***Data Cleaning and Preparation***

Graphical user interface, text, application

Description automatically generated

By converting the GV, the first digit of GVKEY into a numerical variable and compare it with nine, last digit of my student number, earnings dataset is set to only include data for companies with GVKEY starting with a number less than or equal to the last digit of my student number.

Graphical user interface, text

Description automatically generated with medium confidence

The reason that the number of observations does not change is that there is no single digit number greater than nine and thus this condition makes no difference to the dataset.

Text

Description automatically generated

Since this task requires us to use models to forecast earnings per share (EPS) and return on net operating assets (RNOA), we must construct variables of EPS and variables required to calculate RNOA and forecast RNOA according to Soliman (2008).

Graphical user interface, text, application

Description automatically generated

Table

Description automatically generated

Since observations with missing required variables are not useful in forecasting RNOA and EPS, they should be deleted to prevent those observations interfering results of our forecast. By using the do-loop and if statement, we delete observations that miss any one of those required variables. Earnings1’s number of observations significantly decreased to 64433 after deleting observations which are missing any one of those required variables.

Text

Description automatically generated

Since one of our project’s main tasks is to forecast RNOA and calculate its errors, we must construct two variables, next period’s RNOA (RNOAF) and this period’s RNOA (RNOA). We construct these two variables by self-joining dataset Earnings1 to ensure that each observation has next year’s operating income after depreciation (OIADP) and this year’s opening net operating assets (NOA) to calculate RNOAF and RNOA.

Table

Description automatically generated

Since the application of this SQL self-joining code only keeps observations that have the same companies’ previous year’s and next year’s data, the number of observations in the dataset is further reduced to 42201. However, the ridiculous maximum and minimum value of each variable and huge difference between mean and median indicates that the dataset includes some extreme values which will significantly impact the results of our models in following parts. Therefore, the dataset must be winsorized to minimize the extreme values’ impact on our data analysis’s results.

Graphical user interface, text, application

Description automatically generated



A SAS macro is created to winsorize extreme values of variables at top and bottom 1% and then apply it to the dataset. The SAS macro first sort the data in the way we need and then use PROC MEANS to output the 1st percentile and 99th percentile to a dataset called STAT. The new dataset will merge the old dataset with STAT by the sorting method we applied above. It then compares the listed variables with its 1st percentile and 99th percentile, variable observations that is not a missing number and is smaller than the 1st percentile will be set to equal 1st percentile whilst variable observations that are greater than 99th percentile will be set to equal 99th percentile. We then apply the SAS macro to winsorize the constructed variables each year at top and bottom 1%.

Table

Description automatically generated

After winsorization, the dataset’s variables’ maximum and minimum values become more reasonable and the gap between mean and median narrowed, this means that winsorization effectively limits the impact of extreme values on our data analysis. After all the above data preparation processes, the data is now set to be analyzed and be applied in different models.

***Descriptive Statistics Discussions***

This project’s purpose is forecasting EPS and RNOAF using various models, thus RNOAF and EPS are two most important variables. Therefore, we focus on discussion of the descriptive statistics of EPS and RNOAF.

A picture containing text

Description automatically generated

PROC UNIVARIATE function is used to produce descriptive statistics of RNOAF and EPS in the NEWREGRESSION dataset.

*RNOAF’s Descriptive Statistics*

Table

Description automatically generatedTable

Description automatically generated

Distribution of RNOAF’s sample mean is positively skewed which is shown from the fact that mean is greater than median which is greater than mode and its skewness is 1.5453, but we can fairly assume that this data is approximately normally distributed because central limit theorem states that when dataset is large enough (42201 observations in this instance), distribution of sample mean is approximately normally distributed.

Descriptive statistics show that the student t-distribution’s p-value is less than 0.01 and therefore the hypothesis that the RNOAF’s mean is 0 should be rejected at a significance level of 1%, therefore, we are 99% confident that the data’s RNOAF’s mean is not 0 and determine its mean to be 0.20425. Thus, the descriptive statistics shows that on average, companies can generate $0.20425 operating income after depreciation for each dollar of net operating assets.

*EPS’s Descriptive Statistics*

***Table

Description automatically generated***Table

Description automatically generated

Descriptive statistics show that the EPS data’s mean is greater than median which is greater than mode and has skewness of 19.0365, these indicate that the EPS’s sample mean distribution is positively skewed. However, considering that according to central limit theorem, when the dataset is sufficiently large (in this instance, 42201 observations), the sample mean distribution is approximately normally distributed. Therefore, it is fair to assume that EPS’s sample mean distribution is normally distributed.

The student’s t distribution’s p-value is less than the significant level of 1%, we are 99% confident to reject the null hypothesis that the EPS’s sample mean is 0 and determine EPS’s sample mean to be 2.7810. Therefore, on average, each share generates $2.7810 net profit for its holder every year.

***Correlation Coefficient Between Constructed Variables***

Graphical user interface, text, application

Description automatically generated

The PROC CORR function demonstrates the Pearson and Spearman correlation between the Soliman (2008) variables and RNOAF which is an important topic for this project.

Table

Description automatically generated

For both Pearson and Spearman correlation table, the null hypothesis that selling and general expense proportion’s (SGX) correlation with RNOAF is 0 cannot be rejected at the significance level of 1% as its p-value is greater than 1%. For RNOAF’s correlation coefficient with other Soliman (2008) variables, since the p-values are below than 1%, we are 99% confident to conclude that the correlation coefficient is not zero, but is the correlation coefficient values on the tables.

According to Soliman (2008), RNOAF can be forecasted using the eight column variables. Thus, it is not surprising that both Pearson and Spearman correlation indicates that net operating asset turnover (ATO), total accruals (TACC) are shown to have strong correlation with RNOAF whilst inventory proportion (INV) and receivable proportions (AR) are shown to have moderate correlation with RNOAF. However, both Spearman and Pearson correlation coefficient indicates that profit margin (PM) and effective interest rate (ETR) only have a weak correlation with the RNOAF. For this year’s RNOA (RNOA) and RNOAF’s correlation, Pearson correlation coefficient indicates a weak correlation while Spearman correlation indicates a strong correlation. Since Pearson correlation measures the strength of linear relationship between two variables whilst spearman correlation measures the tendency of two variables moving in the same direction, it is likely that RNOA and RNOAF have a strong tendency to move in the same direction but not in a linear form.

***Forecasting EPS with Simple Models and Their Errors***

*Construction of Forecast Models*

Text

Description automatically generated

Graphical user interface, text

Description automatically generated

To construct the forecast model, we first use lag function to create variables needed to set up the condition used to create the random walk model and moving average forecast. Then, by applying several conditions, we created forecast for one, two and three years random walk model and also created forecast for two, three and five years moving average model. We also calculated each models’ forecast errors by subtracting forecast EPS from actual EPS and dividing that difference by actual EPS. This eliminates companies’ scales’ impact on forecast errors.

*Cleaning Forecast Errors’ Descriptive Statistics*

*Text

Description automatically generated with low confidence*

Table

Description automatically generated with medium confidence

Since different models use different conditions, the number of observations being eliminated is different for models of different years. The ridiculously large maximum error and the unusually large gap between mean and median indicates that the descriptive statistics of errors is distorted by extreme values. Therefore, extreme values must be winsorized to reduce their impact on our investigation on different forecast models’ errors.

Table

Description automatically generated

Table

Description automatically generated

After winsorizing extreme values at top and bottom 1%, the maximum values look more reasonable and the gap between mean and median has narrowed. Therefore, the impact of extreme values on the descriptive statistics of the data significantly decreased.

*Selection of Model with Minimal Error*

There are six different simple models being applied in this step, they can be classified into 2 big categories, random walk model and moving average model. Therefore, we first determine which categories’ models produce smaller error. This can be done by comparing performance of the two categories’ models given they have the same information.

Text

Description automatically generated

Table

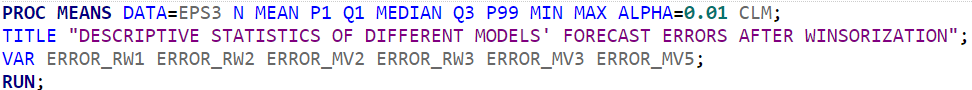
Description automatically generated

Table

Description automatically generated

Since the p-value for both t-tests are less than the significance level of 1%, we are 99% confident that the hypothesis that 2 years and 3 years random walk model’s forecast errors is the same as 2 years and 3 years moving average model’s respective forecast errors is rejected. Therefore, on average, 2 years random walk model’s forecast error is 113% higher than 2 years moving average model’s forecast error whilst 3 years random walk model’s forecast error is 175% higher than the 3 years moving average model’s forecast error.

Therefore, generally, moving average model generates smaller forecast error given that same amount of information is given since moving average model accounts for more information to forecast future EPS than random walk model and is therefore more accurate.



Table

Description automatically generated

The descriptive statistics shows that the 5-years moving average model’s forecast error is the lowest for both lower and upper 99% confidence level for mean. Therefore, consistent with the above, moving average model generates smaller forecast error and the longer the time, the more accurate the moving average model.

However, it is also notable that the one-year random walk model has the lowest 1st percentile, 1st quartile and median forecast errors. Therefore, it is arguable that the one-year random walk model provides the smallest forecast error. Although we have winsorized the forecast errors, some forecast errors still seem to be quite extreme, for example 1-year random walk model’s maximum forecast error is 34 times of this year’s EPS which is quite ridiculous.

Therefore, considering that we have a large dataset of at least 17407 observations, median may be a better measurement of forecast error than mean. On the fact that the 1-year random walk model generates the smallest forecast error, it seems to be the most accurate forecast model within these 6 different models.

***Using OLS Regression to Estimate Soliman (2008) RNOA Forecast Model***

Text

Description automatically generated with low confidence

Table

Description automatically generated Table

Description automatically generated

The OLS regression shows that 60.63% of future RNOA variances can be explained by the eight independent variables in Soliman (2008) RNOA forecast model. However, after the adjusting for the number of observations and variables in the model, the OLS regression shows that only 60.62% of future RNOA variances can be explained by the eight independent variables in Soliman (2008) RNOA forecast model.

By applying a 1% significance level, the independent variable of ETR fails to reject the null hypothesis that its parameter estimate is 0 and therefore should be excluded from the model while all other Soliman (2008) RNOA forecast model variables successfully reject the null hypothesis at 1% significance level and thus should be included in the RNOAF forecast formula.

Therefore, according to OLS regression, the formula to forecast RNOAF should be .

***Using Stepwise OLS Regression to Select RNOA Forecast Model***

Text

Description automatically generated

Table

Description automatically generated

Model R-Square is the proportion of variance for RNOAF that can be explained by the variables entered, for example, in step 3, model R-square of 0.5992 means 59.92% of RNOAF can be explained by the variable ATO, TACC and AR.

Partial R-square is the incremental proportion of variance of RNOAF can be explained by variable entered because of adding in the new variable. For example, in step 3, the partial R-square of 0.0361 means because of adding AR to the model to forecast RNOAF, 3.61% more of variance for RNOAF can be explained by the model.

Theoretically, the more variables we use in a model, the more powerful the model will be in terms of explanatory power. However, the more variables are included in a model, the more complex the model is and therefore more difficult to be acted on. Therefore, we need to select a rule to select variables thereby to help us construct a simpler model without losing too much explanatory power. This is particularly important for this model which is used to estimate financial data which often requires analysis of a large dataset.

Let’s set that only variables with partial R-square larger than 0.6% (which is 1% of the full model’s explanatory power) is allowed to enter the model. Therefore, 5 variables will be eliminated and only ATO, TACC and AR will be selected as independent variables for the RNOAF forecast model.

***Back Testing of Soliman (2008) RNOA Forecast Model and 1-Year Random Walk Model***

Text

Description automatically generated

Text

Description automatically generated

This code uses ranuni function to separate NEWREGRESSION dataset into an estimation and test sample. Then, we apply OLS regression to estimate model using the estimation sample and output the Soliman (2008) RNOA forecast model’s variables’ regression coefficient to EST1. By adding EST1 to the test dataset, we then calculate the forecast RNOAF and Soliman (2008) model’s forecast error as well as the 1-year random walk model’s forecast error. The mean and median of both models’ forecast errors are then output to the dataset, average. The above processes are repeated 100 times to give us a larger dataset and thus allowing us to better assess the accuracy of each of the models.

*Back Testing Results and Discussions*

Graphical user interface, text, application, email

Description automatically generated

Table

Description automatically generated

Since the student-t p-value is less than the significance level of 1%, we are 99% confident to reject the null hypothesis that the mean forecast error of Soliman (2008) regression model is the same as 1-year random walk model’s mean forecast error. Therefore, we can conclude that on average, the mean forecast error of random walk model is 41.25% lower than the mean forecast error of Soliman (2008) regression model.

Table

Description automatically generated

Similarly, the student-t p-value is less than the significance level of 1%, we are 99% confident to reject the null hypothesis that the median forecast error of Soliman (2008) regression model is the same as the median forecast error of 1-year random walk model. Therefore, we can conclude that on average, the median forecast error of random walk model is 16.45% lower than the median forecast error of Soliman (2008) regression model.

A picture containing table

Description automatically generated

Also, the descriptive statistics also shows that for all minimum, lower quartile, median, upper quartile and maximum, 1-year random walk model’s mean and median forecast error is shown to be less than the Soliman (2008) regression model’s respective forecast errors.

Therefore, it is uncontentious that 1-year random walk model’s forecast error is on average smaller than Soliman (2008) regression model’s forecast error. Thus, it is fair to conclude that the 1-year random walk model predicts companies’ future operating income with higher accuracy. Although Soliman (2008) regression model accounts for more variables and information, it is also being affected by the more company specific information which negatively impacts the model’s accuracy. This may be the reason why although the 1-year random walk model seems to be simpler, it still predicts companies’ future operating income with higher accuracy than the complex Soliman (2008) regression model.

(Word Count: 2489)