# 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: March 15, 2015

**Abstract**

We propose a new regression method to select amongst a large group of candidate factors — many of which might be the 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 such as t-ratios. While our main application focuses on asset pricing, the method can be applied in any situation where regression analysis is used in the presence of multiple testing. This includes, for example, the evaluation of investment manager performance as well as time-series prediction of asset returns.

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

* Current Version: March 15, 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 Thomas Flury, Hagen Kim and Marco Rossi. We thank Yong Chen for supplying us with the mutual fund data. We thank Gene Fama and Ken French 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 regression method that allows us to better navigate through 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 random chance. 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 Fama-MacBeth type of regressions?

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 apparent 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%, completely nullifying the 2.0 cutoff in single tests.[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 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 by a factor of 20 ($= 0.99/0.05$). 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 Yong (2014).

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 perform multiple testing and variable selection within a given regression model.

Our paper adds to the recent literature on the multidimensionality of the cross-section of expected returns. Harvey, Liu and Zhu (2015) 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 (2014) 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 Yong, 2014) 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.

Our paper is organized as follows. In the second section, we present our testing framework. In the third section, we first illustrate the insights of our method by examining mutual fund performance evaluation. We then apply our method to the selection of risk factors. 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 $\Psi$. Our framework permits the use of an arbitrary performance measure $\Psi$, 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 indi-

vidually 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 realizations instead of computer-based simulations. Second, we use bootstrap and block bootstrap to approximate the empirical distribution of test statistics. FSW's approach cannot incorporate higher moment dependence or time-series dependence as the simulations will likely fall short of mimicking the true data structure.

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.[4] 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.[5]

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 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., $Y^{e,k}$.

[4] 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.

[5] See Adler, Feldman and Taqqu (1998) for a general treatment.

***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 $\Psi^b_I$, i.e.,

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

Intuitively, $\Psi^b_I$ 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 models 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 simulate the 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.[6] 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 $\Psi_I$ or $\Psi_I^b$ 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 $\Psi_I$, 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.[7]

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

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

[7] 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).

***Step III: Hypothesis Testing and Variable Selection***

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

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

where $(\Psi_I)_{1-\alpha}^B$ 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., $\Psi_I^d$) 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.[8]

## 2.2 GRS and Panel Regression Models

Our method can be adapted to study standard time-series tests such as the Gibbons, Ross and Shanken (GRS, 1989) test. The idea is to 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

[8] We plan to conduct a simulation study and benchmark our results to existing models, including FSW (1997) and Bonferroni/Holm/BHY as in Harvey, Liu and Zhu (2015) and Harvey and Liu (2014a). Different forms of cross-sectional and time-series dependence will be examined to evaluate the performance of our method in comparison with others.

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 and is the basis of the GRS statistic.

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, 2014). 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 has difficulty in answering 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 on average explain more than 80% of both 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 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 nonnormality distortions. In fact, bootstrap based resampling techniques are often recommended as the cure for these sources of bias.

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

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. 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 Equation (5). By subtracting the mean from the time-series of $f_{1t}$, we rewrite Equation (5) as

$$R_{it} - R_{ft} = \underbrace{[a_i + b_{i1}E(f_{1t})]}_{\text{Mean excess return}=E(R_{it}-R_{ft})} + b_{i1} \underbrace{[f_{1t} - E(f_{1t})]}_{\text{Demeaned factor return}} + \epsilon_{it}. \tag{6}$$

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 Equation (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., $b_i$) 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 explains why some factors (e.g., the market factor) can generate large time-series regression $R^2$'s but do little in explaining the cross-section of asset returns.

Another important observation from Equation (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., $a_i$) 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 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. Thus defined 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 mispricing. Candidate statistics include mean/median absolute pricing errors, mean squared pricing errors, and 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 use the regression residual as our new pseudo 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.

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. When multiple testing is the concern and we need to choose from a set of candidate variables, we can rely on the max statistic (in this case, the min statistic since minimizing the average pricing error is the objective) discussed in the previous section to provide inference.

## 2.3 Cross-sectional Regressions

Our method can 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) — one of the most important testing frameworks that allow time-varying risk loadings.

One hurdle in applying our method to FM regressions is the time-varying cross-sectional slopes. In particular, separate cross-sectional regressions are performed in each time period to capture the variability in regression slopes. We test the significance of a factor by looking at the average 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.

Suppose the vectors of residual excess returns are $Y_1, Y_2, \ldots, Y_T$ and the corresponding vectors of risk loadings (i.e., $\beta$'s) for a certain factor are $X_1, X_2, \ldots, X_T$.[10] Suppose there are $n_i$ stocks or portfolios in the cross-section at time $i, i = 1, \ldots, T$. Notice that the number of stocks or portfolios in the cross-section may vary across

[10]The vectors of residual excess returns should be understood as the FM regression residuals based on a pre-selected set of variables.

time so the vectors of excess returns may not have the same length. We start by running the following regressions:

$$\begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_T \end{bmatrix}_{\sum_{i=1}^{T} n_i \times 1} = \begin{bmatrix} \phi_1 \\ \phi_2 \\ \vdots \\ \phi_T \end{bmatrix}_{\sum_{i=1}^{T} n_i \times 1} + \xi_{1\times 1} \cdot \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_T \end{bmatrix}_{\sum_{i=1}^{T} n_i \times 1} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_T \end{bmatrix}_{\sum_{i=1}^{T} n_i \times 1}, \quad (7)$$

where $[\phi_1', \phi_2', \ldots, \phi_T']'$ is the vector of intercepts that are time-dependent, $\xi$ is a scalar that is time-independent, and $[\varepsilon_1', \varepsilon_2', \ldots, \varepsilon_T']'$ is the vector of projected regressors that will be used in the follow-up bootstrap analysis. The components in the vector of intercepts (i.e., $\phi_i, i = 1, \ldots, T$) are the same, so within each period we have the same intercept across stocks or portfolios.

Notice that the above regression pools returns and factor loadings together to estimate a single slope parameter, similar to what we do in predictive regressions. 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 the left-hand size variable out of the right-hand variable, we need to allow for time-varying intercepts as well. Mathematically, the time-dependent intercepts allow the regression residuals to sum up to zero in 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 $\varepsilon$ by its sum of squares $\varepsilon'\varepsilon$ and generate the orthogonalized regressor vectors:

$$X_i^e = \varepsilon_i/(\varepsilon_i'\varepsilon_i), \ i = 1, 2, \ldots, T. \quad (8)$$

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 OLS with these orthogonalized regressor vectors for each period:

$$Y_i = \mu_i + \gamma_i X_i^e + \eta_i, \ i = 1, 2, \ldots, T, \quad (9)$$

where $\mu_i$ is the $n_i \times 1$ vector of intercepts, $\gamma_i$ is the scalar slope for the $i$-th period, and $\eta_i$ is the $n_i \times 1$ vector of residuals. We can show that the following FM null hypothesis holds in-sample:

$$\sum_{i=1}^{T} \gamma_i = 0. \quad (10)$$

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 differs in different 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 cross-sectional regressions.

## 3 Results

### 3.1 Luck versus Skill: A Motivating Example

We first study mutual fund performance to illustrate the two key ingredients in our approach: orthogonalization and sequential selection. For performance evaluation, orthogonalization amounts to setting the in-sample alphas relative to a benchmark model at zero. By doing this, we are able to make inference on the overall performance of mutual funds by bootstrapping from the joint null that all funds generate a mean return of zero. The approach follows FF (2010). We then add something new. We use sequential selection to estimate the fraction of funds that generate nonzero returns, that is, funds that do not follow the null hypothesis.

We obtain the mutual fund data used in Ferson and Yong (2014). Their fund data is from the Center for Research in Security Prices Mutual Fund database. They focus on active, domestic equity funds covering the 1984-2011 period. To mitigate omission bias (Elton, Gruber and Blake, 2001) and incubation and back-fill bias (Evans, 2010), they apply several screening procedures. They limit their tests to funds that have

initial total net assets (TNA) above $10 million and have more than 80% of their holdings in stock in their first year to enter our data. They combine multiple share classes for a fund and use TNA-weighted aggregate share class. We require that a fund has at least twelve months of return history to enter our test. These leave us with a sample of 3716 mutual funds for the 1984-2011 period.[11]

We use the three-factor model in Fama and French (1993) as our benchmark model. For each fund in our sample, we project its returns in excess of the Treasury bill rates onto the three factors and obtain the regression intercept — alpha. We then calculate the t-statistic for the alpha (t($\alpha$)), which is usually called the precision-adjusted alpha.

The estimation strategy is as follows. We start from the overall null hypothesis that all funds generate an alpha of zero. To create this null for the realized data, we subtract the in-sample fitted alphas from fund returns so that each fund exactly has an alpha of zero. We then run bootstrapped simulations to generate a distribution of the cross-section of t($\alpha$)'s. In particular, for each simulation run we randomly sample the time periods and then calculate the cross-section of t($\alpha$)'s. In our simulations, we make sure that the same random sample of months applies to all funds, similar to Fama and French (2010). To draw inference on the null hypothesis, we compare the empirical distribution of the cross-section of t($\alpha$)'s with the realized t($\alpha$) cross-section for the original data. Following White (2000), we focus on the extreme percentiles to provide inference.[12]

The first row of Table 1 presents the results. We focus on the 0.5th and 99.5th percentiles.[13] The 99.5th percentile of t($\alpha$) for the original mutual fund data is 2.688, indicating a single test p-value of 0.4%. Without multiple testing concerns, we would declare the fund with such a high t($\alpha$) significant. With multiple testing, the bootstrapped p-value is 56.4%. This means that, by randomly sampling from the null hypothesis that all funds are generating a zero alpha, the chance for us to observe a 99.5th percentile that is at least 2.688 is 56.4%. We therefore fail to reject the null and conclude that there is no outperforming funds. Our results are consistent with Fama and French (2010), who also found that under the overall null of zero fund returns, there is no evidence for the existence of outperforming funds.

On the other hand, the 0.5th percentile in the data is -4.265 and its bootstrapped p-value is 0.6%. At the 5% significance level, we would reject the null and conclude

[11]We thank Yong Chen for providing us with the mutual fund data used in Ferson and Yong (2014).

[12]White (2000) uses the max statistic. However, the max statistic is problematic for unbalanced panels, like the mutual fund data. This is because funds with a short return history may have few nonrepetitive observations in a bootstrapped sample. This leads to extreme values of test statistics. To alleviate this, we can require a long return history (e.g., five years). But this leaves us with too few funds in the data.

[13]Some funds in our sample have short histories. When we bootstrap, the number of time periods with distinct observations could be even smaller because we may sample the same time period multiple times. The small sample size leads to extreme values of t($\alpha$). To alleviate this problem, we truncate the top and bottom 0.5% of t($\alpha$)'s and focus on the 0.5th and 99.5th percentile.

Table 1: **Tests on Mutual Funds, 1984-2011**

Test results on the cross-section of mutual funds. We use a three-factor benchmark model to evaluate fund performance. "Marginal t($\alpha$)" is the t($\alpha$) cutoff for the original data. "P-value" is the bootstrapped p-value either for outperformance (i.e., the 99.5th percentile) or underperformance (i.e., the rest of the percentiles). "Average t($\alpha$)" calculates the mean of t($\alpha$) conditional on falling above (i.e., the 99.5th percentile) or below (i.e., the rest of the percentiles) the corresponding t($\alpha$) percentiles.

| Percentile | Marginal t($\alpha$) | P-value | Average t($\alpha$) |
|---|---|---|---|
| Outperform | | | |
| 99.5 | 2.688 | 0.564 | 2.913 |
| Underperform | | | |
| 0.5 | -4.265 | 0.006 | -4.904 |
| 1.0 | -3.946 | 0.027 | -4.502 |
| 2.0 | -3.366 | 0.037 | -4.074 |
| 5.0 | -3.216 | 0.041 | -3.913 |
| 8.0 | -2.960 | 0.050 | -3.603 |
| 9.0 | -2.887 | 0.054 | -3.527 |
| 10.0 | -2.819 | 0.055 | -3.458 |

that there exist funds that are significantly underperforming. Until now, we exactly follow the procedure in FF (2010) and basically replicate their main results.

Rejecting the overall null of no performance is only the first step. Our method allows us to sequentially find the fraction of funds that are underperforming. By doing this, we step beyond Fama and French (2010). In particular, we sequentially add back the in-sample fitted alphas to funds. Suppose we add back the alphas for the bottom $q$ percent of funds. We then redo bootstrap to generate a new distribution of the cross-section of t($\alpha$)'s. Notice that this distribution is essentially based on the hypothesis that the bottom $q$ percent of funds are indeed underperforming, their alphas equal their in-sample fitted alphas, and the top $1-q$ percent of funds generate a mean of zero. To test this hypothesis, we compare the $q$-th percentile of t($\alpha$) for the original data with the bootstrapped distribution of the $q$-th percentile.

Table 1 shows the results. When only the bottom 1% are assumed to be underperforming, the bootstrapped p-value for the 1st percentile is 2.7%. This is higher than the p-value for the 0.5th percentile (i.e., 0.6%) but still lower than the 5% significance level. We gradually increase the number of funds that are underperforming. When the bottom 8% of funds are assumed to be underperforming, we exactly achieve a p-value of 5%. We therefore conclude that 8% of mutual funds are significantly underperforming. Among this group of underperforming funds, the average t($\alpha$) is

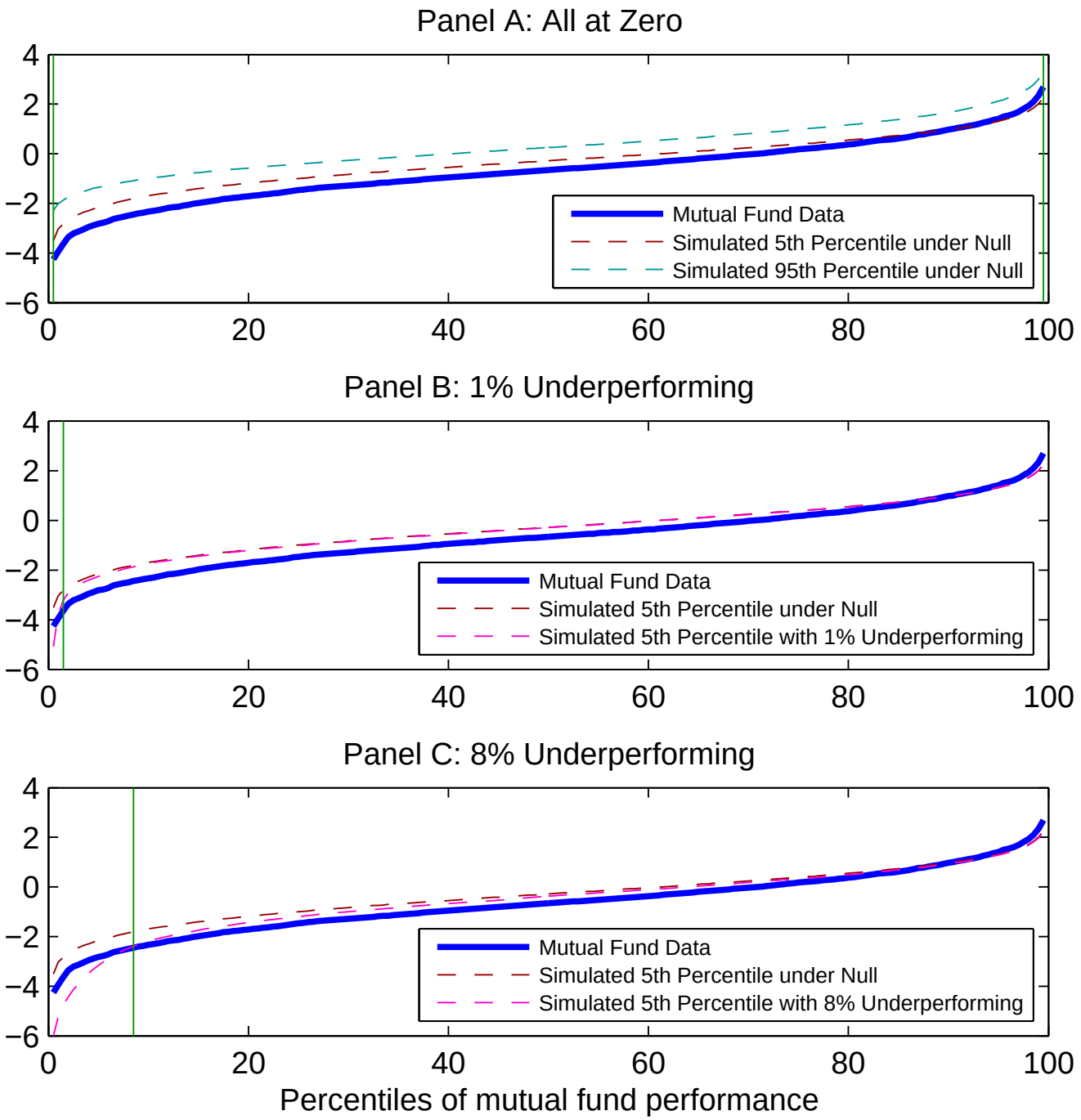
-3.603. The marginal t-ratio is -2.960, which corresponds to a single test p-value of 0.15%. Therefore, due to multiple testing, a fund needs to overcome a more stringent hurdle in order to be considered underperforming.

We compare our results to existing studies. FF (2010) focus on testing under the overall null of no skill and do not offer an estimate of the fraction of underperforming/outperforming funds. They also examine the dispersion in fund alphas. They assume that fund alphas follow a normal distribution and use bootstrap to estimate the dispersion parameter. Their approach is parametric in nature and differs from ours.

Our results are broadly consistent with Ferson and Yong (2014). By refining the false discovery rate approach in Barras, Scaillet and Wermers (2010), they also find that mutual funds are best classified into two groups: one group generates a mean return of zero and the other generates negative returns. However, their estimate the fraction of underperforming funds (20%) differs from ours. They follow Barras, Scaillet and Wermers (2010) and assume a parametric distribution of the means of returns for underperforming funds. We do not need to make such a distributional assumption. Rather, we rely on the in-sample fitted means to estimate this distribution in a nonparametric fashion.

Figure 1 provides a visualization of how our method is implemented. Panel A shows that under the overall null of no outperforming/underperforming funds, the 99.5th percentile of t($\alpha$) falls below the 95% confidence band and the 0.5th percentile falls below the 5% confidence band. In Panel B, we set the alphas for the bottom 1% of funds at their in-sample fitted values. Under this, the simulated 5% confidence band is below the band under the overall null but still above the 1st percentile of t($\alpha$). In Panel C, when the bottom 8% are classified as underperforming funds, the 8th percentile of t($\alpha$) meets the 5% confidence band.

Figure 1: **Estimating the Fraction of Underperforming Funds**



It is important to link what we learn from this example to the insights of our method in general. First, othogonalization amounts to subtracting the in-sample alpha estimate from fund returns in the context of performance evaluation. This helps create "pseudo" funds that behave just like the funds observed in the sample but have an alpha of zero. This insight has recently been applied by the literature to study fund performance. Our focus is regression models. In particular, we orthogonalize regressors such that their in-sample slopes and risk premium estimates exactly equal zero for predictive regressions and Fama-MacBeth regressions, respectively. This orthogonalization is the basis for the follow-up bootstrap exercise.

Second, in contrast to recent papers studying mutual fund performance, we incrementally identify the group of underperforming funds through hypothesis testing. The key is to form a new null hypothesis at each step by including more underperforming funds and test the significance of the marginal fund through bootstrapping. Our procedure is new and provides additional insights to the performance evaluation literature. We now apply this insight to the problem of variable selection in regression models. Two features make our method more attractive compared to existing methods. Through bootstrapping, we avoid the need of appealing to asymptotic theories to provide inference. In addition, our method can easily accommodate different types of summary statistics that are used to differentiate candidate models (e.g., $R^2$, t-statistic, etc.).

## 3.2 Identifying Factors

Next, we provide an example that focuses on risk factors. In principle, we can apply our method to the grand task of sorting out all the risk factors that have been proposed in the literature. 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 an illustrative example. The choice of the test portfolios is the issue. Different test portfolios lead to different results. In contrast, individual stocks avoid the arbitrary portfolio construction. We discuss the possibility of applying our method to individual stocks in the next section. Our illustrative example in this section shows how our method can be applied based on some popular test portfolios and a set of prominent risk factors.

In particular, we apply our panel regression method to ten risk factors that are proposed by Fama and French (2014), Frazzini and Pedersen (2014), Novy-Marx (2013), Pastor and Stambaugh (2003), Carhart (1997), and Asness, Frazzini and Pedersen (2013).[14] For test portfolios, we use the standard 25 size and book-to-market sorted portfolios that are available from Ken French's on-line data library.

[14]Except for the factors in Fama and French (2014), we obtain the data for the rest of the factors from the authors' webpages. Across the ten factors, the liquidity factor in Pastor and Stambaugh

We first provide acronyms for factors. Fama and French (2014) 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. Other factors include 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), and quality minus junk ($qmj$) in Asness, Frazzini and Pedersen (2013). We treat these ten 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 portfolios that are used to estimate the model.[15] Hence, this example should only be viewed as illustrative of a new method. We focus on tests that rely on time-series regressions, similar to Fama and French (2014).

Table 2 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 ten risk factors generate sizable long-short strategy returns. Six of the strategy returns generate t-ratios above 3.0 which is the level advocated by Harvey, Liu and Zhu (2015) to take multiple testing into account. The correlation matrix shows that book-to-market ($hml$) and investment ($cma$) have a correlation of 0.7 and profitability ($rmw$) and quality minus junk ($qmj$) have a correlation of 0.76. These high levels of correlations might pose a challenge for standard estimation approaches. However, our method takes the correlation into account.

To measure the goodness-of-fit of a candidate model, we need a performance metric. Similar to Fama and French (2014), we define four intuitive metrics that capture the cross-sectional goodness-of-fit of a regression model. For a panel regression model with $N$ portfolios, let the regression intercepts be $\{a_i\}_{i=1}^N$ and the cross-sectionally adjusted mean portfolio returns (i.e., the average return on a portfolio minus the average of the cross-section of portfolio returns) be $\{\bar{r}_i\}_{i=1}^N$. The four metrics are the median absolute intercept ($m_1^a$), the mean absolute intercept ($m_1^b$), the mean absolute intercept over the average absolute value of $\bar{r}_i$ ($m_2$), and the mean squared intercept over the average squared value of $\bar{r}_i$ ($m_3$). Fama and French (2014) focus on the last three metrics (i.e., $m_1^b$, $m_2$ and $m_3$). We augment them with $m_1^a$, which we believe is more robust to outliers compared to $m_1^b$.[16] When the null hypothesis of the GRS test is true, the cross-section of intercepts should all be zero and so are the four metrics.

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

(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.

[15]Harvey and Liu (2014b) explore the portfolio selection issue in detail. They advocate the use of individual stocks to avoid subjectivity in the selection of the characteristics used to sort the test portfolios.

[16]See Harvey and Liu (2014b) for an exploration on the relative performances of different metrics.

Table 2: **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 (2014) (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), and quality minus junk (*qmj*) in Asness, Frazzini and Pedersen (2013). 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 | 0.023 | 0.091 | 0.096 | 0.116 | 0.132 |
| 2 | 0.050 | 0.081 | 0.108 | 0.108 | 0.117 |
| 3 | 0.055 | 0.088 | 0.090 | 0.101 | 0.123 |
| 4 | 0.066 | 0.064 | 0.081 | 0.096 | 0.097 |
| Big | 0.050 | 0.057 | 0.052 | 0.063 | 0.068 |

Panel B.1: Factor Returns

| | *mkt* | *smb* | *hml* | *rmw* | *cma* | *bab* | *gp* | *psl* | *mom* | *qmj* |
|---|---|---|---|---|---|---|---|---|---|---|
| Mean | 0.052 | 0.025 | 0.048 | 0.033 | 0.047 | 0.105 | 0.039 | 0.054 | 0.081 | 0.048 |
| t-stat | [2.17] | [1.58] | [3.09] | [2.92] | [4.44] | [5.98] | [3.24] | [2.94] | [3.54] | [3.74] |

Panel B.2: Factor Correlation Matrix

| | *mkt* | *smb* | *hml* | *rmw* | *cma* | *bab* | *gp* | *psl* | *mom* | *qmj* |
|---|---|---|---|---|---|---|---|---|---|---|
| *mkt* | 1.00 | | | | | | | | | |
| *smb* | 0.28 | 1.00 | | | | | | | | |
| *hml* | -0.30 | -0.12 | 1.00 | | | | | | | |
| *rmw* | -0.21 | -0.36 | 0.08 | 1.00 | | | | | | |
| *cma* | -0.40 | -0.11 | 0.70 | -0.11 | 1.00 | | | | | |
| *bab* | -0.09 | -0.02 | 0.40 | 0.26 | 0.32 | 1.00 | | | | |
| *gp* | 0.08 | 0.03 | -0.34 | 0.49 | -0.34 | -0.11 | 1.00 | | | |
| *psl* | -0.05 | -0.04 | 0.03 | 0.04 | 0.03 | 0.06 | 0.03 | 1.00 | | |
| *mom* | -0.14 | -0.05 | -0.15 | 0.10 | 0.01 | 0.18 | 0.01 | -0.03 | 1. 00 | |
| *qmj* | -0.54 | -0.53 | 0.02 | 0.76 | 0.07 | 0.19 | 0.45 | 0.04 | 0.26 | 1.00 |

reason is that, while the othogonalized factor by construction has zero impact on the cross-section of expected returns, it may still affect the error covariance matrix. Since the GRS statistic uses the error covariance matrix to weight the regression intercepts, it changes as the estimate for the covariance matrix changes. We think the GRS statistic is not appropriate in our framework as the weighting function 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 four aforementioned metrics that do not rely on a model-based weighting matrix.

We start by testing whether any of the ten factors is individually significant in explaining the cross-section of expected returns. Panel A in Table 3.2 presents the results. Across the four metrics, the market factor appears to be the best among the candidate factors. For instance, it generates a median absolute intercept of 0.24% per month, much lower than what the other four factors generate. To evaluate the significance of the market factor, we follow our method and orthogonalize the ten factors so they have a zero impact on the cross-section of expected returns in-sample. We then bootstrap to obtain the empirical distributions of the minimums of the test statistics under the three metrics. For instance, the median of the minimal $m_1^a$ in the bootstrapped distribution is 0.56%. Evaluating the minimum statistics based on the real data against the bootstrapped distributions, the corresponding p-values are 3.6% for $m_1^a$, 3.3% for $m_1^b$, 5.8% for $m_2$, and 11.6% for $m_3$. In general, we believe that the first two metrics are more powerful than the rest of the metrics.[17] Based on the first two metrics, we would conclude that the market factor is significant at the 5% significance level.

With the market factor identified as a "true" factor, we include it in the baseline model and continue to test the other nine factors. We orthogonalize these nine factors against the market factor and obtain the residuals. By construction, these residuals maintain the time-series predictability of the original factors but have no impact on the cross-section of pricing errors (because the residuals are mean zero) that are produced when the market factor is used as the single factor. Panel B presents the results. This time, no single factor appears to dominate. In particular, *hml* wins on $m_1^a$ and is beaten by *cma* on the other three metrics. As a result, the minimum test statistic on the real data is generated by *hml* for $m_1^a$, and by *cma* for the other metrics. This lack of consistency across test statistics matters little for our test at the current stage. Indeed, the four bootstrapped p-values are all well below 5% and we conclude with confidence that at least one of the nine factors offers explanatory power for the cross-section of expected returns in addition to the market factor.

But between *hml* and *cma*, which one do we select for the next stage test? We choose *hml* for two reasons. First, it beats *cma* on $m_1^a$, which is more robust to outliers than $m_1^b$. Second, *hml*, as a empirical factor, has a longer history than *cma*. It has an out-of-sample history from 1992.

Notice that in Panel B, with *mkt* as the only factor, the median GRS is 5.468 in the bootstrapped distribution, much larger than 4.291 in Panel A, which is the GRS for the real data with *mkt* as the only factor. This means that by adding one of the demeaned factors (i.e., demeaned *smb*, *hml*, etc.), the GRS becomes much larger. By construction, these demeaned factors have no impact on the intercepts. The only way they can affect GRS is through the error covariance matrix. Hence, the demeaned

[17] See Harvey and Liu (2014b) for the demonstration.

factors make the GRS larger by reducing the error variance estimates. This insight also explains the discrepancy between $m_1^b$ and GRS in Panel A: *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 weights on regression intercepts compared to *bab*. The weighting in GRS does not seem appropriate for model comparison when none of the models is expected to be true. 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 the four metrics that do not depend on the error covariance matrix estimate.

Now with *mkt* and *hml* both in the list of pre-selected variables, we proceed to test the other eight variables. Panel C presents the results. Across the eight variables, *smb* appears to be the best across the first three metrics. However, the corresponding p-values are 5.3%, 13.5% and 8.7%, respectively. Judging at the conventional level, we would declare *smb* as insignificant and stop at this step and declare *mkt* and *hml* as the only two significant factors.

Suppose we also declare *smb* as significant and include it in the list of pre-selected variables, we test the remaining seven factors in Panel D. To be clear, our baseline model has all the factors in Fama and French (1993) and we are testing whether any of the seven recently proposed factors is significant. Our results say no, based on these particular portfolios. Indeed, the bootstrapped p-value for $m_1^a$ — our preferred metric — is 9.4%, which is not sufficient to declare significance at the 5% level. Let us emphasize this is an illustration of our method. A different set of factors might be selected from a different rule for portfolio formation. Nevertheless, our method provides a way to discriminate between nested models.

Table 3: **How Many Factors?**

Testing results on ten risk factors. We use Fama-French size and book-to-market sorted portfolios to test ten risk factors. They are excess market return (*mkt*), size (*smb*), book-to-market (*hml*), profitability (*rmw*), and investment (*cma*)) in Fama and French (2014), 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), and quality minus junk (*qmj*) in Asness, Frazzini and Pedersen (2013). 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 four performance metrics are the median absolute intercept ($m_1^a$), the mean absolute intercept ($m_1$), the mean absolute intercept over the average absolute value of the demeaned portfolio return ($m_2$), and the mean squared intercept over the average squared value of the demeaned portfolio returns ($m_3$). GRS reports the Gibbons, Ross and Shanken (1989) test statistic.

| | | $m_1^a(\%)$ | $m_1^b(\%)$ | $m_2$ | $m_3$ | GRS |
|---|---|---|---|---|---|---|
| Panel A: Baseline = No Factor | | | | | | |
| Real data | *mkt* | **0.240** | **0.259** | **1.485** | **1.656** | 4.291 |
| | *smb* | 0.477 | 0.467 | 2.678 | 4.344 | 4.370 |
| | *hml* | 0.801 | 0.770 | 4.409 | 11.914 | 4.046 |
| | *rmw* | 0.837 | 0.816 | 4.676 | 13.955 | 4.325 |
| | *cma* | 1.004 | 0.935 | 5.355 | 17.883 | 4.238 |
| | *bab* | 0.680 | 0.670 | 3.836 | 8.644 | **3.718** |
| | *gp* | 0.631 | 0.601 | 3.445 | 7.562 | 4.096 |
| | *psl* | 0.688 | 0.660 | 3.781 | 8.864 | 4.292 |
| | *mom* | 0.817 | 0.773 | 4.431 | 12.569 | 4.301 |
| | *qmj* | 1.181 | 1.174 | 6.725 | 29.754 | 5.594 |
| | Min | 0.240 | 0.259 | 1.485 | 1.656 | 3.718 |
| Bootstrap | Median of min | 0.556 | 0.547 | 2.873 | 5.335 | 5.716 |
| | p-value | 0.036 | 0.033 | 0.058 | 0.116 | 0.011 |
| Panel B: Baseline = mkt | | | | | | |
| Real data | *smb* | 0.207 | 0.231 | 1.321 | 1.590 | 4.237 |
| | *hml* | **0.115** | 0.144 | 0.827 | 0.318 | 3.573 |
| | *rmw* | 0.243 | 0.275 | 1.574 | 2.079 | 3.953 |
| | *cma* | 0.124 | **0.132** | **0.754** | **0.192** | 3.434 |
| | *bab* | 0.123 | 0.146 | 0.839 | 0.346 | **3.421** |
| | *gp* | 0.282 | 0.302 | 1.732 | 2.123 | 3.922 |
| | *psl* | 0.223 | 0.251 | 1.438 | 1.549 | 4.060 |
| | *mom* | 0.287 | 0.310 | 1.775 | 2.336 | 4.028 |
| | *qmj* | 0.361 | 0.397 | 2.274 | 4.791 | 4.512 |
| | Min | 0.115 | 0.132 | 0.754 | 0.192 | 3.421 |
| Bootstrap | Median of min | 0.207 | 0.235 | 1.233 | 1.253 | 5.468 |

*Continued on next page*

Table 3 – *Continued from previous page*

| | | $m_1^a(\%)$ | $m_1^b(\%)$ | $m_2$ | $m_3$ | GRS |
|---|---|---|---|---|---|---|
| | p-value | 0.047 | 0.003 | 0.000 | 0.000 | 0.007 |
| Panel C: Baseline = mkt + hml | | | | | | |
| Real data | *smb* | **0.081** | **0.102** | **0.585** | 0.352 | 3.514 |
| | *rmw* | 0.170 | 0.193 | 1.104 | 0.896 | 3.304 |
| | *cma* | 0.126 | 0.131 | 0.752 | 0.227 | 3.431 |
| | *bab* | 0.127 | 0.134 | 0.769 | 0.221 | 3.311 |
| | *gp* | 0.108 | 0.121 | 0.693 | **0.147** | **2.743** |
| | *psl* | 0.116 | 0.141 | 0.809 | 0.300 | 3.380 |
| | *mom* | 0.148 | 0.153 | 0.876 | 0.366 | 3.015 |
| | *qmj* | 0.367 | 0.343 | 1.965 | 3.438 | 3.392 |
| | Min | 0.081 | 0.102 | 0.585 | 0.147 | 2.743 |
| Bootstrap | Median of min | 0.117 | 0.138 | 0.728 | 0.343 | 4.690 |
| | p-value | 0.053 | 0.135 | 0.087 | 0.016 | 0.007 |
| Panel D: Baseline = mkt + hml + smb | | | | | | |
| Real data | *rmw* | 0.083 | 0.097 | 0.556 | 0.236 | 3.068 |
| | *cma* | 0.094 | 0.101 | 0.577 | 0.369 | 3.395 |
| | *bab* | 0.084 | 0.098 | 0.559 | 0.289 | 3.263 |
| | *gp* | **0.062** | **0.084** | **0.480** | **0.168** | 2.682 |
| | *psl* | 0.075 | 0.099 | 0.569 | 0.325 | 3.308 |
| | *mom* | 0.077 | 0.096 | 0.549 | 0.278 | 2.975 |
| | *qmj* | 0.076 | 0.097 | 0.556 | 0.178 | **2.417** |
| | Min | 0.062 | 0.084 | 0.480 | 0.168 | 2.417 |
| Bootstrap | Median of min | 0.083 | 0.112 | 0.596 | 0.353 | 4.643 |
| | p-value | 0.096 | 0.008 | 0.131 | 0.029 | 0.001 |

## 3.3 Individual stocks

Using characteristics-sorted portfolios may be inappropriate because some of these portfolios are also used in the construction of risk factors. Projecting portfolio returns onto risk factors that are themselves functions of these portfolios may bias the results. Ahn, Conrad and Dittmar (2009) propose a characteristics-independent way of forming test portfolios. Ecker (2013) suggests the use of randomly generated portfolios that reduce the noise in individual stocks while at the same time do not bias towards existing risk factors. However, the complexity involved in constructing these test portfolios keeps researchers from applying these methods in practice. As a

result, the majority of researchers still use the readily available characteristics-sorted portfolios as test portfolios.

Our model can potentially bypass this issue by using individual stocks as test assets. Intuitively, individual stocks should provide the most reliable and unbiased source of information in asset pricing tests. The literature has argued that individual stocks are too noisy to provide powerful tests. Our panel regression model allows us to construct test statistics that are robust to noisy firm-level return observations (see Harvey and Liu (2014b)).

# 4 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 and allows for general distributional characteristics, cross-sectional as well as time-series dependency, and a range of test statistics.

Our applications at this point are only illustrative. However, our method is general. It can be used for time-series prediction. The method applies to the evaluation of fund management. Finally, it allows us, in an asset pricing application, to address the problem of lucky factors. In the face of hundreds of candidate variables, some factors will appear significant by chance. Our method provides a new way to separate the factors that are lucky from the ones that explain the cross-section of expected returns.

Finally, while we focus on the asset pricing implications, our technique can be applied to any regression model that faces the problem of multiple testing. Our framework applies to many important areas of corporate finance such as the variables that explain the cross-section of capital structure. 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.*

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.*

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

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

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, 2014, A five-factor asset pricing model, *Working Paper, University of Chicago.*

Ferson, W.E. and Y. Chen, 2014, How many good and bad fund managers are there, really? *Working Paper, USC.*

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.*

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

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.*

Harvey, C.R., Y. Liu and H. Zhu, 2015, ... and the cross-section of expected returns, *Forthcoming, Review of Financial Studies.*
SSRN: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2249314

Harvey, C.R. and Y. Liu, 2014a, Multiple testing in financial economics, *Working Paper, Duke University.*
SSRN: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2358214

Harvey, C.R. and Y. Liu, 2014b, A test of the incremental efficiency of a given portfolio, *Work In Progress, Duke University.*

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

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

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.*

McLean, R.D. and J. Pontiff, 2014, Does academic research destroy stock return predictability? *Working Paper, University of Alberta.*

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.*

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

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

# A Proof for Fama-MacBeth Regressions

The corresponding objective function for the regression model in equation (7) is given by:

$$\mathcal{L} = \sum_{i=1}^{T} [X_i - (\phi_i + \xi Y_i)]'[X_i - (\phi_i + \xi Y_i)]. \tag{11}$$

Taking first order derivatives with respect to $\{\phi_i\}_{i=1}^T$ and $\xi$, respectively, we have

$$\frac{\partial \mathcal{L}}{\partial \phi_i} = \sum_{i=1}^{T} \iota_i' \varepsilon_i = 0, \ i = 1, \ldots, T, \tag{12}$$

$$\frac{\partial \mathcal{L}}{\partial \xi} = \sum_{i=1}^{T} Y_i' \varepsilon_i = 0, \tag{13}$$

where $\iota_i$ is a $n_i \times 1$ vector of ones. Equation (12) says that the residuals within each time period sum up to zero, and equation (13) says that the $Y_i$'s are on average orthogonal to the $\varepsilon_i$'s across time. Importantly, $Y_i$ is not necessarily orthogonal to $\varepsilon_i$ within each time period. As explained in the main text, we next define the orthogonalized regressor $X_i^e$ as the rescaled residuals, i.e.,

$$X_i^e = \varepsilon_i / (\varepsilon_i' \varepsilon_i), \ i = 1, \ldots, T. \tag{14}$$

Solving the OLS equation (9) for each time period, we have:

$$\gamma_i = (X_i^{e\prime} X_i^e)^{-1} X_i^{e\prime} (Y_i - \mu_i), \tag{15}$$

$$= (X_i^{e\prime} X_i^e)^{-1} X_i^{e\prime} Y_i - (X_i^{e\prime} X_i^e)^{-1} X_i^{e\prime} \mu_i, \ i = 1, \ldots, T. \tag{16}$$

We calculate the two components in equation (16) separately. First, notice $X_i^e$ is a rescaled version of $\varepsilon_i$. By equation (12), the second component (i.e., $(X_i^{e\prime} X_i^e)^{-1} X_i^{e\prime} \mu_i$) equals zero. The first component is calculated as:

$$(X_i^{e\prime} X_i^e)^{-1} X_i^{e\prime} Y_i = [(\frac{\varepsilon_i}{\varepsilon_i' \varepsilon_i})'(\frac{\varepsilon_i}{\varepsilon_i' \varepsilon_i})]^{-1} (\frac{\varepsilon_i}{\varepsilon_i' \varepsilon_i})' Y_i, \tag{17}$$

$$= \varepsilon_i' Y_i, \ i = 1, \ldots, T, \tag{18}$$

where we again use the definition of $X_i^e$ in equation (17). Hence, we have:

$$\gamma_i = \varepsilon_i' Y_i, \ i = 1, \ldots, T. \tag{19}$$

Finally, applying equation (13), we have:

$$\sum_{i=1}^{T} \gamma_i = \sum_{i=1}^{T} \varepsilon_i' Y_i = 0.$$

# B 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 $q$.

# C FAQ

## C.1 General Questions

- *Do we want the min (max) or, say, the 5th (95th) percentile? (Section 2)*

  Although the percentiles could be more robust to outliers, we think that the extremes are more powerful test statistics. For instance, suppose there is only one variable that is significant among 1,000 candidate variables. Then we are more likely to miss this variable (i.e., failing to reject the null that all 1,000 variables are insignificant) using the percentiles than using the extremes. Following White (2000), we use the min/max test statistics. Harvey and Liu (2014b) look further into the choice of the test statistics.

- *Can we "test down" for variable selection instead of "testing up" ? (Section 2)*

  Our method does not apply to the "test down" approach. 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 this technical difficulty, we think that "testing up" makes more sense for finance applications. For finance problems, as a prior, we usually do not believe that there should exist hundreds of variables explaining a certain phenomenon. "Testing up" is more consistent with this prior.