Topics in Economics: Financial Data Analytics

Group Assignment

Prof. Giang Nguyen

Lin Yu Chen (1A182G25-5)

Anh Jinho (1A172G01-4)

Part 1 (25 points): Clean data and construct all variables. Winsorize all continuous variables at 1% and 99% to control for outliners. Provide summary statistics (Mean, median, standard deviation, 25th percentile, and 75th percentile.

Table 1:

Statistic	Mean	Median	St. Dev.	Pctl(25)	Pctl(75)
year	2,015.157	2,015	2.584	2,013	2,017
cash_holding	0.214	0.171	0.161	0.097	0.285
total_asset	5.763	5.621	1.743	4.535	6.830
capex_ratio	0.032	0.021	0.033	0.008	0.044
leverage_ratio	0.087	0.046	0.107	0.000	0.137
age	3.956	4.143	0.669	3.611	4.431
sales	5.753	5.630	1.759	4.525	6.888
ebitda	0.093	0.081	0.058	0.053	0.119

Part 2 (25 points): Discuss why these two decisions are important? (Why do we need to use statistical learning to predict the level of cash holdings and dividend payout?)

Corporate cash holdings refer to the cash currently held by the company for expenditure rather than investments. It is important to understand what factors determine the cash holding ratio of firms because holding cash has opportunity costs. Having high cash holding allows the company to meet future contingencies at the expense of profits from projects with positive net present value (Al-Najjar, 2013). In contrast, having low cash holdings can affect the long-term solvency of the

company (Ahmed et al., 2018). Considering that cash is kept at the expense of higher profit, high levels of cash holding may also indicate agency problems between the management and shareholders (Jensen, 1986). The significance of using statistical learning to predict the level of cash holdings can be perceived from two perspectives. From the perspective of the management, using statistical learning will allow them to have a better understanding of the importance of the determinants of corporate cash holdings in crafting the corporate financial policies (Maheshwari & Rao, 2017). Meanwhile, for investors, this allows them to better understand the management's motivation behind holding more cash and the determinants of cash holdings level.

Dividend refers to the amount that the company pays to its shareholders for the capital they invested to fund the firm's activities. Therefore, dividend decisions affect both the wealth of the shareholders as well as the ability of the firm to retain its earnings to fund its succeeding projects. By using statistical models, investors will be able to understand the motivation behind the management's dividend decision and predict their dividend payouts based on the internal financial factors of the firm.

Part 3 (25 points):

a) Model cash holdings as a function of the given predictors. Discuss the estimation results (both statistical and economic significance). Discuss the model fit.

   How are your findings different from the existing literature? Hint: You need to read the existing literature on the determinants of corporate cash holdings and compare your results.

Table 2: Regression Output for Cash Holdings Model

	Dependent variable:
	cash_holding
total_asset	0.025***
	(0.001)
capex_ratio	−0.831***
	(0.025)
leverage_ratio	−0.290***
	(0.008)
age	−0.062***
	(0.001)
sales	−0.051***
	(0.001)
ebitda	0.634***
	(0.015)
Constant	0.604***
	(0.006)
Observations	24,874
R^2	0.427
Adjusted R^2	0.427
Residual Std. Error	0.122 (df = 24867)
F Statistic	3,084.662*** (df = 6; 24867)
Note:	*p<0.1; **p<0.05; ***p<0.01

Model Fit:

Based on the regression result, the Adjusted R-squared is 0.427, which indicates that the model is relatively good at explaining the volatility in cash holding levels of Japanese firms.

Total Asset:

From the results, we can see that the total asset is statistically significant. It has a positive stimate with 0.024606, which shows that total assets have a positive correlation with cash holdings in Japanese corporations. In past researches, we found that the total assets usually have negative correlation in emerging markets, as the negative relation indicates that firms with less total assets face difficulties in accessing the capital markets than larger firms, thus they will benefit from a

larger cash holding (Al-Najjar & Clark, 2017). On the other hand, we also found that in developed countries like the USA, the cash holding for firms with higher total assets is on the increase as the profits of these global companies are also on the rise. It would be heavily taxed to bring these profits back to the U.S, causing these firms to keep more cash overseas. We think that it can be the same case for Japanese corporations. Since more and more Japanese companies are getting globalized, which Japan is also one of the largest economies in the world, big firms might be holding more cash as they keep their profits overseas.

Capital expenditure ratio:

Based on the regression output, capital expenditure ratio is statistically significant and is negatively correlated with cash holdings ratio. This result is consistent with the study of Sher (2014) where he also estimated that there is a negative correlation between the two financial variables . According to his study, a negative correlation between cash holdings and capital expenditure ratio could indicate the extent to which Japanese firms choose to finance their investment spending through cash spending, due to financing constraints or asymmetric information about the nature of investment projects between management and investors.

Leverage ratio:

Based on the regression output, the leverage ratio is statistically significant. It shows that leverage ratio is negatively correlated with cash holdings ratio in Japanese corporations. Using a trade off perspective, we usually expect a company with a higher leverage to have higher cash holding as they are exposed to higher risk of getting bankrupt of financial difficulties (Al-Najjar & Clark, 2017). However, from several empirical studies, we can see that it turns out that the leverage ratio usually has a negative correlation with cash holding, as the leverage ratio of each company is

usually regarded as a proxy for them to issue additional debt (Ozkan & Ozkan, 2004; Guney et al., 2007). Financial Institutions, as a monitoring role, usually pay higher attention to high leveraged corporations, making these firms harder to hoard cash.

Company age:

In the output above, the firm age is statistically significant and has a negative correlation with cash holdings estimated at -0.06. Based on the static trade off theory, it is predicted that older firms have higher access to the capital market and therefore can hold less cash compared to younger firms. According to Heijanto and Budisantosa (2016), most studies that assessed the relationship between cash holdings and firm age concluded that there is a negative correlation between the two financial variables. Hence, this result is consistent with static trade off theory and the existing literature.

Sales & EBITDA:

According to the output, sales has a negative correlation with cash holdings. It can be inferred that Japanese firms use cash as a transaction medium to expand its operations that eventually increase its sales. This therefore decreases the amount of cash held by the firms. On the other hand, EBITDA and cash holdings have a positive correlation, which may imply that a significant portion of the company's earnings are often held by Japanese firms as cash in preparation for future contingencies.

b) Model dividend payout decision as a function of the given predictors. You can use the logistic model and LDA.

Discuss the analysis results (Discuss both statistical and economic significance for the logistic model. For the output of LDA, you can describe how you understand the output.)

Discuss the prediction performance, and how it varies with a given threshold (You can decide the threshold to discuss).

Hint: In (a) and (b), you will use the whole dataset as the training dataset.

Table 3: Logistic Regression Output for Dividend Payout

	Dependent variable:
	dividend_pay
total_asset	0.360***
	(0.037)
capex_ratio	5.567***
	(0.631)
leverage_ratio	−3.698***
	(0.186)
age	0.591***
	(0.034)
sales	0.463***
	(0.034)
ebitda	2.985***
	(0.338)
Constant	−4.740***
	(0.142)
Observations	24,874
Log Likelihood	−7,907.306
Akaike Inf. Crit.	15,828.610
Note:	*p<0.1; **p<0.05; ***p<0.01

Logistic Regression Model

The regression output also indicates that the predictor variables are all statistically significant. Total asset, capital expenditure ratio, firm age, sales, and EBITDA have a positive correlation with dividend payout. On the other hand, only leverage ratio has a negative correlation with dividend payout.

Based on these outputs, it can be concluded that positive financial performance and/or higher firm expenditure are most likely to induce dividend payout. The output indicates the following: (a) an increase in total assets will increase the likelihood of a dividend payout by 0.36; (b) an increase in capital expenditure ratio leads to an increased dividend payout likelihood of 5.57; (c) an increase in sales will increase the likelihood of dividend payout by 0.46; and (d) an increase of EBITDA will translate to an increase the likelihood of dividend payout by 2.98. These relationships might be an indication that increased assets and capital expenditure generally lead to higher profits for Japanese firms, therefore higher residual income for equity holders to claim. Moreover, higher sales and EBITDA seem to have a positive correlation with dividend payout, indicating that these are significant decision-making factors for management in issuing dividend payouts.

Meanwhile, higher debt financing expectedly has a negative impact on dividend payout. Based on the results above, an increase in leverage ratio leads to a decrease in the likelihood of a dividend payout by -3.70. Given that higher leverage leads to higher debt repayment, this consequently makes a delayed dividend payout more likely. This is mainly because of the fact that equity holders have residual claims over a firm's profit; hence, debt repayments are prioritized over dividend payout.

glm.pred	No	Yes
No	1244	485
Yes	2532	20608

The training error rate is 12.2% for the logistic model. The logistic model has a higher prediction rate for corporations that pay a dividend with a 2.29% error rate with only 485 corporations incorrectly labelled out of 20608 corporations. However, the logistic model performs

pretty poorly in corporations that don't pay a dividend. Out of 3776 firms that don't pay dividend, the logistic model got 2532 firms labeled incorrectly, which is a 67% error rate.

Linear Discriminant Analysis

```
> #LDA
> library(MASS)
> lda.fit_b = lda(dividend_pay~total_asset+capex_ratio+leverage_ratio+age+sales+ebitda, data = data)
> lda.fit_b
Call:
lda(dividend_pay ~ total_asset + capex_ratio + leverage_ratio +
    age + sales + ebitda, data = data)

Prior probabilities of groups:
       No       Yes
0.1520061 0.8479939

Group means:
    total_asset capex_ratio leverage_ratio      age    sales     ebitda
No     4.223560  0.02403386     0.10606370 3.489274 4.154756 0.09613499
Yes    6.039244  0.03318080     0.08377902 4.039742 6.039815 0.09250231

Coefficients of linear discriminants:
                     LD1
total_asset     0.0676789
capex_ratio     5.2645131
leverage_ratio -3.1701804
age             0.6391068
sales           0.4091443
ebitda          1.3242407
```

In our LDA model, corporations that pay a dividend are labeled as TRUE which corporations that don't pay a dividend are labeled as FALSE. In this dataset, we can see that 84.25% of the corporations pay a dividend and 15.75% of the corporations don't pay a dividend. Out of all of the predictors, leverage ratio is the only one that has a negative correlation with dividend payment, which has -3.1701804 per unit change for a unit of increase in dividend payment. The reason under this can be pretty intuitive. As we know companies with a high leverage usually has a lower cash holding, it will be harder for them to make a dividend payment for their stockholders.

```
> lda.prob_b = predict(lda.fit_b, type = "response")
> lda.prob.class_b = lda.prob_b$class
>
> contrasts(data$dividend_pay)
    Yes
No    0
Yes   1
> # Confusion Matrix
> table(lda.prob.class_b, data$dividend_pay)

lda.prob.class_b    No   Yes
             No   1254   515
             Yes  2527 20578
>
> # Test error rate
> mean(lda.prob.class_b == data$dividend_pay)
[1] 0.8777036
```

We can see that with the confusion matrix at the threshold of 50%, the training error rate is 12.3%. Although the training error rate is quite high, we can see that LDA did a pretty good job predicting the firms that will pay a dividend, which generated a 2.44% error rate with only 515 corporations incorrectly labelled out of 21093 corporations. However, the results of LDA and the actual results for firms that do not pay dividend is totally unacceptable, which generated a 66.8% error rate with 2527 corporations mislabeled out of only 3781 corporations that don't pay dividend.

```
> #threshold
> glm.pred2=rep("No",24860)
> glm.pred2[glm.probs>.2]="Yes"
>
> # create a confusion table
> table(glm.pred2,data$dividend_pay)

glm.pred2    No   Yes
      No     87     9
      Yes  3694 21084
>
> # % of predicting correctly the movement [1 - this % = training
 error rate]
> mean(glm.pred2==data$dividend_pay)
[1] 0.8511297
>
```

To understand how the threshold will affect the general result of LDA, we lowered the threshold to 20% to observe the results. When we lower the threshold, we can see that the training error rate increased from 12.3% to 14.9%. Although the result of the enter training error rate is higher, the results of LDA prediction improved from a 2.44% error rate to 0.04%, which is almost perfect. However, we can also see the lower threshold of 20% worsened the LDA prediction

performance for firms that do not pay a dividend, moving from the originall 66.8% to an even higher 97.6%.

In this part, the logistic model performs slightly better than the LDA model. The logistic model has an error rate of 12.2% and the LDA model has an error rate of 12.3%. The logistic model performs better in predicting corporations with a dividend payment which its error rate of 2.29% is better than the LDA model which has an error rate of 2.44%. On the other hand, while both are poor in predicting results for corporations without a dividend payment, LDA is slightly better with an error rate of 66.8% compared an 67% error rate for the logistic model.

c) Let's divide the whole sample into training data and test data. The observations after 2016 belong to the test data sample. Assessing prediction accuracy using the logistic regression model and LDA for the case of decision payout in (b).

```
> lda.class_c = lda.pred_c$class
> # Confusion Matrix
> table(lda.class_c, dividend_pay.2016)
           dividend_pay.2016
lda.class_c   No  Yes
        No   587  242
        Yes 1123 9767
>
> # Test error rate
> mean(lda.class_c == dividend_pay.2016)
[1] 0.8835225
```

```
> glm.pred[glm.probs > 0.5] = "Yes"
> table(glm.pred, dividend_pay.2016)
        dividend_pay.2016
glm.pred   No  Yes
     No   550  223
     Yes 1160 9786
> mean(glm.pred == dividend_pay.2016)
[1] 0.8819865
```

From the two models, we can see that the LDA model with an error rate of 11.65% has a better prediction than the logistic model which has an error rate of 11.80%. For predicting companies with dividend, the logistic model is better in terms of performance as it has a lower error rate with 2.22% while LDA has an error rate with 2.41%. However, when it comes to

predicting firms without dividend payment, LDA is better with an error rate of 65.6% than the logistic model which has an error rate of 67.83%.

Part 4 (25 points):

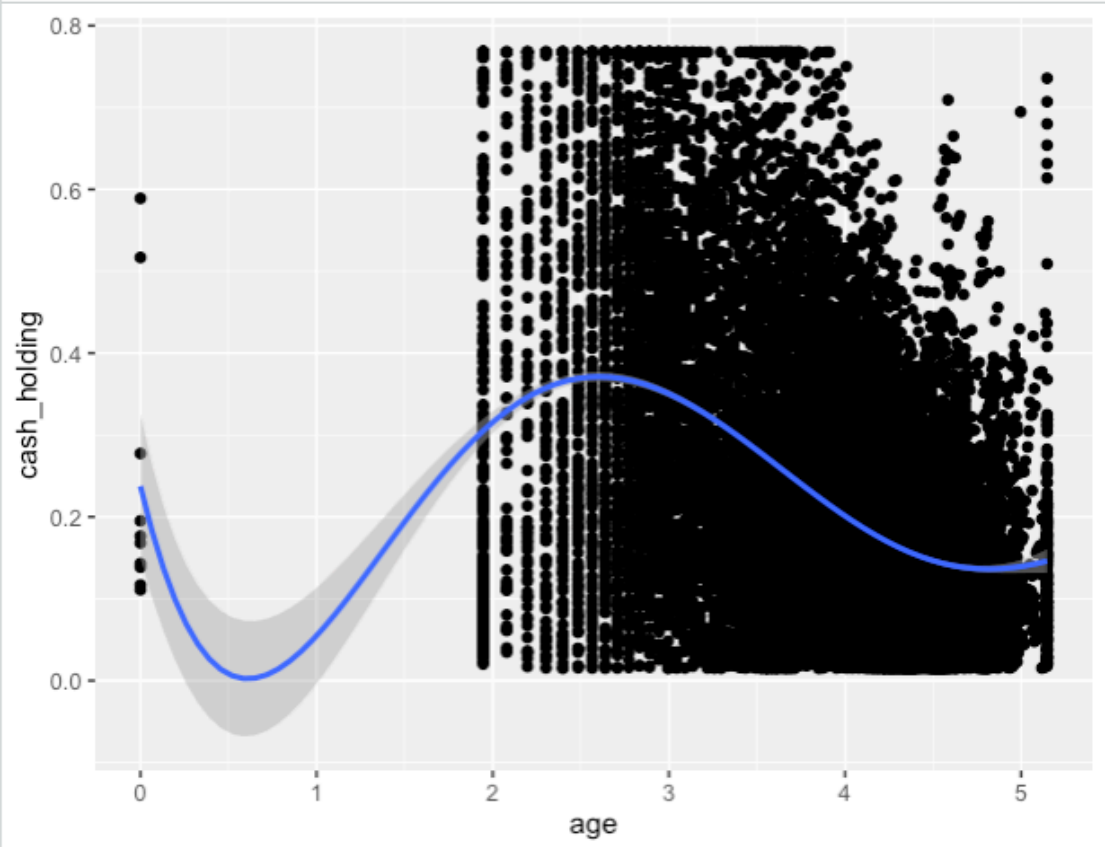
In the previous parts, we did not include polynomial functions of the predictors. Given that we are interested in the non-linear associations of company age (Age).

Find the order of *A*ge that gives us the lowest mean squared errors (MSE) using all cross-validation approaches.

Note: We have 2 corporate decisions, and we use different models in the previous parts. This part requires you to find the flexibility for all models using cross-validation approaches (for example, the validation set approach, LOOC, and 10-Fold Cross-Validation.

We used several different approaches such as k-fold cross validation and leave one out cross validation to test for the most appropriate polynomial regression model to estimate the relationship between cash holding ratio and company age. Several data cleaning and transformation commands were used first, as we had to convert the variables into either logs or they had to be scaled with total assets. We used the sapply function to transform all variables to numeric, and used the na.omit and is.infinite functions to remove all NAs, NaNs, and Infs from our dataset. The model we estimated did not include any control variables, the only regressor was company age and the dependent variable was cash holding. According to the k fold cross validation technique, we used 10 different folds, where each fold was an instance of a training and testing split. We tested 10 different polynomial regression models, with the degree of age ranging from 1

to 10. The MSE and adjusted MSE both showed that the polynomial regression of degree 7 had the lowest MSE.



We also ran the validation set approach and that gave us a RMSE of 3.82. The validation set approach has a higher RMSE since it runs only one iteration with an 80-20 sample split as opposed to the k fold cross validation approach which runs several iterations.

We first tried using loop in order to get the result for Age^n but it took so long time to get the result so we had to adjust it up to 10th power. And we also tried to use LOOCV in the code but it also made us wait for too long to get the results.

Therefore, we had to delete LOOCV and just leave k fold in order to make the code work instantly. In order to perform all the process that we wanted, I think that we will need access to virtual machine (remote server) in order to perform.

```
> # Model performance
> data.frame(RMSE = RMSE(predictions, test.data$age), R2 = R2(predictions,
 test.data$cash_holding))
      RMSE        R2
1 3.819452 0.2301945
```

	delta1	delta2
1	0.01488277	0.01488235
2	0.01489113	0.01489020
3	0.01469636	0.01469536
4	0.01458502	0.01458403
5	0.01457879	0.01457774
6	0.01457957	0.01457846
7	0.01457585	0.01457482
8	0.01457839	0.01457718
9	0.01458205	0.01458041
10	0.01457932	0.01457758

References:

Al-Najjar, B. (2013). The financial determinants of corporate cash holdings: Evidence from some emerging markets. International Business Review(22), 77-88.

Al-Najjar, B., & Clark, E. (2017). Corporate governance and cash holdings in MENA: Evidence from internal and external governance practices. Research in International Business and Finance, 39, 1-12.

Ahmed, et al. (2018). DETERMINANTS OF CORPORATE CASH HOLDINGS: AN EMPIRICAL STUDY OF CHINESE LISTED FIRMS. Corporate Ownership & Control, 15(3), 57-65.

Guney, Y., Ozkan, A., & Ozkan, N. (2007). International evidence on the non-linear impact of leverage on corporate cash holdings. Journal of Multinational Financial Management, 17, 45-60

Jensen, M. (1986). Agency costs of free cash flow, corporate finance, and takeovers. American Economic Review, 76, 323-329.

Ozkan, A., & Ozkan, N. (2004). Corporate cash holdings: An empirical investigation of UK companies . Journal of Banking & Finance, 2103-2134.

Contributions:

Lin Yu Chen: Worked on Parts 1, 2, 3 4

Anh Jinho: Worked on Parts 2 and 4