Replicate several portfolios in Ken French's Data Library

Introduction

The pioneering work of Fama and French have significantly influenced empirical asset pricing through their development of factor models. These models, which articulate the returns to the market factor (Rm-Rf), the size factor (SMB), and the value factor (HML), have been instrumental in academic and practical finance for over three decades. The Fama-French Data Library, that emerged from their seminal paper *Common Risk Factors in the Returns on Stocks and Bonds* (1993), has evolved to become an indispensable asset for researchers and practitioners alike. This library not only facilitates access to the critical stock factors (Mkt, SMB, HML) but also underscores the dynamic nature of financial data and the continual refinement of asset pricing models.

Our project is motivated by the process of updating and refining the inputs and methodologies that underpin the Fama-French factor models. Fama and French's meticulous approach to data corrections and rule changes, as described in their recent work in *Production of U.S. Rm-Rf, SMB, and HML in the Fama-French Data Library* (2023), underscores the complexity and the necessity of maintaining the highest accuracy in financial databases like CRSP (Center for Research in Security Prices) and Compustat. These databases, which provide the foundational data for constructing factor portfolios, undergo continual revisions to correct historical errors and adapt to evolving accounting standards and market realities. The adjustments made to account for such corrections directly impact the computed returns of Rm-Rf, SMB, and HML, thereby affecting the reliability and validity of empirical asset pricing research.

This project aims to replicate the univariate sorts on size, earnings/price (E/P), cashflow/price (CF/P), and dividend yield (D/P), as well as the sorts on operating profitability and investment, using unaggregated CRSP and Compustat data. In doing so, we confront the challenges of accurately linking and processing vast datasets, calculating breakpoints, and sorting stocks into portfolios that reflect the original methodologies of Fama and French.

Data Sources

Our replication is built upon two primary data sources: CRSP and Compustat. CRSP provided stock prices, returns, and share outstanding data, critical for calculating market equity (ME) and returns over time. Compustat offered comprehensive accounting data, including earnings, cash flows, and dividends, necessary for computing E/P, CF/P, and D/P ratios. One of the challenges in utilizing these databases was ensuring accurate linkage between CRSP and Compustat data for each stock. As a result, we employed a combination of CRSP's PERMNO and Compustat's GVKEY identifiers, leveraging the CRSP/Compustat Merged Database when possible to minimize mismatches.

Methodology

This section outlines the methodologies employed in replicating the portfolios based on Fama and French's framework, focusing on Earnings/Price (E/P), Cashflow/Price (CF/P), and the intersection of Size and Operating Profitability.

Portfolios Formed on Earnings/Price (E/P)

Construction

Portfolios were formed at the end of each June, utilizing NYSE breakpoints for sorting. The earnings considered were the total earnings before extraordinary items for the last fiscal year ending in $t-1$. The market equity (ME), defined as the product of price and shares outstanding at the end of December of $t-1$, served as the price denominator.

Portfolio Categories

- **Negative Earnings:** Firms with earnings less than zero were allocated to a distinct "Earnings ¡ 0" portfolio.

- **Percentile Sorts:** The remaining firms were sorted into bottom 30%, middle 40%, and top 30% portfolios based on their E/P ratios, in addition to being classified into quintiles and deciles.

Data Coverage

The methodology was applied to all stocks listed on the NYSE, AMEX, and NASDAQ that met the data availability criteria for ME and earnings before extraordinary items from December of $t-1$ and June of t, spanning from July 1951 to January 2024 for monthly returns, and from 1952 to 2021 for annual returns.

Portfolios Formed on Cashflow/Price (CF/P)

Construction

Similar to E/P portfolios, CF/P portfolios were constructed at the end of each June using NYSE breakpoints. The cashflow measure included earnings before extraordinary items, equity's share of depreciation, and, if available, deferred taxes for the last fiscal year ending in $t-1$.

Portfolio Categories

- **Negative Cashflow:** Firms with a negative cashflow were allocated to a "Cashflow ¡ 0" portfolio.

- **Percentile Sorts:** Other firms were sorted into bottom 30%, middle 40%, and top 30% portfolios based on CF/P ratios, alongside quintile and decile classifications.

Data Coverage

The analysis included stocks from the NYSE, AMEX, and NASDAQ with available ME for December of $t-1$ and June of t, and cashflow data for the calendar year $t-1$, covering annual returns from 1952 to 2021.

Portfolios Formed on Size and Operating Profitability

Construction

Portfolios were constructed at the end of each June, based on the intersections of 5 portfolios formed on operating profitability (OP) and 5 on investment (Inv). OP was calculated as annual revenues minus cost of goods sold, interest expense, and selling, general, and administrative expenses, divided by book equity for the last fiscal year ending in $t-1$. Investment was measured as the change in total assets from the fiscal year ending in $t-2$ to $t-1$, divided by total assets for $t-2$.

Portfolio Categories

The construction resulted in 25 portfolios formed from the intersections of quintile ranks for OP and Inv, utilizing NYSE quintiles for breakpoints.

Data Coverage

The portfolios included all NYSE, AMEX, and NASDAQ stocks with (positive) book equity for $t-1$, total assets data for $t-2$ and $t-1$, non-missing revenues data for $t-1$, and at least one non-missing data point among cost of goods sold, selling, general, and administrative expenses, or interest expense for $t-1$, covering the period from July 1963 to January 2024 for monthly returns, and 1964 to 2023 for annual returns.

Results

This section presents the results of our replication effort, including graphical representations of the portfolios' performance and tabulated data for clearer insights into the analysis.

Graphical Analysis

Figures 1 and 2 show the performance trends of the portfolios formed on Earnings/Price (E/P) and Cashflow/Price (CF/P), respectively, across different time frames.
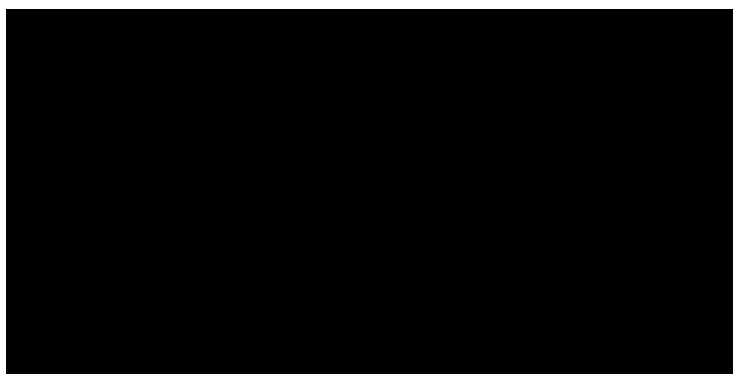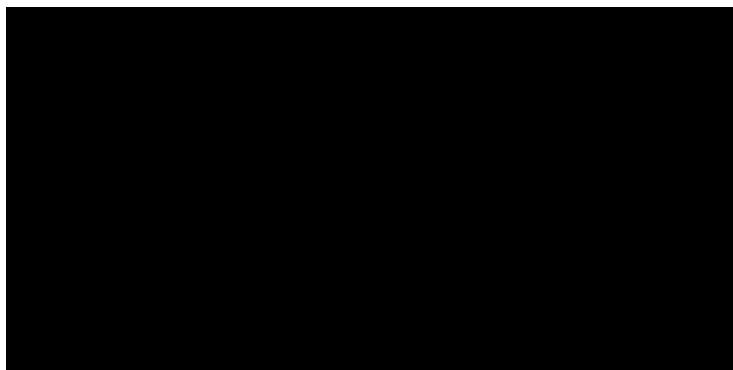

Figure 1: E/P Portfolio Performance Over Time

Figure 2: CF/P Portfolio Performance Over Time

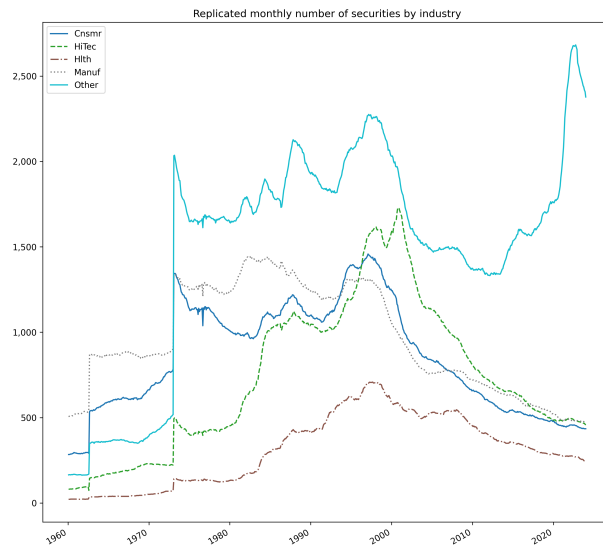## Analyzing Industry Portfolios: Summary Statistics

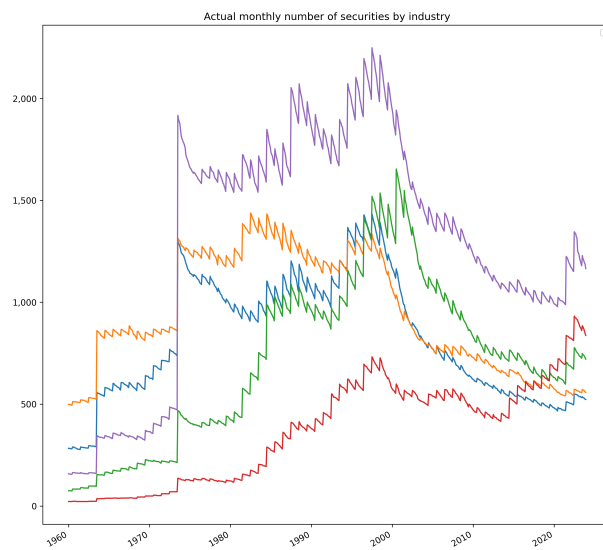Figure 3: Replication for Number of Firms in 5 Industry Portfolios

Figure 4: Actual Number of Firms in 5 Industry Portfolios

Replication of 5 Industry Portfolios Summary Statistics (1960-2024):

industry5	Cnsmr	HiTec	Hlth	Manuf	Other
count	768.0000	768.0000	768.0000	768.0000	768.0000
mean	855.2552	732.3789	314.3685	962.8060	1488.1354
std	314.3435	435.1216	204.4612	308.4889	637.5753
min	284.0000	74.0000	22.0000	472.0000	163.0000
255075max	1458.0000	1729.0000	708.0000	1442.0000	2682.0000

Actual 5 Industry Portfolios Summary Statistics (1960-2024):

	Cnsmr	Manuf	HiTec	Hlth	Other
count	768.0000	768.0000	768.0000	768.0000	768.0000
mean	821.5599	946.5990	716.7826	367.3919	1281.3555
std	296.0788	284.4789	394.7895	248.3433	583.2283
min	277.0000	495.0000	74.0000	22.0000	156.0000
255075max	1433.0000	1438.0000	1654.0000	932.0000	2249.0000

Discussion

[Here we can insert discussion about trends, comparisons to benchmarks, and any anomalies or significant findings observed in the data.]

Challenges in Replication

The replication presented technical and methodological challenges that impacted our ability to precisely mirror their original findings, specifically the calculation of breakpoints and the sorting of stocks into portfolios.

Methodological Interpretation

Despite rigorous documentation, practical replication efforts often encounter challenges in interpreting and applying specific methodological nuances, especially regarding the treatment of outliers or the handling of missing data. These could inadvertently lead to variations in the calculated breakpoints, impacting the composition of the resulting portfolios.

Portfolio Sorting Intricacies

Conditional Sorting and Intersection

A notable challenge in portfolio sorting was managing the conditional sorting criteria and the intersection of various metrics that is highly sensitive to the accurate calculation of each underlying portfolio. Discrepancies in any of these calculations can cascade through the sorting process, affecting the final portfolio composition.

Kahle, K.M. and Walkling, R.A. (1996). *The Impact of Industry Classifications on Financial Research. The Journal of Financial and Quantitative Analysis*, 31(3), p.309.

Final Output and Calculation Breakdown

A significant obstacle we faced in our replication was the challenge associated with the verification of the final output due to the absence of intermediary computational results. This limitation becomes particularly pronounced when attempting to identify and correct discrepancies, as the process lacks transparency without access to detailed breakdowns of each calculation step. Such a situation makes it difficult to pinpoint errors, requiring, each time, a meticulous review of the entire computational procedure to ensure accuracy.

Data Availability and SIC Code Challenges

Furthermore, the replication was compounded by missing data points within the code, which made the construction of accurate functions and the execution of complex calculations more difficult. The evolution and changes in Standard Industrial Classification (SIC) codes over an extensive period (80+ years) presented another layer of complexity.

The discrepancies between the CRSP and Compustat databases regarding a range of values added inconsistencies that significantly impacted the fidelity of replication efforts, as they may introduce variances in the data set that are not present in the original studies. The study on the stability and reliability of financial variables over time by Kahle, K.M. and Walkling, R.A. (1996) in the *Journal of Financial and Quantitative Analysis*, highlights the difficulties in maintaining consistent data classifications over prolonged periods and underscores the potential for classification and data discrepancies to skew results.