Model Building for Prediction of Valuation of Growth Stocks

By Lining Yang

AUTHOR BIO

Lining Yang is an 11th-grade student at Culver Academies with an interest in social science and data science. He plans to study in these fields in college. He has gained valuable experience for machine learning skills as an intern at Eyecool Technology Company. He has demonstrated academic excellence by qualifying for the AIME in AMC, earning the title of AP Scholar with Honor, and receiving the Global Seal of Biliteracy in French. Additionally, he is a member of the Culver Scholar Society. He is passionate about using his skills to make meaningful contributions in his future academic and professional endeavors.

ABSTRACT

It has been a great challenge for investors to evaluate growth stocks because these types of stocks are especially hard to forecast because of their inconsistencies and uncertainty. This paper establishes not only regression models, which include a regular Multivariable Linear Regression (MVLR) and a Fixed Effects (FE) Regression Model but also the application of a Panel Vector Autoregression (PVAR) Model on a dataset aiming to predict the Price-to-Sales Ratio (P/S) with several variables. Excel, Python, and Rstudio environments are used in the process of coding and model building. More specifically, Excel is used to conduct the MVLR model, Python is used for the FE model, and Rstudio is used for the PVAR model. By comparing different models of each type, the most optimized model was determined using a training set. After applying predictions with all types of models using another testing set, the best model is identified to be the best approach to the problem. Ultimately, all three models are compared. PVAR model and FE model are proved to be the more reliable ones out of the three models, because the R-squared for MVLR is 0.1859, lower than that of the FE model, 0.524, and the Mean Absolute Error (MAE) of for PVAR Model and FE model are respectfully 53.2139 and 55.1529. This project serves as a good starting point for future analysis and prediction of financial metrics like P/S ratio and is able to have great significance as it discovers some key points for the prediction of P/S value like its connection with time and entity despite the fact that the models themselves aren't the best fit mainly caused by the limitations of data and variable access.

Keywords: Data science, finance, computer science, mathematics, model establishment, prediction

INTRODUCTION

The stock market has been the most important investment area for American people for a long time (Ludvigson). In the stock market, growth stocks have become the most popular investment choice since the high risks can come with high profits. With great potential, these stocks not only provide investors with the chance of huge earnings but also have a higher chance of bankruptcy as they are mostly start-up companies with no positive cash flow. For investors, a common pursuit is to find reliable models to predict future stock values in hopes of profiting. Nevertheless, the growth stocks were especially hard to forecast. It has been a challenge for investors to decide which start-up companies or promising stocks they should invest in.

However, present valuation models are unable to measure the actual value of startup companies. Some traditional valuation techniques that are widely used in the financial markets are the Dividend Discount Model (DDM) and Discounted Cash Flow (DCF). DDM and DCF are both valuation methods used in finance to estimate the value of an investment based on its expected future cash flows. DDM is Used primarily for valuing stocks of companies that pay dividends. Nonetheless, most growth stocks, especially in their early and mid stages, do not pay dividends because they often reinvest all their profits into further growth. Thus, the DDM becomes less relevant or even inapplicable for such stocks. For DCF, the forecast of the cash flow of the growth stocks in the near term can be particularly challenging.

Therefore, this project aims to build a model that forecasts the price-to-sales(p/s) ratio of a growing stock. The model is built based on different factors that can affect the p/s ratio, including interest rate, gross margin, revenue growth rate, etc. To determine the factors that have a greater influence on the stocks with potential, data are needed at certain turning points of these values along with an analysis of how the company's price-to-sales ratio reacts. The price-to-sales ratio is the variable that will be predicted in the model. The price-to-sales ratio is particularly useful in the context of valuing growth stocks because it focuses on a company's sales rather than earnings, which can be more volatile or negative for growth companies. Growth stocks, especially startups, often have high potential for future increase but may not yet be profitable, which means that their earnings are either very low or negative. Traditional valuation metrics like the price-to-earnings (P/E) ratio become less applicable in these cases since they rely on positive earnings.

First, the project will screen out the independent variables that will be used in the multivariable regression model. The P/E ratio is used in this scenario because it is a decent measurement to recognize a growth stock at its start-up stage since a negative p/e rate indicates a negative earning. Therefore, the stocks chosen all have potential as most start-up companies have no positive earnings or cash flows.

The project establishes a multivariable linear regression with the factors that were found to model the relationship between the variables and the dependent variable, in this case, the p/s ratio while also filtering out the important factors. Moreover, a Fixed Effects model will be established as individual(company) effects are suspected in this case. The last piece of methodology that will be used is Panel Vector Autoregression, which allows variables to correlate with other variables with a certain lag of time in the past. All three models can be useful to predict future trends and determine the intercorrelation between all the independent and dependent variables.

BACKGROUND RESEARCH

The data used in the establishment of the model are found on Yahoo Finance, where growth stocks can be classified by the p/e ratios. The model uses only stocks that have a negative p/e ratio.

Price-to-Sales(p/s) Ratio and Price-to-Earnings(p/e) Ratio are both valuation ratios used in the finance world to assess the value of a company's stock. The Price-to-Earnings Ratio measures the price you pay for every dollar of earnings. It's one of the most widely recognized valuation measures. Price-to-Sales Ratio measures the price you pay for every dollar of sales. It is useful, especially for companies that aren't profitable yet. Therefore, the model predicts the future p/s value for reference of a company's value.

Another technique that will be used is the Panel Vector Autoregression model. A PVAR Model is a statistical method used in econometrics, an area of economics that uses statistical methods to test hypotheses and forecast future trends. This model has been tested to be accurate, simple, and effective over a wide range of sectors in the area of forecasting stocks (Apostolakis). Therefore, it is extremely helpful in the process of forecasting the price-to-sales value of the stocks.

Some factors in this project need defining. These factors were used in the model establishment:

- P/E (Price-to-Earnings Ratio): This is a measure used to value a company that compares its current share price to its per-share earnings. The P/E ratio is calculated by dividing the market value per share by the earnings per share (EPS) (Fernando). A high P/E ratio could mean that a company's stock is over-valued, or investors are expecting high growth rates in the future.
- P/S (Price-to-Sales Ratio): This ratio compares a company's stock price to its revenues. It is calculated by dividing the company's market capitalization by its total sales or revenue over a specific period. This ratio indicates how much investors are willing to pay per dollar of sales.
- Market Value: This refers to the total value of a company's shares in stock. It is calculated by multiplying the current market price of a company's stock by its total number of outstanding shares. Market value changes constantly as the stock price fluctuates.
- Sales Revenue Growth Rate: This measures the pace at which a company's sales revenue is increasing or decreasing, typically reported on an annual basis. It is an indicator of the company's financial health and efficiency in generating revenue (Schmidt).
- Debt to Equity Ratio (D/E): This ratio measures a company's financial leverage and is calculated by dividing its total liabilities by shareholder equity. It indicates the proportion of equity and debt the company is using to finance its assets (Fernando).
- 10-Year Treasury Bond Yield: This is the return on investment for the U.S. government's 10-year treasury bond. The yield reflects the bond's fixed interest rate as a percentage of its price. It's often seen as a benchmark for other interest rates and an indicator of the overall state of the economy.
- Gross Margin: Gross margin represents the proportion of each dollar of revenue that the company retains as gross profit. It is calculated by subtracting the cost of goods sold from revenue, and then dividing that figure by revenue. Gross margin is often expressed as a percentage and indicates how efficiently a company uses labor and supplies in the production process (Bloomenthal).

METHODOLOGY

Multivariable Linear Regression

The model building in this project involves multiple data analyses, thus a great amount of data is needed for creating the model. The data used in this project was collected using the Choice database, from NASDAQ.com. The factors chosen include:

- interest rate (10-year treasury yield)
- gross margin
- sales revenue growth rate
- debt to equity ratio

In the model, those four variables are used due to their significance in evaluating the financial health and growth potential of companies. The 10-year treasury yield is a critical factor in financial markets as it serves as a benchmark for other interest rates, including corporate bonds. Higher treasury yields lead to higher borrowing costs for companies, and make it more expensive to finance growth through debt, which is especially impactful for growth stocks that rely on borrowing for expansion. As treasury yields rise, the attractiveness of riskier investments like growth stocks decline. This could lead to lower valuations of stocks. Gross margin is a key measure of a company's profitability, representing the portion of revenue retained after accounting for the cost of goods sold. A higher gross margin indicates that a company is efficient in its production process and has strong pricing power, which is particularly important for growth stocks that may not yet be profitable. Sales revenue growth is essential in evaluating growth stocks, which are often valued based on future potential rather than current profitability. A higher sales growth rate signals that a company is expanding rapidly, which can indicate future profitability as it scales. The debt-to-equity ratio measures a company's financial leverage by comparing its total liabilities to shareholder equity. Growth companies often take on more debt to finance their expansion, but a high debt-to-equity ratio can also signal higher financial risk. This ratio helps investors understand the balance between risk and growth potential of a certain company, as debt can bring higher potential as well as more instability.

The time frame of each of these data is a quarter year, which means that the data was collected 4 times a year. This is because there will be fewer data points for a year, thus it is easier for data collection and model establishments. The time frame of the data is from 2001/1/1 to 2023/10/18.

For the range of stocks, all stocks listed on NASDAQ were picked first. This is because the NASDAQ stock exchange is known for having a significant number of growth stocks, particularly in the technology sector. As the first stock market online in the US, NASDAQ has attracted a large number of new tech companies (Cabot Wealth Network, 2023). Among all those companies, the ones that have had a negative p/e ratio in December for three years in a row were selected, since that is an obvious indicator of a stock that has potential and will be identified as a growth stock. Since some of the prices of the stocks are highly seasonal, the p/e value in December is more representative. After this, the sample size still needs to be reduced and the next step is that the stocks that follow the criteria were ranked by their market capitalization. The top 100 stocks with the greatest market capitalization were chosen because they are the most representative among all in order to decrease the sample size for model fitting. Then, the metrics needed for these stocks can be extracted from the database for the model establishment. The dataset consists of 3708 data points, including training data and testing data.

Data points from 2001 to 2020 are the training set for the models, and data points from 2021 to 2023 make up the testing set. Using the factors above, a multivariable linear regression model is established to predict future outcomes of the tickers' p/s value.

Multivariable linear regression, also known as multiple linear regression, is a statistical technique used to model the relationship between two or more independent variables and a single dependent variable. The general form of a multivariable linear regression model can be expressed as (Yale):

$$y = \beta 0 + \beta 1 x1 + \beta 2 x2 + \cdots + \beta n x n + \epsilon$$

Where:

- y is the dependent variable (the variable we are trying to predict or explain).
- $x1$, $x2$, …, xn are the independent variables (the variables used to predict y).
- β_0 is the y-intercept.
- β_1, β_2, …, β_n are the coefficients of the independent variables, which represent the change in the dependent variable for a one-unit change in the respective independent variable, assuming all other independent variables are held constant.
- ϵ is the error term, which accounts for the variability in y that cannot be explained by the independent variables.

The initial regression model with all the factors is shown in **Table 1** below with the variables and their according p-values. The p-value is used in hypothesis testing to decide whether to reject the null hypothesis. If the p-value is less than or equal to the chosen significance level of 0.05, then the null hypothesis is rejected. It's a fundamental concept in the field of statistics and is used to determine the statistical significance of the results (NIH). In this situation, the null hypothesis is that there is no relationship between the independent and dependent variables. A p-value of 0.05 means that there's a 5% probability of observing such data if the null hypothesis were actually true. If the P-value is lower than 0.05, it's considered statistically significant, so we can reject the null hypothesis and conclude that there is a relationship between the variables. The p-values for sales revenue growth rate and debt-to-equity ratio were much greater than 0.05. Thus, these two factors would not be included in the next regression model. With two p-values greater than 0.05, it is not considered a good regression model for the prediction of the p/s ratio.

Table 1

All Variables and Their P-values in the Initiation of Multivariable Linear Regression

variables	Coefficients	P-value
Intercept	237.9673	0.0000
sales revenue growth rate	-0.0006	0.3685
debt to equity ratio	-0.0001	0.7641
10-year treasury bond bond yield	-14.4226	0.0017
gross margin	-292.6747	0.0000

Another trial of the multivariable linear regression model has been made with the two factors that contain a p-value less than 0.05 in the previous model. In **Table 2**, it shows that this model is a good fit and both variables have p-values less than 0.05. Other attempts with different combinations of the variables have been made to establish the model, but it can be concluded that the model below is the best fitted as it has the lowest p-value.

Table 2

Variables and Their P-values of the Best Fit Linear Regression Model

variables	Coefficients	P-value
Intercept	237.9378	0.0000
10-year treasury bond bond yield	-14.6003	0.0015
gross margin	-292.201	0.0000

By looking at the residual plot shown in Graph 1 of all the factors, it is clear that the model is not a very good fit for most of the data points, with several outliers that have significantly greater residuals than other data points.

The multivariable analysis was run on the data, and the model was built. The equation is:

$$\hat{PS} = 237.9378 GrossMargin - 14.6003 TenYearTreasuryBondYield - 292.201$$

Fixed Effects Model

The second methodology that will be applied to this dataset is the Fixed Effects Model. A Fixed Effects model is a statistical approach used primarily in panel data analysis, where the data involves observations over multiple time periods for the same entities (such as individuals, firms, countries, etc.). This model is particularly useful for controlling unobserved variables that are constant over time but vary across entities. It is especially useful in this context because Fixed Effects models control for time-invariant unobserved individual-specific effects (Allison). This helps reduce bias in the estimates that can be introduced by omitted variables. In this case, there is probably unobserved interdependence among the factors, as they are all measurements calculated from the company’s statistics. Moreover, not a lot of factors were involved.

Thus, reducing the bias of not including all variables is essential. Another main purpose of applying the FE model is to reduce bias. Since there are 100 companies involved, it is crucial to reduce bias in the estimation of the coefficients by focusing on variation within entities(stocks). In order to use the fixed effect model, it is assumed that there are correlations between the variables used. A Fixed Effects model is basically structured similarly to a linear regression, but with dummy variables to capture individual-specific or time-specific effects. In this approach, add dummy variables to a linear regression model to account for individual or time-specific effects. The model is structured as follows:

$$Yit = \beta 0 + \beta 1Xit + \gamma i + \epsilon it$$

Where:

- Y_{it} is the dependent variable for individual *i* at time *t*.
- X_{it} represents the independent variables.
- γ_i are the individual-specific dummy variables (fixed effects).
- ϵ_{it} is the error term.

As the vector γ_i represents the entity effects for each individual stock, it consists of constants that are different for each stock. The model is fitted with my dataset using Python, the linear model package.

First, for this panel dataset, stock-specific dummy variables representing each stock need to be added to the original dataset. Therefore, 100 columns are added for all the stocks in my dataset, all of them are true or false Boolean values. For each observation data point, it only has a value of True for its respective stock dummy variable and False for all others. **Table 3** and **Table 4** show the updated dataset of the training and testing set. For example, in the first row, the dummy variable AMZN_O, which is the stock for Amazon, has a True value, while all other dummy variables have False because the datapoint is for AMZN_O.

Table 3

Training Data panel with dummy variables

	time	code	PE	PS	...	Z_O	ZG_O	ZI_O	ZS_O
0	2001Q1	AMZN.O	-3.9166	1.2698	...	FALSE	FALSE	FALSE	FALSE
1	2001Q2	AMZN.O	-6.3027	1.7035	...	FALSE	FALSE	FALSE	FALSE
2	2002Q1	AMZN.O	-6.6313	1.7923	...	FALSE	FALSE	FALSE	FALSE
3	2002Q2	AMZN.O	-25.9889	1.7801	...	FALSE	FALSE	FALSE	FALSE
4	2003Q1	AMZN.O	-67.7652	2.5695	...	FALSE	FALSE	FALSE	FALSE
...	...	...	...	...	...	...	...	...	...
2789	2020Q4	EXEL.O	56.2489	6.5299	...	FALSE	FALSE	FALSE	FALSE
2790	2020Q4	NFE.O	-38.4078	24.8506	...	FALSE	FALSE	FALSE	FALSE
2791	2020Q4	IONS.O	-17.5859	10.8825	...	FALSE	FALSE	FALSE	FALSE
2792	2020Q4	APPF.O	36.5562	18.676	...	FALSE	FALSE	FALSE	FALSE
2793	2020Q4	RMBS.O	-44.7212	8.034	...	FALSE	FALSE	FALSE	FALSE

Table 4

Testing Data panel with dummy variables

	time	code	PE	PS	...	Z_O	ZG_O	ZI_O	ZS_O
0	2021Q1	AMZN.O	73.1056	4.0392	...	FALSE	FALSE	FALSE	FALSE
1	2021Q1	TSLA.O	892.1173	20.3962	...	FALSE	FALSE	FALSE	FALSE
2	2021Q1	AMD.O	38.3042	9.7692	...	FALSE	FALSE	FALSE	FALSE
3	2021Q1	PDD.O	-152.4866	18.4027	...	FALSE	FALSE	FALSE	FALSE
4	2021Q1	VRTX.O	20.5114	8.9627	...	FALSE	FALSE	FALSE	FALSE
..	...	...	...	...	...	...	...	...	...
909	2023Q2	BZ.O	70.7134	9.6371	...	FALSE	FALSE	FALSE	FALSE
910	2023Q2	NFE.O	10.1885	2.27	...	FALSE	FALSE	FALSE	FALSE
911	2023Q2	IONS.O	-14.0119	9.3153	...	FALSE	FALSE	FALSE	FALSE
912	2023Q2	APPF.O	-56.7962	11.5256	...	FALSE	FALSE	FALSE	FALSE
913	2023Q2	RMBS.O	20.3384	14.957	...	FALSE	FALSE	FALSE	FALSE

Using this dataset, a regression model is created using all the dummy variables, gross margin, and 10-year treasury bond yield as independent variables and PS as dependent variables. The summary of the result is in **Figure 1**.

Figure 1

Results for the Fixed Effect Model

```
============================== OLSR With Dummies ==============================
                            OLS Regression Results
==============================================================================
Dep. Variable:                     PS   R-squared:                       0.524
```

The model can be mathematically written as:

$$\hat{PS} =- 18.5479 GrossMargin - 1.0710 TenYearTreasuryBondYield + 47.5592 + \gamma i$$

For the entity effects γi, the constant for each stock is calculated, as the result outputs a coefficient and a p-value for each stock. The P-value is used again to verify if it is statistically significant to say that the specific entity has a correlation with the PS value. When p-value < 0.05, it is significant to say that the coefficient is valid, as we can reject the null hypothesis of no correlation. Using this criterion, all variables with a p-value < 0.05 are picked, as shown in **Table 5**.

Table 5

All variables with a p-value that is considered significant

Stocks	Coefficient	p-value
APA_O	-49.1811	0.036
EXAS_O	77.1569	0
LBRDA_O	102.0222	0
LBRDK_O	656.9457	0
LCID_O	610.5532	0
MRNA_O	67.5454	0.02
NBIX_O	59.0724	0.019
SRPT_O	109.8561	0

For all these stocks, they will have a different constant in their equation. For example, EXAS stock will have a constant of 47.5592 + 77.1569 = 124.7161. Therefore, its prediction PS would be -18.5479 GrossMargin - 1.0710 TenYearTreasuryBondYield + 124.7161. The rest of the stocks that don't have a p-value less than 0.05 all use the same constant 47.5592 in their formula.

Panel Vector Autoregression

Panel Vector Autoregression (PVAR) is a statistical method used in econometrics, an area of economics that uses statistical methods to test hypotheses and forecast future trends. The PVAR model is essentially an extension of the Vector Autoregression (VAR) model, which is used to capture the linear interdependencies among multiple time series (Canova). The reason why the model is used is that PVAR allows for the modeling of entity-specific characteristics, which is crucial when variations across different areas of companies are important. It's also beneficial when the focus is on understanding how these variables interact with each other over time within and across these entities. As the dataset used is large and contains data across a long period of time, the PVAR model is appropriate because it provides more reliable results with large datasets. Several assumptions need to be made to make sure the model is a good fit. The math of panel vector autoregression is further explained in the appendix.

Using the variables tested in the previous regression models, 5 PVAR models are established with 5 lag values to compare and optimize the most appropriate lag value.

Moreover, the BIC, AIC, and HQIC values are all calculated to determine the best model by finding the lowest of those values. The criterions of all 5 models are shown in Figure 4 below.

Results

The multivariable linear regression is shown below:

PS=237.9378GrossMargin-14.6003TenYearTreasuryBondYield-292.201

Using the model, it is possible to predict values. In **Table 6** below, the predicted p/s value and the actual values are demonstrated.

Table 6

Predicted PS value using MVLR model

time	code	PS	prediction
2001Q1	AMZN.O	1.2698	89.7756
2001Q2	AMZN.O	1.7035	80.1457
2002Q1	AMZN.O	1.7923	81.868
2002Q2	AMZN.O	1.7801	87.8489
2003Q1	AMZN.O	2.5695	109.0514
2003Q2	AMZN.O	3.1592	113.4814
...			
2023Q2	BZ.O	9.6371	-56.793
2023Q2	NFE.O	2.27	7.6303
2023Q2	IONS.O	9.3153	-105.956
2023Q2	APPF.O	11.5256	5.0509
2023Q2	RMBS.O	14.957	-52.3485

With the Fixed Effects OLS model fit, the equation is shown below:

$$\hat{PS} = -18.5479GrossMargin - 1.0710TenYearTreasuryBondYield + 47.5592 + \gamma i$$

The predicted p/s value and the actual values are shown in **table 7** below.

***Table** 7*
Predicted ps value using Fixed Effects OLS model

time	code	PS	predicted
2001Q1	AMZN.O	1.2698	37.8764
2001Q2	AMZN.O	1.7035	41.7726
2002Q1	AMZN.O	1.7923	37.2193
2002Q2	AMZN.O	1.7801	36.2154
2003Q1	AMZN.O	2.5695	29.3479
...			
2023Q2	BZ.O	9.6371	28.4222
2023Q2	NFE.O	2.27	32.4789
2023Q2	IONS.O	9.3153	25.3264
2023Q2	APPF.O	11.5256	32.3165
2023Q2	RMBS.O	14.957	28.7020

With the Panel Vector Autoregression model fit, the equation is shown below:

PS = 16.5089*lag1GrossMargin+0.8404*lag1PS-1.3471*TenYearTreasuryBondYield-4.4665

The predicted p/s value and the actual values are shown in **Table 8** below.

***Table** 8*
First ten lines of predicted PS value using the PVAR model

	code	time	actual_PS
1	AAL.O	2021Q2	0.735
2	AAL.O	2021Q3	0.543
3	AAL.O	2021Q4	0.389
4	AAL.O	2022Q1	0.397
5	AAL.O	2022Q2	0.202
6	AAL.O	2022Q3	0.173
7	AAL.O	2022Q4	0.169
8	AAL.O	2023Q1	0.197
9	AAL.O	2023Q2	0.222
10	ABNB.O	2021Q2	21.4

MODEL COMPARISON

To find out the optimized model to perform this task, two values are mainly used as a criterion for the effectiveness of the models, r-squared (R^2) and Mean Absolute Error (MAE), particularly because regression models can be measured by its r-squared, while PVAR models are more complex, thus MAE values are needed. The coefficient of determination, R-squared, is a statistical measure in regression analysis that represents the proportion of the variance in the dependent variable that is predictable from the independent variables. It is a key indicator of the goodness of fit of a regression model (CFI). The Mean Absolute Error (MAE) is a measure used to assess the accuracy of a predictive model. It calculates the average magnitude of errors in a set of predictions, without considering their direction. The MAE is a straightforward and intuitive measure that provides a simple average of the absolute differences between predicted and actual values (ScienceDirect). By using these two criteria, models can be compared.

All three models conducted using data points from 2001 to 2020 are validated on the data points from 2021 to 2023. First of all, the r-squared of the Multivariable linear regression and Fixed Effects Model are compared, as they are both similar regression models. R-squared for MVLR is 0.1859, while for the FE model, 0.524. The quite low R-squared suggests that MVLR is probably not the best approach to this problem, while the FE model has a decent R-squared. When applying data to the testing dataset, the MAE for the MVLR model is 88.7662, and for the FE model is 55.1529, both indicating that the model is significantly not a good fit for the data. Still, FE is slightly better. The MAE in this case is calculated by taking the mean of the absolute residual of each datapoint. The absolute residual is the absolute value of the difference between the predicted PS value and the actual PS value. For example, taking the first line of prediction in the FE model on the testing set. The actual PS is 1.2698, and the predicted PS value is 37.8764. The absolute residual is calculated by |37.8764 – 1.2698| = 36.6066. Therefore, the mean absolute error can be determined by taking the mean of the absolute residual of all the data points as the absolute residuals are calculated as I demonstrated. If one model has to be picked from those two, the Fixed Effects Model is picked, because it has both a greater r-squared and a lower MAE, indicating a more fitted model. The MAE value for the PVAR model is 53.2139, which is a bit lower than the FE model. The interpretation of MAE 53.2139 in context is that the predicted Price using the model is 53.2139 times Sales more or less than the actual price. For example, the actual PS is 5, and the prediction is 55. The absolute residual is 50. The predicted price = 55 * sales, and the actual price = 5 *sales. The difference would be 50 * Sales. This is the interpretation of the absolute residual, and MAE is simply the mean of it. Overall, none of the models are good for prediction, while the FE model and PVAR have an advantage over the MVLR model.

CONCLUSION

A conclusion that can be made with variables in this type of dataset is that it is heavily related to time and entities, which means that autoregression approaches and Fixed Effects aspects can be considered in the further improvement of the model, considering that the model is not a very good fit to the dataset when variable numbers are limited. Not only because the variables are limited, but also because of the inconsistent properties of growth stocks' P/S values make it especially hard to find a good model for prediction.

To return to my initial purpose of the project, the model can be used to predict the future P/S ratio of top 100 market cap growth stocks, maybe also applicable to all growth stocks when we already have the predicted gross margin for the company and interest rate for a certain period in the future, even though none of the models are good. There are probably more appropriate models for this scenario.

Another insight from the project is that several companies' stocks with the lowest residuals of the prediction of PS ratio are picked out when the best model PVAR is fitted. These stocks are significantly better predicted by the model than others, shown in **Table 9**.

Table 9

5 stocks with best predictions

	code	time	actual_PS	predicted_PS	absolute_residuals
	<fct>	<fct>	<dbl>	<dbl>	<dbl>
1	WYNN.O	2022Q1	2.46	2.46	0.00482
2	CZR.O	2023Q2	0.961	0.978	0.0177
3	MRVL.O	2023Q2	8.87	8.90	0.0239
4	CZR.O	2022Q3	0.653	0.695	0.0414
5	MELI.O	2023Q1	6.28	6.20	0.0773

As we can tell from the data, these five stocks have minimal residuals, which indicates that the predicted PS values are very close to the actual PS value. This demonstrates that this model is especially good at predicting PS values for those five companies, which suggests that we might be able to apply the model by focusing using it only on these companies despite the fact that the model is not good at predicting PS values for other companies.

FUTURE WORK

First, there are many more variables and factors that have effects on the change of the p/s value of a stock. In this paper, only 4 variables that were believed to be the most important were used to build the model considering that the dataset can be too big to extract when more variables are involved. More work can be done when there is data with more variables and metrics. The model can be greatly improved if there are more variables available considering that the models in this project aren't great.

Secondly, some other regressions can be applied to this dataset to optimize the prediction by making more trials with different models, including Ridge Regression, Lasso Regression, etc.

REFERENCES

Apostolakis, G., & Papadopoulos, A. P. (2018, July 18). Financial stability, monetary stability and growth: A PVAR analysis - Open Economies Review. *SpringerLink*. Springer US. https://link.springer.com/article/10.1007/s11079-018-9507-y

Bloomenthal, A. (n.d.). Gross margin: Definition, example, formula, and how to calculate. *Investopedia*. Retrieved November 17, 2023, from https://www.investopedia.com/terms/g/grossmargin.asp

Fernando, J. (n.d.). Debt-to-equity (D/E) ratio formula and how to interpret it. *Investopedia*. Retrieved November 17, 2023, from https://www.investopedia.com/terms/d/debtequityratio.asp

Fernando, J. (n.d.). P/E ratio definition: Price-to-earnings ratio formula and examples. *Investopedia*. Retrieved November 17, 2023, from https://www.investopedia.com/terms/p/price-earningsratio.asp

Ludvigson, S. (n.d.). How important is the stock market effect on consumption. *Federal Reserve Bank of New York*. Retrieved November 17, 2023, from https://www.newyorkfed.org/medialibrary/media/research/epr/99v05n2/9907ludv.pdf
Schmidt, J. (2023, October 4). Sales revenue. *Corporate Finance Institute*. https://corporatefinanceinstitute.com/resources/accounting/sales-revenue/

Nau, R. (n.d.). Stationarity and differencing of time series data. *Duke University*. Retrieved November 20, 2023, from https://people.duke.edu/~rnau/411diff.htm

MathWorks. (n.d.). Introduction to vector autoregressive (VAR) models. Retrieved November 21, 2023, from https://www.mathworks.com/help/econ/introduction-to-vector-autoregressive-var-models.html

RDocumentation. (n.d.). Pvargmm: GMM estimation of panel VAR models. Retrieved November 21, 2023, from https://rdrr.io/rforge/panelvar/man/pvargmm.html

Bevans, R. (2023, June 22). Akaike information criterion: When & how to use it (Example). *Scribbr*. https://www.scribbr.com/statistics/akaike-information-criterion/

Wikipedia. (2023, June 12). Hannan–Quinn information criterion. *Wikimedia Foundation*. Retrieved from https://en.wikipedia.org/wiki/Hannan%E2%80%93Quinn_information_criterion

Dahiru, T. (2008, June). P-value, a true test of statistical significance? A cautionary note. *Annals of Ibadan Postgraduate Medicine*. U.S. National Library of Medicine. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4111019/#:~:text=The%20P%20value%20is%20defined,groups%20is%20due%20to%20chance

Canova, F. (n.d.). Panel vector autoregressive models: A survey. *European Central Bank*. Retrieved November 15, 2023, from https://www.ecb.europa.eu/pub/pdf/scpwps/ecbwp1507.pdf

Allison, P. (n.d.). Fixed effects regression models. *Google Books*. Retrieved December 5, 2023, from https://books.google.com/books?hl=en&lr=&id=3UxaBQAAQBAJ&oi=fnd&pg=PT12&dq=fixed+effects+model&ots=PLVribRNPR&sig=DvEQJdUI0yA7JoyujQLfNlOAN38#v=onepage&q=fixed%20effects%20model&f=false

Yale University. (n.d.). Multiple linear regression. Retrieved December 5, 2023, from https://www.stat.yale.edu/Courses/1997-98/101/linmult.htm

Indeed. (n.d.). Exogenous vs. endogenous variables (definition and examples). Retrieved December 5, 2023, from https://www.indeed.com/career-advice/career-development/exogenous-variable

ScienceDirect. (n.d.). Mean absolute error - An overview. Retrieved December 5, 2023, from https://www.sciencedirect.com/topics/engineering/mean-absolute-error

Corporate Finance Institute. (2023, November 22). R-squared. https://corporatefinanceinstitute.com/resources/data-science/r-squared/

W3Schools. (n.d.). Data science - Regression table: P-value. Retrieved December 12, 2023, from https://www.w3schools.com/datascience/ds_linear_regression_pvalue.asp#:~:text=The%20P%2Dvalue&text=A%20low%20P%2Dvalue%20(%3C,called%20an%20insignificant%20P%2Dvalue