BUSI722: Final Project Report

Quantitative Investment Strategy

Guide(s): Prof. Shmuel Baruch, Prof. Kerry Back

Submitted by:
Aditi Balaji (ab231)

Contents

Quantitative Investment Strategy

1 Problem Statement

The project entails conducting backtests on two investment strategies, specifically a 150/50 strategy (150% long and 50% short), followed by a detailed report. The report will include descriptions of the strategies, the procedure for backtesting them using historical data from a cloud-based SQL server, and an evaluation of the results against benchmarks like SPY and the three Fama-French factors. It will also discuss the rationale behind choosing one strategy for hypothetical implementation, highlighting its potential advantages. The analysis will utilize 5-6 key features based on the Green Hand and Zhan 2017 paper, with each feature described in detail according to its original research context. Additionally, the model will account for industry variations and trade on a monthly cycle to align with the data frequency.

2 Introduction

In this project, we aim to utilise past data and certain financial features to predict returns which in turn will help us determine the stocks to be bought. The overall process for this project is highlighted in figure 1. The specifics of the project will be explained in furhter details in the upcoming sections.

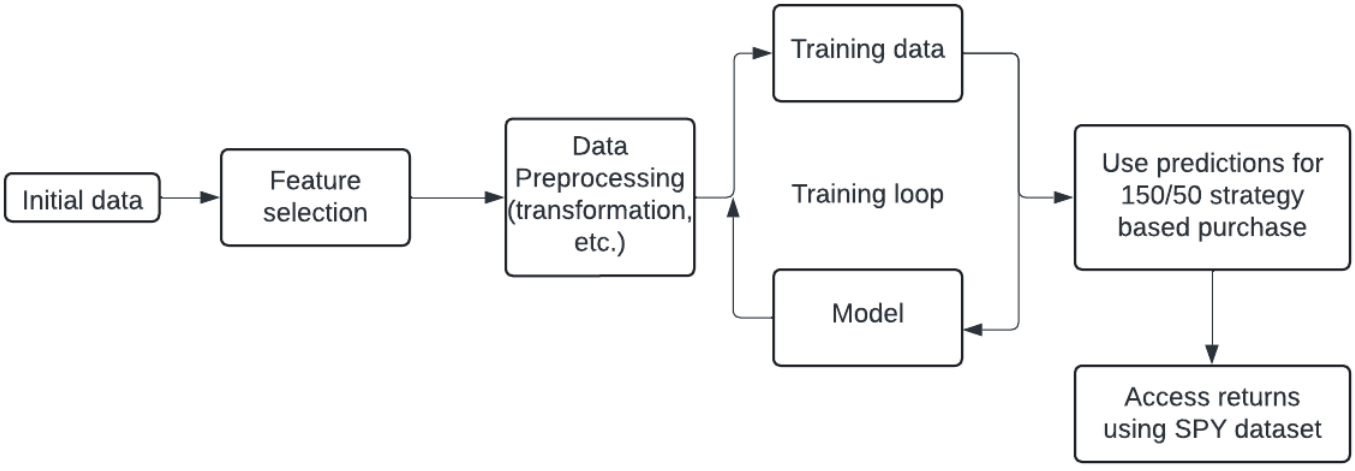


Figure 1: Process overview

3 Dataset

As mentioned in the project description, the dataset used is from the SQL server on the cloud, "mssql-827920.cloudclusters.net:16272". This section highlights dataset extraction, preprocessing and feature selection. As stated earlier, the data used is extracted from the cloud server listed and contains over 105 features. We select relavant features and perform data transformation for those features.

4 Preprocessing

There were 2 major steps in preprocessinh for this dataset.

1. Quantile Transformation: As we learnt and implemented during this course, the data was transformed using quantile tranformation for each month for better results. This is used to normalise skewed data.
2. Industry: siccd codes were used to determine the industry that a particular company/ticker belongs to.

4.1 Feature selection

For selecting features, we used 2 methods - a machine learning approach and a financial approach. All features are extracted from the Green Hand and Zhan 2017 paper [1]. For the machine learning based feature selection, the dataset was fit on a dummy machine learning model and the feature importance curve was extracted. While the main model used was catboost, the same exersice was conducted on linear regression model as well (refer Feature_analysis.ipynb). The following features were selected:

- **mom1m:** This represents the one-month momentum, capturing the past one month's stock performance, which can signal future short-term movements due to persistence in stock prices .
- **mom12m:** Twelve-month momentum indicates the stock's performance over the past year, often used to capture longer-term trends in price movements .
- **mom36m:** The thirty-six-month momentum is used to gauge the performance over the past three years, providing insights into the longer-term market sentiment towards a stock .
- **siccd:** Standard Industrial Classification Code, which categorizes industries based on their primary business activities .

- **ret:** This is the stock return, typically calculated as the percentage change in the stock price over a specified period .
- **sfe:** Scaled forecast earnings, relating analysts' earnings forecasts to the stock price, providing a measure of expected profitability .
- **rsup:** Revenue surprise, which measures the degree to which actual revenues exceed or fall short of analysts' revenue forecasts .
- **ear:** Earnings announcement return, capturing the stock's return around the announcement of earnings, which can indicate how the market reacts to new earnings information .
- **chmom:** Change in six-month momentum, highlighting how momentum has shifted over recent months .
- **idiovol:** Idiosyncratic volatility, which measures the component of a stock's volatility that is not explained by broader market movements, often linked to firm-specific news or events .

5 Model

For this project, Catboost model was used. It is an efficient form of XGBoost model where the categorical variables are accounted for. This proves to be advantages for 'industries' feature. Further, the boosting nature of this model helps accomodate the uncertain trends of stock.

6 Strategy

Three different strategies were implemented in this project. strategy 1 used mom1m feature while strategy 2 implemented mom12m feature and strategy 3 implemented mom36m. This difference helped to study which feature among these 3 significantly affects the returns.

7 Results

The results are as follows:

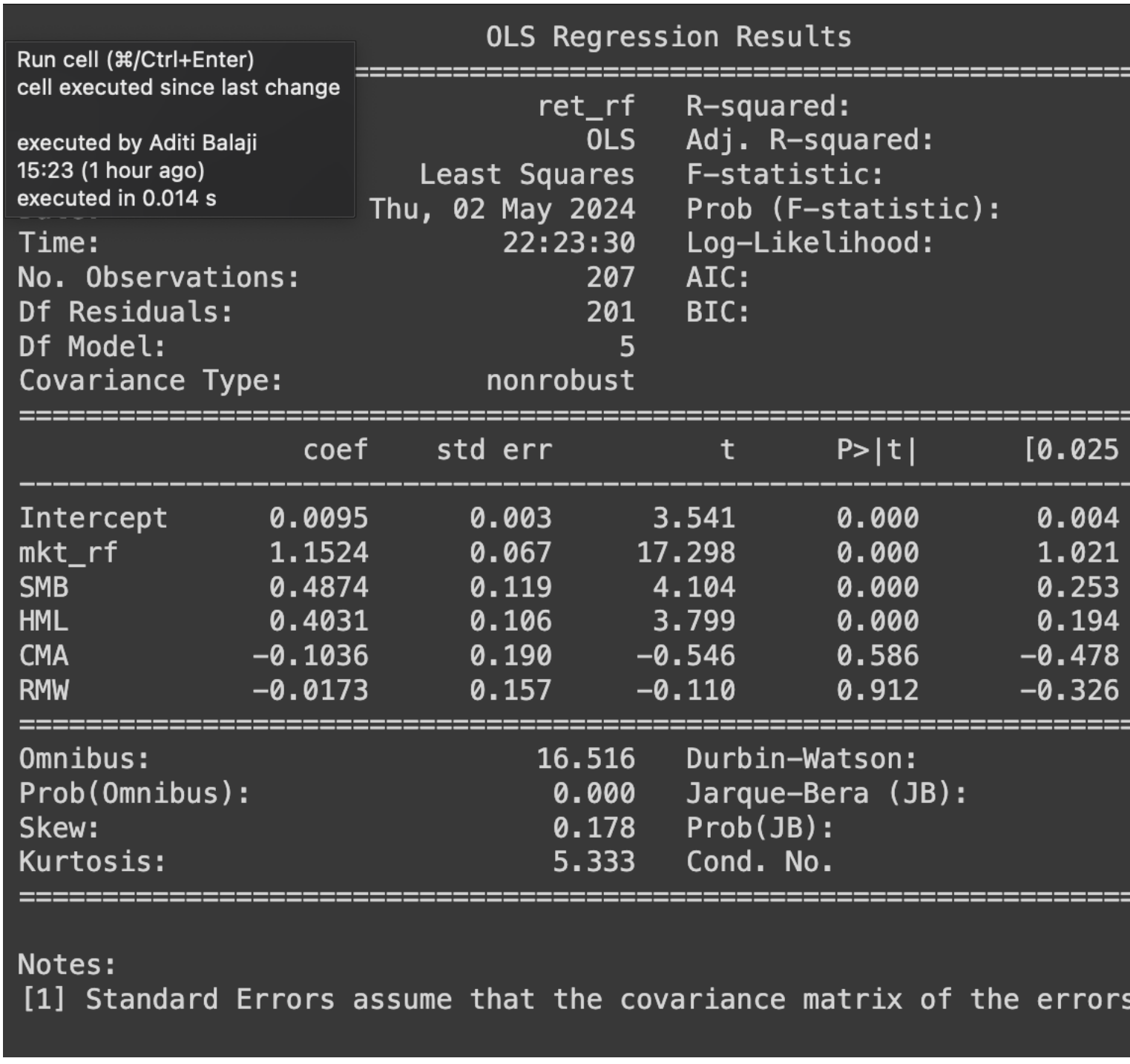
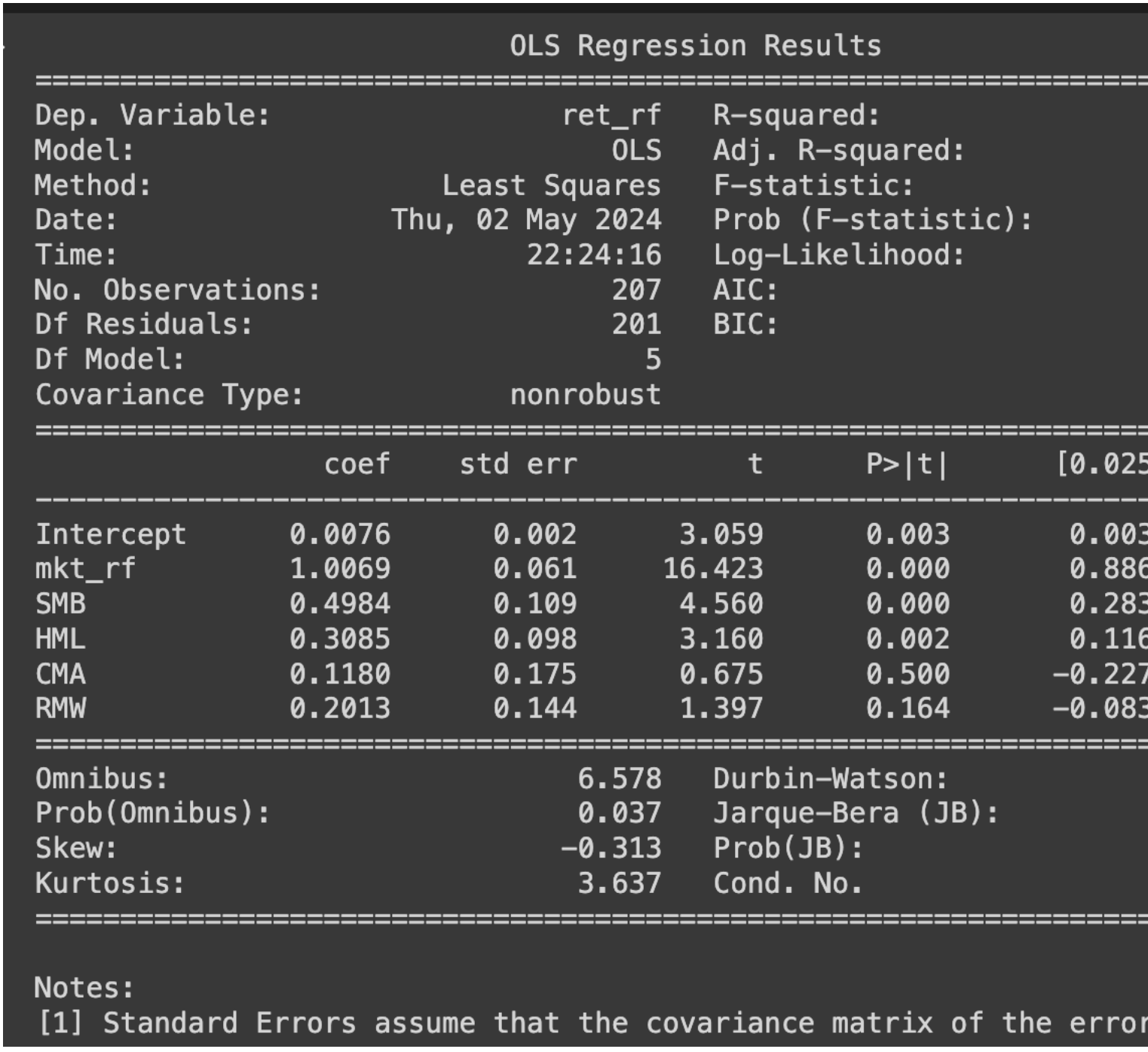


Figure 2: strategy 1



OLS Regression Results

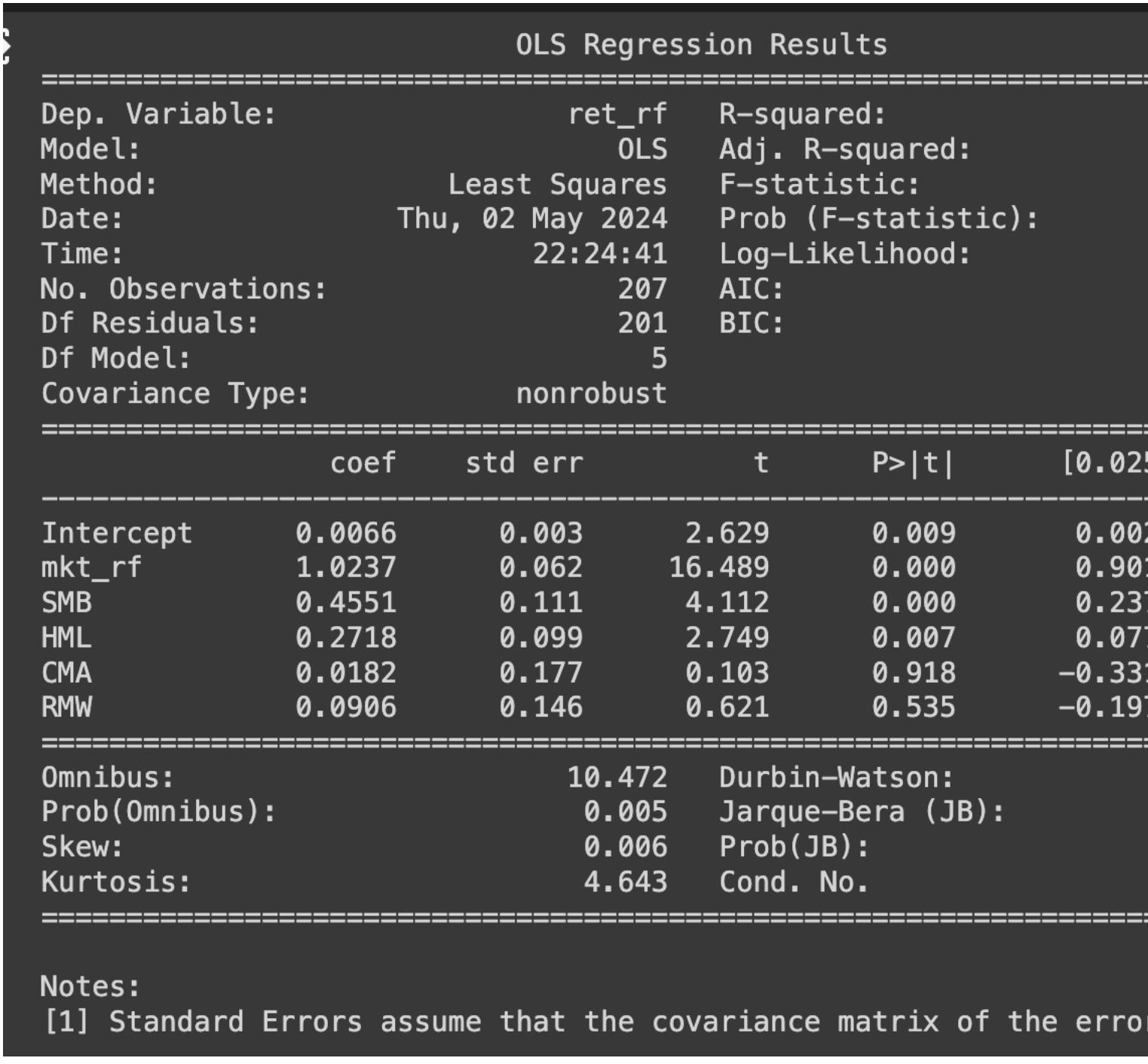
Dep. Variable:	ret_rf	R-squared:	
Model:	OLS	Adj. R-squared:	
Method:	Least Squares	F-statistic:	
Date:	Thu, 02 May 2024	Prob (F-statistic):	
Time:	22:24:16	Log-Likelihood:	
No. Observations:	207	AIC:	
Df Residuals:	201	BIC:	
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025
Intercept	0.0076	0.002	3.059	0.003	0.003
mkt_rf	1.0069	0.061	16.423	0.000	0.886
SMB	0.4984	0.109	4.560	0.000	0.283
HML	0.3085	0.098	3.160	0.002	0.116
CMA	0.1180	0.175	0.675	0.500	-0.227
RMW	0.2013	0.144	1.397	0.164	-0.083

Omnibus:	6.578	Durbin-Watson:	
Prob(Omnibus):	0.037	Jarque-Bera (JB):	
Skew:	-0.313	Prob(JB):	
Kurtosis:	3.637	Cond. No.	

Notes:
[1] Standard Errors assume that the covariance matrix of the error

Figure 3: Strategy 2



OLS Regression Results

Dep. Variable:	ret_rf	R-squared:	
Model:	OLS	Adj. R-squared:	
Method:	Least Squares	F-statistic:	
Date:	Thu, 02 May 2024	Prob (F-statistic):	
Time:	22:24:41	Log-Likelihood:	
No. Observations:	207	AIC:	
Df Residuals:	201	BIC:	
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025
Intercept	0.0066	0.003	2.629	0.009	0.001
mkt_rf	1.0237	0.062	16.489	0.000	0.901
SMB	0.4551	0.111	4.112	0.000	0.237
HML	0.2718	0.099	2.749	0.007	0.077
CMA	0.0182	0.177	0.103	0.918	-0.331
RMW	0.0906	0.146	0.621	0.535	-0.197

Omnibus:	10.472	Durbin-Watson:	
Prob(Omnibus):	0.005	Jarque-Bera (JB):	
Skew:	0.006	Prob(JB):	
Kurtosis:	4.643	Cond. No.	

Notes:
[1] Standard Errors assume that the covariance matrix of the erro

Figure 4: Strategy 3

References

[1] Jeremiah Green, John R. M. Hand, and X. Frank Zhang. *The Characteristics that Provide Independent Information about Average U.S. Monthly Stock Returns*. Mar. 2017. DOI: 10.1093/rfs/hhx019. URL: https://academic.oup.com/rfs/article-abstract/30/12/4389/3091648.