Novy-Marx (2013), Pastor and Stambaugh liquidity (*psl*) in Pastor and Stambaugh (2003), momentum (*mom*) in Carhart (1997), quality minus junk (*qmj*) in Asness, Frazzini and Pedersen (2013), co-skewness (*skew*) in Harvey and Siddique (2000), and common idiosyncratic volatility (*civ*) in Herskovic, Kelly, Lustig, and Van Nieuwerburgh (2014). We treat these 14 factors as candidate risk factors and incrementally select the group of "true" factors. True is in quotation marks because there are a number of other issues such as the original set of factors that we consider. Had we considered a larger set of factors, our results could have been different. We leave these extensions to future research.

[14] The factors in Fama and French (2015a), Hou, Xue and Zhang (2015), Harvey and Siddique (2000) and Herskovic, Kelly, Lustig and Van Nieuwerburgh (2014) are provided by the authors. The factors for the rest of the papers are obtained from the authors' webpages. Across the 14 factors, the liquidity factor in Pástor and Stambaugh (2003) has the shortest length (i.e., January 1968 - December 2012). We therefore focus on the January 1968 to December 2012 period to make sure that all factors have the same sampling period.

3.2 Test Statistics

We focus on test statistics that are economically sensible and statistically sound. Intuitively, a good test statistic in our context should be able to tell the difference in explaining the cross-section of expected returns between a baseline model and an augmented model that adds one additional variable to the baseline model. For the panel regression model, let $\{a_i^b\}_{i=1}^N$ and $\{a_i^g\}_{i=1}^N$ be the cross-section of regression intercepts for the baseline model and the augmented model, respectively. Let $\{s_i^b\}_{i=1}^N$ be the cross-section of standard errors for regression intercepts under the baseline model. Our first test statistic is given by

$$\begin{aligned} SI_{ew}^m &\equiv (\frac{1}{N}\sum_{i=1}^N |a_i^g|/s_i^b - \frac{1}{N}\sum_{i=1}^N |a_i^b|/s_i^b)/\frac{1}{N}\sum_{i=1}^N |a_i^b|/s_i^b, \\ &= \frac{\frac{1}{N}\sum_{i=1}^N (|a_i^g| - |a_i^b|)/s_i^b}{\frac{1}{N}\sum_{i=1}^N |a_i^b|/s_i^b}, \end{aligned}$$

where SI denotes 'scaled intercept', 'ew' denotes equal weighting, and 'm' denotes mean. Intuitively, SI_{ew}^m measures the percentage difference in the absolute regression intercepts scaled by the standard error for the regression intercept under the baseline model. We would expect SI_{ew}^m to be negative if the augmented model improves the baseline model. The significance of the improvement is evaluated against the bootstrapped empirical distribution that is generated under the null hypothesis that the additional variable in the augmented model has zero incremental contribution in explaining the cross-section of expected returns.

While SI_{ew}^m calculates the percentage difference in the scaled mean absolute intercept, it may not be robust to extreme observations in the cross-section, especially when we use individual stocks as test assets. We therefore also consider a robust version that calculates the percentage difference in the scaled median absolute intercept, that is,

$$SI_{ew}^{med} \equiv (median(\{|a_i^g|/s_i^b\}_{i=1}^N) - median(\{|a_i^b|/s_i^b\}_{i=1}^N))/median(\{|a_i^b|/s_i^b\}_{i=1}^N),$$

where $median(\cdot)$ denotes the median of a group of variables and is denoted by a superscript '*med*'.

One key assumption for the validity of our test statistic is that the cross-sectionally averaged $|a_i^g|$ should be smaller than the cross-sectionally averaged $|a_i^b|$ if the additional factor in the augmented model is a true risk factor. At the individual asset level, $|a_i^g|$ will be smaller than $|a_i^b|$ in population (i.e., we have long enough factor and return time-series) if the augmented model is the true underlying factor model. Indeed, if the augmented model is correctly specified, then a_i^g will be zero in population, which is no greater than $|a_i^b|$ under the (incorrectly specified) baseline model.

In reality, we never know whether the augmented model is the true underlying factor model. When the augmented model is misspecified, it is possible that $|a_i^g|$ will be larger than $|a_i^b|$ for certain assets even if the additional factor in the augmented model is a true risk factor. Model misspecification may cause bias in inference not just for our method, but most likely all existing asset pricing models. For instance, it is well-known that Fama-MacBeth regressions are severely biased when the misspecified factor model includes spurious factors (Kan and Zhang, 1999, Bryzgalova, 2014). Similarly, when there is model misspecification, the GRS test will likely reject the factor model, thus unable to identify true risk factors that belong to the underlying true factor model. Interestingly, by simulating individual stock returns under realistic assumptions about the underlying true factor model, Harvey and Liu (2016) find that our method is more robust to model misspecification than existing asset pricing tests. We therefore focus on the above test statistics to provide inference.

There are many reasons for us to consider the scaled intercept instead of the original intercept. First, in a time-series regression model, by thinking of the fitted combination of zero-cost portfolios (that is, factor proxies) as a benchmark index, the scaled intercept is closely related to the *information ratio* of the strategy that takes a long position in the test asset and a short position in the benchmark index.[15] When test assets are not diversified portfolios, information ratio is a better scaled metric to gauge the economic significance of the investment strategy. This is similar to the use of the t-statistic instead of the Jensen's alpha in performance evaluation. The t-statistic of alpha — not alpha itself — tells us how "abnormal" the returns are that are produced by a fund manager.

Second, the use of the scaled intercept takes the heterogeneity in return volatilities into account. Suppose two stocks generate the same regression intercept by fitting a factor model. Then the degree of mispricing by the factor model, as measured by the absolute value of the regression intercept, should be higher for the stock that is less noisy. In other words, we should assign less weight to stocks that are noisier in our panel regression model. This is particularly important when we consider individual stocks as test assets as there is a large amount of heterogeneity in return volatility for individual stocks.

Finally, as mentioned previously, our use of the scaled intercept is consistent with the second principle for bootstrap hypothesis testing in Hall and Wilson (1991). In fact, scaling the intercept by the standard deviation is exactly the recommended transformation in Hall and Wilson (1991) to obtain pivotal statistics.

Another important feature of our test statistics is that we scale the intercepts of the baseline model and the augmented model by the same standard error, that is, the standard error of the estimate of the intercept under the baseline model. This makes sure that our test statistics are exactly zero when the null hypothesis — the candidate factor has zero incremental contribution to explain the cross-section of

[15] See Treynor and Black (1973).

expected returns — is forced to exactly hold in-sample for our procedure. This may not hold under alternative scaling schemes. For example, one might propose the use of the standard errors corresponding to the baseline model and the augmented model to separately scale the regression intercepts under the two models. This does not work in our setup as the orthogonalized candidate factor (e.g., the demeaned market factor), which is constructed to have a zero impact on the regression intercepts, may have non-negligible impacts on the standard errors. As a result, the test statistic will not equal zero at the null hypothesis since the same intercept is scaled by two different standard errors. This makes it difficult to disentangle the cross-sectional impact of a candidate factor from its time-series impact. Our test statistics allow us to single out the cross-sectional contribution of the candidate factor.

What is the difference between our test and the GRS test? The GRS test hypothesizes that the augmented model is true and evaluates the cross-section of $|a_i^g|$ to test this hypothesis. A failure of the test indicates the rejection of the augmented model but tells us little about the individual significance of the additional factor. In contrast, our test hypothesizes that the baseline model is true and uses the reduction in scaled absolute intercepts to evaluate the incremental contribution of the additional factor in the augmented model. As a result, our test is able to tell whether the additional factor is individually significant as a risk factor without having to make a statement about the overall model performance. This is important because given the uncertainty about the underlying true factor model, any given factor model is likely to be misspecified. Our test is more robust to model misspecification compared to the GRS test. We delve further on this point in the next section.

Another difference is that instead of using the entire residual covariance matrix to weight the cross-section of regression intercepts as in the GRS test, we use the individual standard errors for the estimation of the intercepts to weight the cross-section of intercepts. This might seem like a drawback of our test since the GRS test allows one to use the residual covariance matrix to construct portfolios that are mean-variance more efficient than portfolios based on the tested factors alone, thereby improving test power. However, in reality, the instability in the estimation of the covariance matrix may offset the gain in test power. As shown in Fama and French (2015b), when applied to portfolios, the GRS test often implies unrealistically large short positions on certain assets to achieve mean-variance efficiency. This causes trouble in interpreting the GRS test from an economic perspective. Additionally, when we use individual stocks as test assets, the covariance matrix will be poorly measured since the number of assets in the cross-section is larger than the number of periods in the time-series rendering the GRS test inapplicable. Given these concerns, we believe that our test has some advantages over the GRS test when applied to a large cross-section of assets. It allows us to take the residual volatility for each individual asset into account while at the same time avoiding the estimation of the large dimensional residual covariance matrix.[16] We provide a more detailed comparison of our method and the GRS test towards the end of this section.

While we focus on the above test statistics, many other test statistics are feasible. For example, instead of using the scaled intercepts, one may want to use the original intercepts. As another example, we can use value weighting rather than equal weighting. Intuitively, for two stocks that generate the same regression intercept, the mispricing of the factor model should be more economically significant for the stock that has a higher market value. We define alternative test statistics and show the corresponding results in the online appendix of our paper.

The fact that our framework allows us to consider a variety of test statistics demonstrates the flexibility of our bootstrap approach. For example, we usually do not have closed-form asymptotic approximations for test statistics that are based on quantiles (e.g., the median). Our bootstrap-based approach allows us to provide inference for test statistics that rely on the median, which is robust to outliers and therefore instrumental to our application to individual stocks. With a few caveats in mind for the construction of a well-behaved test statistic, our approach is able to provide statistical inference for a variety of test statistics, some of which are of great interest to us from an economic perspective.

Instead of using the equally weighted scaled intercepts, Fama and French (2015a) use the equally weighted absolute intercepts as the heuristic test statistic to evaluate the performance of their investment and profitability factors. Our framework allows us to make precise statements about the statistical significance of their test statistics. However, as shown in the simulation study in the on-line appendix of our paper, our test statistics are much more powerful than their intercepts-based test statistics. We therefore focus on the two aforementioned test statistics (i.e., SI_{ew}^{m} and SI_{ew}^{med}) in our paper.

We can also interpret our test statistics from an investment perspective. However, we postpone such interpretations to later sections, where we discuss the drawbacks the GRS test in more detail.

3.3 Results: Portfolios as Test Assets

We first apply our method to popular test portfolios. In particular, we use the standard 25 size and book-to-market sorted portfolios that are available from Ken French's on-line data library.

[16] One reason for the popularity of the GRS test is that its weighting scheme leads to a test statistic whose distribution does not depend on unknown model parameters under the null hypothesis. As a result, researchers can conveniently refer to the standard F distribution table to perform hypothesis testing. Under our weighting scheme, the distribution of the test statistic under the null hypothesis will depend on model parameters such as the residual volatilities of individual assets. Fortunately, our bootstrap-based framework provides a convenient way to provide inference.

Table 1 presents the summary statistics on portfolios and factors. The 25 portfolios display the usual monotonic pattern in mean returns along the size and book-to-market dimension that we try to explain. The 14 risk factors generate sizable long-short strategy returns. Nine of the strategy returns generate t-ratios above 3.0 which is the level advocated by Harvey, Liu and Zhu (2016) that takes multiple testing into account. The correlation matrix suggests a clustering of some of the factors. There is a 'value' group consisting of book-to-market (*hml*), Fama and French (2015a)'s investment factor (*cma*), and Hou, Xue and Zhang (2015)'s investment factor (*ia*). There is a 'profitability' group consisting of Fama and French (2015a)'s profitability factor (*rmw*), Hou, Xue and Zhang (2015)'s profitability factor (*roe*), and Asness, Frazzini and Pedersen (2013)'s quality minus junk factor (*qmj*). For example, *cma* and *ia* have a correlation of 0.90, and *rmw* and *qmj* have a correlation of 0.76. These high levels of correlations might make it difficult to distinguish the factors within each of groups.

We use the aforementioned test statistics to capture the cross-sectional goodness-of-fit of a regression model. In addition, we also include the standard GRS test statistic. However, our othogonalization design does not guarantee that the GRS test statistic of the baseline model stays the same as the test statistic when we add an othogonalized factor to the model. The reason is that, while the othogonalized factor by construction has zero impact on the cross-section of expected returns, it may still affect the residual covariance matrix. Since the GRS statistic uses the residual covariance matrix to weight the regression intercepts, it changes as the estimate for the covariance matrix. We think the GRS statistic is not appropriate in our framework as its use of the residual covariance matrix to weight the regression intercepts is no longer optimal and may distort the comparison between candidate models. Indeed, for two models that generate the same regression intercepts, the GRS test is biased towards the model that explains a smaller fraction of variance in returns in time-series regressions. To avoid this bias, we focus on the two metrics previously defined that do not rely on a model-based weighting matrix. Again, we postpone a more detailed discussion of the GRS test to later sections.

We start by testing whether any of the 14 factors is individually significant in explaining the cross-section of expected returns. Panel A in Table 2 presents the results. The market factor appears to be the best among the candidate factors. It reduces the mean scaled absolute intercept by 61%, much higher than what the other factors deliver.

To evaluate the significance of the market factor, we follow our method and orthogonalize the 14 factors so they have a zero impact on the cross-section of expected returns in-sample. We bootstrap to obtain the empirical distributions of the individual test statistics. We then evaluate the realized test statistics against these empirical distributions to provide p-values. As shown in Panel A of Table 2, the bootstrapped 5th percentile of SI_{ew}^{m} for the market factor is -0.340. The interpretation is that bootstrapping under the null, i.e., the market factor has no ability to explain the cross-section, produces a distribution of increments to the intercept. At the 5th percentile, there is a percentage reduction in the mean scaled intercept of 34%. The actual factor reduces the mean scaled intercept by more than the 5th percentile, 34%, and we declare it significant. More precisely, by evaluating the 61% reduction against the empirical distribution of SI_{ew}^{m} for the market factor alone, the single-test p-value for the market factor is 0.002.

We can also bootstrap to obtain the empirical distribution of the minimum statistic. In particular, following the bootstrap procedure in Section 2, we resample the time periods. For each bootstrapped sample, we first obtain the test statistic for each of the 14 orthogonalized factors and then record the minimum test statistic across all 14 statistics. The minimum statistic is the the largest intercept reduction among the 14 factors. Since all factors are orthogonalized and therefore have no impact on the cross-section of expected returns, the minimum statistic shows what the largest intercept reduction can be just by chance and therefore controls for multiple testing. It is important that all 14 test statistics are based on the same bootstrapped sample as this controls for test correlations, as emphasized by Fama and French (2010). Lastly, we compare the realized minimum statistic with the bootstrapped distribution of the minimum statistic to provide p-values.

Panel A of Table 2 shows the results on multiple testing as well. In particular, the bootstrapped 5th percentile of SI_{ew}^{m} for the minimum statistic is -0.368. By evaluating the 61% reduction against the empirical distribution of the minimum statistic for multiple testing, the p-value is 0.003. Therefore, the multiple-test p-value is below the 5% cutoff. We therefore also declare the market factor significant from a multiple testing perspective. Across the two metrics we consider, the market factor is the dominating factor and is significant at 5% level, both from a single-test and a multiple-test perspective.

One interesting observation based on Table 2 is that the best factor that is selected may not be the one with the lowest single test p-value. For instance, in Panel A of Table 2 and for SI_{ew}^{m}, the market factor is the first factor that we select despite a lower single test p-value for *civ*. On the surface, this happens because the minimum test statistic picks the factor that has the lowest SI_{ew}^{m} (i.e., highest percentage reduction in the mean scaled absolute intercept), not its p-value. As a result, the market factor, which has a lower SI_{ew}^{m}, is favored over *civ*.

On a deeper level, should we use a minimum test statistic that depends on the p-values instead of the levels of the SI_{ew}^{m}'s? We think not. The use of SI_{ew}^{m} allows us to put weight on both the economic as well as statistical significance. This is especially important for our sequential selection procedure that incrementally identifies the group of true factors. We give a higher priority to a factor that has a large reduction in absolute intercept while passing a certain statistical hurdle than a factor that has a tiny reduction in absolute intercept but having a very small p-value.[17]

[17] Notice that a different scaling of a factor (i.e., long-short portfolio return) to alter the volatility will not change the test statistics or their p-values. This is because we run time-series regressions on the factors. Factor loadings adjust for different scalings. For example, when *mkt* is used as the

After the market factor is declared significant, we continue to identify the second risk factor. This time, *cma* has a multiple testing p-value of 0.001 under SI^{m}_{ew} and less than 0.001 under SI^{med}_{ew}, and is therefore declared significant. Notice that the performance of *hml* is close to that of *cma*. This is not surprising given that *cma* and *hml* are highly correlated (correlation coefficient = 0.71).

After *cma* is identified and included in the baseline model, we continue to search for the third factor. This time, *gp* and *smb* are the best performing factors among the remaining factors under SI^{m}_{ew} and SI^{med}_{ew}, respectively. Overall, across the two test statistics, *smb* seems to be a better performing factor compared to *gp* as it is close to *gp* under SI^{m}_{ew} and a lot better than *gp* under SI^{med}_{ew}. Nonetheless, neither *smb* nor *gp* is significant under multiple testing. We therefore terminate the search and conclude with a two-factor model, i.e., $mkt + cma$.

Overall, our results using equally weighted scaled regression intercepts confirm the idea that *mkt* and *cma*, a factor that is closely related to *hml*, are helpful in explaining the cross-section of returns of Fama-French 25 portfolios. This is not surprising as *hml* and Fama-French 25 portfolios use the same characteristics to sort the cross-section of stocks. What is interesting in our results is that *cma* survives after *mkt* is included. *smb* does not. This suggests that either *smb* is not a true risk factor or the Fama-French 25 portfolios have limited power to identify *smb* as a true risk factor. As we shall see later, the latter explanation seems more plausible.

Our findings seem to be at odds with past studies that also use the Fama-French 25 portfolios to test the market factor. Most of these studies rely on the two-stage Fama-MacBeth regression and find that slope estimates from the second stage regressions are not statistically different from zero. Hence, the market factor is not priced. If one plots the estimated portfolio average returns against the actual average returns, one will see a flat line instead of a 45-degree line as one would expect to see if CAPM holds. In our framework, the market factor is highly significant. Indeed, based on Panel A of 2, the market factor single-handedly reduces the scaled absolute regression intercept by about 60%.

The difference in inference between our approach and past studies stems from the difference in the test design. The test method used by past studies implicitly assumes that CAPM is true, that is, a single-factor model that includes the market factor is the true underlying factor model. However, this assumption is unlikely to hold for the Fama-French 25 portfolios as there may exist other risk factors (e.g., *smb* and *hml*). Suppose the underlying true factor model is a two-factor model that includes the market factor and another factor X. When we run Fama-MacBeth regressions for the market factor, because of the omitted factor X, the relationship between market betas and expected portfolio returns are non-monotonic. As a result, the estimate for the market risk premium will be biased. In fact, depending on what the true model

factor, suppose we have a beta estimate of 1.0 for a certain asset. When $2 \times mkt$ is used, the beta estimate will drop to 0.5, offsetting the scaling on *mkt*. Meanwhile, neither the regression intercept nor its significance will be affected by the scaling.

is, the estimate market risk premium can go anywhere from positive to negative. Harvey and Liu (2016) discuss the bias in Fama-MacBeth regressions when there is model misspecification for the underling factor model. Compared to Fama-MacBeth regressions, our approach is more robust to model misspecification.

While our results based on Fama-French 25 portfolios are interesting, we are reluctant to offer any deeper interpretation given the main drawback of the portfolio approach: tests based on characteristics-sorted portfolios are likely to be biased towards factors that are constructed using the same characteristics. In the next section, we apply our method to individual stocks and hope to provide an unbiased assessment of the 14 risk factors.

Table 1: **Summary Statistics, January 1968 - December 2012**

Summary statistics on portfolios and factors. We report the mean annual returns for Fama-French size and book-to-market sorted 25 portfolios and the five risk factors in Fama and French (2015a) (i.e., excess market return (*mkt*), size (*smb*), book-to-market (*hml*), profitability (*rmw*), and investment (*cma*)), betting against beta (*bab*) in Frazzini and Pedersen (2014), gross profitability (*gp*) in Novy-Marx (2013), Pastor and Stambaugh liquidity (*psl*) in Pastor and Stambaugh (2003), momentum (*mom*) in Carhart (1997), quality minus junk (*qmj*) in Asness, Frazzini and Pedersen (2013), investment (*ia*) and profitability (*roe*) in Hou, Xue and Zhang (2015), co-skewness (*skew*) in Harvey and Siddique (2000), and common idiosyncratic volatility (*civ*) in Herskovic, Kelly, Lustig and Van Nieuwerburgh (2014). We also report the correlation matrix for factor returns. The sample period is from January 1968 to December 2012.

Panel A: Portfolio Returns

	Low	2	3	4	High
Small	0.009	0.078	0.085	0.106	0.120
2	0.039	0.074	0.095	0.101	0.108
3	0.047	0.082	0.082	0.093	0.119
4	0.062	0.061	0.077	0.087	0.090
Big	0.046	0.061	0.053	0.059	0.069

Panel B.1: Factor Returns

	mkt	*smb*	*hml*	*mom*	*skew*	*psl*	*roe*	*ia*	*qmj*	*bab*	*gp*	*cma*	*rmw*	*civ*
Mean	0.052	0.022	0.048	0.081	0.024	0.055	0.068	0.057	0.048	0.105	0.039	0.047	0.033	0.060
t-stat	[2.17]	[1.32]	[3.08]	[3.54]	[1.84]	[2.99]	[5.09]	[5.76]	[3.74]	[5.98]	[3.24]	[4.44]	[2.92]	[3.48]

Panel B.2: Factor Correlation Matrix

	mkt	*smb*	*hml*	*mom*	*skew*	*psl*	*roe*	*ia*	*qmj*	*bab*	*gp*	*cma*	*rmw*	*civ*
mkt	1.00													
smb	0.30	1.00												
hml	-0.32	-0.24	1.00											
mom	-0.14	-0.03	-0.15	1.00										
skew	-0.02	-0.05	0.23	0.03	1.00									
psl	-0.05	-0.04	0.03	-0.03	0.10	1.00								
roe	-0.19	-0.39	-0.11	0.51	0.19	-0.06	1.00							
ia	-0.39	-0.26	**0.69**	0.04	0.15	0.02	0.04	1.00						
qmj	-0.54	-0.54	0.02	0.26	0.13	0.03	**0.68**	0.15	1.00					
bab	-0.09	-0.07	0.40	0.18	0.24	0.06	0.25	0.35	0.19	1.00				
gp	0.08	0.06	-0.34	0.01	-0.01	-0.03	0.34	-0.26	0.45	-0.11	1.00			
cma	-0.41	-0.16	**0.71**	0.01	0.05	0.03	-0.10	**0.90**	0.07	0.32	-0.34	1.00		
rmw	-0.21	-0.42	0.11	0.10	0.27	0.03	**0.68**	0.05	**0.76**	0.26	0.49	-0.08	1.00	
civ	0.17	0.27	0.13	-0.18	0.04	0.05	-0.26	-0.00	-0.28	0.11	-0.00	0.04	-0.10	1.00

Table 2: **Portfolios as Test Assets**

Test results on 14 risk factors using Fama-French size and book-to-market sorted 25 portfolios. (See Table 1 for the definitions of risk factors.). The baseline model refers to the model that includes the pre-selected risk factors. We focus on the panel regression model described in Section 2.2. The two metrics (i.e., SI^{m}_{ew} and SI^{med}_{ew}), which measure the difference in equally weighted scaled mean/median absolute regression intercepts, are defined in Section 3.2. GRS reports the Gibbons, Ross and Shanken (1989) test statistic.

Panel A: Baseline = No factor								Panel B: Baseline = ***mkt***					
		single test			single test				single test			single test	
Factor	SI^{m}_{ew}	$5th$-percentile	p-value	SI^{med}_{ew}	5th-percentile	p-value	GRS	SI^{m}_{ew}	5th-percentile	p-value	SI^{med}_{ew}	5th-percentile	p-value
mkt	**-0.607**	[-0.340]	(0.002)	**-0.672**	[-0.333]	(0.000)	4.290						
smb	-0.209	[-0.243]	(0.072)	-0.108	[-0.257]	(0.215)	4.402	-0.068	[-0.174]	(0.251)	-0.007	[-0.211]	(0.481)
hml	0.189	[-0.100]	(0.999)	0.230	[-0.110]	(0.997)	4.050	-0.434	[-0.260]	(0.000)	-0.397	[-0.302]	(0.009)
mom	0.224	[-0.108]	(0.998)	0.256	[-0.120]	(0.998)	4.302	0.218	[-0.071]	(0.999)	0.210	[-0.113]	(0.985)
skew	-0.014	[-0.040]	(0.195)	0.007	[-0.053]	(0.731)	4.454	-0.116	[-0.085]	(0.025)	-0.134	[-0.117]	(0.039)
psl	0.043	[-0.038]	(0.946)	0.054	[-0.044]	(0.952)	4.286	-0.038	[-0.034]	(0.040)	-0.135	[-0.055]	(0.004)
roe	0.504	[-0.150]	(1.000)	0.470	[-0.144]	(0.999)	4.919	0.375	[-0.106]	(1.000)	0.366	[-0.137]	(0.998)
ia	0.607	[-0.157]	(1.000)	0.637	[-0.164]	(1.000)	4.553	-0.318	[-0.168]	(0.001)	-0.262	[-0.206]	(0.012)
qmj	0.820	[-0.275]	(0.990)	0.806	[-0.273]	(0.983)	5.594	0.560	[-0.134]	(1.000)	0.898	[-0.173]	(1.000)
bab	0.036	[-0.042]	(0.952)	0.030	[-0.055]	(0.908)	**3.718**	-0.442	[-0.154]	(0.000)	-0.447	[-0.179]	(0.000)
gp	-0.042	[-0.037]	(0.039)	0.026	[-0.049]	(0.892)	4.096	0.202	[-0.087]	(1.000)	0.200	[-0.128]	(0.988)
cma	0.450	[-0.143]	(1.000)	0.464	[-0.155]	(0.999)	4.238	**-0.476**	[-0.196]	(0.000)	**-0.500**	[-0.225]	(0.000)
rmw	0.268	[-0.126]	(0.991)	0.273	[-0.124]	(0.987)	4.325	0.055	[-0.056]	(0.991)	0.132	[-0.119]	(0.962)
civ	-0.281	[-0.140]	(0.000)	-0.283	[-0.141]	(0.002)	4.132	-0.219	[-0.094]	(0.001)	-0.099	[-0.128]	(0.088)
		multiple test			multiple test				multiple test			multiple test	
	min	**[-0.368]**	(**0.003**)		**[-0.373]**	(**0.000**)		*min*	**[-0.289]**	(**0.001**)		**[-0.342]**	(**0.000**)

Panel C: Baseline = ***mkt*** + ***cma***						
		single test			single test	
Factor	SI^{m}_{ew}	$5th$-percentile	p-value	SI^{med}_{ew}	5th-percentile	p-value
mkt						
smb	-0.232	[-0.353]	(0.171)	**-0.295**	[-0.454]	(0.188)
hml	0.001	[-0.136]	(0.657)	0.013	[-0.230]	(0.615)
mom	0.091	[-0.067]	(0.981)	0.115	[-0.139]	(0.930)
skew	0.005	[-0.058]	(0.654)	0.093	[-0.134]	(0.896)
psl	-0.027	[-0.028]	(0.054)	0.222	[-0.069]	(0.992)
roe	0.911	[-0.128]	(1.000)	1.271	[-0.228]	(1.000)
ia	0.382	[-0.106]	(1.000)	0.631	[-0.181]	(1.000)
qmj	1.381	[-0.153]	(1.000)	1.857	[-0.242]	(1.000)
bab	0.101	[-0.069]	(0.991)	0.080	[-0.153]	(0.880)
gp	**-0.260**	[-0.073]	(0.061)	-0.084	[-0.104]	(0.081)
cma						
rmw	0.561	[-0.119]	(1.000)	0.644	[-0.188]	(1.000)
civ	-0.160	[-0.100]	(0.013)	-0.214	[-0.211]	(0.049)
		multiple test			multiple test	
	min	**[-0.356]**	(**0.148**)		**[-0.464]**	(**0.253**)

3.4 Why We Abandon the GRS

The GRS test statistic is problematic in our context from a variety of perspectives. For instance, with *mkt* as the only factor in the baseline model and by adding the orthogonalized *smb* to the baseline model, the GRS is 6.039 (not shown in table), much larger than 4.290 in Panel A of Table 2, which is the GRS for the real data with *mkt* as the only factor. This means that by adding the orthogonalized *smb*, the GRS becomes much larger. By construction, the orthogonalized *smb* has no impact on the regression intercepts. The only way it can affect the GRS is through the error covariance matrix. Hence, the orthogonalized factor makes the GRS larger by reducing the error variance estimates. This insight also explains the discrepancy between SI^m_{ew} and the GRS in Panel A of Table 2: *mkt*, which implies a much smaller mean absolute intercept in the cross-section, has a larger GRS than *bab* as *mkt* absorbs a larger fraction of variance in returns in time-series regressions and thereby putting more weight on regression intercepts compared to *bab*.

The weighting in the GRS does not seem appropriate for model comparison when none of the candidate models is expected to be the true model, i.e., the true underlying factor model that fully explains the cross-section of expected returns. Between two models that imply the same time-series regression intercepts, it favors the model that explains *a smaller fraction* of variance in returns. This does not make sense. We choose to focus on our proposed metrics that do not depend on the error covariance matrix estimate.

The way that the GRS test uses the residual covariance matrix to scale regression intercepts is likely to become even more problematic when we use individual stocks as test assets. Given a large cross-section and a limited time-series, the residual covariance matrix will be poorly measured. To make things worse, this covariance matrix needs to be inverted to obtain the weights for intercepts. As a result, the GRS test is likely to be very unstable and potentially distorted when applied to individual stocks.[18]

Our findings about the GRS test resonate with a recent study by Fama and French (2015b). They find that the GRS test often implies unrealistically large short positions on certain assets, which does not make economic sense. To explain their findings, notice that the GRS test can be interpreted as the difference between the Sharpe ratio constructed using both the left-hand side assets and the right-hand side factors (call this Sharpe ratio SR_1) and the Sharpe ratio using only the right-hand side factors (call this Sharpe ratio SR_2). A rejection is found if SR_1 is significantly larger than SR_2. What Fama and French (2015b) find is that certain left-hand side assets need to take extreme short positions in order to achieve SR_1. By imposing short sale constraints, SR_1 is often much smaller, reducing the contribution of the left-hand side assets to the tangency portfolio formed using the right-hand side factors alone. This causes us to question the economic usefulness of the GRS test.

[18] See Gagliardini et al. (2014) for a similar argument.

Our framework provides an economically meaningful approach to evaluate the incremental contribution of SR_1 over SR_2. In a panel regression model, the regression intercepts capture mispricing for the assets in the cross-section. An investor who is trying to exploit this mispricing will be long assets that have positive intercepts and short assets that have negative intercepts. By taking equally-weighted positions in the cross-section, the abnormal return for her portfolio (that is, returns with factor risks purged out) equals the equally weighted absolute intercepts plus a residual component that is the equally weighted average of the regression residuals. When we have a large cross-section — which will be the case when we use individual stocks as test assets — the residual component will be small. Therefore, the equally weighted absolute intercepts captures the abnormal return earned by an investor that tries to exploit the mispricing of the cross-section of assets relative to a factor model.

While we explore the equally weighted absolute intercepts in our on-line appendix, an obvious extension is to take the estimation uncertainty into account by using the standard errors to scale the regression intercepts. This motivates our test statistics (e.g., SI_{ew}^{m}) that are based on the scaled intercepts. As we show in the on-line appendix, our test statistics have substantially higher test power compared to tests that are based on the original intercepts. Finally, an average investor in the economy will invest in proportion to the market capitalizations of assets. Hence, a value-weighted metric may better reflect the economic significance of asset mispricing in the cross-section. We explore this metric in the next section when we use individual stocks as test assets.

3.5 Results: Individual Stocks as Test Assets

Instead of characteristics-sorted portfolios, can we use individual stocks to provide inference? Conventional wisdom says no. Indeed, Black, Jensen and Scholes (1972) and Fama and MacBeth (1973) argue that individual stocks are too noisy to serve as test assets. The GRS test also prohibits the use of individual stocks as the inversion of the large variance-covariance matrix of the return residuals is problematic (see Fama and French, 2010, Gagliardini et al., 2014). Subsequent researchers follow these suggestions and use popular portfolios, in particular the Fama-French 25 portfolios, to test risk factors.

A counter argument is that the use of portfolios may introduce bias and inefficiency in asset pricing tests. Avramov and Chordia (2006) show that the asset pricing implications differ a lot when we use single securities instead of portfolios. Ang, Liu, and Schwarz (2010) argue that the larger dispersion in beta by using individual stocks can potentially enhance the power of the test.[19] Lewellen, Nagel, and Shanken (2010)

[19] Estimation uncertainty for the estimated betas makes the gain in power of the method in Ang, Liu and Schwarz (2010) vanish asymptotically. However, in finite samples, the gain in power by using a larger set of test assets may be non-negligible.

suggest the use of a large number of assets instead of a small number of portfolios to judge model performance.

We believe that the single dominating reason for considering individual stocks is that they provide an unbiased test of risk factors. Tests based on characteristics-sorted portfolios are likely to be biased towards identifying the risk factors that are constructed using the same set of test portfolios. The use of individual stocks guards against the data-snooping bias induced by portfolio-based asset pricing tests, as shown in Lo and MacKinlay (1990). Additionally, if we rephrase the argument in Black, Jensen and Scholes (1972) and Fama and MacBeth (1973) as saying that the high level of noise in individual stocks renders most tests inefficient, then the problem is not about individual stocks themselves, but about the lack of statistical tests that help alleviate the noise in stocks. Our paper provides such a test. In particular, our framework allows us to make inference on test statistics that take stock volatilities and market values into account. It therefore alleviates the noise issue for individual stocks by downweighing the impact of small and noisy stocks. We also use economically motivated test statistics that do not rely on the estimation of the variance-covariance matrix, thereby circumventing the difficulty in using the GRS test.

More specifically, our framework provides a way to overcome many of the challenges in the use of individual stocks. First, we use an unbalanced panel of equity returns, which creates difficulty for standard panel regression models. Our bootstrap-based approach, by allowing the size of the bootstrapped sample to be proportional to the original sample, takes sampling uncertainty (i.e., between two samples drawn from the same population, the inference based on the shorter sample is less accurate than the inference based on the longer sample). Second, there is a large amount of heterogeneity in the cross-section of stocks in terms of firm size, volatility, etc. Our method allows different weighting schemes that take various sources of heterogeneity into account. Finally, given the existence of extreme observations in the returns of individual stocks, our method allows the use of test statistics that are robust to outliers (e.g., median).

3.5.1 A Simulation Study

We first run a simulation study to see whether our tests have statistical power in correctly identifying a true risk factor. This is important given the concern that the high level of noise in individual stock returns might render our test powerless. We do not pursue a full-blown simulation study that investigates every aspect of our procedure but instead focus on the selection of a candidate risk factor that provides additional information to the market factor in explaining the cross-section of expected returns. This is motivated by the fact that the market factor is the single dominating factor that is always selected first in our previous study based on portfolios. It is more interesting to examine the next factor that enters our factor list.

Two features distinguish our simulation study from standard simulation frameworks. First, we take the cross-sectional distribution of factor loadings in the data as given. Second, we bootstrap the realized return residuals to construct the simulated panel of returns. Compared to standard simulation studies that assume a parametric distribution for the factor loadings and/or return residuals, our method brings the simulated data closer to the actual data, and therefore provides a more realistic assessment of test power. Appendix B describes our simulation study in detail.

To benchmark our results against existing methods, we consider two popular beta sorts. The first method is unconditional beta sorts that first sorts stocks based on their unconditional univariate factor loadings estimated over the entire sample, and then forms the long-short portfolio by having a long position in stocks that are in the top beta decile and a short position in stocks that are in the bottom beta decile.[20] The t-statistic of the portfolios returns is used to test the significance of the candidate factor. The second method involves time-varying beta sorts. We estimate factor loadings based on a five-year rolling window and construct a long-short portfolio that we hold out-of-sample for one year. We again use the t-statistic of the portfolio returns to test the significance of the candidate factor.

There are several takeaways from our simulation study. First, compared to the beta sorts, our bootstrap-based tests are more powerful. In particular, when the factor risk premium is similar to what we see in the real data, the power of our tests based on the t-statistics is on average (across different factors) about 10% higher than that based on beta sorts. Additionally, when the number of time periods is about the same as in the real data, the power of our tests is well above 70% across different risk factors. Hence, our tests have power in an absolute sense as well.

Second, we are not necessarily losing test power by considering individual stocks. In particular, we redo our exercise using Fama-French 25 portfolios and find that the average performance of our tests based on portfolios is similar to that of our tests based on individual stocks. The key assumption for our simulation study is that a two-factor model is the true underlying factor model. In reality, the Fama-French 25 portfolios follow a tight factor structure and therefore are likely to favor factors that are correlated with the Fama-French factors (see Lewellen, Nagel, and Shanken, 2010). In contrast, individual stocks can potentially provide unbiased and significantly richer information to identify the true factor model.

3.5.2 Value Weighted Test Statistic

In addition to the previously mentioned test statistics that rely on equally weighted scaled regression intercepts, we consider a value weighted version of them. Value weighting makes economic sense. For two stocks that generate the same regression

[20] Sorts based on multivariate factor loadings yield similar results as the candidate factors are not highly correlated with the market factor.

intercept, the mispricing of the factor model should be more significant economically for the stock that has a higher market value. Our value weighted test statistic therefore uses market values to weight the cross-section of scaled intercepts. In particular, let $\{me_{i,t}\}_{t=1}^{T}$ be the time-series of market equity for stock i, and let $ME_t = \sum_{i=1}^{N} me_{i,t}$ be the aggregate market equity at time t. The test statistic is given by

$$SI_{vw} \equiv \frac{(\sum_{t=1}^{T}\sum_{i=1}^{N}\frac{me_{i,t}}{ME_t}\times|a_i^g|/s_i^b)/T - (\sum_{t=1}^{T}\sum_{i=1}^{N}\frac{me_{i,t}}{ME_t}\times|a_i^b|/s_i^b)/T}{(\sum_{t=1}^{T}\sum_{i=1}^{N}\frac{me_{i,t}}{ME_t}\times|a_i^b|/s_i^b)/T},$$

where vw denotes value weighting. SI_{vw} calculates the percentage difference in the time averaged value-weighted level of mispricing between the augmented model and the baseline model. Our value weighted test statistic takes the time variation in market value into account.

3.5.3 Test Results with Individual Stocks

Table 3 (equal weights) and 4 (value weights) present the results based on individual stocks. Under both weighting schemes and consistent with the results based on the Fama-French 25 portfolios, the market factor is always the first factor selected and is highly significant.

The fact that we always declare the market factor as significant in our framework is not a trivial empirical finding. Although the market factor has a strong theoretical motivation and is probably the first risk factor tested (see Black, Jensen and Scholes, 1972), different papers, by using different testing methods and test assets, often arrive at conflicting conclusions. Therefore, there is no consensus as to whether the market factor is a valid risk factor empirically. Nonetheless, perhaps due to its intuitive appeal and theoretical relevance, most routinely use it for risk adjustment and cost of capital calculation. More recently, by also using individual stocks as test assets and using the Fama-MacBeth cross-sectional test, Jegadeesh and Noh (2014) reject the market factor as a risk factor. Chordia, Goyal, and Shanken (2015) only find weak support for the market factor.

In contrast, our results suggest that the market factor is by far the dominant risk factor in explaining the cross-sectional variation in expected returns, both for well-diversified portfolios and individual stocks. We believe this is due to our use of the panel regression test.

Fama (2015) summarizes the difference between the Fama-MacBeth approach and the panel regression approach. The panel regression approach essentially assumes that the factor risk premium is given by its in-sample estimate, and tries to evaluate what percentage of the expected return is explained by the risk exposure to the factor (i.e., beta times risk premium). The GRS test is one example of a panel regression test.

It focuses on the extreme case where the percentage of the expected return explained by the factor model has to be 100%, that is, the factor model is correctly specified and it is the underlying true factor model. Our test is less extreme than the GRS test in that this percentage does not have to be 100%. A factor could be declared true as long as it explains a significant amount of expected returns for a given cross-section of assets. Indeed, in Table 3 and 4, we show that the market factor single-handedly explains 44% and 20% of expected returns under value weighting and equal weighting, respectively. These numbers are large from an economic perspective.

More precisely, as shown in Harvey and Liu (2016), when there is uncertainty around the underlying true factor model and the factor model being tested is likely misspecified, cross-sectional regressions (e.g., Fama-MacBeth) sometimes generate risk premium estimates that are severely biased and lead to size distortions for hypothesis tests. In contrast, panel regression models provide an attractive setting for testing risk factors.

In Panel B of Table 3, after the market factor is identified, *smb* is the best factor among the remaining factors. The percentage reduction in scaled absolute intercept is 4.1% under SI_{ew}^{m} and 6.2% under SI_{ew}^{med}. The corresponding multiple testing p-values are 0.071 and 0.039. Given that SI_{ew}^{med} is more robust to outliers among the cross-section of scaled intercepts, we put more weight on SI_{ew}^{med} and therefore declare *smb* significant.

In Panel C of Table 3, after both the market factor and *smb* are included in the baseline model, *hml* is the best factor among the remaining factors, reducing the scaled absolute intercept by 1.7% under SI_{ew}^{m} and 4.0% under SI_{ew}^{med}. It also has a significant multiple testing p-value under SI_{ew}^{med} (i.e., 1.8%). We therefore declare it significant and include it in the baseline model. After *mkt*, *smb*, and *hml* are included in the baseline model, none of the remaining factors is significant, as shown in Panel D of Table 3. We therefore terminate our testing procedure and identify the true factor model as *mkt*+*smb*+*hml* using the equally weighted test statistic.

Our results under equal weighting contribute to the literature by identifying *smb* and *hml*, perhaps the two most well-known anomaly variables, as significant risk factors that drive the cross-section of individual stock returns. Our approach distinguishes itself from existing studies in several aspects. First, our method allows us to explore the rich information in the large cross-section of individual stocks while controlling for the noise in individual stock returns and being robust to extreme observations. Second, we do not seek to find the true factor model that completely explains the cross-sectional variation in expected returns. Neither do we impose the existence of such a model in the construction of our test. We try to evaluate the incremental contribution of a factor and sequentially select the list of significant factors. We believe these two features of our model make it advantageous over existing methods and allow us to successfully detect true risk factors.

Contrary to our results, Chordia, Goyal, and Shanken (2015) and Jegadeesh and Noh (2014) also use individual stocks and find that several popular factors (e.g., *smb* and *hml*) that are potentially risk factors do not seem to be priced. Both papers rely on the Fama-MacBeth regression (corrected for errors-in-variables bias) and use OLS in the second stage regression. This method effectively equal weights the cross-section of stocks and is, therefore, consistent with our weighting scheme in Table 3. However, as shown in Harvey and Liu (2016), the key assumption for the inference of the Fama-MacBeth regression to work is that the factor model tested is the true underlying factor model that fully explains the returns of test assets. When there is model misspecification, which is likely to be the case for individual stocks, the Fama-MacBeth approach can be severely biased. Our method is more robust to model misspecification.

Table 3: **Individual Stocks as Test Assets, Equally Weighted Scaled Intercepts**

Test results on 14 risk factors using equally weighted individual stocks. (See Table 1 for the definitions of risk factors). We use individual stocks from CRSP that cover the 1968– 2012 period to test 14 risk factors. A stock needs to have at least 36 monthly observations (either in the original or the bootstrapped sample) to enter our tests. The baseline model refers to the model that includes the pre-selected risk factors. We focus on the panel regression model described in Section 2.2. The two metrics (i.e., SI^{m}_{ew} and SI^{med}_{ew}), which measure the difference in equally weighted scaled mean/median absolute regression intercept, are defined in Section 3.2.

	Panel A: Baseline = No factor						Panel B: Baseline = ***mkt***					
		single test			single test			single test			single test	
Factor	SI^{m}_{ew}	$5th$-percentile	p-value	SI^{med}_{ew}	5th-percentile	p-value	SI^{m}_{ew}	5th-percentile	p-value	SI^{med}_{ew}	5th-percentile	p-value
mkt	**-0.192**	[-0.093]	(0.003)	**-0.206**	[-0.095]	(0.001)						
smb	-0.081	[-0.081]	(0.056)	-0.109	[-0.117]	(0.061)	**-0.041**	[-0.045]	(0.063)	**-0.062**	[-0.052]	(0.032)
hml	0.088	[-0.022]	(0.983)	0.108	[-0.029]	(1.000)	-0.021	[-0.030]	(0.131)	-0.047	[-0.028]	(0.014)
mom	0.091	[-0.034]	(1.000)	0.110	[-0.044]	(1.000)	0.070	[-0.007]	(1.000)	0.089	[-0.012]	(1.000)
skew	-0.008	[-0.031]	(0.278)	-0.002	[-0.034]	(0.478)	-0.004	[-0.009]	(0.167)	-0.003	[-0.013]	(0.319)
psl	0.011	[-0.019]	(0.920)	0.002	[-0.030]	(0.682)	0.001	[-0.004]	(0.409)	-0.003	[-0.012]	(0.237)
roe	0.163	[-0.042]	(0.951)	0.187	[-0.064]	(1.000)	0.142	[-0.019]	(1.000)	0.180	[-0.029]	(1.000)
ia	0.264	[-0.040]	(1.000)	0.291	[-0.048]	(1.000)	0.027	[-0.009]	(0.968)	0.015	[-0.015]	(0.934)
qmj	0.316	[-0.072]	(0.995)	0.358	[-0.090]	(0.998)	0.149	[-0.024]	(0.972)	0.193	[-0.029]	(0.973)
bab	-0.006	[-0.039]	(0.594)	-0.049	[-0.050]	(0.107)	0.018	[-0.010]	(0.983)	-0.014	[-0.017]	(0.181)
gp	0.017	[-0.008]	(0.529)	0.030	[-0.007]	(0.727)	0.023	[-0.005]	(0.961)	0.017	[-0.007]	(0.790)
cma	0.176	[-0.034]	(1.000)	0.199	[-0.035]	(1.000)	-0.012	[-0.013]	(0.057)	-0.031	[-0.019]	(0.027)
rmw	0.116	[-0.011]	(0.986)	0.137	[-0.017]	(0.994)	0.040	[-0.014]	(1.000)	0.048	[-0.020]	(0.975)
civ	-0.096	[-0.044]	(0.023)	-0.130	[-0.062]	(0.031)	-0.018	[-0.018]	(0.052)	-0.049	[-0.030]	(0.021)
		multiple test			multiple test			multiple test			multiple test	
	min	**[-0.109]**	**(0.004)**		**[-0.147]**	**(0.002)**	*min*	**[-0.045]**	**(0.071)**		**[-0.057]**	**(0.039)**

	Panel C: Baseline = ***mkt+smb***						Panel D: Baseline = ***mkt*** + ***smb+hml***					
		single test			single test			single test			single test	
Factor	SI^{m}_{ew}	$5th$-percentile	p-value	SI^{med}_{ew}	5th-percentile	p-value	SI^{m}_{ew}	5th-percentile	p-value	SI^{med}_{ew}	5th-percentile	p-value
mkt												
smb												
hml	**-0.017**	[-0.020]	(0.061)	**-0.040**	[-0.025]	(0.011)						
mom	0.055	[-0.004]	(1.000)	0.076	[-0.010]	(1.000)	0.026	[-0.005]	(1.000)	0.046	[-0.013]	(1.000)
skew	-0.013	[-0.010]	(0.029)	-0.015	[-0.013]	(0.036)	**0.006**	[-0.002]	(0.463)	**-0.001**	[-0.005]	(0.313)
psl	0.011	[-0.002]	(0.945)	0.016	[-0.005]	(0.970)	0.010	[-0.002]	(0.937)	0.007	[-0.005]	(0.771)
roe	0.058	[-0.006]	(0.987)	0.074	[-0.010]	(0.967)	0.072	[-0.004]	(1.000)	0.080	[-0.011]	(1.000)
ia	0.020	[-0.012]	(0.967)	0.008	[-0.013]	(0.719)	0.038	[-0.004]	(0.975)	0.051	[-0.008]	(1.000)
qmj	0.052	[-0.007]	(0.976)	0.061	[-0.008]	(0.998)	0.128	[-0.004]	(0.982)	0.137	[-0.006]	(0.971)
bab	0.016	[-0.010]	(0.896)	-0.014	[-0.013]	(0.043)	0.045	[-0.003]	(0.989)	0.040	[-0.007]	(0.954)
gp	0.022	[-0.003]	(0.972)	0.020	[-0.009]	(0.951)	0.059	[-0.001]	(0.992)	0.055	[-0.006]	(0.984)
cma	0.001	[-0.009]	(0.341)	-0.009	[-0.012]	(0.137)	0.022	[-0.002]	(0.980)	0.023	[-0.005]	(0.967)
rmw	-0.009	[-0.019]	(0.147)	-0.016	[-0.020]	(0.086)	0.036	[-0.002]	(1.000)	0.043	[-0.006]	(0.992)
civ	0.014	[-0.009]	(0.981)	0.003	[-0.019]	(0.615)	0.015	[-0.008]	(0.991)	0.016	[-0.015]	(0.981)
		multiple test			multiple test			multiple test			multiple test	
	min	**[-0.022]**	**(0.122)**		**[-0.027]**	**(0.018)**	*min*	**-0.011**	**(0.997)**		**-0.017**	**(0.932)**

Table 4: **Individual Stocks as Test Assets, Value Weighted Scaled Intercepts**

Test results on 14 risk factors using value weighted individual stocks. (See Table 1 for the definitions of risk factors). We use individual stocks from CRSP that cover the 1968– 2012 period to test 14 risk factors. A stock needs to have at least 36 monthly observations (either in the original or the bootstrapped sample) to enter our tests. The baseline model refers to the model that includes the pre-selected risk factors. We focus on the panel regression model described in Section 2.2. The metric (i.e., SI_{vw}), which measures the difference in value weighted scaled absolute regression intercept, is defined in Section 3.5.2.

Panel A: Baseline = No factor				Panel B: Baseline = ***mkt***			Panel C: Baseline = ***mkt*+*qmj***		
		single test			single test			single test	
Factor	SI_{vw}	$5th$-percentile	p-value	SI_{vw}	5th-percentile	p-value	SI_{vw}	5th-percentile	p-value
mkt	**-0.444**	[-0.258]	(0.000)						
smb	-0.059	[-0.054]	(0.041)	0.018	[-0.042]	(0.831)	0.076	[-0.032]	(0.994)
hml	0.144	[-0.059]	(0.972)	-0.038	[-0.045]	(0.128)	-0.016	[-0.062]	(0.471)
mom	0.153	[-0.064]	(1.000)	0.130	[-0.012]	(1.000)	0.125	[-0.026]	(1.000)
skew	-0.027	[-0.052]	(0.158)	-0.044	[-0.033]	(0.029)	-0.020	[-0.025]	(0.088)
psl	0.035	[-0.023]	(0.970)	0.016	[-0.011]	(0.996)	0.034	[-0.028]	(0.991)
roe	0.105	[-0.043]	(0.993)	-0.079	[-0.043]	(0.021)	0.038	[-0.025]	(0.967)
ia	0.382	[-0.086]	(0.984)	-0.042	[-0.048]	[0.083]	0.078	[-0.047]	(0.935)
qmj	0.363	[-0.112]	(0.892)	**-0.149**	[-0.079]	(0.002)			
bab	-0.048	[-0.035]	(0.026)	-0.088	[-0.049]	(0.006)	**-0.026**	[-0.037]	(0.157)
gp	-0.082	[-0.038]	(0.009)	-0.037	[-0.043]	(0.073)	-0.022	[-0.046]	(0.242)
cma	0.314	[-0.107]	(0.982)	-0.052	[-0.034]	(0.028)	0.019	[-0.038]	(0.941)
rmw	0.045	[-0.014]	(0.942)	-0.146	[-0.066]	(0.019)	0.053	[-0.033]	(1.000)
civ	-0.115	[-0.062]	(0.002)	0.035	[-0.019]	(0.973)	-0.017	[-0.024]	(0.113)
		multiple test			multiple test			multiple test	
	min	**[-0.258]**	**(0.000)**		**[-0.083]**	**(0.004)**		**[-0.069]**	**(0.637)**

Compared to our results based on the Fama-French 25 portfolios, we seem to have power in detecting risk factors by using individual stocks. Indeed, when we use the Fama-French 25 portfolios, only *cma*, a factor closely related to *hml*, is tested as significant besides the market factor. When we use individual stocks, we uncover both *smb* and *hml*. This corroborates the evidence in our simulation study that we are not necessarily losing test power by using individual stocks.

The use of robust test statistics such as SI_{ew}^{med} seems important for tests based on individual stocks. Without SI_{ew}^{med} and by only using SI_{ew}^{m}, we would only be able to declare *smb* marginally significant and *hml* insignificant. Our bootstrap-based framework allows us to make statistical inference on robust test statistics that are otherwise difficult to evaluate under conventional testing frameworks.

Table 4 shows the results with value weighting. Under value weighting, it seems that many of the factors have a larger impact on the cross-section than under equal weighting. First, the economic magnitudes of the test statistics are much larger under value weighting than under equal weighting. For example, when the market factor is included in the baseline model, the reduction of the scaled absolute intercept is 44.4% under value weighting (Table 4) and about 20% under equal weighting (Table 3). Second, after the market factor is selected, more of the remaining candidate factors appear to be able to reduce the cross-section of mispricing under value weighting than under equal weighting. Taken together, our results suggest that many of the discovered factors play a bigger role as risk factors in explaining the expected returns for large stocks than for small stocks.

Under value weighting, the market factor is again the best performing factor. Its multiple testing p-value is less than 0.001, suggesting that the market factor is a highly significant risk factor. After the market factor is identified, the next best factor is *qmj*, which has a multiple testing p-value of 0.004. After both the market factor and *qmj* are identified and included in the baseline model, none of the remaining factors is significant. Indeed, the multiple testing p-value for the next best factor (i.e., *bab*) is 0.637. Therefore, under value weighting, we find a two-factor model that includes *mkt* and *qmj*.

When we sequentially build the factor model, the drop in statistical significance for the best available candidate factor is remarkable. What is equally impressive is the drop in economic significance. For example, the market factor reduces the value-weighted absolute scaled intercept by 44.4%. After the market factor is included in the baseline model, the incremental reduction by the second identified factor (*qmj*) is 14.9%. After both factors are included in the baseline model, the incremental reduction of the next best candidate (*bab*) is only 2.6%. This drop in economic significance gives us confidence in the final model we arrive at.

Although our test picks up *qmj* as the true risk factor, we want to stress that *qmj* is representative of a group of factors, that is, the *profitability group*. This group includes *qmj*, *rmw*, and *roe*. The three factors within the group are highly correlated and have similar performances in our regression test.[21]

Our identification of a profitability factor under value weighting makes economic sense. Papers that propose profitability factors often use theories of firm investment to motivate their findings (e.g., Fama and French, 2015, Hou et al., 2014). Intuitively, larger firms have fewer frictions and therefore can better engage in value maximization — the key assumption for investment theories to work. The fact that our test allows us to value weight the cross-section of intercepts demonstrates the flexibility of our approach. We are able to provide rigorous statistical inference on economically meaningful tests that are otherwise difficult to deal with in traditional testing frameworks.

Chordia, Goyal, and Shanken (2015) use a modified Fama-MacBeth approach that corrects the bias in the return-premium estimation and find weak support for *rmw* as a priced risk factor. They do not consider *qmj*. The support for the profitability factors (both *rmw* and *qmj*) is much stronger in our model than in Chordia, Goyal, Shanken. We think that both value weighting and our panel regression framework contribute to the significance of profitability factors as priced risk factors.

In our on-line appendix, we also perform our tests on the Fama-French 25 portfolios under value weighting. Interestingly, we find a a single-factor model that only includes the market factor. That is, none of the profitability factors (or other factors) is significant in explaining the returns of the Fama-French 25 portfolios. This is suggestive of a gain in power when using individual stocks as opposed to portfolios.

Our results using value weighting have important implications for the current practice of using portfolios as test assets in asset pricing tests. Average portfolio returns are disperse in the cross-section, which is good news for asset pricing tests as it can potentially increase test power. However, the cross-section is small. Indeed, the dispersion of returns of the Fama-French 25 portfolios is largely driven by a few portfolios that are dominated by small stocks. Under equal weighting, current asset pricing tests are likely to identify factors that can explain these few extreme portfolios. This is also consistent with it being relatively easy to data mine a factor that fits (accidentally) these extreme portfolios.[22] This also makes little economic sense as portfolios that cover small stocks are less important than portfolios that cover big stocks to an average investor that invests heavily in big stocks. Our approach provides a new way to take the market value of a portfolio into account when constructing an asset pricing test.

Taken as a whole, our results provide a stronger support for the market factor, conventional factors (i.e., *smb* and *hml*), and profitability factors that are discovered by the literature than concurrent papers that also look at individual stocks. We

[21] For example, *qmj* reduces the scaled absolute intercept by 14.9% and *rmw* is not far from being chosen with 14.6%.

[22] See Lewellen et al. (2010) for a similar argument.

believe this can be attributed to two features of our framework. First, we assume constant factor loadings for our panel regression approach. This may seem restrictive compared to the Fama-MacBeth approach that allows for time-varying factor loadings but can provide more stable parameter estimates. This is especially important for individual stocks. The reduction in estimation uncertainty for factor loadings is likely to outweigh the increase in bias induced by fixed factor loadings (see Section 4 for extensions of our framework to allow for time-varying factor loadings).

Second, running cross-sectional regressions as in the Fama-MacBeth approach is likely to be problematic for individual stocks as extreme observations in the cross-section are frequently observed. Trimming is an ill-advised practice as sometimes large observations provide important information for parameter estimates. In contrast, our panel regression framework focuses on the reduction in regression intercept or the t-statistic of intercept, both of which rely on the entire return time-series and are less affected by a single observation.

3.6 Robustness

In Appendix C, we consider various robustness checks of our results. First, to control for the impact of small stocks, we drop the bottom 10% of stocks in terms of market capitalization in each year and rerun our analysis. The results are similar to our current results that use the entire cross-section of stocks. Second, we explore the Fama-French 49 industry portfolios as an alternative to the Fama-French 25 portfolios. Under equal weighting, we are unable to identify either *smb* or *hml* as risk factors. Under value weighting, investment factors are still significant. Third, to allow for time-series dependence for both stock returns and factor returns, we use block bootstrap instead of independent bootstrap. Fourth, to mitigate the impact of infrequent trading for certain stocks, we consider lagged factor returns and test the combined impact of the original factors and their lags. Neither time-series dependence nor infrequent trading has a material impact on our results. For further details on robustness checks, see Appendix C.

4 Other issues

4.1 Time-varying Factor Loadings

Our application focuses on panel regressions with fixed factor loadings. Our setting is therefore analogous to the environment of the GRS test where asset returns are projected onto factor proxies with constant factor loadings. It is also related to the two-pass cross-sectional regression method with time-invariant factor loadings, see, e.g., Shanken (1992), Jagannathan and Wang (1998), Shanken and Zhou (2007), and Kan, Robotti, and Shanken (2013).

While unconditional models approximate the asset pricing environment in a simple fashion, the true model might be conditional. Therefore, it might seem reasonable to always use a conditional model when possible. This is not true. Even when the true model is conditional, estimation errors for conditional betas may outweigh the gain of correctly specifying the true model, rendering the inference less efficient than an unconditional model specification. See Ghysels (1998).

Having discussed the pros and cons of conditional and unconditional model specifications, we explore two extensions of our panel regression tests that can accommodate time-varying factor loadings. The first extension is to explicitly model the conditioning variables as functions of financial and macroeconomic variables, as in Shanken (1990), Ferson and Harvey (1991), and Lettau and Ludvigson (2001). This effectively introduces new factors that interact the original factors with financial and macroeconomic variables. Our method follows by testing these new factors in addition to the original factors.

The second extension is to use the adapted Fama-MacBeth framework that we laid out in Section 2.3. We show how to modify the Fama-MacBeth framework so that the null hypothesis — the average of the time-series slope coefficients is zero — is exactly achieved in sample. We can then use this framework to incrementally select the list of true risk factors, similar to what we do with panel regressions. This framework allows us to take time-varying factor loadings into account as in the original Fama-MacBeth approach.

In this paper, we focus on unconditional models and leave these extensions to future research.

4.2 Stepwise Model Selection

Our method falls in the realm of stepwise model selection, in particular forward selection for which we sequentially build up the true factor model. Unlike traditional F-tests or R-square procedures, we pay particular attention to the multiple testing issue, making sure that the Wilkinson and Dallal (1981) critique does not apply to our method.

Having said this, we are aware of the issue that the p-value of our overall procedure, however defined, is likely to be a complex function of the p-values of the individual steps. In our simulation study, we also sidestep this issue by only considering the incremental selection of the second risk factor after the market factor is pre-determined. Notice that at each step, our method proposes a self-contained hypothesis testing framework that tests the incremental contribution of a group of candidate factors to a set of pre-determined factors. The p-value represents the multiple testing adjusted statistical significance of the best candidate variable. We leave a more detailed simulation study of our framework, in particular its overall performance in terms of selecting the true model or a model that has a significant overlap with the true model, to future research.

Traditional model selection methods such as the Akaike information criterion (AIC) or the Bayesian information criterion (BIC) are unlikely to work in our framework. We have a large panel of assets. The number of parameters to be estimated is proportional to the size of the cross-section. As a result, the asymptotic approximations that are required by AIC and BIC are unlikely to hold in our framework. Our bootstrap-based method provides a convenient way to make inference for a limited sample.

An alternative approach to forward selection is backward selection. That is, we start with an overall factor model and sequentially eliminate redundant factors. Our method does not apply to backward selection. To see why this is the case, imagine that we have 30 candidate variables. Based on our method, each time we single out one variable and measure how much it adds to the explanatory power of the other 29 variables. We do this 30 times. However, there is no baseline model across the 30 tests. Each model has a different null hypothesis and we do not have an overall null.

Besides the technical difficulty of backward selection, we think that forward selection makes more sense for our application. For the selection of risk factors, as a prior, we usually do not believe that there should exist hundreds of variables explaining a certain phenomenon. Forward selection is consistent with this prior.

4.3 Spurious Factors

Spurious factors in factor models refer to factors that have weak covariance with asset returns. As shown in Kan and Zhang (1999) and Bryzgalova (2014), spurious factors make the usual inference methods problematic as the risk premia in factor models are weakly identified. Kan and Zhang (1999) and Bryzgalova (2014) propose diagnostic tools as well as shrinkage methods to detect and test spurious factors.

Our testing framework departs from the usual two-stage Fama-MacBeth framework or the associated generalized methods of moments (GMM) approach. In particular, we do not need to use the differences in factor loadings in the cross-section to identify factor risk premia, which is the source of the identification problem studied in Kan and Zhang (1999) and Bryzgalova (2014). Our method uses the reduction in absolute regression intercept (i.e., mispricing) as the test metric to gauge the success of a factor model. A spurious factor, by having a factor loading that is close to zero, naturally implies a small reduction in regression intercept and is likely to be identified as a false risk factor in our framework. Harvey and Liu (2016) provide a more detailed discussion of spurious factors and our tests.

4.4 Factor Model Uncertainty and Model Misspecification

To better link our method to the GRS test, one can think of the baseline model as the null hypothesis and the augmented model as the alternative hypothesis. The GRS test hypothesizes that the augmented model is the true underlying factor model and tests the deviations of the absolute regression intercepts from zero under this hypothesis. In this sense, the GRS test works under the alternative hypothesis. In contrast, our method works under the null hypothesis. We assume that the baseline model performs as well as the augmented model, that is, the additional factor in the augmented model has a zero contribution to explaining the cross-section of expected returns. We test whether the augmented model improves on the baseline model under this assumption.

Due to the above difference in the testing framework, our model is likely to be more powerful in identifying risk factors that belong to the true underlying factor model. For example, suppose the baseline model is simply a constant and the augmented model has the market factor as the candidate risk factor. For a given set of test assets and under GRS, we will reject the GRS null hypothesis (that is, CAPM is the true factor model) since there likely exist other factors that also drive asset returns but do not enter our test. As a result, we reject CAPM and conclude that the market factor cannot fully explain the returns of the test assets. This tells us little about whether the market factor is a true risk factor or not. In contrast, in our framework, we are likely to identify the market factor as true since the augmented model significantly improves on the baseline model (that is, a constant) in explaining the cross-section of expected returns.

In general, given the existence of hundreds of factors that are potential candidates for risk factors, there is a large amount of uncertainty around the true underlying factor model. As a result, any given factor model is likely to be misspecified. The use of the GRS test is limited since almost all models will be rejected in the end. Our test is less sensitive to model misspecifications and allows one to sequentially build the factor list. It does not try to make a statement about the underlying true factor model as in the GRS test. However, it tells us which factors are likely to be the members of the underlying true factor model. Harvey and Liu (2016) have a detailed discussion of factor model uncertainty and asset pricing tests.

5 Conclusions

We present a new method that allows researchers to meet the challenge of multiple testing in financial economics. Our method is based on a bootstrap approach and allows for general distributional characteristics, cross-sectional as well as time-series dependency, and a range of test statistics.

We apply our method to the identification of risk factors. Hundred of factors have been proposed in asset pricing to explain the cross-section of expected returns. Some may appear to be significant risk factors just by chance. In addition, there has long been a suspicion in empirical asset pricing research that portfolios sorted by certain characteristics influence the discovery of new factors. We avoid the portfolio sorting critique by applying our technique to an unbalanced panel of individual stocks.

Our results may seem surprising to many. Our analysis points to one dominant factor — the original market factor proposed by Sharpe (1964). When we value weight individual stocks we do find some support for a second factor linked to profitability, however, its contribution is economically small compared to the market factor.

Finally, while we have applied our method to factor discovery in finance, we want to emphasize that our technique can be applied to any regression model in finance or outside of finance that faces the problem of multiple testing. Indeed, there is a growing need for new tools to navigate the vast array of "big data". We offer a new compass.

References

Adler, R., R. Feldman and M. Taqqu, 1998, A practical guide to heavy tails: Statistial techniques and applications, *Birkhäuser.*

Affleck-Graves, J. and B. McDonald, 1989, Nonnormalities and tests of asset pricing theories, *Journal of Finance 44, 889-908.*

Ahn, D., J. Conrad and R. Dittmar, 2009, Basis assets, *Review of Financial Studies 22, 5133-5174.*

Asness, C., A. Frazzini and L. H. Pedersen, 2013, Quality minus junk, *Working Paper.*

Avramov, D., and T. Chordia, 2006. Asset pricing models and financial market anomalies, *Review of Financial Studies 19, 1001-1040.*

Barras, L., O. Scaillet and R. Wermers, 2010, False discoveries in mutual fund performance: Measuring luck in estimated alphas, *Journal of Finance 65, 179-216.*

Beran, R., 1988, Prepivoting test statistics: A bootstrap view of asymptotic refinements, *Journal of the American Statistical Association 83, 682-697.*

Berk, J. B., 1995, A critique of size-related anomalies, *Review of Financial Studies 8, 275-286.*

Bernard, H., B. T. Kelly, H. N. Lustig, and S. Van Nieuwerburgh, 2014, The common factor in idiosyncratic volatility: Quantitative asset pricing implications. *Working Paper.*

Bryzgalova, S., 2014, Spurious factors in linear asset pricing models. *Working Paper, LSE.*

Carhart, M. M., 1997, On persistence in mutual fund performance, *Journal of Finance 52, 57-82.*

Chordia, T., A. Goyal, and J. Shanken, 2015, Cross-sectional asset pricing with individual stocks: Betas versus characteristics, *Working Paper.*

Ecker, F., Asset pricing tests using random portfolios, *Working Paper, Duke University.*

Efron, B. 1987, Better bootstrap confidence intervals, *Journal of the American Statistical Associations 82, 171-185.*

Efron, B. and R. J. Tibshirani, 1993, *An Introduction to the Bootstrap.* New York: Chapman & Hall.

Fama, E. F., 2015, Cross-section versus time-series tests of asset pricing models, *Working Paper.*

Fama, E. F. and J. D. MacBeth, 1973, Risk, return, and equilibrium: Empirical tests, *Journal of Political Economy 81, 607-636.*

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

Fama, E. F. and K. R. French, 2010, Luck versus skill in the cross-section of mutual fund returns, *Journal of Finance 65, 1915-1947.*

Fama, E. F. and K. R. French, 2015a, A five-factor asset pricing model, *Journal of Financial Economics 116, 1-22.*

Fama, E. F. and K. R. French, 2015b, Incremental variables and the investment opportunity set, *Journal of Financial Economics 117, 470-488.*

Ferson, W. E., and Y. Chen, 2015, How many good and bad fund managers are there, really? *Working Paper, USC.*

Ferson, W. E., and C. R. Harvey, 1991, The variation of economic risk premium. *Journal of Political Economy 99, 385–415.*

Foster, F. D., T. Smith and R. E. Whaley, 1997, Assessing goodness-of-fit of asset pricing models: The distribution of the maximal R^2, *Journal of Finance 52, 591-607.*

Frazzini, A. and L. H. Pedersen, 2014, Betting against beta, *Journal of Financial Economics 111, 1-25.*

Gagliardini, P., E. Ossola, and O. Scaillet, 2014. Time-varying risk premium in large cross-sectional equity datasets, *Econometrica, Forthcoming.*

Ghysels, E., 1998, On stable factor structure in the pricing of risk: Do time-varying betas help or not? *Journal of Finance 53, 549–573.*

Gibbons, M. R., S. A. Ross and J. Shanken, 1989, A test of the efficiency of a given portfolio, *Econometrica 57, 1121-1152.*

Graham, J. R., and C. R. Harvey, 2001, The theory and practice of corporate finance: Evidence from the field, *Journal of Financial Economics 60, 187–243.*

Green, J., J. R. Hand and X. F. Zhang, 2013, The remarkable multidimensionality in the cross section of expected US stock returns, *Working Paper, Pennsylvania State University.*

Hall, P., 1988, Theoretical comparison of bootstrap confidence intervals (with Discussion), *Annals of Statistics 16, 927-985.*

Hall, P. and S. R. Wilson, 1991, Two guidelines for bootstrap hypothesis testing, *Biometrics 47, 757-762.*

Harvey, C. R. and Akhtar Siddique, 2000, Conditional skewness in asset pricing tests, *Journal of Finance, 55, 1263-1295.*

Harvey, C. R., Y. Liu and H. Zhu, 2016, ... and the cross-section of expected returns, *Forthcoming, Review of Financial Studies.* SSRN: http://ssrn.com/abstract=2249314

Harvey, C. R. and Y. Liu, 2014, Multiple testing in financial economics, *Working Paper.* SSRN: http://ssrn.com/abstract=2358214

Harvey, C. R. and Y. Liu, 2015, Dissecting luck vs. skill in investment manager performance, *Work In Progress.*

Harvey, C. R. and Y. Liu, 2016, Factor model uncertainty and asset pricing tests, *Work In Progress.*

Hou, K., C. Xue, and L. Zhang, 2014, Digesting anomalies: An investment approach, *Review of Financial Studies, Forthcoming.*

Hinkley, D. V., 1989, Bootstrap significance tests, In *Proceedings of the 47th Session of the International Statistical Institute*, Paris, 29 August - 6 September 1989, 3, 65-74.

Jagannathan, R., and Z. Wang, 1998, An asymptotic theory for estimating beta-pricing models using cross-sectional regression. *Journal of Finance 53, 1285–1309.*

Jegadeesh, N., and J. Noh, 2014, Empirical tests of asset pricing models with individual stocks, *Working Paper.*

Jeong, J., and G. S. Maddala, 1993, A perspective on application of bootstrap methods in econometrics, In G.S. Maddala, C.R. Rao, and H.D. Vinod (eds), *Handbook of Statistics*, Vol. 11. Amsterdam: North Holland, 573-610.

Kan, R., and C. Zhang, 1999, Two-pass tests of asset pricing models with useless factors. *Journal of Finance 54, 203–235.*

Lettau, M., and S. Ludvigson, 2001, Consumption, aggregate wealth, and expected stock returns. *Journal of Finance 56, 815–849.*

Lewellen, J., S. Nagel and J. Shanken, 2010, A skeptical appraisal of asset pricing tests, *Journal of Financial Economics 96, 175-194.*

Li, Q. and G. S. Maddala, 1996, Bootstrapping time series models, *Econometric Reviews 15, 115-195.*

MacKinlay, A. C., 1987, On multivariate tests of the CAPM, *Journal of Financial Economics 18, 341-371.*

MacKinnon, J. G., 2006, Bootstrap methods in econometrics, *Economic Record 82, S2-18.*

Kan, R., C. Robotti, and J. Shanken, 2013, Pricing model performance and the two-pass cross-sectional regression methodology. *Journal of Finance 68, 2617–2649.*

Kosowski, R., A. Timmermann, R. Wermers and H. White, 2006, Can mutual fund "stars" really pick stocks? New evidence from a bootstrap analysis, *Journal of Finance 61, 2551-2595.*

Lo, A. W., and A. C. MacKinlay, 1990, Data-snooping biases in tests of financial asset pricing models, *Review of Financial Studies 3, 431–467.*

McLean, R. D. and J. Pontiff, 2015, Does academic research destroy stock return predictability? *Journal of Finance, Forthcoming.*

Novy-Marx, R., 2013, The other side of value: The gross profitability premium, *Journal of Financial Economics 108, 1-28.*

Pástor, L. and R. F. Stambaugh, 2003, Liquidity risk and expected stock returns, *Journal of Political Economy 111(3).*

Politis, D. and J. Romano, 1994, The Stationary Bootstrap, *Journal of the American Statistical Association 89, 1303-1313.*

Pukthuanthong, K. and R. Roll, 2014, A protocol for factor identification, *Working Paper, University of Missouri.*

Shanken, J., 1990, Intertemporal asset pricing: An empirical investigation, *Journal of Econometrics 45, 99–102.*

Shanken, J., 1992, On the estimation of beta-pricing model, *Review of Financial Studies 5, 1–33.*

Shanken, J., and G. Zhou, 2007, Estimating and testing beta pricing models: Alternative methods and their performance in simulations. *Journal of Financial Economics 84, 40–86.*

Sharpe, W. F., 1964, Capital asset prices: A theory of market equilibrium under conditions of risk, *Journal of Finance 19, 425–442.*

Sullivan, Ryan, Allan Timmermann and Halbert White, 1999, Data-snooping, technical trading rule performance, and the bootstrap, *Journal of Finance 54, 1647-1691.*

Treynor, J. L., and F. Black, 1973, How to use security analysis to improve portfolio selection. *Journal of Business, 66-86.*

Veall, M. R., 1992, Bootstrapping the process of model selection: An econometric example, *Journal of Applied Econometrics 7, 93-99.*

Veall, M. R., 1998, Applications of the bootstrap in econometrics and economic statistics, In D.E.A. Giles and A. Ullah (eds.), *Handbook of Applied Economic Statistics.* New York: Marcel Dekker, chapter 12.

White, Halbert, 2000, A reality check for data snooping, *Econometrica 68, 1097-1126.*

Wilkinson, L., and G. E. Dallal, 1981, Tests of significance in forward selection regression with an F-to-enter stopping rule, *Technometrics 23, 377-380.*

Young, A., 1986, Conditional data-based simulations: Some examples from geometrical statistics, *International Statistical Review 54, 1-13.*

A Proof for Fama-MacBeth Regressions

The corresponding objective function for the regression model in (15) is given by:

$$\mathcal{L} = \sum_{t=1}^{T}[X_t - (\phi_t + \xi Y_t)]'[X_t - (\phi_t + \xi Y_t)]. \tag{19}$$

Taking first order derivatives with respect to $\{\phi_t\}_{t=1}^{T}$ and ξ, respectively, we have

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \phi_t} &= \iota_t'\varepsilon_t = 0,\ t = 1, \ldots, T, \quad (20) \\ \frac{\partial \mathcal{L}}{\partial \xi} &= \sum_{t=1}^{T} Y_t'\varepsilon_t = 0, \quad (21) \end{aligned}$$

where ι_t is a $n_t \times 1$ vector of ones. (20) says that the residuals within each time period sum up to zero, and (21) says that the Y_t's are on average orthogonal to the ε_t's across time. Importantly, Y_t is not necessarily orthogonal to ε_t within each time period. As explained in the main text, we next define the orthogonalized regressor X_t^e as the rescaled residuals, i.e.,

$$X_t^e = \varepsilon_t/(\varepsilon_t'\varepsilon_t),\ t = 1, \ldots, T. \tag{22}$$

Solving the OLS (17) for each time period, we have:

$$\begin{aligned} \gamma_t &= (X_t^{e\prime}X_t^e)^{-1}X_t^{e\prime}(Y_t - \iota_t\mu_t), \quad (23) \\ &= (X_t^{e\prime}X_t^e)^{-1}X_t^{e\prime}Y_t - (X_t^{e\prime}X_t^e)^{-1}X_t^{e\prime}\iota_t\mu_t,\ t = 1, \ldots, T. \quad (24) \end{aligned}$$

We calculate the two components in (24) separately. First, notice X_t^e is a rescaled version of ε_t. By (20), the second component (i.e., $(X_t^{e\prime}X_t^e)^{-1}X_t^{e\prime}\iota_t\mu_t$) equals zero. The first component is calculated as:

$$\begin{aligned} (X_t^{e\prime}X_t^e)^{-1}X_t^{e\prime}Y_t &= [(\frac{\varepsilon_t}{\varepsilon_t'\varepsilon_t})'(\frac{\varepsilon_t}{\varepsilon_t'\varepsilon_t})]^{-1}(\frac{\varepsilon_t}{\varepsilon_t'\varepsilon_t})'Y_t, \quad (25) \\ &= \varepsilon_t'Y_t,\ t = 1, \ldots, T, \quad (26) \end{aligned}$$

where we again use the definition of X_t^e in equation (25). Hence, we have:

$$\gamma_t = \varepsilon_t'Y_t,\ t = 1, \ldots, T. \tag{27}$$

Finally, applying (21), we have:

$$\sum_{t=1}^{T} \gamma_t = \sum_{t=1}^{T} \varepsilon_t' Y_t = 0.$$

B A Simulation Study

A full-blown simulation study that takes all aspects of our method into account (e.g., the error rate for the first factor to be falsely identified, the error rate for the second factor to be falsely identified conditional on the first factor being correctly identified, etc.) is beyond the scope of this paper.[23] Our main goal for this simulation study is to evaluate the power of our bootstrap-based test in correctly identifying a risk factor that has incremental contribution (relative to the market factor) in explaining the cross-section of individual stock returns. This is motivated by the fact that the market factor is always found to be the most significant factor in our empirical study.

We first focus on firms that have a complete return history for the past twenty years. This gives us a balanced panel with $N = 2,732$ firms in the cross-section and $T = 240$ months in time-series. A balanced panel is not required for our method to work. However, we use a balanced panel in our simulation study as it allows us to fix the number of firms in the cross-section. This allows us to better evaluate how the test power changes with the length of the return time series.

We assume that a two-factor model (i.e., the market factor plus a candidate factor denoted as f_t) is the true model. We construct the panel of returns corresponding to the true model by sampling from the real data. In particular, we first project stock returns onto the two factors:

$$R_{it} - R_{ft} = \alpha_i + \beta_{i,m} mkt_t + \beta_{i,f} f_t + \varepsilon_{i,t}.$$

Let $\mathbf{e}_i = [\varepsilon_{i,1}, \varepsilon_{i,2}, \ldots, \varepsilon_{i,T}]'$ denote the vector of factor model residuals for stock i. We collect the cross-section of factor loadings and residuals into matrices $\mathbf{B}$ and $\mathbf{E}$:

$$\begin{aligned} \mathbf{B}_{(2\times N)} &= [[\beta_{1,m}, \beta_{2,m}, \ldots, \beta_{N,m}]', [\beta_{1,f}, \beta_{2,f}, \ldots, \beta_{N,f}]']', \\ \mathbf{E}_{(T\times N)} &= [\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_T]. \end{aligned}$$

We also project the candidate factor f_t onto the market factor mkt_t:

$$f_t = \alpha_f + \beta_m mkt_t + \varepsilon_{f,t}. \tag{28}$$

The magnitude of α_f determines how "true" the candidate factor is after its correlation with the market factor is taken into account. For example, $\alpha_f = 0$ means that the candidate factor is "spanned" by the market factor so it has a zero incremental contribution to explaining the cross-section of expected returns. This constitutes the null hypothesis. Table (B.1) summarizes α_f for all candidate factors for the past

[23] See Harvey, Liu and Zhu (2015) for a discussion on test power when there are multiple hypothesis tests.

twenty years. For our follow-up analysis, we choose to present results for the top five factors based on the ranking of their t-statistics for α_f. Results for the other factors are similar.

Table B.1: **Summary Statistics on α_f, January 1993 - December 2012**

Summary statistics on α_f. We project a candidate factor f_t onto the market factor mkt_t through the regression: $f_t = \alpha_f + \beta_m mkt_t + \varepsilon_{f,t}$. We report the level and the t-statistic of the regression intercept α_f corresponding to the five risk factors in Fama and French (2015a) (i.e., excess market return (*mkt*), size (*smb*), book-to-market (*hml*), profitability (*rmw*), and investment (*cma*)), betting against beta (*bab*) in Frazzini and Pedersen (2014), gross profitability (*gp*) in Novy-Marx (2013), Pastor and Stambaugh liquidity (*psl*) in Pastor and Stambaugh (2003), momentum (*mom*) in Carhart (1997), quality minus junk (*qmj*) in Asness, Frazzini and Pedersen (2013), investment (*ia*) and profitability (*roe*) in Hou, Xue and Zhang (2015), co-skewness (*skew*) in Harvey and Siddique (2000), and common idiosyncratic volatility (*civ*) in Herskovic, Kelly, Lustig and Van Nieuwerburgh (2014).

	smb	*hml*	*mom*	*skew*	*psl*	*roe*	*ia*	*qmj*	*bab*	*gp*	*cma*	*rmw*	*civ*
Mean	0.014	0.047	0.083	0.033	0.078	0.072	0.054	0.078	0.112	0.056	0.056	0.061	0.067
t-stat	[0.52]	[1.86]	[2.08]	[1.63]	[2.48]	[3.60]	[3.40]	[4.45]	[3.63]	[3.13]	[3.56]	[3.11]	[2.28]

To evaluate the test power corresponding to different alternative hypotheses regarding the candidate factor, we assume that the true candidate factor is

$$f_t^A = A \times \alpha_f + \beta_m mkt_t + \varepsilon_{f,t}. \tag{29}$$

By setting A at zero, the factor premium is completely explained by its exposure to the market factor. As a result, the candidate factor has zero incremental explanatory power of the cross-section of expected returns. This constitutes our null hypothesis. The test power corresponding to the null hypothesis tells us the size of the test. By setting A at other values, the alternative hypothesis is true. By changing the magnitude of A, we are able to evaluate the test power corresponding to different levels of factor premiums, which indicate how significant the candidate factor is in offering incremental information to explain the cross-section of expected returns.

Using the factor loadings and return residuals stored in $\mathbf{B}$ and $\mathbf{E}$, we create the panel of returns corresponding to the true model. In particular, for a resampled time index $\{t_j^w\}_{j=1}^T$ and for a given level of A, the panel of excess returns is given by

$$\begin{aligned} rx_{i,j}^w &= \beta_{i,m} mkt_{t_j^w} + \beta_{i,f} f_{t_j^w}^A + \varepsilon_{i,t_j^w} && (30) \\ &= \beta_{i,m} mkt_{t_j^w} + \beta_{i,f}(A \times \alpha_f + \beta_m mkt_{t_j^w} + \varepsilon_{f,t_j^w}) + \varepsilon_{i,t_j^w}, \; j = 1, \ldots, T; \; i = 1, \ldots, N. && (31) \end{aligned}$$

The way we construct the return panel is slightly different from standard simulation methods in that instead of using Gaussian variables to simulate return residuals, we use bootstrapped residuals based on the real data. This allows us to take the non-normality in returns into account while at the same time maintain the dependency among the cross-section of realized return residuals, as emphasized by Fama and French (2010).

Let the simulated return panel corresponding to the w-th resampled time index be $\mathbf{RX}^w$. For this sample, we use our method to make a decision on whether the candidate factor is significant. Let $D_w = 1$ denote the event that the candidate factor is declared significant ($D_w = 0$ denotes otherwise). We bootstrap the time index W ($= 1{,}000$) times and use $\sum_{w=1}^{W} D_w/W$ to approximate the test power.

To compare our approach with alternative testing methods, we consider two popular methods based on beta sorts. The first method is unconditional beta sorts that first sorts stocks based on their unconditional factor loadings estimated over the entire sample, and then forms the long-short portfolio by having a long position in stocks that are in the top beta decile and a short position in stocks that are in the bottom beta decile. The t-statistic of the portfolios returns is used to test the significance of the candidate factor. The second method is conditional beta sorts. We estimate factor loadings based on a five-year rolling window and construct the long-short portfolio that we hold out-of-sample for one year. We again use the t-statistic of the portfolio returns to test the significance of the candidate factor.

To examine how the length of the time series affects test power, we double the length of the time series by creating a new return panel that fixes the cross-section and repeats the time series of the original panel. Essentially, we are assuming that returns are stationary so we can draw their future realizations from their past realizations. Similarly, we create new factor time series. We then follow the aforementioned procedures of the simulation study to examine test power when the sample size of the time series doubles.

Table B.2 and B.3 show the simulation results for $T = 240$ and $T = 480$, respectively. When $T = 240$ and $A = 0$, the significance levels of all tests seem to be controlled at 5%, which is the pre-specified significance level. However, they seem to under-reject the null as the Type I errors of many of these tests are below 5%. This is likely because the size of the time series is small. When T is increased to 480, the significance levels of most tests are higher and are closer to 5%.

When $A = 1.0$ (that is, factor risk premium is the same as the original factor), the power of our tests based on the t-statistics is in general higher than that based on both types of beta sorts. In particular, when $T = 240$, our tests universally dominate those based on beta sorts. The gain in power by using our tests is about 10% on average. However, the gain is not uniform across factors. For example, the power of our tests is similar to that based on the unconditional beta sorts for *ia*, and is about 20% higher than both conditional and unconditional beta sorts for *qmj*. On the other hand, within the four types of tests based on our method, the tests based on the t-statistics are more powerful than those that are based on the intercepts. We therefore favor the t-statistics-based tests when there is inconsistency between results that are based on different tests.

When $T = 480$, which is closer to the size of the 1968-2012 period that we examine in the paper, the power of our tests seems high. For $A = 1.0$ and for tests based on t-statistics, it ranges from 72% (*bab*) to 95% (*qmj*). For our application with the real data, we have $T = 540$. However, we do not have $N = 2,732$ firms that exist throughout the entire sample. The total number of firm-month observations in our sample is about 1.8 times that of the simulation study.[24] We therefore believe that our tests should have a high power for the real applications.

Overall, our simulation results based on individual stocks suggest that our tests, in particular the tests based on t-statistic, have high test power, both in an absolute and relative sense. When the length of the time series is close to our applications, the simulation results show that the power of our tests is well above 70%. It also compares favorably with the tests that are based on either unconditional or conditional beta sorts.

We redo the same simulation exercise based on the Fama-French 25 portfolios, as shown in Table B.4. For $A = 1.0$, our tests based on t-statistics again dominate those are based on beta sorts. The increase in power of our tests EW_T^m relative to the unconditional beta sorts (i.e., the better one between the two beta sorts) is again nonuniform across factors but is on average about 15%.

More interestingly, comparing Table B.4 with B.3, which has a similar number of time periods to Table B.4, we are not necessarily losing power by considering individual stocks. Focusing on our tests based on EW_T^m and when $A = 1.0$, our tests based on individual stocks have a higher power for *roe* and *qmj* than our tests based on the Fama-French 25 portfolios. Overall, across the five factors, our tests based on individual stocks have a similar power to our tests based on the Fama-French 25 portfolios. This seems to be at odds with the conventional thinking that individual stocks are more noisy and thus less informative than portfolios in factor tests. Our t-statistics-based tests, by taking the return volatility into account, seems to be able to detect a true factor as often as tests based on portfolios.

There are several takeaways from our simulation study. First, our tests based on t-statistics seem to be uniformly more powerful than our tests based on alternative statistics. We therefore favor our tests that are based on t-statistics in our applications.

Second, compared to traditional beta sorts, we are not losing power by using our bootstrap-based approach. In fact, we see mild increase in power across a variety

[24] We have slightly more than 20,000 firms in the cross-section. On average, a firm exists for around 10 years. Therefore, our sample size is about 1.8 ($= (20,000 \times 120)/(2732 \times 480)$) times that of the simulation study.

of factors by using both individual stocks and Fama-French 25 portfolios. While bootstrap is not essential for the two-factor exercise we perform in the simulation study, it is key to our tests in the paper that build on the maximum/minimum test statistics. When extreme test statistics are used to adjust for multiple testing, traditional tests (e.g., conditional or unconditional beta sorts) are no longer appropriate as the asymptotic distributions for the extreme test statistics are not known in closed-form. Moreover, we do not know how well the asymptotic distributions work in finite samples. Bootstrap offers a convenient way to provide inference, as shown in White (2000). It is therefore important to show that the bootstrap-based approach has power in the context of application.

Third, we are not losing power by considering individual stocks. The average performance of our tests based on individual stocks are similar to that of our tests based on the Fama-French 25 portfolios. The key assumption for our simulation study is that a two-factor model is the true underlying factor model, either for individual stocks or the Fama-French 25 portfolios. In reality, asset returns could be determined by a more complicated model. Compared to the Fama-French 25 portfolios, which are constructed to maximize the exposure to two existing factors, individual stocks potentially can provide unbiased and significantly richer information to identify the true factor model.

Table B.2: **Test Power, Individual Stocks,** $T = 240$

Test power for risk factors. For a given risk factor f, we project it onto the market return (equation 28) to obtain the regression intercept. We then construct a new factor (f^A) based on equation (29), with A controlling the factor risk premium that is not explained by the market factor. We run $D = 1,000$ sets of simulations. For each set, we bootstrap the sample period to construct the time series for both the new factor given in equation (29) and the market factor, and, by assuming the two-factor model is true, construct a panel of returns through equation (31), where the return innovations are resampled from the return innovations based on the real data. Based on the bootstrapped factors and return panels, we test whether f^A has incremental power to explain the cross-section of expected returns. We calculate the test power by averaging the number of rejections across simulations. We consider six tests. Four of them are EW_I^m, EW_I^d, EW_T^m, and EW_T^d that are explained in Section 3.2. The rest are two beta sorts. The unconditional beta sorts ($Uncond.\beta$) first sorts stocks based on their unconditional factor loadings estimated over the entire sample, and then forms the long-short portfolio by having a long position in stocks that are in the top beta decile and a short position in stocks that are in the bottom beta decile. The t-statistic of the portfolios returns is used to test the significance of the candidate factor. The conditional beta sorts ($Cond.\beta$) first estimates factor loadings based on a five-year rolling window and then constructs the long-short portfolio that we hold out-of-sample for one year. The t-statistic of the portfolio returns is used to test the significance of the candidate factor. The five factors examined are explained in Table 1. We set $T = 240$ and focus on a cross-section of 2,732 firms that have a complete return history for the past twenty years.

Factor	Method	$A = 0$ (null)	$A = 0.5$	$A = 1.0$	$A = 1.5$	$A = 2.0$
roe	EW_I^m	0.017	0.051	0.336	0.561	0.762
	EW_I^d	0.023	0.064	0.288	0.494	0.644
	EW_T^m	0.022	0.114	0.524	0.773	0.865
	EW_T^d	0.027	0.113	0.402	0.706	0.832
	$Cond.\beta$	0.016	0.151	0.397	0.636	0.826
	$Uncond.\beta$	0.008	0.055	0.329	0.709	0.906
ia	EW_I^m	0.016	0.101	0.436	0.710	0.845
	EW_I^d	0.027	0.134	0.450	0.746	0.879
	EW_T^m	0.022	0.171	0.582	0.838	0.969
	EW_T^d	0.029	0.156	0.549	0.810	0.947
	$Cond.\beta$	0.025	0.144	0.435	0.717	0.869
	$Uncond.\beta$	0.014	0.191	0.532	0.827	0.961
qmj	EW_I^m	0.020	0.141	0.558	0.856	0.937
	EW_I^d	0.021	0.139	0.435	0.731	0.895
	EW_T^m	0.018	0.217	0.756	0.958	0.986
	EW_T^d	0.029	0.177	0.654	0.923	0.957
	$Cond.\beta$	0.010	0.181	0.529	0.823	0.961
	$Uncond.\beta$	0.012	0.161	0.578	0.940	0.977
bab	EW_I^m	0.016	0.062	0.286	0.578	0.772
	EW_I^d	0.026	0.077	0.316	0.578	0.754
	EW_T^m	0.024	0.100	0.446	0.756	0.928
	EW_T^d	0.031	0.061	0.437	0.727	0.898
	$Cond.\beta$	0.005	0.114	0.370	0.667	0.858
	$Uncond.\beta$	0.008	0.087	0.302	0.746	0.923
cma	EW_I^m	0.022	0.131	0.462	0.740	0.903
	EW_I^d	0.019	0.159	0.523	0.754	0.916
	EW_T^m	0.037	0.188	0.631	0.903	0.982
	EW_T^d	0.041	0.136	0.557	0.870	0.981
	$Cond.\beta$	0.036	0.155	0.393	0.641	0.817
	$Uncond.\beta$	0.047	0.184	0.521	0.810	0.963

Table B.3: **Test Power, Individual Stocks,** $T = 480$

Test power for risk factors. For a given risk factor f, we project it onto the market return (equation 28) to obtain the regression intercept. We then construct a new factor (f^A) based on equation (29), with A controlling the factor risk premium that is not explained by the market factor. We run $D = 1,000$ sets of simulations. For each set, we bootstrap the sample period to construct the time series for both the new factor given in equation (29) and the market factor, and, by assuming the two-factor model is true, construct a panel of returns through equation (31), where the return innovations are resampled from the return innovations based on the real data. Based on the bootstrapped factors and return panels, we test whether f^A has incremental power to explain the cross-section of expected returns. We calculate the test power by averaging the number of rejections across simulations. We consider six tests. Four of them are EW_I^m, EW_I^d, EW_T^m, and EW_T^d that are explained in Section 3.2. The rest are two beta sorts. The unconditional beta sorts ($Uncond.\beta$) first sorts stocks based on their unconditional factor loadings estimated over the entire sample, and then forms the long-short portfolio by having a long position in stocks that are in the top beta decile and a short position in stocks that are in the bottom beta decile. The t-statistic of the portfolios returns is used to test the significance of the candidate factor. The conditional beta sorts ($Cond.\beta$) first estimates factor loadings based on a five-year rolling window and then constructs the long-short portfolio that we hold out-of-sample for one year. The t-statistic of the portfolio returns is used to test the significance of the candidate factor. The five factors examined are explained in Table 1. We set $T = 480$ by repeating the returns for a cross-section of 2,732 firms that have a complete return history for the past twenty years.

Factor	Method	$A = 0$ (null)	$A = 0.5$	$A = 1.0$	$A = 1.5$	$A = 2.0$
roe	EW_I^m	0.020	0.318	0.733	0.944	0.964
	EW_I^d	0.033	0.295	0.637	0.866	0.965
	EW_T^m	0.024	0.399	0.809	0.963	0.984
	EW_T^d	0.025	0.332	0.766	0.960	0.985
	$Cond.\beta$	0.013	0.163	0.580	0.910	0.994
	$Uncond.\beta$	0.020	0.108	0.631	0.933	0.982
ia	EW_I^m	0.021	0.295	0.776	0.962	0.972
	EW_I^d	0.025	0.291	0.763	0.965	0.990
	EW_T^m	0.037	0.378	0.861	0.974	1.000
	EW_T^d	0.029	0.351	0.834	0.973	1.000
	$Cond.\beta$	0.026	0.274	0.756	0.948	0.990
	$Uncond.\beta$	0.030	0.383	0.849	0.970	0.992
qmj	EW_I^m	0.013	0.439	0.927	0.988	1.000
	EW_I^d	0.028	0.407	0.880	0.973	1.000
	EW_T^m	0.017	0.552	0.952	0.999	1.000
	EW_T^d	0.029	0.486	0.946	1.000	1.000
	$Cond.\beta$	0.011	0.328	0.825	0.987	1.000
	$Uncond.\beta$	0.016	0.286	0.879	0.996	1.000
bab	EW_I^m	0.011	0.241	0.713	0.923	0.993
	EW_I^d	0.033	0.224	0.694	0.908	0.990
	EW_T^m	0.028	0.289	0.801	0.973	0.992
	EW_T^d	0.021	0.247	0.721	0.964	0.995
	$Cond.\beta$	0.015	0.175	0.633	0.935	0.995
	$Uncond.\beta$	0.008	0.147	0.649	0.971	0.997
cma	EW_I^m	0.043	0.311	0.771	0.963	1.000
	EW_I^d	0.033	0.320	0.780	0.952	1.000
	EW_T^m	0.053	0.373	0.849	0.990	1.000
	EW_T^d	0.031	0.359	0.784	0.985	1.000
	$Cond.\beta$	0.056	0.211	0.683	0.918	0.981
	$Uncond.\beta$	0.051	0.329	0.821	0.983	1.000

Table B.4: **Test Power, Fama-French 25 Portfolios, 1968-2012.**

Test power for risk factors. For a given risk factor f, we project it onto the market return (equation 28) to obtain the regression intercept. We then construct a new factor (f^A) based on equation (29), with A controlling the factor risk premium that is not explained by the market factor. We run $D = 1,000$ sets of simulations. For each set, we bootstrap the sample period to construct the time series for both the new factor given in equation (29) and the market factor, and, by assuming the two-factor model is true, construct a panel of returns through equation (31), where the return innovations are resampled from the return innovations based on the real data. Based on the bootstrapped factors and return panels, we test whether f^A has incremental power to explain the cross-section of expected returns. We calculate the test power by averaging the number of rejections across simulations. We consider six tests. Four of them are EW_I^m, EW_I^d, EW_T^m, and EW_T^d that are explained in Section 3.2. The rest are two beta sorts. The unconditional beta sorts ($Uncond.\beta$) first sorts stocks based on their unconditional factor loadings estimated over the entire sample, and then forms the long-short portfolio by having a long position in stocks that are in the top beta decile and a short position in stocks that are in the bottom beta decile. The t-statistic of the portfolios returns is used to test the significance of the candidate factor. The conditional beta sorts ($Cond.\beta$) first estimates factor loadings based on a five-year rolling window and then constructs the long-short portfolio that we hold out-of-sample for one year. The t-statistic of the portfolio returns is used to test the significance of the candidate factor. The five factors examined are explained in Table 1. We focus on the Fama-French 25 portfolios that cover the period 1968–2012.

Factor	Method	$A = 0$ (null)	$A = 0.5$	$A = 1.0$	$A = 1.5$	$A = 2.0$
roe	EW_I^m	0.018	0.496	0.779	0.892	0.950
	EW_I^d	0.047	0.421	0.682	0.837	0.931
	EW_T^m	0.039	0.519	0.793	0.926	0.975
	EW_T^d	0.043	0.507	0.725	0.890	0.956
	$Cond.\beta$	0.027	0.183	0.509	0.817	0.982
	$Uncond.\beta$	0.019	0.156	0.591	0.913	0.983
ia	EW_I^m	0.029	0.609	0.948	1.000	1.000
	EW_I^d	0.031	0.517	0.927	0.998	1.000
	EW_T^m	0.053	0.638	0.970	1.000	1.000
	EW_T^d	0.037	0.569	0.943	1.000	1.000
	$Cond.\beta$	0.014	0.243	0.817	0.992	1.000
	$Uncond.\beta$	0.012	0.301	0.836	0.994	1.000
qmj	EW_I^m	0.031	0.461	0.871	0.973	0.992
	EW_I^d	0.052	0.419	0.832	0.952	0.995
	EW_T^m	0.053	0.517	0.886	0.978	0.997
	EW_T^d	0.047	0.468	0.850	0.962	0.998
	$Cond.\beta$	0.015	0.247	0.752	0.991	0.992
	$Uncond.\beta$	0.009	0.212	0.841	0.993	0.994
bab	EW_I^m	0.027	0.383	0.782	0.905	0.987
	EW_I^d	0.029	0.341	0.778	0.897	0.980
	EW_T^m	0.032	0.465	0.831	0.954	1.000
	EW_T^d	0.029	0.417	0.768	0.939	0.981
	$Cond.\beta$	0.013	0.083	0.312	0.642	0.872
	$Uncond.\beta$	0.009	0.049	0.379	0.681	0.923
cma	EW_I^m	0.018	0.451	0.882	0.997	1.000
	EW_I^d	0.037	0.396	0.880	0.983	1.000
	EW_T^m	0.041	0.559	0.913	0.990	1.000
	EW_T^d	0.061	0.507	0.870	0.996	1.000
	$Cond.\beta$	0.019	0.209	0.673	0.973	1.000
	$Uncond.\beta$	0.039	0.240	0.796	0.991	1.000

C The Block Bootstrap

Our block bootstrap follows the so-called stationary bootstrap proposed by Politis and Romano (1994) and subsequently applied by White (2000) and Sullivan, Timmermann and White (1999). The stationary bootstrap applies to a strictly stationary and weakly dependent time-series to generate a pseudo time series that is stationary. The stationary bootstrap allows us to resample blocks of the original data, with the length of the block being random and following a geometric distribution with a mean of $1/q$. Therefore, the smoothing parameter q controls the average length of the blocks. A small q (i.e., on average long blocks) is needed for data with strong dependence and a large q (i.e., on average short blocks) is appropriate for data with little dependence. We describe the details of the algorithm in this section.

Suppose the set of time indices for the original data is $1, 2, \ldots, T$. For each bootstrapped sample, our goal is to generate a new set of time indices $\{\theta(t)\}_{t=1}^{T}$. Following Politis and Romano (1994), we first need to choose a smoothing parameter q that can be thought of as the reciprocal of the average block length. The conditions that $q = q_n$ needs to satisfies are:

$$0 < q_n \leq 1, q_n \to 0, nq_n \to \infty.$$

Given this smoothing parameter, we follow the following steps to generate the new set of time indices for each bootstrapped sample:

- Step I. Set $t = 1$ and draw $\theta(1)$ independently and uniformly from $1, 2, \ldots, T$.
- Step II. Move forward one period by setting $t = t+1$. Stop if $t > T$. Otherwise, independently draw a uniformly distributed random variable U on the unit interval.
    1. If $U < q$, draw $\theta(t)$ independently and uniformly from $1, 2, \ldots, T$.
    2. Otherwise (i.e., $U \geq q$), set $\theta(t) = \theta(t-1) + 1$ if $\theta(t) \leq T$ and $\theta(t) = 1$ if $\theta(t) > T$.
- Step III. Repeat step II.

For most of our applications, we experiment with different levels of q and show how our results change with respect to the level of q.

D Internet Appendix

D.1 Russell 1000/1500

D.2 Industry Portfolios

D.3 Time Dependence

D.4 Liquidity

Table D.1.1: **Individual Stocks as Test Assets, Big Stocks, Equally Weighted Intercepts**

Test results on 14 risk factors using individual stocks. We use individual stocks from CRSP that cover the 1968– 2012 period to test 14 risk factors. At the beginning of each year, we rank stocks based on their market capitalizations and focus on stocks that have a market capitalization that is above the *10*th percentile of the cross-section of market capitalizations (that is, "big stocks"). A stock needs to have at least 36 monthly observations (either in the original or the bootstrapped sample) to enter our tests. The 14 risk factors are excess market return (*mkt*), size (*smb*), book-to-market (*hml*), profitability (*rmw*), and investment (*cma*) in Fama and French (2015a), betting against beta (*bab*) in Frazzini and Pedersen (2014), gross profitability (*gp*) in Novy-Marx (2013), Pastor and Stambaugh liquidity (*psl*) in Pastor and Stambaugh (2003), momentum (*mom*) in Carhart (1997), quality minus junk (*qmj*) in Asness, Frazzini and Pedersen (2013), investment (*ia*) and profitability (*roe*) in Hou, Xue and Zhang (2015), co-skewness (*skew*) in Harvey and Siddique (2000), and common idiosyncratic volatility in Herskovic, Kelly, Lustig and Van Nieuwerburgh (2014). The baseline model refers to the model that includes the pre-selected risk factors. We focus on the panel regression model described in Section 2.2. The two metrics (i.e., EW_I^m and EW_I^d), which measure the difference in equally weighted mean/median absolute intercepts, are defined in Section 4.2.

Panel A: Baseline = No factor							Panel B: Baseline = ***mkt***					
		single test			single test			single test			single test	
Factor	EW_I^m	5*th*-percentile	*p*-value	EW_I^d	5th-percentile	*p*-value	EW_I^m	5th-percentile	*p*-value	EW_I^d	5th-percentile	*p*-value
mkt	**-0.197**	[-0.226]	(0.116)	**-0.271**	[-0.247]	(0.027)						
smb	-0.040	[-0.096]	(0.275)	-0.056	[-0.075]	(0.136)	**-0.028**	[-0.058]	(0.382)	**-0.009**	[-0.033]	(0.504)
hml	0.186	[-0.056]	(0.995)	0.157	[-0.058]	(0.998)	0.035	[-0.032]	(0.992)	0.003	[-0.030]	(0.716)
mom	0.053	[-0.078]	(1.000)	0.092	[-0.063]	(1.000)	0.003	[-0.018]	(0.621)	0.035	[-0.020]	(1.000)
skew	0.030	[-0.043]	(0.868)	-0.009	[-0.034]	(0.301)	0.026	[-0.008]	(0.975)	-0.005	[-0.021]	(0.456)
psl	0.044	[-0.041]	(0.987)	0.035	[-0.041]	(1.000)	0.009	[-0.017]	(0.718)	0.006	[-0.018]	(0.754)
roe	0.138	[-0.104]	(1.000)	0.085	[-0.087]	(1.000)	0.050	[-0.022]	(1.000)	0.035	[-0.021]	(1.000)
ia	0.385	[-0.079]	(1.000)	0.375	[-0.079]	[1.000]	0.039	[-0.029]	(0.983)	0.005	[-0.031]	(0.732)
qmj	0.395	[-0.103]	(1.000)	0.380	[-0.089]	(1.000)	0.045	[-0.028]	(1.000)	0.032	[-0.030]	(1.000)
bab	0.108	[-0.058]	(1.000)	-0.019	[-0.054]	(0.348)	0.077	[-0.006]	(1.000)	0.020	[-0.022]	(0.932)
gp	0.020	[-0.049]	(0.827)	-0.031	[-0.038]	(0.096)	0.037	[-0.020]	(0.996)	0.009	[-0.022]	(0.915)
cma	0.275	[-0.082]	(0.997)	0.255	[-0.086]	(0.987)	-0.024	[-0.014]	(0.991)	-0.008	[-0.028]	(0.343)
rmw	0.166	[-0.046]	(0.995)	0.081	[-0.035]	(1.000)	0.038	[-0.025]	(1.000)	-0.007	[-0.033]	(0.654)
civ	-0.095	[-0.063]	(0.015)	-0.094	[-0.051]	(0.003)	0.011	[-0.020]	(0.929)	0.010	[-0.018]	(0.895)
		multiple test			multiple test			multiple test			multiple test	
	min	[-0.226]	(0.113)		[-0.247]	(0.021)	*min*	[-0.058]	(0.443)		[-0.044]	(0.951)

Table D.1.2: **Individual Stocks as Test Assets, Big Stocks, Equally Weighted T-Statistics**

Test results on 14 risk factors using individual stocks. We use individual stocks from CRSP that cover the 1968– 2012 period to test 14 risk factors. At the beginning of each year, we rank stocks based on their market capitalizations and focus on stocks that have a market capitalization that is above the *10*th percentile of the cross-section of market capitalizations (that is, "big stocks"). A stock needs to have at least 36 monthly observations (either in the original or the bootstrapped sample) to enter our tests. The 14 risk factors are excess market return (*mkt*), size (*smb*), book-to-market (*hml*), profitability (*rmw*), and investment (*cma*) in Fama and French (2015a), betting against beta (*bab*) in Frazzini and Pedersen (2014), gross profitability (*gp*) in Novy-Marx (2013), Pastor and Stambaugh liquidity (*psl*) in Pastor and Stambaugh (2003), momentum (*mom*) in Carhart (1997), quality minus junk (*qmj*) in Asness, Frazzini and Pedersen (2013), investment (*ia*) and profitability (*roe*) in Hou, Xue and Zhang (2015), co-skewness (*skew*) in Harvey and Siddique (2000), and common idiosyncratic volatility in Herskovic, Kelly, Lustig and Van Nieuwerburgh (2014). The baseline model refers to the model that includes the pre-selected risk factors. We focus on the panel regression model described in Section 2.2. The two metrics (i.e., EW_T^m and EW_T^d), which measure the difference in equally weighted mean/median absolute t-statistics, are defined in Section 4.2.

Panel A: Baseline = No factor							Panel B: Baseline = ***mkt***					
		single test			single test			single test			single test	
Factor	EW_T^m	5*th*-percentile	p-value	EW_T^d	5th-percentile	p-value	EW_T^m	5th-percentile	p-value	EW_T^d	5th-percentile	p-value
mkt	**-0.257**	[-0.234]	(0.039)	**-0.241**	[-0.253]	(0.053)						
smb	-0.038	[-0.050]	(0.698)	-0.056	[-0.080]	(0.079)	-0.007	[-0.031]	(0.419)	-0.030	[-0.039]	(0.147)
hml	0.200	[-0.029]	(1.000)	0.210	[-0.030]	(1.000)	-0.005	[-0.039]	(0.408)	-0.022	[-0.039]	(0.162)
mom	0.092	[-0.046]	(0.972)	0.076	[-0.049]	(0.979)	0.036	[-0.009]	(1.000)	0.031	[-0.014]	(0.927)
skew	-0.009	[-0.042]	(0.215)	0.009	[-0.046]	(0.595)	-0.013	[-0.012]	(0.234)	0.006	[-0.019]	(0.530)
psl	0.039	[-0.038]	(0.973)	0.040	[-0.046]	(0.966)	-0.005	[-0.016]	(0.206)	-0.013	[-0.022]	(0.098)
roe	0.097	[-0.101]	(1.000)	0.093	[-0.123]	(1.000)	0.010	[-0.024]	(0.863)	0.007	[-0.026]	(0.778)
ia	0.431	[-0.047]	(1.000)	0.447	[-0.047]	[1.000]	-0.029	[-0.039]	(0.093)	-0.054	[-0.039]	(0.001)
qmj	0.508	[-0.097]	(1.000)	0.555	[-0.115]	(1.000)	-0.005	[-0.028]	(0.517)	0.005	[-0.039]	(0.752)
bab	-0.053	[-0.068]	(0.096)	-0.102	[-0.075]	(0.027)	-0.012	[-0.025]	(0.182)	-0.043	[-0.031]	(0.013)
gp	-0.026	[-0.048]	(0.241)	-0.017	[-0.056]	(0.322)	0.025	[-0.019]	(0.991)	0.011	[-0.029]	(0.789)
cma	0.319	[-0.070]	(1.000)	0.323	[-0.078]	(1.000)	**-0.032**	[-0.038]	(0.073)	**-0.058**	[-0.037]	(0.012)
rmw	0.128	[-0.024]	(0.992)	0.141	[-0.029]	(0.995)	0.013	[-0.032]	(0.301)	-0.025	[-0.042]	(0.215)
civ	-0.120	[-0.045]	(0.008)	-0.141	[-0.050]	(0.006)	0.011	[-0.016]	(0.948)	-0.001	[-0.028]	(0.643)
		multiple test			multiple test			multiple test			multiple test	
	min	[-0.234]	(0.041)		[-0.253]	(0.055)	*min*	[-0.053]	(0.213)		[-0.060]	(0.039)

Panel C: Baseline = ***mkt*** + ***cma***						
		single test			single test	
Factor	EW_T^m	5*th*-percentile	p-value	EW_T^d	5th-percentile	p-value
mkt						
smb	0.002	[-0.031]	(0.538)	-0.001	[-0.031]	(0.515)
hml	0.013	[-0.009]	(0.491)	0.015	[-0.019]	(0.598)
mom	-0.007	[-0.011]	(0.138)	**-0.007**	[-0.018]	(0.175)
skew	0.011	[-0.012]	(0.708)	0.010	[-0.017]	(0.628)
psl	0.003	[-0.019]	(0.660)	-0.007	[-0.028]	(0.382)
roe	0.014	[-0.028]	(0.940)	0.004	[-0.028]	(0.653)
ia	**-0.007**	[-0.020]	(0.317)	0.001	[-0.026]	(0.618)
qmj	0.029	[-0.014]	(0.968)	0.013	[-0.021]	(0.719)
bab	0.041	[-0.009]	(1.000)	0.017	[-0.020]	(0.871)
gp	0.046	[-0.007]	(1.000)	0.039	[-0.034]	(1.000)
cma						
rmw	0.021	[-0.024]	(0.972)	0.010	[-0.027]	(0.758)
civ	0.014	[-0.016]	(0.913)	0.015	[-0.020]	(0.851)
		multiple test			multiple test	
	min	[-0.041]	(0.842)		[-0.040]	(0.908)

Table D.1.3: **Individual Stocks as Test Assets, Big Stocks, Value Weighted Intercepts/T-Statistics**

Test results on 14 risk factors using individual stocks. We use individual stocks from CRSP that cover the 1968– 2012 period to test 14 risk factors. At the beginning of each year, we rank stocks based on their market capitalizations and focus on stocks that have a market capitalization that is above the *10*th percentile of the cross-section of market capitalizations (that is, "big stocks"). A stock needs to have at least 36 monthly observations (either in the original or the bootstrapped sample) to enter our tests. The 14 risk factors are excess market return (*mkt*), size (*smb*), book-to-market (*hml*), profitability (*rmw*), and investment (*cma*) in Fama and French (2015a), betting against beta (*bab*) in Frazzini and Pedersen (2014), gross profitability (*gp*) in Novy-Marx (2013), Pastor and Stambaugh liquidity (*psl*) in Pastor and Stambaugh (2003), momentum (*mom*) in Carhart (1997), quality minus junk (*qmj*) in Asness, Frazzini and Pedersen (2013), investment (*ia*) and profitability (*roe*) in Hou, Xue and Zhang (2015), co-skewness (*skew*) in Harvey and Siddique (2000), and common idiosyncratic volatility in Herskovic, Kelly, Lustig and Van Nieuwerburgh (2014). The baseline model refers to the model that includes the pre-selected risk factors. We focus on the panel regression model described in Section 2.2. The two metrics (i.e., VW_I and VW_T), which measure the difference in value-weighted absolute intercepts and t-statistics, are defined in Section 4.2.

Panel A: Baseline = No factor							Panel B: Baseline = ***mkt***					
		single test			single test			single test			single test	
Factor	VW_I	5th-percentile	p-value	VW_T	5th-percentile	p-value	VW_I	5th-percentile	p-value	VW_T	5th-percentile	p-value
mkt	**-0.367**	[-0.300]	(0.025)	**-0.702**	[-0.458]	(0.015)						
smb	-0.050	[-0.065]	(0.087)	-0.067	[-0.056]	(0.027)	0.006	[-0.025]	(0.831)	0.056	[-0.093]	(0.848)
hml	0.152	[-0.095]	(1.000)	0.322	[-0.119]	(1.000)	0.022	[-0.028]	(0.996)	-0.008	[-0.077]	(0.235)
mom	0.118	[-0.068]	(1.000)	0.266	[-0.105]	(1.000)	0.047	[-0.008]	(1.000)	0.128	[-0.011]	(1.000)
skew	-0.004	[-0.040]	(0.351)	-0.036	[-0.069]	(0.146)	-0.005	[-0.007]	(0.095)	-0.043	[-0.038]	(0.033)
psl	0.032	[-0.031]	(0.991)	0.060	[-0.053]	(0.951)	0.012	[-0.006]	(1.000)	0.017	[-0.014]	(0.949)
roe	0.100	[-0.077]	(1.000)	0.128	[-0.065]	(1.000)	-0.022	[-0.024]	(0.054)	-0.163	[-0.064]	(0.007)
ia	0.365	[-0.088]	(1.000)	0.728	[-0.151]	(1.000)	0.030	[-0.027]	(1.000)	-0.060	[-0.100]	(0.089)
qmj	0.338	[-0.136]	(1.000)	0.734	[-0.177]	(1.000)	**-0.043**	[-0.034]	(0.014)	**-0.246**	[-0.139]	(0.003)
bab	0.012	[-0.047]	(0.914)	-0.109	[-0.079]	(0.014)	0.010	[-0.013]	(0.992)	-0.103	[-0.052]	(0.002)
gp	-0.066	[-0.043]	(0.013)	-0.161	[-0.056]	(0.003)	-0.016	[-0.030]	(0.372)	-0.026	[-0.038]	(0.106)
cma	0.297	[-0.099]	(1.000)	0.641	[-0.171]	(0.998)	0.020	[-0.022]	(1.000)	-0.054	[-0.065]	(0.070)
rmw	0.071	[-0.026]	(0.994)	0.077	[-0.015]	(0.924)	-0.032	[-0.021]	(0.029)	-0.183	[-0.110]	(0.000)
civ	-0.104	[-0.061]	(0.005)	-0.210	[-0.092]	(0.001)	0.018	[-0.009]	(0.998)	0.063	[-0.033]	(0.991)
		multiple test			multiple test			multiple test			multiple test	
	min	[-0.300]	(0.031)		[-0.458]	(0.021)	*min*	[-0.039]	(0.032)		[-0.155]	(0.001)

Panel C: Baseline = ***mkt*** + ***bab***						
		single test			single test	
Factor	VW_I	5*th*-percentile	p-value	VW_T	5th-percentile	p-value
mkt						
smb	0.035	[-0.020]	(1.000)	0.146	[-0.045]	(1.000)
hml	0.036	[-0.033]	(1.000)	0.023	[-0.068]	(0.873)
mom	0.038	[-0.018]	(1.000)	0.107	[-0.043]	(1.000)
skew	0.022	[-0.008]	(0.463)	-0.008	[-0.021]	(0.118)
psl	0.015	[-0.009]	(0.997)	0.036	[-0.022]	(0.993)
roe	0.009	[-0.013]	(0.988)	0.028	[-0.024]	(0.982)
ia	0.069	[-0.030]	(1.000)	0.088	[-0.071]	(1.000)
qmj						
bab	0.030	[-0.016]	(0.991)	-0.001	[-0.032]	(0.476)
gp	**-0.014**	[-0.035]	(0.438)	**-0.014**	[-0.047]	(0.301)
cma	0.051	[-0.024]	(1.000)	0.031	[-0.052]	(0.937)
rmw	0.010	[-0.011]	(0.977)	0.039	[-0.024]	(0.985)
civ	0.000	[-0.011]	(0.721)	-0.000	[-0.023]	(0.428)
		multiple test			multiple test	
	min	[-0.039]	(0.832)		[-0.090]	(0.867)

Table D.2.1: **49 Industries, Equally Weighted Intercepts**

Test results on 14 risk factors using 49 industry portfolios. We use the 49 industry portfolios that are available on Ken French's on-line data library to test 14 risk factors. The 14 risk factors are excess market return (*mkt*), size (*smb*), book-to-market (*hml*), profitability (*rmw*), and investment (*cma*) in Fama and French (2015a), betting against beta (*bab*) in Frazzini and Pedersen (2014), gross profitability (*gp*) in Novy-Marx (2013), Pastor and Stambaugh liquidity (*psl*) in Pastor and Stambaugh (2003), momentum (*mom*) in Carhart (1997), quality minus junk (*qmj*) in Asness, Frazzini and Pedersen (2013), investment (*ia*) and profitability (*roe*) in Hou, Xue and Zhang (2015), co-skewness (*skew*) in Harvey and Siddique (2000), and common idiosyncratic volatility in Herskovic, Kelly, Lustig and Van Nieuwerburgh (2014). The baseline model refers to the model that includes the pre-selected risk factors. We focus on the panel regression model described in Section 2.2. The two metrics (i.e., EW_I^m and EW_I^d), which measure the difference in equally weighted mean/median absolute intercepts, are defined in Section 4.2.

	Panel A: Baseline = No factor						Panel B: Baseline = ***mkt***					
		single test			single test			single test			single test	
Factor	EW_I^m	5*th*-percentile	*p*-value	EW_I^d	5th-percentile	*p*-value	EW_I^m	5th-percentile	*p*-value	EW_I^d	5th-percentile	*p*-value
mkt	**-0.370**	[-0.342]	(0.0410)	**-0.416**	[-0.350]	(0.023)						
smb	-0.098	[-0.126]	(0.093)	-0.073	[-0.129]	(0.168)	0.011	[-0.029]	(0.887)	0.013	[-0.036]	(0.842)
hml	0.147	[-0.078]	(0.998)	0.162	[-0.083]	(0.999)	0.010	[-0.025]	(0.908)	**0.025**	[-0.041]	(0.942)
mom	0.164	[-0.084]	(0.999)	0.179	[-0.092]	(0.999)	0.021	[-0.010]	(0.996)	0.049	[-0.024]	(0.999)
skew	-0.022	[-0.027]	(0.070)	0.003	[-0.034]	(0.662)	-0.023	[-0.018]	(0.029)	-0.012	[-0.026]	(0.165)
psl	0.016	[-0.019]	(0.916)	0.041	[-0.026]	(0.979)	-0.005	[-0.006]	(0.082)	0.010	[-0.014]	(0.891)
roe	0.207	[-0.074]	(1.000)	0.191	[-0.083]	(1.000)	-0.015	[-0.019]	(0.084)	0.018	[-0.028]	(0.912)
ia	0.397	[-0.107]	(1.000)	0.395	[-0.114]	[1.000]	0.026	[-0.024]	(0.991)	0.025	[-0.038]	(0.950)
qmj	0.402	[-0.168]	(1.000)	0.403	[-0.175]	(1.000)	-0.013	[-0.023]	(0.142)	-0.024	[-0.032]	(0.087)
bab	0.019	[-0.020]	(0.940)	-0.009	[-0.031]	(0.232)	-0.003	[-0.024]	(0.386)	0.005	[-0.036]	(0.701)
gp	-0.069	[-0.040]	(0.012)	-0.082	[-0.053]	(0.010)	0.026	[-0.016]	(1.000)	0.001	[-0.029]	(0.542)
cma	0.334	[-0.116]	(1.000)	0.332	[-0.121]	(1.000)	**-0.020**	[-0.014]	(0.822)	0.034	[-0.032]	(0.985)
rmw	0.091	[-0.047]	(0.998)	0.067	[-0.051]	(0.977)	-0.023	[-0.031]	(0.086)	-0.023	[-0.039]	(0.131)
civ	-0.160	[-0.078]	(0.002)	-0.182	[-0.082]	(0.002)	0.015	[-0.009]	(0.988)	0.028	[-0.020]	(0.977)
		multiple test			multiple test			multiple test			multiple test	
	min	[-0.342]	(0.041)		[-0.350]	(0.023)	*min*	[-0.042]	(0.263)		[-0.060]	(0.466)

Table D.2.2: **49 Industry, Equally Weighted T-Statistics**

Test results on 14 risk factors using 49 industry portfolios. We use the 49 industry portfolios that are available on Ken French's on-line data library to test 14 risk factors. The 14 risk factors are excess market return (*mkt*), size (*smb*), book-to-market (*hml*), profitability (*rmw*), and investment (*cma*) in Fama and French (2015a), betting against beta (*bab*) in Frazzini and Pedersen (2014), gross profitability (*gp*) in Novy-Marx (2013), Pastor and Stambaugh liquidity (*psl*) in Pastor and Stambaugh (2003), momentum (*mom*) in Carhart (1997), quality minus junk (*qmj*) in Asness, Frazzini and Pedersen (2013), investment (*ia*) and profitability (*roe*) in Hou, Xue and Zhang (2015), co-skewness (*skew*) in Harvey and Siddique (2000), and common idiosyncratic volatility in Herskovic, Kelly, Lustig and Van Nieuwerburgh (2014). The baseline model refers to the model that includes the pre-selected risk factors. We focus on the panel regression model described in Section 2.2. The two metrics (i.e., EW_T^m and EW_T^d), which measure the difference in equally weighted mean/median absolute t-statistics, are defined in Section 4.2.

Panel A: Baseline = No factor							Panel B: Baseline = ***mkt***					
		single test			single test			single test			single test	
Factor	EW_T^m	5*th*-percentile	*p*-value	EW_T^d	5th-percentile	*p*-value	EW_T^m	5th-percentile	*p*-value	EW_T^d	5th-percentile	*p*-value
mkt	**-1.023**	[-0.105]	(0.003)	**-1.325**	[-0.147]	(0.003)						
smb	-0.261	[-0.289]	(0.064)	-0.297	[-0.336]	(0.063)	0.083	[-0.118]	(0.854)	0.147	[-0.167]	(0.916)
hml	0.524	[-0.210]	(0.997)	0.533	[-0.235]	(0.996)	0.068	[-0.113]	(0.841)	0.044	[-0.184]	(0.652)
mom	0.571	[-0.249]	(1.000)	0.717	[-0.285]	(1.000)	0.142	[-0.040]	(0.999)	0.326	[-0.105]	(1.000)
skew	-0.084	[-0.090]	(0.060)	-0.110	[-0.117]	(0.056)	-0.114	[-0.075]	(0.018)	-0.136	[-0.119]	(0.035)
psl	0.040	[-0.068]	(0.859)	0.071	[-0.089]	(0.888)	-0.023	[-0.028]	(0.079)	-0.007	[-0.080]	(0.378)
roe	0.639	[-0.205]	(1.000)	0.658	[-0.245]	(0.999)	-0.088	[-0.093]	(0.057)	0.057	[-0.141]	(0.762)
ia	1.339	[-0.271]	(1.000)	1.340	[-0.337]	[1.000]	0.081	[-0.098]	(0.937)	0.196	[-0.162]	(0.980)
qmj	1.614	[-0.441]	(1.000)	1.561	[-0.471]	(1.000)	-0.085	[-0.111]	(0.076)	-0.025	[-0.140]	(0.285)
bab	-0.025	[-0.058]	(0.144)	-0.185	[-0.103]	(0.014)	-0.009	[-0.097]	(0.253)	0.049	[-0.164]	(0.704)
gp	-0.243	[-0.118]	(0.004)	-0.375	[-0.159]	(0.001)	0.104	[-0.045]	(0.998)	0.104	[-0.107]	(0.856)
cma	1.186	[-0.321]	(1.000)	1.243	[-0.364]	(1.000)	**0.000**	[-0.090]	(0.327)	**0.137**	[-0.145]	(0.954)
rmw	0.288	[-0.132]	(0.993)	0.276	[-0.165]	(0.968)	-0.129	[-0.117]	(0.037)	-0.017	[-0.181]	(0.343)
civ	-0.528	[-0.231]	(0.001)	-0.637	[-0.265]	(0.001)	0.075	[-0.048]	(0.982)	0.076	[-0.100]	(0.889)
		multiple test			multiple test			multiple test			multiple test	
	min	[-0.555]	(0.003)		[-0.629]	(0.003)	*min*	[-0.201]	(0.176)		[-0.311]	(0.339)

Table D.2.3: **49 Industry Portfolios, Value Weighted Intercepts/T-Statistics**

Test results on 14 risk factors using 49 industry portfolios. We use the 49 industry portfolios that are available on Ken French's on-line data library to test 14 risk factors. The 14 risk factors are excess market return (*mkt*), size (*smb*), book-to-market (*hml*), profitability (*rmw*), and investment (*cma*) in Fama and French (2015a), betting against beta (*bab*) in Frazzini and Pedersen (2014), gross profitability (*gp*) in Novy-Marx (2013), Pastor and Stambaugh liquidity (*psl*) in Pastor and Stambaugh (2003), momentum (*mom*) in Carhart (1997), quality minus junk (*qmj*) in Asness, Frazzini and Pedersen (2013), investment (*ia*) and profitability (*roe*) in Hou, Xue and Zhang (2015), co-skewness (*skew*) in Harvey and Siddique (2000), and common idiosyncratic volatility in Herskovic, Kelly, Lustig and Van Nieuwerburgh (2014). The baseline model refers to the model that includes the pre-selected risk factors. We focus on the panel regression model described in Section 2.2. The two metrics (i.e., VW_I and VW_T), which measure the difference in value-weighted absolute intercepts, are defined in Section 4.2.

	Panel A: Baseline = No factor						Panel B: Baseline = ***mkt***					
		single test			single test			single test			single test	
Factor	VW_I	5*th*-percentile	*p*-value	VW_T	5th-percentile	*p*-value	VW_I	5th-percentile	*p*-value	VW_T	5th-percentile	*p*-value
mkt	**-0.371**	[-0.306]	(0.018)	**-1.066**	[-0.141]	(0.002)						
smb	-0.062	[-0.076]	(0.083)	-0.178	[-0.220]	(0.078)	0.013	[-0.028]	(0.803)	0.092	[-0.126]	(0.829)
hml	0.124	[-0.064]	(0.997)	0.470	[-0.184]	(0.998)	-0.033	[-0.037]	(0.069)	-0.144	[-0.147]	(0.052)
mom	0.136	[-0.063]	(0.998)	0.488	[-0.202]	(0.997)	0.019	[-0.007]	(0.992)	0.112	[-0.035]	(0.999)
skew	-0.020	[-0.026]	(0.085)	-0.081	[-0.092]	(0.062)	-0.035	[-0.026]	(0.018)	-0.169	[-0.105]	(0.013)
psl	0.015	[-0.016]	(0.903)	0.038	[-0.078]	(0.857)	-0.001	[-0.006]	(0.365)	-0.016	[-0.027]	(0.123)
roe	0.133	[-0.047]	(1.000)	0.414	[-0.157]	(0.999)	-0.062	[-0.033]	(0.004)	-0.293	[-0.133]	(0.004)
ia	0.317	[-0.084]	(1.000)	1.117	[-0.272]	(1.000)	-0.046	[-0.034]	(0.018)	-0.244	[-0.144]	(0.007)
qmj	0.301	[-0.133]	(1.000)	1.258	[-0.329]	(1.000)	-0.084	[-0.041]	(0.002)	**-0.390**	[-0.166]	(0.001)
bab	-0.010	[-0.016]	(0.092)	-0.136	[-0.060]	(0.007)	-0.067	[-0.040]	(0.009)	-0.320	[-0.163]	(0.004)
gp	-0.070	[-0.036]	(0.007)	-0.248	[-0.114]	(0.004)	0.007	[-0.016]	(0.951)	0.051	[-0.043]	(0.755)
cma	0.270	[-0.094]	(1.000)	1.006	[-0.305]	(1.000)	-0.055	[-0.032]	(0.008)	-0.282	[-0.130]	(0.001)
rmw	0.049	[-0.027]	(0.988)	0.159	[-0.058]	(0.975)	**-0.076**	[-0.047]	(0.010)	-0.353	[-0.195]	(0.009)
civ	-0.121	[-0.062]	(0.003)	-0.426	[-0.187]	(0.002)	0.020	[-0.011]	(0.993)	0.100	[-0.043]	(0.988)
		multiple test			multiple test			multiple test			multiple test	
	min	[-0.306]	(0.018)		[-0.428]	(0.002)	*min*	[-0.062]	(0.011)		[-0.272]	(0.009)

	Panel C: Baseline = ***mkt*** + ***bab***					
		single test			single test	
Factor	VW_I	5*th*-percentile	*p*-value	VW_T	5th-percentile	*p*-value
mkt						
smb	**0.044**	[-0.013]	(1.000)	0.259	[-0.070]	(1.000)
hml	0.032	[-0.038]	(0.990)	0.210	[-0.173]	(0.994)
mom	0.027	[-0.015]	(0.999)	0.142	[-0.065]	(0.996)
skew	0.003	[-0.015]	(0.816)	0.031	[-0.062]	(0.827)
psl	-0.004	[-0.007]	(0.108)	-0.036	[-0.033]	(0.046)
roe	0.010	[-0.010]	(0.990)	0.060	[-0.037]	(0.983)
ia	0.040	[-0.031]	(0.998)	0.230	[-0.134]	(1.000)
qmj						
bab	0.020	[-0.024]	(0.973)	**0.148**	[-0.100]	(0.995)
gp	-0.007	[-0.033]	(0.341)	-0.008	[-0.116]	(0.283)
cma	0.019	[-0.029]	(0.929)	0.128	[-0.131]	(0.959)
rmw	0.021	[-0.026]	(0.977)	0.110	[-0.107]	(0.973)
civ	0.029	[-0.011]	(1.000)	0.134	[-0.039]	(1.000)
		multiple test			multiple test	
	min	[-0.048]	(0.795)		[-0.201]	(0.653)

E FAQ

E.1 General Questions

- *Isn't weighting by market cap just another way of creating size portfolios? (Section 2)*

  In our paper, value weighting allows us to take into account the differential impact of a factor across size groups. It could be the significance (e.g., t-statistic) of the factor across size groups. This is different from the return differential across size groups that are caused by the size effect.