Carnegie Mellon University
Statistics & Data Science

Predicting Annual Stock Price

Authors: Tianyou Zheng, Roochi Shah, Tiffany Shiu, Tony Hwang
Advisor: Dr. Joel Greenhouse; Client: Dr. Lars-Alexander Kuehn; Associate: Orhun Gun

Introduction & Background

Our overall objective of the Stock Market Signals project is to **predict the annual stock return for a given company**. The ongoing issue within the project is that the 10K financial statements provided by the Securities and Exchange Commission (SEC) have plenty of missing values. Before our team can tackle the problem of deficient financial statements, we had to first address the problem with the inconsistency of the financial statements themselves. There is no standardized methodology to report a 10K statement per entity in the United States. We overcome this challenge by utilizing Standard Industrial Classification (SIC) codes.

SIC Codes	Industries
1–999	Agriculture, forestry, and fishing
1000–1499	Mining
1500–1799	Construction
2000–3999	Manufacturing
4000–4999	Transportation, communications, electric, gas, and sanitary services
5000–5199	Wholesale trade
5200–5999	Retail trade
6000–6799	Finance, insurance, and real estate
7000–8999	Services
9000–9999	Public administration

Table 1: This table lists the industries corresponding to different SIC industry codes.

Data Pre-Processing

- 900+ features from U.S. firms 1960 – present
- Response variable Y_t: Stock price of a company at time t
- Book Value of Equity (be) was selected as a predictor variable for the model based on its high Pearson correlation coefficient of 0.76 and relatively small amount of missing values
- Notably high proportion of variables with significant fraction of missing values and very low or zero missing values (Figure 1)
- Most variables are missing at the same frequency across most sectors, with the Finance sector as the only exception (Figure 2)

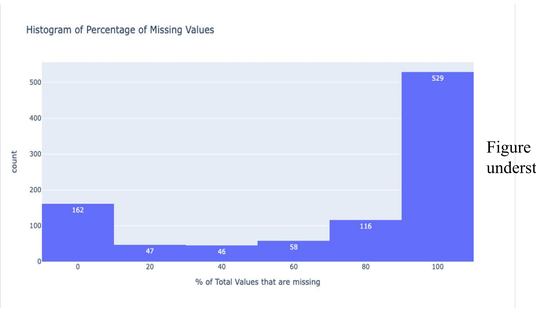

Figure 1: Univariate analysis on Agriculture sector to understand the proportion of total values that are missing

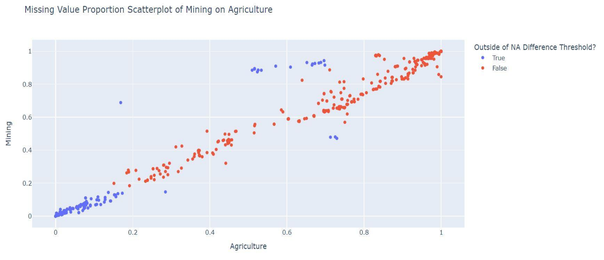

Figure 2: Bivariate analysis on missing value proportion scatterplot of Mining vs Agriculture

Methods

After analyzing the amount of missingness and predictor variables in our dataset, we proceed to perform the following models:

- Model 1: Y(t) ~ X(t-1)
- Model 2: Y(t) ~ Y(t-1) + X(t-1)
- Model 3: log(Y(t)) ~ log(Y(t-1))
- Model 4: log(Y(t)) ~ log(Y(t-1)) + X(t-1)

Models 2, 3, and 4 are autoregressive models, as they include the term Y(t-1), also known as the autoregressive term.

Analysis and Results

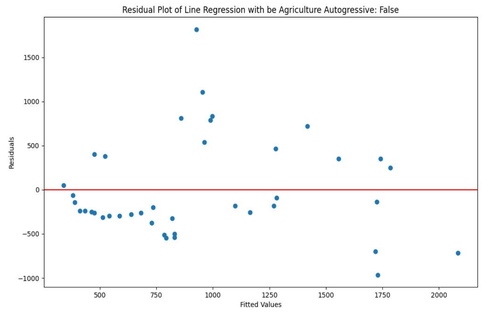

Figure 3: Residual vs fitted plot for Dole Food
Model 1; Y(t) ~ be(t-1)

- From Figure 3, the residuals form a nonlinear shape, so the true relationship is unlikely to be linear

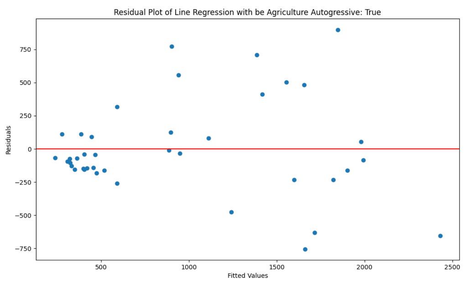

Figure 4: Residual vs fitted plot for Dole Food
Model 2; Y(t) ~ Y(t-1) + be(t-1)

- In Figure 4, for fitted values 0-500, the residuals are clustered close to the zero line but then get farther away from the line with higher fitted values, violating the constant variance assumption

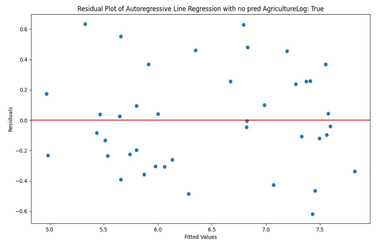

Figure 5: Residual vs fitted plot for Dole Food
Model 3; log(Y(t)) ~ log(Y(t-1))

- Residual plots look very similar between Models 3 and 4

- We see an improvement in the residual plot in Figure 5

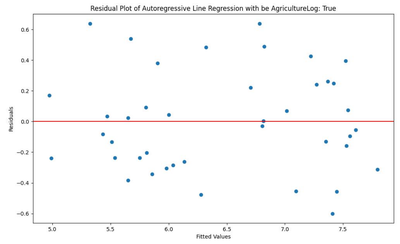

Figure 6: Residual vs fitted plot for Dole Food
Model 4; log(Y(t)) ~ log(Y(t-1) +be(t-1)

- Durbin-Watson statistic for Model 1 shows that there is positive autocorrelation between the residuals

Models	Durbin Watson	R-squared	Statistically Significant Predictors
Model 1	0.842	0.414	be
Model 2	1.982	0.758	y(t-1)
Model 3	2.012	0.871	y(t-1)
Model 4	1.975	0.872	y(t-1)

Table 2: Summary Output for all models

Conclusion

- Models 2-4 fix the autocorrelation issue in Model 1
- The log transformations in Models 3 and 4 show an improvement in the residual plots and model fits
- No noticeable difference between Models 3 and 4 implies that adding be doesn't account for any more of the variability in outcome
- Some next steps could be further analysis on these models with other sectors and imputation due to the amount of missing variables

References

Wharton Research Data Services
Perktold, Josef, et al.
“Statsmodels.stats.stattools.durbin_watson.” Statsmodels, https://www.statsmodels.org/dev/generated/statsmodels.stats.stattools.durbin_watson.html.