BY PAULINA SUEN

Picture credit: Forbes

Estimating yield for a hypothetical bond

Predictive and Visual Analytics for Business

Table of Contents

Introduction

Bonds are loans that investors make to borrowers for a certain period in exchange for interest payments. The certain period is called maturity, bond issuer, whose issue the bond and borrow money from the investor, must pay off the loan at the end of the period.

There are two types of bonds, government bonds and corporate bonds. In this report, bonds refer to corporate bonds, specifically the corporate bonds in US market. The return that investors expected to receive annually during the period till maturity is called bond's yield. This report aims to estimate the bond yield with a range of variables. With the estimation, investors can determine whether a bond is worth to invest.

Regression analysis will be performed with three regression models to estimate the yield of bond. There are four sections in this report including data wrangling, data summary, methodology, and result analysis. It is hypothesized that 9 predictor variables in the dataset are statistically significant to yield and will be utilized to predict the value of yield.

Data Wrangling

Reliable and complete data provide better analysis performance, data wrangling is performed to avoid the negative influence of raw data.

First, two sets of data were used to perform estimation in this report, Dataset A and Dataset B. There are 4,247 observations in Dataset A with 3 variables, Bond_id, Coupon, and Yield. Dataset B has 4,250 observations with 12 variables, Bond_id, Issuer, Maturity, Sector, Currency, Convertible, Moodys_cred_rat, Market_of_Issue, Amount_outstanding, Callable, Putable, and Seniority.

To avoid invalid statistics, duplicates must be removed from the dataset. Two observations were removed from Dataset A and four observations were removed from Dataset B. Then, two datasets were merged by Bond_id, a variable that represent each observation in both datasets. Before merging, the datasets were sorted by Bond_id to make sure all data of each observation are following to its specific Bond_id. The new dataset Dataset C with merged data have 4,254 observations and 14 variables.

Based on the requirements, putable bond and convertible bonds were removed from the dataset, which reduced the observations to 4,245. Observations continue reduced to 4,216 after observations with missing value were removed.

An observation's value that is extremely higher or lower than others is called outlier, which may affect the result of statistical analysis, lead to underestimation or overestimation. Hence, outlier(s) must be found and removed from the dataset before performing analysis. To identify outlier(s), the PROC UNIVARIATE procedure in SAS was run. The result of this procedure provides a statistics summary including the information of extreme observations.

Extreme Observations					
Lowest			Highest		
Value	Bond_id	Obs	Value	Bond_id	Obs
-150.000	1133	1104	13.9867	4150	4112
-0.124	3412	3375	15.8570	3237	3200
-0.100	4192	4154	15.8573	3236	3199
-0.094	4195	4157	1000.0000	1338	1309
-0.091	1448	1418	158775.0000	1583	1553

Table 1: Extreme observations

From Table 1, it is observed that the value of Bond_id 1583 and 1338 are extraordinarily higher than others, while value of Bond_id 1133 are extremely lower than others. This level of unique value is known as extreme outlier, which will seriously decrease the accuracy of the regression model. Besides of extreme outliers, the box plot in Figure 1 indicates that there are many other outliers in the dataset. This box plot of yield was plotted with the data after removing the 3 extreme outliers.

Outlier(s) can be identified by quartile. Any value that exceeds the lower inner fence or upper inner fence are discriminated as outlier. Based on the value of first quartile (Q1) and third quartile (Q3) of the dataset, lower inner fence and upper inner fence are calculated based on the equation Q1 - 1.5* (Q3 - Q1) and Q3 + 1.5 * (Q3 - Q1). After

removing all observations that were recognized as outliers, the dataset remains 3,986 observations.

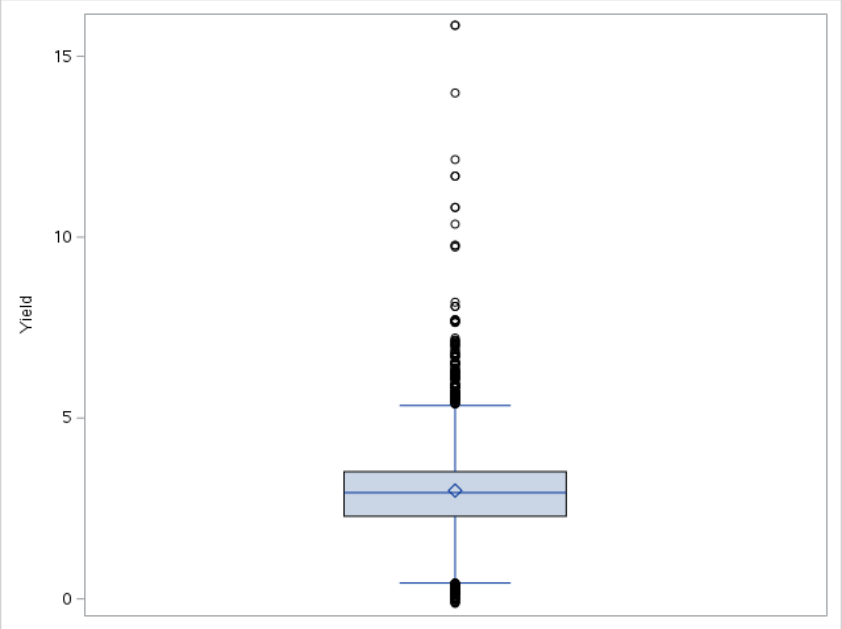


Figure 1: Box Plot of Yield

Based on requirements, several observations are removed and just keep the observations that are denominated in US dollars, that left 3,784 observations in the dataset.

To ensure well performance of regression residuals, several variables are created and added into the dataset based on requirements. They include converting maturity from date into year, converting amount outstanding into billions of USD and natural logarithm. Variable Coupon was scaled by natural logarithm and an assign to a new variable coupon_ln. Categorical data are difficult to use in regression analysis due to its non-numerical nature. Hence, categorical data in the dataset are converted into dummy variables, an artificial variable that stand for distinct categories. Number 0 and 1 were assigned to represent specific level of the category, such as if a bond is callable, it will be represented as 1, otherwise 0. Three categorical variables were represented by 9 different dummy variables based on the requirement, and other 37 dummy variables were created to represent two variables market_of_issue and sector for the purpose of performing regression analysis. Data wrangling is completed at this point, and the finalise dataset that used to perform regression analysis contained 3,784 observations and 64 variables in total.

Data Summary

Summary Statistics

Summary of the dataset will be described in this section. The histogram in Figure 2 shows that the data is normally distributed, it is in bell-shaped with only one peak. It indicates that there are almost the same between mean, median and mode.

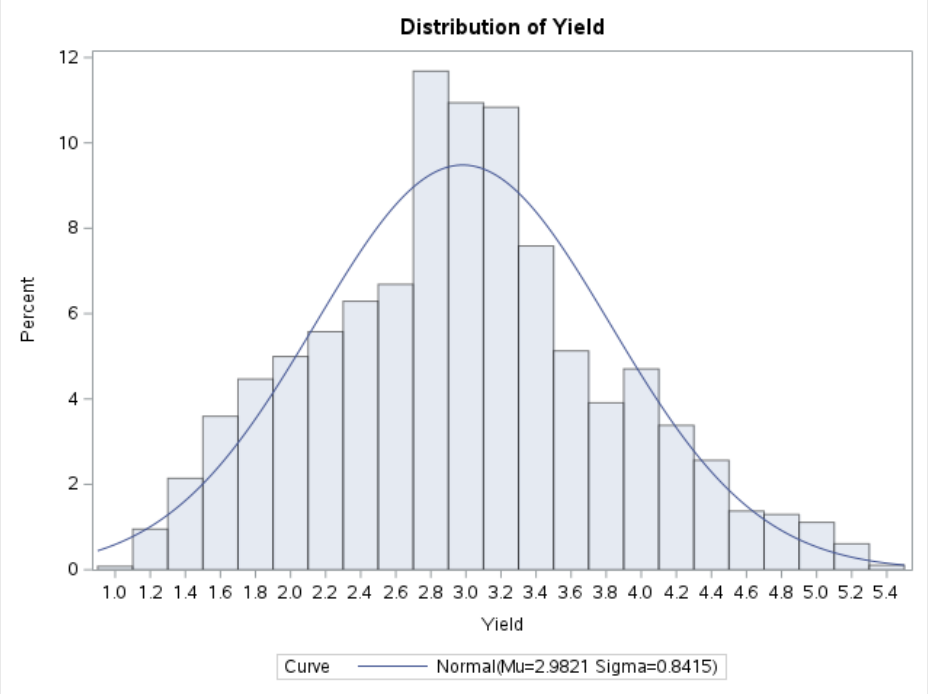


Figure 2: Histogram of yield

This result of normal distribution is supported by the statistics result. Table 2 presents the summary statistics of yield in the dataset. The average yield for 3,784 observation is 2.98, which is approximately equal to the median 2.96 and mode 3.29. While with the standard deviation of 0.841, which tells the distribution of data that about 95% of bond yield is within 1.298% to 4.662% (2.98% +/- 0.841% *2), which is not far from the average value, meaning the data are relatively consistent and reliable. Moreover, the skewness, a measurement of symmetrical distribution, is 0.22 that is between -0.5 and 0.5, suggesting the distribution is almost symmetric. Under normal distribution, there is higher accuracy to observe the population with the sample statistics.

The UNIVARIATE Procedure
Variable: Yield

Moments			
N	3784	Sum Weights	3784
Mean	2.98206973	Sum Observations	11284.114
Std Deviation	0.84148821	Variance	0.7081024
Skewness	0.21850097	Kurtosis	-0.2348773
Uncorrected SS	36328.8633	Corrected SS	2678.75139
Coeff Variation	28.2183552	Std Error Mean	0.01367957

Basic Statistical Measures			
Location		Variability	
Mean	2.982060	Std Deviation	0.84149
Median	2.957000	Variance	0.70810
Mode	3.288000	Range	4.36330
		Interquartile Range	1.07165

Note: The mode displayed is the smallest of 2 modes with a count of 5.

Tests for Location: Mu0=0				
Test	Statistic		p Value	
Student's t	t	217.9837	Pr > |t|	<.0001
Sign	M	1892	Pr >= |M|	<.0001
Signed Rank	S	3580610	Pr >= |S|	<.0001

Quantiles (Definition 5)	
Level	Quantile
100% Max	5.34550
99%	5.04900
95%	4.46660
90%	4.12270
75% Q3	3.47275
50% Median	2.95700
25% Q1	2.40110
10%	1.86070
5%	1.61420
1%	1.29150
0% Min	0.98220

Extreme Observations			
Lowest		Highest	
Value	Obs	Value	Obs
0.9822	1438	5.2784	3463
1.0055	3673	5.3360	2464
1.0499	2995	5.3363	2463
1.1124	3867	5.3410	2007
1.1258	2999	5.3455	2006

Table 2: Summary statistics

In addition, in normal distribution, outlier can be identified by standard deviation and mean, any value smaller or greater than {mean +/- (3 * standard deviation)} is defined as outlier. A procedure has been run in SAS with the standard deviation and mean shown in Table 2 and concluded that this latest dataset has no more outlier.

Data visualization

Bonds in the dataset are issued from three markets, Eurobond, domestic and global. Figure 3 shows the proportion of each market. It indicates that nearly 60% of bonds are from global market, around 28% is from domestic market and less than 20% is Eurobond.

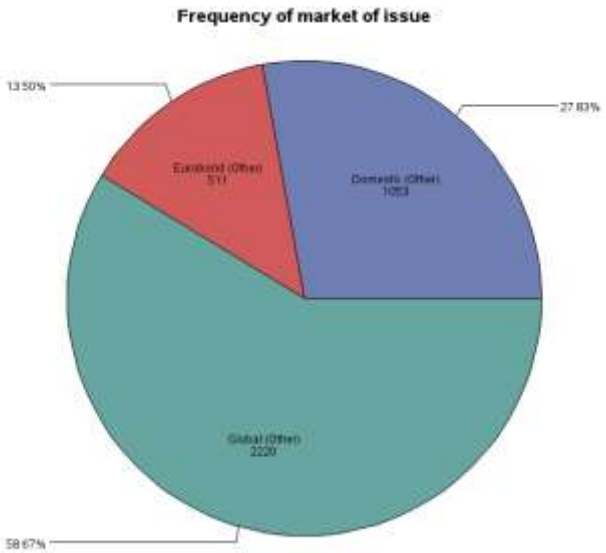


Figure 3: Frequency of market of issue

Issuer may redeem callable bond before the maturity date. It is a bond that allow business to gain benefit based on market interest rate, while its coupon rate is usually higher and benefit the investor. Figure 4 displays the percentage of callable bond in the dataset. Almost 80% of bonds are callable compared to only around 20% are not callable.

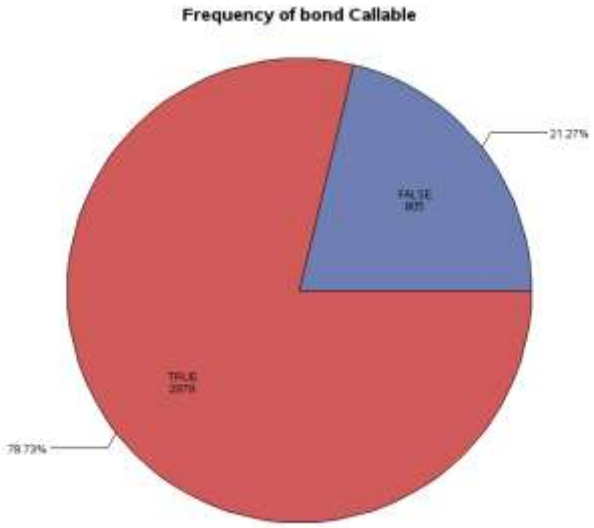


Figure 4: Frequency of bond Callable

Seniority in bonds is related to the priority of a bond investor receive payment when the bond issuer is under bankruptcy. Investors holding secured bonds would be paid off their debt first in the event of the issuer's liquidation, and senior secured bond has the highest priority in payout. In opposite, bond that is senior unsecured are not guaranteed to be paid. It is observed in Figure 5 that some 87% of bonds in this dataset are senior unsecured and no more than 5% of bond are secured.

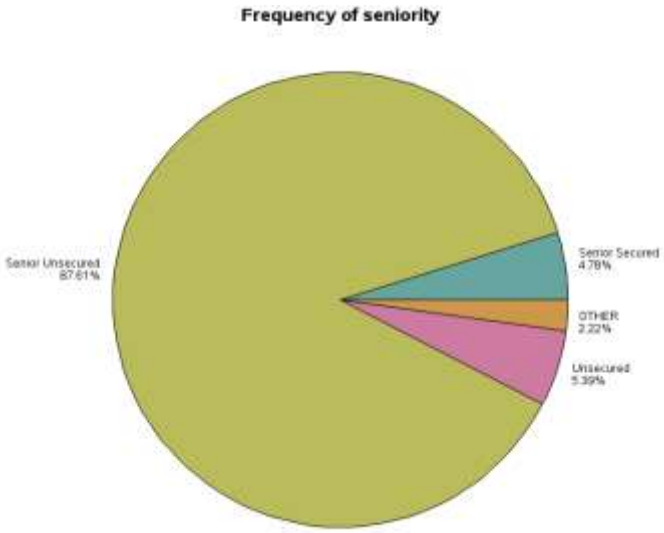
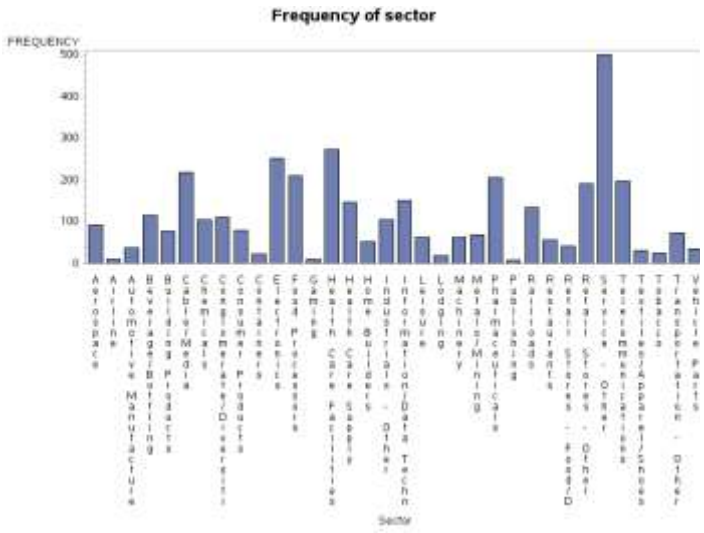


Figure 5: Frequency of seniority

Companies that issued bonds in this dataset are from 34 different sectors, it can be observed in Figure 6 that numerous bonds are issued by companies from Service sector, which is 500 within 3,784 bonds. More than 250 bonds are from Health care facilities and Electronics sector. Airline, Gaming and Publishing sector have less than 1% proportion in this dataset.

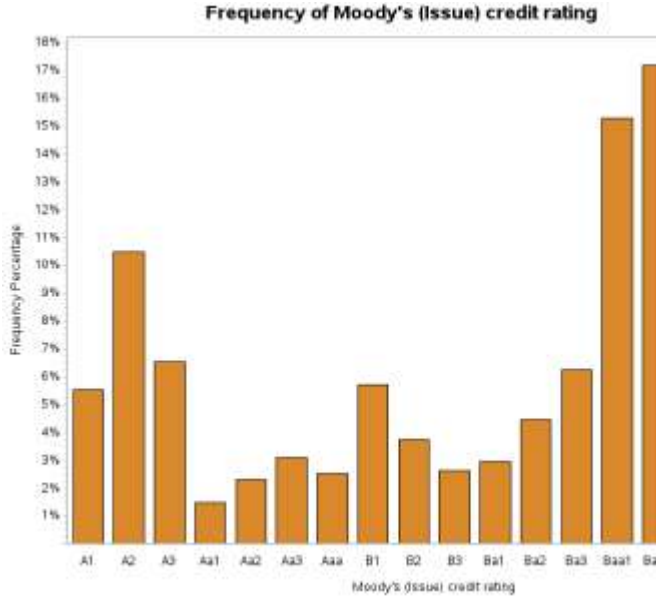


The FREQ Procedure

Sector	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Aerospace	92	2.43	92	2.43
Airline	10	0.26	102	2.70
Automotive Manufacture	38	1.00	140	3.70
Beverage/Bottling	115	3.04	255	6.74
Building Products	78	2.06	333	8.80
Cable/Media	219	5.79	552	14.59
Chemicals	105	2.77	657	17.36
Conglomerate/Diversifi	110	2.91	767	20.27
Consumer Products	80	2.11	847	22.38
Containers	23	0.61	870	22.99
Electronics	253	6.69	1123	29.68
Food Processors	210	5.55	1333	35.23
Gaming	11	0.29	1344	35.52
Health Care Facilities	273	7.21	1617	42.73
Health Care Supply	147	3.88	1764	46.62
Home Builders	53	1.40	1817	48.02
Industrials - Other	105	2.77	1922	50.79
Information/Data Techn	152	4.02	2074	54.81
Leisure	62	1.64	2136	56.45
Lodging	19	0.48	2154	56.93
Machinery	62	1.64	2216	58.56
Metals/Mining	66	1.80	2284	60.36
Pharmaceuticals	206	5.44	2490	65.82
Publishing	8	0.21	2498	66.01
Railroads	135	3.57	2633	69.58
Restaurants	57	1.51	2690	71.09
Retail Stores - Food/D	42	1.11	2732	72.20
Retail Stores - Other	190	5.02	2922	77.22
Service - Other	500	13.21	3422	90.43
Telecommunications	188	5.23	3610	95.67
Textiles/Apparel/Shoes	31	0.82	3651	96.49
Tobacco	24	0.63	3675	97.12
Transportation - Other	73	1.93	3748	99.05
Vehicle Parts	36	0.95	3784	100.00

Figure 6: Frequency of sector

The credit risk of a bonds is ranked by Moody's credit rating, with the scale of running from the highest Aaa to the lowest C. Figure 7 shows the frequency of each level of rating of the bonds. 17% of bonds rank at Baa2, 2% higher than those with Baa1 rating. Few bonds are ranked with the low rating, less than 1% are Caa1 or Caa2 rating.



Frequency of sector

The FREQ Procedure

Moodys_cred_rat	Frequency	Percent	Cumulative Frequency
A1	209	5.52	209
A2	397	10.49	606
A3	248	6.55	854
Aa1	56	1.48	910
Aa2	87	2.30	997
Aa3	117	3.09	1114
Aaa	95	2.51	1209
B1	216	5.71	1425
B2	142	3.75	1567
B3	99	2.62	1666
Ba1	111	2.93	1777
Ba2	169	4.47	1946
Ba3	237	6.26	2183
Baa1	578	15.27	2761
Baa2	650	17.18	3411
Baa3	335	8.85	3746
Caa1	28	0.74	3774
Caa2	10	0.26	3784

Figure 7: Frequency of Moody's (Issue) credit rating

Coupon in bond is the annual interest rate paid for a bond, it is usually a rate in percentage unit. Bond yield generally calculated by coupon and bond price; thus, they are highly related. Scatter plot in Figure 8 presents the positive relationship between coupon and yield since the data points form a direction from bottom left-hand corner to top right-hand corner. The ellipse (the circle in blue colour in Figure 8) also present the

correlation between coupon and yield. The positive relationship between these two variables indicates that whenever one variable increase or decrease, the other variable will have the same movement. In addition, it can be observed that from Figure 8, most coupon rate for bonds in the dataset are around 2 to 6%.

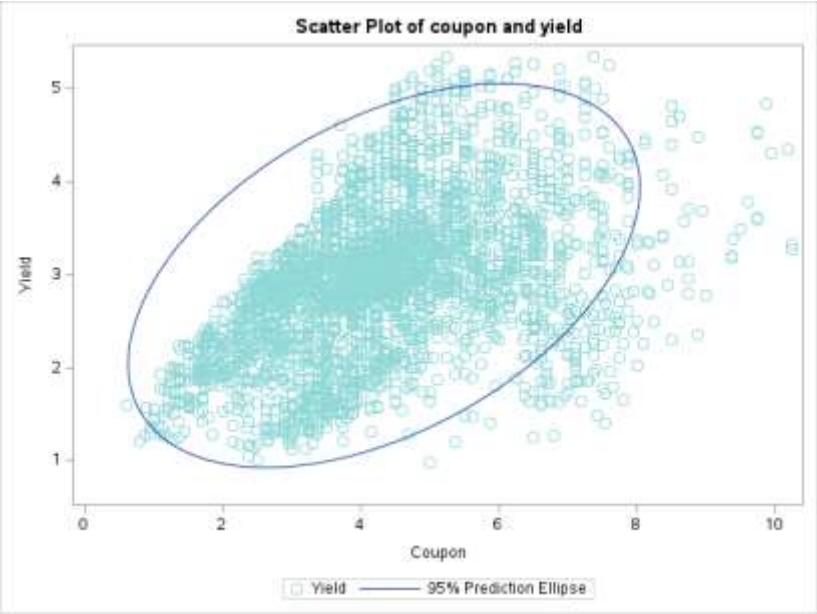


Figure 8: Scatter Plot of coupon and yield

Amount outstanding is the debt that the bond issuers are unpaid. Figure 9 visualizes data point of yield and amount outstanding. To make it looks clear, amount outstanding is calculated as the natural logarithm of amount. The figure shows that the amount basically around 18 to 23, while mostly within 19 to 21. Most data points located at the same position and did not able to form a straight line from one corner to the other corner. The ellipse is wide and round suggests the variables are unlikely related. Yet, the relationship between yield and amount outstanding, especially whether amount outstanding statistically related to yield will be observed in the following section by regression model.

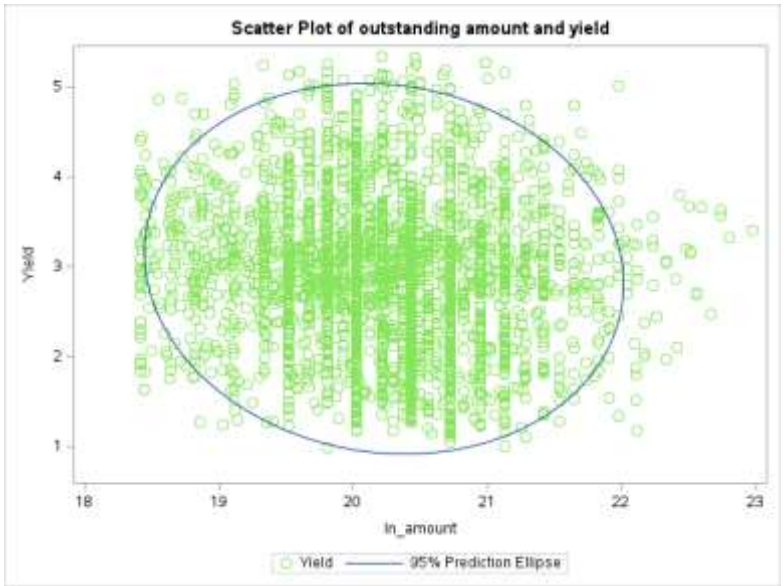


Figure 9: Scatter Plot of amount outstanding and yield

Maturity is the date that a bond is due, and issuer should pay back the principal amount

that investor pay to buy the bonds at the beginning. It is a specific date in the dataset, and it is converted to year of maturity. From Figure 10, it can be observed that most bonds are due within 20 years, specifically they are less than 10 years. Figure 10 also indicates that there is low relationship between year of maturity and yield since the data point basically form a horizontal line. A horizontal line in scatter plot means there is no relationship between the two variables and one variable is just randomly scattered on the grid.

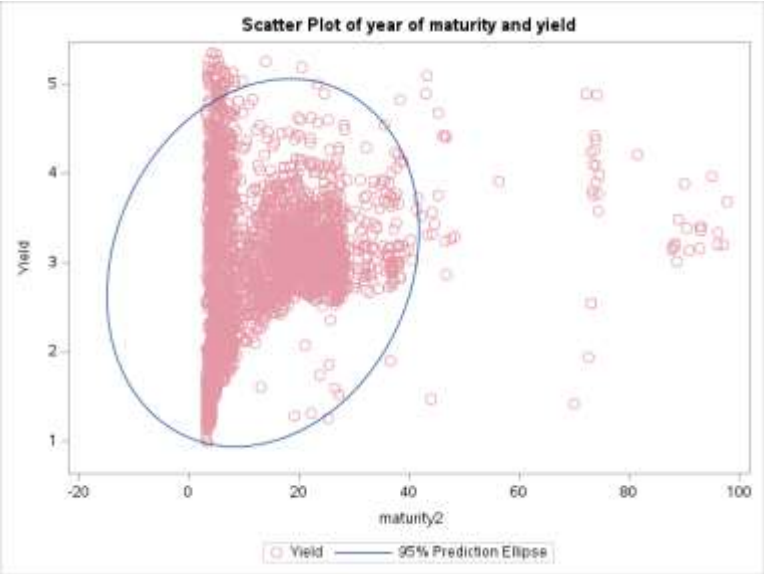


Figure 10: Scatter plot of year of maturity and yield

Methodology

In this report, a regression analysis will be performed to estimate the yield for a Aa2 Moody's (Issue) credit rating callable senior unsecured bond that from global market in Electronics sector, and with 10 years of maturity, 2.5% of coupon rate and $750,000,000 amount outstanding. Three regression models will be built.

All three regression models use 3,784 observations to estimate the response variable **yield**. For predictor variables, there are three discrete variables, including **maturity2** (represents year of maturity), **coupon**, and **amount_outstanding**. There are dummy variables which are converted from categorical variables **seniority**, **Moodys_cred_rat**, **sector**, **callable**, and **market_of_issue**. There are 49 predictor variables in total for each regression model. The main difference between three regression models is the variables represent coupon and amount outstanding.

In the first regression model, amount2, a variable that represents amount outstanding in billion of USD is used as predictor variable. The second regression model uses the natural logarithm of amount outstanding to represent amount outstanding. The third regression model based on the second regression model but changing the coupon variable as the natural logarithm of coupon. Also, since coupon rate is in the unit of percentage, it is divided by 100 while calculating its natural logarithm.

On the other hand, null hypothesis and alternative hypothesis are defined before performing the regression analysis.

H0: There is no relationship between response variable and predictor variables.

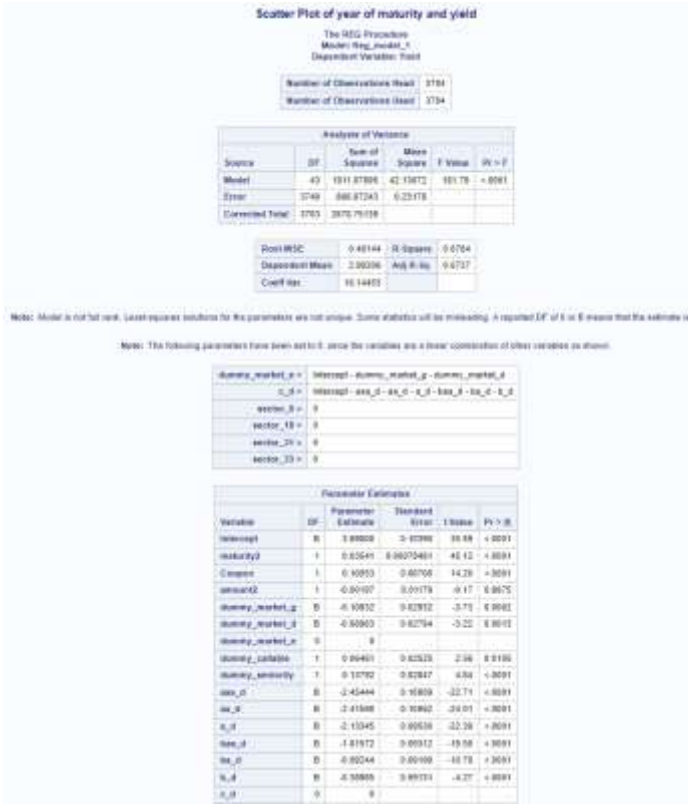
H1: There is relationship between response variable and predictor variables.

Based on the p-value of each regression model, to determine whether null hypothesis can be rejected.

Result and Analysis

This section interprets the result of each regression model and estimate yield for bond with characteristic that mentioned in the Methodology section.

Regression model 1



sector_1	1	0.20414	0.05705	3.58	0.0004
sector_2	1	0.10620	0.10487	1.20	0.2293
sector_3	1	0.30460	0.08320	3.66	0.0003
sector_4	1	0.05704	0.05230	1.10	0.2694
sector_5	1	0.02116	0.06072	0.35	0.7276
sector_6	1	0.09618	0.04217	2.33	0.0199
sector_7	1	0.01683	0.05375	0.31	0.7543
sector_8	0	0			
sector_9	1	0.07388	0.06188	1.19	0.2340
sector_10	1	-0.24788	0.10388	-2.39	0.0171
sector_11	1	0.04187	0.04011	1.04	0.2968
sector_12	1	0.07264	0.04208	1.73	0.0844
sector_13	1	-0.16323	0.18405	-0.89	0.3198
sector_14	1	0.08936	0.04094	2.18	0.0291
sector_15	1	0.01682	0.04753	0.35	0.7234
sector_16	1	0.23705	0.07154	3.31	0.0009
sector_17	1	-0.06658	0.05388	-1.24	0.2165
sector_18	0	0			
sector_19	1	0.01695	0.06712	0.25	0.8008
sector_20	1	0.31153	0.11693	2.67	0.0076
sector_21	1	-0.09053	0.08658	-0.13	0.8901
sector_22	1	0.28304	0.06456	4.38	<.0001
sector_23	1	0.08232	0.04249	1.94	0.0528
sector_24	1	0.31803	0.17258	1.84	0.0654
sector_25	1	-0.13850	0.04998	-2.79	0.0053
sector_26	1	0.03334	0.06918	0.48	0.6295
sector_27	1	0.12723	0.07906	1.61	0.1076
sector_28	1	0.11334	0.04362	2.60	0.0094
sector_29	1	0.10671	0.03442	3.10	0.0020
sector_30	1	0.10913	0.04456	2.45	0.0144
sector_31	0	0			
sector_32	1	0.89992	0.10230	8.84	<.0001
sector_33	0	0			
sector_34	1	0.08484	0.08436	0.77	0.4422

Table 3: Result of regression model 1

Table 3 shows the result of regression model 1. The p-value of the model is less than 0.0001 suggesting the sample data in this model provide sufficient evidence to reject null hypothesis for the entire population, there is relationship between variables in this model. Moreover, the adjusted R-squared is 0.627 implying the predictor variables can explain 62.7% of the variability after the number of predictors in the model is adjusted. Also, coefficient of variation for this model is 16.14, this value describes how the model relative to the predicted value. It compares with other models to determine the best fit of regression model. Although comparing to other two models, model 1's coefficient variance value is slightly smaller, other values suggests that it still a functional model that can use to estimate the response variable, only the accuracy may be less than other models.

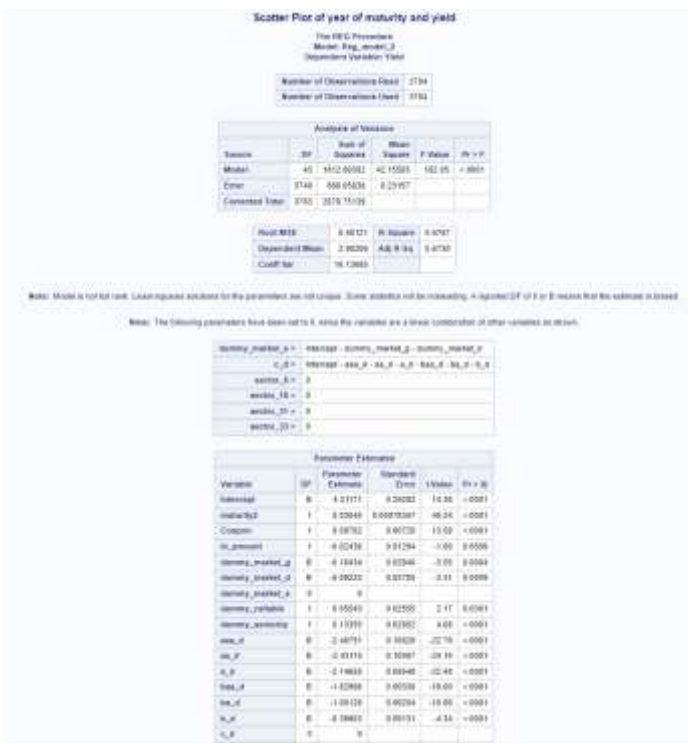
Variables that have a p-value less than significant level 5% mean they are statistically significant to the dependent variable. Table 3 indicates that year of maturity, coupon, market of issue, seniority, and Moody's (Issue) credit rating are all highly related to yield, they are all resulting a p-value lower than significant level 0.1%. Yet, amount outstanding has a p-value higher than 5%, meaning it is statistically insignificant to yield. Generally, independent variable(s) that is not significant should be removed from the model. Amount outstanding of bonds is one of the required characteristics to

estimate the yield for bond, it will be retained. Similarly, for sector Electronics, one of the required characteristics, has a higher than 5% p-value but will be retained in the model and use to estimate yield even it is not statistically related to the dependent variable.

The estimate coefficient of each variable implies the relationship between response variables and predictor variables. When all other predictor variables is equal to 0, the yield for bond is 3.698%. When there is 1% increase in coupon, the value of yield will increase 0.1% on average. Yield will drop 0.002% on average when the amount outstanding increase 1 billion of USD.

Based on the result listed in Table 3, the estimation for yield of the bond is 1.772%. This value is calculated by the regression equation: 3.698 + (0.035 * 10) + (0.100 * 2.5%) + (-0.002 * (750,000,000 / 1,000,000,000)) + (-0.109) + 0.065 + 0.138 + (-2.416) + 0.04187. As mentioned, there are two independent variables are not statistically significant to yield, but it is included in the equation and resulting an overestimation on the predicted value of yield.

Regression model 2



sector_1	1	0.20874	0.05700	3.63	0.0003
sector_2	1	0.18471	0.15478	1.19	0.2328
sector_3	1	0.31008	0.08320	3.73	0.0002
sector_4	1	0.06293	0.05229	1.20	0.2289
sector_5	1	0.01684	0.06072	0.28	0.7783
sector_6	1	0.11015	0.04225	2.61	0.0092
sector_7	1	0.01524	0.05375	0.28	0.7768
sector_8	0	0			
sector_9	1	0.07726	0.06168	1.25	0.2119
sector_10	1	-0.25053	0.10383	-2.41	0.0158
sector_11	1	0.04671	0.04012	1.16	0.2444
sector_12	1	0.07220	0.04206	1.72	0.0852
sector_13	1	-0.18559	0.18368	-1.01	0.3126
sector_14	1	0.09155	0.04090	2.24	0.0253
sector_15	1	0.01590	0.04750	0.33	0.7443
sector_16	1	0.22776	0.07164	3.18	0.0015
sector_17	1	-0.06950	0.05388	-1.29	0.1979
sector_18	0	0			
sector_19	1	0.01794	0.06766	0.27	0.7891
sector_20	1	0.30578	0.11659	2.62	0.0088
sector_21	1	-0.01239	0.06656	-0.19	0.8524
sector_22	1	0.27932	0.06455	4.33	<.0001
sector_23	1	0.08864	0.04248	2.09	0.0370
sector_24	1	0.31322	0.17249	1.82	0.0695
sector_25	1	-0.14248	0.04964	-2.87	0.0041
sector_26	1	0.03704	0.06909	0.54	0.5920
sector_27	1	0.12643	0.07802	1.60	0.1097
sector_28	1	0.11435	0.04359	2.62	0.0087
sector_29	1	0.10320	0.03445	3.00	0.0028
sector_30	1	0.11910	0.04415	2.70	0.0070
sector_31	0	0			
sector_32	1	0.70581	0.18230	8.99	<.0001
sector_33	0	0			
sector_34	1	0.05871	0.08436	0.70	0.4866

Table 4: Result of regression model 2

The result of regression model 2 is presented in Table 4. Same as model 1, the p-value of the model 2 is less than 0.0001, there is enough evidence to reject null hypothesis, proving a relationship between variables in this model. The adjusted R-squared for model 2 is 0.673 implying the predictor variables can explain 67.3% of the variability after the number of predictors in the model is adjusted. This value is higher than model 1 suggesting it is a better fit regression model, and this result is supported by the lower value of coefficient of variation. The main different between model 1 and model 2 is variable amount outstanding. Natural logarithm of the values is used in model 2. From

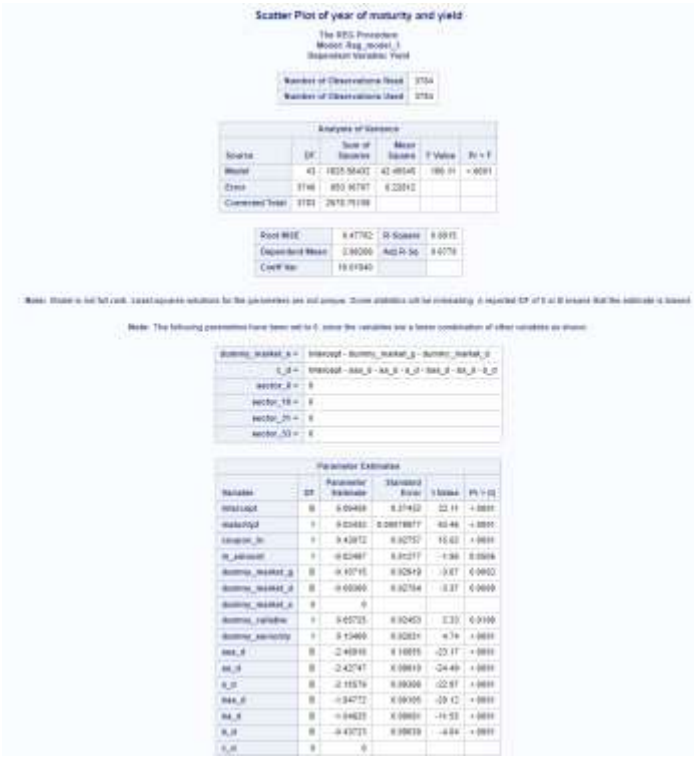
the p-value, it can be observed that the statistically significant for this variable decrease, compared to the value in model 1, which explain why model 2 is a better regression model.

Same as model 1, year of maturity, coupon, market of issue, seniority, and Moody's (Issue) credit rating are all resulting a p-value that lower than significant level 5%, suggesting they are statistically important to yield. Even though model 2 is a better regression model, and the p-value for amount outstanding decrease, it still slightly higher than the significant level, meaning it is still insignificant to yield. This situation also applied to Electronics sector, it is not important to the response variable. Yet, for the reason mentioned before, they will not be removed from the model and will be used in the estimation.

Based on the estimate coefficient, when all other predictor variables are equal to 0, the yield for bond is 4.212%. When there is 1 year increase in maturity, the value of yield will increase 0.035% on average. Yield will drop 0.0244% on average when the amount outstanding increase 1 billion of USD.

Table 4 shows the estimate coefficient of each variable, and this value is used to apply in the regression equation to calculate the yield of bond with the provided characteristics. The result is 2.251% with the regression equation: 4.212 + (0.035 * 10) + (0.098 * 2.5%) + (-0.0244 * (750,000,000 / 1,000,000,000)) + (-0.104) + 0.055 + 0.134+ (-2.431) + 0.047. The value is higher than model 1, but same as before, this value is overestimated since two variables that are not statistically important to yield is included.

Regression model 3



sector_1	1	0.20381	0.05657	3.60	0.0003
sector_2	1	0.17211	0.15363	1.12	0.2627
sector_3	1	0.31975	0.08248	3.88	0.0001
sector_4	1	0.06615	0.05199	1.27	0.2025
sector_5	1	0.01344	0.06026	0.22	0.8235
sector_6	1	0.11096	0.04182	2.65	0.0082
sector_7	1	0.01698	0.05335	0.32	0.7514
sector_8	0	0	.	.	.
sector_9	1	0.10109	0.06145	1.64	0.1001
sector_10	1	-0.23277	0.10307	-2.26	0.0240
sector_11	1	0.04859	0.03982	1.22	0.2224
sector_12	1	0.07270	0.04175	1.74	0.0817
sector_13	1	-0.19188	0.16271	-1.18	0.2384
sector_14	1	0.08348	0.04060	2.06	0.0398
sector_15	1	0.01089	0.04715	0.23	0.8174
sector_16	1	0.23465	0.07110	3.30	0.0010
sector_17	1	-0.06488	0.05345	-1.21	0.2307
sector_18	0	0	.	.	.
sector_19	1	0.01031	0.06057	0.15	0.8799
sector_20	1	0.28621	0.11574	2.47	0.0134
sector_21	1	-0.01038	0.06009	-0.16	0.8752
sector_22	1	0.28625	0.06402	4.47	<.0001
sector_23	1	0.08795	0.04218	2.09	0.0370
sector_24	1	0.34103	0.17122	1.99	0.0464
sector_25	1	-0.14596	0.04927	-2.96	0.0031
sector_26	1	0.02741	0.06856	0.40	0.6893
sector_27	1	0.12859	0.07841	1.64	0.1011
sector_28	1	0.12301	0.04328	2.84	0.0045
sector_29	1	0.10344	0.03418	3.03	0.0025
sector_30	1	0.12331	0.04381	2.81	0.0049
sector_31	0	0	.	.	.
sector_32	1	0.70199	0.10153	6.91	<.0001
sector_33	0	0	.	.	.
sector_34	1	0.06745	0.08372	0.81	0.4205

Table 5: Result of regression model 3

Table 5 indicate the result of regression model 3. The less than 0.0001 p-value of

model 3 provide sufficient evidence to reject null hypothesis, implying variables have correlation in this model. Model 3 have the highest adjusted R-squared compared to the first two model. There is 67.78% of the variability can be explained by the predictor variables. With the highest adjusted R-squared value and lowest coefficient of variation value, model 3 is the best fit regression model within three regression model. Model 3 utilizes the natural logarithm of coupon as the predictor, which is different to model 1 and 2.

Predictor variables that have a lower than significant level 5% p-value include year of maturity, coupon, market of issue, seniority, and Moody's (Issue) credit rating, they are all statistically relevant to yield. Even though amount outstanding variable and Electronics sector variable still have a higher than 5% p-value, the value decrease much compared to other model, especially for amount outstanding variable, it is very close to reach the significant level. As always, these two variables that are statistically not influential to yield will be retained to estimate the dependent variable.

Based on the estimate coefficient, when all other predictor variables are equal to 0, the yield for bond is 6.065%. When there is 1% increase in coupon rate, the value of yield will increase 0.431% on average. Yield will drop 0.025% on average when the amount outstanding increase 1 billion of USD.

In regression model 3, the estimation of yield for bond is 4.105%, calculated by the regression equation 6.065 + (0.034 * 10) + (0.431 * 2.5%) + (-0.025 * ($750,000,000 / 1,000,000,000)) + (-0.107) + 0.057 + 0.134 + (-2.427) + 0.049. Model 3 results the highest value of yield. Although the value is still overestimated due to the insignificant variables, it should be more accurate than the estimation of model 1 and 2 since it is the best fit of regression model.

Conclusion

In conclusion, a dataset with 3,784 observations is used to build three regression model and perform regression analysis to estimate the yield of bond. The dataset is merged by two datasets, multiple outliers observed by the extreme observation and boxplot are removed based on the calculation of upper inner fence and lower inner fence. Several variables are created based on requirements and dummy variables are created to run the regression model for categorical variables.

The 64 variables of 3,784 observation in the dataset is summarized in table and graph. The average yield is 2.98 with 0.841 standard deviation. The data is normally distributed with a bell-shaped histogram and value of skewness that suggests the distribution is symmetric. Variables that are chosen in the regression model are all summarize in the data summary section. It is observed that there is a positive relationship between yield and coupon with data point showing a trend from the bottom left-hand corner to upper right-hand corner in the scatter plot.

The yield of several specific characteristics bonds is required to be estimated. Three regression models are built with 49 predictor variables and yield is set as the response variable. The difference of three models is the variable of amount outstanding and variable coupon, natural logarithm of value is used in different model. All model has a null hypothesis of all independent variables has no relationship with the dependent variable.

The results of all three regression models provide sufficient evidence to reject the null hypothesis, meaning all independent variables have relationship with the dependent variable. The first regression model found number of independent variables statistically significant to yield, although there also some variables are insignificant. The estimated yield for bond is 1.772% with a possibility of overestimation due to two independent variables are relatively not significant. The second regression model with a higher adjusted R-squared value proves it is a better regression model have a similar result as model 1 and the yield for bond is 2.251% also with the possibility of overestimation. The last regression model is the best fit regression model with the highest adjusted R-squared value, it estimates the highest value of yield for bond which is 4.105% with the least overestimation.

The estimation of yield is calculated based on the three regression models and there are still room of improvements for the model. As mentioned, the value of yield may be overestimated. To improve the accuracy, it is recommended to increase the number of sample or remove the variables that are not statistically significant to the response variable.

References

Chen, James., (2020). Maturity Date. Available online: https://www.investopedia.com/terms/m/maturitydate.asp. Accessed 26 August 2023.

Konasani, Venkat & Kadre, Shailendra. (2015). Practical Business Analytics Using SAS. Available online: https://link.springer.com/book/10.1007/978-1-4842-0043-8. Accessed 26 August 2023.

Moody's investors services. Moody's Rating System in Brief. Available online: https://www.moodys.com/sites/products/productattachments/moody%27s%20rating%20system.pdf. Accessed 26 August 2023.

Reserve Bank of Australia. Bonds and the Yield Curve. Available online: https://www.rba.gov.au/education/resources/explainers/bonds-and-the-yield-curve.html#:~:text=A%20bond's%20yield%20is%20the,the%20price%20of%20the%20bond. Accessed 26 August 2023.