JOURNAL OF FINANCIAL AND QUANTITATIVE ANALYSIS Vol. 54, No. 4, Aug. 2019, pp. 1713–1758
COPYRIGHT 2018, MICHAEL G. FOSTER SCHOOL OF BUSINESS, UNIVERSITY OF WASHINGTON, SEATTLE, WA 98195
doi:10.1017/S0022109018000947

https://doi.org/10.1017/S0022109018000947 Published online by Cambridge University Press

Best of the Best: A Comparison of Factor Models

Shamim Ahmed, Ziwen Bu, and Daniel Tsvetanov*

Abstract

We compare major factor models and find that the Stambaugh and Yuan (2016) 4-factor model is the overall winner in the time-series domain. The Hou, Xue, and Zhang (2015) q-factor model takes second place and the Fama and French (2015) 5-factor model and the Barillas and Shanken (2018) 6-factor model jointly take third place. The pairwise cross-sectional R^2 and the multiple model comparison tests show that the Hou et al. (2015) q-factor model, the Fama and French (2015) 5-factor and 4-factor models, and the Barillas and Shanken (2018) 6-factor model take equal first place in the horse race.

I. Introduction

Starting with the classic capital asset pricing model of Sharpe (1964) and Lintner (1965), the finance literature has been in search for a model that explains the cross section of expected returns on assets. This had led to the development of a wide variety of factor pricing models over the past few years. Although these models are merely approximations of reality (Barillas and Shanken (2018), Gospodinov, Kan, and Robotti (2013), Kan and Robotti (2009), and Kan, Robotti, and Shanken (2013)), it is important from an academic and practitioner perspective to know which model provides the best overall description of asset returns. For example, there is ample evidence, both empirical and anecdotal, that portfolio managers most often use the capital asset pricing model and a variety of multi-factor models to compute expectations of returns (see, among others, Ang (2014), Brealey, Myers, and Allen (2016), Fischer and Wermers (2012), Gitman and Mercurio (1982), Graham and Harvey (2001), Grinold and Kahn (1995), and Jagannathan and Meier (2002)). In relation to this, Fama and French (2016) compare the performance of the recently proposed 5-factor model of Fama and French (2015), along with models that use subsets of its factors. Along this line of research,

*Ahmed (corresponding author), shamim.ahmed@nottingham.ac.uk, Bu, ziwen.bu@nottingham.ac.uk, University of Nottingham Business School; and Tsvetanov, D.Tsvetanov@uea.ac.uk, University of East Anglia Norwich Business School. We thank an anonymous referee, Abhay Abhyankar, Sanjay Banerji, Hendrik Bessembinder (the editor), Weimin Liu, Xiaoquan Liu, and Albert Menkveld for useful comments and suggestions.

Hou, Xue, and Zhang (2017a) and Stambaugh and Yuan (2016) also investigate the performance of several prominent asset pricing models. All of these studies examine the relative performance of factor pricing models in time-series regressions using various test asset portfolios, but without the use of any formal statistical procedure that takes into account the sampling and model misspecification uncertainty. However, Barillas and Shanken (2017), Harvey and Liu (2017), Kan and Robotti (2009), and Kan et al. (2013), among others, emphasize that the practice of identifying the best model(s) by simply comparing point estimates of pricing performance metrics can be misleading.

In this article, we go beyond the vast majority of empirical studies by statistically comparing the performance of a much larger array of both classic and new-generation asset pricing models in order to explain several prominent return anomalies. Our array comprises of the capital asset pricing model of Sharpe (1964) and Lintner (1965), the Fama and French (1993) 3-factor model, the Fama and French (1993) and Carhart (1997) 4-factor model, the Fama and French (1993) and Pástor and Stambaugh (2003) 4-factor model, the Asness and Frazzini (2013) 3-factor model, the Hou et al. (2015) q-factor model, the Fama and French (2015) 5-factor model, the 4-factor model of Fama and French (2015) that drops the "high minus low" value factor, the Stambaugh and Yuan (2016) 4-factor model, and the Barillas and Shanken (2018) 6-factor model. Although recent research (see Harvey, Liu, and Zhu (2016), McLean and Pontiff (2016)) documents hundreds of potential traded and nontraded factors that explain the cross-section of stock returns, we limit ourselves to these traded factor models for two reasons. Firstly, these models have remained the most prominent, if not the best, models over the years. Importantly, most of our chosen models are frequently used for risk-adjustment purposes in the empirical asset pricing literature (see, e.g., Bali, Cakici, and Whitelaw (2011), Bali, Engle, and Murray (2016), Gu (2016), Hirshleifer, Hsu, and Li (2018), and Kuehn, Simutin, and Wang (2017)). Secondly, focusing only on return factor models allows us to explore their performance in both the time-series and the cross-sectional domains. Moreover, given the plethora of models, investigating the 10 prominent factor models keeps the model comparison exercise at a reliably manageable level. Hence, our empirical examination of the relative model performance can be viewed as a search for the best of the best factor pricing models.

As in many recent papers (see, among others, Fama and French (2015), (2016), (2018), Hou, Karolyi, and Kho (2011), Hou et al. (2015), (2017a), and Stambaugh and Yuan (2016)), we begin by applying our chosen factor models to the data, and informally comparing them using a comprehensive set of pricing performance metrics estimated in the time-series domain. We then employ the pairwise test for equality of cross-sectional regression R^2s introduced by Kan et al. (2013). A key advantage of the sample cross-sectional R^2 test is that the testing procedure takes into account the impact of potential model misspecification, and therefore ensures robust and valid statistical inference regarding relative performance. This also enables us to observe any differences between our informal findings on relative model performance in the time-series domain and those of the cross-sectional statistical evidence. However, the pairwise model comparison may not always unambiguously determine the best factor model when multiple

https://doi.org/10.1017/S0022109018000947 Published online by Cambridge University Press

https://doi.org/10.1017/S0022109018000947 Published online by Cambridge University Press

models are involved in the horse race (Barillas, Kan, Robotti, and Shanken (2017), Gospodinov et al. (2013), Kan et al. (2013)). To address this concern, we also conduct a multiple model comparison, following Kan et al. (2013), which identifies whether a "benchmark" asset pricing model significantly outperforms a set of alternative models in terms of the cross-sectional regression R^2.

To investigate the strengths and weaknesses of competing factor pricing models, choosing test assets based on numerous predictive signals (see, e.g., Green, Hand, and Zhang (2014)) is a challenge. In this regard, we follow prior empirical studies (see, among others, Fama and French (1997), (2008), (2016), Feng, Giglio, and Xiu (2017), Kan et al. (2013), Kozak, Nagel, and Santosh (2018), and Stambaugh and Yuan (2016)), particularly the suggestion of Lewellen, Nagel, and Shanken (2010), while keeping our analysis parsimonious. We choose subsets of test assets representative of major anomaly categories, such as investment, profitability, intangibles, and trading frictions, documented in the literature (see Green et al. (2014), Hou et al. (2015), (2017a), Hou, Xue, and Zhang (2017b), and references therein). Specifically, the sets of test assets used in this article are the value-weighted monthly excess returns on: the 25 size-STR (short-term reversal) portfolios, the 25 size-CI (abnormal capital investment) portfolios, the 25 size-DR (distress risk) portfolios, the 25 size-β (market beta) portfolios, the 35 size-NI (net share issues) portfolios, the 25 size-MAX (lottery demand) portfolios, the 25 size-RVar (residual variance) portfolios, the 25 size-AC (accruals) portfolios, the 30 IND (industry) portfolios, and the 8 D10−1 (high minus low decile) portfolios. These test asset portfolios capture a vast cross-section of return anomalies; at the same time, a majority of them are readily available in the public domain, as described in the next section. More importantly, many of our test asset portfolios pose a greater challenge to existing asset pricing models (see, e.g., Fama and French (2016)). Different from Hou et al. (2017a) and Stambaugh and Yuan (2016), our rationale for using the eight sets of independent two-way sorted portfolios is that they allow us to see how well the competing factor models accommodate anomaly returns across size groups. In fact, asset pricing models have systematic problems explaining average returns on small stocks (see Fama and French (1993), (1996), (2015), (2016), (2018)).

Our time-series analysis using point estimates of an exhaustive set of pricing performance measures demonstrates that the Stambaugh and Yuan (2016) 4-factor model is the overall winner among all the asset pricing models in explaining anomalies. For example, the model outperforms, though informally, the competing models by producing the smallest point estimates for the average absolute alpha and a variant of the alpha dispersion metric for half of the sets of test asset portfolios. Considering the metric measuring the dispersion (second moment) of the alphas due to sampling error, we see that the Stambaugh and Yuan (2016) 4-factor model delivers the largest point estimate in four out of the ten sets of test portfolios. Moreover, the model generates the least number of statistically significant alphas for six portfolio sets. The Hou et al. (2015) q-factor model does a reasonably good job in accommodating anomaly returns, and takes second place in the horse race. For example, the q-factor model produces insignificant alphas for all 25 of the size-DR portfolios. The model also offers the largest point estimate for the dispersion (second moment) of the alphas, due to sampling

error, in three sets of test asset portfolios. Among the remaining competing models, the Fama and French (2015) 5-factor model and the Barillas and Shanken (2018) 6-factor model perform almost equally well, and take equal third place. Both deliver the largest point estimate for the average time-series regression R^2 in five out of the ten sets of portfolios. Looking at the 42 small size portfolios across the eight anomaly variables, we find that all of our asset pricing models have difficulties explaining their returns, an observation consistent with prior studies.

When the model comparison exercises are carried out based on pairwise tests of the equality of the sample cross-sectional R^2s and multiple model comparison tests, a different yet statistically valid picture emerges. Overall, we find that the Hou et al. (2015) q-factor model, the Fama and French (2015) 5-factor model, the 4-factor model of Fama and French (2015) that drops the value factor, and the Barillas and Shanken (2018) 6-factor model, all perform almost equally well, and thus jointly take first place in the horse race. More importantly, these asset pricing models are never statistically dominated at the 5% level in our multiple model comparison analysis. The Stambaugh and Yuan (2016) 4-factor model, which is the top performer in the time-series analysis, now takes the next place. Different from most studies in the time-series domain (see, e.g., Fama and French (1996), (2016), Hou et al. (2015), (2017a)), we also find that the capital asset pricing model of Sharpe (1964) and Lintner (1965) performs reasonably well. All of these findings remain robust, irrespective of cross-sectional regression methodologies and normal and sequential tests for nonnested models.

Taken together, our statistical evidence from the cross-sectional analysis reinforces the argument of Barillas and Shanken (2017), Harvey and Liu (2017), and Kan and Robotti (2009) that the common practice of identifying the superior model(s) by informally comparing pricing performance metrics can lead to erroneous conclusions. Importantly, the empirical results in this article also contribute to the growing literature on assessing the performance of asset pricing models, and have valuable implications for practical applications, including capital budgeting, equity valuation, quantitative investment management, and fund performance evaluation. For example, the Fama and French (1993) and Carhart (1997) 4-factor model is the current workhorse model used to evaluate mutual fund performance (see, among others, Ang (2014), Fama and French (2010), and Fischer and Wermers (2012)). Given our cross-sectional statistical evidence on the superior performance of the Hou et al. (2015) q-factor model, the Fama and French (2015) 5-factor and 4-factor models, and the Barillas and Shanken (2018) 6-factor model, they can all be applied in this area.

Our study differs from several recent papers that compare model performance, as we employ misspecification robust statistical tests on a much larger array of factor pricing models in the cross-sectional domain. Some of these papers include Fama and French (2016), Hou et al. (2017a), and Stambaugh and Yuan (2016). For instance, Fama and French (2016) show that the 5-factor model of Fama and French (2015) accommodates several return anomalies not explained by the Fama and French (1993) 3-factor model. The authors however, compare the performance of the Fama and French (2015) 5-factor model only with that of models that use subsets of those same five factors. Stambaugh and Yuan (2016)

https://doi.org/10.1017/S0022109018000947 Published online by Cambridge University Press

https://doi.org/10.1017/S0022109018000947 Published online by Cambridge University Press

compare the performance of their version of the 4-factor model with that of only three other models, namely the Fama and French (1993) 3-factor model, the Fama and French (2015) 5-factor model, and the Hou et al. (2015) q-factor model. Using one-way (univariate) sorted anomaly portfolios in the time-series regressions, the authors show that their 4-factor model, also known as the mispricing factor model, performs the best. Likewise, Hou et al. (2017a) show that the q-factor model of Hou et al. (2015) outperforms the Fama and French (2015) 5-factor model in explaining a wide array of asset pricing anomalies. However, their list of models in the horse race excludes the newly proposed 4-factor model of Stambaugh and Yuan (2016). More recently, Barillas and Shanken (2018) develop a Bayesian framework for comparing return factor models in terms of their posterior probabilities. Yet this framework requires formulating priors on the pricing errors, and therefore empirical results can be sensitive. In this context, the pairwise cross-sectional R^2 equality test of Kan et al. (2013) that we use in this paper is much simpler to implement where both traded and nontraded factor pricing models can be evaluated, and allows for fairly reliable statistical inferences to be drawn. Finally, Barillas et al. (2017) compare eight asset pricing models based on their maximum squared Sharpe ratios. But our model comparison analysis utilizing return anomalies also shows the strengths and weaknesses of the ten prominent asset pricing models.

The remainder of the paper is organized as follows: Section II describes the factor pricing models, the sets of return anomaly portfolios that are used as test assets, and the metrics for comparing model performance in explaining these anomalies. Section III discusses the results from the time-series regressions, while Section IV summarizes the results from the cross-sectional regressions. Finally, Section V concludes the paper. The Supplementary Material contains further details on test assets, robustness tests, and additional results.

II. Factor Models and Anomaly Portfolios

A. Competing Models

We examine the ability of ten different factor pricing models to explain the average excess returns of portfolios sorted on several well-known anomaly variables. These models are: i) the capital asset pricing model (CAPM) of Sharpe (1964) and Lintner (1965), which only includes a market factor; ii) the Fama and French (1993) 3-factor (FF3) model, which extends the CAPM by including empirically motivated size and value factors; iii) the Fama and French (1993) and Carhart (1997) 4-factor (FFC) model, which adds a momentum factor to the FF3 model; iv) the Fama and French (1993) and Pástor and Stambaugh (2003) 4-factor (FFPS) model, which combines a traded liquidity factor with those of the FF3 model; v) the Asness and Frazzini (2013) 3-factor (FFAF) model, which adds a more "timely" version of the value factor to the market and size factors of the FF3 model; vi) the Hou et al. (2015) q-factor (HXZ) model, comprising of market, size, investment, and profitability factors; vii) the Fama and French (2015) 5-factor (FF5) model, consisting of market, size, value, profitability, and investment factors; viii) the 4-factor (FF4) model, which drops the value factor of the FF5 model; ix) the Stambaugh and Yuan (2016) 4-factor (SY4) model, which

comprises of market, size, and two mispricing factors; and x) the Barillas and Shanken (2018) 6-factor (BS6) model, comprising of the market, size, and momentum factors of the FF5 model, the profitability and investment factors of the HXZ model, and the value factor of the FFAF model. The time-series regression specifications of these factor models are as follows:

$$
\begin{aligned}
(1)\quad r^e_{i,t} &= \alpha_{i,\text{CAPM}} + \beta_{i,\text{MKT}}\text{MKT}_t + \epsilon_{i,t},\\
(2)\quad r^e_{i,t} &= \alpha_{i,\text{FF3}} + \beta_{i,\text{MKT}}\text{MKT}_t + \beta_{i,\text{SMB}^*}\text{SMB}^*_t + \beta_{i,\text{HML}}\text{HML}_t + \epsilon_{i,t},\\
(3)\quad r^e_{i,t} &= \alpha_{i,\text{FFC}} + \beta_{i,\text{MKT}}\text{MKT}_t + \beta_{i,\text{SMB}^*}\text{SMB}^*_t\\
&\quad + \beta_{i,\text{HML}}\text{HML}_t + \beta_{i,\text{UMD}}\text{UMD}_t + \epsilon_{i,t},\\
(4)\quad r^e_{i,t} &= \alpha_{i,\text{FFPS}} + \beta_{i,\text{MKT}}\text{MKT}_t + \beta_{i,\text{SMB}^*}\text{SMB}^*_t\\
&\quad + \beta_{i,\text{HML}}\text{HML}_t + \beta_{i,\text{LIQ}}\text{LIQ}_t + \epsilon_{i,t},\\
(5)\quad r^e_{i,t} &= \alpha_{i,\text{FFAF}} + \beta_{i,\text{MKT}}\text{MKT}_t + \beta_{i,\text{SMB}^*}\text{SMB}^*_t + \beta_{i,\text{HML}^m}\text{HML}^m_t + \epsilon_{i,t},\\
(6)\quad r^e_{i,t} &= \alpha_{i,\text{HXZ}} + \beta_{i,\text{MKT}}\text{MKT}_t + \beta_{i,\text{ME}}r_{\text{ME},t}\\
&\quad + \beta_{i,\text{I/A}}r_{\text{I/A},t} + \beta_{i,\text{ROE}}r_{\text{ROE},t} + \epsilon_{i,t},\\
(7)\quad r^e_{i,t} &= \alpha_{i,\text{FF5}} + \beta_{i,\text{MKT}}\text{MKT}_t + \beta_{i,\text{SMB}}\text{SMB}_t\\
&\quad + \beta_{i,\text{HML}}\text{HML}_t + \beta_{i,\text{RMW}}\text{RMW}_t + \beta_{i,\text{CMA}}\text{CMA}_t + \epsilon_{i,t},\\
(8)\quad r^e_{i,t} &= \alpha_{i,\text{FF4}} + \beta_{i,\text{MKT}}\text{MKT}_t + \beta_{i,\text{SMB}}\text{SMB}_t\\
&\quad + \beta_{i,\text{RMW}}\text{RMW}_t + \beta_{i,\text{CMA}}\text{CMA}_t + \epsilon_{i,t},\\
(9)\quad r^e_{i,t} &= \alpha_{i,\text{SY4}} + \beta_{i,\text{MKT}}\text{MKT}_t + \beta_{i,\text{SMB}_M}\text{SMB}_{M,t}\\
&\quad + \beta_{i,\text{MGMT}}\text{MGMT}_t + \beta_{i,\text{PERF}}\text{PERF}_t + \epsilon_{i,t},\\
(10)\quad r^e_{i,t} &= \alpha_{i,\text{BS6}} + \beta_{i,\text{MKT}}\text{MKT}_t + \beta_{i,\text{SMB}}\text{SMB}_t + \beta_{i,\text{I/A}}r_{\text{I/A},t}\\
&\quad + \beta_{i,\text{ROE}}r_{\text{ROE},t} + \beta_{i,\text{HML}^m}\text{HML}^m_t + \beta_{i,\text{UMD}}\text{UMD}_t + \epsilon_{i,t},
\end{aligned}
$$

where $r^e_{i,t} = r_{i,t} - r_{f,t}$ is the period t return on asset i in excess of the risk-free rate; MKT_t, SMB^*_t (small minus big), and HML_t (high minus low) are, respectively, the market, size, and value factors of Fama and French (1993); LIQ_t is the traded liquidity factor of Pástor and Stambaugh (2003); HML^m_t is the value factor of Asness and Frazzini (2013); $r_{\text{ME},t}$, $r_{\text{I/A},t}$, and $r_{\text{ROE},t}$ are, respectively, the size, investment, and profitability factors of Hou et al. (2015); SMB_t, RMW_t (robust minus weak), CMA_t (conservative minus aggressive), and UMD_t (up minus down) are, respectively, the size, profitability, investment, and momentum factors of Fama and French (2015), (2016); and $\text{SMB}_{M,t}$, MGMT_t, and PERF_t are the size and two mispricing factors of Stambaugh and Yuan (2016), respectively.

Before proceeding to the factor model regression analysis in the subsequent sections, we now briefly describe the factors of interest. Fama and French (1993) construct the market factor, MKT, as the return on the value-weighted portfolio of all U.S.-based common stocks in the Center for Research in Security Prices database minus the 1-month Treasury bill rate from Ibbotson Associates. The SMB^* and HML factors come from independent 2×3 sorts of stocks on size (i.e., equity market capitalization) and the book-to-market equity (B/M) ratio. The intersections of the two size and three B/M groups produce six value-weighted portfolios. The original size factor, SMB^*, is the average of the returns on the three small stock portfolios minus the average of the returns on the three

https://doi.org/10.1017/S0022109018000947 Published online by Cambridge University Press

https://doi.org/10.1017/S0022109018000947 Published online by Cambridge University Press

big stock portfolios. The value factor, HML, is the average of the returns on the two high B/M portfolios minus the average of the returns on the two low B/M portfolios. Adopting a seemingly small modification to the HML value factor of the FF3 model, Asness and Frazzini (2013) construct a more "timely" version of the value factor, denoted HML^m, based on B/M rankings that use the most recent monthly stock price in the denominator.

Pástor and Stambaugh (2003) sort stocks into deciles of liquidity risk exposure estimates and construct their liquidity factor, LIQ, as the difference in value-weighted returns between the top and bottom decile portfolios. The excess return factors constructed by Hou et al. (2015) use independent $2 \times 3 \times 3$ sorts of stocks on size, investment measured by the investment-to-assets ratio (I/A), and profitability measured by return on equity (ROE). The intersections of the two size, three I/A, and three ROE groups produce 18 value-weighted portfolios. The size factor, r_{ME}, is the average of the nine small stock portfolio returns minus the average of the nine big stock portfolio returns. The investment factor, $r_{\text{I/A}}$, is the average of the six low I/A portfolio returns minus the average of the six high I/A portfolio returns. Similarly, the profitability factor, r_{ROE}, is the average of the returns on the six high ROE portfolios minus the average of the returns on the six low ROE portfolios. A feature of these excess return factors that distinguishes them from those of other asset pricing model factors is that they are constructed using stocks of nonfinancial firms with a nonnegative B/M only.

Departing from the original size factor SMB* of the FF3 model, Fama and French (2015), (2016) construct their size factor SMB as the average of the value-weighted returns on the nine small stock portfolios of the three independent 2×3 sorts, minus the average of the value-weighted returns on the nine big stock portfolios. The profitability, investment, and momentum factors are, respectively, from independent 2×3 sorts of stocks on size and operating profitability (OP), independent 2×3 sorts of stocks on size and investment (INV) measured by asset growth, and independent 2×3 sorts of stocks on size and momentum, denoted MOM (measured as the cumulative average return over the past 12 months, skipping the most recent month's return). In each case, the intersections of the sorts produce six value-weighted portfolios. The profitability factor, RMW, is then the average of the returns on the two high OP portfolios minus the average of the returns on the two low OP portfolios. Likewise, CMA is the average of the two low INV portfolio returns minus the average of the two high INV portfolio returns. The momentum factor, UMD, is the average of the two up (i.e., high value of MOM) portfolio returns minus the average of the two down (i.e., low value of MOM) portfolio returns. Fama and French (2015), (2016) show that their value factor, HML, is redundant in explaining the cross-section of the average U.S. stock returns. We, therefore, include the FF4 model, on the grounds of parsimony.[1] It is worth highlighting that although the FF4 model includes return factors that are designed to capture risk premiums similar to those of factors in the HXZ model, the factor constructions are very different in these models. For example,

[1] Xing (2008) introduces an investment growth factor and shows that the pricing factor contains information similar to that of the HML factor. Recently, Hou et al. (2015) claim that the value factor, HML, is a noisy version of their investment factor, $r_{\text{I/A}}$.

the profitability factor r_{ROE} is derived from monthly sorts on ROE, whereas RMW is obtained from annual sorts on OP.

Stambaugh and Yuan (2016) introduce two mispricing factors from a set of 11 prominent return anomalies documented in the literature (see, e.g., Stambaugh, Yu, and Yuan (2015)). The first mispricing factor, denoted MGMT, makes use of information from a cluster of anomaly variables, including NI, composite equity issues, AC, net operating assets, asset growth, and I/A. The second mispricing factor, PERF, comes from another cluster of anomalies, comprising of financial distress, O-score bankruptcy probability, momentum, gross profitability, and the return on assets. In particular, the authors construct two composite mispricing measures, P1 and P2, on a monthly basis for each stock by averaging its rankings with respect to the anomaly measures within each of the aforementioned clusters. Their next step entails the formation of six value-weighted portfolios from independent 2×3 sorts of stocks on market capitalization and a composite mispricing measure P1. The authors repeat the two-way sorting, with the exception of the second sort, which is on the composite mispricing measure P2. The mispricing factor MGMT, related to firm management, is then the average of the two low P1 (underpriced) portfolio returns minus the average of the two high P1 (overpriced) portfolio returns. The other mispricing factor, PERF, which is related more to firm performance, is the average of the returns on the two low P2 portfolios minus the average of the returns on the two high P2 portfolios. In addition to these factors, Stambaugh and Yuan (2016) construct their version of the size factor, denoted as SMB_M, using stocks that are least likely to be mispriced. Notably, the size factor SMB_M in a given month consists of stocks that are not used in forming either of the mispricing factors. As a result, the construction of SMB_M differs significantly from that of the SMB^* and SMB factors.

We source data on the risk-free rate, the MKT, SMB^*, SMB, HML, UMD, RMW, and CMA factors from the Internet Data Library maintained by Kenneth R. French (http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html). The data on the r_{ME}, $r_{\text{I/A}}$, and r_{ROE} factors is obtained from the Web site of Lu Zhang (https://sites.google.com/site/theqfactormodel/?pli=1). while the data on the LIQ factor is from the Web site of Ľuboš Pástor (http://faculty.chicagobooth.edu/lubos.pastor/research/). We collect data on the size factor SMB_M, and the two mispricing factors, MGMT and PERF, from the Web site of Yu Yuan (http://www.saif.sjtu.edu.cn/facultylist/yyuan/). The data on the alternative version of the value factor, HML^m, is sourced from the AQR Data Library (https://www.aqr.com/library/data-sets/the-devil-in-hmls-details-factors-monthly). Finally, our sample period spans Jan. 1968 to Dec. 2016. To conserve space, we provide factor correlations in Table IA1 of the Supplementary Material.

B. Anomaly Portfolios

To identify the strengths and weaknesses of our asset pricing models, we make use of several prominent return anomalies targeted in prior studies (see Bali et al. (2011), Fama and French (1997), (2016), (2018), Feng et al. (2017), Hou et al. (2017a), Kozak et al. (2018), and Linnainmaa and Roberts (2018)). The list comprises of STR (see Lehmann (1990), Jegadeesh (1990)), CI (see Titman,

https://doi.org/10.1017/S0022109018000947 Published online by Cambridge University Press

https://doi.org/10.1017/S0022109018000947 Published online by Cambridge University Press

Wei, and Xie (2004)), DR (see Bharath and Shumway (2008), Campbell, Hilscher, and Szilagyi (2008)), market β (see Black, Jensen, and Scholes (1972), Fama and MacBeth (1973), and Frazzini and Pedersen (2014)), NI (see Ikenberry, Lakonishok, and Vermaelen (1995), Loughran and Ritter (1995)), MAX (see Bali et al. (2011)), RVar (see Ang, Hodrick, Xing, and Zhang (2006), Fu (2009)), and AC (see Sloan (1996)). Our test assets are from independent two-way sorts of NYSE-, AMEX-, and NASDAQ-listed ordinary common stocks on size and each of the return anomaly variables from the above list.[2] In particular, the sets of test assets used in the empirical analysis include the value-weighted monthly excess returns on the 25 size-STR portfolios, the 25 size-CI portfolios, the 25 size-DR portfolios, the 25 size-β portfolios, the 35 size-NI portfolios, the 25 size-MAX portfolios, the 25 size-RVar portfolios, and the 25 size-AC portfolios. As a further assessment of our factor pricing models, we make use of the value-weighted monthly excess returns on the 30 IND (see Fama and French (1997)) portfolios and the 8 D10−1 portfolios associated with one-way sorts on STR, CI, DR, market β, NI, MAX, RVar, and AC.[3] The inclusion of these return anomaly portfolios is also in line with the advice of Lewellen et al. (2010), to improve the power of asset pricing tests by considering portfolios, in the test assets, that are sorted on market β, industry, or other firm-level characteristics.

The majority of our test assets are two-way independent sorted portfolios. The rationale is that they allow us to examine both the variation in anomaly returns and the explanatory power of the different factor pricing models across size groups. In fact, anomaly patterns are much stronger for small (microcap) stocks, and the workhorse asset pricing models in the literature have systematic problems in explaining their average excess returns (Fama (1998), Fama and French (2012), (2015), (2016), (2018)). Moreover, our rationale for using the value-weighted portfolio excess returns follows from Fama (1998), who argues that value-weighting more accurately captures the total wealth effects experienced by investors. We source data on STR, market β, NI, RVar, AC, and IND portfolio returns once again from Kenneth R. French's Web site. The data sources of the remaining anomaly portfolio returns are provided in the Supplementary Material.

C. Model Performance Measures

Our assessment of the performance of each return factor model in the time-series regressions is based on a battery of metrics commonly used in the empirical asset pricing literature (see Fama and French (1993), (2015), (2016), (2018), Hou et al. (2015), (2017a), and Stambaugh and Yuan (2016)). The first metric is the F-statistic of the Gibbons, Ross, and Shanken (1989) test, henceforth denoted as the GRS statistic. The null hypothesis states that the cross-section of alphas (i.e., all intercepts in time-series regressions of test assets on a return factor model) are jointly indistinguishable from 0. Although popular in the literature, the Gibbons et al. (1989) test is not without shortcomings. For example, the power of the test decreases with the number of assets. The second metric used to evaluate

[2]We thank the referee for suggesting some of the test asset portfolios used in the empirical analysis.

[3]To conserve space, we omit definitions of the anomaly variables and details of the portfolio construction in this article. These are available in Fama and French (2008), (2016) and in the Supplementary Material.

https://doi.org/10.1017/S0022109018000947 Published online by Cambridge University Press

competing factor models is the average absolute value of the alphas, denoted $A|\alpha_i|$. In the time-series factor regression, alpha is viewed as a measure of model mispricing or a test asset's deviation from the model. Our third metric, $A|\alpha_i|/A|\bar{r}_i|$, is the ratio of the average absolute value of the alphas to the average absolute value of $\bar{r}_i$. We compute $\bar{r}_i$ as the average excess return on an anomaly portfolio i minus the value-weighted average excess return on the market portfolio. The fourth metric that we employ is $A\alpha_i^2/A\bar{r}_i^2$, which represents the ratio of the average squared alpha to the average squared value of $\bar{r}_i$. Both $A|\alpha_i|/A|\bar{r}_i|$ and $A\alpha_i^2/A\bar{r}_i^2$ measure the dispersion of the alphas produced by a given asset pricing model relative to the dispersion of the average excess returns on test assets. Consequently, low values of these ratios indicate a better performance of the model. Our fifth metric for model performance evaluation is $As^2(\alpha_i)/A\alpha_i^2$, which is the ratio of the average variance estimate of the sampling errors of the estimated alphas to $A\alpha_i^2$. More precisely, $As^2(\alpha_i)/A\alpha_i^2$ shows the proportion of the dispersion (second moment) of the alpha estimates that is due to sampling error rather than to the dispersion of the true alphas. By construction, a higher value of the ratio suggests better model performance. Another metric that we estimate is $A(R^2)$, which denotes the average value of the time-series regression R^2 adjusted for degrees of freedom.

We next compute two additional metrics as in Barillas and Shanken (2017) and Fama and French (2018). These are the maximum squared Sharpe ratio, denoted by $\mathrm{SH}^2(\alpha)$, used for the alphas of test assets relative to a given factor model, and the maximum squared Sharpe ratio of a model's factor(s), denoted as $\mathrm{SH}^2(f)$. The economic mispricing metric, $\mathrm{SH}^2(\alpha)$, is also known as the Sharpe ratio improvement metric. In fact, Gibbons et al. (1989) show that

$$\mathrm{SH}^2(\alpha) \quad = \quad \mathrm{SH}^2(f, r^e) - \mathrm{SH}^2(f) \quad = \quad \alpha' \Sigma^{-1} \alpha, \tag{11}$$

where r^e is the set of test asset (excess) returns, f is the excess return factor(s) of a given model, α is the cross-section of alphas, Σ is the time-series regression residual (invertible) covariance matrix, and $\mathrm{SH}^2(f, r^e)$ is the maximum squared Sharpe ratio of the ex post tangency portfolio formed by combining the test assets with the factor(s).[4] A lower value of $\mathrm{SH}^2(\alpha)$ implies a lower squared Sharpe ratio improvement from exploiting mispricing (nonzero alphas) by a given factor model, that is, better model performance. Fama and French (2018) note that the sample estimates of $\mathrm{SH}^2(f)$ are biased upward. In particular, the bias is likely to be larger in smaller samples and for models with more traded factors. Hence, we also compute a bias-free version of the maximum squared Sharpe ratio of the factor(s), denoted $\mathrm{SH}_B^2(f)$, as the average of $\mathrm{SH}^2(f)$ from 100,000 full-sample bootstrap simulation runs. Harvey and Liu (2017), however, argue that a factor model may still perform poorly in explaining the cross-section of average returns, even if it generates a large point estimate of $\mathrm{SH}^2(f)$.

Consistent with many recent papers in the empirical asset pricing literature (see, among others, Fama and French (1996), (2015), (2016), Hou et al. (2015), (2017a), and Stambaugh and Yuan (2016); and references therein), we first

[4] Although a negative relation between $\mathrm{SH}^2(\alpha)$ and $\mathrm{SH}^2(f)$ is expected, the relation may not be perfect across different sets of test assets (Fama and French (2018)).

https://doi.org/10.1017/S0022109018000947 Published online by Cambridge University Press

informally compare the performance of different factor models, by simply eyeballing the point estimates of the above metrics. Barillas and Shanken (2017), however, show that identifying a superior model (or, equivalently, ranking models) based on such an informal, but common, procedure can be misleading. Relating to this, Harvey and Liu (2017) also caution that when two competing factor models both generate nonzero alphas, the use of the GRS statistic to heuristically determine the relative performance of the two models is inappropriate. Furthermore, in the excluded factor time-series regressions, the use of the Gibbons et al. (1989) test is not suited for a valid statistical inference on the comparison of nonnested models (Barillas and Shanken (2017), Barillas et al. (2017), and Fama and French (2018)). Taking these important issues into account, we then examine whether a competing factor model is significantly better than another candidate model. To do so, we adopt the pairwise test for equality of the cross-sectional regression R^2 introduced by Kan et al. (2013). The cross-sectional R^2 test, described in the following section, overcomes many of the issues of model comparisons that arise in the traditional alpha-based framework. In addition, this testing procedure accounts for the impact of potential model misspecification, and therefore allows for robust and valid statistical inference. However, the pairwise model comparison is open to the criticism that the process of searching over alternative models can lead to an overstatement of statistical significance (Barillas et al. (2017), Kan et al. (2013)). To address this issue, we also adopt the multiple model comparison test, which determines whether a given factor pricing model outperforms a set of alternative models in terms of the cross-sectional regression R^2.

D. Tests for Cross-Sectional R^2s Comparisons

1. Pairwise Tests of Equality of Cross-Sectional R^2s

In this section, we briefly discuss the Kan et al. (2013) test of model comparison based on the second-pass cross-sectional regression R^2s of two asset pricing models.[5] Let ρ_A^2 and ρ_B^2 be the cross-sectional R^2s of two competing models A and B, respectively. For nested models, the test H_0 is used to statistically verifying whether they have the same explanatory power: $\rho_A^2 = \rho_B^2$. Yet the H_0 test: $\rho_A^2 = \rho_B^2$ is fairly complicated for models that are nonnested. This is due to three possible asymptotic distributions for $\widehat{\rho}_A^2 - \widehat{\rho}_B^2$ under the null hypothesis. Firstly, perhaps the pricing factors that are not common to the two nonnested models are irrelevant for explaining the cross-sectional variation in average (excess) returns. In this case, the models will have the same pricing errors and the same overall goodness of fit in the population. Secondly, the two models may produce different pricing errors, yet still have identical cross-sectional R^2s. Finally, it is also possible that the two nonnested models are both correctly specified (i.e., $\rho_A^2 = \rho_B^2 = 1$), although their factors differ. Given the aforementioned three distinct possibilities, the test of H_0: $\rho_A^2 = \rho_B^2$ for nonnested models entails a complicated sequential

[5] The cross-sectional regression specification for a given factor model is

$$\mu_r^e \quad = \quad C\lambda,$$

where μ_r^e is the mean of r^e, $C = [1_N, \text{cov}[r^e, f']]$, $\lambda = (C'WC)^{-1}C'W\mu_r^e$, W is the weighting matrix, and N is the number of test assets. For more details, see Kan et al. (2013).

https://doi.org/10.1017/S0022109018000947 Published online by Cambridge University Press

procedure (described in the Supplementary Material). An alternative to the sequential testing procedure is to simply perform the normal test of H_0: $0 < \rho_A^2 = \rho_B^2 < 1$. This approach implicitly rules out the possibility that the additional factors in each model are completely irrelevant for explaining the variation in average returns. More importantly, it assumes that both models are misspecified, since asset pricing models are merely approximations of the true data generating process (Kan et al. (2013)).

In this article, we perform both the sequential and the normal tests when comparing nonnested models. For brevity, we only report the results based on the normal test. The results based on the sequential test are provided in Tables IA12 through IA21 of the Supplementary Material, which are qualitatively similar to those reported in this article. Moreover, we compute both the ordinary least squares (OLS) R^2 and generalized least squares (GLS) cross-sectional regression R^2s. The use of the OLS is more relevant for explaining the average (excess) returns of test assets, while the use of the GLS R^2 is more relevant from an investment perspective (Kandel and Stambaugh (1995), Kan et al. (2013), Lewellen et al. (2010)).

2. Multiple Model Comparison

We follow Kan et al. (2013) to conduct the multiple model comparison test. Let $\delta \equiv (\delta_2, \ldots, \delta_p)$, where $\delta_i \equiv \rho_1^2 - \rho_i^2$ for model $i = 2, \ldots, p$. We test the null hypothesis that the benchmark factor model 1 performs at least as well as the competing factor models 2 to p, that is, H_0: $\delta \geq 0_u$ with $u = p - 1$. The alternative hypothesis is that some model has a higher population cross-sectional R^2 than the benchmark factor model 1. The multiple model comparison test is based on the sample counterpart of δ, $\hat{\delta} \equiv (\hat{\delta}_2, \ldots, \hat{\delta}_p)$, where $\hat{\delta}_i \equiv \hat{\rho}_1^2 - \hat{\rho}_i^2$. We assume that $0 < \rho_i^2 < 1$ for all model i, such that $\hat{\delta}$ has an asymptotically normal distribution, with mean δ and covariance matrix $\Sigma_{\hat{\delta}}$. The test statistic is then constructed by solving the following quadratic programming problem:

$$\min_{\delta} (\hat{\delta} - \delta)' \hat{\Sigma}_{\hat{\delta}}^{-1} (\hat{\delta} - \delta) \quad \text{s.t.} \quad \delta \geq 0_u, \tag{12}$$

where $\hat{\Sigma}_{\hat{\delta}}$ is a consistent estimator of $\Sigma_{\hat{\delta}}$. The likelihood ratio test of the null hypothesis is given as

$$\text{LR} = T(\hat{\delta} - \tilde{\delta})' \hat{\Sigma}_{\hat{\delta}}^{-1} (\hat{\delta} - \tilde{\delta}), \tag{13}$$

where $\tilde{\delta}$ is the optimal solution to the problem given by equation (3). In order to conduct statistical inferences, we obtain asymptotically valid p-values following the numerically efficient procedure outlined in Kan et al. (2013).

For the multiple nonnested model comparison, we first remove the alternative factor pricing models i that are nested by the benchmark model. If any of the remaining alternatives are nested by another alternative factor model, we remove the "smaller" model. Finally, we also remove any alternative models that nest the benchmark factor model. However, to conduct the nested multiple model comparison, we cannot use the likelihood ratio test, since $\hat{\delta}$ is not asymptotically normally distributed. In this case, we form a single expanded model, including all the factors contained in the models that nest the benchmark model. Hence, the null

hypothesis that the benchmark model has the same OLS (GLS) cross-sectional regression R^2 as these alternative models can be tested using the approach for pairwise nested model comparisons in Section II.D.1.

III. Time-Series Results

Table 1 reports the factor model performance in the time-series (absolute) tests on the different sets of anomaly portfolios. To save space, we provide average monthly excess returns and alphas for portfolios, relative to each of the models, and the associated Newey and West (1987)-adjusted t-statistics in Tables IA2 through IA11 of the Supplementary Material.[6] Throughout this section, our discussion focuses on the best-performing asset pricing model.

TABLE 1

Performance of Factor Models in Time-Series (Absolute) Tests

Table 1 reports the summary statistics for time-series (absolute) tests of ten different factor models: the capital asset pricing model (CAPM) of Sharpe (1964) and Lintner (1965); the Fama and French (1993) 3-factor (FF3) model; the Fama and French (1993) and Carhart (1997) 4-factor (FFC) model; the Fama and French (1993) and Pástor and Stambaugh (2003) 4-factor (FFPS) model; the Asness and Frazzini (2013) 3-factor (FFAF) model, which combines their value factor with the market and size factors of the FF3 model; the Hou et al. (2015) q-factor (HXZ) model; the Fama and French (2015) 5-factor (FF5) model; the 4-factor (FF4) model that excludes the value factor from the FF5 model; the Stambaugh and Yuan (2016) 4-factor (SY4) model; and the Barillas and Shanken (2018) 6-factor (BS6) model, which includes the market, size, and momentum factors from the FF5 model, the profitability and investment factors from the HXZ model, and the value factor from the FFAF model. The test assets are the value-weighted monthly excess returns on anomaly portfolios: the 25 size-STR (short-term reversal) portfolios, the 25 size-CI (abnormal capital investment) portfolios, the 25 size-DR (distress risk) portfolios, the 25 size-β (market beta) portfolios, the 35 size-NI (net share issues) portfolios, the 25 size-MAX (lottery demand) portfolios, the 25 size-RVar (residual variance) portfolios, the 25 size-AC (accruals) portfolios, the 30 IND (industry) portfolios, and the 8 D10−1 (high minus low decile) portfolios. For each of the factor models, GRS denotes the F-statistic of the Gibbons et al. (1989) test on the null hypothesis that the alphas (i.e., intercepts in time-series regressions) across a given set of anomaly portfolios are jointly equal to 0, p(GRS) is the p-value associated with the GRS statistic (i.e., F-statistic), $A|\alpha_i|$ is the average absolute value of the alphas, $A|\alpha_i|/A|\overline{r}_i|$ is the ratio of the average absolute value of the alphas to the average absolute value of $\overline{r}_i$, estimated as the average excess return on an anomaly portfolio i minus the value-weighted average excess return on the market portfolio, $A\alpha_i^2/A\overline{r}_i^2$ is the ratio of the average squared alpha to the average squared value of $\overline{r}_i$, $As^2(\alpha_i)/A\alpha_i^2$ is the ratio defined as the average variance estimate of the sampling errors of the estimated alphas to $A\alpha_i^2$, $A(R^2)$ is the average value of the time-series regression R^2 adjusted for degrees of freedom, $SH^2(f)$ is the maximum squared Sharpe ratio of the factor(s), $SH_B^2(f)$ is the average $SH^2(f)$ from 100,000 full-sample bootstrap simulation runs, and $SH^2(\alpha)$ is the maximum squared Sharpe ratio for the alphas across a given set of anomaly portfolios. The sample period is from Jan. 1968 to Dec. 2016.

Model	GRS	p(GRS)	$A\lvert\alpha_i\rvert$	$\frac{A\lvert\alpha_i\rvert}{A\lvert\overline{r}_i\rvert}$	$\frac{A\alpha_i^2}{A\overline{r}_i^2}$	$\frac{As^2(\alpha_i)}{A\alpha_i^2}$	$A(R^2)$	$SH^2(f)$	$SH_B^2(f)$	$SH^2(\alpha)$
Panel A. 25 Size-STR Portfolios										
CAPM	4.147	0.000	0.202	0.842	0.708	0.241	0.752	0.012	0.012	0.732
FF3	4.373	0.000	0.155	0.647	0.558	0.164	0.866	0.043	0.032	0.709
FFC	4.403	0.000	0.191	0.793	0.750	0.119	0.875	0.086	0.054	0.803
FFPS	4.502	0.000	0.150	0.625	0.546	0.169	0.866	0.056	0.046	0.704
FFAF	4.444	0.000	0.144	0.599	0.532	0.162	0.872	0.027	0.026	0.746
HXZ	3.887	0.000	0.176	0.733	0.707	0.146	0.866	0.182	0.118	0.729
FF5	4.041	0.000	0.160	0.668	0.550	0.165	0.874	0.107	0.078	0.749
FF4	4.045	0.000	0.160	0.666	0.544	0.173	0.869	0.107	0.060	0.685
SY4	4.902	0.000	0.230	0.956	1.097	0.096	0.870	0.233	0.112	0.877
BS6	3.830	0.000	0.163	0.677	0.531	0.179	0.883	0.237	0.147	0.746

(continued on next page)

[6] Standard in the literature, we consider an alpha to be statistically significant if the corresponding t-statistic is at least 2.

https://doi.org/10.1017/S0022109018000947 Published online by Cambridge University Press

TABLE 1 (continued)
Performance of Factor Models in Time-Series (Absolute) Tests

Model	GRS	p(GRS)	$A\|\alpha_i\|$	$\frac{A\|\alpha_i\|}{A\|\bar{r}_i\|}$	$\frac{A\alpha_i^2}{A\bar{r}_i^2}$	$\frac{As^2(\alpha_i)}{A\alpha_i^2}$	$A(R^2)$	$SH^2(f)$	$SH^2_B(f)$	$SH^2(\alpha)$
Panel B. 25 Size-CI Portfolios										
CAPM	2.797	0.000	0.183	0.743	0.610	0.280	0.755	0.012	0.012	0.597
FF3	2.196	0.001	0.081	0.329	0.121	0.660	0.885	0.043	0.032	0.587
FFC	3.011	0.000	0.156	0.635	0.381	0.197	0.895	0.086	0.054	0.899
FFPS	2.233	0.001	0.079	0.323	0.123	0.658	0.886	0.056	0.046	0.575
FFAF	2.463	0.000	0.091	0.371	0.148	0.483	0.894	0.027	0.026	0.645
HXZ	2.339	0.000	0.145	0.591	0.382	0.239	0.885	0.182	0.118	0.633
FF5	1.997	0.003	0.082	0.335	0.145	0.547	0.893	0.107	0.078	0.559
FF4	1.981	0.003	0.081	0.330	0.142	0.585	0.888	0.107	0.060	0.556
SY4	1.987	0.003	0.119	0.485	0.247	0.394	0.881	0.233	0.112	0.631
BS6	2.355	0.000	0.097	0.397	0.215	0.361	0.904	0.237	0.147	0.785
Panel C. 25 Size-DR Portfolios										
CAPM	0.953	0.530	0.113	0.623	0.406	0.973	0.734	0.012	0.012	0.485
FF3	2.074	0.002	0.130	0.714	0.758	0.255	0.867	0.043	0.032	0.545
FFC	1.948	0.004	0.106	0.583	0.353	0.499	0.881	0.086	0.054	0.772
FFPS	2.032	0.002	0.129	0.710	0.728	0.269	0.868	0.056	0.046	0.539
FFAF	1.984	0.003	0.105	0.577	0.506	0.327	0.880	0.027	0.026	0.521
HXZ	1.518	0.052	0.082	0.452	0.257	0.912	0.854	0.182	0.118	0.484
FF5	1.807	0.010	0.108	0.591	0.493	0.393	0.875	0.107	0.078	0.493
FF4	1.573	0.039	0.115	0.632	0.562	0.398	0.855	0.107	0.060	0.493
SY4	1.900	0.006	0.093	0.513	0.298	0.755	0.869	0.233	0.112	0.595
BS6	2.716	0.000	0.123	0.673	0.606	0.303	0.889	0.237	0.147	0.627
Panel D. 25 Size-β Portfolios										
CAPM	2.387	0.000	0.236	1.187	1.248	0.208	0.738	0.012	0.012	0.533
FF3	1.788	0.011	0.129	0.646	0.425	0.272	0.870	0.043	0.032	0.662
FFC	1.341	0.125	0.082	0.411	0.186	0.632	0.873	0.086	0.054	0.667
FFPS	1.917	0.005	0.135	0.678	0.469	0.249	0.870	0.056	0.046	0.663
FFAF	2.149	0.001	0.185	0.929	0.758	0.161	0.857	0.027	0.026	0.598
HXZ	1.827	0.009	0.076	0.381	0.160	0.831	0.866	0.182	0.118	0.536
FF5	1.857	0.007	0.083	0.415	0.179	0.621	0.881	0.107	0.078	0.855
FF4	1.837	0.008	0.080	0.402	0.175	0.662	0.874	0.107	0.060	0.620
SY4	1.217	0.216	0.063	0.319	0.103	1.457	0.861	0.233	0.112	0.424
BS6	2.193	0.001	0.112	0.563	0.318	0.388	0.882	0.237	0.147	0.650
Panel E. 35 Size-NI Portfolios										
CAPM	4.639	0.000	0.244	1.003	1.047	0.196	0.727	0.012	0.012	1.300
FF3	4.115	0.000	0.148	0.606	0.557	0.175	0.863	0.043	0.032	1.406
FFC	3.653	0.000	0.134	0.551	0.451	0.224	0.864	0.086	0.054	1.381
FFPS	4.057	0.000	0.148	0.608	0.569	0.173	0.864	0.056	0.046	1.397
FFAF	4.477	0.000	0.183	0.752	0.757	0.132	0.856	0.027	0.026	1.369
HXZ	3.133	0.000	0.104	0.428	0.223	0.503	0.861	0.182	0.118	1.268
FF5	3.324	0.000	0.110	0.450	0.241	0.386	0.877	0.107	0.078	1.466
FF4	3.319	0.000	0.110	0.453	0.243	0.395	0.874	0.107	0.060	1.374
SY4	2.780	0.000	0.109	0.446	0.251	0.496	0.858	0.233	0.112	1.180
BS6	3.404	0.000	0.124	0.511	0.314	0.341	0.874	0.237	0.147	1.297
Panel F. 25 Size-MAX Portfolios										
CAPM	7.867	0.000	0.318	1.142	1.296	0.094	0.754	0.012	0.012	0.947
FF3	7.165	0.000	0.221	0.792	0.741	0.083	0.862	0.043	0.032	0.965
FFC	6.725	0.000	0.208	0.748	0.614	0.100	0.866	0.086	0.054	1.054
FFPS	7.157	0.000	0.220	0.788	0.737	0.085	0.863	0.056	0.046	0.955
FFAF	7.673	0.000	0.269	0.967	1.020	0.060	0.857	0.027	0.026	1.003
HXZ	6.097	0.000	0.152	0.547	0.324	0.214	0.861	0.182	0.118	0.867
FF5	6.377	0.000	0.147	0.526	0.331	0.166	0.885	0.107	0.078	0.992
FF4	6.356	0.000	0.145	0.521	0.325	0.178	0.878	0.107	0.060	0.895
SY4	5.445	0.000	0.141	0.507	0.275	0.265	0.863	0.233	0.112	0.888
BS6	6.074	0.000	0.150	0.537	0.261	0.235	0.884	0.237	0.147	0.982

(continued on next page)

https://doi.org/10.1017/S0022109018000947 Published online by Cambridge University Press

https://doi.org/10.1017/S0022109018000947 Published online by Cambridge University Press

TABLE 1 (continued)
Performance of Factor Models in Time-Series (Absolute) Tests

Model	GRS	p(GRS)	$A\|\alpha_i\|$	$\frac{A\|\alpha_i\|}{A\|\bar{r}_i\|}$	$\frac{A\alpha_i^2}{A\bar{r}_i^2}$	$\frac{As^2(\alpha_i)}{A\alpha_i^2}$	$A(R^2)$	$SH^2(f)$	$SH^2_B(f)$	$SH^2(\alpha)$
Panel G. 25 Size-RVar Portfolios										
CAPM	6.835	0.000	0.361	1.157	1.384	0.078	0.746	0.012	0.012	1.149
FF3	6.738	0.000	0.240	0.770	0.989	0.055	0.865	0.043	0.032	1.216
FFC	6.019	0.000	0.221	0.710	0.660	0.079	0.870	0.086	0.054	1.207
FFPS	6.678	0.000	0.243	0.779	0.996	0.055	0.865	0.056	0.046	1.203
FFAF	7.514	0.000	0.305	0.978	1.306	0.041	0.855	0.027	0.026	1.222
HXZ	5.623	0.000	0.123	0.395	0.213	0.273	0.863	0.182	0.118	1.038
FF5	6.049	0.000	0.139	0.445	0.405	0.113	0.893	0.107	0.078	1.194
FF4	6.058	0.000	0.137	0.440	0.406	0.122	0.881	0.107	0.060	1.123
SY4	5.813	0.000	0.116	0.373	0.232	0.271	0.866	0.233	0.112	1.072
BS6	6.427	0.000	0.143	0.458	0.228	0.214	0.892	0.237	0.147	1.023
Panel H. 25 Size-AC Portfolios										
CAPM	3.865	0.000	0.147	0.821	0.657	0.485	0.759	0.012	0.012	0.725
FF3	3.935	0.000	0.113	0.634	0.458	0.263	0.902	0.043	0.032	0.877
FFC	3.485	0.000	0.110	0.614	0.394	0.316	0.902	0.086	0.054	0.882
FFPS	3.781	0.000	0.109	0.613	0.434	0.279	0.902	0.056	0.046	0.869
FFAF	3.870	0.000	0.116	0.650	0.477	0.249	0.901	0.027	0.026	0.833
HXZ	3.693	0.000	0.134	0.749	0.570	0.238	0.904	0.182	0.118	0.805
FF5	4.186	0.000	0.117	0.656	0.468	0.251	0.910	0.107	0.078	0.910
FF4	4.032	0.000	0.118	0.664	0.475	0.249	0.909	0.107	0.060	0.846
SY4	2.818	0.000	0.095	0.530	0.328	0.495	0.889	0.233	0.112	0.812
BS6	4.750	0.000	0.137	0.768	0.608	0.218	0.909	0.237	0.147	0.964
Panel I. 30 IND Portfolios										
CAPM	1.311	0.127	0.150	1.173	1.630	0.669	0.578	0.012	0.012	0.650
FF3	2.284	0.000	0.206	1.604	2.527	0.409	0.614	0.043	0.032	0.672
FFC	2.395	0.000	0.174	1.356	1.928	0.551	0.619	0.086	0.054	0.808
FFPS	2.339	0.000	0.217	1.692	2.870	0.361	0.618	0.056	0.046	0.667
FFAF	1.774	0.007	0.183	1.425	2.260	0.449	0.614	0.027	0.026	0.692
HXZ	2.395	0.000	0.207	1.618	2.387	0.489	0.618	0.182	0.118	0.675
FF5	2.644	0.000	0.243	1.899	3.033	0.340	0.643	0.107	0.078	0.622
FF4	2.282	0.000	0.249	1.944	3.170	0.332	0.633	0.107	0.060	0.623
SY4	1.971	0.002	0.180	1.401	1.774	0.691	0.615	0.233	0.112	0.684
BS6	4.530	0.000	0.298	2.327	4.545	0.254	0.641	0.237	0.147	0.735
Panel J. 8 D10 −1 Portfolios										
CAPM	9.964	0.000	0.547	0.647	0.557	0.079	0.192	0.012	0.012	0.052
FF3	12.860	0.000	0.577	0.683	0.545	0.057	0.361	0.043	0.032	0.057
FFC	10.194	0.000	0.464	0.549	0.332	0.090	0.403	0.086	0.054	0.044
FFPS	12.805	0.000	0.580	0.686	0.557	0.056	0.363	0.056	0.046	0.058
FFAF	14.790	0.000	0.623	0.737	0.668	0.046	0.368	0.027	0.026	0.075
HXZ	5.940	0.000	0.262	0.310	0.109	0.308	0.381	0.182	0.118	0.022
FF5	8.317	0.000	0.362	0.429	0.189	0.154	0.411	0.107	0.078	0.019
FF4	8.183	0.000	0.365	0.432	0.192	0.161	0.386	0.107	0.060	0.020
SY4	4.347	0.000	0.259	0.307	0.120	0.280	0.414	0.233	0.112	0.033
BS6	7.153	0.000	0.270	0.320	0.133	0.240	0.431	0.237	0.147	0.045

A. Size-STR (Short-Term Reversal) Portfolios

Starting with the 25 size-STR portfolios in Panel A of Table 1, we find that all the models are comfortably rejected by the Gibbons et al. (1989) test at conventional significance levels. The p-values corresponding to the GRS statistics round to 0 to at least three decimal places. However, the BS6 model outperforms, though not statistically, all other factor models, as it generates the smallest point estimate of the GRS statistic. The next best asset pricing model turns out to be the HXZ model, followed by the FF5 model, in terms of their GRS statistic. The point estimate of the GRS statistic generated by the FF4 model is marginally higher than those of the HXZ and FF5 models. The SY4 model appears to be the

worst-performing model, followed by the FFPS model, in terms of the magnitude of the GRS statistic. The average absolute value of 0.144% per month for alphas in the FFAF model is the smallest among all factor pricing models. The value of 0.599 for the ratio $A|\alpha_i|/A|\bar{r}_i|$ produced by the FFAF model is also the smallest among all the competing models. A point estimate of 0.599 for $A|\alpha_i|/A|\bar{r}_i|$ implies that the dispersion of alphas is approximately 60% as large as the dispersion of the average excess returns on our anomaly portfolios. The BS6 model appears to be the top performer when looking at the point estimate of 0.531 for the ratio $A\alpha_i^2/A\bar{r}_i^2$, which is the smallest among all the competing factor models. In terms of the magnitude of $A(R^2)$, the BS6 model once again turns out to be the best-performing asset pricing model, followed by the FFC model, while the CAPM is the worst performer, followed jointly by the FF3, FFPS, and HXZ models.

We notice that almost all of our models outperform the CAPM by generating lower point estimates for the $A|\alpha_i|$ and $A|\alpha_i|/A|\bar{r}_i|$ metrics. Yet the point estimate of the ratio $As^2(\alpha_i)/A\alpha_i^2$ for the CAPM tells a different story. For example, approximately 24% of the dispersion (second moment) of the alpha estimates for the CAPM is due to sampling error, compared to approximately 10% for the SY4 model. Notably, approximately 18% of the second moment of the alpha estimates for the BS6 model is due to sampling error, and the remaining 82% is due to dispersion in the true alphas. However, the six factors of the model generate the largest point estimates of 0.237 and 0.147, respectively, for $\mathrm{SH}^2(f)$ and $\mathrm{SH}_B^2(f)$. The SY4 model factors also perform well, generating a large $\mathrm{SH}^2(f)$ of 0.233. In terms of $\mathrm{SH}^2(\alpha)$, the FF4 model produces the smallest point estimate of 0.685. As with the $\mathrm{SH}^2(f)$ metric, after adjusting for an upward bias, the point estimates of $\mathrm{SH}_B^2(f)$ drop by as much as 52% (see, e.g., the SY4 model).

In Table IA2 of the Supplementary Material, all models generate a statistically significant alpha for the microcap extreme STR portfolio. This result shows that microcap extreme STR stocks are a huge problem for our asset pricing models. In the smallest size quintile, all but the FFC, HXZ, SY4, and BS6 models generate a statistically significant alpha for only one portfolio, which is the lowest number across all models. We also notice that the BS6 model outperforms other models in explaining average excess returns, by generating just five significant alphas. In contrast, both the FFC and the SY4 models generate 11 significant alphas for portfolios; the highest across all the factor models. To sum up, the point estimates of more than half of our performance metrics favor the BS6 model as the best asset pricing model.

B. Size-CI (Abnormal Capital Investment) Portfolios

We now focus on the 25 size-CI portfolios in Panel B of Table 1. The Gibbons et al. (1989) test easily rejects all the asset pricing models. Yet the rejection is the weakest for the FF4 model, which generates a GRS statistic of 1.981 (p-value = 0.003). In contrast, the FFC model produces the largest GRS statistic of 3.011 (p-value = 0.000). We find that the average absolute alpha produced by the FFPS model is only 0.079% per month, the smallest among all the competing factor models. The CAPM generates the largest point estimate for $A|\alpha_i|$, which is 0.183% per month. When looking at the two dispersion measures of alphas, $A|\alpha_i|/A|\bar{r}_i|$ and $A\alpha_i^2/A\bar{r}_i^2$, we find that the FFPS and FF3 models generate the

https://doi.org/10.1017/S0022109018000947 Published online by Cambridge University Press

https://doi.org/10.1017/S0022109018000947 Published online by Cambridge University Press

smallest values, of 0.323 and 0.121, respectively. The FF3 model also produces the largest point estimate of the ratio $As^2(\alpha_i)/A\alpha_i^2$, given as 0.660. This indicates that two-thirds of the second moment of the alpha estimates for the model is due to sampling error and only one-third is due to dispersion in the true alphas. We further notice that the BS6 model produces the largest point estimate for $A(R^2)$, at approximately 90%. However, in terms of the magnitude of $\mathrm{SH}^2(\alpha)$, the FF4 model is the best performer, generating the smallest value for the metric among all the asset pricing models.

When examining the alphas in Table IA3 of the Supplementary Material, we see that the BS6 model generates a statistically significant alpha for just two portfolios. Out of 25 portfolios, three demonstrate significant alphas in the FFAF, while four portfolios in the FF3, FFPS, HXZ, FF5, FF4, and SY4 models. In contrast, the FFC model generates 14 significant alphas, which is the highest number among all the factor pricing models. Moreover, all but the CAPM and the FFC model have a statistically significant alpha for one out of five portfolios in the smallest size quintile. In summary, our results suggest that, despite rejection based on the Gibbons et al. (1989) test, the FF3, FFPS, and FF4 models are the three best factor pricing models, accommodating the average excess returns on size and CI sorted portfolios.

C. Size-DR (Distress Risk) Portfolios

The results for the 25 size-DR portfolios in Panel C of Table 1 indicate that all but the CAPM and the HXZ model are rejected by the Gibbons et al. (1989) test at the 5% level of significance. The CAPM generates the smallest GRS statistic of 0.953 (p-value = 0.530). The model also generates the largest value, 0.973, for the ratio $As^2(\alpha_i)/A\alpha_i^2$. Yet the HXZ model shows the best performance by producing the smallest value of 0.082% per month for $A|\alpha_i|$, the smallest values of 0.452 and 0.257, respectively, for the alpha dispersion metrics $A|\alpha_i|/A|\bar{r}_i|$ and $A\alpha_i^2/A\bar{r}_i^2$, and the smallest value of 0.484 for $\mathrm{SH}^2(\alpha)$. The model also generates the second smallest GRS statistic of 1.518. The BS6 model, which is the model that best explains size and STR sorted portfolios, now shows a superior performance only in terms of producing the largest point estimate for $A(R^2)$. Both the FF5 and SY4 models demonstrate a reasonably strong performance when judged by their $A|\alpha_i|$, $A|\alpha_i|/A|\bar{r}_i|$, $A\alpha_i^2/A\bar{r}_i^2$, and $A(R^2)$ estimates. Coincidentally, the point estimates of $\mathrm{SH}^2(\alpha)$ produced by the FF5 and FF4 models are the same, at 0.493.

Our results in Table IA4 of the Supplementary Material show that only the CAPM and the HXZ model generate statistically insignificant alphas across the 25 size-DR portfolios. Most of the remaining compared factor pricing models have problems explaining average returns on the highest distress quintile portfolios. The exceptions are the FFC and SY4 models. We also see that the FFPS model performs the worst in terms of generating the least number of significant alphas for portfolios. In particular, the model produces eight alphas that are statistically distinguishable from 0. Taken together, all of the above results suggest that the HXZ model provides the best description of average size-DR portfolio excess returns.

D. Size-β (Market Beta) Portfolios

We observe in Table 1 that the rejections of our factor pricing models by the Gibbons et al. (1989) test are weakest for the 25 size-β portfolios. In fact, the null hypothesis of zero alphas cannot be rejected for the FFC and SY4 models at conventional levels of significance (Panel D). The SY4 model generates a smaller GRS statistic compared to the FFC model. The average absolute value of the alphas is 0.063% per month for the SY4 model, which is smaller than those of other asset pricing models. The SY4 model also produces the smallest values of 0.319, 0.103, and 0.424 for pricing performance metrics $A|\alpha_i|/A|\bar{r}_i|$, $A\alpha_i^2/A\bar{r}_i^2$, and $\mathrm{SH}^2(\alpha)$, respectively. When judged by the point estimates of the GRS statistic, $A|\alpha_i|$, $A|\alpha_i|/A|\bar{r}_i|$, $A\alpha_i^2/A\bar{r}_i^2$, and $A(R^2)$, the CAPM turns out to be the worst performer among all the factor models.

Looking at the alphas reported in Table IA5 of the Supplementary Material, we find that the SY4 model outperforms all other factor pricing models. All alphas for the 25 size-β portfolios relative to the model are statistically indistinguishable from 0. Although the FFC model comfortably passes the Gibbons et al. (1989) test, the model generates three portfolio alphas that are statistically significant. Notably, 13 out of 25 portfolios have a significant alpha relative to the FFAF model. In contrast, the HXZ, FF5, and FF4 models generate just two significant alphas each. To summarize, our results indicate that the SY4 model best captures the average excess returns on the 25 size-β portfolios.

E. Size-NI (Net Share Issues) Portfolios

The Gibbons et al. (1989) test on the 35 size-NI portfolios, shown in Panel E of Table 1, rejects all the asset pricing models at conventional levels of significance. This implies that all our models are incomplete descriptions of average excess returns on these anomaly portfolios. However, the SY4 model produces the smallest GRS statistic of 2.780 (p-value = 0.000), followed by 3.133 (p-value = 0.000) for the HXZ model. Despite rejections, both these models appear to perform well based on the majority of the metrics. The average absolute alphas from the two models are 0.109% and 0.104% per month. The FF5 and FF4 models also produce small values of $A|\alpha_i|$, both at 0.110% per month. The dispersion of the alpha estimates for the HXZ model is approximately 43% as large as the dispersion of average excess returns on size-NI portfolios, whereas it is approximately 45% for the FF5, FF4, and SY4 models. The HXZ model has an edge over the SY4 model when judged by its $A\alpha_i^2/A\bar{r}_i^2$ ratio of 0.223. The corresponding point estimate is 0.251 for the SY4 model. Both the HXZ and the SY4 models show a superior performance to other factor models, by producing larger values, of 0.503 and 0.496 respectively, for the ratio $As^2(\alpha_i)/A\alpha_i^2$. However, the $A(R^2)$ values for the FF5 and FF4 models are marginally higher than those for the HXZ and SY4 models. When it comes to $\mathrm{SH}^2(\alpha)$, the SY4 model outperforms all the competing asset pricing models by generating the smallest value of 1.180. The HXZ model lines up as the next best performer, with a value of 1.268.

The results reported in Table IA6 of the Supplementary Material show that the SY4 model outperforms other candidate factor models by producing just three significant alphas. Moreover, only one out of these three significant alphas is located in the smallest size quintile. The HXZ model is the next best performer, as it

https://doi.org/10.1017/S0022109018000947 Published online by Cambridge University Press

https://doi.org/10.1017/S0022109018000947 Published online by Cambridge University Press

generates five alphas that are statistically distinguishable from 0. Conversely, in both the CAPM and the FFAF model, 15 out of 35 anomaly portfolios have statistically significant alphas. These results show that the SY4 and HXZ models are the top two performers in tests on the 35 size-NI portfolios, though the former model has a marginal edge over the latter when judged by the number of significant alphas.

F. Size-MAX (Lottery Demand) Portfolios

In Panel F of Table 1, we see that the Gibbons et al. (1989) test on the 25 size-MAX portfolios strongly rejects all of our competing models. Although rejected by the test, the SY4 model stands out as the best, based on the majority of the performance metrics. For example, the average absolute alpha relative to the model is 0.141% per month, which is smaller than that for any other factor pricing model. The SY4 model also produces the smallest value, of 0.507, and the second smallest value, of 0.275, respectively, for the ratios $A|\alpha_i|/A|\bar{r}_i|$ and $A\alpha_i^2/A\bar{r}_i^2$. Approximately 27% of the dispersion (second moment) of the alpha estimates for the model can be attributed to sampling error. Judged by the same $As^2(\alpha_i)/A\alpha_i^2$ ratio, the BS6 model turns out to be the next best model, with a value of 0.235. For this, the FFAF model emerges as the worst-performing asset pricing model, with a value of 0.060. We see that the FF5 and BS6 models deliver high and nearly identical values of $A(R^2)$. When examining the value of $\mathrm{SH}^2(\alpha)$, it can be observed that the HXZ model generates the smallest value of 0.867. The corresponding value for the SY4 model is slightly higher, at 0.888.

Examining the portfolio alphas, reported in Table IA7 of the Supplementary Material, we find that all of our models have difficulty explaining average excess returns on microcap portfolios. The HXZ model produces a statistically significant alpha for four portfolios, two of which are located in the microcap quintile. Out of 25 portfolios, six have significant alphas in the SY4 model, while eight in the FF5, FF4, and BS6 models. The FFAF model is ranked at the bottom, producing a statistically significant alpha for 15 portfolios. Overall, these results suggest that the SY4 model has a superior ability to accommodate, though not completely, the average excess returns on the 25 size-MAX portfolios.

G. Size-RVar (Residual Variance) Portfolios

The results in Panel G of Table 1, for the 25 size-RVar portfolios, show that the Gibbons et al. (1989) test strongly rejects all the asset pricing models. Notably, we find that the GRS statistics are generally large for the models, an observation similar to that reported in Fama and French (2016). The SY4 model generates the smallest average absolute alpha, of 0.116% per month, and the smallest point estimate, of 0.373, for the ratio $A|\alpha_i|/A|\bar{r}_i|$. Conversely, the HXZ model produces the smallest value, of 0.213, for the ratio $A\alpha_i^2/A\bar{r}_i^2$ and the largest value, of 0.273, for the ratio $As^2(\alpha_i)/A\alpha_i^2$. We also notice that the FF5 model ranks first among the factor pricing models when judged by the point estimate of $A(R^2)$.

The anomaly portfolio alphas, reported in Table IA8 of the Supplementary Material, show that the HXZ model produces a statistically significant alpha for four portfolios, with just two of which located in the smallest size quintile. The SY4 model also produces four significant alphas, yet three are located in the

smallest size quintile. We observe an improvement for the FF4 model over its performance in the 25 size-MAX portfolios. In particular, the model now produces four significant alphas and with just one located in the microcap quintile. The FF5 and BS6 models also show similar improvements in performance based on this criterion. The CAPM, however, ranks last among our factor pricing models. Out of 25 portfolios, 17 have significant alphas relative to the model. It is worth noting that across the ten sets of anomaly portfolios, this is the highest number of statistically significant alphas generated by any model. In summary, both the HXZ and the SY4 models are the best performing, but with imperfections, in explaining the average excess returns on the 25 size-RVar portfolios.

H. Size-AC (Accruals) Portfolios

We can see that from Panel H of Table 1 the Gibbons et al. (1989) test easily rejects all the models considered for the 25 size-AC portfolios. Although the competing models are incomplete descriptions of average excess returns on these portfolios, the SY4 model delivers the best performance on more than half of the metrics. For example, it produces the smallest GRS statistic, at 2.818, the smallest average absolute alpha, at 0.095% per month, and the smallest point estimate, at 0.530, for the alpha dispersion metric $A|\alpha_i|/A|\overline{r}_i|$. Among the other factor models, there is a noteworthy improvement in performance for the CAPM when we consider $As^2(\alpha_i)/A\alpha_i^2$, which is given as 0.485. The corresponding point estimate of the metric is 0.495 for the SY4 model, which is the largest value across all the asset pricing models tested on AC anomaly portfolios. We also find that the CAPM delivers the smallest value, at 0.725, for $\text{SH}^2(\alpha)$.

Out of 25 accruals anomaly portfolios reported in Table IA9 of the Supplementary Material, only three have statistically significant alphas relative to the SY4 model. The CAPM produces four significant alphas, none of which are located in the smallest size quintile. The performance of the FF3, FF5, and FF4 models do not bode well, since each of these models produce ten alphas that are statistically distinguishable from 0. This is in line with the evidence provided by Fama and French (2016). We also notice that the majority of our asset pricing models have difficulties in explaining the average excess returns of the biggest size stocks and the stocks in the highest AC quintile.

I. IND (Industry) Portfolios

Turning now to the 30 IND portfolios, the results for which are reported in Panel I of Table 1, we find that the Gibbons et al. (1989) test rejects almost all of the factor models at conventional significance levels. The sole exception is the CAPM, which produces a GRS statistic of 1.311 (p-value = 0.127). Moreover, informally comparing the competing models based on the point estimates of our performance metrics reveals that the CAPM is the top contender in the horse race. Overall, the SY4 model takes second place, while the BS6 model ranks last among our candidate factor models. In particular, the average absolute alpha is 0.150% per month for the CAPM, which is lower than the value of 0.180% per month for the SY4 model. Likewise, the point estimates of the ratios $A|\alpha_i|/A|\overline{r}_i|$ and $A\alpha_i^2/A\overline{r}_i^2$ are, respectively, 1.173 and 1.630 for the CAPM, which are much smaller than those of 1.401 and 1.774 for the SY4 model. However, on the

https://doi.org/10.1017/S0022109018000947 Published online by Cambridge University Press

basis of $A(R^2)$ and the ratio $As^2(\alpha_i)/A\alpha_i^2$, the SY4 model shows a marginal edge over the CAPM. Notably, the CAPM performs marginally better than the SY4 model when judged on the point estimate of $SH^2(\alpha)$. On the same criterion, the FF5 model performs the best, as it generates the smallest point estimate of 0.622. The model also produces the largest value for $A(R^2)$, which is consistent with Stambaugh and Yuan (2016), who informally compare factor models, such as the CAPM, FF3, HXZ, FF5, and SY4 models, using the 30 IND portfolios.

Table IA10 in the Supplementary Material reports that the CAPM and the SY4 model are the top performers. Both these models produce alphas for four portfolios that are statistically distinguishable from 0. In contrast, the BS6 model produces as many as 11 significant alphas. The number of significant alphas is ten for both the FF5 and the FF4 models, and seven for the FFC, FFAF, and HXZ models. The superior performance of the CAPM in producing the least number of significant alphas for industry portfolios is also consistent with the evidence by Fama and French (1997), who compare the model with the FF3 model.

J. D10−1 (High Minus Low Decile) Portfolios

We next consider the set of D10−1 portfolios corresponding to eight anomalies (i.e., STR, CI, DR, market β, NI, MAX, RVar, and AC) and examine the abilities of our competing asset pricing models to explain the average excess returns on these long-short portfolios. Panel J in Table 1 reports that all models are strongly rejected by the Gibbons et al. (1989) test. An informal analysis of the performance metrics suggests that the SY4 model is the best, though still incomplete, description of average excess returns on these anomaly portfolios. The average absolute alpha for portfolios is 0.259% per month in the model, which is the smallest value for the metric across all models. The $A|\alpha_i|/A|\bar{r}_i|$ ratio of 0.307 for the SY4 model is small compared with those of competing factor models. The HXZ model also does a good job in explaining the set of anomaly long-short spreads. The model generates a value of 0.262% per month for $A|\alpha_i|$, which is marginally higher than that for the SY4 model. About 31% of the dispersion (second moment) of the alpha estimates for the model is due to sampling error. All other asset pricing models fare poorly on this dispersion metric. We also find that the BS6 model generates the largest point estimate for $A(R^2)$, while the FF5 model produces the smallest point estimate for $SH^2(\alpha)$. Overall, the FFAF model performs the worst in the horse race among all models.

The results in Table IA11 of the Supplementary Material suggest that both the HXZ and the SY4 models are superior to other competing models, as each of these models generate just two significant alphas. In particular, the HXZ model generates a statistically significant alpha for the D10−1 portfolios corresponding to the NI and AC anomalies, while the SY4 model generates a significant alpha for the D10−1 portfolios sorted on STR and AC. The BS6 model also performs well in explaining average returns on anomaly portfolios. The model generates significant alphas for the high-minus-low portfolios that are sorted on DR, NI, and AC. Consistent with the findings in Table 1 (Panel J), the FFAF model ranks last. Out of eight anomaly long-short portfolios, seven demonstrate significant alphas relative to the model. Finally, none of our asset pricing models accommodate average return on the D10−1 portfolio corresponding to the AC anomaly.

https://doi.org/10.1017/S0022109018000947 Published online by Cambridge University Press

K. Overall Model Performance in Time-Series Regressions

To facilitate discussion in this section, Table 2 summarizes the overall performance of each of the return factor models in the time-series regressions presented in Table 1 and Tables IA2 through IA11 of the Supplementary Material. Columns 1 to 7 report the number of the sets of anomaly portfolios across which a given asset pricing model performs the best. More precisely, the criteria for overall performance evaluation include the number of times a competing model generates the smallest point estimates of the GRS statistic, $A|\alpha_i|$, $A|\alpha_i|/A|\bar{r}_i|$, $A\alpha_i^2/A\bar{r}_i^2$, and $SH^2(\alpha)$, and the largest point estimates of the $As^2(\alpha_i)/A\alpha_i^2$ and $A(R^2)$ metrics. In columns 8 to 17, we provide the number of statistically significant alphas relative to a given factor model tested on each set of the anomaly portfolios. The last column of Table 2 reports, for each asset pricing model, the number of significant alphas out of 42 small size portfolios from the eight anomaly variables, STR, CI, DR, market β, NI, MAX, RVar, and AC.

TABLE 2

Overall Performance of Factor Models in Time-Series (Absolute) Tests

Table 2 reports the overall performance in time-series (absolute) tests of ten different factor models: the capital asset pricing model (CAPM) of Sharpe (1964) and Lintner (1965); the Fama and French (1993) 3-factor (FF3) model; the Fama and French (1993) and Carhart (1997) 4-factor (FFC) model; the Fama and French (1993) and Pástor and Stambaugh (2003) 4-factor (FFPS) model; the Asness and Frazzini (2013) 3-factor (FFAF) model, which combines their value factor with the market and size factors of the FF3 model; the Hou et al. (2015) q-factor (HXZ) model; the Fama and French (2015) 5-factor (FF5) model; the 4-factor (FF4) model that excludes the value factor from the FF5 model; the Stambaugh and Yuan (2016) 4-factor (SY4) model; and the Barillas and Shanken (2018) 6-factor (BS6) model, which includes the market, size, and momentum factors from the FF5 model, the profitability and investment factors from the HXZ model, and the value factor from the FFAF model. NGRS, $N_A|\alpha_i|$, $N_A|\alpha_i|/A|\bar{r}_i|$, $N_A\alpha_i^2/A\bar{r}_i^2$, $N_As^2(\alpha_i)/A\alpha_i^2$, $N_A(R^2)$, and $NSH^2(\alpha)$ are the number of the sets of anomaly portfolios across which a given factor model produces; the smallest F-statistic of the Gibbons et al. (1989) test, the smallest value of $A|\alpha_i|$, the smallest value of $A|\alpha_i|/A|\bar{r}_i|$, the smallest value of $A\alpha_i^2/A\bar{r}_i^2$, the largest value of $As^2(\alpha_i)/A\alpha_i^2$, the largest value of $A(R^2)$, and the smallest value of $SH^2(\alpha)$, respectively. NSTR, NCI, NDR, Nβ, NNI, NMAX, NRVar, NAC, NIND, and ND10−1 are the number of statistically significant alphas (i.e., with a Newey and West (1987)-adjusted t-statistic $\geq$ 2.00), across the 25 size-STR (short-term reversal) portfolios, the 25 size-CI (abnormal capital investment) portfolios, the 25 size-DR (distress risk) portfolios, the 25 size-β (market beta) portfolios, the 35 size-NI (net share issues) portfolios, the 25 size-MAX (lottery demand) portfolios, the 25 size-RVar (residual variance) portfolios, the 25 size-AC (accruals) portfolios, the 30 IND (industry) portfolios, and the 8 D10−1 (high minus low decile) portfolios relative to a given factor model, respectively. NMicro is the number of statistically significant alphas (i.e., with a Newey and West (1987)-adjusted t-statistic $\geq$ 2.00) on the 42 smallest size (i.e., microcap) portfolios relative to a given factor model. The sample period is from Jan. 1968 to Dec. 2016. See also notes to Table 1.

Model	NGRS	$N_A\|\alpha_i\|$	$N\frac{A\|\alpha_i\|}{A\|\bar{r}_i\|}$	$N\frac{A\alpha_i^2}{A\bar{r}_i^2}$	$N\frac{As^2(\alpha_i)}{A\alpha_i^2}$	$N_A(R^2)$	$NSH^2(\alpha)$	NSTR	NCI	NDR	Nβ	NNI	NMAX	NRVar	NAC	NIND	ND10−1	NMicro
CAPM	2	1	1	1	2	0	1	10	9	0	12	15	14	17	4	4	5	15
FF3	0	0	0	1	1	0	0	6	4	7	7	11	12	13	10	9	6	16
FFC	0	0	0	0	0	0	0	11	14	4	3	13	14	14	8	7	5	19
FFPS	0	1	1	0	0	0	0	6	4	8	10	11	11	14	7	9	6	16
FFAF	0	1	1	0	0	0	0	7	3	6	13	15	15	14	9	7	7	19
HXZ	1	2	2	4	3	0	2	8	4	0	2	5	4	4	7	7	2	12
FF5	0	0	0	0	0	5	2	9	4	5	2	7	8	5	10	10	6	12
FF4	1	0	0	0	0	0	2	8	4	3	2	7	8	4	10	10	6	12
SY4	5	5	5	2	4	0	2	11	4	3	0	3	6	4	3	4	2	11
BS6	1	0	0	2	0	5	1	5	2	7	4	7	8	5	5	11	3	10

Starting with the Gibbons et al. (1989) test, we see that the SY4 model outperforms all the competing models by producing the smallest GRS statistic in five out of the ten sets of portfolios. The CAPM ranks second in our list of models, as it generates the smallest GRS statistic in two sets of portfolios. The results also show that the HXZ, FF4, and BS6 models each produce the smallest GRS statistic in one set of portfolios. For the metric $A|\alpha_i|$, which measures the unexplained average absolute excess return, the SY4 model generates the smallest point estimate in five out of the ten sets of portfolios, while the HXZ model

https://doi.org/10.1017/S0022109018000947 Published online by Cambridge University Press

https://doi.org/10.1017/S0022109018000947 Published online by Cambridge University Press

delivers the smallest value in two sets of portfolios. The SY4 model continues to outperform its nine competing models by generating the smallest value for the ratio $A|\alpha_i|/A|\bar{r}_i|$, and the largest value for the ratio $As^2(\alpha_i)/A\alpha_i^2$, in five and four sets of test asset portfolios, respectively. In contrast, the HXZ model generates the smallest $A|\alpha_i|/A|\bar{r}_i|$ ratio and the largest $As^2(\alpha_i)/A\alpha_i^2$ ratio two and three times, respectively. The HXZ model ranks first among the models when assessed by the point estimate of the alpha dispersion metric, $A\alpha_i^2/A\bar{r}_i^2$. The model generates the smallest point estimate in four out of the ten sets of portfolios. When assessed by the point estimate of $A(R^2)$, we find that the FF5 and BS6 models perform better than the other models in explaining average excess returns. Both these models deliver the largest value for the metric in five out of the ten sets of portfolios. For the $\mathrm{SH}^2(\alpha)$ criterion, the HXZ, FF5, FF4, and SY4 models turn out to be the best performers, as each of these models generate the smallest point estimate for the metric in two sets of test asset portfolios.

Comparing the number of statistically significant alphas generated by each of our factor models, we find the SY4 model is the best performer. The model delivers the least number of significant alphas for the 25 size-β portfolios, the 35 size-NI portfolios, the 25 size-RVar portfolios, the 25 size-AC portfolios, the 30 IND portfolios, and the 8 D10−1 portfolios. The HXZ, BS6, FF4, and FF5 models take second, third, fourth, and fifth places, respectively. The SY4 model also turns out to be one of the top performers in explaining the average excess returns on portfolios with small stocks. Out of 42 small size portfolios, the number of significant alphas is 11 in the SY4 model, 10 in the BS6 model, and 12 in the HXZ, FF5, and FF4 models. In contrast, both the FFC and the FFAF models generate statistically significant alphas for 19 portfolios, the highest number of significant alphas for small size portfolios produced by any model. The corresponding number for the CAPM is 15, and 16 for the FF3 and FFPS models. This clearly suggests that all of our asset pricing models have some difficulty in explaining returns on small stocks. Overall, the time-series results in Tables 1 and 2 indicate that the SY4 model is the best model, in that it outperforms other models on most performance metrics. The HXZ model takes second place, while the FF5 and BS6 models, performing almost equally well, take joint third place. Stambaugh and Yuan (2016) also report a similar ranking when comparing only the FF3, HXZ, FF5, and SY4 models, although their test assets in time-series regressions are different from those employed in this study.

Common to many recent papers (see, e.g., Fama and French (2015), (2016), Hou et al. (2015), (2017a)), the above time-series analysis of the relative performance of factor models can be criticized on the grounds that it are conducted without any use of a formal statistical criterion. The ranking of models might differ when their relative performance is tested statistically. In the following section, we investigate whether a competing asset pricing model is significantly better than another candidate model by employing the pairwise test for equality of the cross-sectional regression R^2s. The pairwise test accounts for the impact of potential model misspecification, and therefore enables us to draw valid statistical inferences on the relative performance of our competing models. Finally, we conduct a multiple model comparison analysis in order to identify whether a given asset

pricing model significantly outperforms a set of alternative models in terms of the OLS (GLS) cross-sectional regression R^2.

IV. Cross-Sectional Results[7]

A. Size-STR (Short-Term Reversal) Portfolios

Table 3 reports pairwise tests of equality for the cross-sectional regression R^2s for the ten return factor models, estimated using the value-weighted monthly excess returns on the 25 size-STR portfolios. For nonnested models, we present results of the normal test described in Section II.D.1. The results based on the OLS cross-sectional regressions in Panel A show that the CAPM is outperformed by some of the asset pricing models. That is, the CAPM generates a significantly lower cross-sectional R^2 than the FF3, FFAF, FF4, and BS6 models. This is not the case for the HXZ, FF5, and SY4 models, although the differences between the OLS and CAPM R^2 still exceed more than 50 percentage points. This result also shows that the common practice of simply comparing the point estimate of a given performance metric can be misleading when identifying a superior asset pricing model. The BS6 model, which is the best model in the time-series regressions for this set of test asset portfolios, now significantly dominates only the CAPM. We find that the FF5, FF4, SY4, and BS6 models generate higher OLS cross-sectional regression R^2s compared to the HXZ model, yet the differences are not statistically distinguishable from 0. Likewise, the FF5 model generates a higher cross-sectional R^2 compared to the FF4 and SY4 models, yet the differences are not significant, even at the 10% level.

In Panel B of Table 3, which reports equality of R^2s test results based on the GLS cross-sectional regressions, we find that the FF3, FFC, FFPS, FFAF, FF4, and BS6 models outperform the CAPM. The relative performance of other factor pricing models is similar to that observed for the OLS cross-sectional regressions shown in Panel A. To summarize, the pairwise cross-sectional regression R^2 equality test results in Panels A and B favor the FF3, FFAF, FF4, and BS6 models over the remaining asset pricing models when judged by the number of times they significantly dominate another model. However, when focusing just on the SY4, HXZ, FF5, and BS6 models, which are the top four performing models in the time-series regressions, we find none are significantly better than the other three in pairwise tests. Thus, from this point of view, the performances of these four recently proposed asset pricing models are not different from each other when tested on the 25 size-STR portfolios.

Panel A of Table 14 reports multiple model comparison tests of the OLS and GLS cross-sectional R^2s for all factor models, where we differentiate between the nested and the nonnested models. The results based on both the OLS and the GLS cross-sectional regressions, which take into account the process of searching over alternative models, suggest that no asset pricing model is significantly dominated by the others at the 5% level. Notably, we cannot reject the CAPM, which is

[7]Throughout this section, we reject H_0: $\rho_A^2 = \rho_B^2$ if the corresponding p-value is at most 0.05.

https://doi.org/10.1017/S0022109018000947 Published online by Cambridge University Press

TABLE 3

Tests of Equality of Cross-Sectional R^2s: 25 Size-STR Portfolios

Table 3 reports the pairwise tests of equality of the ordinary least squares (OLS) and generalized least squares (GLS) cross-sectional regression R^2s for ten different factor models: the capital asset pricing model (CAPM) of Sharpe (1964) and Lintner (1965); the Fama and French (1993) 3-factor (FF3) model; the Fama and French (1993) and Carhart (1997) 4-factor (FFC) model; the Fama and French (1993) and Pástor and Stambaugh (2003) 4-factor (FFPS) model; the Asness and Frazzini (2013) 3-factor (FFAF) model, which combines their value factor with the market and size factors of the FF3 model; the Hou et al. (2015) q-factor (HXZ) model; the Fama and French (2015) 5-factor (FF5) model; the 4-factor (FF4) model that excludes the value factor from the FF5 model; the Stambaugh and Yuan (2016) 4-factor (SY4) model; and the Barillas and Shanken (2018) 6-factor (BS6) model, which includes the market, size, and momentum factors from the FF5 model, the profitability and investment factors from the HXZ model, and the value factor from the FFAF model. The test assets are the value-weighted monthly excess returns on the 25 size-STR (short-term reversal) portfolios. Panel A (Panel B) shows the difference between the OLS (GLS) cross-sectional regression R^2s of the factor models in row i and column j, denoted as $\hat{\rho}_i^2 - \hat{\rho}_j^2$, and the corresponding p-value (in parentheses) calculated using the Newey and West (1987)-adjusted t-statistic for the test of H_0: $\rho_i^2 = \rho_j^2$. The 1-tailed (2-tailed) p-values in the cases of comparing nested (nonnested) models are computed under the assumption that the factor models are potentially misspecified. The sample period is from Jan. 1968 to Dec. 2016.

Model	FF3	FFC	FFPS	FFAF	HXZ	FF5	FF4	SY4	BS6
Panel A. OLS									
CAPM	−0.577	−0.582	−0.622	−0.487	−0.510	−0.627	−0.573	−0.553	−0.660
	(0.046)	(0.062)	(0.132)	(0.021)	(0.128)	(0.098)	(0.049)	(0.066)	(0.023)
FF3		−0.006	−0.046	0.090	0.066	−0.050	0.004	0.024	−0.083
		(0.649)	(0.155)	(0.195)	(0.693)	(0.319)	(0.972)	(0.667)	(0.259)
FFC			−0.040	0.096	0.072	−0.045	0.010	0.030	−0.077
			(0.554)	(0.203)	(0.638)	(0.357)	(0.925)	(0.540)	(0.300)
FFPS				0.135	0.112	−0.005	0.050	0.069	0.000
				(0.170)	(0.564)	(0.945)	(0.729)	(0.408)	(0.617)
FFAF					−0.023	−0.140	−0.086	−0.066	−0.173
					(0.878)	(0.113)	(0.475)	(0.432)	(0.087)
HXZ						−0.117	−0.062	−0.042	−0.149
						(0.494)	(0.528)	(0.774)	(0.411)
FF5							0.055	0.074	−0.033
							(0.295)	(0.284)	(0.431)
FF4								0.020	−0.087
								(0.852)	(0.465)
SY4									−0.107
									(0.238)
Panel B. GLS									
CAPM	−0.213	−0.221	−0.304	−0.193	−0.115	−0.277	−0.190	−0.163	−0.474
	(0.009)	(0.025)	(0.040)	(0.004)	(0.180)	(0.083)	(0.046)	(0.063)	(0.014)
FF3		−0.008	−0.091	0.020	0.097	−0.064	0.023	0.049	−0.261
		(0.567)	(0.171)	(0.787)	(0.449)	(0.506)	(0.838)	(0.527)	(0.161)
FFC			−0.083	0.028	0.106	−0.056	0.031	0.058	−0.252
			(0.512)	(0.600)	(0.454)	(0.565)	(0.792)	(0.448)	(0.151)
FFPS				0.111	0.189	0.027	0.114	0.141	−0.170
				(0.454)	(0.259)	(0.835)	(0.458)	(0.348)	(0.401)
FFAF					0.078	−0.084	0.003	0.030	−0.280
					(0.588)	(0.470)	(0.978)	(0.704)	(0.101)
HXZ						−0.161	−0.074	−0.048	−0.358
						(0.318)	(0.391)	(0.724)	(0.095)
FF5							0.087	0.113	−0.197
							(0.143)	(0.405)	(0.135)
FF4								0.026	−0.284
								(0.834)	(0.096)
SY4									−0.310
									(0.129)

nested by the remaining nine factor models that we consider, although some of these models generate a cross-sectional R^2 that is three to four times higher.[8]

[8] Kan et al. (2013) provide similar evidence for the CAPM when tested on the 25 size-B/M portfolios and five industry portfolios.

https://doi.org/10.1017/S0022109018000947 Published online by Cambridge University Press

B. Size-CI (Abnormal Capital Investment) Portfolios

When using the value-weighted monthly excess returns on the 25 size-CI portfolios as test assets, it can be seen in Panel A of Table 4 that all but the FFC, SY4, and BS6 models have significantly higher R^2s compared to the CAPM. In

TABLE 4

Tests of Equality of Cross-Sectional R^2s: 25 Size-CI Portfolios

Table 4 reports the pairwise tests of equality of the ordinary least squares (OLS) and generalized least squares (GLS) cross-sectional regression R^2s for ten different factor models: the capital asset pricing model (CAPM) of Sharpe (1964) and Lintner (1965); the Fama and French (1993) 3-factor (FF3) model; the Fama and French (1993) and Carhart (1997) 4-factor (FFC) model; the Fama and French (1993) and Pástor and Stambaugh (2003) 4-factor (FFPS) model; the Asness and Frazzini (2013) 3-factor (FFAF) model, which combines their value factor with the market and size factors of the FF3 model; the Hou et al. (2015) q-factor (HXZ) model; the Fama and French (2015) 5-factor (FF5) model; the 4-factor (FF4) model that excludes the value factor from the FF5 model; the Stambaugh and Yuan (2016) 4-factor (SY4) model; and the Barillas and Shanken (2018) 6-factor (BS6) model, which includes the market, size, and momentum factors from the FF5 model, the profitability and investment factors from the HXZ model, and the value factor from the FFAF model. The test assets are the value-weighted monthly excess returns on the 25 size-CI (abnormal capital investment) portfolios. Panel A (Panel B) shows the difference between the OLS (GLS) cross-sectional regression R^2s of the factor models in row i and column j, denoted as $\hat{\rho}_i^2 - \hat{\rho}_j^2$, and the corresponding p-value (in parentheses) calculated using the Newey and West (1987)-adjusted t-statistic for the test of H_0: $\rho_i^2 = \rho_j^2$. The 1-tailed (2-tailed) p-values in the cases of comparing nested (nonnested) models are computed under the assumption that the factor models are potentially misspecified. The sample period is from Jan. 1968 to Dec. 2016.

Model	FF3	FFC	FFPS	FFAF	HXZ	FF5	FF4	SY4	BS6
Panel A. OLS									
CAPM	−0.352 (0.023)	−0.482 (0.150)	−0.365 (0.036)	−0.301 (0.034)	−0.403 (0.048)	−0.352 (0.043)	−0.352 (0.036)	−0.442 (0.052)	−0.486 (0.141)
FF3		−0.130 (0.035)	−0.013 (0.398)	0.051 (0.230)	−0.051 (0.431)	0.000 (0.984)	0.001 (0.982)	−0.090 (0.225)	−0.134 (0.151)
FFC			0.117 (0.200)	0.181 (0.077)	0.079 (0.298)	0.130 (0.136)	0.131 (0.128)	0.040 (0.449)	−0.003 (0.924)
FFPS				0.064 (0.258)	−0.038 (0.532)	0.013 (0.686)	0.013 (0.693)	−0.077 (0.326)	0.000 (0.228)
FFAF					−0.102 (0.243)	−0.051 (0.304)	−0.050 (0.336)	−0.141 (0.114)	−0.185 (0.076)
HXZ						0.050 (0.330)	0.051 (0.287)	−0.039 (0.513)	−0.083 (0.323)
FF5							0.001 (0.831)	−0.090 (0.181)	−0.133 (0.140)
FF4								−0.090 (0.164)	−0.134 (0.135)
SY4									−0.044 (0.463)
Panel B. GLS									
CAPM	−0.273 (0.002)	−0.507 (0.009)	−0.300 (0.010)	−0.127 (0.045)	−0.383 (0.008)	−0.297 (0.036)	−0.287 (0.013)	−0.467 (0.004)	−0.474 (0.032)
FF3		−0.233 (0.025)	−0.026 (0.359)	0.146 (0.072)	−0.110 (0.378)	−0.023 (0.704)	−0.014 (0.875)	−0.193 (0.179)	−0.201 (0.224)
FFC			0.207 (0.242)	0.379 (0.051)	0.124 (0.425)	0.210 (0.164)	0.220 (0.150)	0.040 (0.719)	0.032 (0.586)
FFPS				0.172 (0.097)	−0.084 (0.501)	0.003 (0.969)	0.013 (0.896)	−0.167 (0.290)	−0.175 (0.323)
FFAF					−0.256 (0.079)	−0.169 (0.122)	−0.160 (0.169)	−0.339 (0.043)	−0.347 (0.067)
HXZ						0.087 (0.451)	0.096 (0.339)	−0.084 (0.530)	−0.091 (0.490)
FF5							0.010 (0.604)	−0.170 (0.218)	−0.178 (0.247)
FF4								−0.180 (0.176)	−0.187 (0.206)
SY4									−0.008 (0.941)

https://doi.org/10.1017/S0022109018000947 Published online by Cambridge University Press

https://doi.org/10.1017/S0022109018000947 Published online by Cambridge University Press

fact, the OLS cross-sectional regression R^2 differences between these models and the CAPM exceed as much as 40 percentage points (see, e.g., the HXZ model). We observe that there are several occasions on which a large R^2 difference does not lead to the rejection of the null hypothesis of equal R^2s for two competing factor pricing models. For example, the OLS cross-sectional R^2 of the BS6 model is higher than that of the FF5 model by approximately 13 percentage points, yet the null hypothesis still cannot be rejected (p-value $= 0.140$). A qualitatively similar finding emerges in Panel B, where the GLS cross-sectional regression R^2 of the BS6 model is higher than that of the FF5 model by approximately 18 percentage points, and still the difference in R^2 is statistically insignificant at conventional levels. We also find that the CAPM produces a cross-sectional R^2 that is statistically lower than those of the competing models estimated by the GLS regressions. Furthermore, in Panel B, both the FFC and the SY4 models now show an improved performance. For example, the SY4 model significantly outperforms the FFAF model. It is worth emphasizing that the HXZ, FF5, SY4, and BS6 models are never statistically dominated at the 5% level in Panels A and B. Overall, our pairwise test results, based on the OLS and GLS cross-sectional regression R^2s, indicate that the FFC model performs best in explaining the average excess returns on the 25 size-CI portfolios. The FF3, FFPS, and FF4 models, the three best factor models identified informally for this set of test asset portfolios in the time-series regressions, no longer perform well in the horse race. Examining the multiple model comparison test results, reported in Panel B of Table 14, we find that the FFAF (GLS) model is rejected at the 5% significance level. Despite the lower cross-sectional regression R^2 of the CAPM, we fail to reject the null hypothesis that it performs at least as well as all other asset pricing models.

C. Size-DR (Distress Risk) Portfolios

In Table 5, we report the results of the factor pricing model performances in the 25 size-DR portfolios. The pairwise tests of equality of the OLS cross-sectional regression R^2s in Panel A reveal that no model statistically outperforms, nor is outperformed, by any of the other models at conventional significance levels. Hence, the 25 size-DR portfolios present the biggest challenge for all the asset pricing models. A similar picture emerges when model performance is analyzed on the basis of differences in the GLS cross-sectional regression R^2s (Panel B). Here, the R^2 differences exceed by as much as 26 percentage points, yet are still statistically insignificant. These findings from pairwise tests are consistent with those of the multiple model comparison tests in Panel C of Table 14. All of the p-values are above the conventional levels of significance, which suggests that none of the factor pricing models is outperformed by the rest of the models. Clearly, the above observation contrasts with that from the time-series regressions (see Panel C of Table 1), where the HXZ model is the best performer, informally, in the horse race.

D. Size-β (Market Beta) Portfolios

Table 6 reports the results with test assets as the 25 size-β portfolios. In Panel A, we notice that only the HXZ, FF5, FF4, and SY4 models offer a somewhat superior performance, by generating significantly higher OLS cross-sectional R^2s.

TABLE 5

Tests of Equality of Cross-Sectional R^2s: 25 Size-DR Portfolios

Table 5 reports the pairwise tests of equality of the ordinary least squares (OLS) and generalized least squares (GLS) cross-sectional regression R^2s for ten different factor models: the capital asset pricing model (CAPM) of Sharpe (1964) and Lintner (1965); the Fama and French (1993) 3-factor (FF3) model; the Fama and French (1993) and Carhart (1997) 4-factor (FFC) model; the Fama and French (1993) and Pástor and Stambaugh (2003) 4-factor (FFPS) model; the Asness and Frazzini (2013) 3-factor (FFAF) model, which combines their value factor with the market and size factors of the FF3 model; the Hou et al. (2015) q-factor (HXZ) model; the Fama and French (2015) 5-factor (FF5) model; the 4-factor (FF4) model that excludes the value factor from the FF5 model; the Stambaugh and Yuan (2016) 4-factor (SY4) model; and the Barillas and Shanken (2018) 6-factor (BS6) model, which includes the market, size, and momentum factors from the FF5 model, the profitability and investment factors from the HXZ model, and the value factor from the FFAF model. The test assets are the value-weighted monthly excess returns on the 25 size-DR (distress risk) portfolios. Panel A (Panel B) shows the difference between the OLS (GLS) cross-sectional regression R^2s of the factor models in row i and column j, denoted as $\hat{\rho}_i^2 - \hat{\rho}_j^2$, and the corresponding p-value (in parentheses) calculated using the Newey and West (1987)-adjusted t-statistic for the test of H_0: $\rho_i^2 = \rho_j^2$. The 1-tailed (2-tailed) p-values in the cases of comparing nested (nonnested) models are computed under the assumption that the factor models are potentially misspecified. The sample period is from Jan. 1968 to Dec. 2016.

Model	FF3	FFC	FFPS	FFAF	HXZ	FF5	FF4	SY4	BS6
Panel A. OLS									
CAPM	−0.217	−0.297	−0.219	−0.251	−0.246	−0.252	−0.187	−0.270	−0.298
	(0.495)	(0.561)	(0.573)	(0.454)	(0.536)	(0.611)	(0.658)	(0.513)	(0.712)
FF3		−0.081	−0.002	−0.034	−0.029	−0.035	0.029	−0.053	−0.081
		(0.235)	(0.831)	(0.489)	(0.782)	(0.684)	(0.682)	(0.571)	(0.502)
FFC			0.078	0.047	0.051	0.045	0.110	0.027	−0.001
			(0.569)	(0.607)	(0.519)	(0.744)	(0.365)	(0.761)	(0.967)
FFPS				−0.032	−0.027	−0.033	0.032	−0.051	0.000
				(0.597)	(0.810)	(0.706)	(0.674)	(0.604)	(0.522)
FFAF					0.005	−0.001	0.063	−0.019	−0.048
					(0.951)	(0.988)	(0.405)	(0.786)	(0.584)
HXZ						−0.006	0.059	−0.024	−0.052
						(0.962)	(0.413)	(0.781)	(0.550)
FF5							0.065	−0.018	−0.046
							(0.232)	(0.790)	(0.711)
FF4								−0.083	−0.111
								(0.410)	(0.364)
SY4									−0.028
									(0.722)
Panel B. GLS									
CAPM	−0.086	−0.242	−0.088	−0.133	−0.148	−0.102	−0.058	−0.149	−0.256
	(0.431)	(0.362)	(0.650)	(0.278)	(0.452)	(0.779)	(0.800)	(0.430)	(0.632)
FF3		−0.156	−0.002	−0.047	−0.062	−0.016	0.028	−0.063	−0.170
		(0.191)	(0.861)	(0.396)	(0.610)	(0.805)	(0.647)	(0.605)	(0.432)
FFC			0.154	0.109	0.094	0.140	0.184	0.093	−0.014
			(0.453)	(0.521)	(0.579)	(0.483)	(0.371)	(0.538)	(0.818)
FFPS				−0.045	−0.060	−0.015	0.029	−0.062	−0.169
				(0.479)	(0.638)	(0.834)	(0.646)	(0.622)	(0.445)
FFAF					−0.015	0.030	0.074	−0.017	−0.124
					(0.878)	(0.707)	(0.415)	(0.876)	(0.518)
HXZ						0.046	0.090	−0.001	−0.108
						(0.721)	(0.421)	(0.990)	(0.550)
FF5							0.044	−0.047	−0.154
							(0.404)	(0.637)	(0.494)
FF4								−0.091	−0.198
								(0.496)	(0.351)
SY4									−0.107
									(0.539)

In particular, each of these factor pricing models outperforms the CAPM. The differences in R^2 between these models and the CAPM exceed more than 70 percentage points. We emphasize that the relatively good performance of the SY4 model in Panel A is consistent with that observed for the time-series regressions

https://doi.org/10.1017/S0022109018000947 Published online by Cambridge University Press

TABLE 6

Tests of Equality of Cross-Sectional R^2s: 25 Size-β Portfolios

Table 6 reports the pairwise tests of equality of the ordinary least squares (OLS) and generalized least squares (GLS) cross-sectional regression R^2s for ten different factor models: the capital asset pricing model (CAPM) of Sharpe (1964) and Lintner (1965); the Fama and French (1993) 3-factor (FF3) model; the Fama and French (1993) and Carhart (1997) 4-factor (FFC) model; the Fama and French (1993) and Pástor and Stambaugh (2003) 4-factor (FFPS) model; the Asness and Frazzini (2013) 3-factor (FFAF) model, which combines their value factor with the market and size factors of the FF3 model; the Hou et al. (2015) q-factor (HXZ) model; the Fama and French (2015) 5-factor (FF5) model; the 4-factor (FF4) model that excludes the value factor from the FF5 model; the Stambaugh and Yuan (2016) 4-factor (SY4) model; and the Barillas and Shanken (2018) 6-factor (BS6) model, which includes the market, size, and momentum factors from the FF5 model, the profitability and investment factors from the HXZ model, and the value factor from the FFAF model. The test assets are the value-weighted monthly excess returns on the 25 size-β (market beta) portfolios. Panel A (Panel B) shows the difference between the OLS (GLS) cross-sectional regression R^2s of the factor models in row i and column j, denoted as $\hat{\rho}_i^2 - \hat{\rho}_j^2$, and the corresponding p-value (in parentheses) calculated using the Newey and West (1987)-adjusted t-statistic for the test of H_0: $\rho_i^2 = \rho_j^2$. The 1-tailed (2-tailed) p-values in the cases of comparing nested (nonnested) models are computed under the assumption that the factor models are potentially misspecified. The sample period is from Jan. 1968 to Dec. 2016.

Model	FF3	FFC	FFPS	FFAF	HXZ	FF5	FF4	SY4	BS6
Panel A. OLS									
CAPM	−0.693 (0.071)	−0.785 (0.099)	−0.710 (0.096)	−0.554 (0.113)	−0.743 (0.039)	−0.749 (0.044)	−0.745 (0.034)	−0.750 (0.045)	−0.778 (0.072)
FF3		−0.091 (0.063)	−0.016 (0.321)	0.139 (0.200)	−0.050 (0.539)	−0.056 (0.427)	−0.052 (0.524)	−0.057 (0.575)	−0.084 (0.368)
FFC			0.075 (0.351)	0.231 (0.155)	0.041 (0.569)	0.035 (0.590)	0.040 (0.563)	0.034 (0.682)	0.007 (0.864)
FFPS				0.156 (0.202)	−0.034 (0.612)	−0.039 (0.484)	−0.035 (0.590)	−0.041 (0.662)	0.000 (0.426)
FFAF					−0.189 (0.234)	−0.195 (0.205)	−0.191 (0.225)	−0.196 (0.231)	−0.224 (0.171)
HXZ						−0.006 (0.865)	−0.001 (0.965)	−0.007 (0.938)	−0.034 (0.569)
FF5							0.004 (0.594)	−0.001 (0.987)	−0.028 (0.642)
FF4								−0.006 (0.937)	−0.033 (0.587)
SY4									−0.027 (0.767)
Panel B. GLS									
CAPM	−0.218 (0.021)	−0.443 (0.029)	−0.246 (0.047)	−0.070 (0.347)	−0.337 (0.013)	−0.325 (0.028)	−0.323 (0.013)	−0.429 (0.009)	−0.448 (0.051)
FF3		−0.225 (0.051)	−0.028 (0.365)	0.148 (0.175)	−0.119 (0.372)	−0.107 (0.324)	−0.105 (0.373)	−0.212 (0.249)	−0.230 (0.255)
FFC			0.198 (0.320)	0.373 (0.102)	0.106 (0.537)	0.118 (0.546)	0.120 (0.542)	0.014 (0.940)	−0.004 (0.954)
FFPS				0.175 (0.165)	−0.091 (0.476)	−0.079 (0.465)	−0.077 (0.505)	−0.184 (0.321)	−0.202 (0.325)
FFAF					−0.267 (0.100)	−0.255 (0.089)	−0.253 (0.091)	−0.359 (0.054)	−0.377 (0.088)
HXZ						0.012 (0.883)	0.014 (0.859)	−0.093 (0.544)	−0.111 (0.468)
FF5							0.002 (0.784)	−0.105 (0.450)	−0.123 (0.517)
FF4								−0.107 (0.425)	−0.125 (0.509)
SY4									−0.018 (0.916)

using the same set of anomaly portfolios. The superior performance of the HXZ, FF5, FF4, and SY4 models can also be seen in Panel B, where pairwise tests of equality for R^2s are based on the GLS cross-sectional regressions. Consistent with Panel A, the CAPM again shows up as the worst-performing asset pricing model.

https://doi.org/10.1017/S0022109018000947 Published online by Cambridge University Press

Here, it is additionally significantly outperformed by the FF3, FFC, and FFPS models. Overall, the OLS and GLS cross-sectional results show that the HXZ, FF5, FF4, and SY4 models do a relatively better job in explaining the average excess returns on size and market β sorted portfolios. Yet looking at the multiple model comparison tests in Panel D of Table 14, we find that only the FFAF (GLS) model is rejected at the 5% level. Once again, the CAPM is not rejected at conventional levels of significance. That is, it performs at least as well as all others when assessed within the multiple model comparison framework.

E. Size-NI (net share issues) Portfolios

The pairwise tests of equality for factor model R^2s based on the 35 size-NI portfolios in Table 7 show that the SY4 model is the best overall performer, followed by the HXZ model. In Panel A, from the results based on the OLS cross-sectional regression R^2s, it is observed that the SY4 model significantly dominates the CAPM, FF3, FFPS, and FFAF models. The HXZ model, however, fails to perform significantly better than the FFPS model in terms of R^2. The BS6 model now demonstrates an improved relative performance for this set of anomaly portfolios compared with those observed in Tables 3 through 6. The model outperforms the CAPM, FF3, and FFAF models by generating significantly higher OLS cross-sectional regression R^2s. However, in the pairwise test, when comparing the BS6 and FFPS models, we observe that the BS6 model is statistically outperformed by the FFPS model. We further notice that all but the FFC model easily outperform the CAPM. However, the FFC model clearly dominates both the FF3 and the FFAF models. The results based on the GLS cross-sectional regression R^2s show a better performance for the HXZ model compared with that observed in Panel A. Now, the HXZ model lines up as the best-performing asset pricing model, along with the SY4 model. However, the relative performances of the FF5, FF4, and BS6 models now fall short slightly. In particular, these models are no longer able to generate a significantly higher GLS cross-sectional R^2 compared to the FF3 model. In Panels A and B, it can also be seen that the HXZ, FF5, and SY4 models are not outperformed in any of our pairwise R^2 tests. Furthermore, the relative performance of the SY4 model is consistent with that observed in the time-series regressions.

Turning now to the multiple model comparison test results, reported in Panel E of Table 14, we see that the CAPM (OLS) is rejected at the 5% level of significance. Moreover, for the OLS-based comparisons, the FF3 and FFAF models are statistically dominated. The GLS-based comparison tests show that the FF3, FFPS, and FFAF models are dominated at the 5% level. Notably, the CAPM (GLS) marginally misses rejection of the null hypothesis with a p-value of 0.051. As for the remaining asset pricing models, the corresponding p-values suggest that none can be rejected in the multiple model comparison tests of the OLS and GLS R^2s.

F. Size-MAX (Lottery Demand) Portfolios

Moving to the 25 size-MAX portfolios, in Table 8, we find that the CAPM is outperformed by most other asset pricing models. In particular, the model shows a poor performance relative to all but the FF5 model in Panel A, in terms of the

https://doi.org/10.1017/S0022109018000947 Published online by Cambridge University Press

TABLE 7

Tests of Equality of Cross-Sectional R^2s: 35 Size-NI Portfolios

Table 7 reports the pairwise tests of equality of the ordinary least squares (OLS) and generalized least squares (GLS) cross-sectional regression R^2s for ten different factor models: the capital asset pricing model (CAPM) of Sharpe (1964) and Lintner (1965); the Fama and French (1993) 3-factor (FF3) model; the Fama and French (1993) and Carhart (1997) 4-factor (FFC) model; the Fama and French (1993) and Pástor and Stambaugh (2003) 4-factor (FFPS) model; the Asness and Frazzini (2013) 3-factor (FFAF) model, which combines their value factor with the market and size factors of the FF3 model; the Hou et al. (2015) q-factor (HXZ) model; the Fama and French (2015) 5-factor (FF5) model; the 4-factor (FF4) model that excludes the value factor from the FF5 model; the Stambaugh and Yuan (2016) 4-factor (SY4) model; and the Barillas and Shanken (2018) 6-factor (BS6) model, which includes the market, size, and momentum factors from the FF5 model, the profitability and investment factors from the HXZ model, and the value factor from the FFAF model. The test assets are the value-weighted monthly excess returns on the 35 size-NI (net share issues) portfolios. Panel A (Panel B) shows the difference between the OLS (GLS) cross-sectional regression R^2s of the factor models in row i and column j, denoted as $\hat{\rho}_i^2 - \hat{\rho}_j^2$, and the corresponding p-value (in parentheses) calculated using the Newey and West (1987)-adjusted t-statistic for the test of H_0: $\rho_i^2 = \rho_j^2$. The 1-tailed (2-tailed) p-values in the cases of comparing nested (nonnested) models are computed under the assumption that the factor models are potentially misspecified. The sample period is from Jan. 1968 to Dec. 2016.

Model	FF3	FFC	FFPS	FFAF	HXZ	FF5	FF4	SY4	BS6
Panel A. OLS									
CAPM	−0.438	−0.641	−0.463	−0.315	−0.645	−0.607	−0.596	−0.660	−0.665
	(0.006)	(0.055)	(0.016)	(0.017)	(0.004)	(0.004)	(0.005)	(0.003)	(0.020)
FF3		−0.203	−0.025	0.124	−0.206	−0.169	−0.158	−0.222	−0.227
		(0.009)	(0.349)	(0.047)	(0.011)	(0.025)	(0.031)	(0.015)	(0.020)
FFC			0.178	0.326	−0.004	0.034	0.045	−0.019	−0.024
			(0.070)	(0.010)	(0.964)	(0.734)	(0.677)	(0.792)	(0.700)
FFPS				0.148	−0.182	−0.144	−0.133	−0.197	0.001
				(0.064)	(0.052)	(0.109)	(0.134)	(0.045)	(0.037)
FFAF					−0.330	−0.292	−0.282	−0.345	−0.350
					(0.004)	(0.008)	(0.013)	(0.003)	(0.004)
HXZ						0.038	0.048	−0.015	−0.020
						(0.434)	(0.365)	(0.817)	(0.689)
FF5							0.011	−0.053	−0.058
							(0.412)	(0.404)	(0.416)
FF4								−0.064	−0.069
								(0.427)	(0.415)
SY4									−0.005
									(0.922)
Panel B. GLS									
CAPM	−0.094	−0.278	−0.101	−0.029	−0.307	−0.226	−0.223	−0.336	−0.329
	(0.007)	(0.013)	(0.108)	(0.207)	(0.000)	(0.004)	(0.001)	(0.000)	(0.016)
FF3		−0.184	−0.008	0.065	−0.214	−0.132	−0.130	−0.242	−0.235
		(0.019)	(0.605)	(0.128)	(0.023)	(0.075)	(0.068)	(0.018)	(0.053)
FFC			0.177	0.249	−0.029	0.052	0.055	−0.058	−0.051
			(0.151)	(0.067)	(0.797)	(0.702)	(0.689)	(0.589)	(0.513)
FFPS				0.072	−0.206	−0.125	−0.122	−0.235	−0.227
				(0.155)	(0.041)	(0.132)	(0.127)	(0.027)	(0.066)
FFAF					−0.278	−0.197	−0.194	−0.307	−0.300
					(0.006)	(0.014)	(0.017)	(0.003)	(0.018)
HXZ						0.081	0.084	−0.029	−0.021
						(0.229)	(0.223)	(0.729)	(0.742)
FF5							0.003	−0.110	−0.103
							(0.681)	(0.180)	(0.329)
FF4								−0.113	−0.105
								(0.196)	(0.333)
SY4									0.007
									(0.931)

OLS cross-sectional regressions. We also observe that in Panel A, the BS6 model demonstrates a somewhat better performance in explaining the average excess returns on portfolios. In addition to outperforming the CAPM, the model produces a significantly higher cross-sectional R^2 compared to the FFAF model. Looking

https://doi.org/10.1017/S0022109018000947 Published online by Cambridge University Press

TABLE 8

Tests of Equality of Cross-Sectional R^2s: 25 Size-MAX Portfolios

Table 8 reports the pairwise tests of equality of the ordinary least squares (OLS) and generalized least squares (GLS) cross-sectional regression R^2s for ten different factor models: the capital asset pricing model (CAPM) of Sharpe (1964) and Lintner (1965); the Fama and French (1993) 3-factor (FF3) model; the Fama and French (1993) and Carhart (1997) 4-factor (FFC) model; the Fama and French (1993) and Pástor and Stambaugh (2003) 4-factor (FFPS) model; the Asness and Frazzini (2013) 3-factor (FFAF) model, which combines their value factor with the market and size factors of the FF3 model; the Hou et al. (2015) q-factor (HXZ) model; the Fama and French (2015) 5-factor (FF5) model; the 4-factor (FF4) model that excludes the value factor from the FF5 model; the Stambaugh and Yuan (2016) 4-factor (SY4) model; and the Barillas and Shanken (2018) 6-factor (BS6) model, which includes the market, size, and momentum factors from the FF5 model, the profitability and investment factors from the HXZ model, and the value factor from the FFAF model. The test assets are the value-weighted monthly excess returns on the 25 size-MAX (lottery demand) portfolios. Panel A (Panel B) shows the difference between the OLS (GLS) cross-sectional regression R^2s of the factor models in row i and column j, denoted as $\hat{\rho}_i^2 - \hat{\rho}_j^2$, and the corresponding p-value (in parentheses) calculated using the Newey and West (1987)-adjusted t-statistic for the test of H_0: $\rho_i^2 = \rho_j^2$. The 1-tailed (2-tailed) p-values in the cases of comparing nested (nonnested) models are computed under the assumption that the factor models are potentially misspecified. The sample period is from Jan. 1968 to Dec. 2016.

Model	FF3	FFC	FFPS	FFAF	HXZ	FF5	FF4	SY4	BS6
Panel A. OLS									
CAPM	−0.313	−0.383	−0.315	−0.212	−0.345	−0.366	−0.354	−0.359	−0.393
	(0.007)	(0.039)	(0.011)	(0.023)	(0.009)	(0.056)	(0.012)	(0.007)	(0.034)
FF3		−0.069	−0.002	0.101	−0.032	−0.052	−0.041	−0.046	−0.080
		(0.091)	(0.691)	(0.108)	(0.572)	(0.424)	(0.469)	(0.301)	(0.231)
FFC			0.067	0.170	0.037	0.017	0.028	0.023	−0.011
			(0.213)	(0.061)	(0.458)	(0.829)	(0.656)	(0.554)	(0.830)
FFPS				0.103	−0.030	−0.050	−0.039	−0.044	0.001
				(0.088)	(0.603)	(0.445)	(0.497)	(0.342)	(0.238)
FFAF					−0.133	−0.153	−0.142	−0.147	−0.181
					(0.123)	(0.070)	(0.080)	(0.081)	(0.040)
HXZ						−0.020	−0.009	−0.014	−0.048
						(0.764)	(0.824)	(0.760)	(0.369)
FF5							0.011	0.007	−0.028
							(0.544)	(0.931)	(0.661)
FF4								−0.005	−0.039
								(0.932)	(0.456)
SY4									−0.034
									(0.554)
Panel B. GLS									
CAPM	−0.107	−0.219	−0.110	−0.024	−0.229	−0.245	−0.238	−0.293	−0.263
	(0.003)	(0.032)	(0.040)	(0.371)	(0.004)	(0.058)	(0.015)	(0.002)	(0.169)
FF3		−0.113	−0.003	0.082	−0.122	−0.138	−0.132	−0.186	−0.156
		(0.092)	(0.682)	(0.093)	(0.243)	(0.167)	(0.145)	(0.053)	(0.121)
FFC			0.109	0.195	−0.009	−0.025	−0.019	−0.074	−0.044
			(0.357)	(0.157)	(0.924)	(0.838)	(0.881)	(0.416)	(0.651)
FFPS				0.086	−0.119	−0.135	−0.128	−0.183	−0.153
				(0.110)	(0.251)	(0.172)	(0.148)	(0.058)	(0.126)
FFAF					−0.205	−0.221	−0.214	−0.269	−0.239
					(0.048)	(0.015)	(0.010)	(0.016)	(0.017)
HXZ						−0.016	−0.010	−0.064	−0.034
						(0.777)	(0.889)	(0.474)	(0.328)
FF5							0.007	−0.048	−0.018
							(0.681)	(0.607)	(0.746)
FF4								−0.055	−0.025
								(0.594)	(0.670)
SY4									0.030
									(0.742)

at the pairwise test results in Panel B, based on the GLS cross-sectional regressions, we find that the relative performance of the CAPM improves marginally. Now it has no significant differences compared to the FFAF and BS6 models, in addition to the FF5 model, in terms of R^2. An improvement in performance of

https://doi.org/10.1017/S0022109018000947 Published online by Cambridge University Press

the HXZ, FF5, FF4, and SY4 models can also be observed, as these models now generate statistically higher GLS cross-sectional regression R^2s compared to the FFAF model. Taken together, the pairwise test results in Panels A and B indicate that the HXZ, FF4, SY4, and BS6 models are the top performers among the ten factor pricing models being compared. Finally, focusing on the OLS- and GLS-based multiple model comparison test results in Panel F of Table 14, we find that only the FFAF (GLS) model is rejected at the 5% level of significance.

G. Size-RVar (Residual Variance) Portfolios

The CAPM performs the worst when we conduct the pairwise cross-sectional regression R^2 equality tests for our asset pricing models using monthly excess returns on the 25 size-RVar portfolios, as reported in Table 9. On the other hand, both the FF5 and the FF4 models consistently exhibit the best overall performance, although neither is statistically better than the HXZ or BS6 model in terms of the cross-sectional R^2. The most striking result in Table 9 is that both the FF5 and FF4 models now statistically dominate the SY4 model (see Panel B). The GLS cross-sectional regression R^2 differences with the SY4 model exceed more than 25 percentage points. In Panel A, it is also observable that the HXZ and BS6 models are jointly ranked, following the FF5 and FF4 models, in terms of the number of times they dominate other asset pricing models. Between these two models, however, the BS6 model is statistically dominated by the FFPS model in the pairwise test of equality of cross-sectional R^2s. The relative performance of the BS6 model improves markedly in Panel B, where it is no longer dominated by the FFPS model. In fact, we observe the opposite: the FFPS model is now statistically outperformed by the BS6 model. On the other hand, the performance of the HXZ model drops slightly, as it fails to generate a statistically higher GLS cross-sectional R^2 compared to the FF3 model. In Panel B, we further notice that the performance of the SY4 model is at a minimum in terms of statistically outperforming other factor pricing models, in that it is significantly better than only one other model, namely the CAPM. Finally, for the OLS-based multiple model comparisons in Panel G of Table 14, it can be seen that the FF3, FFPS, and FFAF models are rejected at the 5% level. The number of model rejections increases when we consider the GLS-based multiple model comparison tests. Now, the FFC and SY4 models are also dominated at the 5% level of significance. These results on the performance of the factor pricing models contrast with those in the time-series regressions for the 25 size-RVar portfolios.

H. Size-AC (Accruals) Portfolios

The R^2 equality tests using the 25 size-AC portfolios in Table 10 show a best overall performance for the FF4 model. The FFC and FF5 models also show a modest performance in the horse race, as they generate significantly higher OLS cross-sectional regression R^2s compared to the FF3 model and the CAPM, respectively. No remaining competing asset pricing models statistically dominate, nor are dominated, by any of the other models at the 5% level. We notice that both the FFC and the FF5 models fail to outperform any of the other models in Panel B, which shows the GLS cross-sectional regressions. Consistent with the findings in Panel A, we see that the HXZ, FF5, SY4, and BS6 models neither

https://doi.org/10.1017/S0022109018000947 Published online by Cambridge University Press

TABLE 9

Tests of Equality of Cross-Sectional R^2s: 25 Size-RVar Portfolios

Table 9 reports the pairwise tests of equality of the ordinary least squares (OLS) and generalized least squares (GLS) cross-sectional regression R^2s for ten different factor models: the capital asset pricing model (CAPM) of Sharpe (1964) and Lintner (1965); the Fama and French (1993) 3-factor (FF3) model; the Fama and French (1993) and Carhart (1997) 4-factor (FFC) model; the Fama and French (1993) and Pástor and Stambaugh (2003) 4-factor (FFPS) model; the Asness and Frazzini (2013) 3-factor (FFAF) model, which combines their value factor with the market and size factors of the FF3 model; the Hou et al. (2015) q-factor (HXZ) model; the Fama and French (2015) 5-factor (FF5) model; the 4-factor (FF4) model that excludes the value factor from the FF5 model; the Stambaugh and Yuan (2016) 4-factor (SY4) model; and the Barillas and Shanken (2018) 6-factor (BS6) model, which includes the market, size, and momentum factors from the FF5 model, the profitability and investment factors from the HXZ model, and the value factor from the FFAF model. The test assets are the value-weighted monthly excess returns on the 25 size-RVar (residual variance) portfolios. Panel A (Panel B) shows the difference between the OLS (GLS) cross-sectional regression R^2s of the factor models in row i and column j, denoted as $\hat{\rho}_i^2 - \hat{\rho}_j^2$, and the corresponding p-value (in parentheses) calculated using the Newey and West (1987)-adjusted t-statistic for the test of H_0: $\rho_i^2 = \rho_j^2$. The 1-tailed (2-tailed) p-values in the cases of comparing nested (nonnested) models are computed under the assumption that the factor models are potentially misspecified. The sample period is from Jan. 1968 to Dec. 2016.

Model	FF3	FFC	FFPS	FFAF	HXZ	FF5	FF4	SY4	BS6
Panel A. OLS									
CAPM	−0.296	−0.481	−0.314	−0.096	−0.542	−0.580	−0.580	−0.520	−0.579
	(0.006)	(0.071)	(0.019)	(0.162)	(0.002)	(0.017)	(0.008)	(0.013)	(0.013)
FF3		−0.185	−0.018	0.200	−0.246	−0.285	−0.284	−0.224	−0.283
		(0.048)	(0.409)	(0.069)	(0.034)	(0.012)	(0.011)	(0.054)	(0.010)
FFC			0.167	0.385	−0.061	−0.099	−0.099	−0.039	−0.098
			(0.136)	(0.025)	(0.291)	(0.166)	(0.160)	(0.417)	(0.152)
FFPS				0.218	−0.228	−0.267	−0.266	−0.206	0.003
				(0.089)	(0.061)	(0.024)	(0.023)	(0.094)	(0.017)
FFAF					−0.446	−0.485	−0.484	−0.424	−0.483
					(0.006)	(0.003)	(0.003)	(0.011)	(0.004)
HXZ						−0.039	−0.038	0.022	−0.037
						(0.388)	(0.360)	(0.489)	(0.341)
FF5							0.001	0.060	0.002
							(0.863)	(0.207)	(0.963)
FF4								0.060	0.001
								(0.180)	(0.972)
SY4									−0.059
									(0.241)
Panel B. GLS									
CAPM	−0.028	−0.051	−0.030	−0.006	−0.266	−0.450	−0.449	−0.170	−0.392
	(0.220)	(0.400)	(0.495)	(0.778)	(0.001)	(0.006)	(0.001)	(0.040)	(0.013)
FF3		−0.023	−0.003	0.021	−0.238	−0.422	−0.422	−0.143	−0.364
		(0.431)	(0.725)	(0.452)	(0.069)	(0.012)	(0.010)	(0.273)	(0.019)
FFC			0.020	0.045	−0.215	−0.399	−0.398	−0.119	−0.341
			(0.718)	(0.475)	(0.051)	(0.016)	(0.015)	(0.228)	(0.037)
FFPS				0.024	−0.235	−0.420	−0.419	−0.140	−0.361
				(0.422)	(0.075)	(0.011)	(0.010)	(0.277)	(0.021)
FFAF					−0.259	−0.444	−0.443	−0.164	−0.386
					(0.035)	(0.007)	(0.006)	(0.183)	(0.012)
HXZ						−0.184	−0.184	0.095	−0.126
						(0.137)	(0.127)	(0.190)	(0.250)
FF5							0.001	0.280	0.058
							(0.878)	(0.035)	(0.556)
FF4								0.279	0.057
								(0.031)	(0.559)
SY4									−0.222
									(0.131)

statistically dominate, nor are dominated, by any of the competing factor models. This evidence contrasts with the good performance of the SY4 model in the time-series regressions using the same set of anomaly portfolios. Further, analyzing the

https://doi.org/10.1017/S0022109018000947 Published online by Cambridge University Press

TABLE 10

Tests of Equality of Cross-Sectional R^2s: 25 Size-AC Portfolios

Table 10 reports the pairwise tests of equality of the ordinary least squares (OLS) and generalized least squares (GLS) cross-sectional regression R^2s for ten different factor models: the capital asset pricing model (CAPM) of Sharpe (1964) and Lintner (1965); the Fama and French (1993) 3-factor (FF3) model; the Fama and French (1993) and Carhart (1997) 4-factor (FFC) model; the Fama and French (1993) and Pástor and Stambaugh (2003) 4-factor (FFPS) model; the Asness and Frazzini (2013) 3-factor (FFAF) model, which combines their value factor with the market and size factors of the FF3 model; the Hou et al. (2015) q-factor (HXZ) model; the Fama and French (2015) 5-factor (FF5) model; the 4-factor (FF4) model that excludes the value factor from the FF5 model; the Stambaugh and Yuan (2016) 4-factor (SY4) model; and the Barillas and Shanken (2018) 6-factor (BS6) model, which includes the market, size, and momentum factors from the FF5 model, the profitability and investment factors from the HXZ model, and the value factor from the FFAF model. The test assets are the value-weighted monthly excess returns on the 25 size-AC (accruals) portfolios. Panel A (Panel B) shows the difference between the OLS (GLS) cross-sectional regression R^2s of the factor models in row i and column j, denoted as $\hat{\rho}_i^2 - \hat{\rho}_j^2$, and the corresponding p-value (in parentheses) calculated using the Newey and West (1987)-adjusted t-statistic for the test of H_0: $\rho_i^2 = \rho_j^2$. The 1-tailed (2-tailed) p-values in the cases of comparing nested (nonnested) models are computed under the assumption that the factor models are potentially misspecified. The sample period is from Jan. 1968 to Dec. 2016.

Model	FF3	FFC	FFPS	FFAF	HXZ	FF5	FF4	SY4	BS6
Panel A. OLS									
CAPM	−0.360 (0.093)	−0.639 (0.201)	−0.382 (0.144)	−0.335 (0.121)	−0.494 (0.079)	−0.584 (0.048)	−0.551 (0.025)	−0.489 (0.106)	−0.680 (0.202)
FF3		−0.279 (0.007)	−0.022 (0.568)	0.025 (0.504)	−0.134 (0.092)	−0.224 (0.100)	−0.191 (0.099)	−0.129 (0.179)	−0.320 (0.083)
FFC			0.257 (0.185)	0.304 (0.143)	0.145 (0.420)	0.055 (0.717)	0.088 (0.556)	0.150 (0.412)	−0.041 (0.459)
FFPS				0.047 (0.557)	−0.112 (0.340)	−0.202 (0.168)	−0.169 (0.203)	−0.107 (0.340)	0.001 (0.098)
FFAF					−0.159 (0.083)	−0.249 (0.080)	−0.216 (0.078)	−0.154 (0.154)	−0.345 (0.081)
HXZ						−0.090 (0.393)	−0.057 (0.507)	0.005 (0.958)	−0.186 (0.259)
FF5							0.033 (0.196)	0.095 (0.514)	−0.096 (0.466)
FF4								0.062 (0.636)	−0.129 (0.324)
SY4									−0.191 (0.275)
Panel B. GLS									
CAPM	−0.080 (0.103)	−0.301 (0.066)	−0.106 (0.328)	−0.077 (0.158)	−0.140 (0.092)	−0.212 (0.059)	−0.158 (0.033)	−0.174 (0.092)	−0.352 (0.111)
FF3		−0.222 (0.054)	−0.026 (0.510)	0.003 (0.891)	−0.061 (0.430)	−0.132 (0.235)	−0.078 (0.290)	−0.095 (0.334)	−0.272 (0.145)
FFC			0.196 (0.352)	0.225 (0.245)	0.161 (0.410)	0.089 (0.658)	0.144 (0.468)	0.127 (0.449)	−0.050 (0.432)
FFPS				0.029 (0.700)	−0.035 (0.777)	−0.106 (0.452)	−0.052 (0.658)	−0.068 (0.556)	−0.246 (0.227)
FFAF					−0.064 (0.491)	−0.135 (0.241)	−0.081 (0.350)	−0.097 (0.342)	−0.275 (0.142)
HXZ						−0.072 (0.471)	−0.017 (0.814)	−0.034 (0.738)	−0.211 (0.239)
FF5							0.055 (0.185)	0.038 (0.779)	−0.140 (0.431)
FF4								−0.017 (0.890)	−0.194 (0.273)
SY4									−0.178 (0.276)

multiple model comparison test results in Panel H of Table 14, we find that just the FF3 (OLS) model is statistically dominated by the others.

https://doi.org/10.1017/S0022109018000947 Published online by Cambridge University Press

I. IND (Industry) Portfolios

We see in Panel A of Table 11 that no asset pricing model statistically outperforms, nor is outperformed, by any of the other competing models. All of the p-values are much higher than the 10% level of significance. The statistical

TABLE 11

Tests of Equality of Cross-Sectional R^2s: 30 IND Portfolios

Table 11 reports the pairwise tests of equality of the ordinary least squares (OLS) and generalized least squares (GLS) cross-sectional regression R^2s for ten different factor models: the capital asset pricing model (CAPM) of Sharpe (1964) and Lintner (1965); the Fama and French (1993) 3-factor (FF3) model; the Fama and French (1993) and Carhart (1997) 4-factor (FFC) model; the Fama and French (1993) and Pástor and Stambaugh (2003) 4-factor (FFPS) model; the Asness and Frazzini (2013) 3-factor (FFAF) model, which combines their value factor with the market and size factors of the FF3 model; the Hou et al. (2015) q-factor (HXZ) model; the Fama and French (2015) 5-factor (FF5) model; the 4-factor (FF4) model that excludes the value factor from the FF5 model; the Stambaugh and Yuan (2016) 4-factor (SY4) model; and the Barillas and Shanken (2018) 6-factor (BS6) model, which includes the market, size, and momentum factors from the FF5 model, the profitability and investment factors from the HXZ model, and the value factor from the FFAF model. The test assets are the value-weighted monthly excess returns on the 30 IND (industry) portfolios. Panel A (Panel B) shows the difference between the OLS (GLS) cross-sectional regression R^2s of the factor models in row i and column j, denoted as $\hat{\rho}_i^2 - \hat{\rho}_j^2$, and the corresponding p-value (in parentheses) calculated using the Newey and West (1987)-adjusted t-statistic for the test of H_0: $\rho_i^2 = \rho_j^2$. The 1-tailed (2-tailed) p-values in the cases of comparing nested (nonnested) models are computed under the assumption that the factor models are potentially misspecified. The sample period is from Jan. 1968 to Dec. 2016.

Model	FF3	FFC	FFPS	FFAF	HXZ	FF5	FF4	SY4	BS6
Panel A. OLS									
CAPM	−0.044 (0.782)	−0.046 (0.936)	−0.045 (0.941)	−0.046 (0.773)	−0.168 (0.644)	−0.280 (0.645)	−0.206 (0.635)	−0.210 (0.550)	−0.183 (0.907)
FF3		−0.001 (0.927)	−0.001 (0.953)	−0.001 (0.949)	−0.123 (0.615)	−0.235 (0.427)	−0.162 (0.636)	−0.166 (0.526)	−0.139 (0.645)
FFC			0.001 (0.971)	0.000 (0.996)	−0.122 (0.615)	−0.234 (0.428)	−0.161 (0.626)	−0.164 (0.531)	−0.138 (0.636)
FFPS				−0.001 (0.985)	−0.123 (0.600)	−0.235 (0.416)	−0.161 (0.622)	−0.165 (0.516)	0.000 (0.604)
FFAF					−0.122 (0.601)	−0.234 (0.414)	−0.161 (0.634)	−0.164 (0.511)	−0.138 (0.640)
HXZ						−0.112 (0.387)	−0.039 (0.805)	−0.042 (0.708)	−0.016 (0.884)
FF5							0.073 (0.456)	0.070 (0.478)	0.096 (0.619)
FF4								−0.004 (0.987)	0.023 (0.781)
SY4									0.027 (0.888)
Panel B. GLS									
CAPM	−0.087 (0.381)	−0.102 (0.520)	−0.087 (0.591)	−0.080 (0.404)	−0.113 (0.498)	−0.164 (0.518)	−0.161 (0.361)	−0.104 (0.562)	−0.170 (0.671)
FF3		−0.015 (0.574)	0.000 (0.967)	0.007 (0.769)	−0.027 (0.711)	−0.077 (0.525)	−0.074 (0.549)	−0.017 (0.845)	−0.083 (0.564)
FFC			0.015 (0.786)	0.022 (0.750)	−0.012 (0.916)	−0.062 (0.664)	−0.059 (0.682)	−0.002 (0.986)	−0.068 (0.628)
FFPS				0.007 (0.761)	−0.026 (0.708)	−0.077 (0.524)	−0.074 (0.549)	−0.017 (0.845)	−0.083 (0.563)
FFAF					−0.033 (0.622)	−0.084 (0.482)	−0.081 (0.502)	−0.024 (0.769)	−0.090 (0.529)
HXZ						−0.051 (0.489)	−0.048 (0.534)	0.009 (0.886)	−0.057 (0.645)
FF5							0.003 (0.819)	0.060 (0.509)	−0.006 (0.964)
FF4								0.057 (0.574)	−0.009 (0.939)
SY4									−0.066 (0.682)

https://doi.org/10.1017/S0022109018000947 Published online by Cambridge University Press

https://doi.org/10.1017/S0022109018000947 Published online by Cambridge University Press

inferences about model performance also remain the same when we look at the pairwise tests of equality for the GLS cross-sectional regression R^2s in Panel B. Clearly, these observations suggest that the 30 IND portfolios pose a serious challenge for all of our factor pricing models. When model performance is analyzed, in Panel I of Table 14, using the multiple model comparison tests, we once again see that no models is rejected at conventional significance levels. This empirical evidence contrasts with that from the time-series regressions, in Panel I of Table 1, where an informal comparison of performance metrics suggests that the CAPM is the best model.

J. D10−1 (High Minus Low Decile) Portfolios

Table 12 reports the results based on the set of D10−1 portfolios corresponding to eight anomalies. The pairwise tests of equality for the OLS cross-sectional regression R^2s in Panel A show that most of the models significantly outperform the CAPM. The two exceptions are the FFPS and FFAF models. Although both these models generate a higher cross-sectional R^2 compared to the CAPM, the differences are statistically indistinguishable from 0 (at the 5% level). A somewhat similar picture emerges in Panel B, with the GLS cross-sectional regressions, where we see that only the FFPS model fails to outperform the CAPM. We further find, in Panels A and B, that the FF3, FFC, FFPS, FFAF, HXZ, FF5, FF4, SY4, and BS6 models are never statistically dominated in any of our pairwise tests of equality for cross-sectional regression R^2s. These results on the relative performance of our factor pricing models contrast with those in the time-series regressions for the same set of anomaly long-short portfolios. Finally, in Panel J of Table 14, we see that the FFAF model is statistically dominated at the conventional 5% level when the GLS-based multiple model comparison tests are considered.[9]

K. Overall Model Performance in Cross-Sectional Regressions

We summarize the pairwise cross-sectional regression R^2 equality test results from Tables 3 through 12 for all the asset pricing models in Table 13. In Panel A, for the OLS-based results, it can be seen that the FF4 model offers the best overall performance, in that it statistically outperforms competing asset pricing models the most times in a majority of the sets of test asset portfolios. The FF5 model takes second place in the horse race. The HXZ and BS6 models, which perform almost equally well, turn out jointly to be the next best models. It also appears that the SY4 model fails to secure a position in the top three places among the ten different return factor models. Analyzing the GLS cross-sectional regression based test results in Panel B, we once again find that the FF4 and FF5 models secure the first and the second place, respectively. Yet the SY4 model now takes joint third place, along with the HXZ and BS6 models. We emphasize that in Panels A and B, the HXZ, FF5, and FF4 models are the only asset pricing models that are never statistically outperformed at the 5% level in any of our pairwise R^2 tests. The results from the sequential tests for nonnested models reported in the Supplementary Material also corroborate our findings in Table 13.

[9]We do not report results for the CAPM in Panel J of Table 14, as there are more factors in the expanded model than the cross-sectional data points. In this case, no unique solution exists for the second-pass regression.

TABLE 12

Tests of Equality of Cross-Sectional R^2s: 8 D10−1 Portfolios

Table 12 reports the pairwise tests of equality of the ordinary least squares (OLS) and generalized least squares (GLS) cross-sectional regression R^2s for ten different factor models: the capital asset pricing model (CAPM) of Sharpe (1964) and Lintner (1965); the Fama and French (1993) 3-factor (FF3) model; the Fama and French (1993) and Carhart (1997) 4-factor (FFC) model; the Fama and French (1993) and Pástor and Stambaugh (2003) 4-factor (FFPS) model; the Asness and Frazzini (2013) 3-factor (FFAF) model, which combines their value factor with the market and size factors of the FF3 model; the Hou et al. (2015) q-factor (HXZ) model; the Fama and French (2015) 5-factor (FF5) model; the 4-factor (FF4) model that excludes the value factor from the FF5 model; the Stambaugh and Yuan (2016) 4-factor (SY4) model; and the Barillas and Shanken (2018) 6-factor (BS6) model, which includes the market, size, and momentum factors from the FF5 model, the profitability and investment factors from the HXZ model, and the value factor from the FFAF model. The test assets are the value-weighted monthly excess returns on the 8 D10−1 (high minus low decile) portfolios. Panel A (Panel B) shows the difference between the OLS (GLS) cross-sectional regression R^2s of the factor models in row i and column j, denoted as $\hat{\rho}_i^2 - \hat{\rho}_j^2$, and the corresponding p-value (in parentheses) calculated using the Newey and West (1987)-adjusted t-statistic for the test of H_0: $\rho_i^2 = \rho_j^2$. The 1-tailed (2-tailed) p-values in the cases of comparing nested (nonnested) models are computed under the assumption that the factor models are potentially misspecified. The sample period is from Jan. 1968 to Dec. 2016.

Model	FF3	FFC	FFPS	FFAF	HXZ	FF5	FF4	SY4	BS6
Panel A. OLS									
CAPM	−0.690 (0.026)	−0.861 (0.035)	−0.832 (0.291)	−0.543 (0.051)	−0.889 (0.011)	−0.914 (0.009)	−0.817 (0.008)	−0.972 (0.007)	−0.986 (0.039)
FF3		−0.171 (0.071)	−0.142 (0.198)	0.147 (0.172)	−0.199 (0.176)	−0.224 (0.204)	−0.127 (0.203)	−0.281 (0.195)	−0.296 (0.195)
FFC			0.029 (0.877)	0.318 (0.192)	−0.028 (0.817)	−0.053 (0.733)	0.043 (0.798)	−0.111 (0.324)	−0.126 (0.267)
FFPS				0.289 (0.246)	−0.057 (0.740)	−0.082 (0.690)	0.014 (0.940)	−0.140 (0.500)	0.017 (0.482)
FFAF					−0.346 (0.121)	−0.371 (0.099)	−0.275 (0.125)	−0.429 (0.112)	−0.444 (0.111)
HXZ						−0.025 (0.837)	0.071 (0.412)	−0.083 (0.554)	−0.098 (0.518)
FF5							0.097 (0.155)	−0.058 (0.580)	−0.072 (0.528)
FF4								−0.154 (0.389)	−0.169 (0.381)
SY4									−0.015 (0.642)
Panel B. GLS									
CAPM	−0.607 (0.005)	−0.669 (0.010)	−0.687 (0.111)	−0.461 (0.016)	−0.797 (0.001)	−0.823 (0.002)	−0.782 (0.001)	−0.883 (0.000)	−0.913 (0.014)
FF3		−0.062 (0.310)	−0.080 (0.389)	0.146 (0.181)	−0.191 (0.186)	−0.216 (0.268)	−0.175 (0.216)	−0.276 (0.197)	−0.306 (0.187)
FFC			−0.018 (0.913)	0.207 (0.308)	−0.129 (0.215)	−0.154 (0.507)	−0.114 (0.533)	−0.214 (0.272)	−0.244 (0.234)
FFPS				0.226 (0.336)	−0.110 (0.537)	−0.136 (0.593)	−0.095 (0.653)	−0.196 (0.410)	−0.226 (0.353)
FFAF					−0.336 (0.134)	−0.361 (0.110)	−0.321 (0.103)	−0.422 (0.112)	−0.452 (0.105)
HXZ						−0.025 (0.880)	0.015 (0.896)	−0.085 (0.520)	−0.116 (0.438)
FF5							0.040 (0.356)	−0.060 (0.620)	−0.090 (0.519)
FF4								−0.101 (0.459)	−0.131 (0.422)
SY4									−0.030 (0.639)

However, looking at the multiple model comparison test results, summarized in Table 15, we see that only the HXZ, FF5, FF4, and BS6 models are never rejected at the 5% level of significance. We highlight that this observation holds regardless of whether the OLS- or GLS-based multiple model comparison tests

https://doi.org/10.1017/S0022109018000947 Published online by Cambridge University Press

TABLE 13

Summary of Factor Model Performance: Tests of Equality of Cross-Sectional R^2s

Table 13 reports the summary of the pairwise tests for equality of the ordinary least squares (OLS) and generalized least squares (GLS) cross-sectional regression R^2s for ten different factor models: the capital asset pricing model (CAPM) of Sharpe (1964) and Lintner (1965); the Fama and French (1993) 3-factor (FF3) model; the Fama and French (1993) and Carhart (1997) 4-factor (FFC) model; the Fama and French (1993) and Pástor and Stambaugh (2003) 4-factor (FFPS) model; the Asness and Frazzini (2013) 3-factor (FFAF) model, which combines their value factor with the market and size factors of the FF3 model; the Hou et al. (2015) q-factor (HXZ) model; the Fama and French (2015) 5-factor (FF5) model; the 4-factor (FF4) model that excludes the value factor from the FF5 model; the Stambaugh and Yuan (2016) 4-factor (SY4) model; and the Barillas and Shanken (2018) 6-factor (BS6) model, which includes the market, size, and momentum factors from the FF5 model, the profitability and investment factors from the HXZ model, and the value factor from the FFAF model. The test assets are the value-weighted monthly excess returns on anomaly portfolios: the 25 size-STR (short-term reversal) portfolios, the 25 size-CI (abnormal capital investment) portfolios, the 25 size-DR (distress risk) portfolios, the 25 size-β (market beta) portfolios, the 35 size-NI (net share issues) portfolios, the 25 size-MAX (lottery demand) portfolios, the 25 size-RVar (residual variance) portfolios, the 25 size-AC (accruals) portfolios, the 30 IND (industry) portfolios, and the 8 D10−1 (high minus low decile) portfolios. Panel A (Panel B) shows, for each set of anomaly portfolios, the number of times a given factor model produces a significantly (i.e., at the 5% level) higher OLS (GLS) cross-sectional regression R^2 relative to another competing factor model. See also notes to Table 3.

Model	Size-STR	Size-CI	Size-DR	Size-β	Size-NI	Size-MAX	Size-RVar	Size-AC	IND	D10−1
Panel A. OLS										
CAPM	0	0	0	0	0	0	0	0	0	0
FF3	1	1	0	0	2	1	1	0	0	1
FFC	0	1	0	0	2	1	2	1	0	1
FFPS	0	1	0	0	2	1	2	0	0	0
FFAF	1	1	0	0	1	1	0	0	0	0
HXZ	0	1	0	1	3	1	3	0	0	1
FF5	0	1	0	1	3	0	4	1	0	1
FF4	1	1	0	1	3	1	4	1	0	1
SY4	0	0	0	1	4	1	2	0	0	1
BS6	1	0	0	0	3	2	3	0	0	1
Panel B. GLS										
CAPM	0	0	0	0	0	0	0	0	0	0
FF3	1	1	0	1	1	1	0	0	0	1
FFC	1	2	0	1	2	1	0	0	0	1
FFPS	1	1	0	1	0	1	0	0	0	0
FFAF	1	1	0	0	0	0	0	0	0	1
HXZ	0	1	0	1	4	2	2	0	0	1
FF5	0	1	0	1	2	1	6	0	0	1
FF4	1	1	0	1	2	2	6	1	0	1
SY4	0	2	0	1	4	2	1	0	0	1
BS6	1	1	0	0	2	1	5	0	0	1

are employed. We also find that the SY4 model is rejected in one out the ten sets of test asset portfolios. Although these findings are slightly different from those summarized in Table 13, they can be rationalized by the fact that the pairwise model comparisons do not take into account the process of searching across alternative factor pricing models. Consequently, the pairwise testing procedure can lead to an overstatement of statistical significance (see Barillas et al. (2017), Kan et al. (2013)).

Taken together, our cross-sectional empirical results summarized in Tables 13 and 15 suggest that the HXZ, FF5, FF4, and BS6 models perform about equally well and jointly take first place in the horse race. The SY4 model, which is the top performer in the time-series regressions, is now the next best asset pricing model. Importantly, we also notice that the CAPM performs reasonably good (evidence that is different from most empirical studies based on time-series regressions (see, among others, Fama and French (1996), (2016), Hou et al. (2015), (2017a), and Stambaugh and Yuan (2016))). These cross-sectional statistical findings on the relative performance of the return factor models are different from our informal findings in the time-series regressions, reported in Section III. They can

https://doi.org/10.1017/S0022109018000947 Published online by Cambridge University Press

https://doi.org/10.1017/S0022109018000947 Published online by Cambridge University Press

TABLE 14
Multiple Model Comparison Tests

Table 14 reports the multiple model comparison tests from the ordinary least squares (OLS) and generalized least squares (GLS) cross-sectional regression R^2s for ten different factor models: the capital asset pricing model (CAPM) of Sharpe (1964) and Lintner (1965); the Fama and French (1993) 3-factor (FF3) model; the Fama and French (1993) and Carhart (1997) 4-factor (FFC) model; the Fama and French (1993) and Pástor and Stambaugh (2003) 4-factor (FFPS) model; the Asness and Frazzini (2013) 3-factor (FFAF) model, which combines their value factor with the market and size factors of the FF3 model; the Hou et al. (2015) q-factor (HXZ) model; the Fama and French (2015) 5-factor (FF5) model; the 4-factor (FF4) model that excludes the value factor from the FF5 model; the Stambaugh and Yuan (2016) 4-factor (SY4) model; and the Barillas and Shanken (2018) 6-factor (BS6) model, which includes the market, size, and momentum factors from the FF5 model, the profitability and investment factors from the HXZ model, and the value factor from the FFAF model. $\hat{\rho}^2$ denotes the cross-sectional regression R^2 of the benchmark model; u is the number of alternative factor models in each multiple nonnested model comparison; LR is the value of the likelihood ratio statistic with the corresponding p-value reported in the adjacent column; v is the number of factor models that nest the benchmark model; $\hat{\rho}^2_M - \hat{\rho}^2$ is the difference between the cross-sectional regression R^2 of the expanded factor model (M) and the cross-sectional regression R^2 of the benchmark model with the corresponding p-value reported in the adjacent column. The sample period is from Jan. 1968 to Dec. 2016. See also notes to Table 13.

	OLS							GLS						
Benchmark	$\hat{\rho}^2$	u	LR	p-Value	v	$\hat{\rho}^2_M - \hat{\rho}^2$	p-Value	$\hat{\rho}^2$	u	LR	p-Value	v	$\hat{\rho}^2_M - \hat{\rho}^2$	p-Value
Panel A. 25 Size-STR Portfolios														
CAPM	0.216				9	0.744	0.440	0.008				9	0.734	0.749
FF3	0.793	5	1.281	0.463	2	0.048	0.438	0.221	5	2.448	0.283	2	0.110	0.267
FFC	0.798	6	1.274	0.601				0.229	6	2.875	0.309			
FFPS	0.838	6	0.196	0.650				0.312	6	0.780	0.433			
FFAF	0.703	6	3.245	0.140				0.201	6	3.891	0.164			
HXZ	0.726	6	0.861	0.298				0.123	6	4.007	0.087			
FF5	0.843	6	0.663	0.716				0.285	6	2.736	0.245			
FF4	0.788	6	0.636	0.559	1	0.055	0.295	0.198	6	3.993	0.143	1	0.087	0.143
SY4	0.769	6	1.598	0.402				0.171	6	3.106	0.180			
BS6	0.876	6	0.000	0.858				0.482	6	0.000	0.757			
Panel B. 25 Size-CI Portfolios														
CAPM	0.341				9	0.615	0.328	0.000				9	0.832	0.291
FF3	0.693	5	1.144	0.500	2	0.133	0.070	0.274	5	2.242	0.288	2	0.251	0.057
FFC	0.823	6	0.009	0.808				0.507	6	0.000	0.841			
FFPS	0.706	6	1.017	0.494				0.300	6	1.622	0.333			
FFAF	0.642	6	1.577	0.224				0.128	6	7.392	0.020			
HXZ	0.744	6	0.819	0.515				0.384	6	0.689	0.514			
FF5	0.694	6	1.320	0.520				0.297	6	2.408	0.282			
FF4	0.693	6	1.381	0.500	1	0.001	0.831	0.287	6	2.838	0.230	1	0.010	0.604
SY4	0.783	6	0.487	0.652				0.467	6	0.130	0.733			
BS6	0.827	6	0.000	0.812				0.475	6	0.301	0.681			
Panel C. 25 Size-DR Portfolios														
CAPM	0.522				9	0.395	0.961	0.238				9	0.399	0.992
FF3	0.738	5	0.398	0.525	2	0.101	0.459	0.324	5	1.048	0.454	2	0.173	0.363
FFC	0.819	6	0.002	0.801				0.480	6	0.053	0.784			
FFPS	0.741	6	0.309	0.546				0.326	6	0.785	0.498			
FFAF	0.772	6	0.252	0.824				0.371	6	0.458	0.722			
HXZ	0.768	6	0.357	0.602				0.386	6	0.372	0.607			
FF5	0.774	6	0.116	0.663				0.341	6	0.537	0.576			
FF4	0.709	6	0.575	0.473	1	0.065	0.232	0.297	6	1.478	0.370	1	0.044	0.404
SY4	0.792	6	0.114	0.829				0.388	6	0.411	0.652			
BS6	0.820	6	0.000	0.896				0.495	6	0.000	0.762			
Panel D. 25 Size-β Portfolios														
CAPM	0.046				9	0.856	0.226	0.001				9	0.651	0.559
FF3	0.739	5	0.825	0.548	2	0.093	0.108	0.218	5	2.043	0.261	2	0.233	0.082
FFC	0.830	6	0.000	0.868				0.444	6	0.003	0.803			
FFPS	0.755	6	0.796	0.596				0.246	6	1.435	0.362			
FFAF	0.600	6	1.888	0.157				0.071	6	7.298	0.023			
HXZ	0.789	6	0.327	0.695				0.337	6	0.681	0.603			
FF5	0.795	6	0.259	0.838				0.325	6	0.683	0.607			
FF4	0.790	6	0.305	0.810	1	0.004	0.594	0.323	6	0.756	0.586	1	0.002	0.784
SY4	0.796	6	0.155	0.667				0.430	6	0.011	0.781			
BS6	0.823	6	0.029	0.834				0.448	6	0.000	0.857			

(continued on next page)

TABLE 14 (continued)
Multiple Model Comparison Tests

	OLS							GLS						
Benchmark	$\hat{\rho}^2$	u	LR	p-Value	v	$\hat{\rho}^2_M - \hat{\rho}^2$	p-Value	$\hat{\rho}^2$	u	LR	p-Value	v	$\hat{\rho}^2_M - \hat{\rho}^2$	p-Value
Panel E. 35 Size-NI Portfolios														
CAPM	0.081				9	0.798	0.028	0.014				9	0.589	0.051
FF3	0.520	5	7.043	0.023	2	0.209	0.023	0.107	5	7.819	0.017	2	0.184	0.070
FFC	0.722	6	0.151	0.659				0.292	6	0.462	0.489			
FFPS	0.545	6	4.396	0.079				0.115	6	6.259	0.039			
FFAF	0.396	6	13.180	0.001				0.043	6	13.793	0.001			
HXZ	0.726	6	0.154	0.792				0.321	6	0.165	0.772			
FF5	0.688	6	0.870	0.523				0.240	6	2.511	0.229			
FF4	0.678	6	0.829	0.488	1	0.011	0.412	0.237	6	2.326	0.244	1	0.003	0.681
SY4	0.741	6	0.010	0.862				0.350	6	0.000	0.863			
BS6	0.746	6	0.000	0.899				0.342	6	0.007	0.858			
Panel F. 25 Size-MAX Portfolios														
CAPM	0.309				9	0.636	0.763	0.000				9	0.808	0.612
FF3	0.623	5	1.509	0.409	2	0.111	0.138	0.107	5	4.866	0.060	2	0.134	0.215
FFC	0.692	6	0.046	0.810				0.220	6	0.632	0.435			
FFPS	0.625	6	1.599	0.426				0.110	6	4.583	0.085			
FFAF	0.522	6	4.390	0.060				0.024	6	11.365	0.003			
HXZ	0.655	6	0.824	0.545				0.229	6	1.149	0.500			
FF5	0.675	6	0.195	0.636				0.245	6	0.281	0.675			
FF4	0.664	6	0.543	0.598	1	0.011	0.544	0.238	6	0.318	0.656	1	0.007	0.681
SY4	0.668	6	0.433	0.679				0.293	6	0.000	0.837			
BS6	0.703	6	0.000	0.862				0.263	6	0.107	0.832			
Panel G. 25 Size-RVar Portfolios														
CAPM	0.293				9	0.671	0.282	0.000				9	0.754	0.302
FF3	0.589	5	6.676	0.020	2	0.188	0.066	0.028	5	12.002	0.002	2	0.041	0.576
FFC	0.774	6	1.766	0.316				0.051	6	9.876	0.008			
FFPS	0.607	6	5.467	0.039				0.031	6	11.818	0.003			
FFAF	0.389	6	10.582	0.002				0.007	6	17.716	0.000			
HXZ	0.835	6	0.964	0.609				0.266	6	2.812	0.206			
FF5	0.874	6	0.000	0.858				0.450	6	0.000	0.821			
FF4	0.873	6	0.000	0.868	1	0.001	0.863	0.450	6	0.000	0.827	1	0.001	0.878
SY4	0.813	6	1.399	0.490				0.171	6	5.890	0.039			
BS6	0.872	6	0.002	0.895				0.392	6	0.364	0.545			
Panel H. 25 Size-AC Portfolios														
CAPM	0.041				9	0.864	0.529	0.041				9	0.707	0.899
FF3	0.401	5	3.176	0.203	2	0.303	0.039	0.121	5	3.241	0.236	2	0.268	0.118
FFC	0.679	6	0.558	0.497				0.343	6	0.641	0.452			
FFPS	0.422	6	2.304	0.234				0.147	6	1.724	0.296			
FFAF	0.376	6	3.007	0.158				0.118	6	3.315	0.209			
HXZ	0.535	6	1.350	0.463				0.182	6	1.720	0.355			
FF5	0.625	6	0.496	0.605				0.253	6	0.643	0.508			
FF4	0.592	6	0.856	0.554	1	0.033	0.196	0.199	6	1.304	0.409	1	0.055	0.185
SY4	0.530	6	1.292	0.403				0.215	6	1.352	0.421			
BS6	0.721	6	0.000	0.865				0.393	6	0.000	0.866			
Panel I. 30 IND Portfolios														
CAPM	0.084				9	0.561	0.968	0.008				9	0.357	0.984
FF3	0.128	5	1.019	0.328	2	0.002	0.993	0.094	5	0.527	0.567	2	0.016	0.841
FFC	0.130	6	1.013	0.401				0.109	6	0.277	0.584			
FFPS	0.129	6	1.169	0.563				0.094	6	0.681	0.701			
FFAF	0.130	6	1.106	0.403				0.087	6	0.800	0.624			
HXZ	0.251	6	1.287	0.519				0.121	6	0.659	0.690			
FF5	0.364	6	0.000	0.834				0.172	6	0.002	0.807			
FF4	0.290	6	0.000	0.737	1	0.073	0.456	0.169	6	0.006	0.801	1	0.003	0.819
SY4	0.294	6	0.609	0.544				0.111	6	0.468	0.545			
BS6	0.267	6	0.269	0.515				0.177	6	0.000	0.719			

(continued on next page)

be rationalized using two possible reasons. Firstly, in the time-series regression, a return factor receives a zero pricing error in each sample because the factor

https://doi.org/10.1017/S0022109018000947 Published online by Cambridge University Press

https://doi.org/10.1017/S0022109018000947 Published online by Cambridge University Press

TABLE 14 (continued)
Multiple Model Comparison Tests

	OLS							GLS						
Benchmark	$\hat{\rho}^2$	u	LR	p-Value	v	$\hat{\rho}^2_M - \hat{\rho}^2$	p-Value	$\hat{\rho}^2$	u	LR	p-Value	v	$\hat{\rho}^2_M - \hat{\rho}^2$	p-Value
Panel J. 8 D10 −1 Portfolios														
CAPM	0.004				9	—	—	0.067				9	—	—
FF3	0.694	5	2.243	0.224	2	0.177	0.322	0.674	5	2.706	0.186	2	0.084	0.704
FFC	0.865	6	1.178	0.441				0.736	6	2.144	0.265			
FFPS	0.836	6	0.500	0.480				0.754	6	0.941	0.370			
FFAF	0.547	6	4.399	0.056				0.528	6	5.055	0.048			
HXZ	0.893	6	0.362	0.730				0.865	6	0.616	0.694			
FF5	0.918	6	0.369	0.675				0.890	6	0.429	0.580			
FF4	0.822	6	0.769	0.562	1	0.097	0.155	0.849	6	0.672	0.578	1	0.040	0.356
SY4	0.976	6	0.202	0.813				0.950	6	0.218	0.783			
BS6	0.991	6	0.000	0.858				0.980	6	0.000	0.860			

TABLE 15
Summary of Factor Model Performance: Multiple Model Comparison Tests

Table 15 reports the summary of the multiple model comparison tests of the ordinary least squares (OLS) and generalized least squares (GLS) cross-sectional regression R^2s for ten different factor models: the capital asset pricing model (CAPM) of Sharpe (1964) and Lintner (1965); the Fama and French (1993) 3-factor (FF3) model; the Fama and French (1993) and Carhart (1997) 4-factor (FFC) model; the Fama and French (1993) and Pástor and Stambaugh (2003) 4-factor (FFPS) model; the Asness and Frazzini (2013) 3-factor (FFAF) model, which combines their value factor with the market and size factors of the FF3 model; the Hou et al. (2015) q-factor (HXZ) model; the Fama and French (2015) 5-factor (FF5) model; the 4-factor (FF4) model that excludes the value factor from the FF5 model; the Stambaugh and Yuan (2016) 4-factor (SY4) model; and the Barillas and Shanken (2018) 6-factor (BS6) model, which includes the market, size, and momentum factors from the FF5 model, the profitability and investment factors from the HXZ model, and the value factor from the FFAF model. For each set of anomaly portfolios, X (✓) denotes a rejection of the null hypothesis, at the 5% level, that the benchmark model given in column 1 performs at least as well as all other competing nested (nonnested) factor models. See also notes to Tables 13 and 14.

Model	Size-STR	Size-CI	Size-DR	Size-β	Size-NI	Size-MAX	Size-RVar	Size-AC	IND	D10−1
Panel A. OLS										
CAPM					X					
FF3					X ✓		✓	X		
FFC										
FFPS							✓			
FFAF					✓		✓			
HXZ										
FF5										
FF4										
SY4										
BS6										
Panel B. GLS										
CAPM										
FF3					✓		✓			
FFC							✓			
FFPS					✓		✓			
FFAF		✓		✓	✓	✓	✓			✓
HXZ										
FF5										
FF4										
SY4							✓			
BS6										

risk premium is estimated as the sample mean of the factor, whereas the cross-sectional regression minimizes the sum of squares of all the pricing errors (see Cochrane (2005), p. 244). Secondly, in the cross-sectional approach, we compare the different factor pricing models by relying on formal statistical methods, and therefore take into account the sampling variability. Moreover, the cross-sectional

https://doi.org/10.1017/S0022109018000947 Published online by Cambridge University Press

statistical tests allow for potential model misspecifications, that is, population deviations from the model. We acknowledge that although model comparisons can be sensitive to the test assets, the findings from the tests based on cross-sectional regressions are fairly robust across the various sets of test asset portfolios. Our pairwise R^2 equality test and the multiple model comparison test results reinforce the argument of Barillas and Shanken (2017) and Harvey and Liu (2017) in that factor model comparisons conducted informally, by eyeballing point estimates of pricing performance metrics, can be misleading.

V. Conclusion

This article compares the performance of ten different asset pricing models in explaining the average excess returns on various sets of prominent anomaly portfolios. Our list of models comprises the capital asset pricing model of Sharpe (1964) and Lintner (1965), the Fama and French (1993) 3-factor model, the Fama and French (1993) and Carhart (1997) 4-factor model, the Fama and French (1993) and Pástor and Stambaugh (2003) 4-factor model, the Asness and Frazzini (2013) 3-factor model, the Hou et al. (2015) q-factor model, the Fama and French (2015) 5-factor model, the 4-factor model of Fama and French (2015) that drops the value factor, the Stambaugh and Yuan (2016) 4-factor model, and the Barillas and Shanken (2018) 6-factor model. An informal analysis based on a comprehensive set of performance metrics in the time-series domain shows that the Stambaugh and Yuan (2016) 4-factor model provides the best, though still incomplete, description of average monthly excess returns on a majority of our test asset portfolios. The Hou et al. (2015) q-factor model takes second place, while the Fama and French (2015) 5-factor model and the Barillas and Shanken (2018) 6-factor model jointly take third place in the horse race.

However, a different picture emerges when the relative performance of all the models is tested statistically, using the pairwise cross-sectional regression R^2 equality test and the multiple model comparison test. We find that the Hou et al. (2015) q-factor model, the Fama and French (2015) 5-factor model, the 4-factor model of Fama and French (2015) that drops the value factor, and the Barillas and Shanken (2018) 6-factor model, perform almost equally well and take equal first place. The Stambaugh and Yuan (2016) 4-factor model, which is the top performer in the time-series regressions, turns out to be the next best model. Unlike previous empirical studies, we also find that the capital asset pricing model of Sharpe (1964) and Lintner (1965) performs reasonably well. Our cross-sectional findings remain robust, irrespective of regression methodologies and sequential tests for nonnested models. Our results also indicate that the common practice of identifying the superior model(s) by informally comparing pricing performance metrics can sometimes be misleading. Taken together, our empirical findings have important implications for practitioners who need to evaluate capital budgeting, equity valuation, investment management, and fund performance.

Supplementary Material

Supplementary Material for this article is available at https://doi.org/10.1017/S0022109018000947.

References

Ang, A. *Asset Management: A Systematic Approach to Factor Investing*. New York, NY: Oxford University Press (2014).

Ang, A.; R. J. Hodrick; Y. Xing; and X. Zhang. "The Cross-Section of Volatility and Expected Returns." *Journal of Finance*, 61 (2006), 259–299.

Asness, C., and A. Frazzini. "The Devil in HML's Details." *Journal of Portfolio Management*, 39 (2013), 49–68.

Bali, T. G.; N. Cakici; and R. F. Whitelaw. "Maxing Out: Stocks as Lotteries and the Cross-Section of Expected Returns." *Journal of Financial Economics*, 99 (2011), 427–446.

Bali, T. G.; R. F. Engle; and S. Murray. *Empirical Asset Pricing: The Cross Section of Stock Returns*. New Jersey: John Wiley & Sons, Inc. (2016).

Barillas, F.; R. Kan; C. Robotti; and J. Shanken. "Model Comparison with Sharpe Ratios." Working Paper, Emory University (2017).

Barillas, F., and J. Shanken. "Which Alpha?" *Review of Financial Studies*, 30 (2017), 1316–1338.

Barillas, F., and J. Shanken. "Comparing Asset Pricing Models." *Journal of Finance*, 73 (2018), 715–754.

Bharath, S. T., and T. Shumway. "Forecasting Default with the Merton Distance to Default Model." *Review of Financial Studies*, 21 (2008), 1339–1369.

Black, F.; M. C. Jensen; and M. Scholes. "The Capital Asset Pricing Model: Some Empirical Tests." In *Studies in the Theory of Capital Markets*, M. C. Jensen, ed. New York, NY: Praeger (1972), 79–121.

Brealey, R. A.; S. C. Myers; and F. Allen. *Principles of Corporate Finance*. New York, NY: McGraw-Hill Education (2016).

Campbell, J. Y.; J. Hilscher; and J. Szilagyi. "In Search of Distress Risk." *Journal of Finance*, 63 (2008), 2899–2939.

Carhart, M. M. "On Persistence in Mutual Fund Performance." *Journal of Finance*, 52 (1997), 57–82.

Cochrane, J. H. *Asset Pricing*. Princeton, NJ: Princeton University Press (2005).

Fama, E. F. "Market Efficiency, Long-Term Returns, and Behavioral Finance." *Journal of Financial Economics*, 49 (1998), 283–306.

Fama, E. F., and K. R. French. "Common Risk Factors in the Returns on Stocks and Bonds." *Journal of Financial Economics*, 33 (1993), 3–56.

Fama, E. F., and K. R. French. "Multifactor Explanations of Asset Pricing Anomalies." *Journal of Finance*, 51 (1996), 55–84.

Fama, E. F., and K. R. French. "Industry Costs of Equity." *Journal of Financial Economics*, 43 (1997), 153–193.

Fama, E. F., and K. R. French. "Dissecting Anomalies." *Journal of Finance*, 63 (2008), 1653–1678.

Fama, E. F., and K. R. French. "Luck versus Skill in the Cross-Section of Mutual Fund Returns." *Journal of Finance*, 65 (2010), 1915–1947.

Fama, E. F., and K. R. French. "Size, Value, and Momentum in International Stock Returns." *Journal of Financial Economics*, 105 (2012), 457–472.

Fama, E. F., and K. R. French. "A Five-Factor Asset Pricing Model." *Journal of Financial Economics*, 116 (2015), 1–22.

Fama, E. F., and K. R. French. "Dissecting Anomalies with a Five-Factor Model." *Review of Financial Studies*, 29 (2016), 69–103.

Fama, E. F., and K. R. French. "Choosing Factors." *Journal of Financial Economics*, 128 (2018), 234–252.

Fama, E. F., and J. D. MacBeth. "Risk, Return, and Equilibrium: Empirical Tests." *Journal of Political Economy*, 81 (1973), 607–636.

Feng, G.; S. Giglio; and D. Xiu. "Taming the Factor Zoo." Working Paper, Yale University (2017).

Fischer, B. R., and R. Wermers. *Performance Evaluation and Attribution of Security Portfolios*. Oxford, UK: Academic Press (2012).

Frazzini, A., and L. H. Pedersen. "Betting Against Beta." *Journal of Financial Economics*, 111 (2014), 1–25.

Fu, F. "Idiosyncratic Risk and the Cross-Section of Expected Stock Returns." *Journal of Financial Economics*, 91 (2009), 24–37.

https://doi.org/10.1017/S0022109018000947 Published online by Cambridge University Press

Gibbons, M. R.; S. A. Ross; and J. Shanken. "A Test of the Efficiency of a Given Portfolio." *Econometrica*, 57 (1989), 1121–1152.
Gitman, L. J., and V. A. Mercurio. "Cost of Capital Techniques Used by Major U.S. Firms: Survey and Analysis of Fortune's 1000." *Financial Management*, 14 (1982), 21–29.
Gospodinov, N.; R. Kan; and C. Robotti. "Chi-Squared Tests for Evaluation and Comparison of Asset Pricing Models." *Journal of Econometrics*, 173 (2013), 108–125.
Graham, J. R., and C. R. Harvey. "The Theory and Practice of Corporate Finance: Evidence from the Field." *Journal of Financial Economics*, 60 (2001), 187–243.
Green, J.; J. R. M. Hand; and X. F. Zhang. "The Remarkable Multidimensionality in the Cross-Section of Expected U.S. Stock Returns." Working Paper, Yale University (2014).
Grinold, R. C., and R. N. Kahn. *Active Portfolio Management: Quantitative Theory and Applications*. New York, NY: McGraw-Hill Education (1995).
Gu, L. "Product Market Competition, R&D Investment, and Stock Returns." *Journal of Financial Economics*, 119 (2016), 441–455.
Harvey, C. R., and Y. Liu. "Lucky Factors." Working Paper, Duke University (2017).
Harvey, C. R.; Y. Liu; and H. Zhu. "… and the Cross-Section of Expected Returns." *Review of Financial Studies*, 29 (2016), 5–68.
Hirshleifer, D.; P.-H. Hsu; and D. Li. "Innovative Originality, Profitability, and Stock Returns." *Review of Financial Studies*, 31 (2018), 2553–2605.
Hou, K.; G. A. Karolyi; and B.-C. Kho. "What Factors Drive Global Stock Returns?" *Review of Financial Studies*, 24 (2011), 2527–2574.
Hou, K.; C. Xue; and L. Zhang. "Digesting Anomalies: An Investment Approach." *Review of Financial Studies*, 28 (2015), 650–705.
Hou, K.; C. Xue; and L. Zhang. "A Comparison of New Factor Models." Working Paper, Ohio State University (2017a).
Hou, K.; C. Xue; and L. Zhang. "Replicating Anomalies." Working Paper, Ohio State University (2017b).
Ikenberry, D.; J. Lakonishok; and T. Vermaelen. "Market Underreaction to Open Market Share Repurchases." *Journal of Financial Economics*, 39 (1995), 181–208.
Jagannathan, R., and I. Meier. "Do We Need CAPM for Capital Budgeting?" *Financial Management*, Winter (2002), 5–27.
Jegadeesh, N. "Evidence of Predictable Behavior of Security Returns." *Journal of Finance*, 45 (1990), 881–898.
Kan, R., and C. Robotti. "Model Comparison Using the Hansen–Jagannathan Distance." *Review of Financial Studies*, 22 (2009), 3449–3490.
Kan, R.; C. Robotti; and J. Shanken. "Pricing Model Performance and the Two-Pass Cross-Sectional Regression Methodology." *Journal of Finance*, 68 (2013), 2617–2649.
Kandel, S., and R. F. Stambaugh. "Portfolio Inefficiency and the Cross-Section of Expected Returns." *Journal of Finance*, 50 (1995), 157–184.
Kozak, S.; S. Nagel; and S. Santosh. "Interpreting Factor Models." *Journal of Finance*, 73 (2018), 1183–1223.
Kuehn, L.-A.; M. Simutin; and J. J. Wang. "A Labor Capital Asset Pricing Model." *Journal of Finance*, 72 (2017), 2131–2178.
Lehmann, B. N. "Fads, Martingales, and Market Efficiency." *Quarterly Journal of Economics*, 105 (1990), 1–28.
Lewellen, J.; S. Nagel; and J. Shanken. "A Skeptical Appraisal of Asset Pricing Tests." *Journal of Financial Economics*, 96 (2010), 175–194.
Linnainmaa, J. T., and M. R. Roberts. "The History of the Cross-Section of Stock Returns." *Review of Financial Studies*, 31 (2018), 2606–2649.
Lintner, J. "The Valuation of Risk Assets and the Selection of Risky Investments in Stock Portfolios and Capital Budgets." *Review of Economics and Statistics*, 47 (1965), 13–37.
Loughran, T., and J. R. Ritter. "The New Issues Puzzle." *Journal of Finance*, 50 (1995), 23–51.
McLean, D. R., and J. Pontiff. "Does Academic Research Destroy Stock Return Predictability?" *Journal of Finance*, 71 (2016), 5–32.
Newey, W. K., and K. D. West. "A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix." *Econometrica*, 55 (1987), 703–708.
Pástor, Ľ., and R. F. Stambaugh. "Liquidity Risk and Expected Stock Returns." *Journal of Political Economy*, 111 (2003), 642–685.
Sharpe, W. F. "Capital Asset Prices: A Theory of Market Equilibrium under Conditions of Risk." *Journal of Finance*, 19 (1964), 425–442.
Sloan, R. G. "Do Stock Prices Fully Reflect Information in Accruals and Cash Flows About Future Earnings?" *Accounting Review*, 71 (1996), 289–315.

https://doi.org/10.1017/S0022109018000947 Published online by Cambridge University Press

Stambaugh, R. F.; J. Yu; and Y. Yuan. "Arbitrage Asymmetry and the Idiosyncratic Volatility Puzzle." *Journal of Finance*, 70 (2015), 1903–1948.

Stambaugh, R. F., and Y. Yuan. "Mispricing Factors." *Review of Financial Studies*, 30 (2016), 1270–1315.

Titman, S.; K. C. J. Wei; and F. Xie. "Capital Investments and Stock Returns." *Journal of Financial and Quantitative Analysis*, 39 (2004), 677–700.

Xing, Y. "Interpreting the Value Effect Through the *Q*-Theory: An Empirical Investigation." *Review of Financial Studies*, 21 (2008), 1767–1795.

https://doi.org/10.1017/S0022109018000947 Published online by Cambridge University Press