Handelshøyskolen BI

GRA 19703 Master Thesis

Thesis Master of Science 100% - B

Predefinert informasjon

Startdato:	09-01-2023 09:00 CET	**Termin:**	202310
Sluttdato:	03-07-2023 12:00 CEST	**Vurderingsform:**	Norsk 6-trinns skala (A-F)
Eksamensform:	T		
Flowkode:	202310||11378||IN00||B||T		
Intern sensor:	(Anonymisert)		

Deltaker

Navn:	Lars Veseth Kristoffersen og Ole Gunnar Bruheim

Informasjon fra deltaker

Tittel *:	Unveiling the Dynamics of Stock Illiquidity: Exploring Key Stock Characteristics and their Significance
Navn på veileder *:	Stephen Walter Szaura

Inneholder besvarelsen konfidensielt materiale?:	Nei	**Kan besvarelsen offentliggjøres?:**	Ja

Gruppe

Gruppenavn:	(Anonymisert)
Gruppenummer:	8
Andre medlemmer i gruppen:	

Unveiling the Dynamics of Stock Illiquidity: Exploring Key Stock Characteristics and their Significance

Master Thesis

by

Lars Veseth Kristoffersen and Ole Gunnar Bruheim

MSc in Business

Bergen, July 03, 2023

Supervisor:

Stephen Walter Szaura, Accociate Professor

Bi Norwegian Business School

This thesis is a part of the MSC program at BI Norwegian Business School. The school takes no responsibility for the methods used, results found, or conclusions drawn.

Contents

Abstract

This paper explores the use of machine learning models to predict what characteristics affect illiquidity in stocks using historical data. The paper uses thirteen different regressions, exploring the effects of 43 characteristics. The regressions are run with and without the variable bid-ask spread. The in-sample findings suggest that the oracle, group lasso and enet regressions are outperforming the OLS regression both with and without bid-ask spread. Bid-ask spread is seen to be the variable with the highest correlation in the out of sample analysis. The regressions without bid-ask spread show more variance in the results also showing the variables BM and VolMkt to be most correlated. Concluding that the bid-ask spread is the most correlated characteristic.

Acknowledgements

We would like the express our deepest gratitude to our thesis supervisor, Associate Professor Stephen Walter Szaura, from the department of finance at BI Norwegian Business School. He has provided us with invaluable guidance with your expertise and through insightful feedback and constructive criticism. We deeply appreciate his contribution.

Furthermore we deeply appreciate Dacheng Xiu's invaluable support in our regression analysis. It was with great skill that Mr. Xiu developed the Matlab script that we utilized for our regression work.

1.0 Introduction

Since the inception of competitive stock exchanges, speculators and analysts have wanted higher returns in constantly changing stock/market trading conditions. In this thesis, we will measure what characteristics correlate with/affect stock illiquidity the best. The thesis provides results suggesting that bid-ask spread is the most significant, and without the bid-ask spread, the BM and VolMkt are the most significant characteristics. Stock illiquidity measures the relative ease or difficulty of trading a particular stock. It is calculated by dividing the average daily volume of the investment by the absolute value of the daily return of the asset. An illiquidity value suggests it is more difficult to trade the asset, as it may take longer to find a buyer or seller, and the bid-ask spread is larger. This can make buying or selling large quantities of the asset more challenging. Illiquidity between stock returns and trading volume occurs when a stock cannot comfortably and swiftly be sold or exchanged for cash without a considerable loss in value. Knowing about this is important because transaction costs are an essential part of trading, and getting a better understanding of what stock and what predicts illiquidity based on characteristics using non-linear relationships. Since we expect the bid-ask spread to have the highest correlation, we will also have regressions without the bid-ask spread. The reason for using machine learning techniques is that stock illiquidity is non-linear, and the impact of characteristics is non-linear, so linear regression may not be appropriate. Research on this topic is essential to literature because it increases the amount and takes a new perspective. Specifically, we will ask:

“What are the key stock characteristics that significantly influence stock illiquidity?"

Amihud (2002) measures the ease or difficulty of trading a particular financial asset. This information can prove valuable to investors and traders who want to evaluate the liquidity of various assets and determine how it could affect their trading strategies. The transaction costs are dependent on the market type and trade size. One challenge experienced by major players is that their high market influence results in high transaction costs. In such cases, splitting the trade into smaller orders and executing them over time is advisable.

On the other hand, over-the-counter markets require one to call a dealer on the phone to complete the trade. In this case, it is best to trade in chunks worth the dealer's time and get competitive bids by calling several dealers. (Amihud, 2002)

Liquidity is an essential stock characteristic to consider when investing or trading because it can impact the price at which an asset can be bought or sold and the speed at which trades can be executed. Highly liquid assets are generally easier to trade and may be more suitable for investors who need to buy or sell large quantities of the investment quickly. On the other hand, less liquid assets may be more challenging to trade. They may be more suitable for investors who are willing to accept the additional risk and reduced trading flexibility in exchange for the potential for higher returns. Amihud illiquidity is just one of many measures that can be used to assess liquidity. For our thesis, we measure price impact, in which Amihud has been shown to be the best. It can also be useful for understanding the liquidity characteristics of different assets and how they may impact trading and investment decisions.

When selecting stocks, it is crucial to consider stock illiquidity for various reasons. Firstly, low trading volumes make it difficult for investors to quickly enter or exit positions, resulting in significant price impacts and a need for counterparties to transact with. Secondly, illiquid stocks often have wider bid-ask spreads, leading to higher transaction costs and reducing potential profits for investors. Additionally, limited trading activity can result in significant price fluctuations, making it challenging to accurately assess the stock's actual value, heightening the investment risk. Understanding stock illiquidity is a vital measure to consider when choosing stocks.

Additionally, illiquid stocks often have less information avaliable. Analyst Coverage, financial news, and market participants often overlook them. More information is required to perform a thorough fundamental analysis and gain insights into the company's prospects, amplifying investment uncertainty. Lastly, illiquid stocks may not be suitable for investors with short investment horizons or those who require liquidity. These stocks restrict the ability to adjust portfolios or capitalize on new investment opportunities as they arise. While illiquid stocks can present unique investment opportunities, they necessitate careful consideration of the factors above, along with higher risk tolerance and potential limitations in executing trades.

Machine learning has been studied in numerous papers. Gu et al. (2020) explore the application of machine learning techniques in asset pricing. They try to investigate if machine learning models can give insights into asset pricing and the accuracy of return forecasts. Machine learning can uncover non-linear relationships through learning algorithms that traditional methods cannot. In our thesis, we will use similar machine learning techniques and regression as Gu et al. (2020) to find the different characteristics that predict the illiquidity in stocks. The selected machine learning techniques for the thesis, includes Simple OLS R2, PCR R2, PLS R2, Lasso R2, Ridge R2, Enet R2, Oracle R2, and Group Lasso R2, are applicable due to their diverse capabilities in regression analysis. They cover a range of scenarios such as linear regression (OLS), addressing multicollinearity (PCR, PLS), variable selection (Lasso, Oracle), regularization (Ridge, Enet), and handling grouped variables (Group Lasso). This comprehensive set of techniques allows us to explore different aspects of the data, such as model interpretability, predictive performance, variable importance, and handling specific challenges like multicollinearity or irregular grouping.

When choosing a stock to invest in, it is important focusing on different factors that affect the return. Through several research articles such as (Amihud, 2002), illiquidity has become a measure of how to predict share returns. The relationship between stock returns and illiquidity is displayed through higher returns for stocks with higher illiquidity. Also, stocks that are expected to be sensitive to changes in liquidity should concede a higher return to compensate for the risk. Understanding and using the relationship between return and illiquidity makes this an essential measure in choosing a stock with a high expected return concerning risk. For our thesis's chosen characteristics, we selected them to assess their predictive power for illiquidity and diversified our choices to ensure representative research. The results show the importance of the bid-ask spread and its most significant characteristics in the regressions. When the regressions are run without the bid-ask spread, there is more variance in our variables. BM and VolMkt are the most distinct characteristics in the regressions without the bid-ask spread. The regressions show that the Oracle, Enet, and Lasso regressions are more distinct than the OLS regressions.

The rest of this thesis has a six-part structure. Section two will contain the related literature. The third section is where we will present the methodology of our research, and the part is used to go into detail and explain why and how our study is done. The fourth section will present the data used in our research. Section five is where we will analyze the results. The last section, section 6, will proceed to the conclusion of our master thesis. After the decision and master's thesis, we will find a bibliography and appendix.

2.0 Literature Review and Expectations

Amihud (2002) is widely known as the measure of stock illiquidity. Plenty of research uses Amihud’s method, and plenty of academic papers are based on his theory. The paper examines the relationship between stock illiquidity, the difficulty in buying or selling a stock, and future stock returns. The paper looks at both cross-sectional and time-series effects of illiquidity on stock returns.

According to Amihud (2002), stocks with higher illiquidity levels tend to have higher returns, referred to as the "illiquidity premium." Investors need a higher return to offset illiquid stocks' potential risk and trading costs.

In general, Amihud (2002) emphasizes the significance of illiquidity in assessing the performance of stocks and the stock market. The illiquidity measures presented in the paper could aid us in investigating our research question regarding the factors that impact illiquidity. The critical difference between our paper and the one cited is that it examines the influence of illiquidity on returns by measuring it first and then determining the correlation. On the other hand, we will utilize the measure to consider the characteristics that contribute to illiquidity and evaluate their impact on it.

The Amihud illiquidity measure bases itself on the average ratio of the absolute return of the day (Riyd) divided by the day's volume (VOLDivyd). This calculation can be reformulated into an equation using machine learning techniques to determine the characteristics that impact illiquidity. By measuring these variables, we can determine the level of illiquidity.

Goyenko (et al.,2009) examine whether the standard liquidity measures can fully capture the liquidity of a market or asset. The article proposes that these measures may only sometimes be reliable liquidity indicators and that additional research is required for more precise measurements. Goyenko (et al., 2009) also explore the significance of liquidity in empirical finance by introducing and testing novel liquidity measures compared to commonly used measures in previous studies.

Unlike previous studies that analyze the correlation between security returns and liquidity measures, (Goyenko et al., 2009) take a unique approach by evaluating the connection between the suggested liquidity measures and tangible transaction costs. Previous research tends to rely on liquidity measures based on daily return and volume data as a proxy for investor liquidity and transaction costs without conducting direct tests on their relationship with actual trading costs. In contrast, (Goyenko et al., 2009) aims to test whether the proposed liquidity proxies accurately represent transaction costs by comparing them against factual trading data such as effective and realized spreads.

Using new and widely employed measures in the literature, running horse races of annual and monthly estimates of each measure against liquidity benchmarks (Goyenko et al., 2009). Benchmarks are effective spread, realized spread, and price impact based on Trade and Quote (TAQ) and Rule 605 data. They are finding that the new effective/realized spread measures win most horse races, while the Amihud paper (Amihud, Y. 2002) measure does best in price impact. Goyenko et al.(2009) created a proxy based on (Amihud, 2002), and the proxy proved the best results in measuring price impact and proving that the measure is the most relevant illiquidity measure for our thesis.

Dynamic trading involves analyzing whether traders can accurately forecast the returns on their traded assets. The research (Garleanu & Pedersen, 2013), has demonstrated that a threshold-based approach is the best trading strategy when considering transaction costs and predictable returns. This strategy involves purchasing (or selling) assets only when their returns exceed (or fall below) a specific threshold.

Garleanu and Pedersen's (2013) paper delves into the realm of dynamic trading and assesses the proficiency of traders in predicting asset returns. Through their research, it has been discovered that a threshold-based approach is the most effective trading strategy, considering predictable returns and transaction costs. This approach entails purchasing assets solely when their returns surpass a particular threshold and vending them when returns decline below another threshold. Garleanu and Pedersen's (2013) paper presents a unique perspective on trading in a dynamic environment with predictable returns and transaction costs. The paper's insights are valuable to traders, investors, and researchers seeking to understand how to navigate such markets effectively.

Kaniel et al. (2023) conducted a study utilizing machine learning methods to predict the performance of mutual funds based on various attributes. The study found that the flow and momentum of funds were the primary factors for outperformance, while the characteristics of the stocks held by the funds had little impact. Interestingly, the study also revealed interaction effects between fund characteristics and investor sentiment, highlighting the significance of non-linear models in capturing complex relationships. These discoveries have significant implications for delegation theories in the mutual fund market and stress the potential benefits of avoiding underperforming funds.

Gu (et al.,, 2020) study aims to assess and compare multiple machine-learning approaches in predicting asset returns. The results highlight the potential of machine learning techniques to improve our understanding of asset prices.

Interestingly, the study has uncovered that "shallow" learning techniques, which utilize models with fewer layers or simpler architectures, tend to perform better than "deep" learning techniques (Gu et al., 2020). This finding differs from observations in other areas, such as computer vision or bioinformatics. It is likely due to the limited data availability and the low signal-to-noise ratio inherent in asset pricing problems.

In addition, the research emphasizes the usefulness of machine learning techniques in predicting returns for more extensive, more easily tradable stocks and building optimized portfolios (Gu et al., 2020). By utilizing machine learning capabilities, investors and portfolio managers can make better-informed choices that may increase returns.

Gu (et al., 2020) study has analyzed various machine learning models and has identified a small but significant set of predictive signals that consistently emerge as solid indicators of asset returns. These signals include price trends like return reversal and momentum, measures of stock liquidity, volatility, and valuation ratios. These findings offer valuable insights into the factors that drive asset prices and can aid in refining economic models and enhancing risk measurement techniques.

Not only does machine learning advance our understanding of asset pricing, but it also has practical implications for return prediction. By minimizing approximation and estimation errors, machine learning techniques can provide more precise measurements of risk premiums. This leads to easier identification of reliable economic mechanisms behind asset pricing phenomena, allowing for the development of more robust economic models and informed portfolio choices.
(Gu et al., 2020)

Gu (et al., 2020) research findings show that machine learning is increasingly important in financial technology (fintech). The proven success of machine learning in predicting asset returns justifies its use throughout the architecture of fintech applications. These findings are valuable for academic research and have practical implications for portfolio management, investment strategies, and decision-making in the financial industry.

Jensen et al.'s (2022) research combine machine learning with portfolio selection to improve investment portfolios by considering transaction expenses and expected returns based on security features. Their technique significantly expands the implementable efficient frontier and offers a new perspective on the importance of securities when optimizing a portfolio. Their proposed method offers a practical and effective way to optimize investment portfolios by factoring in expected returns and transaction costs.

3.0 Methodology

We utilize the widely known illiquidity measure from Amihud (2002) to determine which characteristics affect illiquidity. Contrary to the Amihud (2002) measure, which aims to find the correlation of illiquidity to stock returns, we utilize machine learning and change the equation to fit our research.

$$ILLIQ_{iy} = 1/D_{iy} \sum_{T=1}^{Diy} |\mathrm{R}_{iyd}| \,/VOLD_{ivyd,} \tag{1}$$

The equation is the measure of illiquidity used by (Amihud, 2022). The calculation of illiquidity, which we can reformulate into an equation using machine learning techniques and use to find which characteristics that affect illiquidity. Amihud (2002) uses the average ratio of the absolute return of the day (Riyd) against the volume of the day (VOLDivyd). Illiquidity is measured using these variables. Where Diy is the number of days for which data are available for stock I in the year.

When finding characteristics affecting illiquidity, we will use machine learning. Gu et al. (2020) explore different measures using machine learning. Since we will use machine learning to find characteristics that affect illiquidity, we will combine the machine learning techniques with illiquidity measures. Gu et al. (2020) follow a linear predictive regression.

$$L(\theta) = \frac{1}{NT} \sum_{i=1}^{N} (r_{i,t+1} - g(z_{i,t};\theta))^2 \tag{2}$$

We then combine the equation to measure illiquidity with the regression model for machine learning and can then estimate the equation. In the paper of Amihud, he explores the relationship between illiquidity and returns. At the same time, we want to redefine this and use the measure instead to find the characteristics that affect illiquidity. We will put in the equation used in (Amihud, 2002) and replace the returns (Gu et al.,2020)

The simple linear model assumes that the expected value of a variable can be represented as a linear combination of the predictor variables and a parameter vector. Gu et al.(2020) show that this model is limited because it does not allow for interactions or non-linear effects between predictors.

A least squares objective function is used to estimate the model following Freedman (2009), which minimizes the difference between the actual and predicted values from the linear model. Using Freedman's (2009) approach is convenient because it provides analytical estimates and does not require complex optimization or computation. The estimated parameters are referred to as the pooled OLS estimator.

$$L(\theta)=\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}((\frac{1}{D_{iy}}\sum_{T=1}^{D_{iy}}\left|R_{iyd}\right|/VOLD_{ivyd})_{t+1}-g(z_{i,t}-g(z_{i,t};\theta))^{2} \quad (3)$$

The equation uses a standard least squares objective function. We will use the equation that we have to use machine learning while measuring. The predictors also cannot allow us to affect each other, and if they affect each other, the research findings might be wrong or be affected by not the factors we find.

The equation (Amihud,2002)(1) is our dependent variable in all the regressions. The term for equation (1) is $ILLIQ_{iy}$ will therefore be the variable we change the different r variables to.

$$Lw(\theta)=\frac{1}{NT}\sum_{t=1}^{N}\sum_{t=1}^{T}w_{i,t}(ILLIQ_{iy,t+1}-g(z_{i,t};\theta))^{2} \quad (4)$$

The weighted least squares objective function is an alternative to the traditional one used in regression analysis. It involves assigning weights to each observation based on their statistical or economic significance. By doing so, the econometrician can emphasize specific comments over others, thereby improving the model’s predictive performance. Gu et al.(2020) provide two variations of weighted least squares: one based on the number of stocks available at a given time and the other based on the equity market value of each store. These variations allow for equal weighting or value weighting of the squared loss of stocks.

The motivation behind weighted least squares is to address the issue of heavy-tailed distributions often observed in financial data. Gu et al.(2020) refer to heavy tails as the presence of extreme observations or outliers that can disproportionately impact the model's predictions. The convexity of the traditional least squares objective function places a high emphasis on significant errors, making the model sensitive to outliers. Therefore we will deploy a robust loss function.

$$L_H(\theta) = \frac{1}{NT}\sum_{t=1}^{N}\sum_{t=1}^{T} H\ (ILLIQ_{iy,t+1} - g(z_{i,t};\theta))^2 \qquad (5)$$

The Huber robust loss function is commonly used in machine learning to handle heavy-tailed observations. Huber (1964) combines squared loss for relatively small errors and total defeat for rather large mistakes. The tuning parameter ξ controls the transition point between the two loss functions and can be optimized from the data. Huber (1981) tells of the advantage of the Huber loss function: it is more robust to extreme observations, providing more stable forecasts than the traditional least squares method.

The H equation for robustness is defined as:

$$H(x;\xi) = \{^{x^2, if\,|\mathrm{x}|<\xi;}_{2\xi|\mathrm{x}|-\xi^2, if\,|\mathrm{x}|>\xi.} \qquad (6)$$

There are challenges using a basic linear model when there are many predictors compared to the number of observations. In such situations, the linear model performs poorly as it starts with proper noise instead of the actual signal (Gu et al., 2020). This is a common problem when predicting returns, where helpful information is often mixed with much noise.

To overcome noise issues, regularization techniques are employed. Following (Tibshirani, 1996), regularization involves adding a penalty term to the model's objective function, encouraging simpler models with fewer predictors. Although this penalty may worsen the model's performance on the data used for training (in-sample), it helps the model generalize better to new, unseen data (out-of-sample) by reducing the impact of noise while still capturing the critical patterns.

$$\mathrm{L}(\theta;\cdot) = \mathrm{L}(\theta) + \phi(\theta;\cdot) \tag{7}$$

The specific regularization method used in this paper is called the net elastic penalty. Zou & Hastie (2004) explains how it has two parameters, λ and ρ, which control the amount and type of penalty applied. Further, the elastic net combines two well-known regularization methods: the lasso and ridge regression. The lasso sets some coefficients to precisely zero, effectively performing variable selection and keeping only the most important predictors. Tibshirani (1996) tells how on the other hand, ridge regression shrinks the coefficients toward zero without setting any to precisely zero, which helps prevent coefficients from becoming too large.

$$\phi(\theta;\lambda,\rho) = \lambda(1-\rho)\sum_{j=1}^{P}\left|\theta_j\right| + \frac{1}{2}\lambda\rho\sum_{j=1}^{P}\theta_j^2 \tag{8}$$

An adaptive optimization process determines the best values for λ and ρ. The tuning parameters are optimized by evaluating the model's performance on a separate validation sample (Huber,1981). This allows the model to find the right balance between simplicity and capturing the signal.

Dimension reduction techniques, such as Principal Components Regression (PCR) and Partial Least Squares (PLS), address this issue by forming linear combinations of predictors (Gu et al.,2020). These techniques help reduce noise and decorrelate highly dependent predictors, thus isolating the signal in the predictors more effectively. (Geladi & Kowalski, 1986)

PCR involves a two-step process. In the first step, Principal Components Analysis (PCA) combines the predictors into smaller linear combinations that preserve the covariance structure among the predictors (Livak & Schmittgen, 2001). A few leading components are used in standard regression in the second step. PCR achieves regularization by setting coefficients on low-variance parts to zero.

Papers such as (Mullis & Faloona, 1987) (Livak & Schmittgen, 2001), and (Gu et al.,2020) have discussed different drawbacks of PCR. One drawback of PCR is that it does not explicitly consider the forecasting objective when reducing dimensionality. It condenses the data into components based on the covariation among predictors without considering their association with future returns.

In contrast, (PLS) directly exploits predictors' and forecast targets' covariation (Gu et al.,2020). PLS regression proceeds by estimating each predictor's univariate return prediction coefficient using (OLS) (Wold, 1966). These coefficients reflect the sensitivity of returns to each predictor. PLS then averages all predictors into a single aggregate component, giving higher weights to stronger univariate predictors and lower consequences to weaker ones. This dimension reduction is performed considering the forecasting objective (Geladi & Kowalski, 1986). To create more than one predictive component, the target and predictors are orthogonalized concerning previously constructed features, and the process is repeated until the desired number of PLS components is achieved. (Gu et al.,2020)

Implementing both PCR and PLS starts with a vectorized version of the linear model and rearranges it into a matrix form. Following Gu et al. (2020), the dimension-reduced predictor set is represented by a matrix ΩK, and the forecasting model is written as a regression of the target variable on ZΩK, where Z is the original predictor matrix. The choice of the combination weights in PCR and PLS is determined by solving optimization problems that either maximize the variation retained by the predictors (PCR) or the predictive association with the target (PLS). The coefficient vector θK is estimated using OLS regression.

When using Pls, we change the original equations.

$$ILLIQ_{iy} = 1/D_{iy} \sum_{T=1}^{Diy} |\mathrm{R}_{iyd}|\,/VOLD_{ivyd,} \quad {}_{i,t+1} = z'_{i,t}\theta + \epsilon_{i,t+1} \tag{9}$$

$$ILLIQ_{iyi,t+1} = z'_{i,t}\theta + \epsilon_{i,t+1} \tag{10}$$

The equation becomes:

$$ILLIQ = Z\theta + E \tag{11}$$

The A is the NTx1 matrix created by R_(i,t+1). Z is a matrix with dimensions NT×P, where NT represents the number of observations and P represents the number of predictors (De Jong, 1993). Each row of Z corresponds to a stacked predictor zi,t. On the other hand, E is a vector of residuals with dimensions NT×1, where each element represents the residual ϵi,t+1 for a particular observation.

PCR and PLS are techniques used to reduce the dimensionality of a set of predictors. The goal is to transform many predictors (dimension P) into a smaller group of linear combinations (dimension K) (Gewaldi, 1986) (Mullis & Faloona, 1987).

In PCR and PLS, the predictors are combined to capture the essential information while discarding some less relevant or redundant information as found in (De Jong, 1993). This condensation process helps simplify the data and makes it more manageable for analysis.
The resulting condensed set of predictors is used to build a forecasting model. This model represents the relationship between the predictors and the target variable. (Gu et al., 2020)

$$\mathrm{A} = (\mathrm{Z}\Omega\mathrm{K})\theta\mathrm{K} + \tilde{\mathrm{E}} \tag{12}$$

ΩK represents a matrix with dimensions P×K, where each column (w1, w2, ..., wK) corresponds to a set of weights for creating the predictive components. These weights determine the linear combinations used to generate the reduced version of the original location of predictors, denoted as ZΩK (De Jong, 1993). In this condensed version, the predictors are transformed into a lower-dimensional representation

PCR employs an iterative process to determine the combination weights, ΩK. At each iteration, the jth linear combination is calculated to solve the equation:

$$\begin{gathered}\mathrm{wj} = \arg\max \mathrm{wVar}(\mathrm{Zw}),\\ \mathrm{s.t.\,w'w} = 1,\\ \mathrm{Cov}(\mathrm{Zw},\mathrm{Zwl}) = 0, \mathrm{l} = 1,2,\ldots,\mathrm{j}-1.\end{gathered} \tag{13}$$

PCR aims to find the best linear combinations of variables (Z) that closely resemble the entire predictor set. The focus is on capturing common variation rather than the forecasting objective. The PCR algorithm efficiently computes ΩK using the singular value decomposition of Z. De Jong's (1993) PLS research seeks, in contrast, the K linear combinations of Z that are highly predictive of the forecast target. The weights used for constructing each PLS component solve a different objective.

$$\begin{gathered}w_j = arg\overset{Max}{\underset{w}{}} Cov^2(R, Zw),\\ s.t.\ w'w = 1,\ Cov(Zw, Zwl) = 0,\\ l = 1,2,\ldots,j-1.\end{gathered} \tag{14}$$

$$\begin{gathered}ILLIQ_{i,t+1} - ILLIQ^{\wedge}{}_{i,t+1} =\\ \left(g*(z_{i,t}) - g(z_{i,t};\theta)\right) + \left(g(z_{i,t};\theta) - g(z_{i,t};\theta^{\wedge})\right) + (\epsilon_{i,t+1})\end{gathered} \tag{15}$$

.

The subsequent subsections introduce nonparametric models of g(·) with varying degrees of flexibility, accompanied by regularization techniques to address overfitting. (Hastie, Tibshirani, & Friedman, 2009) ((Faraway, 2005)

The model we analyze enhances the basic linear form by incorporating a spline series expansion with K terms of the predictors.

$$g(z;\theta,p(.)) = \sum_{j=1}^{p} p(z_j)'^{\theta_j} \tag{16}$$

The model we examine incorporates a vector of basis functions p(·) and parameter matrix θ=(θ1,θ2,…,θN) to represent a spline series expansion. The spline functions we consider are of order two, specifically (1, z, (z-c1)², (z-c2)², …, (z-cK-2)²), where c1, c2, ..., cK-2 denote the knots. Various options exist for selecting spline functions, but we adopt this specific spline series.

The generalized linear model, incorporating higher-order terms additively, allows for forecasting using *standard* estimation methods. Our analysis employs a least squares objective function, with or without Huber robustness modification. To handle the increased number of model parameters due to series expansion, we employ penalization using a specialized function called the group lasso, which controls degrees of freedom.

$$\phi(\theta;\lambda,K) = \lambda\sum_{j=1}^{P}(\sum_{k=1}^{K}\theta_{j,k}^{2})^{1/2} \tag{17}$$

The group lasso method selects all or none of the K spline terms associated with a specific characteristic (Tibshirani, 1996). This penalty is incorporated into the objective function, accommodating both least squares and robust Huber objectives. Yuan & Lin (2006) see that the group lasso employs accelerated proximal gradient descent, similar to the elastic net, with two tuning parameters, λ and K.

We deeply appreciate Dacheng Xiu's invaluable support in our regression analysis. It was with great skill that Mr. Xiu developed the Matlab script that we utilized for our regression work. The exceptional quality of the script is a testament to his expertise and unwavering commitment to his craft. We are grateful to him for sharing his knowledge and making his code accessible through his official website (Xiu, 2019). Mr. Xiu's work has greatly facilitated our research, and we extend our heartfelt gratitude for his significant contribution.

The code conducts regression analysis for each predictor variable using various regression methods. It removes the current predictor variable from the independent variable matrix and calculates the out-of-sample R-squared value for each regression method. The results are stored in a matrix.

The code visualizes the results through a heatmap, which displays the out-of-sample R-squared values for each regression method and predictor variable. The heatmap allows for a quick comparison of the performance of different regression methods across predictors.

Additionally, the code generates bar plots in two layouts. The first layout presents the results for six regression methods, while the second layout shows the results for the remaining seven methods. The bar plots provide a clear overview of the R-squared values associated with each predictor variable for the different regression methods.

This methodology allows researchers or analysts to assess the performance of multiple regression methods and identify the most effective predictor variables. The visualization techniques help understand the relationships between predictors and the quality of regression models.

4.0 Data

4.1 Data Sources

The dataset on which our thesis is based is taken from Chen & Zimmerman (2023). Our dataset comprises 204 distinct stock characteristics, from which we have selected 43 variables for our study. The characteristics are created based on CRPS/composted of US stocks. Our stock dataset covers the years 1986-2021 and includes financial ratios, volatility measures, and other metrics. Some values are missing and need to be handled carefully for accurate results. We addressed these missing values by creating monthly averages or filling in where possible. However, some data points had to be removed, resulting in a more limited dataset. We needed to compute the average of the available numbers within each period. By calculating these averages, we obtained representative values to fill in the NaN entries corresponding to the respective periods. This approach ensured that missing values were substituted with plausible estimates based on the available data.

However, it is essential to note that there were instances where no values existed for companies during specific periods. Consequently, we had to remove these rows, reducing the overall size of the dataset. Removing these rows had a cascading effect, impacting the dataset's dimensions and decreasing it from nearly 5,000,000 rows to just under 2,800,000.

4.2 Description of Variables

Table 1 describes the variables used in our empirical investigation and regression analysis. The part is split between dependent- and independent variables.

5.0 Main results

In this part, we plan to present empirical evidence which shows which characteristics affect illiquidity through the various regressions. We offer empirical evidence through the period 1986-2021. The characteristics are explained in Table 1.

This section presents the empirical results of our study, which aimed to identify and analyze the characteristics that impact illiquidity in various financial contexts. By employing a rigorous research design and advanced econometric techniques, we sought to unravel new insights into illiquidity determinants, contributing to academic literature and practical applications in financial markets.

These characteristics encompass a range of factors, including firm-specific attributes, market-related variables, investor behavior, and macroeconomic indicators. By examining a wide array of potential drivers of illiquidity, we sought to capture the complex interplay of factors shaping liquidity conditions in different market environments.

5.1 In-sample analysis

In-sample analysis, or in-sample testing or evaluation, refers to assessing a model's performance on the same data set used for training (Bishop, 2006). This analysis compares the model's predictions against the training data's actual outcomes or target variables.

Bishop's (2006) in-sample analysis aims to understand how well the model fits the training data and how accurately it can predict the outcomes already seen. By evaluating the model's performance on the training data, we can assess its ability to capture the underlying patterns and relationships within the data and identify any issues such as overfitting or underfitting.

Sample analysis indicates how well the model has learned from the training data, but it may need to reflect its performance on new, unseen data. This is because the model has already "seen" the training data and might have learned to memorize the patterns specific to that data, leading to overly optimistic results.
Table 2 shows the R2 values for various regression models. The Oracle model has the highest value at 0.1250, indicating the best performance. Other models have lower values, ranging from -0.1193 to 0.1224. Some models have additional adjustments, but their R2 values remain similar to standard ones, which means that, at best, over 85% of illiquidity is explained by factors other than our characteristics.

When the bid-ask spread variable is removed from the regression analysis, the resulting models yield varying outcomes. The Simple OLS and Simple OLS + H models exhibit weak fits, as indicated by their negative R2 values of -0.0873, shown in Table 2. Conversely, the PCR model displays a better fit, showing an increased R2 value of 0.0648. However, the PLS model indicates a poorer fit, with a more negative R2 value of -0.2465. R2 values are explained following (Chatterjee & Hadi, 2012)(Gu et al.,2020) (Plonsky & Ghanbar, 2018) and the results. The Lasso, Ridge, Enet, and Group Lasso models maintain low R2 values around 0.0160, highlighting their weak fits without bid-ask spread. The Oracle model's fit remains unchanged at 0.1250, while Group Lasso and Group Lasso + H models consistently exhibit low fits, with their R2 values hovering around 0.0146.

5.2 Out-of-sample analysis

In data analysis and machine learning, evaluating the performance of a model on unseen data is a critical step in assessing its effectiveness and generalization capabilities. An out-of-sample analysis allows us to test the model's ability to make accurate predictions on independent data points, providing insights into its applicability and reliability (Bishop, 2006).

Bishop's (2006) paper shows that the purpose of analyzing out-of-sample data is to evaluate how well a model performs on new, unseen data. It helps assess the model's ability to generalize and make accurate predictions beyond the data it was trained on. We can gain insights into real-world applicability, reliability, and performance by testing the model on out-of-sample data. Following Gu et al. (2020), the out-of-sample is the most important for our thesis, and the trained model will be the most relevant models and important for our results. Therefore, the out-of-sample analysis is the main focus.

5.3 Results regressions

The MATLAB script reads and preprocesses data, calculates variable importance, and visualizes the results using a heatmap and multiple layouts of bar plots. It defines labels for the plots, creates figures, and arranges the bar plots in a tiled layout. The script creates four different layouts of bar plots, each with a different subset of methods to plot. For each layout, a figure is created using the figure function. A tiled layout is established using the tilted layout function with two rows and one column. For the heatmap, a figure is created using the figure function.

The heatmap is generated using the images function, with r2_oos_mat as the input data.

Simple OLS

This study aimed to conduct a simple Ordinary Least Squares (OLS) regression analysis on a given dataset shown in Figure 2 and Table 3. Including all the variables, the initial regression yielded a coefficient of determination (R-squared) value of 0.12071, indicating that the independent variables explain approximately 12% of the variability in the dependent variable (Chatterjee & Hadi, 2012). Among the variables considered, the Bid-ask spread was identified as the most important predictor, as its exclusion resulted in the most significant impact on the regression results. Furthermore, when excluding each variable one at a time, the variables High52 and VolMkt emerged as dominant factors, as their exclusion had the most significant influence on the regression outcomes. On the other hand, the variable std_turn was the least important, as its exclusion did not significantly alter the regression results. Interestingly, no noticeable impact on the regression was observed when employing the Huber loss function. These findings emphasize the crucial role of bid-ask spread, High52, and VolMkt in explaining the dependent variable, suggesting that std_turn may have limited explanatory power in this analysis.

PCR

Using (PCR) is to establish the correlation between a dependent variable and a set of independent variables(Mullis & Faloona, 1987). The PCR analysis was conducted on all variables, resulting in a coefficient of determination (R-squared) value of -0.11934, as shown in Figure 3 and Table 3. This value indicates that the independent variables collectively account for around 12% of the variability observed in the dependent variable.

After thoroughly analyzing multiple variables, we can see that BM had the most significant influence on the dependent variable, thus making it the most important predictor. Interestingly, in certain scenarios where specific variables were removed from the equation, it was revealed that AccrualsBM, BetaFP, and PriceDelayRsq played a dominant role in the results obtained from the regression. The absence of these variables had the most significant impact on the regression outcomes. These findings indicate the essential role of these variables in explaining the dependent variable within the PCR model context.

Conversely, the variable bid-ask spread was deemed the least important, implying that its inclusion or exclusion had minimal impact on the regression outcomes. This suggests that bid-ask spread may not have a strong relationship with the dependent variable or may be redundant in the presence of other influential variables.

In conclusion, the results of the PCR analysis indicate that several variables, namely BM, AccrualsBM, BetaFP, and PriceDelayRsq, significantly explain the dependent variable. It was noted, however, that the bid-ask spread variable may need to be more relevant in this particular analysis.

PLS

We run PLS regression to investigate the correlation between a set of independent variables and a dependent variable. The PLS analysis was conducted with all variables and shown in Figure 3 and Table 3. It yielded an R-squared value of -0,08250, indicating that the independent variables collectively account for roughly 8% of the variation in the dependent variable.

Bid-ask spread was the most significant predictor of all the variables analyzed, implying that it considerably impacted the dependent variable. Interestingly, when each variable was removed individually, PriceDelayRsq and ShareVol were identified as the dominant factors, as their exclusion resulted in the most significant changes in the regression results. These results emphasize the crucial role of PriceDelayRsq and ShareVol in explaining the dependent variable in the PLS model.

Conversely, the variable Accruals were determined to be the least important, indicating that its inclusion or exclusion had minimal effect on the regression outcomes. This suggests that Accruals may have a weak relationship with the dependent variable or may be less influential than other model variables.

In summary, the PLS regression analysis revealed the significance of variables such as bid-ask spread, PriceDelayRsq, and ShareVol in explaining the dependent variable. Additionally, the regression indicated that Accruals may have little importance in the context of this study.

Lasso

Including all variables, the Lasso analysis resulted in a coefficient of determination (R-squared) value of 0.12237, indicating that the independent variables collectively explain approximately 12% of the variability in the dependent variable shown in Figure 4 and Table 3.

Among the variables considered, bid-ask spread was identified as the most important predictor, implying that it significantly impacted the dependent variable. Interestingly, when excluding each variable one at a time, the variables High52 and VolMkt emerged as dominant factors, as their exclusion had the most significant influence on the regression results. These findings highlight the crucial role of High52 and VolMkt in explaining the dependent variable within the context of the Lasso model.

On the other hand, the variable std_turn was determined to be the least important, indicating that its inclusion or exclusion had minimal effect on the regression outcomes. This suggests that std_turn may need more relevance in explaining the dependent variable when considering the other variables included in the model.

Additionally, applying the Huber loss function, the regression stays the same. This could imply that the Lasso model is relatively strong and not significantly affected by outliers or influential observations.

In summary, the Lasso regression analysis demonstrated the significance of variables such as bid-ask spread, High52, and VolMkt in explaining the dependent variable. Furthermore, it indicated that std_turn might have little importance in this analysis, while the robustness of the model was observed through the Huber loss function.

Ridge

The Ridge analysis, utilizing all variables, yielded a coefficient of determination (R-squared) value of 0.12062, indicating that the independent variables collectively explain approximately 12% of the variability in the dependent variable given in Figure 5 and Table 3.

Among the variables considered, bid-ask spread was identified as the most important predictor, implying that it significantly impacted the dependent variable. Intriguingly, when excluding each variable one at a time, the variables MaxRet and BetaTailRisk emerged as dominant factors, as their exclusion had the most significant influence on the regression results. These findings underscore the critical role of MaxRet and BetaTailRisk in explaining the dependent variable within the context of the Ridge model.

Conversely, the variable std_turn was determined to be the least important, suggesting that its inclusion or exclusion had minimal effect on the regression outcomes. This indicates that std_turn may need more relevance in explaining the dependent variable when considered alongside the other variables in the model.

Employing the Huber loss function did not affect the regression results. This indicates that the Ridge model is relatively strong and is not significantly influenced by outliers or influential observations.

In summary, the Ridge regression analysis demonstrated the significance of variables such as bid-ask spread, MaxRet, and BetaTailRisk in explaining the dependent variable. Furthermore, it indicated that std_turn might have little importance in this analysis, while the robustness of the model was observed through the application of the Huber loss function.

Enet

The Elastic Net analysis, incorporating all variables, resulted in a coefficient of determination (R-squared) value of 0.12204, indicating that the independent variables collectively explain approximately 12% of the variability in the dependent variable shown in Figure 6 and Table 3.

Among the variables considered, bid-ask spread emerged as the most important predictor, suggesting that it holds influence over the dependent variable. Notably, when excluding each variable one at a time, the variables High52 and VolMkt were found to be most dominant, as their exclusion had the most significant impact on the regression results. These findings highlight the roles of High52 and VolMkt in explaining the dependent variable within the context of the Elastic Net model.

Conversely, the variable std_turn was identified as the least important, indicating that its inclusion or exclusion had minimal effect on the regression outcomes. This indicates that std_turn may need more relevance in explaining the dependent variable when considered alongside the other variables included in the model. Moreover, using the Huber loss function did not significantly influence the results. Implying that the Elastic Net model is relatively insensitive to outliers or influential observations.

Oracle

The Oracle analysis shown in Figure 6 and Table 3, incorporating all variables, resulted in a coefficient of determination (R-squared) value of 0.12499, indicating that the independent variables collectively explain approximately 12.5% of the variability in the dependent variable.

Among the variables considered, roaq was identified as the most important predictor, suggesting that it significantly influences the dependent variable. Intriguingly, when excluding each variable one at a time, the variable Accruals emerged as the most dominant factor, as its exclusion had the most significant impact on the regression results. These findings underscore the role of Accruals in explaining the dependent variable within the context of the Oracle model.

Conversely, the variable betaVIX was determined to be the least important, indicating that its inclusion or exclusion had minimal effect on the regression outcomes. This suggests that betaVIX may need more relevance in explaining the dependent variable when considered alongside the other variables included in the model.

Group Lasso

The Group Lasso analysis, using all variables shown in Figure 7 and Table 3, yielded a coefficient of determination (R-squared) value of 0.05242, indicating that the independent variables collectively explain approximately 5.2% of the variability in the dependent variable.

Among the variables considered, bid-ask spread emerged as the most important predictor, suggesting that it holds a significant influence over the dependent variable. Interestingly, every other variable in the model had the same value, implying that they collectively contribute to the regression outcome but do not possess individual importance.

Furthermore, applying the Huber loss function did not affect the regression results, as shown in Figure 8. This indicates that the Group Lasso model is robust to outliers or influential observations.

5. 4 Discussion regressions

When examining various regression models, it becomes clear that the bid-ask spread significantly impacts the dependent variable in most cases. This finding is consistent across most regression methods, suggesting that the bid-ask spread is a crucial factor in explaining the variability in the dependent variable. However, this rule has a few exceptions, such as PCR and Oracle regression. Interestingly, the results of these two regression methods differ, with Roaq performing best in Oracle regression and BM performing best in PCR. The results imply that other variables may be necessary depending on the regression approach.

Roaq emerges as the most significant predictor in Oracle regression, while BM takes precedence in PCR regression. This finding underscores the importance of considering different factors when performing regression analysis, as different variables may have varying degrees of influence depending on the method employed. It is worth noting that the bid-ask spread consistently exhibits stronger correlations than other variables, which is unsurprising given that it measures liquidity.

Overall, the dominance of bid-ask spread across multiple regression models reinforces its significance in explaining the dependent variable. This finding highlights the importance of liquidity considerations in financial analysis and enhances our understanding of the factors that affect market dynamics. By examining the performance of different variables across multiple regression methods, we can gain a more comprehensive understanding of the factors that drive financial markets.

After conducting a regression analysis, it was observed that incorporating the Huber loss function did not substantially impact the initial results. The results suggest that the regression model was already quite resilient to outliers and other influential observations (Huber,1964) Therefore, the original regression model could withstand outliers effectively if the results are not significantly altered by integrating this robust regression method.

Through our analysis, we have found several findings that must be addressed. Firstly, removing variables individually does indeed affect the regression results, indicating that each variable plays a role in contributing to the overall explanatory power to some degree. However, it was also determined that the bid-ask spread factor significantly impacts the dependent variable, revealing its crucial role in predicting the model's outcome. Nevertheless, relying solely on one predictor may decrease the model's prediction accuracy, suggesting that a more comprehensive approach may be necessary.

Interestingly we found that the Lasso, Enet, and Oracle regressions were proven to be better than the OLS regressions. This can be explained by Lasso, Enet, and Oracle regressions have advantages over OLS regression in situations with many predictors, potential multicollinearity, or a need for variable selection. They provide more robust and interpretable models, leading to improved prediction accuracy and enhanced understanding of the underlying relationships between predictors and the response variable, which is also suggested through the results of the different regressions in our thesis.

Furthermore, the coefficient of determination (R-squared) values were relatively low, hovering around 0.20, implying that none of the variables, individually or collectively, can effectively predict the model's outcome with high accuracy. This low explanatory power raises questions about the ability to accurately predict illiquidity based on the variables considered in this study. Although certain variables have a more significant impact than others, the limited data and variables available in this research limit the generalizability of the findings. With a larger dataset and a more comprehensive range of variables, making solid predictions about illiquidity is easier.

Regressions without bid-ask spread

The new R-squared values without the "bid-ask spread" variable are generally lower than the original R-squared values, as shown in Table 4. This suggests that removing the "bid-ask spread" variable has negatively impacted the models' ability to explain the variance in the dependent variable.

Among the models, the Simple OLS, Simple OLS + H, Ridge, and Ridge+H models show lower R-squared values in both the original and new calculations, indicating that the removal of the variable has worsened their performance. The Lasso, Lasso+H, Enet, and Enet+Huber loss models also exhibit lower R-squared values in the new calculations.

On the other hand, the PCR model shows a higher R-squared value without the "bid-ask spread" variable, indicating improved performance after removing it. However, the PLS, Oracle, Group Lasso, and Group Lasso+H models still show lower R-squared values in the new calculations, suggesting a weaker fit even without the "bid-ask spread" variable.

The removal of the "bid-ask spread" variable has generally led to lower R-squared values, indicating a decrease in the model's ability to explain the dependent variable's variance. This is more proof that the bid-ask spread is the most significant variable, causing the regressions to show a lower variance that explains the Amihud illiquidity. The most important element in the new regressions is how poorly OLS performs compared to the other regressions shown in Table 4. Therefore, having less significant explanatory power and other regressions such as Lasso, Oracle, and Enet will be more relevant to explore.

Bid-ask spread out-of-sample analysis

Not surprisingly, the models were worse when removing the bid-ask spread, but interestingly the models gave more variance in the different characteristics in the different out-of-sample regressions.

In Figure 10, the Ols regression, the VolMkt, Sharevol, and BM were the most distinct characteristics. All the characteristics were negative, and the VolmMkt was the most significant at -0,1545.

The PCR regression in Figure 11 can both have positive and negative values. When the value is negative, it means that when the characteristic value increases, the dependent variable will decrease, and when it is positive, the dependent variable increases. Therefore, the most prolific characteristics are based on the highest negative or positive values (Livak & Schmittgen, 2001). BM and RDS, with both negative numbers, have the highest values but still low explanatory value at -0,0751.

PlS can also be negative and have the same inverse relationship as PCR. The most prolific characteristics were AccrualsBM and PriceDelayRsq. While still having a low R2 value given in Figure 11. Having a value of -0,07203

Lasso regression, there is some positive and some negative R2 in Figure 12. Further showing the small inverse relationship BM and VolMkt have on Amihud illiquidity since the values are still small and explain -0,0447.

The ridge regression in Figure 13 VolMkt and BM are still the most prolific factors with BM -0,1125 and VolMKT -0,1150 from Table 5. They are still showing an inverse relationship with Amihud's illiquidity. Ridge regression shows higher values than previous regressions.

The Enet regression also shows low values in Figure 14, continuing that BM and VolMkt are still the most significant and continuing the inverse relationship with the dependent variable with a value of -0,0448 and -0,0213 from Table 5.

The oracle regression is constant with no particular characteristic being different shown in Figure 15.

While Group Lasso, also shown figures 15 and 16, shows all characteristics as the same with MaxRet with values of 0,0023 from Table 5.

The out-of-sample and in-sample analysis results without the bid-ask spread show more variance in the different regressions. The regressions such as Oracle, Enet, and Lasso are also proven to be more accurate than the OLS regression in the in-sample regression. This shows that the fit for the other regressions performs better than the OLS, as seen in Table 4. Different characteristics are shown to be more significant through the regressions. The regressions show clearly that VolMkt and BM are the most significant in the most regressions but also in the regressions that are proven to be most significant.

These results vary based on the specific characteristics chosen in this thesis. Alternative results could have been obtained had different variables been chosen or if the set of characteristics had been expanded. The study only considered 43 characteristics, a limited subset of all the potential factors that could impact Amihud's illiquidity. It would be beneficial to incorporate a larger number of variables to attain a more comprehensive understanding and improve the accuracy of predictions. By doing so, additional characteristics could be explored, offering deeper insights and potentially leading to more precise predictions. The reason for not including more characteristics was that the dataset had Na/NaN values, and including more characteristics would make our dataset smaller.

Our research shows that illiquidity cannot be predicted effectively using the investigated variables. The lack of explanatory power, reliance on a single predictor, and limitations imposed by data constraints all highlight the need for further research and a more comprehensive approach to accurately predict and understand illiquidity in financial markets. The results could also be because the characteristics tested do not significantly affect illiquidity. More extensive research must provide valuable insights into this complex phenomenon.

5.5 Robustness of the analysis

One potential criticism of the thesis is handling the dataset's NaN (missing) values. Taking the average across periods and filling in NaN values can introduce biases and potentially distort the results. Averaging across periods assumes that the missing values have a similar distribution to the available data, which may only sometimes be the case. This method can overlook significant variations and patterns in the data.

Moreover, removing rows where filling in NaN values was impossible is another aspect that could be criticized. This significant reduction in sample size raises concerns about the representativeness and generalizability of the findings. It is crucial to consider the potential impact of such data loss on the statistical power and validity of the results.

Another point of criticism is the dominance of the bid-ask spread variable in every regression model. If bid-ask spread consistently emerges as a significant predictor and dominates the results, it raises questions about the robustness and reliability of the findings. This dominance could overshadow the effects of other variables and lead to an overemphasis on bid-ask spread in interpreting the results. It is essential to thoroughly investigate the reasons behind this dominance and assess whether it is reasonable or if it could be attributed to data peculiarities or model misspecifications.

Exploring alternative methods for handling missing values is recommended to address these criticisms, such as multiple imputation techniques that consider the underlying patterns and relationships in the data. Additionally, sensitivity analyses can be conducted to assess the impact of the missing data and evaluate the robustness of the results when different imputation or deletion strategies are employed.

Regarding the dominance of bid-ask spread, it is crucial to evaluate the reasons behind its strong influence and assess its economic and theoretical significance. This could involve examining the underlying relationships between bid-ask spread and the other variables and considering potential confounding factors or omitted variables that might explain the observed dominance. Further analysis, such as regression diagnostics and model selection techniques, can help determine whether the results are reliable and if the power of bid-ask spread is justified. Overall, these criticisms highlight the need for careful consideration and transparent reporting of the data handling techniques, as well as a thorough evaluation of the impact of dominant variables on the overall results. Addressing these concerns can strengthen the validity and reliability of the thesis findings and enhance the trustworthiness of the research outcomes.

6.0 Conclusion

In conclusion, the regression analysis conducted using the Oracle and Group Lasso models provided insights into the factors influencing the variability in the dependent variable. The findings indicated that while the independent variables collectively explain a portion of the variability, the overall explanatory power was relatively low. The R-squared values hovered around 0.20, suggesting that the variables examined in the study, individually and collectively, were not highly effective in accurately predicting the outcome. Our out-of-sample analysis shows how few of the characteristics explain Amihud illiquidity. This may prove that the characteristics do not affect the illiquidity or only the illiquidity in a limited way.

Among the variables considered, the bid-ask spread consistently emerged as a significant predictor across multiple regression methods. This explains its crucial role in explaining the variability in the dependent variable and highlights the importance of liquidity considerations in financial analysis. However, the dominance of bid-ask spread raised questions about the robustness and reliability of the findings. It was crucial to thoroughly investigate the reasons behind this dominance and assess whether it was reasonable due to data peculiarities or model misspecifications.

The analysis also proved that the characteristics show low explanatory value to the Amihud illiquidity, making it hard for practical use.

While the bid-ask spread exhibited strong correlations with the dependent variable, it was essential to recognize that relying solely on one predictor may decrease the model's prediction accuracy. The results indicated the need for a more comprehensive approach that includes a larger number of variables to enhance the accuracy of predictions and gain a deeper understanding of the underlying dynamics.

The regressions without the bid-ask spread also proved to show some characteristics that are more significant such as BM and VolMkt shown to have the highest values in the regressions. Lasse, Oracle and enet are the regression proved to be the best with their R2 values. Proving that over these regressions that BM and VolMkt are the most significant in these regressions.

7.0 Bibliography

Amihud, Y. (2002). Illiquidity and stock returns: Cross-section and time-series effects. Journal of Financial Markets, 5(1), 31–56.

Barker, M., & Rayens, W. (2003). Partial least squares for discrimination. Journal of Chemometrics, 17(3), 166-173.

Bishop, C.M. (2006) Pattern Recognition and Machine Learning. Springer, Berlin.

Chatterjee, S., & Hadi, A. S. (2012). Regression Analysis by Example (5th ed.). John Wiley & Sons.

Chen, A. Zimmerman, T. (2023). Open Source Asset Pricing. Featured Stock-level Signal Datasets. https://www.openassetpricing.com/data/

Cont, R., & Moussa, A. (2013). Liquidity, transaction costs, and market efficiency. Journal of Financial Markets, 16(1), 1-32.

De Jong, S., 1993. SIMPLS: an alternative approach to partial least squares regression. Chemometrics and Intelligent Laboratory Systems, 18: 251–263.

Freedman, D. A. (2009). Ordinary least squares. Wiley Interdisciplinary Reviews: Computational Statistics, 1(1), 65-70.

Gârleanu, N., & Pedersen, L. H. (2013). Dynamic Trading with Predictable Returns and Transaction Costs. Journal of Finance, 68(6), 2309-2340.

Geladi, P., & Kowalski, B. R. (1986). Partial Least-Squares Regression: A Tutorial. Analytica Chimica Acta, 185, 1-17.

Goyenko, Ruslan & Holden, Craig & Trzcinka, Charles. (2009). Do Liquidity Measures Measure Liquidity?. Journal of Financial Economics. 92. 153-181. 10.1016/j.jfineco.2008.06.002.

Gu, S., Kelly, B., & Xiu, D. (2020). Empirical asset pricing via machine learning. The Review of Financial Studies, 33(5), 2223-2273. https://doi.org/10.1093/rfs/hhaa009.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: Data mining, inference, and prediction (2nd ed.). Springer.
Faraway, J. J. (2005). Linear models with R. CRC Press.

Huber, P. J. (1964). Robust estimation of a location parameter. The Annals of Mathematical Statistics, 35(1), 73-101.

Huber, P. J. (1981). Least absolute deviations: Theory, applications, and algorithms. In F. L. Hodges Jr., P. R. Nelson, & N. C. Kenkel (Eds.), Studies in Econometrics, Time Series, and Multivariate Statistics (pp. 301-334). Academic Press.

Jensen, T. I., Kelly, B. T., Malamud, S., & Pedersen, L. H. (2022). Machine Learning and the Implementable Efficient Frontier. Swiss Finance Institute. Swiss Finance Institute Research Paper Series No. 22-63.

Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: A review and recent developments. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 374(2065), 20150202.

Kaniel, R, Z Lin, M Pelger and S Van Nieuwerburgh (eds) (2023), “DP18129 Machine-Learning the Skill of Mutual Fund Managers”, CEPR Press Discussion Paper No. 18129. https://cepr.org/publications/dp18129

Livak, K. J., & Schmittgen, T. D. (2001). Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method. Methods, 25(4), 402-408.

Martens, H., & Martens, M. (2001). Multivariate analysis of quality: An introduction. John Wiley & Sons.

Mullis, K. B., & Faloona, F. A. (1987). Specific synthesis of DNA in vitro via a polymerase-catalyzed chain reaction. Methods in Enzymology, 155, 335-350.

Plonsky, L., & Ghanbar, H. (2018). Multiple regression in L2 research: A methodological synthesis and guide to interpreting R2 values. The Modern Language Journal, 102(4), 713-731.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological), 58(1), 267-288.

Wold, H. (1966). Estimation of Principal Components and Related Models by Iterative Least Squares. In Multivariate Analysis (pp. 391-420). Academic Press.

Wold, S., Ruhe, A., Wold, H., & Dunn III, W. J. (1984). The collinearity problem in linear regression. The partial least squares (PLS) approach to generalized inverses. SIAM Journal on Scientific and Statistical Computing, 5(3), 735-743.

Xiu, D. (2019, June 24). ML_Codes. https://github.com/xiubooth/ML_Codes.

Yan, X., Su, X., & Ma, X. (2009). Linear regression analysis: Theory and computing. Statistics Surveys, 3, 45-78.

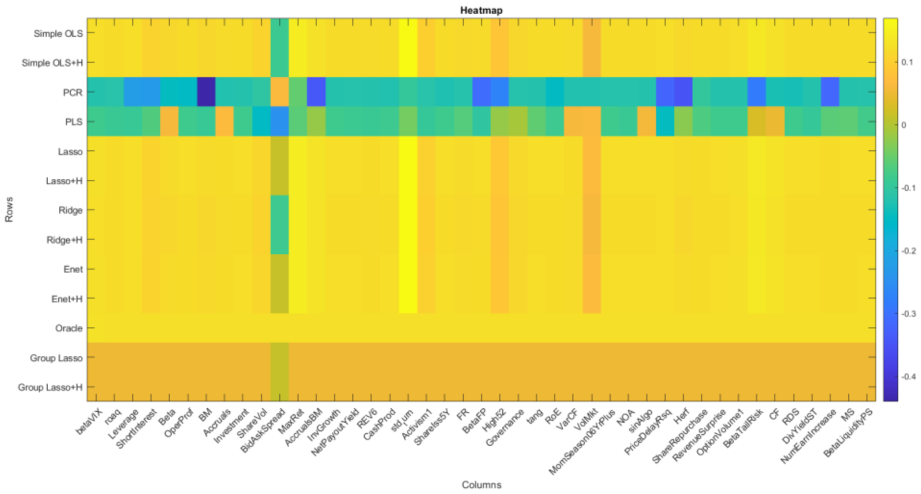
Yuan, M., & Lin, Y. (2006). Model selection and estimation in regression with grouped variables. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 68(1), 49-67.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. Journal of the royal statistical society: series B (statistical methodology), 67(2), 301-320.

Appendix:

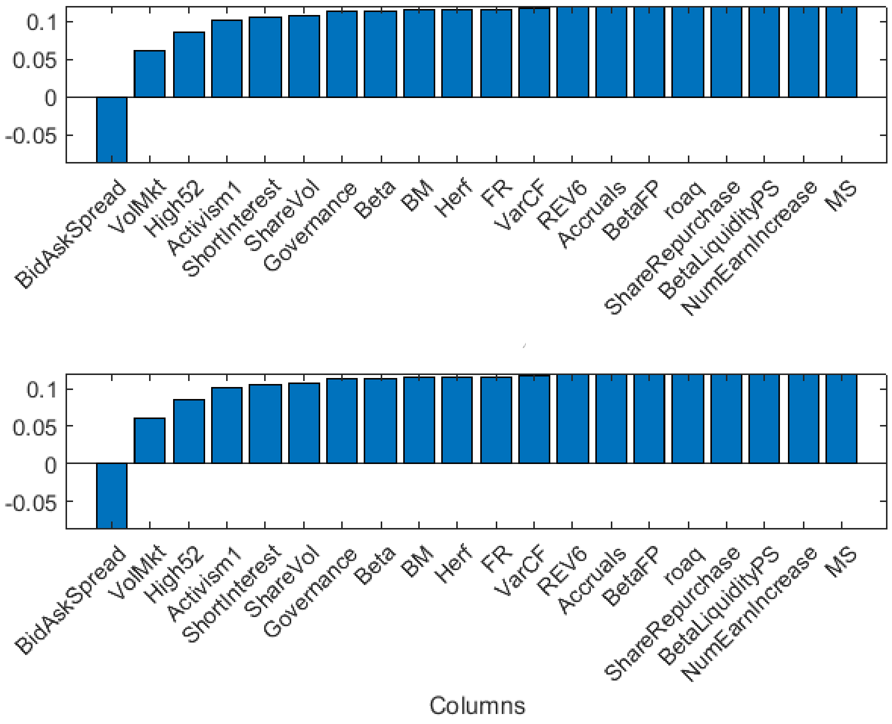
Figures:

Figure 1 - Heatmap



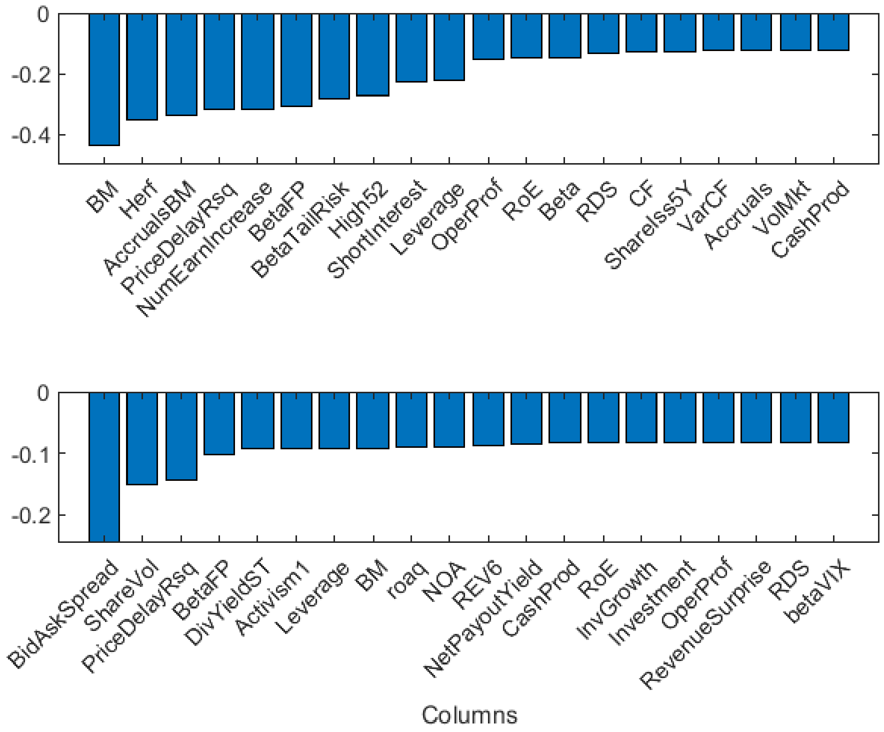
This figure shows a heatmap consisting of the 13 regression methods and the 43 stock characteristics.

Figure 2 – Simple OLS and Simple OLS with Huber loss



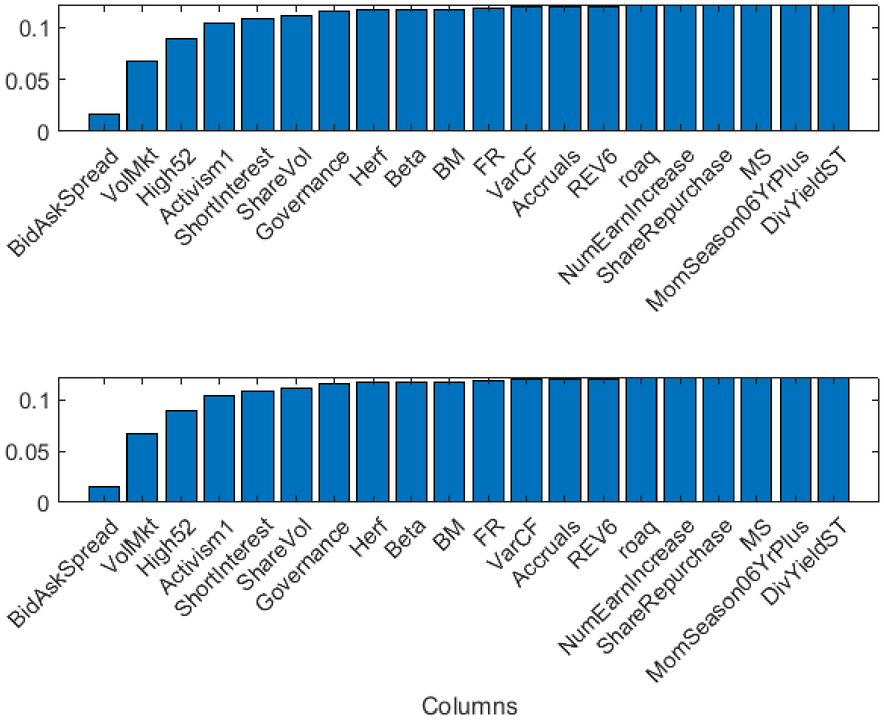
This figure shows the regressions Simple OLS and Simple OLS with Huber loss function and the 20 most significant stock characteristics. The graph illustrates the relationship between the characteristics removed (x-axis) and the corresponding R2 values (y-axis).

Figure 3 PCR and PLS



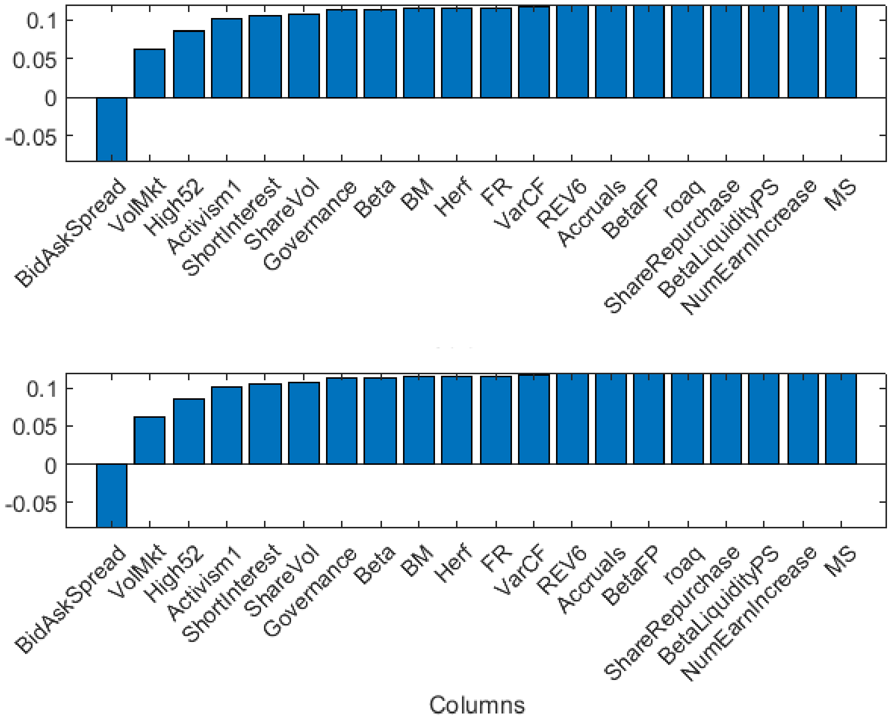
This figure shows the PCR and PLS regressions and the 20 most significant stock characteristics. The graph illustrates the relationship between the characteristics removed (x-axis) and the corresponding R2 values (y-axis).

Figure 4 Lasso and Lasso with Huber loss



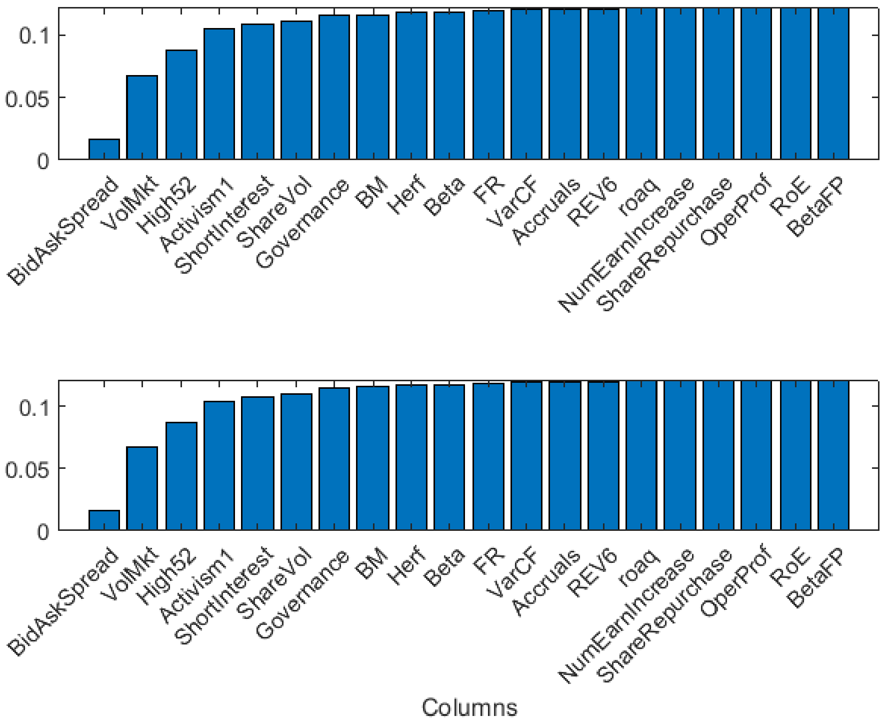
This figure shows Lasso and Lasso's regressions with the Huber loss function and the 20 most significant stock characteristics. The graph illustrates the relationship between the characteristics removed (x-axis) and the corresponding R2 values (y-axis).

Figure 5 Ridge and Ridge with Huber loss



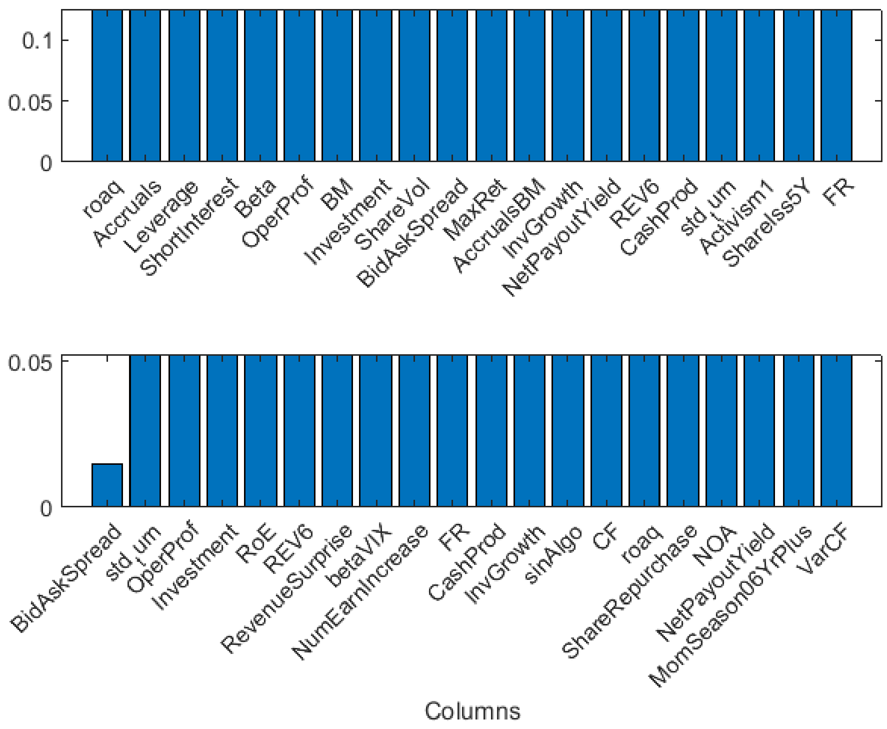
This figure shows Ridge and Ridge regressions with Huber loss function and the 20 most significant stock characteristics. The graph illustrates the relationship between the characteristics removed (x-axis) and the corresponding R2 values (y-axis).

Figure 6 Enet and Enet with Huber loss



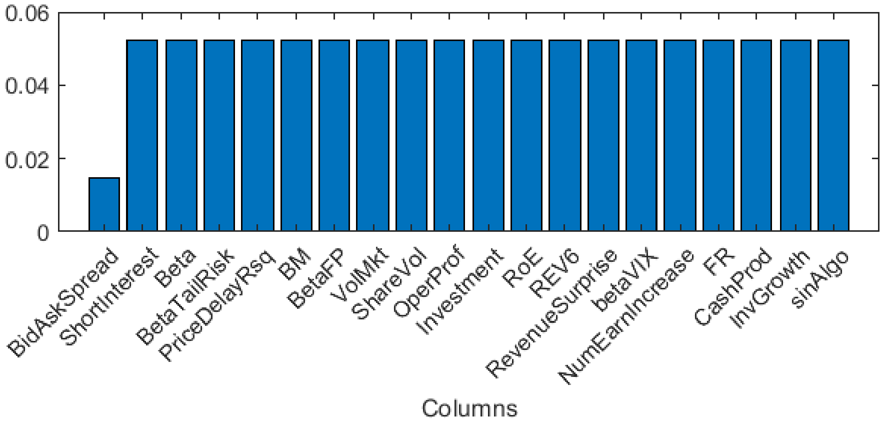
This figure shows Enet and Enet's regressions with the Huber loss function and the 20 most significant stock characteristics. The graph illustrates the relationship between the characteristics removed (x-axis) and the corresponding R2 values (y-axis).

Figure 7 Oracle and Group Lasso



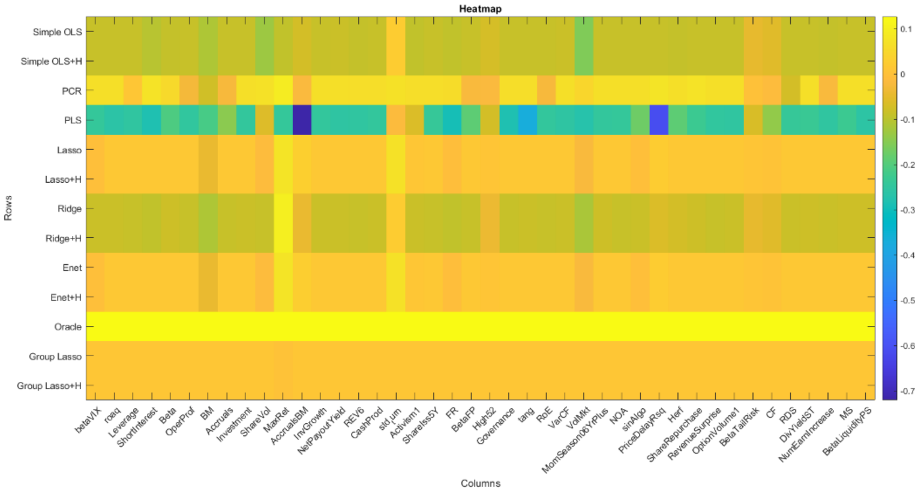
This figure shows the regressions of Oracle and Group Lasso and the 20 most significant stock characteristics. The graph illustrates the relationship between the characteristics removed (x-axis) and the corresponding R2 values (y-axis).

Figure 8 Group Lasse with Huber loss



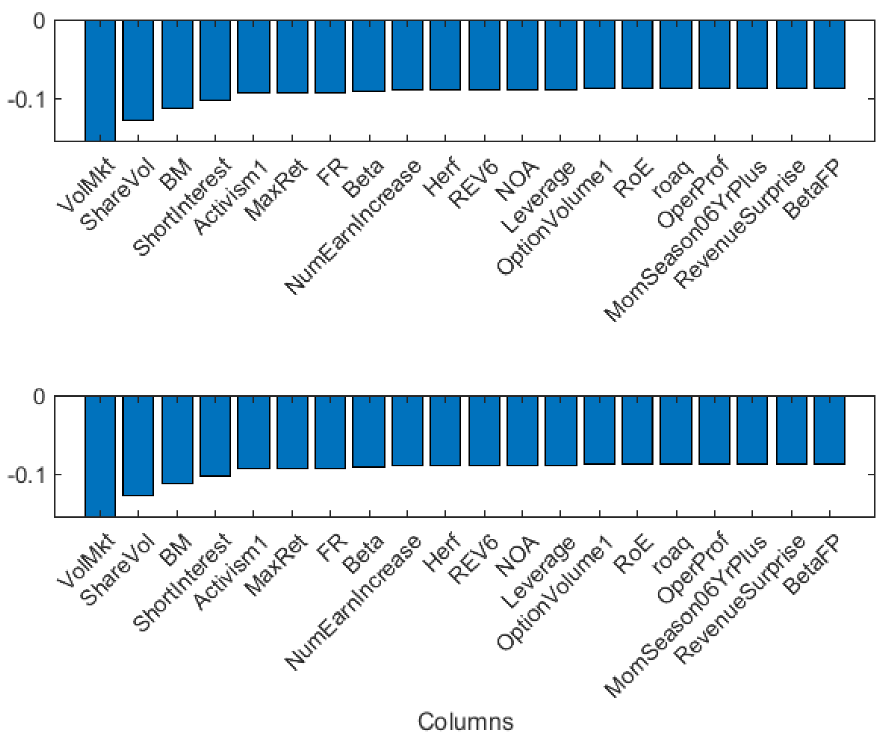
The figure shows the regression Group Lasso with the Huber loss function and the 20 most significant stock characteristics. The graph illustrates the relationship between the characteristics removed (x-axis) and the corresponding R2 values (y-axis).

Figure 9 – Heatmap without bid-ask spread



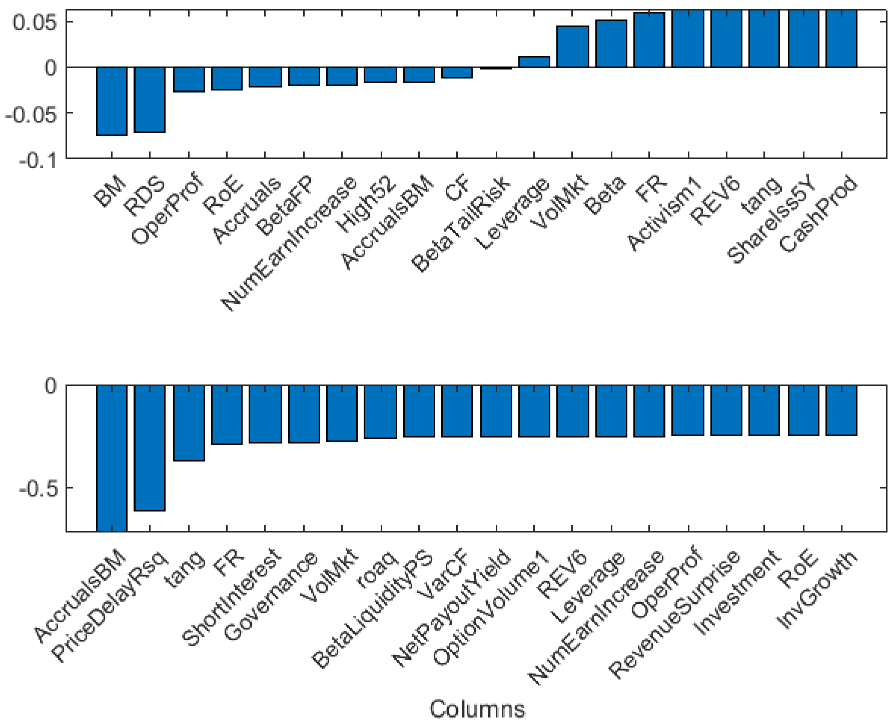
This figure shows a heatmap consisting of the 13 regression methods and the 42 stock characteristics when excluding the bid-ask spread characteristic.

Figure 10 Simple OLS and Simple OLS with Huber loss function excluding bid-ask spread



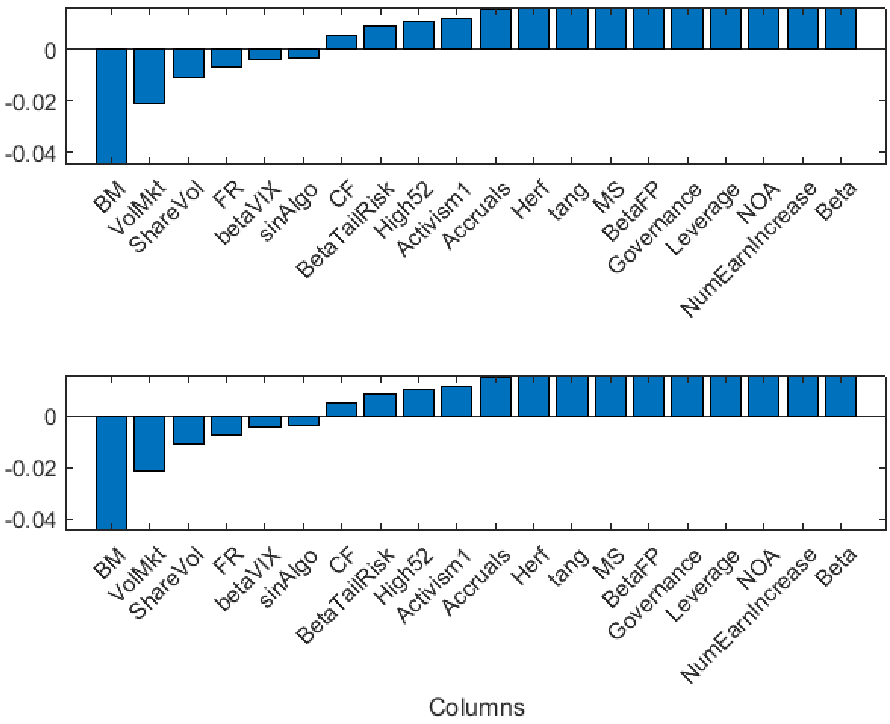
This figure shows the regressions Simple OLS and Simple OLS with Huber loss function and the 20 most significant stock characteristics when excluding the bid-ask spread characteristic. The graph illustrates the relationship between the characteristics removed (x-axis) and the corresponding R2 values (y-axis).

Figure 11 PCR and PLS excluding bid-ask spread



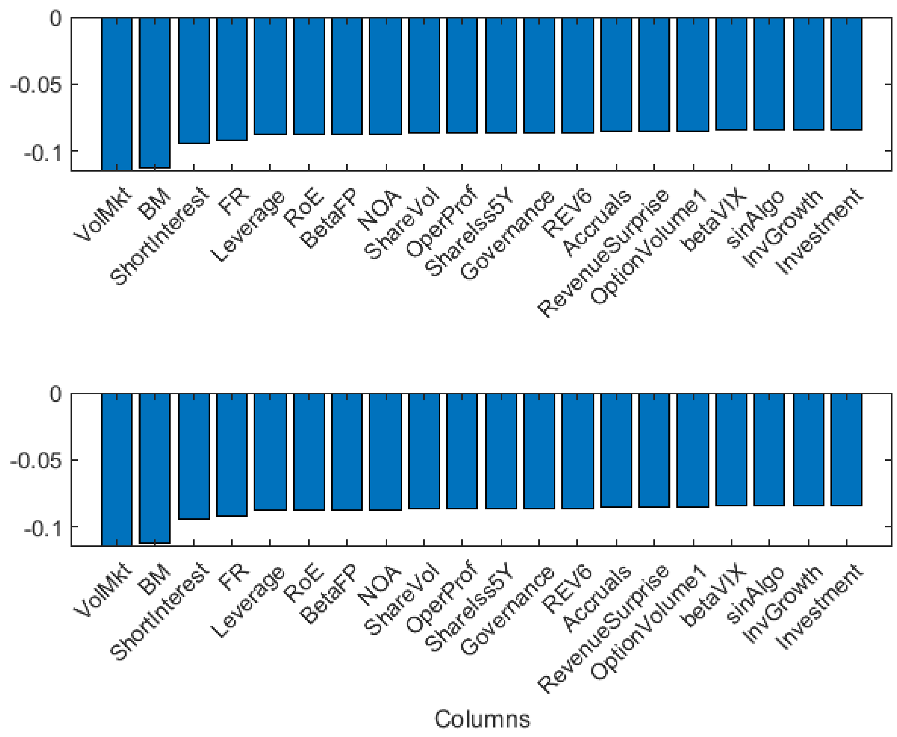
This figure shows the PCR and PLS regressions and the 20 most significant stock characteristics when excluding the bid-ask spread characteristic. The graph illustrates the relationship between the characteristics removed (x-axis) and the corresponding R2 values (y-axis).

Figure 12 Lasso and Lasso with Huber loss excluding bid-ask spread



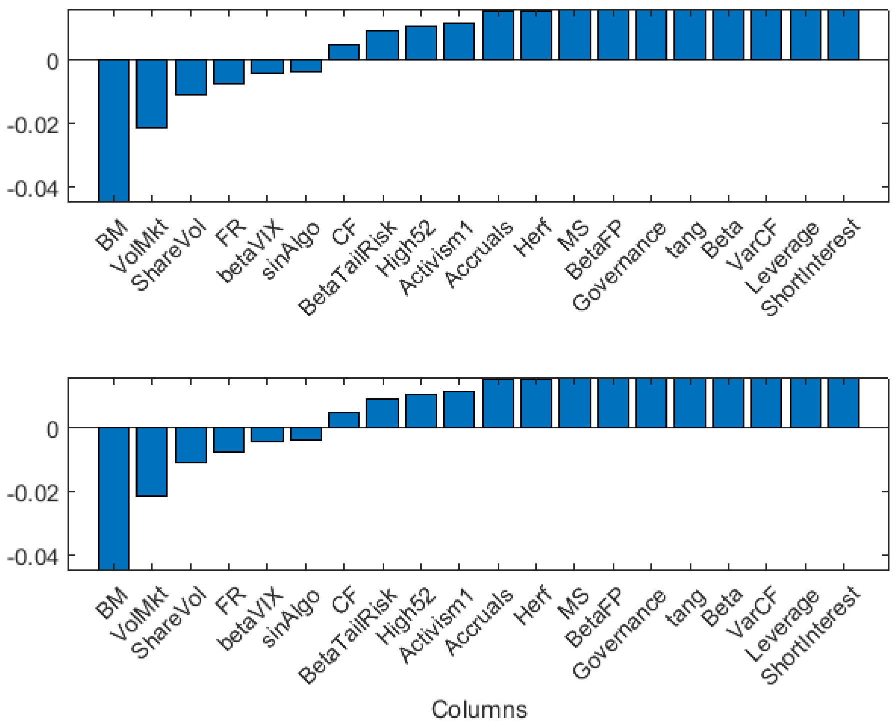
This figure shows the regressions Lasso and Lasso with Huber loss function and the 20 most significant stock characteristics when excluding the bid-ask spread characteristic. The graph illustrates the relationship between the characteristics removed (x-axis) and the corresponding R2 values (y-axis).

Figure 13 Ridge and Ridge with Huber loss excluding bid-ask spread



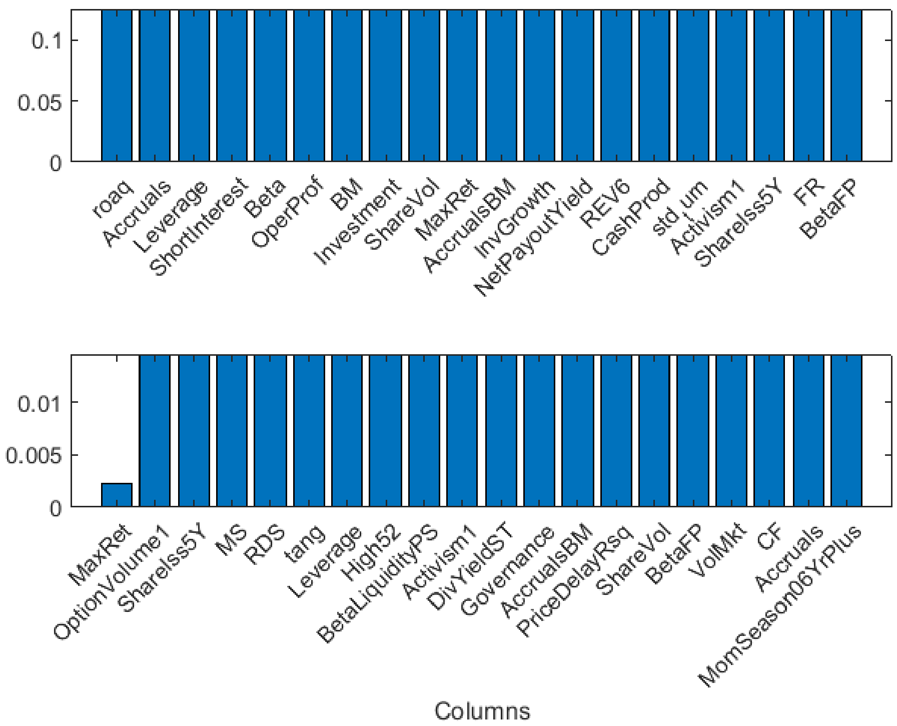
This figure shows the regressions Ridge and Ridge with Huber loss function and the 20 most significant stock characteristics when excluding the bid-ask spread characteristic. The graph illustrates the relationship between the characteristics removed (x-axis) and the corresponding R2 values (y-axis).

Figure 14 Enet and Enet with Huber loss excluding bid-ask spread



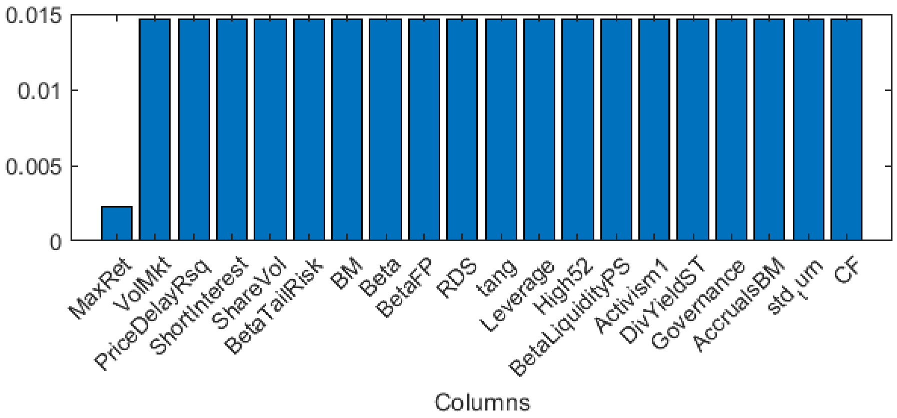
This figure shows Enet and Enet's regressions with the Huber loss function and the 20 most significant stock characteristics when excluding the bid-ask spread characteristics. The graph illustrates the relationship between the characteristics removed (x-axis) and the corresponding R2 values (y-axis).

Figure 15 Oracle and Group Lasso excluding bid-ask spread



This figure shows the regressions of Oracle and Group Lasso and the 20 most significant stock characteristics when excluding the bid-ask spread characteristics. The graph illustrates the relationship between the characteristics removed (x-axis) and the corresponding R2 values (y-axis).

Figure 10 Group Lasso with Huber loss excluding bid-ask spread



The figure shows the regression Group Lasso with the Huber loss function and the 20 most significant stock characteristics when excluding the bid-ask spread characteristics. The graph illustrates the relationship between the characteristics removed (x-axis) and the corresponding R2 values (y-axis).

Tables

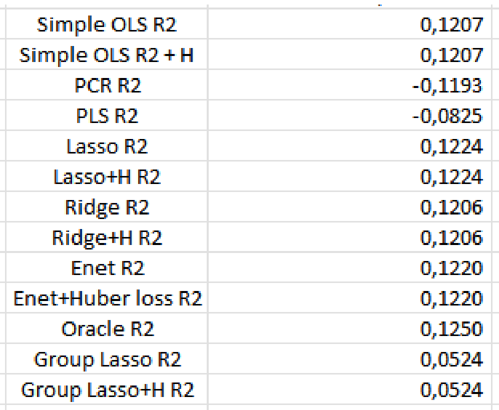
Table 1

betaVIX	betaVIX: Beta concerning the VIX index	High52	52-week high
Roaq	Return on assets quality	Governance	Governance score
Leverage	Debt-to-equity ratio	tang	Tangibility of assets
ShortInterest	Short interest ratio	RoE	Return on equity
Beta	Beta coefficient measuring stock's sensitivity to market movements	VarCF	Variability of cash flow
OperProf	Operating profit margin	VolMkt	Market volatility
BM	The ratio of book value to the market value of equity	MomSeason06YrPlus	Momentum seasonality
Accruals	Earnings accruals	NOA	Net operating assets
Investment	Level of investment	sinAlgo	Algorithmic trading score
ShareVol	Share volume traded	PriceDelayRsq	Price delay R-squared
BidAskSpread	Difference between bid and ask prices	Herf	Herfindahl-Hirschman index
MaxRet	Maximum return	ShareRepurchase	Share repurchase activity
AccrualsBM	Interaction between accruals and book-to-market ratio	RevenueSurprise	Surprise in revenue
InvGrowth	Investment growth rate	OptionVolume1	Option volume traded
NetPayoutYield	Net payout yield	BetaTailRisk	Beta tail risk
REV6	Revenue over six months	CF	Cash flow
CashProd	Cash production	RDS	Research and development expenses
std_turn	Standardized turnover	DivYieldST	Short-term dividend yield
Activism1	Activism score	NumEarnIncrease	Number of earnings increases
ShareIss5Y	Share issuance in the last five years	MS	Market sensitivity

FR	Quality of financial reporting	BetaLiquidityPS	Beta liquidity factor
BetaFP	Beta factor pricing		

Table 1 shows the 43 stock characteristics with a short description of them.

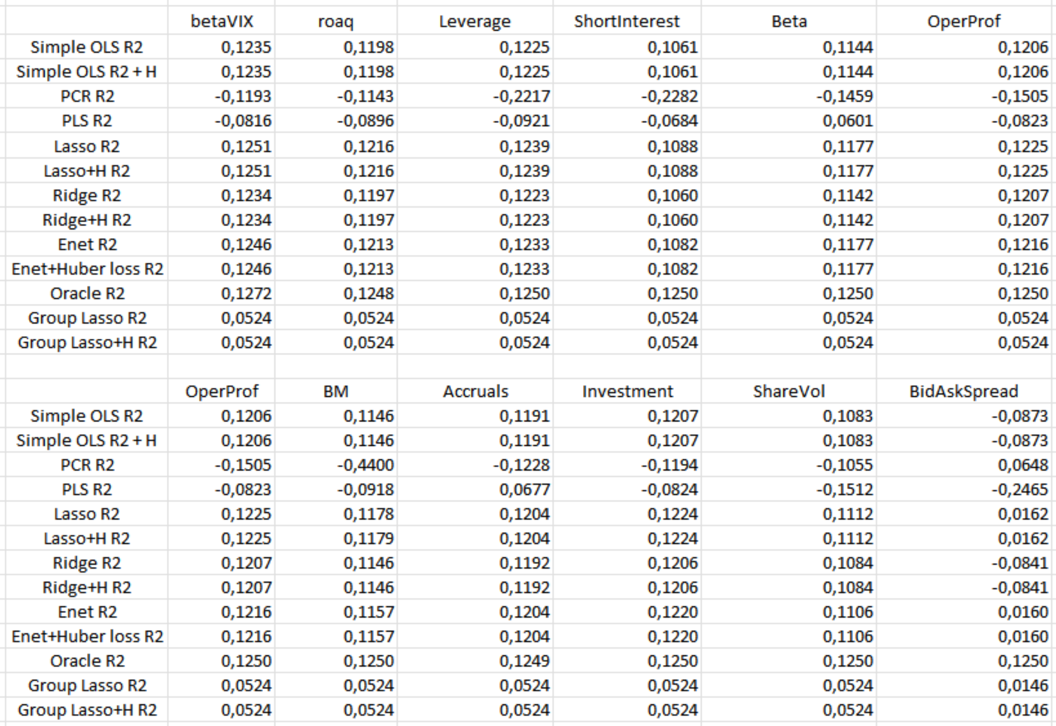
Table 2 – In sample values



Simple OLS R2	0,1207
Simple OLS R2 + H	0,1207
PCR R2	-0,1193
PLS R2	-0,0825
Lasso R2	0,1224
Lasso+H R2	0,1224
Ridge R2	0,1206
Ridge+H R2	0,1206
Enet R2	0,1220
Enet+Huber loss R2	0,1220
Oracle R2	0,1250
Group Lasso R2	0,0524
Group Lasso+H R2	0,0524

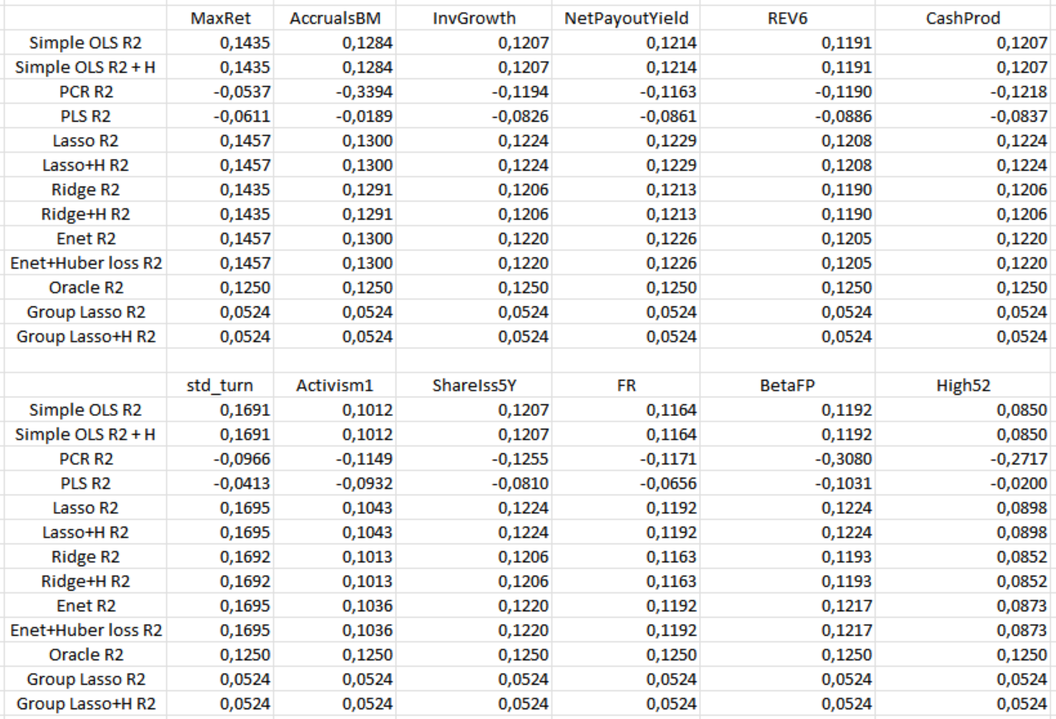
This table contains In sample values.

Table 3 – Out of sample values



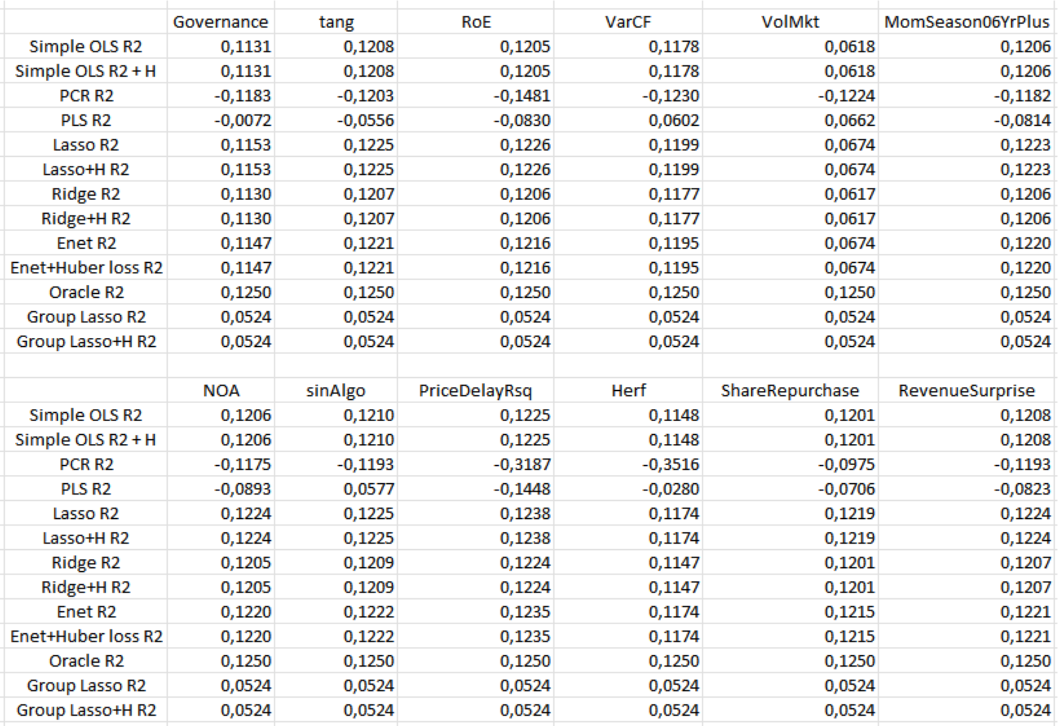
	betaVIX	roaq	Leverage	ShortInterest	Beta	OperProf
Simple OLS R2	0,1235	0,1198	0,1225	0,1061	0,1144	0,1206
Simple OLS R2 + H	0,1235	0,1198	0,1225	0,1061	0,1144	0,1206
PCR R2	-0,1193	-0,1143	-0,2217	-0,2282	-0,1459	-0,1505
PLS R2	-0,0816	-0,0896	-0,0921	-0,0684	0,0601	-0,0823
Lasso R2	0,1251	0,1216	0,1239	0,1088	0,1177	0,1225
Lasso+H R2	0,1251	0,1216	0,1239	0,1088	0,1177	0,1225
Ridge R2	0,1234	0,1197	0,1223	0,1060	0,1142	0,1207
Ridge+H R2	0,1234	0,1197	0,1223	0,1060	0,1142	0,1207
Enet R2	0,1246	0,1213	0,1233	0,1082	0,1177	0,1216
Enet+Huber loss R2	0,1246	0,1213	0,1233	0,1082	0,1177	0,1216
Oracle R2	0,1272	0,1248	0,1250	0,1250	0,1250	0,1250
Group Lasso R2	0,0524	0,0524	0,0524	0,0524	0,0524	0,0524
Group Lasso+H R2	0,0524	0,0524	0,0524	0,0524	0,0524	0,0524

	OperProf	BM	Accruals	Investment	ShareVol	BidAskSpread
Simple OLS R2	0,1206	0,1146	0,1191	0,1207	0,1083	-0,0873
Simple OLS R2 + H	0,1206	0,1146	0,1191	0,1207	0,1083	-0,0873
PCR R2	-0,1505	-0,4400	-0,1228	-0,1194	-0,1055	0,0648
PLS R2	-0,0823	-0,0918	0,0677	-0,0824	-0,1512	-0,2465
Lasso R2	0,1225	0,1178	0,1204	0,1224	0,1112	0,0162
Lasso+H R2	0,1225	0,1179	0,1204	0,1224	0,1112	0,0162
Ridge R2	0,1207	0,1146	0,1192	0,1206	0,1084	-0,0841
Ridge+H R2	0,1207	0,1146	0,1192	0,1206	0,1084	-0,0841
Enet R2	0,1216	0,1157	0,1204	0,1220	0,1106	0,0160
Enet+Huber loss R2	0,1216	0,1157	0,1204	0,1220	0,1106	0,0160
Oracle R2	0,1250	0,1250	0,1249	0,1250	0,1250	0,1250
Group Lasso R2	0,0524	0,0524	0,0524	0,0524	0,0524	0,0146
Group Lasso+H R2	0,0524	0,0524	0,0524	0,0524	0,0524	0,0146



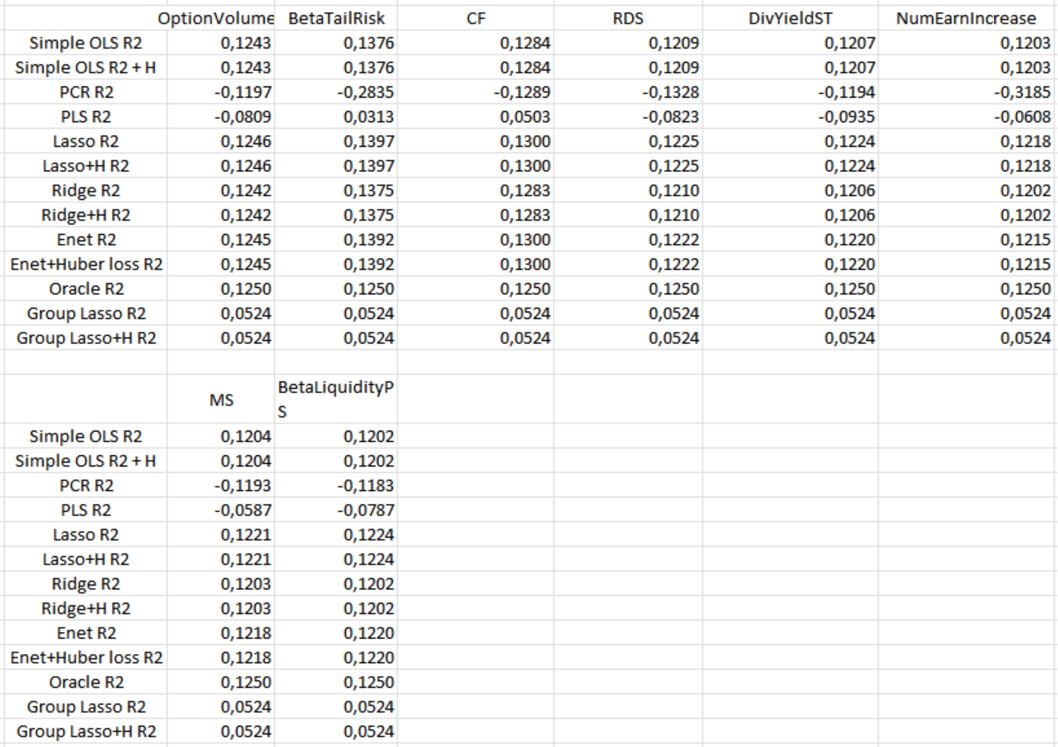
	MaxRet	AccrualsBM	InvGrowth	NetPayoutYield	REV6	CashProd
Simple OLS R2	0,1435	0,1284	0,1207	0,1214	0,1191	0,1207
Simple OLS R2 + H	0,1435	0,1284	0,1207	0,1214	0,1191	0,1207
PCR R2	-0,0537	-0,3394	-0,1194	-0,1163	-0,1190	-0,1218
PLS R2	-0,0611	-0,0189	-0,0826	-0,0861	-0,0886	-0,0837
Lasso R2	0,1457	0,1300	0,1224	0,1229	0,1208	0,1224
Lasso+H R2	0,1457	0,1300	0,1224	0,1229	0,1208	0,1224
Ridge R2	0,1435	0,1291	0,1206	0,1213	0,1190	0,1206
Ridge+H R2	0,1435	0,1291	0,1206	0,1213	0,1190	0,1206
Enet R2	0,1457	0,1300	0,1220	0,1226	0,1205	0,1220
Enet+Huber loss R2	0,1457	0,1300	0,1220	0,1226	0,1205	0,1220
Oracle R2	0,1250	0,1250	0,1250	0,1250	0,1250	0,1250
Group Lasso R2	0,0524	0,0524	0,0524	0,0524	0,0524	0,0524
Group Lasso+H R2	0,0524	0,0524	0,0524	0,0524	0,0524	0,0524

	std_turn	Activism1	ShareIss5Y	FR	BetaFP	High52
Simple OLS R2	0,1691	0,1012	0,1207	0,1164	0,1192	0,0850
Simple OLS R2 + H	0,1691	0,1012	0,1207	0,1164	0,1192	0,0850
PCR R2	-0,0966	-0,1149	-0,1255	-0,1171	-0,3080	-0,2717
PLS R2	-0,0413	-0,0932	-0,0810	-0,0656	-0,1031	-0,0200
Lasso R2	0,1695	0,1043	0,1224	0,1192	0,1224	0,0898
Lasso+H R2	0,1695	0,1043	0,1224	0,1192	0,1224	0,0898
Ridge R2	0,1692	0,1013	0,1206	0,1163	0,1193	0,0852
Ridge+H R2	0,1692	0,1013	0,1206	0,1163	0,1193	0,0852
Enet R2	0,1695	0,1036	0,1220	0,1192	0,1217	0,0873
Enet+Huber loss R2	0,1695	0,1036	0,1220	0,1192	0,1217	0,0873
Oracle R2	0,1250	0,1250	0,1250	0,1250	0,1250	0,1250
Group Lasso R2	0,0524	0,0524	0,0524	0,0524	0,0524	0,0524
Group Lasso+H R2	0,0524	0,0524	0,0524	0,0524	0,0524	0,0524



	Governance	tang	RoE	VarCF	VolMkt	MomSeason06YrPlus
Simple OLS R2	0,1131	0,1208	0,1205	0,1178	0,0618	0,1206
Simple OLS R2 + H	0,1131	0,1208	0,1205	0,1178	0,0618	0,1206
PCR R2	-0,1183	-0,1203	-0,1481	-0,1230	-0,1224	-0,1182
PLS R2	-0,0072	-0,0556	-0,0830	0,0602	0,0662	-0,0814
Lasso R2	0,1153	0,1225	0,1226	0,1199	0,0674	0,1223
Lasso+H R2	0,1153	0,1225	0,1226	0,1199	0,0674	0,1223
Ridge R2	0,1130	0,1207	0,1206	0,1177	0,0617	0,1206
Ridge+H R2	0,1130	0,1207	0,1206	0,1177	0,0617	0,1206
Enet R2	0,1147	0,1221	0,1216	0,1195	0,0674	0,1220
Enet+Huber loss R2	0,1147	0,1221	0,1216	0,1195	0,0674	0,1220
Oracle R2	0,1250	0,1250	0,1250	0,1250	0,1250	0,1250
Group Lasso R2	0,0524	0,0524	0,0524	0,0524	0,0524	0,0524
Group Lasso+H R2	0,0524	0,0524	0,0524	0,0524	0,0524	0,0524

	NOA	sinAlgo	PriceDelayRsq	Herf	ShareRepurchase	RevenueSurprise
Simple OLS R2	0,1206	0,1210	0,1225	0,1148	0,1201	0,1208
Simple OLS R2 + H	0,1206	0,1210	0,1225	0,1148	0,1201	0,1208
PCR R2	-0,1175	-0,1193	-0,3187	-0,3516	-0,0975	-0,1193
PLS R2	-0,0893	0,0577	-0,1448	-0,0280	-0,0706	-0,0823
Lasso R2	0,1224	0,1225	0,1238	0,1174	0,1219	0,1224
Lasso+H R2	0,1224	0,1225	0,1238	0,1174	0,1219	0,1224
Ridge R2	0,1205	0,1209	0,1224	0,1147	0,1201	0,1207
Ridge+H R2	0,1205	0,1209	0,1224	0,1147	0,1201	0,1207
Enet R2	0,1220	0,1222	0,1235	0,1174	0,1215	0,1221
Enet+Huber loss R2	0,1220	0,1222	0,1235	0,1174	0,1215	0,1221
Oracle R2	0,1250	0,1250	0,1250	0,1250	0,1250	0,1250
Group Lasso R2	0,0524	0,0524	0,0524	0,0524	0,0524	0,0524
Group Lasso+H R2	0,0524	0,0524	0,0524	0,0524	0,0524	0,0524

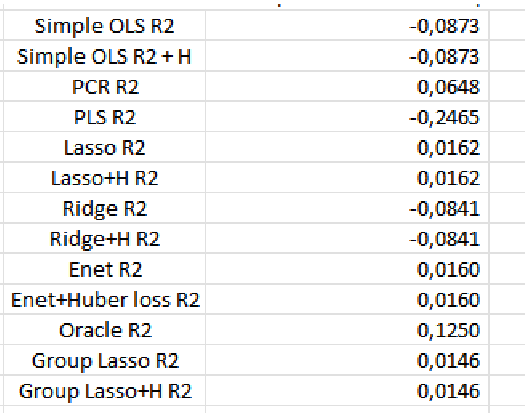


	OptionVolume	BetaTailRisk	CF	RDS	DivYieldST	NumEarnIncrease
Simple OLS R2	0,1243	0,1376	0,1284	0,1209	0,1207	0,1203
Simple OLS R2 + H	0,1243	0,1376	0,1284	0,1209	0,1207	0,1203
PCR R2	-0,1197	-0,2835	-0,1289	-0,1328	-0,1194	-0,3185
PLS R2	-0,0809	0,0313	0,0503	-0,0823	-0,0935	-0,0608
Lasso R2	0,1246	0,1397	0,1300	0,1225	0,1224	0,1218
Lasso+H R2	0,1246	0,1397	0,1300	0,1225	0,1224	0,1218
Ridge R2	0,1242	0,1375	0,1283	0,1210	0,1206	0,1202
Ridge+H R2	0,1242	0,1375	0,1283	0,1210	0,1206	0,1202
Enet R2	0,1245	0,1392	0,1300	0,1222	0,1220	0,1215
Enet+Huber loss R2	0,1245	0,1392	0,1300	0,1222	0,1220	0,1215
Oracle R2	0,1250	0,1250	0,1250	0,1250	0,1250	0,1250
Group Lasso R2	0,0524	0,0524	0,0524	0,0524	0,0524	0,0524
Group Lasso+H R2	0,0524	0,0524	0,0524	0,0524	0,0524	0,0524

	MS	BetaLiquidityPS
Simple OLS R2	0,1204	0,1202
Simple OLS R2 + H	0,1204	0,1202
PCR R2	-0,1193	-0,1183
PLS R2	-0,0587	-0,0787
Lasso R2	0,1221	0,1224
Lasso+H R2	0,1221	0,1224
Ridge R2	0,1203	0,1202
Ridge+H R2	0,1203	0,1202
Enet R2	0,1218	0,1220
Enet+Huber loss R2	0,1218	0,1220
Oracle R2	0,1250	0,1250
Group Lasso R2	0,0524	0,0524
Group Lasso+H R2	0,0524	0,0524

Contains the out of sample values from regressions and every stock characteristic.

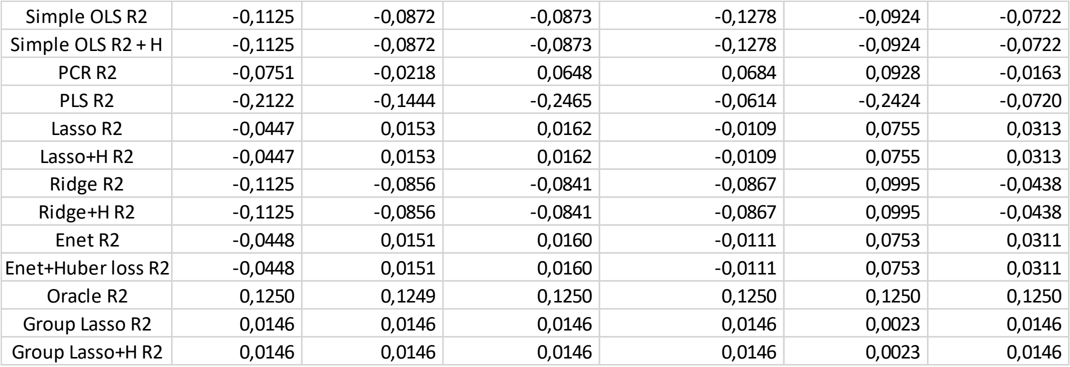
Table 4 – In sample values excluding bid-ask spread



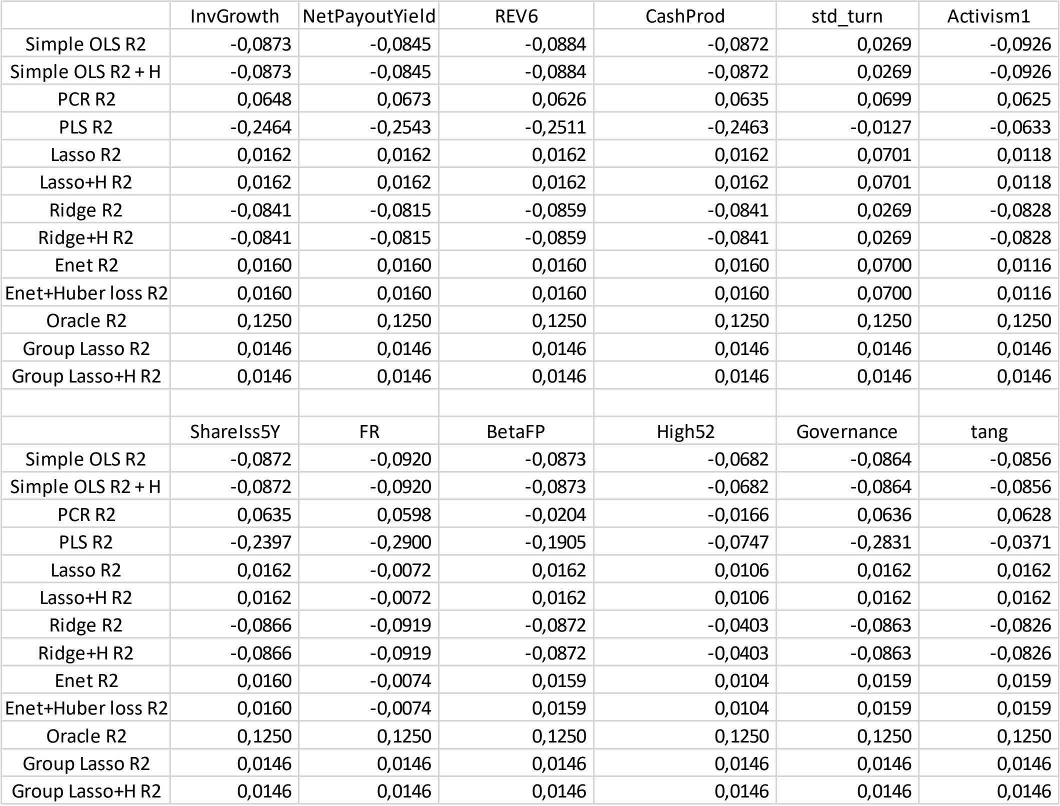
Simple OLS R2	-0,0873
Simple OLS R2 + H	-0,0873
PCR R2	0,0648
PLS R2	-0,2465
Lasso R2	0,0162
Lasso+H R2	0,0162
Ridge R2	-0,0841
Ridge+H R2	-0,0841
Enet R2	0,0160
Enet+Huber loss R2	0,0160
Oracle R2	0,1250
Group Lasso R2	0,0146
Group Lasso+H R2	0,0146

This table contains In sample values when excluding bid-ask spread

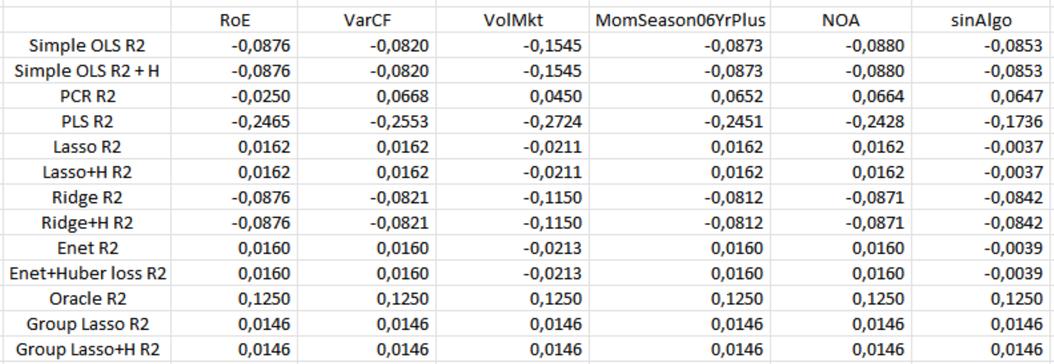
Table 5 – Out of sample values excluding bid-ask spread



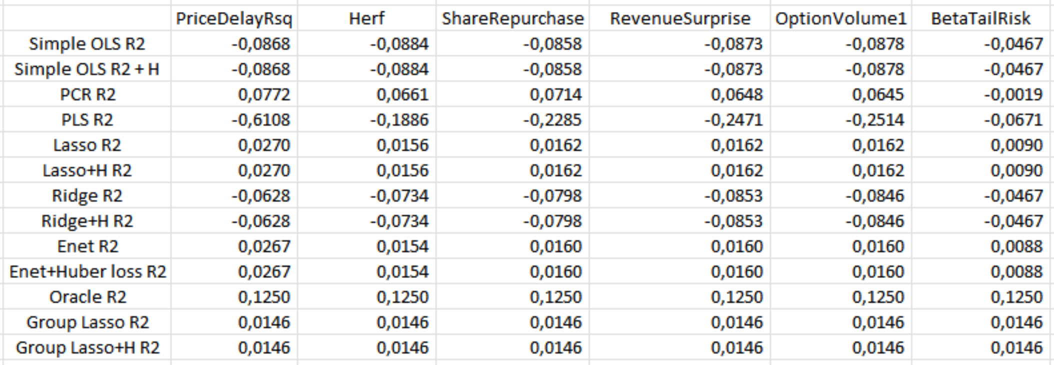
Simple OLS R2	-0,1125	-0,0872	-0,0873	-0,1278	-0,0924	-0,0722
Simple OLS R2 + H	-0,1125	-0,0872	-0,0873	-0,1278	-0,0924	-0,0722
PCR R2	-0,0751	-0,0218	0,0648	0,0684	0,0928	-0,0163
PLS R2	-0,2122	-0,1444	-0,2465	-0,0614	-0,2424	-0,0720
Lasso R2	-0,0447	0,0153	0,0162	-0,0109	0,0755	0,0313
Lasso+H R2	-0,0447	0,0153	0,0162	-0,0109	0,0755	0,0313
Ridge R2	-0,1125	-0,0856	-0,0841	-0,0867	0,0995	-0,0438
Ridge+H R2	-0,1125	-0,0856	-0,0841	-0,0867	0,0995	-0,0438
Enet R2	-0,0448	0,0151	0,0160	-0,0111	0,0753	0,0311
Enet+Huber loss R2	-0,0448	0,0151	0,0160	-0,0111	0,0753	0,0311
Oracle R2	0,1250	0,1249	0,1250	0,1250	0,1250	0,1250
Group Lasso R2	0,0146	0,0146	0,0146	0,0146	0,0023	0,0146
Group Lasso+H R2	0,0146	0,0146	0,0146	0,0146	0,0023	0,0146



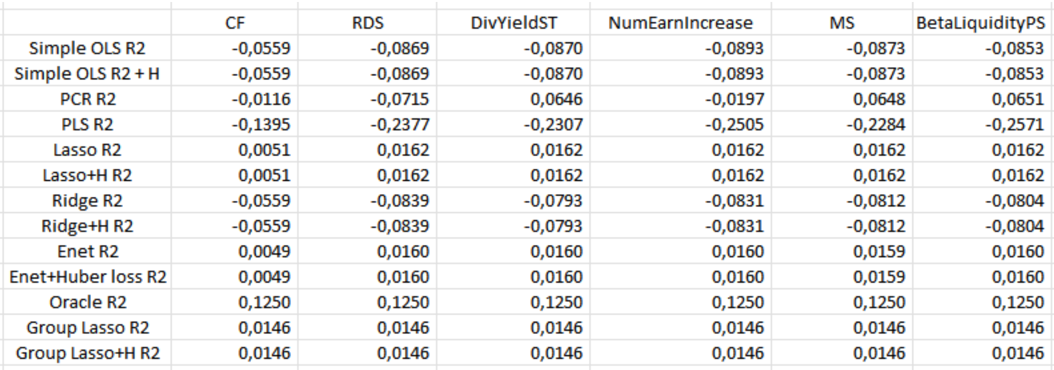
	InvGrowth	NetPayoutYield	REV6	CashProd	std_turn	Activism1
Simple OLS R2	-0,0873	-0,0845	-0,0884	-0,0872	0,0269	-0,0926
Simple OLS R2 + H	-0,0873	-0,0845	-0,0884	-0,0872	0,0269	-0,0926
PCR R2	0,0648	0,0673	0,0626	0,0635	0,0699	0,0625
PLS R2	-0,2464	-0,2543	-0,2511	-0,2463	-0,0127	-0,0633
Lasso R2	0,0162	0,0162	0,0162	0,0162	0,0701	0,0118
Lasso+H R2	0,0162	0,0162	0,0162	0,0162	0,0701	0,0118
Ridge R2	-0,0841	-0,0815	-0,0859	-0,0841	0,0269	-0,0828
Ridge+H R2	-0,0841	-0,0815	-0,0859	-0,0841	0,0269	-0,0828
Enet R2	0,0160	0,0160	0,0160	0,0160	0,0700	0,0116
Enet+Huber loss R2	0,0160	0,0160	0,0160	0,0160	0,0700	0,0116
Oracle R2	0,1250	0,1250	0,1250	0,1250	0,1250	0,1250
Group Lasso R2	0,0146	0,0146	0,0146	0,0146	0,0146	0,0146
Group Lasso+H R2	0,0146	0,0146	0,0146	0,0146	0,0146	0,0146
	ShareIss5Y	FR	BetaFP	High52	Governance	tang
Simple OLS R2	-0,0872	-0,0920	-0,0873	-0,0682	-0,0864	-0,0856
Simple OLS R2 + H	-0,0872	-0,0920	-0,0873	-0,0682	-0,0864	-0,0856
PCR R2	0,0635	0,0598	-0,0204	-0,0166	0,0636	0,0628
PLS R2	-0,2397	-0,2900	-0,1905	-0,0747	-0,2831	-0,0371
Lasso R2	0,0162	-0,0072	0,0162	0,0106	0,0162	0,0162
Lasso+H R2	0,0162	-0,0072	0,0162	0,0106	0,0162	0,0162
Ridge R2	-0,0866	-0,0919	-0,0872	-0,0403	-0,0863	-0,0826
Ridge+H R2	-0,0866	-0,0919	-0,0872	-0,0403	-0,0863	-0,0826
Enet R2	0,0160	-0,0074	0,0159	0,0104	0,0159	0,0159
Enet+Huber loss R2	0,0160	-0,0074	0,0159	0,0104	0,0159	0,0159
Oracle R2	0,1250	0,1250	0,1250	0,1250	0,1250	0,1250
Group Lasso R2	0,0146	0,0146	0,0146	0,0146	0,0146	0,0146
Group Lasso+H R2	0,0146	0,0146	0,0146	0,0146	0,0146	0,0146



	RoE	VarCF	VolMkt	MomSeason06YrPlus	NOA	sinAlgo
Simple OLS R2	-0,0876	-0,0820	-0,1545	-0,0873	-0,0880	-0,0853
Simple OLS R2 + H	-0,0876	-0,0820	-0,1545	-0,0873	-0,0880	-0,0853
PCR R2	-0,0250	0,0668	0,0450	0,0652	0,0664	0,0647
PLS R2	-0,2465	-0,2553	-0,2724	-0,2451	-0,2428	-0,1736
Lasso R2	0,0162	0,0162	-0,0211	0,0162	0,0162	-0,0037
Lasso+H R2	0,0162	0,0162	-0,0211	0,0162	0,0162	-0,0037
Ridge R2	-0,0876	-0,0821	-0,1150	-0,0812	-0,0871	-0,0842
Ridge+H R2	-0,0876	-0,0821	-0,1150	-0,0812	-0,0871	-0,0842
Enet R2	0,0160	0,0160	-0,0213	0,0160	0,0160	-0,0039
Enet+Huber loss R2	0,0160	0,0160	-0,0213	0,0160	0,0160	-0,0039
Oracle R2	0,1250	0,1250	0,1250	0,1250	0,1250	0,1250
Group Lasso R2	0,0146	0,0146	0,0146	0,0146	0,0146	0,0146
Group Lasso+H R2	0,0146	0,0146	0,0146	0,0146	0,0146	0,0146



	PriceDelayRsq	Herf	ShareRepurchase	RevenueSurprise	OptionVolume1	BetaTailRisk
Simple OLS R2	-0,0868	-0,0884	-0,0858	-0,0873	-0,0878	-0,0467
Simple OLS R2 + H	-0,0868	-0,0884	-0,0858	-0,0873	-0,0878	-0,0467
PCR R2	0,0772	0,0661	0,0714	0,0648	0,0645	-0,0019
PLS R2	-0,6108	-0,1886	-0,2285	-0,2471	-0,2514	-0,0671
Lasso R2	0,0270	0,0156	0,0162	0,0162	0,0162	0,0090
Lasso+H R2	0,0270	0,0156	0,0162	0,0162	0,0162	0,0090
Ridge R2	-0,0628	-0,0734	-0,0798	-0,0853	-0,0846	-0,0467
Ridge+H R2	-0,0628	-0,0734	-0,0798	-0,0853	-0,0846	-0,0467
Enet R2	0,0267	0,0154	0,0160	0,0160	0,0160	0,0088
Enet+Huber loss R2	0,0267	0,0154	0,0160	0,0160	0,0160	0,0088
Oracle R2	0,1250	0,1250	0,1250	0,1250	0,1250	0,1250
Group Lasso R2	0,0146	0,0146	0,0146	0,0146	0,0146	0,0146
Group Lasso+H R2	0,0146	0,0146	0,0146	0,0146	0,0146	0,0146



	CF	RDS	DivYieldST	NumEarnIncrease	MS	BetaLiquidityPS
Simple OLS R2	-0,0559	-0,0869	-0,0870	-0,0893	-0,0873	-0,0853
Simple OLS R2 + H	-0,0559	-0,0869	-0,0870	-0,0893	-0,0873	-0,0853
PCR R2	-0,0116	-0,0715	0,0646	-0,0197	0,0648	0,0651
PLS R2	-0,1395	-0,2377	-0,2307	-0,2505	-0,2284	-0,2571
Lasso R2	0,0051	0,0162	0,0162	0,0162	0,0162	0,0162
Lasso+H R2	0,0051	0,0162	0,0162	0,0162	0,0162	0,0162
Ridge R2	-0,0559	-0,0839	-0,0793	-0,0831	-0,0812	-0,0804
Ridge+H R2	-0,0559	-0,0839	-0,0793	-0,0831	-0,0812	-0,0804
Enet R2	0,0049	0,0160	0,0160	0,0160	0,0159	0,0160
Enet+Huber loss R2	0,0049	0,0160	0,0160	0,0160	0,0159	0,0160
Oracle R2	0,1250	0,1250	0,1250	0,1250	0,1250	0,1250
Group Lasso R2	0,0146	0,0146	0,0146	0,0146	0,0146	0,0146
Group Lasso+H R2	0,0146	0,0146	0,0146	0,0146	0,0146	0,0146

Contains the out of sample from regressions when excluding the bid-ask spread characteristic.