Intra-daily 51 Financial Factors Database

Guilherme Masuko
Advisor: Marcelo C. Medeiros

March 2023

Abstract

This paper had as main goal to report how we will create an intra daily basis of some of the key financial factors for the US market. In Section 2 we theoretically show how we form 51 factors. The final database will have 51 financial factors in addition to the returns of stocks listed on the US stock exchange NYSE. The main novelty in this database is the intra-daily frequency of these variables.

During the scope of the article, we discussed the use of dimensionality reduction tools that will be needed in the context of model selection. We motivated this discussion by bringing the next steps that include verifying whether these factors help in the predictive power of high-frequency returns.

1 Introduction

Within the predictive world of finance, financial economists try to find methods and models that can increasingly more accurately predict stock returns. There are several approaches and tools used in financial analysis. These techniques are used to analyze historical data and try to predict future stock market movements.

On the one hand, the literature walked for a long time trying to discover financial factors that could better explain the returns or that would increase the predictive power of models that already have significant factors when making forecasts of returns. The search for factors that explain the cross section of expected stock returns began a long time ago with its first model, the Capital Asset Pricing Model (CAPM) introduced by Sharpe (1964), Lintner (1965) and Mossin (1966) independently , based on previous work by Markowitz (1952), using only the market factor to explain stocks' excess returns.

The factors are formed through portfolios where there is a long and a short position. For example, to create the market factor, we use a long position on the market index of interest and short on the risk-free interest rate, that is, market return minus risk-free asset return.

Looking for more factors that increase the predictive power, the well-known Fama and French models emerged, respectively, the three-factor model in Fama e French (1992) and the five-factor model in Fama e French (2015).

The first model aggregates the factors related to size, SMB (Small minus Big), and value, HML (High minus Low), based respectively on the variables Market Capitalization and Book-to-Market ratio, keeping the market factor at your model. In the second model, the authors add two more factors related to profitability and investment. Defined analogously to the SMB and HML factors, the RMW (Robust minus Weak) factors are constructed from the difference between the returns of firms with robust (high) and weak (low) operating profitability and the CMA investment factor (Conservative minus Aggressive) constructed from the difference between the returns of companies that invest conservatively and companies that invest aggressively. Another factor related to momentum was also added to the Fama and French three-factor model by Carhart (1997). MOM (Monthly Momentum Factor) can be calculated by subtracting the equal weighted average of the lowest performing companies from the equal weighted average of the highest performing companies.

Now, with hundreds of financial factors released in the literature, the difficulty has become to tame this high dimensionality of factors.

On the other hand, there is still a great use of traditional methods, such as the use

of autoregressive regression (AR). This is what is done in Chinco, Clark-Joseph e Ye (2019), which is the article most closely linked to this work. In the article, the authors use the AR regression with three lags as a reference model in their main specifications, although they demonstrate that the number of lags does not matter with the problem addressed. The objective of the article is to show that the identification of predictors in finance using only the intuition of researchers would only work for long-term stable predictors.

In high-frequency finance with modern, large, fast, and complex financial markets, the process of selecting predictors requires efforts that go beyond researchers' intuition. And perhaps these long-term stable predictors are not suitable to explain intraday returns because they are not able to capture the effects that affect stock prices in the middle of a day. The paper's results show that statistical model selection tools can increase the predictive power of one-minute-ahead forecasts of benchmark models using out-of-sample fit and forecast-implied Sharpe ratios as measures of quality.

For model selection, the authors make use of LASSO (Least Absolute Shrinkage and Selection Operator) using three lags of all returns as candidate predictors. They ague the use of a dimension reduction tool because there are many predictor candidates. For example, in the month of January 2003 our sample contains $4500+$ stocks. By using three lags, we then have $3 \cdot 4500+ = 13500+$ candidate predictors. It would take more than 34 trading days to do a simple one-minute OLS (Ordinary Least Squares) estimation (each day has 390 minutes/observations) and test the predictors.

With this high dimensionality of predictor candidates, a reduction tool is necessary and a researcher's intuition does not seem to be the right path for this scale. LASSO fits as a solution because it is able to identify unexpected and short-term predictors, suitable for an intraday financial predictive model. The initial hypothesis for using LASSO is to bet on sparsity, that is, if among the $13500+$ candidate predictors, only a few predictors, say S, actually help to predict the returns of an asset, then we can leverage to use LASSO, because it would be needed only a little more than S observations to make predictions, this means that, since we don't have to worry about the weak estimators, LASSO can estimate the remaining parameters with far fewer observations. Thus, if there are only S important predictors at each point in time, LASSO is adequate for estimating unexpected short-term parameters. We explain more about how LASSO works in Appendix A.1.

Although the paper obtains results that refer to gains in predictive power by combining LASSO with Benchmarks models, the selection of predictors is made based only on return lags. The factors literature is neglected and therefore, this work was motivated to verify what happens when we add to this predictive model a base that

contains relevant financial factors.

This summer paper therefore aims to explain how the database will be created and report the next steps after that. The work is divided into two more sections. Section 2 addresses how the databases have been built so far and how theoretically the factors will be built. Section 3 will motivate what the next steps will be after having the base ready.

2 Data

The first database of this work is TAQ Returns. The TAQ Returns Database is a financial database that contains information about returns for all stocks listed on the New York Stock Exchange (NYSE). We have around ten thousand assets entering and leaving the database from January 2003 to December 2020.

The TAQ Returns database provides return data at different frequencies, ranging from daily frequency to minute and second frequencies. For example, you can get return data for intervals such as 30, 15, 5, and 1 minute, as well as 30, 15, and 5 seconds. This allows for a more detailed and accurate analysis of stock performance over different time frames.

The second database was built using two different sets of data. The first contains data on PERMNO, TAQ_TICKER and some firm characteristics. The second dataset also has the PERMNO column, which is the variable we connect the two datasets with, and several columns that give us the percentiles of companies on where they are located based on financial factors. Companies are grouped into ten percentiles. Thus, each firm receives values between 1 and 10 for each factor. Next, we show how each financial factor was constructed.

1. **Size** ($size$)

   Based on Fama e French (1993) we can built the first two factors, $size$ and $value$.

$$size = \mathrm{ME}_{\mathrm{Jun}}$$

   They construct six portfolios, $S/L, S/M, S/H, B/L, B/M, B/H$ from the intersections of two ME (Market Equity - price times shares) and three $\mathrm{BE/ME}$ (Ratio of Book Equity to Market Equity) groups. The size breakpoint for year t is the median NYSE market equity at the end of June of year t, breaking the first group

ME into two, S (Small) and B (Big). The SMB (Small minus Big) factor is, therefore, the average return on the three small portfolios minus the average return on the three big portfolios.

$$SMB = \frac{1}{3}\left(S/L + S/M + S/H\right) - \frac{1}{3}\left(B/L + B/M + B/H\right)$$

Rebalanced annually.

2. **Value (annual)** ($value$)

   Follows Fama e French (1993).

$$value = \frac{\mathrm{BE}}{\mathrm{ME}}$$

   The $\mathrm{BE/ME}$ group is broken into three book-to-market equity groups based on the breakpoints until the 30th $\mathrm{BE/ME}$ percentile is L (Low), from 30th until 70th percentile is M (Middle) and finally the top 30th percentile H (High). $\mathrm{BE/ME}$ for June of year t is the book equity for the last fiscal year end in $t-1$ divided by ME for December of $t-1$. The HML (High minus Low) factor is the average return on the two high portfolios minus the average return on the two low portfolios.

$$HML = \frac{1}{2}\left(S/H + B/H\right) - \frac{1}{2}\left(S/L + B/L\right)$$

   Rebalanced annually.

3. **Gross Profitability** ($prof$)

   Follows Novy-Marx (2013).

$$prof = \frac{\mathrm{GP}}{\mathrm{AT}}$$

   Profitability is measured by the ratio of a firm's gross profits (revenues minus cost of goods sold) to its assets. The gross profitability factor is constructed using

the basic methodology employed in the construction of HML. The strategy are long (short) firms in the top (bottom) tertile by NYSE breaks on the profitability sorting variable. The returns of this portfolio are value-weighted and rebalanced annually at the end of June based on its performance over the first 11 months of the preceding year and gross profits-to-assets. The PMU (Profitable minus Unprofitable) factor is

$$PMU = P - U$$

where GP is gross profits and AT is total assets, P and U are, respectively, the value-weighted portfolio of the top 30th percentile and the bottom 30th percentile.

4. **Value-Profitability** ($valprof$).

   Follows Novy-Marx (2013).

$$valprof = \text{rank}(value) + \text{rank}(prof)$$

   Sum of ranks in univariate sorts on book-to-market and profitability. Annual book-to-market and profitability values are used for the entire year. Rebalanced annually.

   The strategy that the author consider is construct within the 500 largest non-financial stocks for which gross profits-to-assets and book-to-market are both available. Each year these stocks are ranked on both their gross profits-to-assets and book-to-market ratios, from 1 (lowest) to 500 (highest). At the end of each June the strategy buys one dollar of each of the 150 stocks with the highest (H) combined profitability and value ranks, and it shorts one dollar of each of the 150 stocks with the lowest (L) combined ranks.

$$HML = H - L$$

5. **Piotroski's F-score** ($F\text{-}score$)

   Follows Piotroski (2000).

$$F\text{-}score = 1_{\text{IB}>0} + 1_{\Delta\text{ROA}>0} + 1_{\text{CFO}>0} + 1_{\text{CFO}>\text{IB}} + 1_{\Delta\text{DTA}<0|\text{DLTT}=0|\text{DLTT}_{-12}=0} + 1_{\Delta\text{ATL}>0} + 1_{\text{EqIss}\leq 0} + 1_{\Delta\text{GM}>0} + 1_{\Delta\text{ATO}>0}$$

where IB is income before extraordinary items, ROA is income before extraordinary items scaled by lagged total assets, CFO is cash flow from operations, DTA is total long-term debt scaled by total assets, DLTT is total long-term debt, ATL is total current assets scaled by total current liabilities, EqIss is the difference between sales of common stock and purchases of common stock recorded on the cash flow statement, GM equals one minus the ratio of cost of goods sold and total revenues, and ATO equals total revenues, scaled by total assets. Rebalanced annualy.

6. **Debit Issuance** ($debtiss$)

   Follows Spiess e Affleck-Graves (1999).

$$debtiss = 1_{\text{DLTISS}\leq 0}$$

   Binary variable equal to one if long-term debt issuance indicated in statement of cash flow. Updated annually.

7. **Share Repurchases** ($repurch$)

   Follows Ikenberry, Lakonishok e Vermaelen (1995).

$$repurch = 1_{\text{PRSTKC}>0}$$

   Binary variable equal to one if repurchase of common or preferred shares indicated in statement of cash flow. Updated annually.

8. **Share Issuance (annual)** ($nissa$)

   Follows Pontiff e Woodgate (2008).

$$nissa = \frac{\text{shrout}_{\text{Jun}}}{\text{shrout}_{\text{Jun}-12}}$$

where shrout is the number of shares outstanding. Change in real number of shares outstanding from past June to June of the previous year. Excludes changes in shares due to stock dividends and splits, and companies with no changes in shrout.

9. **Accruals** ($accruals$)

   Follows Sloan (1996).

$$accruals = \frac{\Delta\text{ACT} - \Delta\text{CIIE} - \Delta\text{LCT} + \Delta\text{DLC} + \Delta\text{TXP} - \Delta\text{DP}}{(\text{AT} + \text{AT}_{-12})/2}$$

   where ΔACT is the annual change in total current assets, ΔCHE is the annual change in total cash and short-term investments, ΔLCT is the annual change in current liabilities, ΔDLC is the annual change in debt in current liabilities, ΔTXP is the annual change in income taxes payable, ΔDP is the annual change in depreciation and amortization, and $(\text{AT} + \text{AT}_{-12})/2$ is average total assets over the last two years. Rebalanced annually.

10. **Asset Growth** ($growth$)

    Follows Cooper, Gulen e Schill (2008).

$$growth = \frac{\text{AT}}{\text{AT}_{-12}}$$

    Rebalanced annually.

11. **Asset Turnover** ($aturnover$)

    Follows Soliman (2008) and Novy-Marx (2013).

$$aturnover = \frac{\text{SALE}}{\text{AT}}$$

    Sales to total assets. Rebalanced annually.

    Portfolios are constructed using a quintile sort, where Low (L) represents the lowest 20th percentile and High (H) represents the highest 20th percentile based on New York Stock Exchange (NYSE) break points, and are rebalanced each year at the end of June

$$HML = H - L$$

12. **Gross Margins** ($gmargins$).

    Follows Novy-Marx (2013).

$$gmargins = \frac{\text{GP}}{\text{SALE}}$$

    where GP is gross profits and SALE is total revenues. Rebalanced annually.

    Portfolios are constructed using a quintile sort, where Low (L) represents the lowest 20th percentile and High (H) represents the highest 20th percentile based on New York Stock Exchange (NYSE) break points, and are rebalanced each year at the end of June.

$$HML = H - L$$

13. **Dividend Yield** ($divp$)

    Follows Naranjo, Nimalendran e Ryngaert (1998).

$$divp = \frac{\text{Div}}{\text{ME}_{\text{Dec}}}$$

    Dividend scaled by price. Both are measured in December of the year $t-1$ or $t-2$ (for returns in months prior to July). Rebalanced annually.

14. **Earnings/Price** (ep)

    Follows Basu (1977).

$$ep = \frac{\text{IB}}{\text{ME}_{\text{Dec}}}$$

    Net income scaled by market value of equity. Updated annually.

15. **Cash Flow / Market Value of Equity** (cfp)

Follows Lakonishok, Shleifer e Vishny (1994).

$$cfp = \frac{\text{IB} + \text{DP}}{\text{ME}_{\text{Dec}}}$$

Net income plus depreciation and amortization, all scaled by market value of equity measured at the same date. Updated annually.

16. **Net Operating Assets** (noa)

    Follows Hirshleifer et al. (2004).

$$noa = \frac{(\text{AT} - \text{CHE}) - (\text{AT} - \text{DLC} - \text{DLTT} - \text{MIB} - \text{PSTK} - \text{CEQ})}{\text{AT}_{-12}}$$

    where AT is total assets, CHE is cash and short-term investments, DLC is debt in current liabilities, DLTT is long term debt, MIB is non-controlling interest, PSTK is preferred capital stock, and CEQ is common equity. Updated annually.

17. **Investment** (inv).

    Follows Chen, Novy-Marx e Zhang (2011).

$$inv = \frac{\Delta\text{PPEGT} + \Delta\text{INVT}}{\text{AT}_{-12}}$$

    where ΔPPEGT is the annual change in gross total property, plant, and equipment, ΔINVT is the annual change in total inventories, and AT_{-12} is lagged total assets. Rebalanced annually, uses the full period.

18. **Investment-to-Capital** ($invcap$).

    Follows Xing (2008).

$$invap = \frac{\text{CAPX}}{\text{PPENT}}$$

    Investment to capital is the ratio of capital expenditure (Compustat item CAPX) over property, plant, and equipment (Compustat item PPENT).

19. **Invetment Growth** ($growth$).

Follows Xing (2008).

$$growth = \frac{\text{CAPX}}{\text{CAPX}_{-12}}$$

Investment growth is the percentage change in capital expenditure (Compustat item (CAPX)

20. **Sales Growth** $(sgrouth)$.

    Follows Lakonishok, Shleifer e Vishny (1994).

$$sgrowth = \frac{\text{SALE}}{\text{SALE}_{-12}}$$

    Sales growth is the percent change in net sales over turnover (Compustat item SALE).

21. **Leverage** (lev).

    Follows Bhandari (1988).

$$lev = \frac{\text{AT}}{\text{ME}_{\text{Dec}}}$$

    Market leverage is the ratio of total assets (Compustat item AT) over the market value of equity. Both are measured in December of the same year.

22. **Return on Assets (annual)** $(roma)$.

    Follows Chen, Novy-Marx e Zhang (2011).

$$roaa = \frac{\text{IB}}{\text{AT}}$$

    Net income scaled by total assets. Updated annually.

23. **Return on Equity (annual)** $(roea)$.

    Follows Haugen e Baker (1996).

$$roea = \frac{\text{IB}}{\text{BE}}$$

Net income scaled by book value of equity. Updated annually.

24. **Sales-to-Price** (sp).

    Follows Jr, Mukherji e Raines (1996).

$$sp = \frac{\text{SALE}}{\text{ME}_{\text{Dec}}}$$

    Total revenues divided by stock price. Updated annually.

25. **Growth in LTNOA** ($gltnoa$).

    Follows Fairfield, Whisenant e Yohn (2003).

$$gltnoa = \text{GRNOA} - \text{ACC}$$

    Growth in Net Operating Assets minus Accruals, where

$$\begin{aligned}
\text{GRNOA} &= \text{NOA} - \text{NOA}_{-12} \\
\text{NOA} &= \frac{\text{RECT} + \text{INVT} + \text{ACO} + \text{PPENT} + \text{INTAN} + \text{AO} - \text{AP} - \text{LCO} - \text{LO}}{\text{AT}} \\
\text{ACC} &= \frac{\Delta\text{RECT} + \Delta\text{INVT} + \Delta\text{ACO} - \Delta\text{AP} - \Delta\text{LCO} - \text{DP}}{(\Delta\text{AT}/2)} \\
\Delta\text{RECT} &= \text{RECT} - \text{RECT}_{-12} \\
\Delta\text{INVT} &= \text{INVT} - \text{INVT}_{-12} \\
\Delta\text{ACO} &= \text{ACO} - \text{ACO}_{-12} \\
\Delta\text{AP} &= \text{AP} - \text{AP}_{-12} \\
\Delta\text{LCO} &= \text{LCO} - \text{LCO}_{-12} \\
\Delta\text{AT} &= \text{AT} - \text{AT}_{-12}
\end{aligned}$$

    where RECH = Receivables, INVT = Total Inventory, ACO = Current Assets, AP = Accounts Payable, LCO = Current Liabilities (Other), DP = Depreciation

and Amortization, AT = Assets, PPENT = Property, Plant, and Equipment (net), INTAN = Intangible Assets, AO = Assets (Other), LO = Liabilities (Other). Updated annually.

26. **Momentum (6m)** (mom).

    Follows Jegadeesh e Titman (1993).

$$mom = \sum_{l=2}^{7} r_{t-l}$$

    Cumulated past performance in the previous 6 months by skipping the most recent month. Rebalanced monthly.

27. **Industry Momentum** ($indmom$).

    Follows Moskowitz e Grinblatt (1999).

$$indmom = \text{rank}\left(\sum_{l=1}^{6} r_{t-l}^{\text{ind}}\right)$$

    In each month, the Fama and French 49 industries are ranked on their value-weighted past 6-months performance. Rebalanced monthly.

28. **Value-Momentum** ($valmom$).

    Follows Novy-Marx (2013).

$$valmom = \text{rank}(value) + \text{rank}(mom)$$

    Sum of ranks in univariate sorts on book-to-market and momentum. Annual book-to-market values are used for the entire year. Rebalanced monthly.

29. **Value-Momentum-Profitability** ($malmomprof$).

    Follows Novy-Marx (2013).

$$valmomprof = \text{rank}(value) + \text{rank}(prof) + \text{rank}(mom)$$

Sum of ranks in umivariate sorts on book-to-market, profitability, and momentum. Annual book-to-market and profitability values are used for the entire year. Rebalanced monthly.

30. **Short Interest** ($shortint$).

    Follows Dechow, Kothari e Watts (1998).

$$shortint = \frac{\text{Shares Shorted}}{\text{Shares Outstanding}}$$

    Updated monthly.

31. **Momentum (1 year)** ($mom12$).

    Follows Jegadeesh e Titman (1993).

$$mom12 = \sum_{l=2}^{12} r_{t-l}$$

    Cumulated past performance in the previous year by skipping the most recent month. Rebalanced monthly.

32. **Momentum-Reversal** ($momrev$).

    Follows Jegadeesh e Titman (1993).

$$momrev = \sum_{d=14}^{19} r_{t-l}$$

    Buy and hold returns from $t-19$ to $t-14$. Updated monthly.

33. **Long-term Reversals** ($lrrev$).

    Follows Bondt e Thaler (1985).

$$lrrev = \sum_{l=13}^{60} r_{t-l}$$

    Cumulative returns from $t-60$ to $t-13$. Updated monthly.

34. **Value (monthly)** ($valuem$).

    Follows Asness e Frazzini (2013).

$$valuem = \frac{\text{BEQ}_{-3}}{\text{ME}_{-1}}$$

    Book-to-market ratio using the most up-to-date prices and book equity (appropriately lagged). Rebalanced monthly.

35. **Share Issuance (monthly)** ($nissm$).

    Follows Pontiff e Woodgate (2008).

$$nissm = \frac{\text{shrout}_{t-13}}{\text{shrout}_{t-1}}$$

    where shrout is the number of shares outstanding. Change in real number of shares outstanding from $t-13$ to $t-1$. Excludes changes in shares due to stock dividends and splits, and companies with no changes in shrout.

36. **PEAD (SUE)** (sue).

    Follows Foster, Olsen e Shevlin (1984).

$$sue = \frac{\text{IBQ} - \text{IBQ}_{-12}}{\sigma_{\text{IBQ}_{-24}:\text{IBQ}_{-3}}}$$

    where IBQ is income before extraordinary items (updated quarterly), and $\sigma_{\text{IBQ}_{-24}:\text{IBQ}_{-3}}$ is the standard deviation of IBQ in the past two years skipping the most recent quarter. Earnings surprises are measured by Standardized Unexpected Earnings (SUE), which is the change in the most recently announced quarterly earnings per share from its value announced four quarters ago divided by the standard deviation of this change in quarterly earnings over the prior eight quarters. Rebalanced monthly.

37. **Return on Book Equity** (roe).

    Follows Chen, Novy-Marx e Zhang (2011).

$$roe = \frac{\text{IBQ}}{\text{BEQ}_{-3}}$$

where IBQ is income before extraordinary items (updated quarterly), and BEQ is book value of equity. Rebalanced monthly.

38. **Return on Market Equity** ($rome$).

    Follows Chen, Novy-Marx e Zhang (2011).

$$rome = \frac{\text{IBQ}}{\text{ME}_{-4}}$$

    where IBQ is income before extraordinary items (updated quarterly), and ME is market value of equity. Rebalanced monthly.

39. **Return on Assets** (roa).

    Follows Chen, Novy-Marx e Zhang (2011).

$$roa = \frac{\text{IBQ}}{\text{ATC}_{-3}}$$

    Net income scaled by total assets. Updated quarterly.

40. **Short-term Reversal** ($strev$).

    Follows Jegadeesh (1990).

$$strev = r_{t-1}$$

    Return in the previous month. Updated monthly.

41. **Idiosyncratic Volatility** ($ivol$).

    Follows Ang et al. (2006).

$$ivol = \text{std}\left(R_{i,t} - \beta_i R_{M,t} - s_i SMB_t - h_i HML_t\right)$$

The standard deviation of the residual from firm-level regression of daily stock returns on the daily innovations of the Fama and French three-factor model using the estimation window of three months. Lagged one month.

42. **Beta Arbitrage** ($beta$).

    Follows Cooper, Gulen e Schill (2008).

$$beta = \beta_{t-60:t-1}$$

    Beta with respect to the CRSP equal-weighted return index. Estimated over the past 60 months (minimum 36 months) using daily data and lagged one month. Updated monthly.

43. **Seasonality** ($season$).

    Follows Heston e Sadka (2008).

$$season = \sum_{l=1}^{5} r_{t-l\,\mathrm{x}\,12}$$

    Average monthly return in the same calendar month over the last 5 years. As an example, the average return from prior Octobers is used to predict returns this October. The firm needs at least one year of data to be included in the sample. Updated monthly.

44. **Industry Relative Reversals** ($indrrev$).

    Follows Da, Liu e Schaumburg (2014).

$$indrrev = r_{-1} - r_{-1}^{ind}$$

    where r is the return on a stock and r^ind is return on its industry. Difference between a stocks' prior month's return and the prior month's return of its industry (based on the Fama and French 49 industries). Updated monthly.

45. **Industry Relative Reversals (Low Volatility)** ($indrrevlv$).

    Follows Da, Liu e Schaumburg (2014).

$$indrrevlv = r_{-1} - r_{-1}^{ind}$$

if vol ¡ NYSE median, where r is the return on a stock and rind is return on its industry. Difference between a stocks' prior month's return and the prior month's return of its industry (based on the Fama and French 49 industries). Only stocks with idiosyncratic volatility lower than the NYSE median for month are included in the sorts. Updated monthly.

46. **Industry Momentum-Reversal** ($indmomrev$).

    Follows Moskowitz e Grinblatt (1999).

$$indmomrev = \text{rank}(indmom) + \text{rank}(indrrevlv)$$

    Sum of Fama and French 49 industries ranks on industry momentum and industry relative reversals (low vol). Rebalanced monthly.

47. **Composite Issuance** ($ciss$).

    Follows Daniel e Titman (2006).

$$ciss = \log\left(\frac{\text{ME}_{t-13}}{\text{ME}_{t-60}}\right) - \sum^{6} 0_{l=13} r_{t-l}$$

    where r is the log return on the stock and ME is total market equity. Updated monthly.

48. **Price** ($price$).

    Follows Blume e Husic (1973).

$$price = \log\left(\frac{\text{ME}}{\text{shrout}}\right)$$

    where ME is market equity and shrout is the number of shares outstanding. Log of stock price. Updated monthly.

49. **Firm Age** (age).

Follows Barry e Brown (1984).

$$age = \log(1 + \text{number of months since listing})$$

The number of months that a firm has been listed in the CRSP database.

50. **Share Volume** ($shvol$).

    Follows Datar, Naik e Radcliffe (1998).

$$shvol = \frac{1}{3}\sum_{i=1}^{3}\frac{\text{volume}_{t-i}}{\text{shrout}_t}$$

    Average number of shares traded over the previous three months scaled by shares out-standing. Updated monthly.

51. **Cash flow duration** (dur).

    Follows ?.

$$dur = \frac{\sum_t \text{PV}_0\left(t \times \text{CF}_t\right)}{P_0}$$

    Present value of expected cashflows. Cashflows' components (from clean surplus identity, ROE and book equity growth) are forecasted using AR(1). Sums are discounted using a constant discount rate. Rebalanced monthly.

This database, which we will call Factors, contains daily data from January 2005 to December 2019.

Figure 1 presents a graph that illustrates the intersection of firms between the Factors Database and the TAQ Returns Database, during the period from January 2005 to December 2019. The graph depicts the evolution of the number of stocks present in each of the bases, as well as the quantity of stocks in which there is a match in both bases and that will be used for the formation of the portfolios.

Using the TAQ Returns and Factors databases, it is possible to create portfolios of various financial factors that are widely used in the literature. However, the construction of these portfolios can vary greatly between studies. To ensure standardization in the creation of these portfolios, we adopted four different strategies. The first two

strategies consist of forming portfolios by longing (shorting) stocks that are above (below) the median. The difference between these strategies will be the way of weighting the assets, which can be equal weighted or value weighted. The other two methodologies will use the 30th and 70th percentiles as breakpoints for stock selection, that is, by longing (shorting) stocks in the upper tertile (lower tertile).

3 Conclusion

The main objective of this summer paper was to document the data manipulation process. We describe how we create portfolios related to some of the key financial factors documented in the literature. The final database contains daily intraday stock returns and financial factor portfolios.

In Section 1, we briefly mentioned the goals after finalizing the database. The intention is to model through LASSO which predictor candidates among stock returns and portfolios of financial factors help to improve the predictive power of high-frequency asset pricing. Along the same line, we will be able to verify whether, unlike Chinco, Clark-Joseph e Ye (2019), when including financial factors there is an increase in predictive power.

Next steps include adding more robust methods to evaluate which predictor candidates increase predictive power, such as the Double Selection LASSO featured in Feng, Giglio e Xiu (2020). This procedure considers model selection errors by selecting predictors that not only help explain the cross-section of expected returns, but are also useful in mitigating the problem of omitted variable bias.

Furthermore, as LASSO is a tool whose predictor selection is based purely on a statistical rule, we would like to check whether there is economic meaning for the selected predictors through twitter data.

References

ANG, A. et al. The cross-section of volatility and expected returns. *The journal of finance*, Wiley Online Library, v. 61, n. 1, p. 259–299, 2006.

ASNESS, C.; FRAZZINI, A. The devil in hml's details. *The Journal of Portfolio Management*, Institutional Investor Journals Umbrella, v. 39, n. 4, p. 49–68, 2013.

BARRY, C. B.; BROWN, S. J. Differential information and the small firm effect. *Journal of financial economics*, Elsevier, v. 13, n. 2, p. 283–294, 1984.

BASU, S. Investment performance of common stocks in relation to their price-earnings ratios: A test of the efficient market hypothesis. *The journal of Finance*, Wiley Online Library, v. 32, n. 3, p. 663–682, 1977.

BHANDARI, L. C. Debt/equity ratio and expected common stock returns: Empirical evidence. *The journal of finance*, Wiley Online Library, v. 43, n. 2, p. 507–528, 1988.

BLUME, M. E.; HUSIC, F. Price, beta, and exchange listing. *The Journal of Finance*, JSTOR, v. 28, n. 2, p. 283–299, 1973.

BONDT, W. F. D.; THALER, R. Does the stock market overreact? *The Journal of finance*, Wiley Online Library, v. 40, n. 3, p. 793–805, 1985.

CARHART, M. M. On persistence in mutual fund performance. *The Journal of finance*, Wiley Online Library, v. 52, n. 1, p. 57–82, 1997.

CHEN, L.; NOVY-MARX, R.; ZHANG, L. An alternative three-factor model. *Available at SSRN 1418117*, 2011.

CHINCO, A.; CLARK-JOSEPH, A. D.; YE, M. Sparse signals in the cross-section of returns. *The Journal of Finance*, Wiley Online Library, v. 74, n. 1, p. 449–492, 2019.

COOPER, M. J.; GULEN, H.; SCHILL, M. J. Asset growth and the cross-section of stock returns. *the Journal of Finance*, Wiley Online Library, v. 63, n. 4, p. 1609–1651, 2008.

DA, Z.; LIU, Q.; SCHAUMBURG, E. A closer look at the short-term return reversal. *Management science*, INFORMS, v. 60, n. 3, p. 658–674, 2014.

DANIEL, K.; TITMAN, S. Market reactions to tangible and intangible information. *The Journal of Finance*, Wiley Online Library, v. 61, n. 4, p. 1605–1643, 2006.

DATAR, V. T.; NAIK, N. Y.; RADCLIFFE, R. Liquidity and stock returns: An alternative test. *Journal of financial markets*, Elsevier, v. 1, n. 2, p. 203–219, 1998.

DECHOW, P. M.; KOTHARI, S. P.; WATTS, R. L. The relation between earnings and cash flows. *Journal of accounting and Economics*, Elsevier, v. 25, n. 2, p. 133–168, 1998.

FAIRFIELD, P. M.; WHISENANT, J. S.; YOHN, T. L. Accrued earnings and growth: Implications for future profitability and market mispricing. *The accounting review*, v. 78, n. 1, p. 353–371, 2003.

FAMA, E. F.; FRENCH, K. R. The cross-section of expected stock returns. *the Journal of Finance*, Wiley Online Library, v. 47, n. 2, p. 427–465, 1992.

FAMA, E. F.; FRENCH, K. R. Common risk factors in the returns on stocks and bonds. *Journal of financial economics*, Elsevier, v. 33, n. 1, p. 3–56, 1993.

FAMA, E. F.; FRENCH, K. R. A five-factor asset pricing model. *Journal of financial economics*, Elsevier, v. 116, n. 1, p. 1–22, 2015.

FENG, G.; GIGLIO, S.; XIU, D. Taming the factor zoo: A test of new factors. *The Journal of Finance*, Wiley Online Library, v. 75, n. 3, p. 1327–1370, 2020.

FOSTER, G.; OLSEN, C.; SHEVLIN, T. Earnings releases, anomalies, and the behavior of security returns. *Accounting Review*, JSTOR, p. 574–603, 1984.

HAUGEN, R. A.; BAKER, N. L. Commonality in the determinants of expected stock returns. *Journal of financial economics*, Elsevier, v. 41, n. 3, p. 401–439, 1996.

HESTON, S. L.; SADKA, R. Seasonality in the cross-section of stock returns. *Journal of Financial Economics*, Elsevier, v. 87, n. 2, p. 418–445, 2008.

HIRSHLEIFER, D. et al. Do investors overvalue firms with bloated balance sheets? *Journal of Accounting and Economics*, Elsevier, v. 38, p. 297–331, 2004.

IKENBERRY, D.; LAKONISHOK, J.; VERMAELEN, T. Market underreaction to open market share repurchases. *Journal of financial economics*, Elsevier, v. 39, n. 2-3, p. 181–208, 1995.

JEGADEESH, N. Evidence of predictable behavior of security returns. *The Journal of finance*, Wiley Online Library, v. 45, n. 3, p. 881–898, 1990.

JEGADEESH, N.; TITMAN, S. Returns to buying winners and selling losers: Implications for stock market efficiency. *The Journal of finance*, Wiley Online Library, v. 48, n. 1, p. 65–91, 1993.

JR, W. C. B.; MUKHERJI, S.; RAINES, G. A. Do sales–price and debt–equity explain stock returns better than book–market and firm size? *Financial Analysts Journal*, Taylor & Francis, v. 52, n. 2, p. 56–60, 1996.

LAKONISHOK, J.; SHLEIFER, A.; VISHNY, R. W. Contrarian investment, extrapolation, and risk. *The journal of finance*, Wiley Online Library, v. 49, n. 5, p. 1541–1578, 1994.

LINTNER, J. Security prices, risk, and maximal gains from diversification. *The journal of finance*, JSTOR, v. 20, n. 4, p. 587–615, 1965.

MARKOWITZ, H. Portfolio selection. *Journal of Finance*, v. 7, n. 1, p. 77–91, 1952.

MOSKOWITZ, T. J.; GRINBLATT, M. Do industries explain momentum? *The Journal of finance*, Wiley Online Library, v. 54, n. 4, p. 1249–1290, 1999.

MOSSIN, J. Equilibrium in a capital asset market. *Econometrica: Journal of the econometric society*, JSTOR, p. 768–783, 1966.

NARANJO, A.; NIMALENDRAN, M.; RYNGAERT, M. Stock returns, dividend yields, and taxes. *The Journal of Finance*, Wiley Online Library, v. 53, n. 6, p. 2029–2057, 1998.

NOVY-MARX, R. The other side of value: The gross profitability premium. *Journal of financial economics*, Elsevier, v. 108, n. 1, p. 1–28, 2013.

PIOTROSKI, J. D. Value investing: The use of historical financial statement information to separate winners from losers. *Journal of Accounting Research*, JSTOR, p. 1–41, 2000.

PONTIFF, J.; WOODGATE, A. Share issuance and cross-sectional returns. *The Journal of Finance*, Wiley Online Library, v. 63, n. 2, p. 921–945, 2008.

SHARPE, W. F. Capital asset prices: A theory of market equilibrium under conditions of risk. *The journal of finance*, Wiley Online Library, v. 19, n. 3, p. 425–442, 1964.

SLOAN, R. G. Do stock prices fully reflect information in accruals and cash flows about future earnings? *Accounting review*, JSTOR, p. 289–315, 1996.

SOLIMAN, M. T. The use of dupont analysis by market participants. *The accounting review*, v. 83, n. 3, p. 823–853, 2008.

SPIESS, D. K.; AFFLECK-GRAVES, J. The long-run performance of stock returns following debt offerings. *Journal of Financial Economics*, Elsevier, v. 54, n. 1, p. 45–73, 1999.

XING, Y. Interpreting the value effect through the q-theory: An empirical investigation. *The Review of Financial Studies*, Society for Financial Studies, v. 21, n. 4, p. 1767–1795, 2008.

A Appendix

A.1 LASSO

LASSO is a categorized procedure within the set of penalized regressions. Penalized Least Squares is a punishment procedure that adds to the OLS regression a component that penalizes weak coefficients. The Penalty Least Squares estimator is obtained through

$$\widehat{\boldsymbol{\beta}}(\lambda) = \underset{\boldsymbol{\beta}\in\mathcal{B}}{\arg\min} \left[\sum_{i=1}^{n} \left(Y_i - \boldsymbol{\beta}' \boldsymbol{X}_i\right)^2 + \sum_{j=1}^{p} p_\lambda \left(|\beta_j| ; \boldsymbol{\alpha},\ \text{data}\ \right) \right]$$

where $p_\lambda\left(|\beta_j| ; \boldsymbol{\alpha}, \text{data}\right)$ is a non-negative penalty function indexed by the regularization parameter λ and could rely on both the data and additional hyperparameters. For example,

- Ridge

  $p_\lambda \left(|\beta_j| ; \boldsymbol{\alpha}, \text{data}\ \right) = \lambda \left|\beta_j\right|^2$

- LASSO

  $p_\lambda \left(|\beta_j| ; \boldsymbol{\alpha}, \text{data}\ \right) = \lambda \left|\beta_j\right|$

- LASSO and Ridge combined

  $p_\lambda \left(|\beta_j| ; \boldsymbol{\alpha}, \text{data}\ \right) = \alpha\lambda \left|\beta_j\right| + (1-\alpha)\lambda \left|\beta_j\right|^2$

The regularization parameter λ controls the number of parameters in the model. If $\lambda = \infty$, then no parameters enter the model, and if $\lambda = 0$, then the parameters are simply OLS estimators.

A.2 Figures

Figure 1: Number of stocks in each Database