Predicting 3M Corporation's Weekly Stock Return using Linear Regression

Jay Kwon

Metis Data Science and Engineering (flex program)

Module 2 – Linear Regression and Web Scraping

November 10, 2021

Goal:

- Create a Linear Regression model to predict the weekly returns for 3M corporation's stock.
  1. optimize the performance of the model.
  2. gain insights on how to predict stock returns, including which features are most relevant

Motivation:

- insights gained would be beneficial to anyone interested in stock returns including:
  - **investors** (individuals, wealth management firms, hedge funds, pension funds)
  - **financial news** (Bloomberg, CNBC, Wall Street Journal)
  - **ratings agencies** (S&P, Moody's, Fitch)
  - **financial services**

DATASET:

- Target: Weekly return of 3M stock derived from historic prices scraped from https://finance.yahoo.com/quote/MMM/
- **Feb, 2010 to Feb, 2020 : 522 weeks**
- Features derived from historic price and volume: previous week's price, weekly avg volume traded
- Features from other data sources (Federal Reserve): inflation, unemployment, Fed Funds rate, and US treasury rate

- **Why 3M Corporation?** Less volatile, easier to predict
  - large, mature company with long history (founded 1902, part of S&P 500), not as volatile as a small growing company
  - Industrials sector is not as volatile as other sectors like Tech

Tools used:

- Selenium, BS, Numpy, Pandas, Matplotlib, Seaborn, **StatsModels**, **scikit-learn**

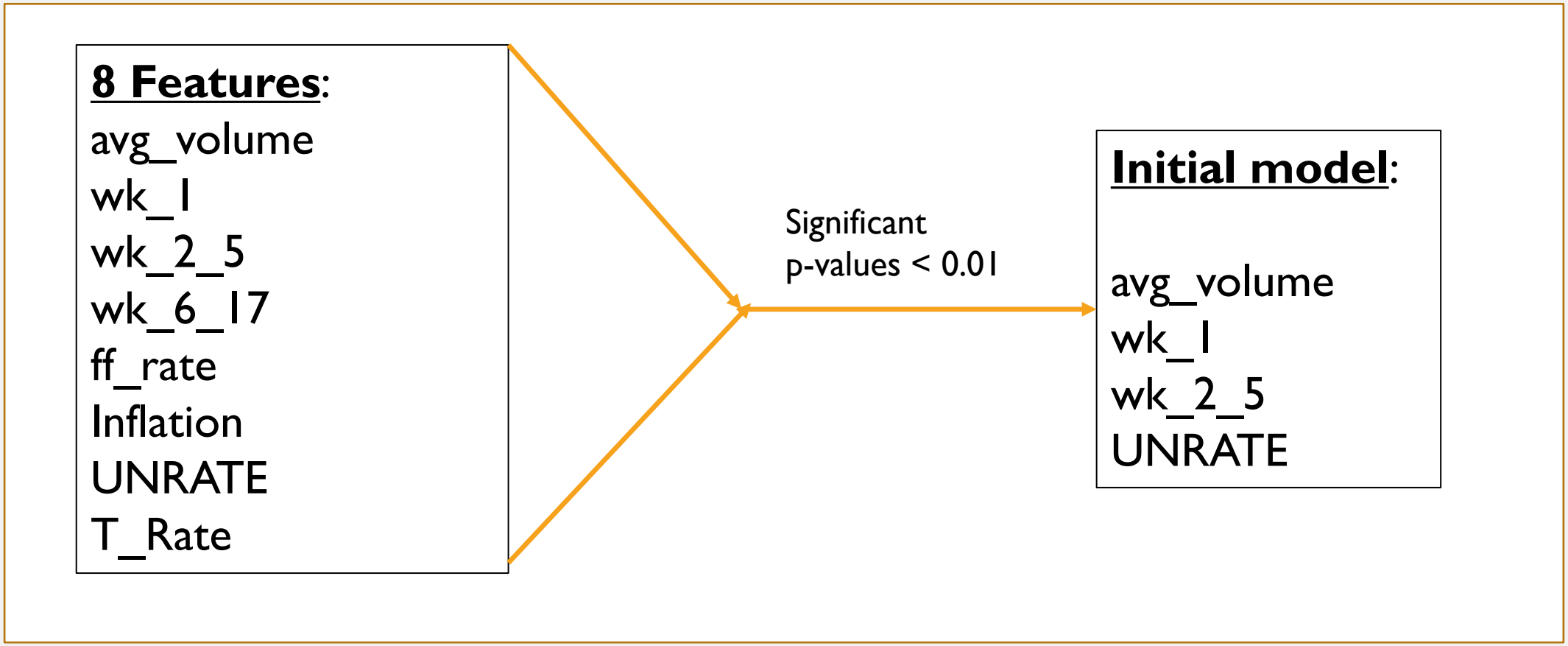
HISTORIC TREND OF STOCK MARKET

* Standard and Poor's 500 Index (source: Google)

Period of our Data: Feb 2010 to Feb 2020

- want to avoid extreme events that are hard to predict
- after 2008 financial crisis, before 2020 Covid-19 pandemic
- relatively stable bull market

Initial Model:



Takeaways:

- Recent historical returns may be more valuable predictors compared to older historical returns.
- Fed Funds Rate, Inflation, and 1-yr Treasuries are not good predictors of returns

Performance and Feature Interpretation:

OLS Regression Results

Dep. Variable:	wk_return	R-squared:	0.106
Model:	OLS	Adj. R-squared:	0.098
Method:	Least Squares	F-statistic:	12.28
Date:	Wed, 10 Nov 2021	Prob (F-statistic):	2.87e-11
Time:	01:19:21	Log-Likelihood:	1181.3
No. Observations:	522	AIC:	-2351.
Df Residuals:	516	BIC:	-2325.
Df Model:	5		
Covariance Type:	nonrobust		

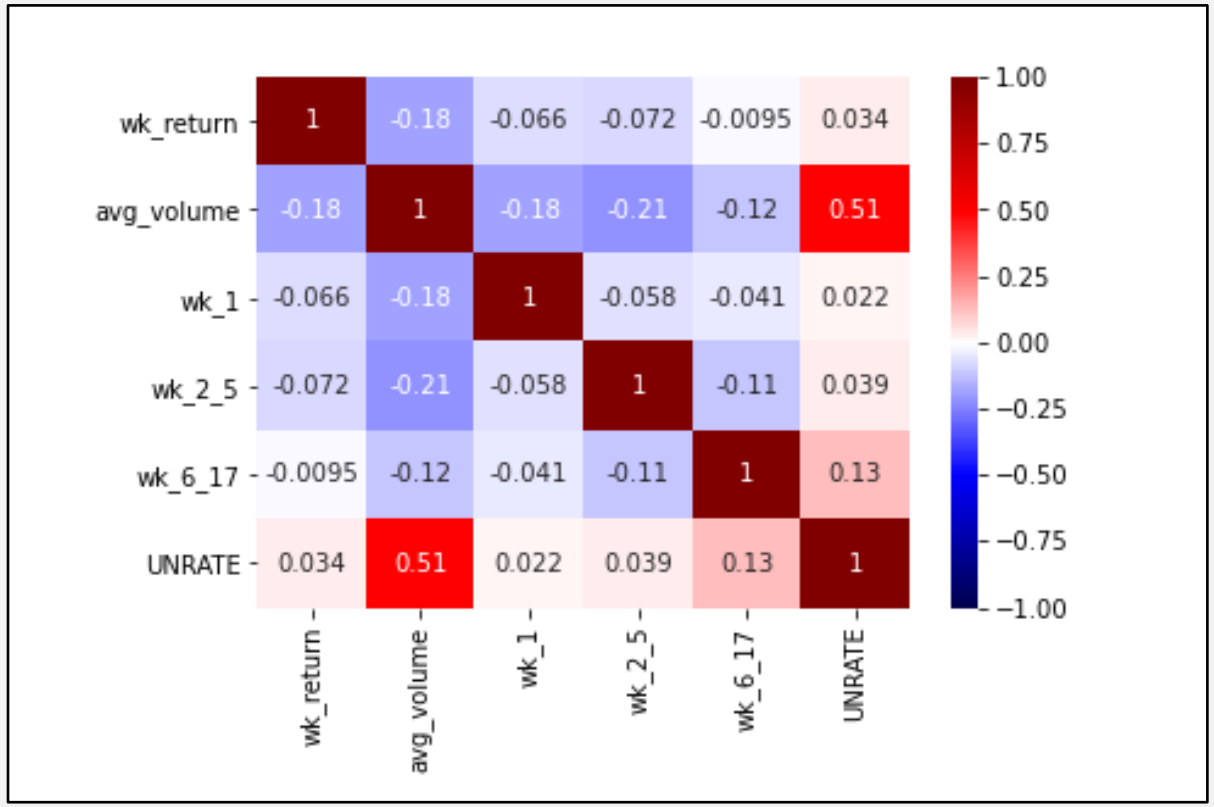
	coef	std err	t	P>|t|	[0.025	0.975]
const	0.0100	0.004	2.639	0.009	0.003	0.018
avg_volume	-9.562e-09	1.3e-09	-7.372	0.000	-1.21e-08	-7.01e-09
wk_1	-0.1629	0.044	-3.691	0.000	-0.250	-0.076
wk_2_5	-0.4021	0.095	-4.235	0.000	-0.589	-0.216
wk_6_17	-0.4676	0.176	-2.662	0.008	-0.813	-0.123
UNRATE	0.0034	0.001	5.072	0.000	0.002	0.005

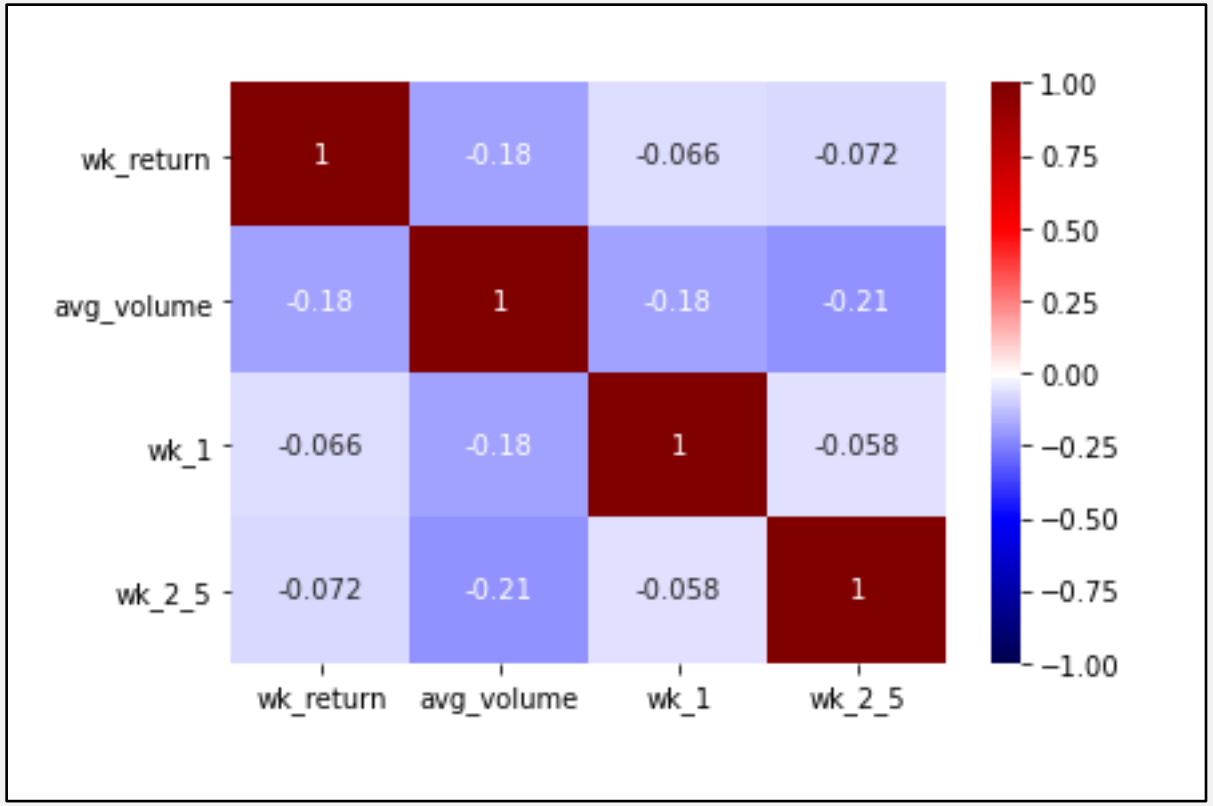
- Adj R-squared of 0.10 means that our features explain 10% of the variance of our target
- In the domain of finance, an R-squared of 0.20 is considered stellar

High condition number (not shown) suggest multicollinearity

- When average volume traded was high last week, the weekly return this week will be low
- When returns have been high in previous weeks, the weekly return this week will be low
- When unemployment is high, the returns will be high

Fixing Multicollinearity





- Problem: correlation between features: Average volume and Unemployment rate
- Keep Average Volume (stronger correlation of -0.18)

- Upon removing UNRATE, wk_6_17 became insignificant and had to be removed

- Our final model has 3 features, with avg_volume having the strongest correlation to our target.

Performance on Unseen Data

- Our final model has 3 features: previous week return, -2 to -5 week average return, and previous week's average volume (highest correlation to target, -0.18)
  - → R-squared: 0.059 Adj. R-squared: 0.053
  - → this is on the entire 522 rows that it was trained on

- 5-fold Cross Validation on final model
  - → R-squares: 0.041, -0.16, -0.075, -0.068, -0.012
  - → avg. R-square: **-0.055**
  - → significantly smaller than 0.059 we saw above.

- A negative R-squared of -0.055 means that **our model is performing worse than if we were to use the simple average** of the previous weekly stock returns (naïve model).

- Perhaps party affected by 20% less data used to train the model
- 522 rows total is a limited amount of data
- wide ranged in R-squared values calculated in the cross validation suggests that more of the same type of data is necessary

CONCLUSIONS AND RECOMMENDATIONS:

Conclusions:

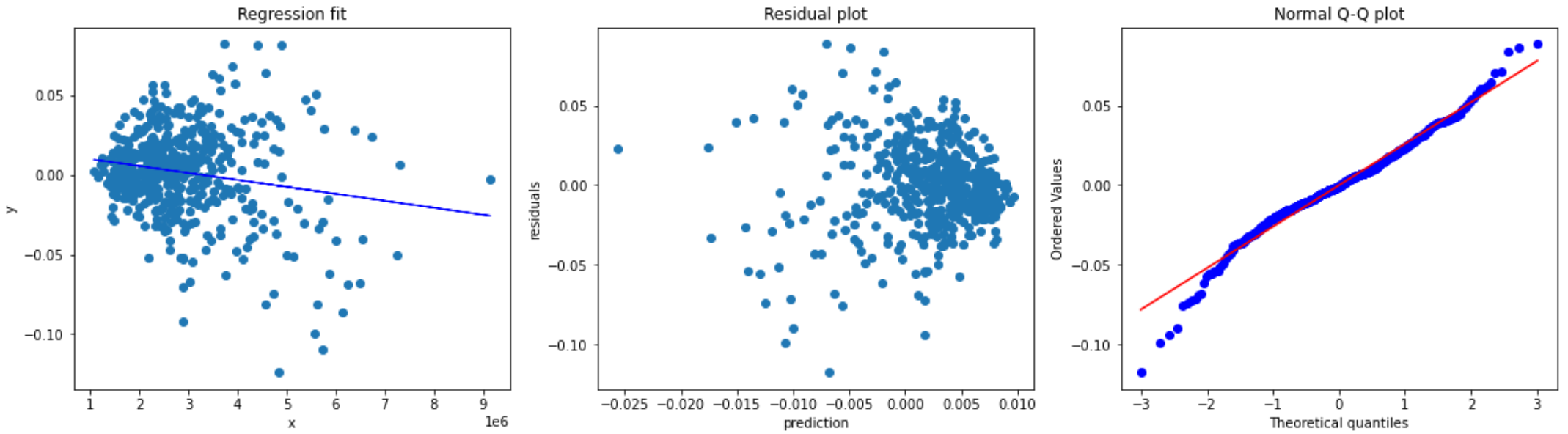
- Previous week's average volume traded may be weakly, negatively correlated with future return
- Macro factors including inflation, fed funds rate, unemployment are poor predictors of future return
- Previous returns are poor predictors of future return
- Weekly returns may limit the amount of data too much for the features and target examined.

Recommendations:

- Look for more company specific features, metrics derived from financial statements and the news such as expected and actual revenue, and earnings per share
- Find a way to obtain more data to train and test the models on: e.g. use daily returns or try to predict cross sectionally across many stocks instead of longitudinally over time
- Try modeling sectors or indices instead of individual stocks.

Supplemental:

- Further regression assumption check on avg_volume feature:



- RMSE of 5-fold cross validation of our final model:
- 5-fold cross validation RMSE's: [0.035 0.021 0.023 0.02 0.031]
- 5-fold cross validation RMSE mean: 0.026