Cross-Sectional Regression Models for Financial Returns

Heman Kolla, kollah@rpi.edu

ITWS 4600-02: Data Analytics

Dr. Ahmed Eleish

Dec 12, 2025

Abstract and Introduction

This project explores a common question posed by all financial analysts: what financial features best predict a firm's return. This cannot be answered by standard time-series regression models. Instead, leveraging financial features and returns for many years, this project employs cross-sectional regression. In this methodology, each firm-month observation is treated as an independent data point with financial feature input used to predict the output, which are firms' contemporaneous return.

The motivation behind this project originates from both practical finance applications and prior work with quantum machine learning (QML). In previous work, I implemented a quantum neural network that performed a similar regression task. However, the resulting model exhibited poor predictive abilities. As such, the aim of this project is to validate a classical regression model to establish a baseline for quantum cross-sectional regression in future work.

The hypothesis is that, despite the expansive financial features in the data, there will be little to no statistically significant evidence that the model can predict returns. In applying two different classical regression techniques, Multiple Linear Regression & Support Vector Regression, the project aims to confirm or deny this hypothesis, providing insight into the predictive limits of cross-sectional regression for financial returns.

Data Description & Preliminary Analysis

A. Data Description

The dataset used in this project was sourced from two financial research publications in ML-based approaches to empirical asset pricing **[1][2]**. These sources provide firm-level financial characters (referred to above and hereafter as financial features) that have been rigorously validated in prior literature as being

the top 20 most informative predictors of realized stock returns. See Table 1 for a detailed list of financial features and their corresponding identifiers in the dataset. This dataset was selected for this project in part based on its broad set of financial features that are suitable for cross-sectional analysis of stock return predictions. Additionally, its extensive use in prior academic work lets it serve as a reliable baseline when aiming to reproduce results.

Table 1: Financial Features. Adapted from [1].

Acronym (from dataset)	Financial Feature Name	Source
mom1m	1-month momentum	[2]
mom12m	12-month momentum	[2]
chmom	Change in 6-month momentum	[2]
indmom	Industry momentum	[2]
mom36m	36 month momentum	[2]
turn	Share turnover	[2]
mvel1	Size	[2]
dolvol	Dollar trading volume	[2]
ill	Illiquidity	[2]
zerotrade	Zero trading days	[2]
baspread	Bid-ask spread	[2]
retvol	Return volatility	[2]
idiovol	Idiosyncratic return volatility	[2]
beta	Beta	[2]
betasq	Beta squared	[2]
ep	Earnings to price	[2]
sp	Sales to price	[2]
agr	Asset growth	[2]
nincr	Number of earnings increases	[2]
return (t-1)	Past return	[1]

The dataset contains stock measurements for approximately 9,420 firms observed over a 20-year period, which spans 228 periods at a monthly frequency. In total, the data consists of 891,616 rows, or firm-month observations. For each observation the dataset includes the date, 20 financial features detailed in Table 1, and a target variable, which is defined as the firm's realized stock in that current observation date.

Each observation additionally has a field "permno", which is a permanent numerical identifier assigned by the Center for Research in Security Prices (CRSP). This enables each firm to be uniquely identified in the dataset across time, even under varying scenarios such as M&A (mergers and acquisitions) or name changes.

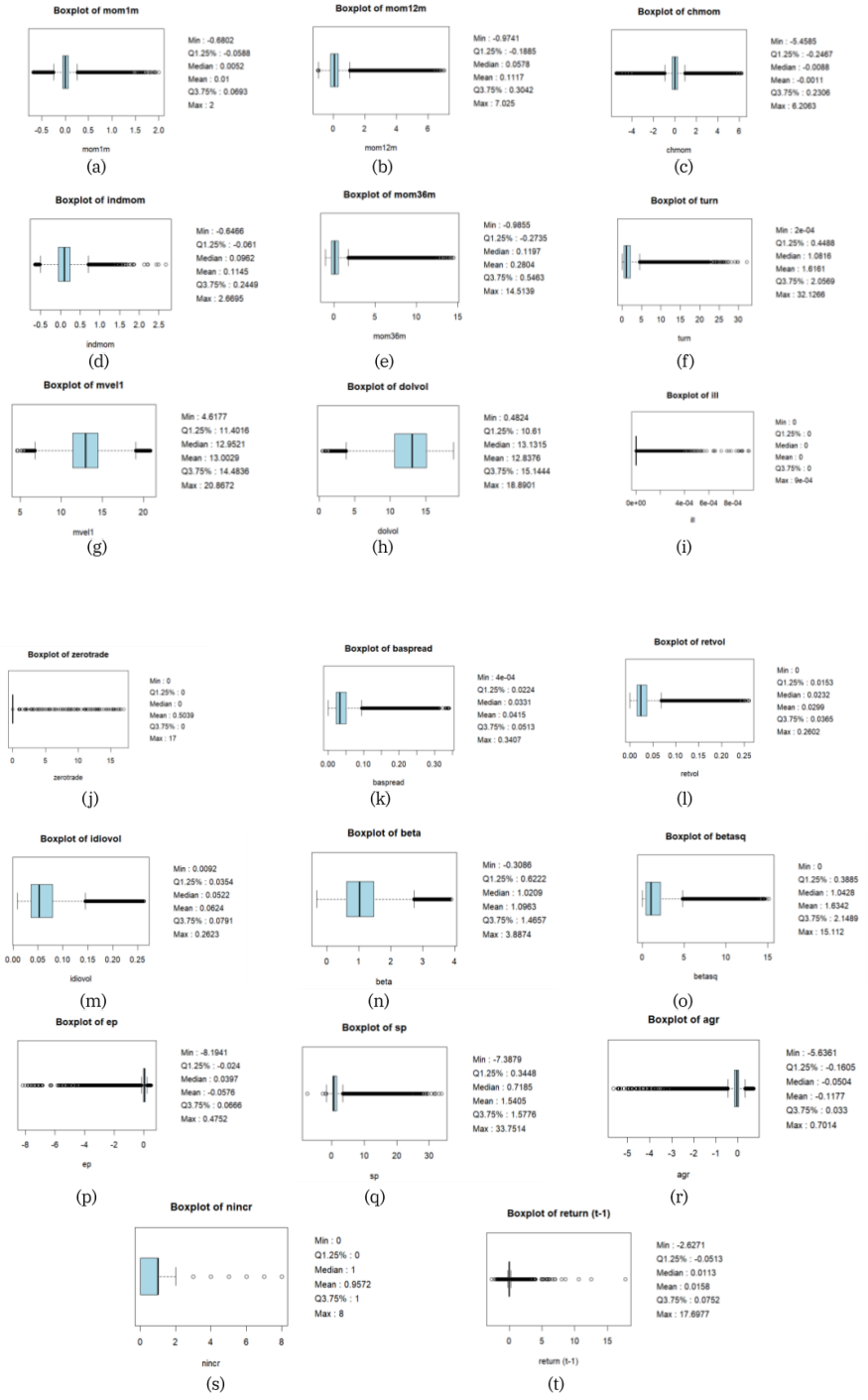
In the cross-sectional framework of this project, each of these firm-month observations is treated as an independent data point, with financial features becoming inputs used to predict contemporaneous returns.

The included features form a broad set of widely-recognized indicators of stock returns. These features can be further grouped into several categories, including: momentum based measures, liquidity and trading activity metrics, risk and volatility measures, and valuation-related ratios. Momentum-based measures are measures of past returns over varying windows of time. Liquidity-related variables like bid-ask spread and illiquidity reflect the disparities with the market when a market participant such as the firm makes a decision involving it. Risk measures quantify both firm-level risk (i.e. return volatility) and systematic market-level risks (i.e. beta). Valuation-related ratios capture information related to a firm's balance sheet.

B. Preliminary Analysis

With respect to preliminary analysis, initial boxplots and summary statistics were generated for all 20 financial feature variables to understand distributional properties prior to any transformations or modeling decisions. The following in Fig. 1 specify the min, Q1, media, Q3, max, mean, and outliers for each feature individually. The visualizations display noticeable differences in scale, degrees of skewness, and quantity of outliers, motivating significant transformations and preprocessing

Figure 1: Preliminary Boxplots & Summary Statistics for Financial Features



Exploratory Analysis

A. Data Cleaning

Prior to conducting any exploratory analysis, the dataset first underwent cleaning to ensure preserved data retained no missing values (NA or NaN) across the 20 feature variables and the target return. This process ensured compatibility when training the downstream regression models, and aggressively reduced the dataset to do so.

The singular intentional exception made was for the first month in the sample. The variable corresponding to lagged return ("return (t-1)") in these dates is undefined by design as no prior months exist in the dataset.

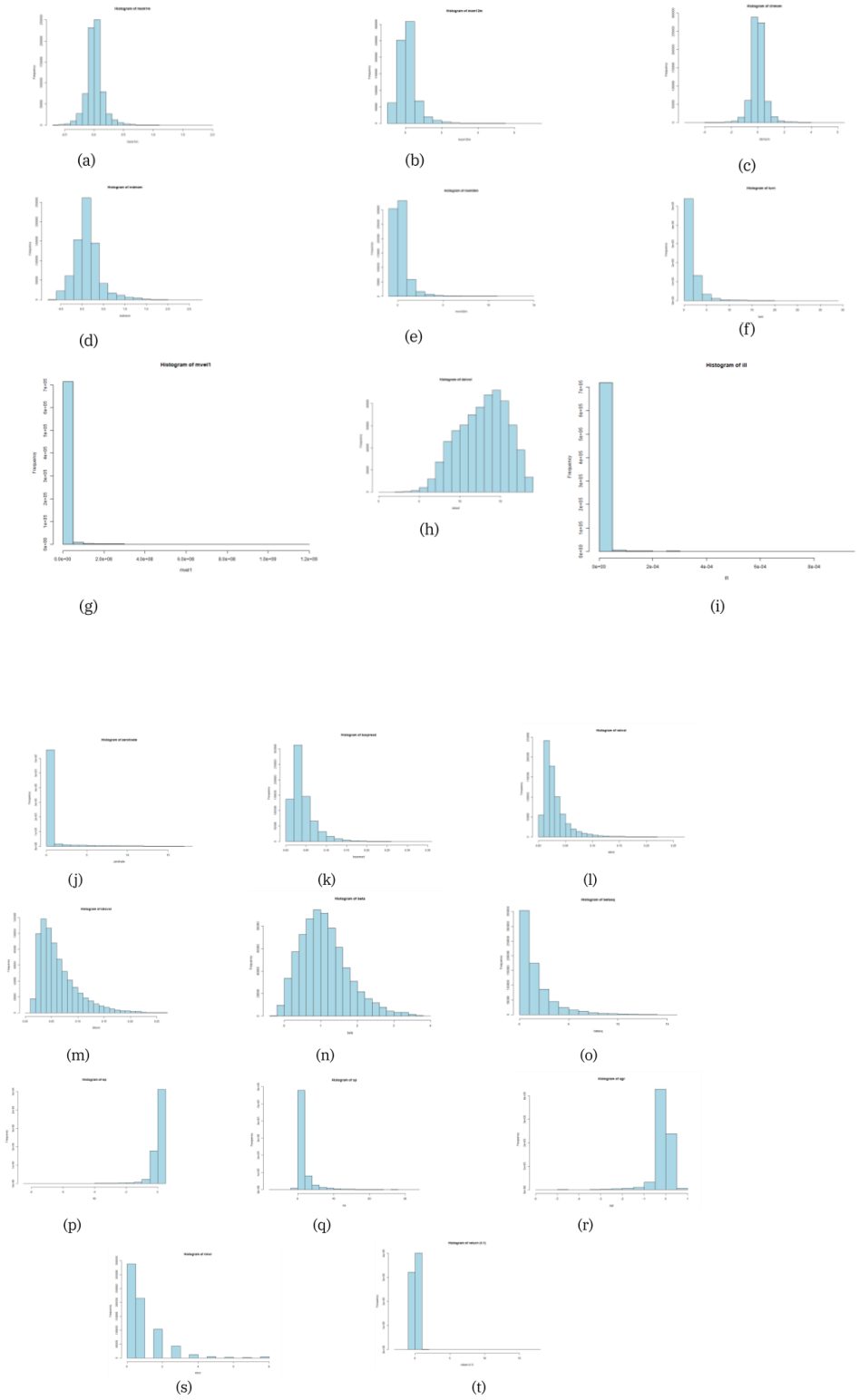
The next phase of cleaning looked at outliers, which were revealed in Fig. 1 during preliminary analysis. From a purely data analytics and machine learning perspective, such observations would impact the final summary statistics and influence the training of downstream models.

However, to be consistent with the methodology from **[1]**, outliers were not cleaned from the dataset. From a financial perspective, extreme observations correspond to specific economic events (i.e. market stress, firm-specific news, etc). As such, removing those observations would be intentionally disregarding valuable financial information. Therefore, this decision to retain outliers prioritizes economic fidelity. The negative effects of outliers were instead addressed through data transformations and modeling choices.

B. Distributional Analysis & Data Transformations

Following data cleaning, exploratory analysis can be performed. However, the initial boxplots generated are no longer accurate to the working dataframe. Consequently, to understand the marginal distributions of the financial features, I generated frequency histograms for each of the 20 financial features using the cleaned data in Fig. 2. Note that all features in Fig. 1 and Fig. 2 follow the sequential ordering of Table 1.

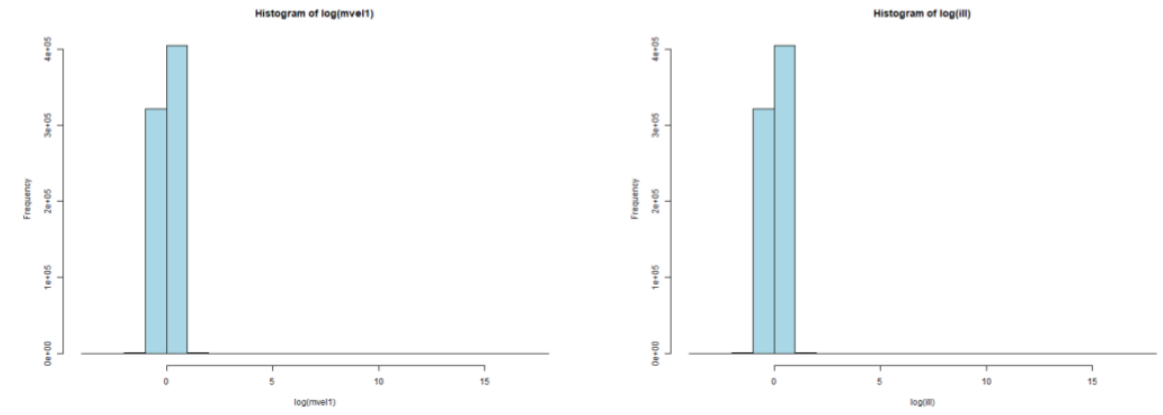
Figure 2: Frequency Histograms for Financial Features



The histograms reveal several features' distributions to be moderately skewed. Of the skewed distributions, there are two variables in particular—market value (Fig. 1g) and illiquidity (Fig. 1i)—that exhibit a strong right skew, having most data points fall into one bin despite the extreme range on the x-axis. In addition to making the associated figures less insightful, this creates concerns for downstream modeling as extreme values can dominate loss functions.

To address this, logarithmic transformations are applied to illiquidity ("ill") and market value ("mvel1"). In doing so, the x-axis is scaled down, which lets observations spread across more bins. As modeled in Fig. 3, the transformed frequency histograms exhibit improved x-axis scaling that resolve the prior concern.

Figure 3: Frequency Histograms for log(mvel1) and log(ill)



(a) (b)

C. Assessing Correlation & Multicollinearity

The intuitive first approach to understand the relationship between a financial feature and return is to plot a scatterplot. However, with over 700,000 observations, such plots would only reveal visual blobs, being insightful. As a result, we examine feature-to-target relationships based on a correlation analysis instead.

Table 2: Feature-to-Return Correlation

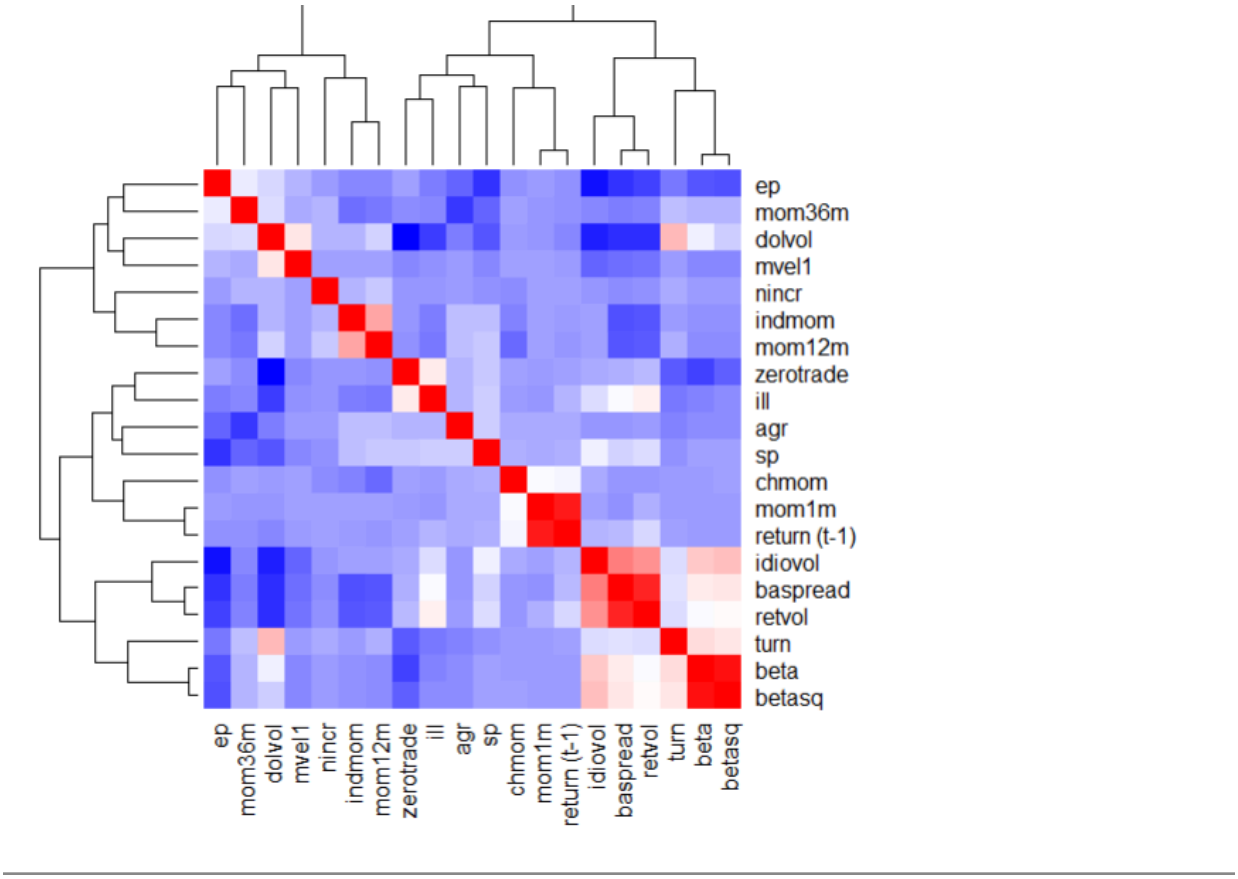
Feature	Correlation with Return (to the nearest ten-thousandth
ill	0.0799
baspread	0.0788
retvol	0.0743
dolvol	-0.0631
idiovol	0.0499
sp	0.0385
turn	-0.0356
ep	-0.0312
mom36m	-0.0285
ar	-0.0279
mom1m	-0.0268
zerotrade	0.0623
mom12m	-0.0208
chmom	-0.0139
indmom	-0.0131
mvel1	-0.0114
beta	-0.0069
betasq	-0.0068
nincr	0.0021
return (t-1)	NA

The low correlation with return in Table 2 for most features indicate that no single feature exhibits strong correlation, and by extension, linear predictive power, with returns. This is consistent with our expectations and reinforces the

necessity for a specifically cross-sectional regression modeling approach to test our hypothesis.

Secondly, a feature to-feature correlation analysis was performed to identify collinearity among predictors. The results were plotted in Figure 4.

Figure 4: Feature-to-Feature Correlation Matrix as a Heatmap



In this heatmap depiction of Fig. 4, red represents a positive correlation while blue represents a negative correlation. The boldness of the color represents the strength of the correlation.

Fig. 4 reveals numerous intuitive patterns. Risk-related measures exhibit stronger positive correlations with one another. Similarly, momentum-based measures exhibit the same internal correlation. These patterns confirm the prior grouping of variables into economic categories—momentum, risk, liquidity, and

valuation. It is important to note, though, that such correlation does not hold for all groupings, and Fig. 4 must be referred to for pre-modeling correlation.

This leads us to the conclusion that, while some degree of multicollinearity is present, it less reflects data quality issues and more reflects the grouping based on economic characteristics.

D. Sources of Uncertainty & Bias

There do exist sources of uncertainty and possible bias with the design decisions made. For one, the choice of retaining specific lagged return NA values on first month observations can introduce survivorship bias. This is because the data cleaning phase is not applied uniformly to the entire dataframe. Secondly, the choice to retain extreme outliers, though some transformed, is still likely to increase variance in model estimates despite the economic motivations. This may also exaggerate the effects of survivorship bias.

Model Development & Application of Models

A. Multiple Linear Regression

To serve as a baseline model, I implemented a Multiple Linear Regression (MLR) model to examine the linear predictive power of the combined set of financial features. This facilitates direct benchmarking with the factor models from **[1]**, and allows for easier interpretability of the results.

The model itself is specified to have contemporaneous stock return as the target feature (dependent variable) with the full set of 20 financial features being the input (independent variables). Following the data cleanup and transformations outlined in the exploratory analysis section, a combined dataframe was then constructed.

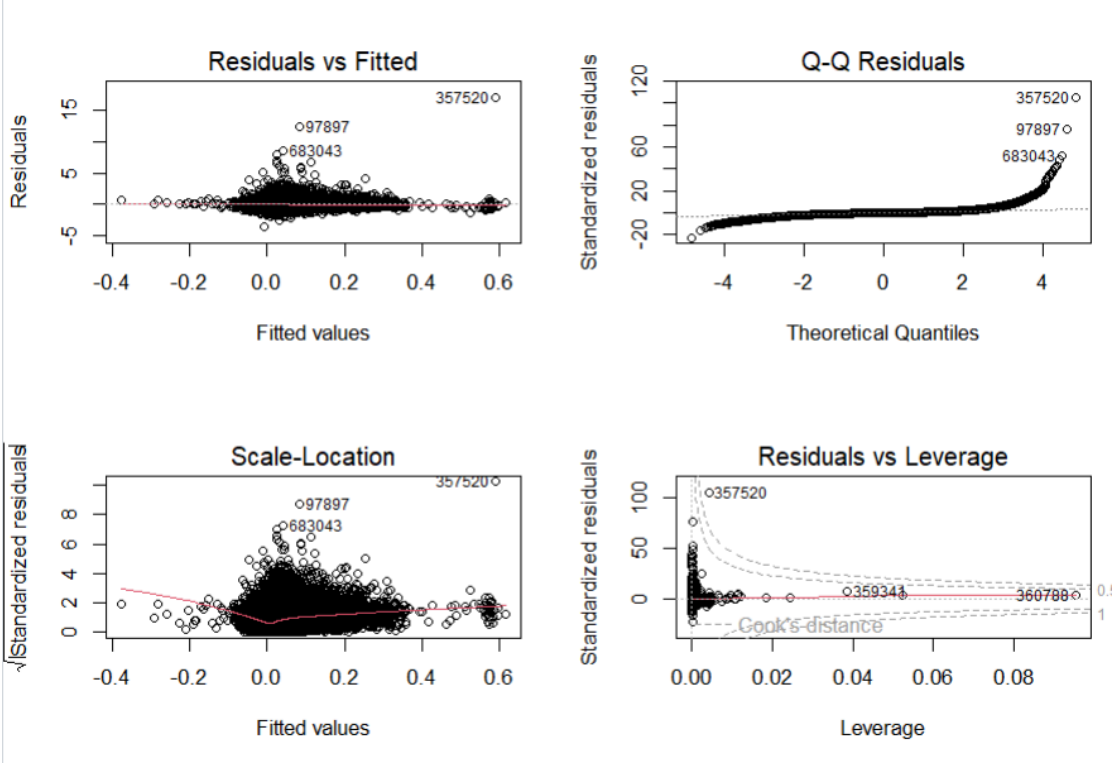
Then, the model was estimated using the default Ordinary Least Squares (OLS) method without additional regularization or feature selection. This provides a benchmark against potential optimizations made to the model. The summary statistics are displayed in Fig. 5.

Figure 5: R Summary Output for MLR Model

```
Call:
lm(formula = y ~ ., data = df_model_one)

Residuals:
    Min      1Q  Median      3Q     Max
-3.6927 -0.0649 -0.0007  0.0618 17.1069

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   8.181e-03  1.640e-03   4.988 6.11e-07 ***
mom1m         1.767e-02  3.398e-03   5.201 1.98e-07 ***
mom12m       -2.324e-03  4.221e-04  -5.507 3.65e-08 ***
chmom        -1.183e-03  3.765e-04  -3.143  0.00167 **
indmom        3.077e-03  7.646e-04   4.024 5.71e-05 ***
mom36m       -1.829e-03  2.141e-04  -8.543  < 2e-16 ***
turn         -2.569e-03  1.317e-04 -19.512  < 2e-16 ***
mvel1         1.182e-11  1.039e-11   1.138  0.25521
dolvol       -3.692e-04  1.157e-04  -3.191  0.00142 **
ill           5.614e+02  1.200e+01  46.770  < 2e-16 ***
zerotrade    -6.234e-04  1.252e-04  -4.978 6.42e-07 ***
baspread      1.873e-01  1.582e-02  11.837  < 2e-16 ***
retvol        8.249e-02  1.941e-02   4.250 2.14e-05 ***
idiovol       4.049e-02  8.304e-03   4.876 1.08e-06 ***
beta          5.937e-03  9.603e-04   6.182 6.33e-10 ***
betasq       -3.440e-03  3.262e-04 -10.545  < 2e-16 ***
ep           -1.690e-03  5.415e-04  -3.121  0.00180 **
sp            9.745e-04  7.957e-05  12.248  < 2e-16 ***
agr           8.570e-03  5.236e-04  16.368  < 2e-16 ***
nincr         1.061e-03  1.527e-04   6.947 3.74e-12 ***
return..t.1. -3.581e-02  2.975e-03 -12.035  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1633 on 723180 degrees of freedom
Multiple R-squared:  0.0113,    Adjusted R-squared:  0.01127
F-statistic: 413.3 on 20 and 723180 DF,  p-value: < 2.2e-16
```

The model yielded an R-squared value of 0.0113. In context, this means that approximately 1.13% of the variance in monthly stock returns can be explained by the 20 financial features. The R-squared value being this low indicates that the MLR model, based on statistical evidence, performs poorly. To further assess model quality, we can examine the residuals, plotted in Fig. 6.

Figure 6: Residual Diagnostic Plots for MLR Model



The residual plots reflect several issues. First, the upper left diagram in Fig. 6 reports the presence of extreme residual values, which come from the outliers intentionally retained for economic integrity. As expected the tradeoff for economic integrity did negatively impact the model. Additionally the residual plot did not showcase randomness, and instead had a softly linear pattern centered at 0. Paired, these observations from Fig. 6 reaffirms the MLR model's poor performance from Fig. 5.

Combined, the Multiple Linear Regression model confirms the hypothesis since there is little statistical evidence to suggest that the set of financial features is a good predictor of monthly stock returns.

From a financial perspective, however, this outcome is entirely consistent with findings in prior literature in cross-sectional asset pricing. Monthly stock returns are recognized to be exceptionally small with most industry standard models typically yielding R-squared values from 0.01 to 0.035. Given this, the MLR model's 0.0113 R-squared value is financially viable, and performs similar to existing cross-sectional regression models belonging to larger banking and investment firms. Additionally, the

author of **[1]**, Dr. Clark reports R-squared values between 0.01 and 0.02, placing the MLR model in the acceptable range for theoretical standards that I am trying to replicate.

However, given the statistically insignificant evidence, I explore potential optimization by once again checking for multicollinearity. This time though, I do so after model training using the Variance Inflation Factor (VIF). VIF is a measure of how much of the variance of a feature's predictions is affected/inflated by correlations with another feature. Higher values correspond to greater multicollinearity and reduced reliability of individual feature estimates of return. See the VIF scores in Table 3.

Table 3: Variance Inflation Factors (VIFs) for each Financial Feature

Financial Feature	Variance Inflation Factor (VIF) to the nearest ten-thousandth
mom1m	6.4787
mom12m	1.4840
chmom	1.1139
indmom	1.4183
mom36m	1.1966
turn	2.2911
mvel1	14.6603
dolvol	18.0660
ill	1.2602
zerotrade	1.5856
baspread	5.7655
retvol	5.2415
idiovol	2.6543
beta	10.8254
betasq	10.1594
ep	1.3938
sp	1.2078
agr	1.1472
nincr	1.0207
return (t-1)	6.6648

Table 3 displays elevated VIF scores for many features. Moreover, the features that are conceptually related are shown to have similar values. This is exemplified by “beta” and “betasq” having VIF scores of 10.8254 and 10.1594, respectively, which make sense as one can be derived from the other.

More broadly, the clustering of similar VIF values reflecting their economic similarities reaffirms the groupings identified during feature-to-feature correlation in exploratory analysis. It is important to note that this demonstrated the preservation of correlation across predictors both before and after model regression.

Regardless, the multicollinearity concerns with reliability, displayed by high VIF values, persist. So, to address these concerns, I tested an optimization to perform a Principal Component Regression (PCR) model. The workflow was modified to scale the set of 20 input features and perform a Principal Component Analysis (PCA) reduction on the financial features. From there, the resulting principal components replaced the set of features in the model, whose summary statistics are displayed in Fig. 7.

Figure 7: R Summary Output for PCR Model

```
Call:
lm(formula = y ~ ., data = pc_data)

Residuals:
    Min      1Q  Median      3Q     Max
-3.7273 -0.0650 -0.0012  0.0623 17.3692

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.0149047  0.0001923  77.513  < 2e-16 ***
PC1         -0.0044797  0.0001016 -44.098  < 2e-16 ***
PC2         -0.0066947  0.0001188 -56.331  < 2e-16 ***
PC3          0.0037262  0.0001353  27.545  < 2e-16 ***
PC4          0.0004362  0.0001406   3.102  0.00192 **
PC5          0.0023939  0.0001750  13.683  < 2e-16 ***
PC6          0.0037709  0.0001851  20.375  < 2e-16 ***
PC7          0.0009844  0.0001956   5.033 4.84e-07 ***
PC8          0.0010404  0.0001979   5.258 1.46e-07 ***
PC9         -0.0019743  0.0002003  -9.856  < 2e-16 ***
PC10        -0.0002600  0.0002123  -1.225  0.22072
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1635 on 723190 degrees of freedom
Multiple R-squared:  0.009098,  Adjusted R-squared:  0.009085
F-statistic:   664 on 10 and 723190 DF,  p-value: < 2.2e-16
```

The PCR model yielded an R-squared value of 0.009098, which is lower than the 0.0113 by the MLR model. This indicates that, in this setting, implementing PCA did not improve model quality and instead removed important information encoded in the original set of features. As a result, this attempt at optimization is disregarded and retaining the prior MLR workflow results in a more informative regression model for return.

B. Support Vector Regression (SVR)

The second model employed is a Support Vector Regressor (SVR), which takes on a nonlinear modeling approach to see if it can provide better predictive power of returns given the same feature set. SVRs are suited to high-dimensional input settings as they employ kernel functions. As such, this is a natural extension beyond MLR.

Similar to the MLR, the SVR was trained on a cleaned and transformed dataframe, using the 20 financial features as predictors (independent variables) and contemporaneous monthly return as the target (dependent variable). Due to time and resource constraints with training SVRs on very large datasets, I performed random sampling to select 20,000 observations from the dataframe.

Before training the model, the workflow was modified to include some additional steps. All input features were scaled to have a standardized mean of 0 and unit variance. This step is vital as the optimization performed in an SVR would have otherwise viewde the magnitude of a feature's measurements to correspond with the feature's weight, which is inaccurate. After this is controlled for, the subsample was then split into training and testing sets using a 80/20 split.

Lastly, I had one more design choice to make before training the SVR—the kernel. Typically, the radial basis function (RBF) kernel is the default setting. However, due to it being computationally expensive, I went with a linear kernel as the computational cheaper alternative.

The model was then trained, and predictions were generated on the test split. Since I scaled the inputs to the SVR, I then unscaled the return predictions. Then, the model performance was assessed using a manual calculation of the R-squared metric. As the SVR workflow introduced an element of stochasticity with random sampling, I conducted multiple training runs. Two representative models' computed R-squared metrics are reported in Fig. 8 and Fig. 9.

Figure 8: Reported R-squared of SVR (Representative Trial 1)

```
> r2 <- 1 - sum((y_test_orig - y_pred_orig)^2) / sum((y_test_orig - mean(y_test_orig))^2)
> cat("SVR R-squared:", r2, "\n")
SVR R-squared: -0.004481713
```

Figure 9: Reported R-squared of SVR (Representative Trial 2)

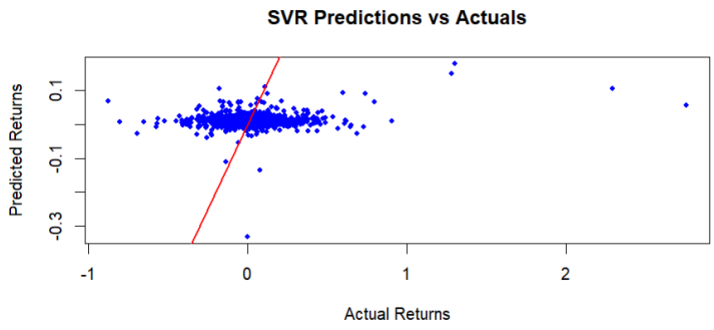
```
> r2 <- 1 - sum((y_test_orig - y_pred_orig)^2) / sum((y_test_orig - mean(y_test_orig))^2)
> cat("SVR R-squared:", r2, "\n")
SVR R-squared: 0.02224772
```

In Trial 1, the SVR reported an R-squared value of -0.004482. A negative R-squared indicates worse performance than a naive benchmark, and is said to fail to capture and predictive information from the input feature set about the target feature.

In Trial 2, the SVR reported an R-squared value of 0.022248, which exceeds the MLR model's R-squared of 0.0113. This indicates that the second model trained on a different sample of 20,000 firm-month observations still performs poorly, but performed better than the SVR model from Trial 1.

Despite the improved performance, such variation in reported R-squared values raises model reliability concerns. This can largely be attributed to the significantly smaller sample size, letting sampling bias heavily influence model performance. To further assess model quality we can plot SVR predicted returns for the better performing SVR model from Trial 1 in Fig. 10.

Figure 10: SVR (Trial 1) Predicted versus Actual Returns



The plot reaffirms the poor SVL model performance statistically, seeing as the cluster of points (observations) does not follow the slope of the 45-degree reference line. With little correspondence between predicted and actual returns in Fig. 10, it is shown that, even with positive R-squared values reported, the SVR model still performs poorly.

Taken together, the SVR results from R-squared metrics in Fig. 8 and Fig. 9, and the plot in Fig. 10, the model does not perform well and does not perform reliably. As such, the SVR model confirms that there is little statistical significance suggesting that the set of 20 financial features can predict monthly stock returns.

Conclusions & Discussion

This project investigated cross-sectional regression models for financial returns using classical techniques. The data was sourced from **[1]** and **[2]**, including monthly observations for 9,420 firms over a 20-year period. Each observation represents a snapshot of a single firm in a single month, and included 20 financial features (independent variables) and the contemporaneous return (dependent variable). The purpose of this project was to evaluate the hypothesis that there would be little to no statistically significant evidence that the feature set would be able to predict contemporaneous return.

Over the course of the project, the data underwent significant transformation. Initial preprocessing cleaned missing or invalid values, and additionally performed logistic transformations on select input features (“ill” and “mvel1”). This was then used to train two different models.

Model 1 was Multiple Linear Regression (MLR) and used Ordinary Least Squares (OLS) to regress returns on all 20 features. This model reported an R-squared value of 0.0113 and had a residual plot that lacked the identifiers of model fitness. As such, we concluded from Model 1 that the hypothesis was confirmed. In testing optimizations for this model, we looked towards Variance Inflation Factors (VIFs) and found high multicollinearity in the model, which was

not desirable. Consequently, we tested Principal Component Regression (PCR), which yielded a less, worse R-squared value of 0.00908. As such, we concluded that the optimization to PCR implementing Principal Component Analysis (PCA) was a bad design choice.

Model 2 was Support Vector Regression (SVR) and used a linear kernel to regress returns on all 20 features. This model had an inherently different approach than MLR as it aimed to use kernel functions to identify nonlinear relationships. One big design consideration with this model was its computational expensiveness. As it required significant time and resources that were unavailable, the dataset was sampled down to 20,000 for each trial, and a linear kernel was used in place of the default radial kernel. This model reported an R-squared value of 0.022248 in one trial and -0.004482 in another. This demonstrated poor model performance with consistently low R-squared values and unreliable model performance given negative R-squared value. Plotting predicted versus actual returns displayed poor results as well. As such, we also confirmed from Model 2 that the hypothesis was true.

Together, both models support the initial hypothesis that there is no statistically significant evidence for the feature set being a good predictor of returns. However, in a financial context, this project's findings are in the expected ranges or industry (0.01-0.035) and theoretical (0.01-0.02 **[1]**) standards for cross-sectional regression models.

With respect to future works, several improvements would be made. The primary of which is seeking high-performance computing resources. A full sample-training of nonlinear SVR models would give this project a more definitive answer on its predictive ability, and potentially disprove the initial hypothesis. This could look like moving away from R's primarily single-core execution model to CUDA-based code to leverage GPU parallelization to increase time efficiency, thereby allowing for larger samples to be executed. Though I cannot claim this would change our findings definitively, such an approach would better leverage the quantity of observations in the dataset, potentially providing additional insights.

References

[1] Clark, Brian J., Sai Palepu, and Akhtar R. Siddique. *Firm Complexity and Information Asymmetry: Evidence from ML-Based Complexity to Measure Information Processing Costs.* SSRN, 18 Mar 2020, https://papers.ssrn.com/spl3/papers.cfm?abstract_id=4763575.

[2] Gu, Shihao, Bryan T. Kelly, and Dacheng Xiu. "Emprical Asset Pricing via Machine Learning" *The Review of Financial Studies*, vol. 33, no. 5, 2022, pp.2223-2273. https://academic.oup.com/rfs/article/33/5/2223/5758276

[3] Meyer, David, Evgenia Dimitriadou, Kurt Hornik, Andreas Weingessel, and Friedrish Leisch. *e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071).* R Foundation for Statistical Computing, https://cran.r-project.org/package=e1071

[4] R Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, 2024, Vienna, Austria, www.r-project.org

[5] Wickham, Hadley. *Ggplot2: Elegant Graphics for Data Analysis.* Springer, 2016, https://ggplot2.tidyverse.org