Do Cross-Sectional Predictors Contain Systematic Information?

JOSEPH ENGELBERG, R. DAVID MCLEAN, JEFFREY PONTIFF,

AND MATTHEW C. RINGGENBERG[*]

Forthcoming, Journal of Financial and Quantitative Analysis

October 2021

ABSTRACT

Firm-level variables that predict cross-sectional stock returns, such as price-to-earnings and short interest, are often averaged and used to predict market returns. Using various samples of cross-sectional predictors and accounting for the number of predictors and their interdependence, we find only weak evidence that cross-sectional predictors make good time-series predictors, especially out-of-sample. The results suggest that cross-sectional predictors do not generally contain systematic information.

Keywords: Return predictability, data snooping, statistical bias, market risk premium.
JEL Code: G00, G14, L3, C1

I. Introduction

Is the market risk premium predictable? Financial research has strived to answer this question going back to at least to Dow (1920), and a large number of variables have been used to predict the time-series of stock returns. A growing number of papers with time-series predictors begin with a cross-sectional predictor and ask whether that cross-sectional variable can be aggregated and then used to predict market returns. Such papers presume that cross-sectional predictors contain systematic information. For example, Baker and Wurgler (2000) cite the extensive market-timing literature, which relates equity issuance to returns in the cross-section, and suggest "issuers try to time *both* their idiosyncratic return and the market return." They find that an aggregated capital raising variable predicts market returns. Similarly, Lewellen (2004) finds that aggregated price-to-fundamental ratios based on dividends, book-to-market, and earnings predict market returns. Robert Shiller's famous CAPE ratio is based on the idea that aggregated price-to-earnings ratios predicts market returns. Hirshleifer, Hou, and Teoh (2009) aggregate Sloan's (1996) accrual anomaly and use it to predict the market. Baker and Wurgler (2007) incorporate several firm-level variables to create an investor sentiment index, and show that it predicts market returns.

Drawing on this large literature, we examine market return predictability from cross-sectional variables through the lens of multiple hypothesis testing. After all, there are many cross-sectional predictors that *could be chosen* in order to predict market returns. Published papers present individual hypothesis tests for predictors that *were chosen*. This paper asks: do cross-sectional variables generally aggregate to make good time-series predictors? Using various samples of cross-sectional predictors and incorporating a multiple hypothesis testing framework, we find that the answer is "no."

We begin by creating a sample of time-series predictors constructed from the population of cross-sectional predictors documented in the literature. We use a sample of 140 cross-sectional predictors that is essentially the population of firm-level characteristics

1

that have demonstrated cross-sectional predictability in the academic literature.[1] Of these 140 predictors, 26 have already been aggregated and used to predict market returns in a published study.[2] These papers are well-cited, having received over 12,000 Google Scholar citations in total.

It is impossible to know the true set of variables examined by researchers. Statistically significant findings are more likely to be published, so the 26 variables in the aforementioned papers represent a lower-bound on the set of predictors considered by researchers. We begin our analysis be examining this sample; we then extend our analyses to examine other subsets motivated by economic theory, as well as the entire sample of 140 predictors that *could* have been considered. Our analyses thus examine the minimum and maximum number of cross-sectional variables that could have been used to generate time-series variables.

We begin by taking each of the 140 firm-level predictors and calculating monthly cross-sectional averages to get a single, monthly value. For each predictor, we construct equal-weighted and value-weighted averages; the resulting database has 280 different predictive variables (140 equal-weighted and 140 value-weighted).[3] To examine the market-level predictability of these variables, we perform both in-sample and out-of-sample tests. Like other papers in the time-series literature, our in-sample tests use the entire sample of data and estimate a single parameter estimate from a time-series regression of the market risk premium on the predictor. Our out-of-sample tests consist of expanding, rolling-window regressions that use only information available at each point in time to examine whether a predictor is useful for forecasting the market's risk premium.

At first glance, it appears that many cross-sectional predictors are good market-level predictors in-sample. When we examine the predictors already studied by the existing

---

[1] This builds on the 97-predictor list used in McLean and Pontiff (2016) and Engelberg, McLean, and Pontiff (2018).

[2] Table IA.VI in the Internet Appendix lists these 26 cross-sectional predictors along with the 23 time-series papers that each was featured in.

[3] We follow existing practice when constructing aggregate predictors from cross-sectional data: although some researchers consider equal-weights (for example, Goyal and Santa-Clara, 2003, and Hirshleifer, Hou, and Teoh, 2009), value-weights are most common. For completeness, we present results with both types of weighting.

literature, we find that 20% of them predict 1-year market returns in an ordinary least squares (OLS) regression with coefficients that are significant at the 10% level or better, and 8% of them are significant at the 1% level.[4] The strength of this result is strongly related to the horizon of predictability: when considering 1-month market returns, the number of significant variables falls to 12% at the 10% level and 2% at the 1% level.

We then consider two predictor subgroups that can clearly be motivated by economic theory. The first is a subsample of *Valuation* predictors, which are variables that are based on ratios of market prices to fundamentals. Many *Valuation* predictors have received attention in the market risk premium literature, including the dividend-to-price and earnings-to-price ratios. Moreover, since virtually all valuation ratios should be a function of discount rates, theory suggests that they should all work in a time-series setting (Kelly and Pruitt, 2013, and Lewellen, 2004). We also form a subsample of *Opinion* predictors (e.g., institutional trading, analyst upgrades) which can be motivated with the sentiment explanation of Baker and Wurgler (2006) or with the information explanation of Seyhun (1988). However, despite the economic motivation for these subsamples, we find only weak in-sample predictability for the *Valuation* and *Opinion* sub-categories. In fact, the *Valuation* and *Opinion* sub-categories perform worse than the sample of predictors from the existing literature, suggesting the predictors in the existing literature are a special subset of all cross-sectional variables.

We also examine a third subsample, *Best Cross-Sectional*, which consists of the ten predictors with the highest cross-sectional *t*-statistic. A number of papers have found evidence that cross-sectional return predictability is smaller in recent periods (e.g., McLean and Pontiff (2016), Green, Hand, and Zhang (2017)), so it is possible that the weak performance of some of our predictors (like those in the *Valuation* and *Opinion* subsamples) results from weakness of the underlying cross-sectional. To account for this

---

[4] There are 51 variables since we create both equal-weighted and value-weighted predictors for each of the 26 variables in the existing literature and one of the resulting 52 possible variables is dropped because it is non-stationary. While we do not find that 51 out of 51 are significant, it is important to note these are re-analyses, not replications, in the language of Welch (2019) because we do not use the same sample or code used in the original papers. Moreover, we examine two versions of each predictor (equal-weighted and value-weighted) while many of the original papers only examine one of these.

3

possibility, we examine a subset of predictors formed from the cross-sectional variables with the best performance. Yet, once again, for this subsample we find only weak evidence of return predictability.

The 51 predictors from the existing literature that we discuss above represent a lower-bound on the set of variables actually considered. Harvey, Liu, and Zhu (2015) note that there is a publication bias in Finance, because journals are less likely to publish results that are not statistically significant. Chen (2020) notes that selective reporting of results makes it difficult to know the true set of variables considered, and this has implications for statistical inference. We therefore examine all 280 possible predictors constructed from the 140 cross-sectional variables in the literature – this set represents an upper bound on the cross-sectional predictors that could have been used to make time-series predictors. Out of the 280 predictive variables, 253 of them are stationary and exhibit sufficient time series variation.[5] Of these 253 predictors, 43 (17%) predict 1-year market returns in an ordinary least squares (OLS) regression with coefficients that are significant at the 10% level or better, and 14 (6%) are significant at the 1% level. Again, the strength of this result is strongly related to the horizon of predictability: when considering 1-month market returns, only 27 of the 253 predictors (11%) are significant at the 10% level and 7 (3%) are significant at the 1% level. We also find that several cross-sectional predictors–such as *Asset Turnover* and *Z-Score*–are among the best performers for market predictability, but have yet to be documented in this literature. For example, value-weighted *Asset Turnover,* which has not been previously proposed as a predictor of market risk premia, predicts the market risk premium with an R-squared of 17.8% at the one-year horizon.

Since we examine a large number of predictors, we expect that some variables will appear significant by chance. Moreover, many of these variables are related to each other so the tests we conduct are not independent. To address both the number of tests we conduct as well as the dependency among the tests, we perform the Romano and Wolf (2016)

---

[5] We test each of the 280 predictors for a unit root using an Augmented Dickey-Fuller (1979) test. If we fail to reject the null that a variable is non-stationary, we then calculate the first-difference. If we fail to reject the null that the first-differenced variable is non-stationary, we drop the variable. We also drop variables that aggregate to form a variable that does not exhibit time-series variation and we filter variables that should not aggregate according to economic logic. See Section II for details.

resampling-based stepdown procedure to compute adjusted p-values that control the family-wise error rate while accounting for the number of tests and dependence. Among methods that control the family-wise error rate, the Romano and Wolf (2016) procedure is preferable to the well-known Bonferroni (Dunn, 1961) or Holm (1979) tests because it considers the dependence structure across multiple tests and thus has more power (Romano and Wolf, 2005, 2016).[6] Our paper is first to apply this procedure to predicting the equity market risk premium. In robustness tests, we also estimate adjusted p-values while controlling the false discovery rate (Benjamini and Yekutieli, 2001). This approach is less conservative than controlling the family-wise error rate (Romano, Shaikh, and Wolf (2008)).[7] We reach the same conclusion with both approaches.

When we apply the Romano and Wolf (2016) stepdown procedure to the 51 predictors from the existing literature, we find weaker evidence that cross-sectional predictors contain systematic information. While 10 are significant at the 1-year horizon using single hypothesis testing, only 3 are significant at the 1-year horizon using the Romano and Wolf procedure. The results suggest the predictors in the existing literature may be a result of selective testing and reporting. Of course, it is not possible to know the true set of cross-sectional variables considered, and Chen (2020) shows that accounting for this could raise or lower the bar for statistical significance. To account for this, we then examine our other samples using the Romano and Wolf procedure. For the *Valuation* and *Opinion* subsamples, as well as the *Best Cross-Sectional* subsample, we find weak evidence of return predictability using multiple testing methods.

We then turn to our full sample of predictors. Using in-sample regressions, when we examine the full set of 253 predictors and perform the Romano and Wolf (2016)

---

[6] Romano, Shaikh, and Wolf (2008) examine simulation evidence for a variety of multiple testing techniques and find that the Romano and Wolf (2005) procedure has good power, especially relative to methods that do not account for the dependence of the individual test statistics. While some other methods that control generalized error rates have even better power, this comes at the cost of having higher false rejection rates. They state, "It appears that when the number of strategies is in the thousands, controlling the [family-wise error rate] becomes too stringent." Since we examine a maximum of 269 strategies, we focus on controlling the family-wise error rate.

[7] Romano, Shaikh, and Wolf (2008) show that false discovery rate methods generally exhibit fewer false negatives at the cost of allowing more false positives.

stepdown procedure we are unable to reject the null of no-predictability at the 1% level for any predictor at any horizon. At the 1-month and 3-month horizons, no variable is significant even at the 10% level. At the 12-month horizon only two variables, *Z-Score* and *Asset Turnover*, have Romano and Wolf p-values that are less than 5%. When we examine more permissive p-values based on controlling the false discovery rate, there are only a few more significant predictors. We find that one predictor is significant at the 5% level at the 1-month horizon and eight (out of 253) are significant at the 12-month horizon. In short, we find most of the cross-sectional variables that were statistically significant when examined in isolation are no longer significant when examined in the context of all cross-sectional predictors. However, a few variables, like *Z-Score* and *Asset Turnover*, do still show evidence of return predictability.

Goyal and Welch (2008) argue that out-of-sample regressions serve as a useful diagnostic. Accordingly, we turn to out-of-sample forecasting regressions – again, we start with the predictors already examined in the literature, and then examine other samples of predictors including the sub-groups motivated by economic theory and the set of all 140 cross-sectional variables that could have been considered. In all samples, we find that things look even bleaker for cross-sectional predictors when we consider out-of-sample tests. While 11 out of the 51 predictors from the existing literature are significant at the 12-month horizon, none are significant at the 1-month horizon. Similarly, among the 253 stationary predictors we examine in out-of-sample regressions, 3% significantly predict market returns at the 1-month horizon and 17% predict market returns at the 12-month horizon (compared with 11% and 17% in-sample). Moreover, once we adjust the individual p-values using the Romano and Wolf (2016) stepdown procedure, we no longer find evidence against the null at the 10% level for any predictor at any horizon. For example, in our out-of-sample tests, *Asset Turnover* again appears to be a strong predictor. It exhibits positive out-of-sample R-squared values at every forecasting horizon, with an impressive R-squared of 17% at the one-year horizon. However, the corresponding Romano and Wolf (2016) adjusted p-value is not statistically significant at any of the usual levels. We find

similar results when controlling the false discovery rate. Taking all of the results together, our findings suggest that cross-sectional predictors do not contain systematic information.

Our paper makes a number of contributions. First, we provide context for existing papers that propose a particular cross-sectional predictor should be transformed into a time-series one. Existing results in the literature that have economically large coefficients and impressive t-statistics could be the result of chance if enough predictors are considered. Our findings show that cross-sectional predictors as a group do not make good time-series predictors. The literature's elevation of individual predictors should consider the economic motivation behind each predictor, and the results should be interpreted in light of our data mining critique.

Second, our results provide new insight into the nature of return predictability, both in the cross-section and the time-series. By aggregating cross-sectional predictors into time series variables, we are able to understand whether cross-sectional variables contain information about the systematic components of returns. Our results suggest they do not. This is surprising, as Wen (2009) finds that asset growth can be aggregated to predict market returns, and several studies (e.g., Hou, Xue, and Zhang, 2009) find that factor models with an investment factor can explain many cross-sectional anomalies. Taken together, these papers suggest that most anomalies should aggregate to predict market returns, yet we find little evidence that they do.

Finally, we contribute to the extensive equity return premium literature. While Goyal and Welch (2008) show that 14 popular time-series variables do not significantly predict returns in out-of-sample tests, subsequent papers have documented evidence of return predictability using firm-level variables aggregated across stocks (e.g., Hirshleifer, Hou, and Teoh (2009), Rapach, Ringgenberg, and Zhou (2016)). Our results extend these findings by showing that several other cross-sectional predictors can be aggregated to form good time series predictors. However, we find that many of these predictors are no longer significant after adjusting for multiple hypothesis testing. Our findings emphasize the importance of considering the impact of data snooping bias when examining time-series return predictability.

The remainder of this paper proceeds as follows: Section I briefly describes the existing literature and outlines the theoretical relation between cross-sectional predictive variables and time series return predictability. Section II describes the data used in this study. Section III characterizes our findings and Section IV concludes.

II. Background

Financial researchers have examined the predictability of stock returns for over a century (e.g., Gibson, 1906) and a large literature has documented evidence of predictability in the cross-section of stock returns. A separate literature has examined the predictability of the equity risk-premium using time-series predictive variables. To date, these two literatures have evolved relatively independently. We connect these two literatures.

A. Time Series Return Predictability

A number of papers find in-sample evidence of time series return predictability, but out-of-sample evidence is rare, suggesting that many predictors are the result of data snooping (i.e., overfitting). For example, Bossaerts and Hillion (1999) use model selection criteria from the statistics literature to choose candidate predictors, which allows them to partially avoid data snooping biases, yet they find that the resulting predictors are unable to forecast out-of-sample returns. Similarly, Goyal and Welch (2008) examine 14 popular predictors from the existing literature and find that they fail to forecast the equity risk premium in out-of-sample tests. Cooper and Gulen (2006) note that researchers have many different choices regarding the specification of predictability tests, including the predictor variables, the estimation periods, and the assets being forecasted. They perform specification searches across these parameters and find that return predictability is highly sensitive to these parameter choices. More recently, Bartsch, Dichtl, Drobetz, and Neuhierl (2017) examine a wide variety of possible permutations of the predictors in Goyal and Welch (2008) and the technical predictors in Neely, Rapach, Tu, and Zhou (2014) and estimate that most out-of-sample performance for these variables is from data snooping.

8

In light of the poor performance of many predictive variables in out-of-sample tests, researchers have focused on developing methodologies that are robust to data snooping concerns (see Harvey, Liu, and Saretto, 2020). Foster, Smith, and Whaley (1997) develop a procedure to account for data snooping biases when evaluating the fit of predictive regressions. White (2000) develops a reality check bootstrap (RCB) to account for data snooping biases that result from specification searches, and Sullivan, Timmermann, and White (1999) apply the RCB procedure to a set of technical trading rules. While the White RCB procedure determines whether the best predictor among a group is statistically significant after adjusting for data snooping biases, Romano and Wolf (2005, 2016) show how to adjust the p-value for each individual predictor to account for data snooping biases. We are the first to apply the Romano and Wolf stepdown procedure to a large set of predictive variables derived from existing academic studies. In robustness checks, we also examine adjusted p-values that control the false discovery rate (Benjamini and Yekutieli, 2001) which is less conservative than the Romano and Wolf procedure.

B. Cross-sectional Return Predictability

While the literature on time series return predictability has generally found that most predictors fail to perform in out-of-sample tests, a large literature finds evidence of return predictability in the cross-section of stocks. More recently, a number of papers re-affirm earlier studies that find return predictability in the cross-section of stocks. As examples, using quintile portfolios, McLean and Pontiff (2016) find that 88% of firm-level predictors generate *t*-statistics greater than 1.5 in the sample period used by the original study. Chen and Zimmerman (2019) replicate 319 firm-level predictors and find that only three fail to reproduce the statistical significance of the original study. Some studies question the breadth of predictability and whether implementable portfolio strategies are possible. Hou, Xue, and Zhang (2020) use a sample of 452 anomalies and conclude that most fail to replicate. Chen and Zimmerman reconcile their results with Hou, Xue, and Zhang (2020) by noting that Hou et al. eliminate microcap stocks from their sample (60% of CRSP), that many of Hou et al.'s anomalies are not unique and are based on the same firm-level variable

9

measured over different horizons or intervals, and that many of Hou et al.'s anomalies are unique to their paper and were never significant to begin with. The fact that Hou et al. find that eliminating microcaps weakens return-predictability is consistent with many studies (see Barberis and Thaler (2003) or Pontiff (2006) for a thorough review) which find that anomalies are weaker in large stocks and strongest in small stocks that are more costly to hold and trade (Pontiff, 1996).

There is though, an evolving discussion centered on the extent to which cross-sectional predictability is the outcome of data mining and whether the literature has overlooked multiple testing issues. Harvey, Liu and Zhu (2016) conclude that the majority of findings in financial economics papers are likely false, and that t-statistic hurdles should be raised to 3. McLean and Pontiff (2016) examine 97 anomalies and find that at most 25% of anomaly returns can be explained by data mining. Linnainmaa and Roberts (2018) conclude that the majority of accounting-based anomalies are the outcome of data mining.

Yet other recent studies find that data mining is likely not the first-order cause of anomaly returns. McLean and Pontiff (2016) find that anomaly returns are 50% lower post-publication, an effect they attribute to both data mining and arbitrage informed by the original publication. Importantly, they reject the null that anomaly returns are zero post-publication. Jacobs and Muller (2019), find that anomaly returns are, on average, significant outside of the US, and do not decline post-publication as they do in the US. Chen and Zimmerman (2019) find that publication bias only accounts for about 12% of anomaly returns. Chen (2021) argues that it would take 10,000 researchers hundreds of years to produce the large t-statistics reported in the anomaly literature.

Overall, the existence, and nature, of cross-sectional return predictability is an important debate that has yet to be resolved. However, none of the existing studies examine the relation between cross-sectional predictors and aggregate market returns, which is the focus of our paper.

C. The Information in Cross-Sectional Variables

While there are extensive literatures on both time series predictability and cross-sectional predictability, with a few notable exceptions they have evolved independently. Several papers show that firm-level anomalies aggregate to market-wide predictors. As we mention earlier, to the best of our knowledge 26 predictors have been aggregated and used to predict market returns in a published study. Table 1 lists these 26 cross-sectional predictors along with the 23 time-series papers that each was featured in. Some examples include Pontiff and Schall (1998) with book-to-market ratios, Campbell and Shiller (1988) with P/E ratios, and Chordia et al. (2002) with insider trading. More recently, Hirshleifer, Hou, and Teoh (2009) find that firm-level accruals and cash-flow, when aggregated across stocks, contain information about market returns, and Wen (2019) shows that aggregate asset growth predicts market returns. Finally, Rapach, Ringgenberg and Zhou (2016) show that firm-level short interest aggregates to form one of the strongest known predictors of market returns.

D. The Information in Non-Systematic Predictors

In this paper, our goal is to understand the relation between cross-sectional return predictability and time-series return predictability. While it may seem natural that cross-sectional return predictors should aggregate to generate time-series return predictors, it is possible to have one without the other.[8] Specifically, cross-sectional variables could predict returns because they forecast the systematic component of returns or the non-systematic component of returns. As such, cross-sectional return predictors do not necessarily aggregate to form good time series predictors. To see this, define a variable $X_{i,t}^{Non-Syst}$ as a non-systematic predictor if it forecasts the *non-systematic* portion of stock returns for

---

[8] Indeed, several existing papers find that firm-level relations do not hold at the aggregate level. Kothari, Lewellen, and Warner (2006) document a negative relation between returns and earnings surprise at the aggregate level, in contrast to the positive relation documented at the firm-level. Similarly, Hirshleifer, Hou, and Teoh (2009) find that the relation between accruals and returns changes sign between firm-level and aggregate-level analyses.

11

asset i on date $t+1$. Without loss of generalization,[9] we can express the return on asset i using the market model:

(1)
$$R_{i,t} = R_f + \beta_i(R_{m,t} - R_f) + \varepsilon_{i,t},$$

where $R_{i,t}$ is the return on stock i on date t, R_f is the risk-free rate, $R_{m,t}$ is the market return, and $\varepsilon_{i,t}$ is the portion of stock i's return that is orthogonal to the market's return. We define a non-systematic predictor as a variable $X_{i,t-1}^{Non-Syst}$ that satisfies $\gamma_1 \neq 0$ in a linear regression of the form: [10]

(2)
$$\widehat{\varepsilon_{i,t}} = \gamma_0 + \gamma_1 X_{i,t-1}^{Non-Syst} + \omega_{i,t},$$

where $\widehat{\varepsilon_{i,t}}$ is the abnormal return from the market model (Sharpe (1964); Lintner (1965)). In other words, a non-systematic predictor, by definition, forecasts the portion of asset i's return that is *not* explained by aggregate market movements. However, while $X_{i,t-1}^{Non-Syst}$ contains information about individual stock returns, it will not aggregate to generate market return predictability. To see this, average equation (2) across assets all N stocks in the economy and multiply both sides by $\frac{mc_{i,t}}{\sum_i^N mc_{i,t}}$, where $mc_{i,t}$ is the market capitalization of stock i on date t:

(3)
$$\frac{mc_{i,t}}{\sum_i^N mc_{i,t}}\sum_{i=1}^N[R_{i,t} - R_f - \beta_i(R_{m,t} - R_f)] = \overline{\gamma_0} + \overline{\gamma_1}\overline{X_{t-1}^{Non-Syst}},$$

where the bar above a variable denotes the value-weighted mean. It is simple to show that the left-hand side of equation (3) is equal to zero. Thus, the value-weighted variable $\overline{X_{t-1}^{Non-Syst}}$ contains *no* information about aggregate market returns.

E. The Information in Systematic Predictors

While non-systematic predictors contain no information about aggregate market returns, it is possible to have a variable that predicts information in the cross-section that

---

[9] In Section B of the Internet Appendix we show the logic in this section generalizes to *any* model with non-systematic and systematic components including the Fama-and French (1992) 3-factor model and the Fama-and French (2014) 5-factor model.

[10] For simplicity, we ignore the sign of the abnormal return and define an anomaly as any variable that predicts abnormal returns in either direction.

contains information about the aggregate risk-premium. Define a variable $X_{i,t-1}^{syst}$ as a systematic predictor if it forecasts the *systematic* portion of stock returns for asset i on date *t+1*. Thus, define a systematic predictor as a variable $X_{i,t-1}^{syst}$ that satisfies $\gamma_1 \neq 0$ in a linear regression of the form:

$$(4) \qquad \beta_i\left(R_{m,t} - R_f\right) = \gamma_0 + \gamma_1 X_{i,t-1}^{syst} + \omega_{i,t}.$$

Because the market beta is 1, it is easy to show that the left-hand side of equation (4) implies a direct linear relation between the predictor variable and the market risk-premium. Notice also that this relation goes in both directions: if a time-series predictor is constructed from individual assets, it *must* contain information about the systematic portion of individual asset returns.[11,12] This implication provides additional economic information to test the validity of proposed predictors. In other words, when evaluating predictors constructed from individual characteristics, we should focus on the subsample of individual characteristics that contain information about individual asset returns. Accordingly, in the rest of the paper, we examine the aggregate information in a set of 140 predictors that have been previously shown to contain information about individual asset returns (McLean and Pontiff, 2016).

III. Data

To examine the relation between cross-sectional anomaly variables and the equity risk premium, we combine daily data from the Center for Research in Security Prices ("CRSP") and Compustat over the period 1926 through 2017.

---

[11] In the market model (Sharpe (1964); Lintner (1965)), the systematic portion of stock returns reflects compensation for bearing systematic risk. However, outside of the market model, a common component of returns could exist that is not related to systematic risk (for example, consumer sentiment). Similar to the systematic portion of returns in the market model, such a variable could be related to the cross-section of returns and, since it has a common component, it could aggregate to contain information about the equity risk premium.

[12] Theoretically, it is possible that investors switch from gathering systematic information to gathering idiosyncratic information depending on economic conditions as in Kacperczyk, Van Nieuwerburgh, and Veldkamp (2016). As a result, some variables could contain idiosyncratic information at times, and systematic information at other times.

We calculate the equity risk premium as the log return on the S&P 500 index minus the log return on a one-month Treasury bill as in Goyal and Welch (2008).[13] We construct aggregate time series variables for each of the 97 cross-sectional anomalies in McLean and Pontiff (2016) and we supplement this dataset with 43 additional variables from the extant literature to arrive at 140 candidate predictors.[14] As we explain in the Introduction, we explore various subgroups within the 140. Table 1 contains summary statistics for the cross-sectional variables we use to form time-series predictors. Of the 140 variables, 15 are classified as *Valuation* predictors and 22 are classified as *Opinion* predictors. *Best Cross-Sectional*. This subcategory consists of the ten predictors with the greatest cross-sectional *t*-statistic, using the sample period in the original paper. By definition, the *Best Cross-Sectional* predictors exhibit strong cross-sectional predictive power; indeed, the mean cross-sectional *t*-statistic for these predictors is approximately 9. Predictors from existing papers consist of the 26 cross-sectional variables that have been used to predict market returns in a published study.

We construct two time-series predictors from each cross-sectional variable based on the equal-weighted average and value-weighted average. Unlike with cross-sectional estimation, consideration of stationarity is crucial in estimation of the market risk premium (for example, Campbell, 1991, and Hodrick, 1992). We test each time-series predictor for a unit root using an Augmented Dickey-Fuller (1979) test. Because some of the resulting time-series variables are non-stationary, we proceed as follows: if we reject the null that the raw (untransformed) variable is non-stationary, we use it in our tests. If not, we calculate the first-difference for each variable; if we reject the null that it is non-stationary then we use it, otherwise the variable is excluded.

Some of our cross-sectional predictors should (theoretically) aggregate to form a variable that is constant across time (e.g., CAPM beta). Accordingly, after constructing time-series versions of each variable, we apply a manual filter to drop variables that should not aggregate to form an economically meaningful variable and we also examine each

---

[13] We download this data from Amit Goyal's website (http://www.hec.unil.ch/agoyal/).
[14] See the Internet Appendix for an overview of the construction of these 140 variables.

14

variable's time-series standard deviation and exclude predictors that exhibit little to no time-series variation.[15] Of our original 280 different predictive variables (140 equal-weighted and 140 value-weighted variables), 253 survive the screening procedures discussed above. We use these 253 variables in our main regressions.

To verify that the Romano and Wolf (2016) adjusted p-values are not sensitive to the set of variables we consider, we also examine three alternate sets of candidate predictors: one that has more variables (269 variables) and two that have fewer (137 and 51 variables, respectively). The first set expands the list of predictors to 269 variables by omitting the manual filter discussed above and adding stochastic detrending as a way to avoid non-stationarity. Specifically, if we reject the null that the raw (untransformed) variable is non-stationary, we use it in our tests. If not, we calculate deviations from a linear trend model for each variable; if we reject the null that it is non-stationary then we use it. If not, we calculate the first-difference for each variable; if we reject the null that it is non-stationary then we use it, otherwise we drop the variable. For the linear trend, we estimate a model of the form:

$$(5) \qquad\qquad x_t = a + bt + u_t \text{ for } t = 1, \ldots, T,$$

for each predictor variable x_t and time period t. We take the fitted residual, $\hat{u}_t$, as our de-trended measure. By construction, $\hat{u}_t$ has a mean of zero and we normalize it to have a standard deviation of one.[16] This set represents the maximum number of candidate variables possible; we use all of the candidate predictors that we can with minimal assumptions (i.e., they must have time-series variation and pass the unit root test).

One potential criticism is that there may be only a few potent predictor variables. In this view of the world, a large sample such as ours might be padded with obviously unlikely candidates, and a multiple hypothesis test would have low statistical power. To address this potential criticism, we consider two smaller sets of predictors. One set contains 137

---

[15] We calculate a measure of variation for each variable as the ratio of its time-series standard deviation to the absolute value of its time-series mean, and we drop variables with a ratio less than 0.06. See the Internet Appendix for a list of variables dropped using the manual filter.

[16] For our in-sample analyses, we estimate the linear trend model using all available data. For our out-of-sample analyses, we estimate the trend model only using data available at each point in time to avoid a look-ahead bias.

predictors and uses only the raw version of each variable (i.e., it does not use first-differencing or stochastic detrending). This selection reflects the notion that authors, referees, and editors avoid time-series variables that require extra manipulation. For this set, we again apply the manual filter to drop predictors that exhibit little to no time-series variation. The other set represents the minimum number of candidate variables possible; we use only the predictors that have *already* been examined in the existing time-series literature. This set contains 51 variables and, by definition, contains variables that the literature views as reasonable candidate variables.

IV. Results

In this section, we examine whether cross-sectional predictors, in general, contain information about the equity risk premium. We start by examining in-sample tests that use the entire sample of data and estimate a single parameter estimate from a time-series regression of the market risk premium on the predictor. We then examine out-of-sample tests that use rolling regressions to test whether a variable is useful for predicting the future equity risk premium, using only information available at each date.

A. In-Sample Tests

As noted in Goyal and Welch (2008), "It is unreasonable to propose a model if the [in-sample] performance is insignificant, regardless of its [out-of-sample] performance." As discussed in Section II, we start with 140 variables from the existing literature and, using these, we form 268 candidate predictor variables. The sample length for each predictive variable depends on data availability. Some predictors have data available as far back as 1926, while other variables have samples that start more recently. In our in-sample tests, the length of the time series varies depending on data availability.

For each variable, we run predictive regression models of the form:

(6) $$r_{t:t+h} = \alpha + \beta x_t + \varepsilon_{t:t+h} \text{ for } t = 1, \ldots, T - h,$$

where $r_{t:t+h} = (1/h)(r_{t+1} + \cdots + r_{t+h})$, r_t is the continuously compounded S&P 500 return for month t from CRSP including dividends and excess of the monthly risk-free rate from

16

Goyal and Welch (2008), x_t is the predictor variable, and h denotes the forecast horizon. We examine four different forecast horizons: one month ahead, one quarter ahead, one-half year ahead, and one year ahead (i.e., $h = 1$, 3, 6, or 12). For each predictor at each forecast horizon, the regression is estimated using all available data, leading to a single parameter estimate (β) that measures the predictive ability of the candidate variable at that horizon.[17]

The results are shown in Figure 1 and Tables 2 and 3. In Figure 1, we provide a visual display of the relation between cross-sectional predictors and their time-series counterpart. For each predictor, the horizontal axis plots the cross-sectional t-statistic, while the vertical axis plots the time-series t-statistic. Panel A displays results for 1-month time-series regressions, while panels B, C, and D display results for 3-month, 6-month, and 12-month regressions.[18] If there is a direct mapping from cross-sectional predictability to time-series predictability, then at a minimum, the signs of the two results should be the same. This would imply that all of the results should either be in quadrant I (the top right of each panel) or quadrant III (the bottom left). While a number of results are in quadrant I and III, they represent about 58% of all observations at the 12-month horizon. In other words, there are a number of results in quadrant II (top left) and IV (bottom right) suggesting that many time-series predictors have the opposite sign of their cross-sectional counterpart. In each panel, we also plot a linear trendline; if time-series predictors have the same t-statistics as their cross-sectional counterpart, the trendline should have a 45-degree slope increasing from the left of the figure to the right. In all panels, the line does slope up from left to right, indicating a relation between time-series and cross-sectional predictors, but the line is flatter than 45 degrees. However, the line does become steeper as the forecasting horizon increases. At the 1-month horizon, the Pearson correlation coefficient between the time-series and cross-sectional predictors is 0.175 while at the 12-month horizon the correlation rises to 0.289.

---

[17] In untabulated results, we estimate equation (6) using the weighted least squares method of Johnson (2019). This estimator does not affect our conclusions.
[18] We only display results for value-weighted time series predictors in Figure 1, however, in unreported results the results are similar for equal-weighted predictors.

To formally examine the time-series information in cross-sectional predictors, we next turn to the regression results. Table 2 provides a summary of the performance of the candidate predictors, broken out by various sub-categories. We report the fraction of predictors that are statistically significant at the 10% level or better using two-sided t-statistics computed using a stationary block bootstrap (Politis and Romano, 1994) with 1,000 draws to account for the Stambaugh (1999) bias and the fact that the model uses overlapping observations when h > 1 (Goetzmann and Jorion (1993); Hodrick (1992); Nelson and Kim (1993)).[19]

Panel A displays the results from the four different sets of predictors: (1) "Predictors from Existing Papers" examines only predictors that have *already* been examined in the time-series literature; (2) "Raw Predictors" expands the list to include *all* predictors that are stationary in raw form (i.e., before applying transformations like the first difference) and it imposes filters to remove variables that should not aggregate and/or do not exhibit time-series variation; (3) "Raw Predictors + First Diff." is similar to (2) except it adds the first-difference transformation (i.e., if we fail to reject the null that the raw variable is non-stationary, we then calculate the first difference and include it if it passes the unit root test); (4) "All Possible Predictors" examines as many predictors as possible (i.e., we only require that they have time-series variation and pass the unit root test). Across all four sets, we find evidence of return predictability. In our main specification ("Raw Predictors + First Diff."), on average 13% are statistically significant at the 10% level or better across the various forecasting horizons. This number generally increases as the forecast horizon increases from 27 (11%) at the one-month horizon to 43 (17%) at the 12-month horizon. The numbers are similar for the other sets of predictors. These results run counter to explanations that the results are hinged on the initial set of variables being over- or under-aggressive.

---

[19] Specifically, we resample the original data using a stationary block bootstrap with a mean block size of 5, however our conclusions are robust to alternate block sizes of 10, 25, and 50. To avoid the overlapping dependent variable issue, we have also tried aggregating the independent variable rather than the dependent variable as suggested by Cochrane (1991) and Jegadeesh (1991); our general conclusions - about the lack of statistically significant predictability after accounting for the number of tests run - do not change under any of these alternate approaches.

In Panel B, we examine three sub-categories of the 253 variables in our main specification. The first two sub-categories, *Valuation* and *Opinion*, are motivated by economic theory. The *Valuation* sub-category is composed of variables that are a function of discount rates, so theory suggests that they should be related to returns in a time-series setting (Kelly and Pruitt, 2013, and Lewellen, 2004). The *Opinion* subcategory of predictors consists of variables like institutional trading and analyst upgrades, which can be motivated with the sentiment explanation of Baker and Wurgler (2006) or with the information explanation of Seyhun (1988). Finally, we examine a third subcategory, *Best Cross-sectional*, defined as the 10 most statistically significant anomalies in the cross-sectional literature. Several papers find a reduction of cross-sectional return predictability in recent periods (e.g., McLean and Pontiff (2016), Green, Hand, and Zhang (2017)). If the cross-sectional predictors we start with are only weakly related to returns in the cross-section, then they may lead to weak performance in our time-series tests. Accordingly, we examine a subcategory that focuses on the best performers in the cross-sectional literature.

When we examine the three sub-categories defined above, the results are similar. Interestingly, the results are generally weaker for the *Valuation* and *Opinion* categories than for the entire set of predictors. Since valuation ratios may be a function of discount rates, they are arguably the most likely cross-sectional predictors to work in a time-series setting (Kelly and Pruitt, 2013). Indeed, several different *Valuation* predictors have been studied in the existing market risk premium literature, notably the dividend-to-price and earnings-to-price ratios. Our results suggest these predictors are an exception, rather than the norm. Across all horizons, only 8% of *Valuation* predictors and 8% of *Opinion* predictors are statistically significant, versus 13% across all predictors.

When we examine the results for the "best" cross-sectional variables, there is some evidence of predictability, but again, the results are weaker than when we examine all possible predictors. For the *Best Cross-Sectional* predictors (those with the top ten highest cross-sectional *t*-statistics), 10% are statistically significant at the one-year horizon. The results are generally consistent with the idea that the best cross-sectional variables forecast returns because they contain information about the systematic component of returns. As a

19

result, these variables also aggregate to form good time-series predictors. However, the fact that the subset of *Best Cross-Sectional* predictors is not stronger than the entire set suggests that good cross-sectional predictors are not necessarily good time-series predictors.

Finally, we examine the in-sample results broken out by the different methodologies used to construct the aggregate predictor variables. Specifically, in Panel C we present results for value- or equal-weighted average predictors. The findings are largely consistent across the two methodologies: across all predictors, 13% are statistically significant at the one-month horizon when value-weighted and 8% are statistically significant at the one-month horizon when equal-weighted. As the horizon extends, the value-weighted predictors appear to perform slightly better, but the results are generally similar across the two groups. For example, across all predictors, 20% are statistically significant at the annual horizon when value-weighted versus 14% when equal-weighted. On the surface, before a deeper consideration of multiple hypothesis testing is considered, Table 2 suggests that some cross-sectional anomaly variables predict the market risk premium.

B. Multiple Hypothesis Testing

The results in Table 2 examine as many as 269 different regression models, so the results are subject to concerns about data snooping. Put differently, with 269 different regression models, some tests will likely be statistically significant due to type I errors. Fortunately, a growing literature shows how to adjust p-values to account for the number of models considered.

White (2000) develops a reality check bootstrap (RCB) to correct for data-snooping. While a number of approaches exist to adjust p-values for the bias that results from multiple hypothesis testing, the White approach has several desirable properties. First, it uses a bootstrap procedure to estimate the dependence structure of the p-values across all considered models. In contrast, the Bonferroni, Dunn (1961) and Holm (1979) approaches assume the worst-case dependence structure. This causes them to be overly conservative in that they do not reject the null hypothesis enough when the null is false. Because it

estimates the actual dependence structure, the White RCB has greater power than the Bonferroni and Holm methods. Second, the White procedure is particularly suited to the application of return predictability regressions because the procedure uses a loss function that compares the performance of a predictor to a benchmark model. Economic theory suggests the equity risk premium should be positive; as a result, a strategy that simply predicts positive returns all the time would frequently be correct. Accordingly, the return predictability literature often compares the forecast accuracy of a predictive variable to the so-called "prevailing mean" model that uses the prevailing mean return as the forecast of next period's equity risk premium. The White RCB then uses a loss function that compares the mean squared forecast error for each predictor to the mean squared forecast error from a benchmark model that uses the prevailing mean return.

Despite these advantages, the White procedure does have a drawback: if the null is rejected, it indicates that the *best* predictor examined is better than the benchmark, but it does not reveal whether the other predictors are better. Put differently, the White procedure does not test whether the second-best predictor, or the kth best predictor, is better than the benchmark. Accordingly, Romano and Wolf (2005, 2016) develop a step-down procedure that extends the White procedure to calculate whether individual predictors are better than the benchmark.[20] The resulting procedure controls the familywise error rate and provides p-values for each individual predictor.

This is the first paper to use this procedure to evaluate the predictability of the market risk premium. The closest related papers are Sullivan, Timmerman, and White (1999) and Chordia, Goyal, and Saretto (2020). Sullivan, Timmerman, and White (1999) apply the White procedure to examine the performance of technical trading strategies in predicting the equity risk premium. Sullivan et al. find no evidence that technical trading strategies generate portfolio performance that out-performs a benchmark. More recently, Chordia, Goyal, and Saretto (2020) examine the performance of trading strategies in the cross-

---

[20] See Romano and Wolf (2016) for a detailed discussion of the procedure and see Section C of the Internet Appendix for a detailed overview of our implementation of the procedure. We thank Allan Timmermann for helpful conversations about the White (2000) and Romano and Wolf (2005) procedures.

section and they use several methods to correct for multiple hypothesis testing bias, including the Romano and Wolf (2005, 2016) procedure. They find that most strategies studied in the extant literature are not significant after adjusting for multiple hypothesis testing. Our paper unifies these two literatures by providing the first evidence on the performance of time-series predictors formed from cross-sectional variables.

To further explore the robustness of our findings, in the Internet Appendix we also examine p-values calculated using the Benjamini and Yekutieli (2001) procedure, which controls the false discovery rate while allowing for arbitrary dependence. On average, false discovery rate methods have better power to reject false null hypotheses, but this comes at a cost: they are more likely to reject true null hypotheses. In other words, the Benjamini and Yekutieli (2001) procedure is less conservative than the Romano and Wolf (2016) procedure as it allows for more false positives.[21] Nonetheless, in all analyses our main conclusions are unchanged when we use the Benjamini and Yekutieli (2001) procedure in place of the Romano and Wolf (2016) procedure.

B.2 In-Sample Results Adjusted for Multiple Hypothesis Testing

Table 3 reports detailed estimates for individual predictors that are statistically significant before adjusting for multiple testing. For brevity, we present results for only our main specification, which examines 253 candidate predictors, however the results are similar for other specifications. The table presents coefficient estimates, R-squared values, and both raw and Romano and Wolf (2016) adjusted p-values. Here we reach the main conclusion of the paper: most of the cross-sectional variables that appear to be statistically significant when examined in isolation are no longer statistically significant when examined in the context of all cross-sectional predictors.

For example, at the 1-month and 3-month horizons, none of the predictors remain significant at the 10% level when computing Romano and Wolf (2016) adjusted p-values compared to 27 when computing individual p-values (Table 2). At the 6-month horizon,

---

[21] Specifically, the Romano and Wolf (2016) procedure controls the probability of having *any* false positives while the Benjamini and Yekutieli (2001) procedure controls the *expected proportion* of false positives, so it allows for more false positives as you consider more predictors.

there are 2 predictors that are significant at the 10% level (compared to 36 in Table 2) and at the 12-month horizon there are 2 predictors significant at the 10% level (compared to 43 in Table 2). Moreover, the predictors with remarkable statistical significance when examined in isolation, no longer appear to be so stellar when examined among the set of 253 predictors.

In terms of economic significance, a number of predictors are noteworthy. *Z-score* and *Asset Turnover*, with R-squared values of 4.6% and of 4.7% shown in column 5 of Panel B, are among the best predictors at the three-month horizon. Moreover, several other predictors have R-squared values exceeding 3% including *Percent Operating Accrual* and *Target Price.* Although these R-squared values might seem small in absolute magnitude, Zhou (2010) notes that R-squared values from predictive regressions are typically small in absolute magnitude as stock returns are difficult to forecast. Accordingly, Zhou (2014) builds on the insights of Ross (2005) to construct a mathematical bound on the maximum R-squared that can exist under no arbitrage conditions. Using consumption growth rates as a state variable, Zhou's bounds imply a maximum R-squared at the monthly horizon of between 0.079% and 0.177%[22] and Huang and Zhou (2017) show that the quarterly R-squared is bound by at most 3.74%, depending on the specification, and in most cases it is less than 3%. In light of this, the return predictability of some Table 3 predictors is economically large.[23]

In addition, we note that many of the best predictors at the 3-month horizon are also good predictors at the six-month and twelve-month horizons. At the twelve-month horizon, a number of the predictors have impressive R-squared values of 10% or greater including

---

[22] The bounds developed in Zhou (2010) depend on the choice of a state variable. We do not take a stance on state variables in this paper, we simply note that the R-squared values we document appear economically meaningful relative to the example bounds presented in Zhou (2010).

[23] We also construct predictors based on sets of cross-sectional predictors using principal component analysis (PCA). PCA requires non-missing observations for each predictor; as a result, we are not able to utilize the entire set of predictors until 1999. In untabulated results, the first principal component extracted from all equal-weighted predictors and value-weighted predictors exhibits strong return predictability over the period 1999 to 2017, before adjusting for multiple testing. However, if we estimate the PCA using the set of variables with data starting in 1980, the evidence becomes mixed; we find weak evidence of return predictability for value-weighted predictors, but no evidence of return predictability for equal-weighted predictors. We thank Bryan Kelly for this suggestion.

23

Z-score, *Asset Turnover*, and *Percent Operating Accrual*. For example, *Z-score* exhibits a remarkable R-squared of 18.5% at the twelve-month horizon. However, despite some impressive results for individual predictors, after applying the Romano and Wolf (2005, 2016) procedure only *Z-score* and *Asset Turnover* are statistically significant.

Table 4 summarizes the results after applying the Romano and Wolf (2005, 2016) procedure. In Panel A, our main specification ("Raw Predictors + First Diff.") exhibits 0, 0, 2, and 2 statistically significant predictors at the 1- month, 3- month, 6- month, and 12-month horizons. One possible critique is that these results include predictors that should not have been included, and this causes a failure to reject a false null hypothesis (a type II error). To address this concern, we also present results for a variety of different sets of predictors and subsamples. In our main specification we use only the raw or first-differenced version of each variable, yet some existing papers in the time-series return predictability literature examine stochastically detrended variables. Accordingly, in Panel A we examine "All Possible Predictors" which add predictors that require stochastic detrending. Consistent with Rapach et al. (2016) we find that detrended short interest is one of the best predictors at the 3-month, 6-month, and 12-month horizons.[24] However, after applying the Romano and Wolf procedure, we find that none of the 269 predictors are statistically significant at the 1-month and 3-month horizons and only 3 of the variables are statistically significant at the 12-month horizon.

Similarly, we examine two other sets of predictors ("Raw Predictors" and "Predictors from Existing Papers") and find little evidence of return predictability. In sum, the three alternate sets of predictors in Panel A provide bounds on the possible set of variables to consider. At a minimum, we know the profession has examined the 51 variables in the "Predictors from Existing Papers" so these variables have to be included in a multiple testing framework. At a maximum, the 269 predictors in "All Possible Predictors" represent all the cross-sectional variables that could possibly be examined. The fact that our results are unchanged across these two extremes shows that our conclusions

---

[24] Detailed in-sample results for the set of "All Possible Predictors" are shown in Table IA.III of the Internet Appendix and out-of-sample results are shown in Table IA.IV of the Internet Appendix.

are not sensitive to the number of variables considered. Moreover, when we examine the results by sub-category in Panel B, the results again show zero statistically significant predictors at the 1-month horizon and between 0 and 1 statistically significant predictors at the 12-month horizon.

Finally, Table IA.I summarizes the results after applying the Benjamini and Yekutieli (2001) procedure. As expected, there are more statistically significant predictors in Table IA.I than Table 4, but the overall conclusion is similar. Even though the Benjamini and Yekutieli (2001) procedure is more permissive than the Romano and Wolf (2016) procedure, there are relatively few predictors that remain significant. For example, when we examine "Predictors from Existing Papers," we find no evidence of predictability at the 1-month and 3-month horizons, and only 3 of the 60 predictors are significant at the 6-month horizon and 4 of the 60 predictors are significant at the 12-month horizon. Overall, across a wide variety of samples and methodologies, our conclusion remains unchanged: there is only weak in-sample evidence that cross-sectional predictors contain information about the systematic portion of returns.

C. Out-of-Sample Tests

A number of papers note that in-sample tests may overstate predictability due to the use of information that was not known ex-ante (e.g., Cooper, Gutierrez, Marcum (2005), Goyal and Welch (2008)). Accordingly, in this section, we revisit each of our tests using out-of-sample forecasting regressions. We again run predictive regressions of the form:

(7) $$r_{t:t+h} = \alpha_t + \beta_t x_t + \varepsilon_{t:t+h} \text{ for } t = 1, \ldots, T - h,$$

where $r_{t:t+h} = (1/h)(r_{t+1} + \cdots + r_{t+h})$, r_t is the excess return on the S&P500, and x_t is the predictor variable. However, now we estimate the model separately for each time period, using only information that was available at each date. As such, we estimate new parameter estimates for α_t and β_t at each point in time. If the relation between the predictor variable and the equity risk premium is stable over time, then this out-of-sample approach should produce the same results as the in-sample analysis discussed in Section III.A. If the relation

25

between the predictor variable and the equity risk premium is not stable, the out-of-sample tests may lead to a different conclusion.

As previously discussed, the sample length for each predictive variable depends on data availability. For the out-of-sample tests, we use the first 10 years of data to train the model before we make our first forecast. As before, we start by summarizing the results across all specifications. Table 5 provides a summary of the performance of the candidate predictors. To make inferences, we calculate the out-of-sample R_{OS}^2 statistic defined as in Campbell and Thompson (2008).[25] To calculate the out-of-sample R_{OS}^2, we use the prevailing mean equity risk premium, at each date, as our benchmark model and we use the Clark and West (2007) statistic to assess statistical significance.[26] Panel A summarizes statistical significance across the four different sets of predictors: (1) "Predictors from Existing Papers"; (2) "Raw Predictors"; (3) "Raw Predictors + First Diff."; and (4) "All Possible Predictors." Recall that in Table 2, the in-sample evidence was strongest at the annual horizons, however even at the 1-month horizon approximately 11% of the "Raw Predictors + First Diff." predictors were statistically significant at the 10% level or better. In contrast, the out-of-sample evidence is weaker. At the one-month horizon, less than 3% of the variables are statistically significant.

In Panel B, we examine three sub-categories of the 253 variables in our main specification (*Best Cross-sectional, Opinion,* and *Valuation*) and the results do not look much better. Again, the *Valuation* and *Opinion* sub-categories appear to be worse than the full sample of predictors, despite the economic motivation for *Valuation* predictors and their popularity in the extant literature. Moreover, while the *Best Cross-Sectional* category continues to show some evidence of predictability at longer horizons, none of the predictors is significant at the one-month horizon.

---

[25] We use the unconstrained out-of-sample R-squared from Campbell and Thompson (2008) (i.e., we do not impose any sign restrictions).

[26] Formally, we test the null hypothesis that the mean square forecast error (MSFE) from the baseline model is less than or equal to the MSFE from the predictive model versus the alternative hypothesis that the MSFE from the benchmark model is greater than the MSFE from the predictive model ($H_0: R_{OS}^2 \leq 0$ against ($H_A: R_{OS}^2 > 0$).

The remaining panels of Table 5 examine the out-of-sample results broken out by the different methodologies used to construct the aggregate predictor variables. Specifically, Panel C examines the results when we calculate the aggregate predictor using a value-weighted average or an equal-weighted average, respectively. Again, the results look similar regardless of the methodology used to construct the predictors. Only 4 of 128 value-weighted predictors are significant at the one-month horizon and only 3 of 125 equal-weighted predictors are significant. Overall, the results show that cross-sectional variables contain some systematic information at longer horizons, but not at short horizons. Taken together, it is tempting to conclude that some cross-sectional predictors can be used to form aggregate predictors, suggesting they contain information about the systematic component of returns. However, our out-of-sample analyses consider more than 253 different specifications. In the next section, we revisit our results after accounting for the number of hypotheses tested.

D.1. Multiple Hypothesis Testing

As before, we ask whether our out-of-sample tests show evidence of return predictability after accounting for possible data snooping biases. To do this, we again use the Romano and Wolf (2016) procedure.[27] The procedure follows a similar process to the in-sample procedure, except we estimate rolling regressions and compare the prediction from these regressions to a rolling average market risk premium. Section B of the Internet Appendix provides a detailed overview of the procedure.

D.2 Out-of-Sample Results Adjusted for Multiple Hypothesis Testing

Table 6 displays detailed estimates for individual predictors that are statistically significant before adjusting for multiple testing. The table presents the time-series mean of each coefficient estimate, the Campbell and Thompson (2008) out-of-sample R-squared,

---

[27] The Romano and Wolf procedure expands the White (2000) RCB to adjust p-values for each individual predictor. White (2000) shows this procedure is valid for both in-sample and out-of-sample tests and Sullivan, Timmermann, and White (1999) apply the White (2000) procedure to out-of-sample forecasting regressions.

and both raw and Romano and Wolf (2016) adjusted p-values. Interestingly, the out-of-sample R-squared values in Table 6 highlight many of the same predictors that performed well in Table 3 in the in-sample analyses. Analogous to Table 3, we report the time-series mean of the betas from these regressions in addition to the out-of-sample R^2_{OS} statistic. Indeed, *Z-score* and *Asset Turnover* have out-of-sample R-squared values that are positive and economically meaningful at the three-month, six-month, and twelve-month horizons. Once again, *Asset Turnover* exhibits impressive results with an out-of-sample R-squared value of 17.5% at the twelve-month horizon. Because the out-of-sample R^2_{OS} statistic measures the proportional reduction in mean squared forecast error that results from using the predictor (relative to the benchmark model), its magnitude is also useful for interpreting the economic significance of these findings. For the variables listed above, the out-of-sample R^2_{OS} statistic suggests economically large return predictability.

However, many of these conclusions change after adjusting for multiple testing. Table 7 summarizes the results. We fail to reject the null of no predictably at all forecasting horizons for all predictors using the Romano and Wolf (2016) stepdown procedure. In Panel A the results are consistent across all of the alternate samples (*Predictors from Existing Papers, Raw Predictors, Raw Predictors + First Diff.,* and *All Possible Predictors*). As a result, our conclusions are not sensitive to the number of variables considered in the Romano and Wolf calculation. In Panels B and C, we again find no predictability. The conclusions are not significantly different when we examine false discovery rates, shown in Table IA.II of the internet appendix. At the 1-month and 3-month horizons, none of the predictors are significant in any of the samples. At longer horizons, we find limited evidence of predictability. For example, at the 12-month horizon, we find that 2 out of the predictors from the existing literature remain significant. Overall, the results in this subsection echo the conclusions of subsection B. Many predictors exhibit strong out-of-sample *t*-statistics. However, once multiple hypothesis testing is considered, the results are weaker and we find little evidence that cross-sectional predictors, in general, contain systematic information.

28

V. Conclusion

There is a large literature examining the cross-sectional determinants of stock returns. Similarly, many time series variables have been proposed as predictors of the equity risk premium. While these literatures have evolved largely independently, at least 26 papers have proposed certain cross-sectional variables as candidates for time-series predictability. Using various samples of cross-sectional predictors and accounting for the number of predictors and their interdependence, we examine the link between cross-sectional and time-series predictability. Our analyses provide new information on the nature of return predictability.

We find plenty of evidence that, in isolation, certain cross-sectional variables make great time series predictors. Some of these variables, like *Z-Score* and *Asset Turnover*, have never been proposed as time-series variables. However, these results largely disappear once we account for the data snooping bias arising from the plethora of predictive variables considered. Moreover, when we examine out-of-sample forecasting regressions, we continue to find little evidence of return predictability.

If each of our 140 cross-sectional predictors were examined by different econometricians, it is likely that several articles would be written discussing the powerful in-sample time-series information in cross-sectional variables. Claims of predictability in these articles would likely be bolstered by out-of-sample Goyal and Welch (2008) tests. Indeed, several such articles exist. In this paper, we take a different approach. Once we consider the set of all existing cross-sectional variables documented in the extant literature, the difference in conclusions is stark. The evidence no longer suggests that cross-sectional variables contain information about the systematic component of returns.

29

References

Baker, Malcolm and Jeffrey Wurgler. "The Equity Share in New Issues and Aggregate Stock Returns." *The Journal of Finance,* 55, (2000), 2219–2257.

Bakshi, Gurdip, George Panayotov, and Georgios Skoulakis. "Improving the predictability of real economic activity and asset returns with forward variances inferred from option portfolios." *Journal of Financial Economics,* 100, (2011), 475–495.

Bartsch, Viktoria-Sophie, Hubert Dichtl, Wolfgang Drobetz, and Andreas Neuhierl. "Data Snooping in Equity Premium Prediction." *Journal of International Forecasting*, 37, (2021), 72-94..

Bay, Raymond. "Anomalies in relationships between securities' yields and yield-surrogates." *Journal of Financial Economics,* 6, (1987), 103–126.

Benjamini, Yoav and Daniel Yekutieli. "The control of the false discovery rate in multiple testing under dependency." *Annals of Statistics,* 29, (2001), 1165–1188.

Berk, Jonathan B. "A Critique of Size-Related Anomalies." *Review of Financial Studies,* 8, (1995), 275–286.

Bossaerts, Peter, and Pierre Hillion. "Implementing Statistical Criteria to Select Return Forecasting Models: What Do We Learn?" *Review of Financial Studies,* 12, (1999), 405–428.

Campbell, John. "A Variance Decomposition for Stock Returns." *Economic Journal,* 101, (1999), 157–179.

Campbell, John Y. and Robert J. Shiller. "The Dividend-Price Ratio and Expectations of Future Dividends and Discount Factors" *Review of Financial Studies*, 1, (1988), 195–228.

Campbell, John, and Samuel Thompson. "Predicting excess stock returns out of sample: can anything beat the historical average?" *Review of Financial Studies,* 21,(2008), 1509–1531.

Chen, Andrew Y. and Todd Zimmerman. "Publication Bias and the Cross-Section of Stock Returns." *Review of Asset Pricing Studies,* 10, (2020), 249-289.

Chen, Andrew Y. "The Limits of p-Hacking: Some Thought Experiments." *Journal of Finance*, 76, (2020), 2447-2480.

Chordia, Tarun, Richard Roll and Avanidhar Subrahmanyam. "Order Imbalance, Liquidity and Market Returns." *Journal of Financial Economics,* 65, (2002) 111–130.

Chordia, Tarun, Amit Goyal and Alessio Saretto. "Anomalies and False Rejections", *Review of Financial Studies*, 33, (2020), 2134–2179.

Clark, T.E., West, K.D., "Approximately normal tests for equal predictive accuracy in nested models." *Journal of Econometrics,* 138, (2007), 291–311.

Cochrane, John H. "Volatility Tests and Efficient Markets: A Review Essay." *Journal of Monetary Economics,* 27, (1991), 463–485.

Cochrane, John H. "The Dog That Did Not Bark: A Defense of Return Predictability, *Review of Financial Studies* 21, (2008), 1533–1575.

Cooper, Michael, and Hyseyin Gulen, "Is Time-Series-Based Predictability Evident in Real Time?" *Journal of Business,* 79, (2008), 1263–1292.

Cowles, Alfred. "Can Stock Market Forecasters Forecast?" *Econometrica*, 1, (1933), 309–324.

Dickey, D.A., and Wayne Fuller. "Distribution of the estimators for autoregressive time series with a unit root." *Journal of the American Statistical Association,* 74, (1979) 427–431.

Dow, Charles. "Scientific Stock Speculation." *The Magazine of Wall Street*, (1920).

Dunn, Olive Jean. "Multiple comparisons among means." *Journal of the American Statistical Association*, 56, (1961), 52–64.

Engelberg, Joseph, R. David McLean, and Jeffrey Pontiff. "Anomalies and News." *Journal of Finance,* 73, (2018), 1971-2001.

Fama, Eugene. "Market efficiency, long-term returns and behavioral finance." *Journal of Financial Economics*, 49, (1998), 56–63.

Foster, F. Douglas, Tom Smith, and Robert E. Whaley. "Assessing Goodness-of-Fit of Asset Pricing Models: The Distribution of the Maximal R^2." *Journal of Finance*, 52, (1997), 591–607.

Gibson, Thomas. "The Pitfalls of Speculation". (1906), The Moody Corporation, New York.

Green, J., J. R. Hand and X. F. Zhang. "The Characteristics that Provide Independent Information about Average U.S. Monthly Stock Returns." *Review of Financial Studies*, 30, (2017), 4389-4436.

Hirshleifer, David, Kewei Hou, and Siew Hong Teoh. "Accruals, Cash Flows, and Aggregate Stock Returns," *Journal of Financial Economics,* 91, (2009), 389-406.

Goetzmann, William, Jorion, Philippe. "Testing the predictive power of dividend yields." *Journal of Finance*, 48, (1993), 663–679.

Goyal, Amit, and Pedro Santa-Clara. "Idiosyncratic Risk Matters!" *Journal of Finance,* 58, (2003), 975–1007.

Goyal, Amit, and Ivo Welch. "A Comprehensive Look at the Empirical Performance of Equity Premium Prediction." *Review of Financial Studies,* 21, (2008), 1455–1508.

Goyal, Amit, and Ivo Welch, "Anomalies and False Rejections." *Review of Financial Studies,* 33, (2020), 2134-2179.

Hansen, Peter Reinhard. "A Test for Superior Predictive Ability." *Journal of Business & Economic Statistics,* 23, (2005), 365–380.

Harvey, Campbell R., Yan Liu, and Heqing Zhu. "…and the Cross-Section of Expected Returns." *Review of Financial Studies,* 29, (2016), 5–68.

Harvey, Campbell R., and Yan Liu, "Lucky Factors." *Journal of Financial Economics,* 141, (2021) 413-435.

Harvey, Campbell R., and Yan Liu. "False (and Missed) Discoveries in Financial Economics." *Journal of Finance* 75, (2020) 2503-2553.

Harvey, Campbell R.,Yan Liu, and Alessio Saretto. "An Evaluation of Alternative Multiple Testing Methods for Finance Applications." *Review of Asset Pricing Studies*, (2020) 10, 199–248.

Hirshleifer, David, Kewei Hou, and Siew Hong Teoh. "Accruals, cash flows, and aggregate stock returns." *Journal of Financial Economics,* 91, (2009), 389–406.

Hodrick, Robert. "Dividend yields and expected stock returns: alternative procedures for inference and measurement." *Review of Financial Studies,* 5, (1992), 357–386.

Hou, Kewei, Chen Xue, and Lu Zhang. "Digesting anomalies: An investment approach." *Review of Financial Studies,* 28, (2014) 650-705.

Huang, Dashan and Guofu Zhou. "Upper Bounds on Return Predictability." *Journal of Financial and Quantitative Analysis,* 52, (2017), 401–425.

Jegadeesh, Narasimhan "Seasonality in Stock Price Mean Reversion: Evidence from the U.S. and the U.K." *Journal of Finance* 46, (1991) 1427–44.

Johnson, Travis. "A fresh look at return predictability using a more efficient estimator." *Review of Asset Pricing Studies,* 9, (2019), 1–46.

Jondeau, Eric, Qunzi Zhang and Xiaoneng Zhu. "Average skewness matters." *Journal of Financial Economics,*134, (2019) 29–47.

Kacperczyk, M. S. Van Nieuwerburgh, and L. Veldkamp, "A rational theory of mutual funds' attention allocation." *Econometrica* 84, (2016), 571–626.

Lewellen, Johnathan, "Predicting returns with financial ratios." *Journal of Financial Economics* 74, (2004), 209–235.

Linnainmaa, Juhani and Michael Roberts. "The History of the Cross-Section of Stock Returns." *The Review of Financial Studies* 31, (2018), 2606–2649.

Lintner, John. "The Valuation of Risky Assets and the Selection of Risky Investments in Stock Portfolios and Capital Budgets." *The Review of Economics and Statistics*, 47, (1965), 13–37.

Lo, Andrew W., and A. Craig MacKinlay. "Data-Snooping Bias in Tests of Financial Asset Pricing Models." *Review of Financial Studies,* 3, (1990), 431–467.

McLean, R. David, and Jeffrey Pontiff. "Does Academic Publication Destroy Stock Return Predictability?" *Journal of Finance,* 71, (2016), 5–32.

Neely, Christopher, David Rapach, Jun Tu, and Goufu Zhou. "Forecasting the Equity Risk Premium: The Role of Technical Indicators." *Management Science,* 60, (2014), 1772–1791.

Nelson, Charles and Myung Kim "Predictable stock returns: the role of small sample bias." *Journal of Finance* 48, (1993), 641–661.

Politis, Dimitirs and Joseph Romano. "The Stationary Bootstrap." *Journal of the American Statistical Association,* 89, (1994), 1303–1313.

Pontiff, Jeffrey and Lawrence Schall. "Book-to-market as a predictor of market returns." *Journal of Financial Economics,* 49, (1998), 141–160.

33

Rapach, David E., Matthew C. Ringgenberg and Guofu Zhou, "Aggregate Short Interest and Return Predictability" *Journal of Financial Economics,* 121 (2016), 46–65.

Romano, Joseph P. and Michael Wolf. "Stepwise multiple testing as formalized data snooping." *Econometrica,* 73, (2005), 1237–1282.

Romano, Joseph and Azzem Shaikh and Michael Wolf. "Formalized Data Snooping Based on Generalized Error Rates." *Econometric Theory* 24, (2008), 404–447.

Romano, Joseph and Michael Wolf. "Efficient computation of adjusted p-values for resampling-based stepdown multiple testing." *Statistics and Probability Letters* 113, (2016), 38–40.

Ross, Stephen. Neoclassical Finance. Princeton University Press, (2005).

Schwert, G. William. "Anomalies and market efficiency." In: Constantinides, G.M., Harris, M., Stultz, R. (Eds.), Handbook of the Economics of Finance, 1B. Elsevier, Amsterdam, (2003), pp. 939–974.

Seyhun, H. Nejat, "The Information Content of Aggregate Insider Trading." *The Journal of Business*, 61, (1988), 1–24.

Sharpe, William. "Capital Asset Prices: A theory of market equilibrium under conditions of risk" *Journal of Finance,* 19, 425–442.

Stambaugh, Robert. "Predictive regressions." *Journal of Financial Economic*s, 54, (1999), 375–421.

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

Welch, Ivo. "Reproducing, Extending, Updating, Replicating, Reexamining, and Reconciling." *Critical Finance Review*, 8, (2019), 301-304.

Wen, Quan. "Asset Growth and Stock Market Returns: A Time-Series Analysis" *Review of Finance,* 23, (2019), 599–628.

White, Halbert. "A Reality Check for Data Snooping." *Econometrica,* 68, (2000), 1097–1126.

Zhou, Goufu. "How much stock return predictability can we expect from an asset pricing model?" *Economic Letters,* 108, (2010), 184–186.

Figure 1
In-Sample Relation between Cross-Sectional and Time-Series Predictors

The figure displays a plot of the relation between cross-sectional and time-series predictors. For each predictor, we plot the *t*-statistic from a cross-sectional regression (shown on the horizontal axis) and the *t*-statistic from value-weighted in-sample time series regressions (shown on the vertical axis). Panel A examines 1-month time-series regressions, while panels B, C, and D examine 3-month, 6-month, and 12-month regressions, respectively. In each panel, the diagonal black line denotes a linear trendline.

35
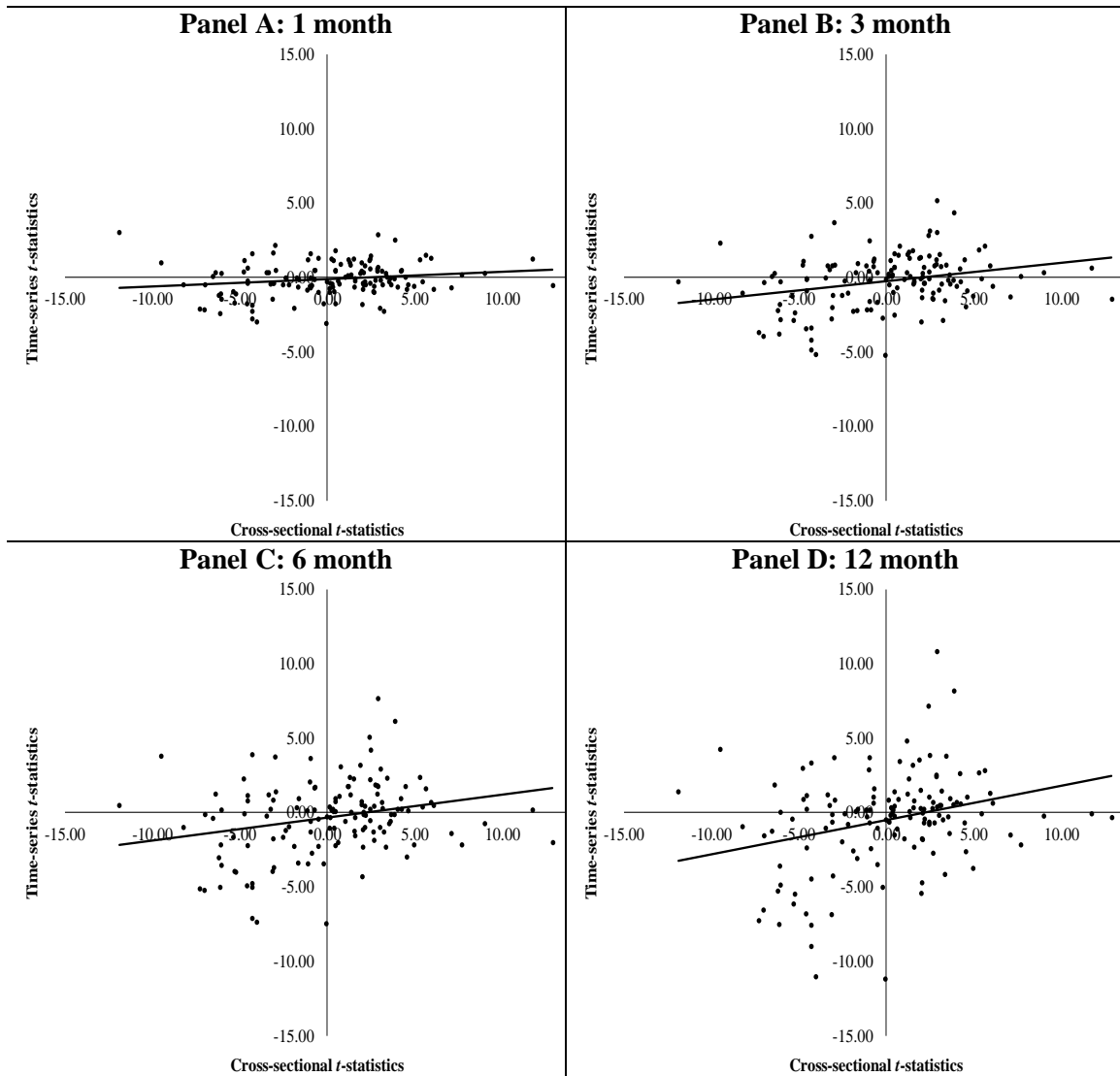
Table 1
Summary Statistics

The table displays summary statistics across all of the cross-sectional variables from which we construct time-series predictors. To construct time-series predictors out of cross-sectional variables, we calculate the value-weighted and equal-weighted mean across all stocks on each date. The first row displays statistics for variables already examined in the existing literature on time-series return predictability (Predictors from Existing Papers), the second row shows statistics for all possible variables derived from the cross-sectional literature (*All Possible*), and the remaining rows examine sub-samples formed on the ten most statistically significant cross-sectional predictors (*Best Cross-Sectional*) and two different groupings from the categories in McLean and Pontiff (2016): (1) *Opinion* and (2) *Valuation*. We display the mean and median values of cross-sectional t-statics as well as the number of citations for each predictor from Google Scholar as of 2018.

Type	N	Cross-Sectional *t*-statistic		Citations	
		Mean	**Median**	**Mean**	**Median**
Predictors from Existing Papers	26	4.26	3.26	3,312	1,531
All Possible	140	3.03	2.71	1,432	517
Best Cross-sectional	10	8.38	7.96	854	740
Opinion	22	2.97	2.48	798	689
Valuation	15	3.14	2.88	2262	273

36

Table 2

Summary of In-Sample Performance using Unadjusted p-values

The table displays a count of the number of predictive variables that are statistically significant at the 10% level or better, as a fraction of the total number of variables examined. We calculate statistical significance using bootstrap p-values. For each anomaly, we estimate an in-sample predictive regression of the form:

$$r_{t:t+h} = \alpha + \beta x_t + \varepsilon_{t:t+h} \text{ for } t = 1, \ldots, T - h,$$

where $r_{t:t+h} = (1/h)(r_{t+1} + \cdots + r_{t+h})$, r_t is the continuously compounded S&P 500 return for month t from CRSP including dividends and excess of the monthly risk-free rate from Goyal and Welch (2008), h indicates the forecast horizon in months, and x_t is one of the 140 predictor variables. To construct time-series predictors out of cross-sectional predictors, we calculate the value-weighted and equal-weighted mean across all stocks on each date resulting in 280 possible predictors. In Panel A, we consider four different definitions: (1) *Predictors from Existing Papers* uses only those variables that are used in the existing literature on time-series return predictability. (2) *Raw Predictors* examines every possible variable for which we reject the null that the raw variable is non-stationary. (3) *Raw Predictors + First Diff.* examines every possible variable however if a variable is not stationary in raw form, we then examine whether it is non-stationary in first-differenced form. If we fail to reject the null that the first differenced variable is non-stationary, we drop the variable. (4) All Possible Predictors examines every possible variable. If we fail to reject the null that a variable is non-stationary, we calculate deviations from a linear trend model. If we fail to reject the null that the linearly detrended variable is non-stationary, we calculate the first-difference. If we fail to reject the null that the first differenced variable is non-stationary, we drop the variable. In Panel B, we examine subsamples of the variables in *Raw Predictors + First Diff.* formed on the ten most statistically significant cross-sectional predictors (*Best Cross-Sectional*), and two different groupings based on the categories in McLean and Pontiff (2016): (1) *Opinion* and (2) *Valuation*. In Panel C we examine equal-weighted vs. value-weighted predictors for the variables in *Raw Predictors + First Diff.*

| | **Return Horizon (h)** | | | |
Predictive Variable	**h=1**	**h=3**	**h=6**	**h=12**
Panel A: Candidate Predictors (number significant / total examined)				
Predictors from Existing Papers	6/51	7/51	6/51	10/51
Raw Predictors	19/137	21/137	30/137	37/137
Raw Predictors + First Diff.	27/253	27/253	36/253	43/253
All Possible Predictors	34/269	43/269	45/269	54/269
Panel B: By Subcategory (number significant / total examined)				
Best Cross-sectional	1/20	1/20	1/20	2/20
Opinion	4/38	3/38	2/38	3/38
Valuation	0/24	0/24	2/24	6/24
Panel C: By Aggregation Method (number significant / total examined)				
Equal-Weighted Predictors	10/125	8/125	14/125	17/125
Value-Weighted Predictors	17/128	19/128	22/128	26/128

Table 3
Best In-Sample Predictive Regression Results using Romano and Wolf p-values

The table reports the ordinary least squares estimate of β, p-values, and the R^2 statistic from in-sample predictive regression models of the form:

$$r_{t:t+h} = \alpha + \beta x_t + \varepsilon_{t:t+h} \text{ for } t = 1, \ldots, T-h,$$

where $r_{t:t+h} = (1/h)(r_{t+1} + \cdots + r_{t+h})$, r_t is the continuously compounded S&P 500 return for month t from CRSP including dividends and excess of the monthly risk-free rate from Goyal and Welch (2008), h indicates the forecast horizon in months, and x_t is the predictor variable shown in columns 2 and 9. For each horizon, we run 253 regressions using the variables in the *Raw Predictors + First Diff.* set of predictors. Panel A displays results for the 1-month horizon, Panel B shows the 3-month horizon, Panel C shows the 6-month horizon, and Panel D shows the 12-month horizon. Within each panel, predictors are sorted by their Romano and Wolf p-value and then their unadjusted p-value. We report all predictors that have unadjusted p-values less than 10% for a given horizon. Unadjusted p-values are shown in columns 6 and 13 and Romano and Wolf (2016) adjusted p-values are shown in columns 7 and 14.

1	2	3	4	5	6	7	8	9	10	11	12	13	14
		EW or			Raw	RW			EW or			Raw	RW
Rank	Predictor	VW	$\widehat{\beta}$	R^2	p-Value		Rank	Predictor	VW	$\widehat{\beta}$	R^2	p-Value	
Panel A: 1 Month Horizon													
1	Z-Score	VW	-0.0056	1.6%	0.00	0.83	15	Amihud'S Measure	EW	0.0026	0.4%	0.03	1.00
2	Asset Turnover	VW	0.0053	1.5%	0.00	0.89	16	Cash Flow Variance	EW	0.0034	0.6%	0.07	1.00
3	Z-Score	EW	-0.0052	1.4%	0.00	0.91	17	ΔRec. + Accrual	EW	-0.0046	1.1%	0.03	1.00
4	LT Reversal	VW	-0.0046	0.7%	0.01	0.92	18	Spreads	EW	0.0026	0.4%	0.06	1.00
5	ΔNC Op. Assets	VW	-0.0036	0.7%	0.03	0.98	19	Capex Growth	VW	-0.0032	0.5%	0.07	1.00
6	Stock Split	EW	-0.0026	0.4%	0.10	1.00	20	SEO	VW	-0.0042	0.9%	0.01	1.00
7	ΔTax to Assets	EW	0.0055	1.6%	0.00	1.00	21	Asset Growth	VW	-0.0035	0.7%	0.02	1.00
8	Volume / MV	EW	-0.0044	1.0%	0.05	1.00	22	Moment Lt Reversal	VW	0.0025	0.2%	0.06	1.00
9	Coskewness	EW	0.0036	0.7%	0.04	1.00	23	Sustainable Growth	VW	-0.0031	0.5%	0.06	1.00
10	Secure / Total Debt	VW	0.0040	0.9%	0.03	1.00	24	Coskewness	VW	0.0037	0.8%	0.07	1.00
11	Spinoffs	VW	-0.0040	0.9%	0.05	1.00	25	% Operating Accrual	VW	-0.0038	0.9%	0.10	1.00
12	Pension Funding	VW	-0.0035	0.7%	0.01	1.00	26	Forecast Dispersion	VW	0.0032	0.6%	0.04	1.00
13	IPOs	VW	0.0035	0.7%	0.05	1.00	27	ΔRec. + Accrual	VW	-0.0049	1.3%	0.05	1.00
14	Target Price	EW	-0.0054	1.7%	0.05	1.00							
Panel B: 3 Month Horizon													
1	Z-Score	VW	-0.0055	4.6%	0.00	0.33	15	Spreads	VW	0.0028	1.2%	0.06	0.99
2	Asset Turnover	VW	0.0056	4.7%	0.00	0.33	16	Abn. Analyst Intense	EW	-0.0026	1.0%	0.01	1.00
3	LT Reversal	VW	-0.0048	2.2%	0.02	0.41	17	Reverse Split	EW	-0.002	0.6%	0.08	1.00
4	Z-Score	EW	-0.0046	3.1%	0.00	0.69	18	ΔTax to Assets	EW	0.0044	2.9%	0.00	1.00
5	SEO	VW	-0.0045	3.1%	0.00	0.72	19	Volume	VW	-0.002	0.4%	0.09	1.00
6	Target Price	EW	-0.0062	6.0%	0.01	0.84	20	Inventory Growth	VW	-0.0023	0.8%	0.08	1.00
7	ΔNC Op. Assets	VW	-0.0033	1.8%	0.04	0.84	21	Secure / Total Debt	VW	0.0021	0.7%	0.08	1.00
8	Asset Growth	VW	-0.0036	2.1%	0.01	0.85	22	Dividend Omission	EW	0.0026	1.0%	0.07	1.00
9	Coskewness	VW	0.0036	2.1%	0.04	0.85	23	Spreads	EW	0.0024	0.9%	0.05	1.00
10	% Operating Accrual	VW	-0.0045	3.4%	0.05	0.93	24	Reverse Split	VW	0.0028	1.3%	0.08	1.00
11	Sustainable Growth	VW	-0.0033	1.7%	0.03	0.95	25	Accruals	VW	0.0029	1.3%	0.07	1.00
12	Cash Flow Variance	EW	0.0033	1.7%	0.06	0.99	26	Share Issues PW	VW	-0.0031	1.4%	0.09	1.00
13	Capex Growth	VW	-0.0032	1.6%	0.06	0.99	27	Moment Lt Reversal	VW	0.0022	0.4%	0.08	1.00
14	ΔRec. + Accrual	VW	-0.0037	2.1%	0.06	0.99							

Table 3 - Continued

1	2	3	4	5	6	7	8	9	10	11	12	13	14
		EW or VW	$\hat{\beta}$	R^2	Raw	RW			EW or VW	$\hat{\beta}$	R^2	Raw	RW
Rank	Predictor	VW			p-Value		Rank	Predictor	VW			p-Value	
Panel C: 6 Month Horizon													
1	Asset Turnover	VW	0.0059	9.7%	0.00	0.07	19	Volume Trend	EW	0.0032	3.0%	0.08	0.94
2	Z-Score	VW	-0.0057	9.1%	0.00	0.10	20	Spreads	EW	0.0026	2.0%	0.01	0.94
3	LT Reversal	VW	-0.0049	4.6%	0.00	0.13	21	Reverse Split	VW	0.0029	2.6%	0.01	0.94
4	Z-Score	EW	-0.0044	5.2%	0.00	0.46	22	Moment-reverse	VW	-0.0026	1.2%	0.07	0.94
5	Asset Growth	VW	-0.0037	3.9%	0.00	0.60	23	ΔSales-ΔInventory	EW	-0.0026	2.0%	0.04	0.95
6	Spreads	VW	0.0035	3.7%	0.00	0.61	24	Share Issues PW	VW	-0.0028	2.1%	0.07	0.95
7	ΔNC Op. Assets	VW	-0.0032	3.1%	0.03	0.64	25	R&D/MV	EW	0.0026	2.0%	0.07	0.96
8	Sustainable Growth	VW	-0.0036	3.8%	0.01	0.64	26	Cash Flow Variance	EW	0.0029	2.6%	0.05	0.96
9	% Operating Accrual	VW	-0.0047	6.7%	0.02	0.66	27	Dividend Omission	EW	0.0021	1.3%	0.09	0.99
10	SEO	VW	-0.0038	4.0%	0.01	0.73	28	Org. Capital	EW	0.0023	1.5%	0.07	0.99
11	Target Price	EW	-0.0052	7.0%	0.01	0.90	29	Profit Margin	VW	0.0025	1.7%	0.09	0.99
12	ΔTax to Assets	EW	0.0048	6.0%	0.00	0.92	30	Abn. Analyst Intense	EW	-0.0012	0.4%	0.05	1.00
13	ΔCapex-ΔInd Capex	EW	-0.0029	2.4%	0.00	0.92	31	Share Issues DT	EW	-0.0021	1.2%	0.05	1.00
14	Capex Growth	VW	-0.0031	3.0%	0.06	0.92	32	Cash to Assets	VW	0.0014	0.6%	0.04	1.00
15	Accruals	VW	0.0029	2.4%	0.03	0.92	33	ΔTax to Assets	VW	0.0028	2.0%	0.09	1.00
16	ΔCapex-ΔInd Capex	VW	-0.0029	2.5%	0.02	0.92	34	Share Issues DT	VW	-0.0017	0.8%	0.06	1.00
17	Price	VW	-0.0028	1.4%	0.09	0.92	35	ΔSales-ΔSG&A	EW	0.0019	1.0%	0.09	1.00
18	Coskewness	VW	0.0027	2.2%	0.06	0.93	36	Cash Flow Variance	VW	0.002	1.2%	0.10	1.00
Panel D: 12 Month Horizon													
1	Z-Score	VW	-0.0058	18.5%	0.00	0.02	23	ΔSales-ΔInventory	EW	-0.0022	2.8%	0.03	0.95
2	Asset Turnover	VW	0.0058	17.8%	0.00	0.03	24	Moment-reverse	VW	-0.0022	1.7%	0.06	0.95
3	LT Reversal	VW	-0.0044	7.3%	0.00	0.13	25	R&D/MV	EW	0.0022	2.5%	0.08	0.97
4	Z-Score	EW	-0.0043	9.8%	0.00	0.28	26	Cash Flow Variance	EW	0.0023	3.0%	0.05	0.97
5	Sustainable Growth	VW	-0.0039	8.3%	0.00	0.30	27	Reverse Split	VW	0.0022	2.8%	0.03	0.97
6	ΔNC Op. Assets	VW	-0.0035	6.8%	0.01	0.31	28	Exchange Switch	VW	-0.0019	2.1%	0.04	0.97
7	Asset Growth	VW	-0.0037	7.7%	0.00	0.36	29	Coskewness	VW	0.0019	2.1%	0.06	0.97
8	Spreads	VW	0.0035	7.2%	0.00	0.37	30	Share Issues DT	EW	-0.0018	1.8%	0.03	0.99
9	Price	VW	-0.0036	4.1%	0.02	0.45	31	E/P	EW	-0.0017	1.7%	0.08	0.99
10	% Operating Accrual	VW	-0.0046	12.2%	0.01	0.47	32	Amihud'S Measure	EW	0.0018	1.8%	0.09	0.99
11	Capex Growth	VW	-0.0034	7.1%	0.02	0.63	33	Hybrid Covar. Risk	VW	0.0018	2.0%	0.10	0.99
12	ΔCapex-ΔInd Capex	EW	-0.003	5.2%	0.00	0.70	34	M/B and Accruals	VW	-0.002	2.3%	0.09	0.99
13	Spreads	EW	0.0027	4.2%	0.00	0.77	35	Stock Split	EW	-0.0017	1.8%	0.08	1.00
14	ΔCapex-ΔInd Capex	VW	-0.0028	4.6%	0.00	0.77	36	ΔTax to Assets	EW	0.0029	4.0%	0.01	1.00
15	Org. Capital	EW	0.003	4.9%	0.01	0.79	37	Coskewness	EW	0.001	0.6%	0.03	1.00
16	Employee Growth	VW	-0.0027	4.3%	0.04	0.80	38	E/P	VW	0.0008	0.4%	0.04	1.00
17	Volume Trend	EW	0.003	5.0%	0.03	0.86	39	CF/MV	VW	0.0007	0.3%	0.06	1.00
18	LT Reversal	EW	-0.0025	2.3%	0.05	0.86	40	ΔSales-ΔSG&A	EW	0.0016	1.5%	0.09	1.00
19	Share Issues PW	VW	-0.0027	4.0%	0.06	0.87	41	Δ # Institut. Owners	VW	-0.0015	1.3%	0.09	1.00
20	ΔAsset Turn	VW	0.0028	4.2%	0.02	0.87	42	ΔPrice Forecast	VW	-0.0011	0.5%	0.10	1.00
21	Target Price	EW	-0.0043	8.7%	0.03	0.92	43	Moment Lt Reversal	VW	0.0015	0.8%	0.03	1.00
22	SEO	VW	-0.0026	3.6%	0.08	0.93							

Table 4
Summary of In-Sample Performance using Romano and Wolf p-values
The table displays a count of the number of predictive variables that are statistically significant at the 10% level or better, as a fraction of the total number of variables examined. We calculate statistical significance using Romano and Wolf adjusted p-values. For each anomaly, we estimate an in-sample predictive regression of the form:

$$r_{t:t+h} = \alpha + \beta x_t + \varepsilon_{t:t+h} \text{ for } t = 1, \ldots, T-h,$$

where $r_{t:t+h} = (1/h)(r_{t+1} + \cdots + r_{t+h})$, r_t is the continuously compounded S&P 500 return for month t from CRSP including dividends and excess of the monthly risk-free rate from Goyal and Welch (2008), h indicates the forecast horizon in months, and x_t is one of the 140 predictor variables. To construct time-series predictors out of cross-sectional predictors, we calculate the value-weighted and equal-weighted mean across all stocks on each date resulting in 280 possible predictors. In Panel A, we consider four different definitions: (1) *Predictors from Existing Papers* uses only those variables that are used in the existing literature on time-series return predictability. (2) *Raw Predictors* examines every possible variable for which we reject the null that the raw variable is non-stationary. (3) *Raw Predictors + First Diff.* examines every possible variable however if a variable is not stationary in raw form, we then examine whether it is non-stationary in first-differenced form. If we fail to reject the null that the first differenced variable is non-stationary, we drop the variable. (4) All Possible Predictors examines every possible variable. If we fail to reject the null that a variable is non-stationary, we calculate deviations from a linear trend model. If we fail to reject the null that the linearly detrended variable is non-stationary, we calculate the first-difference. If we fail to reject the null that the first differenced variable is non-stationary, we drop the variable. In Panel B, we examine subsamples of the variables in *Raw Predictors + First Diff.* formed on the ten most statistically significant cross-sectional predictors (*Best Cross-Sectional*), and two different groupings based on the categories in McLean and Pontiff (2016): (1) *Opinion* and (2) *Valuation*. In Panel C we examine equal-weighted vs. value-weighted predictors for the variables in *Raw Predictors + First Diff.*

	Return Horizon (h)			
Predictive Variable	**h=1**	**h=3**	**h=6**	**h=12**
Panel A: Candidate Predictors (number significant / total examined)				
Predictors from Existing Papers	0/51	0/51	1/51	3/51
Raw Predictors	0/137	0/137	2/137	2/137
Raw Predictors + First Diff.	0/253	0/253	2/253	2/253
All Possible Predictors	0/269	0/269	1/269	3/269
Panel B: By Subcategory (number significant / total examined)				
Best Cross-sectional	0/20	0/20	1/20	1/20
Opinion	0/38	0/38	0/38	0/38
Valuation	0/24	0/24	0/24	1/24
Panel C: By Aggregation Method (number significant / total examined)				
Equal-Weighted Predictors	0/125	0/125	0/125	0/125
Value-Weighted Predictors	0/128	0/128	3/128	2/128

40

Table 5
Summary of Out-of-Sample Performance using Unadjusted p-values

The table displays a count of the number of predictive variables that are statistically significant at the 10% level or better, as a fraction of the total number of variables examined. We calculate statistical significance using bootstrap p-values. For each anomaly, we estimate an out-of-sample predictive regression of the form:

$$r_{t:t+h} = \alpha + \beta x_t + \varepsilon_{t:t+h} \text{ for } t = 1, \ldots, T - h,$$

where $r_{t:t+h} = (1/h)(r_{t+1} + \cdots + r_{t+h})$, r_t is the continuously compounded S&P 500 return for month t from CRSP including dividends and excess of the monthly risk-free rate from Goyal and Welch (2008), h indicates the forecast horizon in months, and x_t is one of the 140 predictor variables. We estimate expanding rolling window regressions using only information available on each date. To construct time-series predictors out of cross-sectional predictors, we calculate the value-weighted and equal-weighted mean across all stocks on each date resulting in 280 possible predictors. In Panel A, we consider four different definitions: (1) *Predictors from Existing Papers* uses only those variables that are used in the existing literature on time-series return predictability. (2) *Raw Predictors* examines every possible variable for which we reject the null that the raw variable is non-stationary. (3) *Raw Predictors + First Diff.* examines every possible variable however if a variable is not stationary in raw form, we then examine whether it is non-stationary in first-differenced form. If we fail to reject the null that the first differenced variable is non-stationary, we drop the variable. (4) All Possible Predictors examines every possible variable. If we fail to reject the null that a variable is non-stationary, we calculate deviations from a linear trend model. If we fail to reject the null that the linearly detrended variable is non-stationary, we calculate the first-difference. If we fail to reject the null that the first differenced variable is non-stationary, we drop the variable. In Panel B, we examine subsamples of the variables in *Raw Predictors + First Diff.* formed on the ten most statistically significant cross-sectional predictors (*Best Cross-Sectional*), and two different groupings based on the categories in McLean and Pontiff (2016): (1) *Opinion* and (2) *Valuation*. In Panel C we examine equal-weighted vs. value-weighted predictors for the variables in *Raw Predictors + First Diff.*

Predictive Variable	Return Horizon (*h*)			
	h=1	*h*=3	*h*=6	*h*=12
Panel A: Candidate Predictors (number significant / total examined)				
Predictors from Existing Papers	0/51	5/51	9/51	11/51
Raw Predictors	4/137	24/137	32/137	36/137
Raw Predictors + First Diff.	7/253	30/253	39/253	44/253
All Possible Predictors	7/269	32/269	41/269	45/269
Panel B: By Subcategory (number significant / total examined)				
Best Cross-sectional	0/20	2/20	3/20	3/20
Opinion	1/38	3/38	1/38	0/38
Valuation	0/24	0/24	0/24	1/24
Panel C: By Aggregation Method (number significant / total examined)				
Equal-Weighted Predictors	3/125	10/125	16/125	17/125
Value-Weighted Predictors	4/128	20/128	23/128	27/128

Table 6
Best Out-of-Sample Predictive Regression Results using Romano and Wolf p-values

The table reports the mean of the ordinary least squares estimate of β, p-values, and the Campbell and Thompson (2008) R^2_{OS} statistic from out-of-sample predictive regression models of the form:

$$r_{t:t+h} = \alpha + \beta x_t + \varepsilon_{t:t+h} \text{ for } t = 1, \ldots, T-h,$$

where $r_{t:t+h} = (1/h)(r_{t+1} + \cdots + r_{t+h})$, r_t is the continuously compounded S&P 500 return for month t from CRSP including dividends and excess of the monthly risk-free rate from Goyal and Welch (2008), h indicates the forecast horizon in months, and x_t is the predictor variable in the first column. $\hat{\beta}$ (column 4) is the time-series mean of the coefficient estimates for each predictor. The Campbell and Thompson R^2_{OS} statistic (columns 5 and 12) is calculated as 1 minus the proportional reduction in mean squared forecast error (MSFE) at the h-month horizon for a predictive regression forecast of the S&P 500 log excess return based on the predictor variable in the first column vis-a-vis the prevailing mean benchmark forecast. For each horizon, we run 253 out-of-sample regressions using the variables in the *Raw Predictors + First Diff.* set of predictors. Panel A displays results for the 1-month horizon, Panel B shows the 3-month horizon, Panel C shows the 6-month horizon, and Panel D shows the 12-month horizon. Within each panel, predictors are sorted by their Romano and Wolf p-value and then their unadjusted p-value. We report all predictors that have unadjusted p-values less than 10% for a given horizon. Unadjusted p-values are shown in columns 6 and 13 and Romano and Wolf (2016) adjusted p-values are shown in columns 7 and 14.

1	2	3	4	5	6	7	8	9	10	11	12	13	14
		EW or VW	$\hat{\beta}$	R^2	Raw	RW			EW or VW	$\hat{\beta}$	R^2	Raw	RW
Rank	Predictor				p-Value		Rank	Predictor				p-Value	
Panel A: 1 Month Horizon													
1	Asset Turnover	VW	0.0070	0.3%	0.04	1.00							
2	ΔRec + Accrual	VW	-0.0029	1.1%	0.04	1.00							
3	Z-score	EW	-0.0043	0.5%	0.04	1.00							
4	Z-score	VW	-0.0049	0.7%	0.05	1.00							
5	ΔTax to Assets	EW	0.0065	-0.9%	0.06	1.00							
6	Pension Funding	VW	-0.0043	-0.4%	0.08	1.00							
7	Coskewness	EW	0.0045	0.0%	0.09	1.00							
Panel B: 3 Month Horizon													
1	Idio Risk	VW	-0.0015	1.11	0.00	0.99	16	Z-Score	EW	-0.0039	1.43	0.04	1.00
2	Volume	EW	-0.0005	0.57	0.00	1.00	17	Exchange Switch	VW	-0.0034	-0.91	0.04	1.00
3	Asset Turnover	VW	0.0073	3.83	0.00	1.00	18	Δ # Institut. Owners	VW	0.0036	-13.43	0.04	1.00
4	ST Reversal	EW	-0.0002	0.39	0.00	1.00	19	Cash Flow Variance	VW	0.0042	-0.03	0.05	1.00
5	ST Reversal	VW	-0.0007	0.30	0.00	1.00	20	LT Reversal	VW	-0.0066	-0.61	0.05	1.00
6	Z-Score	VW	-0.0048	3.25	0.01	1.00	21	Cash Flow Variance	EW	0.0020	1.09	0.05	1.00
7	Lag Moment	VW	-0.0013	0.79	0.01	1.00	22	Age	EW	0.0163	-106.82	0.07	1.00
8	Max	VW	-0.0014	1.15	0.01	1.00	23	Moment Lt Reversal	VW	0.0021	0.34	0.08	1.00
9	Abn. Analyst Intense	EW	-0.0073	-0.11	0.02	1.00	24	Age	VW	0.0019	-0.07	0.08	1.00
10	Lag Moment	EW	-0.0010	0.61	0.02	1.00	25	Reverse Split	VW	0.0048	-0.10	0.08	1.00
11	ΔNC Op. Assets	VW	-0.0013	1.63	0.02	1.00	26	ΔTax to Assets	EW	0.0051	2.18	0.08	1.00
12	Price	VW	-0.0040	0.97	0.02	1.00	27	ΔRec. + Accrual	VW	-0.0030	1.45	0.09	1.00
13	Volume	VW	-0.0018	0.69	0.03	1.00	28	Capex Growth	VW	-0.0024	1.29	0.09	1.00
14	Size	EW	0.0000	0.30	0.03	1.00	29	Max	EW	-0.0001	0.36	0.09	1.00
15	Coskewness	VW	0.0055	1.43	0.03	1.00	30	Moment-reverse	VW	-0.0009	0.50	0.10	1.00

Table 6 - Continued

1	2	3	4	5	6	7	8	9	10	11	12	13	14
		EW or VW			Raw	RW			EW or VW			Raw	RW
Rank	Predictor		$\hat{\beta}$	R^2	p-Value		Rank	Predictor		$\hat{\beta}$	R^2	p-Value	
Panel C: 6 Month Horizon													
1	ST Reversal	EW	0.0000	0.28	0.00	0.92	21	Cash Flow Variance	EW	0.0018	2.39	0.03	0.86
2	Volume	EW	-0.0003	0.90	0.00	1.00	22	LT Reversal	EW	-0.0028	1.45	0.03	1.00
3	Idio Risk	VW	-0.0020	1.56	0.00	0.98	23	Exchange Switch	VW	-0.0041	-1.94	0.04	1.00
4	Size	VW	0.0004	0.39	0.00	1.00	24	Reverse Split	VW	0.0043	-1.32	0.04	1.00
5	ST Reversal	VW	0.0003	0.34	0.00	1.00	25	Spreads	VW	0.0080	-2.32	0.04	1.00
6	Asset Turnover	VW	0.0075	9.63	0.00	1.00	26	Idio Risk	EW	-0.0001	1.31	0.04	1.00
7	Moment Lt Reversal	VW	0.0008	1.18	0.00	1.00	27	Moment-reverse	EW	-0.0011	1.19	0.04	1.00
8	Volume	VW	-0.0017	0.69	0.00	1.00	28	Age	VW	0.0019	-0.10	0.05	1.00
9	Moment-reverse	VW	-0.0024	2.19	0.00	1.00	29	Coskewness	VW	0.0046	1.42	0.05	1.00
10	Price	VW	-0.0036	1.88	0.01	1.00	30	Dividends	EW	0.0000	1.01	0.06	1.00
11	Max	EW	-0.0014	0.80	0.01	1.00	31	Z-Score	EW	-0.0041	1.91	0.06	1.00
12	Max	VW	-0.0024	1.46	0.01	1.00	32	ΔCapex-ΔInd Capex	VW	-0.0037	0.02	0.06	1.00
13	Size	EW	0.0005	0.23	0.01	1.00	33	Price	EW	-0.0023	0.48	0.06	1.00
14	Z-Score	VW	-0.0050	7.37	0.01	0.95	34	Dividends	VW	0.0000	0.97	0.06	1.00
15	Lag Moment	VW	-0.0015	1.94	0.01	1.00	35	ΔCapex-ΔInd Capex	EW	-0.0038	-0.07	0.06	1.00
16	LT Reversal	VW	-0.0064	1.15	0.01	1.00	36	Capex Growth	VW	-0.0026	3.07	0.07	1.00
17	Spreads	EW	0.0120	-1.61	0.02	1.00	37	ΔSales-ΔInventory	EW	-0.0042	-4.29	0.08	1.00
18	ΔNC Op. Assets	VW	-0.0009	3.32	0.02	1.00	38	Asset Growth	VW	0.0003	0.97	0.09	1.00
19	Lag Moment	EW	-0.0012	1.57	0.02	1.00	39	Abn. Analyst Intense	EW	-0.0043	-0.66	0.09	1.00
20	Cash Flow Variance	VW	0.0057	0.26	0.03	1.00							
Panel D: 12 Month Horizon													
1	Volume	EW	0.0000	0.25	0.00	0.95	23	Max	EW	-0.0003	-0.31	0.03	1.00
2	Volume	VW	-0.0002	0.15	0.00	0.99	24	Dividends	EW	0.0000	1.69	0.03	1.00
3	Lag Moment	VW	-0.0020	4.36	0.00	1.00	25	Dividends	VW	0.0000	1.68	0.03	1.00
4	Size	VW	0.0009	-0.19	0.00	1.00	26	Z-Score	VW	-0.0053	15.01	0.03	0.91
5	Moment Lt Reversal	EW	0.0009	3.85	0.00	0.98	27	ΔNC Op. Assets	VW	-0.0008	7.35	0.03	1.00
6	Moment Lt Reversal	VW	0.0013	4.00	0.00	0.94	28	Age	VW	0.0021	-1.17	0.03	1.00
7	Reverse Split	VW	0.0028	-2.50	0.00	1.00	29	Exchange Switch	VW	-0.0040	-2.26	0.04	1.00
8	Lag Moment	EW	-0.0011	2.90	0.00	1.00	30	Moment-reverse	EW	-0.0006	1.06	0.05	1.00
9	Price	VW	-0.0047	2.99	0.00	1.00	31	Sustainable Growth	VW	-0.0021	2.31	0.05	1.00
10	Moment-reverse	VW	-0.0020	2.70	0.00	1.00	32	ΔCapex-ΔInd Capex	EW	-0.0025	2.46	0.05	1.00
11	Spreads	EW	0.0096	1.51	0.00	1.00	33	ΔCapex-ΔInd Capex	VW	-0.0030	0.41	0.07	1.00
12	Max	VW	-0.0013	1.08	0.00	1.00	34	Cash Flow Variance	VW	0.0062	-0.76	0.07	1.00
13	Asset Turnover	VW	0.0079	17.46	0.00	1.00	35	Org. Capital	EW	0.0027	4.62	0.07	1.00
14	ST Reversal	VW	0.0011	-0.95	0.01	1.00	36	Volume / MV	VW	-0.0012	0.02	0.09	1.00
15	ST Reversal	EW	0.0010	-1.01	0.01	1.00	37	Price	EW	-0.0030	-0.83	0.09	1.00
16	Spreads	VW	0.0067	1.58	0.01	1.00	38	Volume Trend	EW	0.0031	2.61	0.09	1.00
17	Size	EW	0.0010	-1.30	0.01	1.00	39	Momentum	EW	0.0007	-1.55	0.09	1.00
18	LT Reversal	VW	-0.0052	6.09	0.01	1.00	40	Asset Growth	VW	0.0001	3.07	0.09	1.00
19	LT Reversal	EW	-0.0025	3.61	0.01	1.00	41	ΔSales-ΔInventory	EW	-0.0024	-4.57	0.09	1.00
20	Idio Risk	VW	0.0005	-0.84	0.01	1.00	42	Coskewness	VW	0.0039	0.18	0.09	1.00
21	Capex Growth	VW	-0.0034	7.55	0.02	1.00	43	% Operating Accrual	VW	-0.0046	9.75	0.10	1.00
22	Cash Flow Variance	EW	0.0013	4.44	0.02	0.75	44	IPOs	VW	-0.0018	-1.26	0.10	1.00

43

Table 7
Summary of Out-of-Sample Performance using Romano and Wolf p-values

The table displays a count of the number of predictive variables that are statistically significant at the 10% level or better, as a fraction of the total number of variables examined. We calculate statistical significance using Romano and Wolf adjusted p-values. For each anomaly, we estimate an out-of-sample predictive regression of the form:

$$r_{t:t+h} = \alpha + \beta x_t + \varepsilon_{t:t+h} \text{ for } t = 1, \ldots, T - h,$$

where $r_{t:t+h} = (1/h)(r_{t+1} + \cdots + r_{t+h})$, r_t is the continuously compounded S&P 500 return for month t from CRSP including dividends and excess of the monthly risk-free rate from Goyal and Welch (2008), h indicates the forecast horizon in months, and x_t is one of the 140 predictor variables. We estimate expanding rolling window regressions using only information available on each date. To construct time-series predictors out of cross-sectional predictors, we calculate the value-weighted and equal-weighted mean across all stocks on each date resulting in 280 possible predictors. In Panel A, we consider four different definitions: (1) *Predictors from Existing Papers* uses only those variables that are used in the existing literature on time-series return predictability. (2) *Raw Predictors* examines every possible variable for which we reject the null that the raw variable is non-stationary. (3) *Raw Predictors + First Diff.* examines every possible variable however if a variable is not stationary in raw form, we then examine whether it is non-stationary in first-differenced form. If we fail to reject the null that the first differenced variable is non-stationary, we drop the variable. (4) All Possible Predictors examines every possible variable. If we fail to reject the null that a variable is non-stationary, we calculate deviations from a linear trend model. If we fail to reject the null that the linearly detrended variable is non-stationary, we calculate the first-difference. If we fail to reject the null that the first differenced variable is non-stationary, we drop the variable. In Panel B, we examine subsamples of the variables in *Raw Predictors + First Diff.* formed on the ten most statistically significant cross-sectional predictors (*Best Cross-Sectional*), and two different groupings based on the categories in McLean and Pontiff (2016): (1) *Opinion* and (2) *Valuation*. In Panel C we examine equal-weighted vs. value-weighted predictors for the variables in *Raw Predictors + First Diff.*

Predictive Variable	Return Horizon (h)			
	h=1	h=3	h=6	h=12
Panel A: Candidate Predictors (number significant / total examined)				
Predictors from Existing Papers	0/51	0/51	0/51	0/51
Raw Predictors	0/137	0/137	0/137	0/137
Raw Predictors + First Diff.	0/253	0/253	0/253	0/253
All Possible Predictors	0/269	0/269	0/269	0/269
Panel B: By Subcategory (number significant / total examined)				
Best Cross-sectional	0/20	0/20	0/20	0/20
Opinion	0/38	0/38	0/38	0/38
Valuation	0/24	0/24	0/24	0/24
Panel C: By Aggregation Method (number significant / total examined)				
Equal-Weighted Predictors	0/125	0/125	0/125	0/125
Value-Weighted Predictors	0/128	0/128	0/128	0/128