

**BC2406-Analytics 1**

**Lab Quiz 2nd Submission**

**Khoo Ju Seng Bryant U1420978E**

1. **Question 1**

**Data Exploration**

The aim of doing Data Exploration is to understand the given dataset, and compare the data to the business problem. I will be using the following questions as the guidelines for Data exploration

1. Is the data sufficient?

* Firstly, run the summary function to find out if there are many NA values for dvc and ib. There are NA values in dvc and ib as shown below:

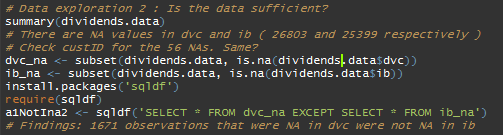


Fig. 1 Checking NA values for dvc and ib

* Using a package called ‘sqldf’, I was able to run a command that resembles sql’s select function to find out how many rows were NA in dvc but not in NA. The findings can be seen in Figure 1.

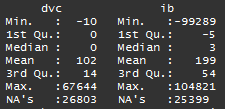


Fig 2. NA values for dvc and ib

1. Does the data provide predictive value?

* I used ggplot 2 to plot a scatter plot with a smooth line, to see the general datapoints and relationship between the two columns.

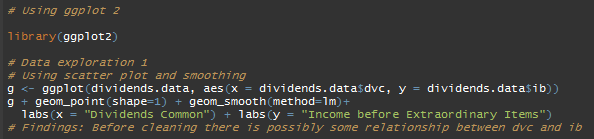


Fig. 3 R-code Using gg plot

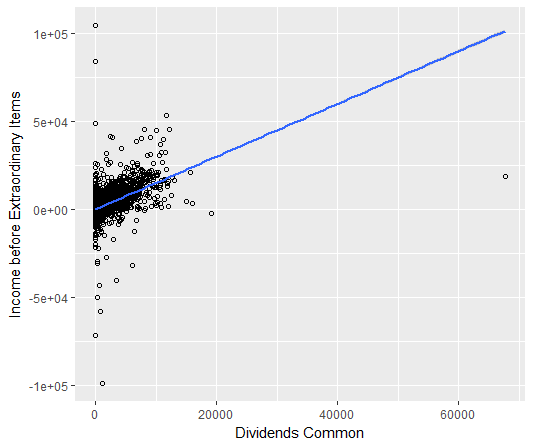


Fig 4. Scatter plot of ib against dvc

* Findings: before data cleaning, we can already see that there **is possibly** a relationship between dvc and ib. **However, there are still outliers.**

1. How is the data quality and are there any anomalies?

* After doing summary(), I noticed that there were negative dvc values.
* I ran the following code to find out how many rows were negative, and which rows are they:

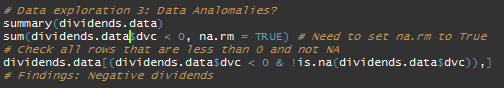


Fig 5. R-code to find explore rows

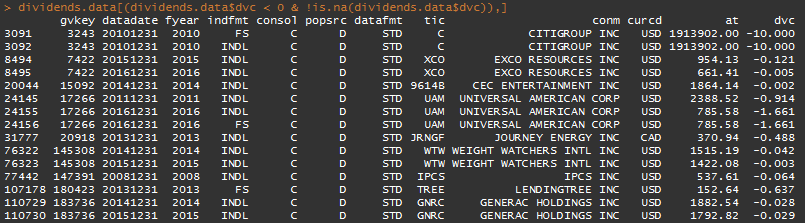


Fig 6. Results for code ran in Fig 5.

1. **Question 2**

**Data Cleaning**

1. Keep records with Stock Exchange Code = 11, 12, or 14.
   * I ran the following R-script, using the subset function on the original data *diviends.data* with the above conditions for the Stock Exchange Code (exchg) to obtain cleaned data *dividends.data.2a* for question 2a.
   * The number of records after execution is 70,161 rows.

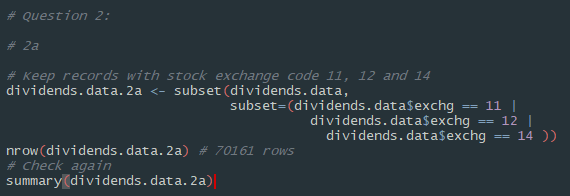


Fig 7. R-code for data cleaning in 2a

1. Keep records with currency = USD.
   * I ran the following R-script, using the subset function, based on the condition that currency is USD on the cleaned data *diviends.data.2a* to obtain cleaned data *dividends.data.2c* for question 2c.
   * The number of records after execution of 2a and 2b is 68,876 rows.

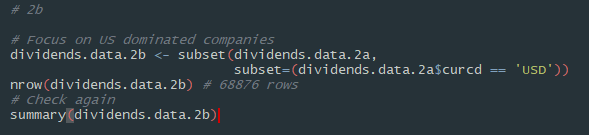


Fig 8. R-code for data cleaning in 2b

1. Drop records for SIC sector code from 6000 to 6999 and 4900 to 4999
   * I ran the following R-script, using the subset function, based on the above condition on the cleaned data *diviends.data.2b* to obtain cleaned data *dividends.data.2c* for question 2c.
   * The number of records after execution of 2a, 2b and 2c is 35,791 rows.

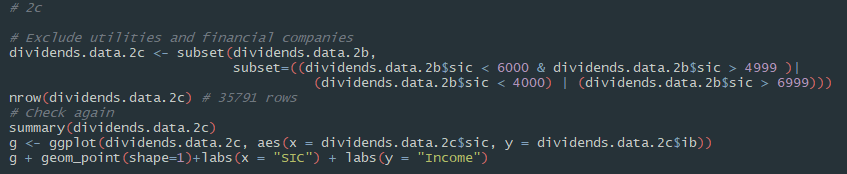


Fig 9. R-code for data cleaning in 2c

* To double check that I cleaned the correct range for SIC, in case of any mistakes in my conditions I used gg plot to plot a scatterplot. From the scatter plot, it seems that the subset function removed the right region of SIC values (highlighted in red below).

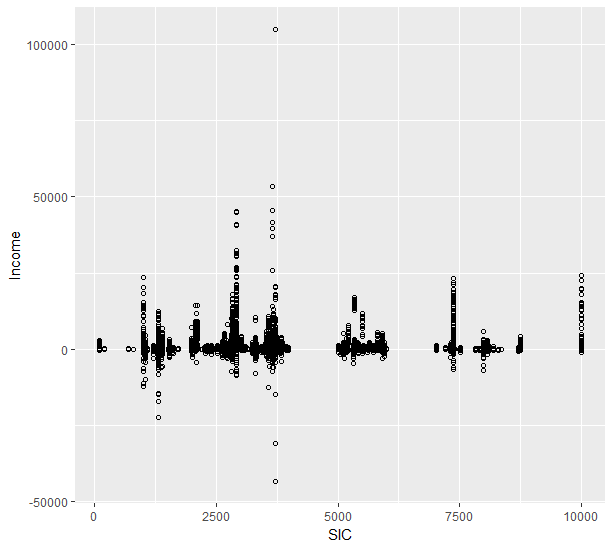


Fig 10. Scatterplot obtained from ggplot

1. Delete observations that do not report income (ib=NA)
   * I ran the following R-script, using the subset function, based on the above condition on the cleaned data *diviends.data.2c* to obtain cleaned data *dividends.data.2d* for question 2d.
   * The number of records after execution of 2a, 2b, 2c and 2d is 34,195 rows.

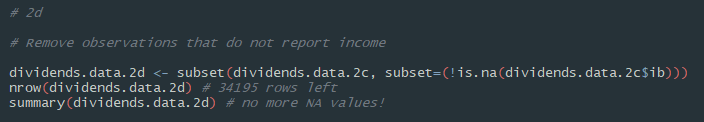


Fig 11. R-code to remove ib’s NA rows in data

1. What other data cleaning steps should be executed that I consider necessary and obvious?
   * I firstly removed NA values for dvc and negative values for dvc.
   * Also, there were no more NA values for ib.
   * I also removed NA values for a few more independent variables X (spi, idit and costat) that I felt would be used later on.

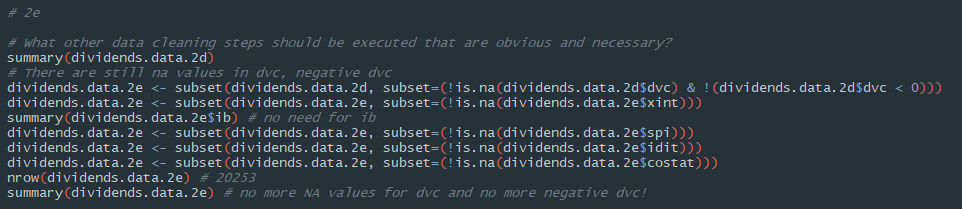


Fig 12. Remove NA and neg. dvc and NA for X variables

* Next, I realised that there are still 1355 rows is NA values in the independent variable prstkc. I ran the following code:

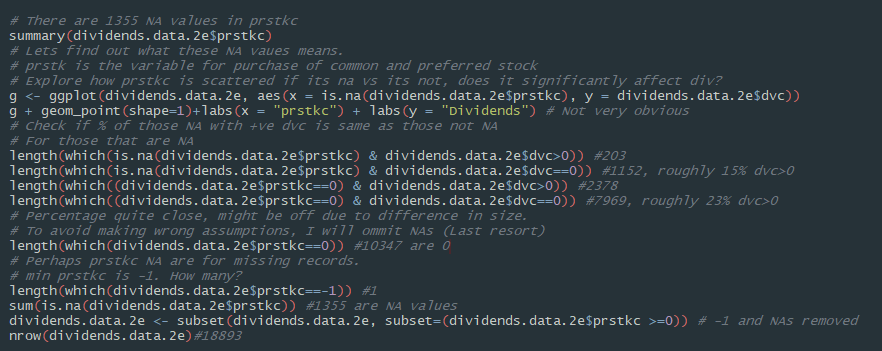


Fig 13. Removing prstkc NA and -1 values

* I firstly plotted a scatterplot to look at how the rows with NA prstkc values were spread out in terms of dividends (dvc), but the results were not very obvious

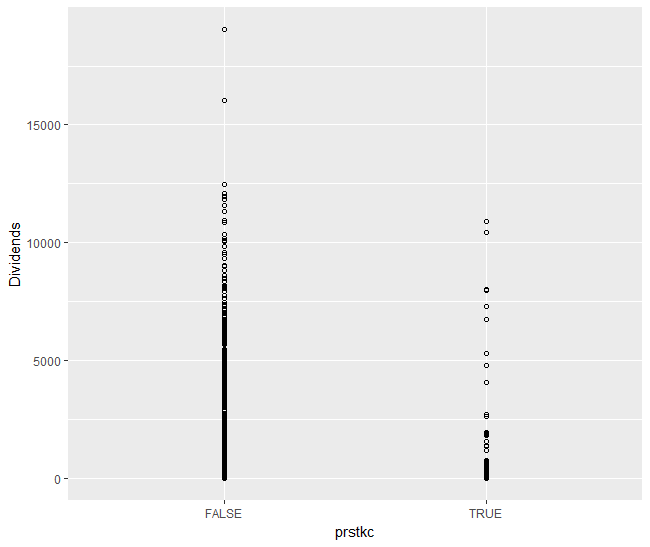


Fig 14. Dvc against is.na(prstkc)

* Hence, I decided to check if the percentage of those with NA and have +ve dvc values is the same as those that are not NA. And the percentage seems to be quite close.
* To avoid making the wrong assumptions for the NA values, I decided to omit these rows as a last resort.
* After removing values with prstkc being NA and -1, I had **18,893** rows left.
* One final data cleaning step at this stage is to convert years to factors. However, one thing that I took note of is that there is only one observation with fyear == ‘2017’. This will pose as a problem later on when I do regression and CART. However, I decided to leave the row for question 3, and it will be removed in later parts when needed.



Fig 15. Convert fyear to factors

1. **Question 3**

**Generating summary table:**

* I ran the following code to generate the summary table

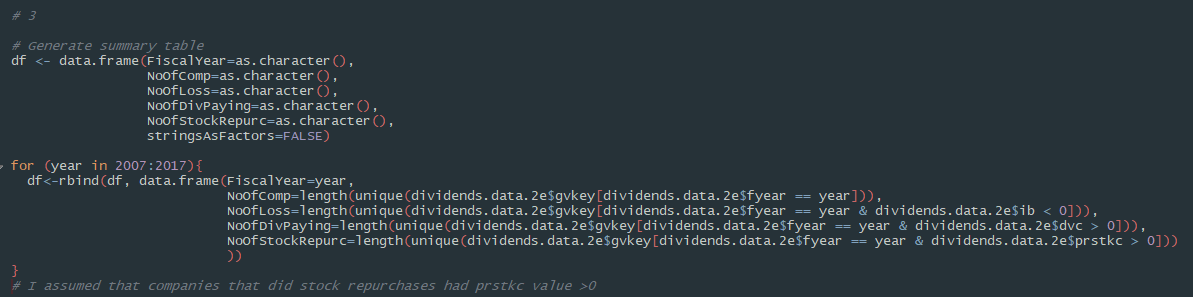


Fig 16. R-code to generate summary table

* I initialised a new dataframe and its columns.
* A for loop is used to iterate through years 2007 to 2017 and data for each year as rows.
* I assumed that stock repurchase meant positive prstkc value.
* The resulting summary table is as follows:

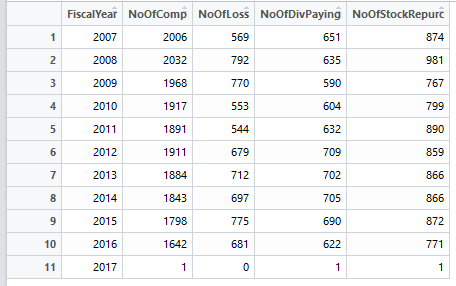


Fig 17. Summary Table (qn3)

* To observe the number of firms that are making losses, paying dividends and repurchasing stock, I found it hard to visualize from the table. **Hence instead of deciding on whether the number increased or decreased, I decided to observe the general trend over the years using ggplot.**
* Since the number of companies in each row is different, I decided to plot the following graphs.
  + Num. of Comp. that Loss / Num of Comp. against Fiscal Year
  + Num. of Div Paying Comp. / Num of Comp. against Fiscal Year
  + Num. of Comp. thatStock Repurchase / Num of Comp. against Fiscal Year
* Hence, the Y values would be a ratio ranging from 0 to 1. I will be using bar plots to represent these values.
* We also need to take note that year 2017 is only represented by 1 observation. It would be good to not take the year’s value too seriously.

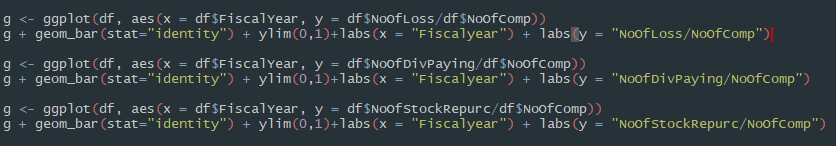


Fig 18. R-code for ggplot

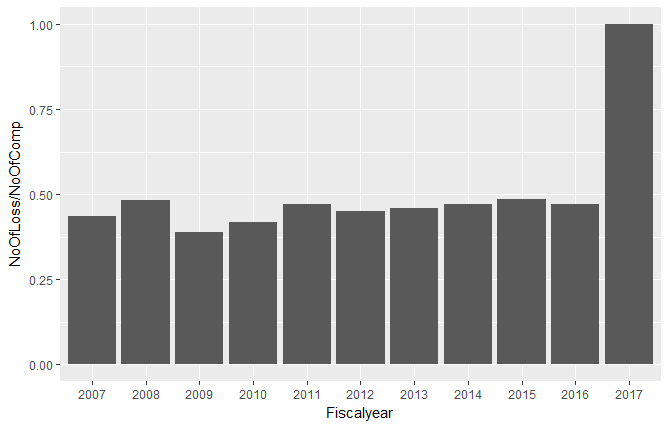


Fig 19. Ratio of companies suffering from loss

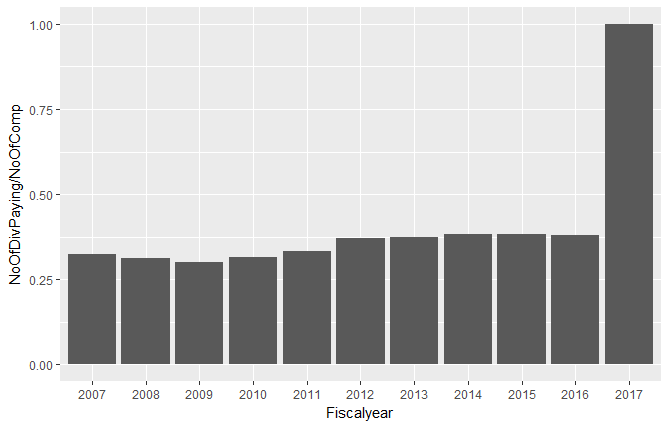


Fig 20. Ratio of companies paying dividends

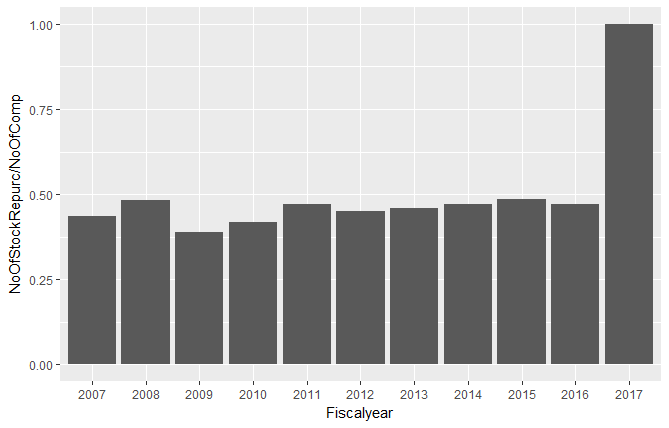


Fig 21. Ratio of companies repurchasing stock

* From figures 19 to 21, we can see that generally the ratio of companies that fall under the 3 categories remained quite stable from 2007 to 2016.
* Furthermore, I observed that there is a dip in 2009 in all of the bar plots.

1. **Question 4: Is dividend pay-out policy affected by earnings**
2. **Definition of earnings.**
   * I defined the earnings for each observation in the dataset as the income before extraordinary items (ib) column. Positive ib would indicate that the company is making profit, negative would mean a loss for that period.

Once earnings is defined, we can conduct the following methods in 4b to 4e to find out if dividends payout policies are affected by earnings. In order to fully explore the relationship between dividends and earnings, I included other variables that I deemed relevant.

The 9 Variables that I use for model selection:

* fyear (factor)
* at
* ib
* idit
* spi
* xint
* exchg
* costat
* sic

1. **Linear Regression.**
   * In before performing linear regression, I performed model selection to pick the best model. I used the regsubsets() function to find the best subsets with BIC, adjusted r-squared and cp as criterias.
   * Best subset with BIC had 7 variables and adjr^2 and cp had 9.

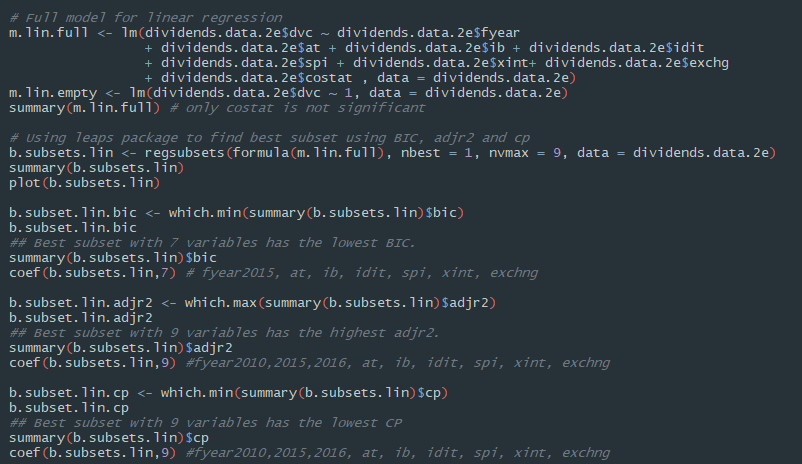


Fig 22. R-code for regsubset() for linear regression model

* Next, I used stepwise selection, with direction set as “foward”.
* The results are 7 variables as follows:
  + Variables that are most significant: at, ib, spi, fyear2015, xint, exchng, idit
  + Variables that are also significant: fyear 2010m fyear2016, fyear2017

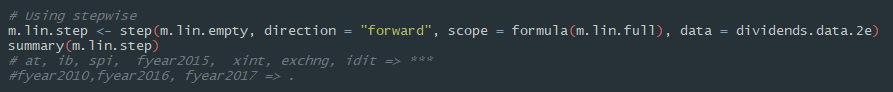


Fig 33. step-wise for linear model

* The first model I will be using is that with 7 variables from BIC and step-wise.
* After performing train test split, I ran the following code. The results in the form of RMSE for train and testsets have been commented into the code for convenience.

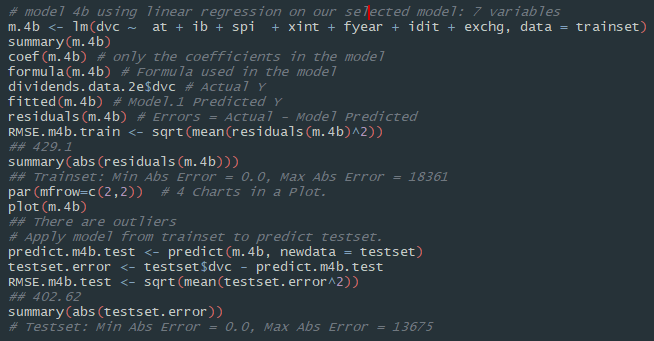


Fig 34. Linear Regression with 7 variables

* When we look at the coefficients of the regression model,

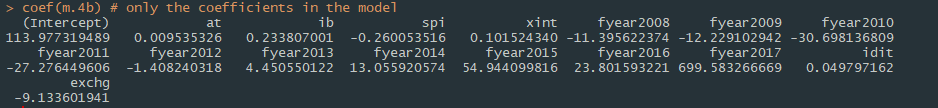


Fig 35. Linear regression model coeff with 7 var.

* It appears that ib has a coefficient of 0.233. **Based on this regression model, increase in roughly 4 dollars in earnings will result in a one dollar increase in dividends payout**.
* Next, I attempted the same steps with a model with only the most significant variables.
* When comparing this model to the previous model, model m.4b.2 has a slightly smaller RMSE value for the test set.
* I deduced that the model with lesser variables would be the better one for the following reasons.
  + It could perform better than the one with 2 more variables.
  + Having lesser variables in the model reduces the chances of overfitting.
* For the coefficients, it seems that the coefficient for ib in this model is very close to that from the previous model. Hence I would draw the same conclusion from it.

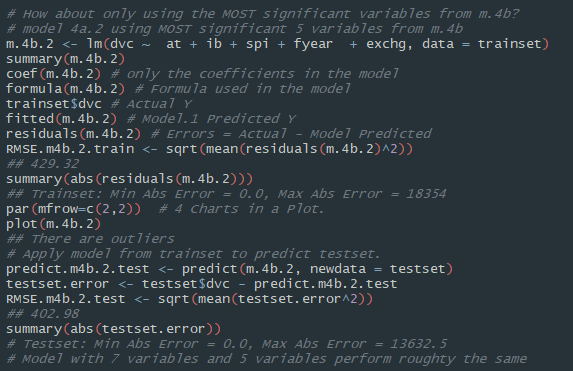


Fig 36. Linear Regression with 5 Most Sig. variables

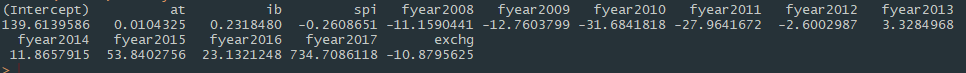


Fig 37. Coefficients for 5 variables

1. **Logistic Regression**
   * Before performing logistics regression, I ran the following code to discretize the dvc column.
   * Observations with 0 dvc will be factored as “0” for companies with no dividend pay-outs and those with dvc>0 will be factored as “1” for companies with dividends pay-outs during that period.

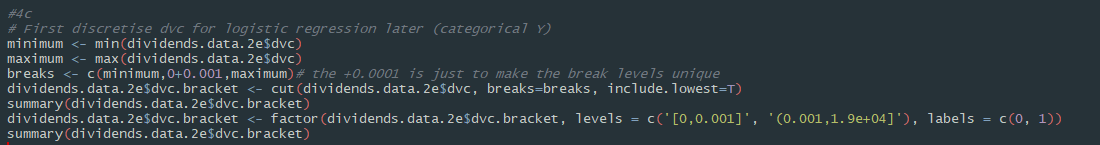


Fig 38. Discretizing Y variable

* Next, I used regsubsets() and step() for model selection.
* The model selected based BIC had the least number of variables 7, stepwise had 8 significant variables, and both adjr^2 and cp had 9 variables (full model)
* I decided to run logistics regression twice, once on the model with 8 significant variables, and another time with only the most significant variables.



Fig 39. Model selection for Logistic Regression

* Logistics regression was performed on the trainset and testset from 4b. Variables used were exchg, costat, ib, spi, xint, idit, fyear and at (based on stepwise selection)



Fig 40. Logistic Regression with above 7 var.

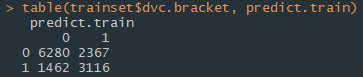


Fig 41. Confusion matrix for trainset

* For the train set, I obtained the above confusion matrix in Fig 41.
* There were 2367 false positives, and 1462 false negatives.
* Misclassification rate is 0.29

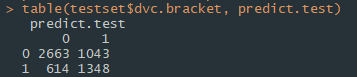


Fig 42. Confusion matrix for testset

* For the test set, I obtained the above confusion matrix in Fig 42.
* There were 1043 false positives and 614 false negatives.
* Misclassification rate is 0.292
* Next, I repeated the steps for only the most significant variables (\*\*\*):
  + exchg, costat, ib, spi, xint

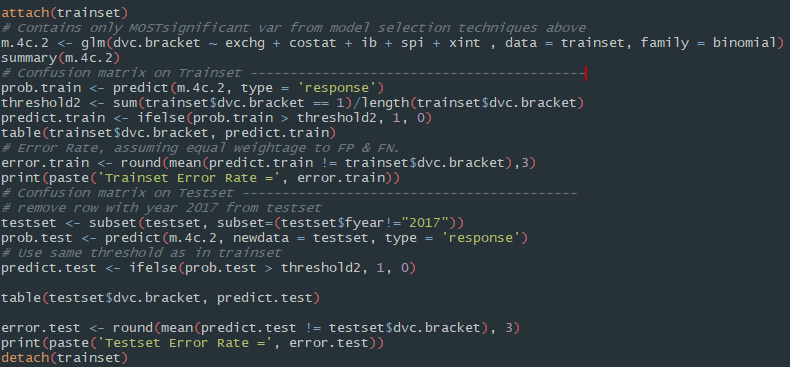


Fig 43. Logistic Regression with most sig. var.

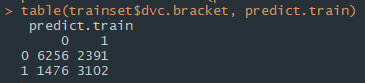


Fig 44. Confusion matrix for train set

* For the train set, I obtained the above confusion matrix in Fig 44.
* There were 2391 false positives, and 1476 false negatives.
* Misclassification rate is 0.292

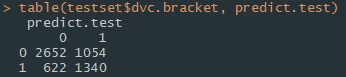


Fig 45. Confusion matrix for test set

* For the test set, I obtained the above confusion matrix in Fig 45.
* There were 1054 false positives, and 622 false negatives.
* Misclassification rate is 0.296

Since the error rates for logistic regression models with 7 or 5 variables are quite identical, I decided to use the one with 5 variables to reduce the likelihood of overfitting.

Going back to the question on whether dividend payout policy is affected by earnings, I used Odds Ratio.



Fig 46. Finding odds ratio for ib

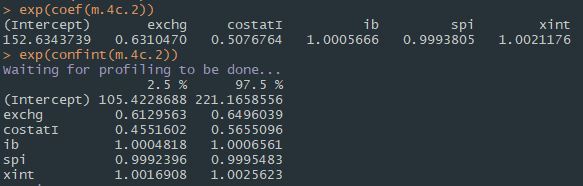


Fig 47. Odds ratios and confidence intervals

* Based on the odds ratio obtained, for every 1 unit increase in earnings (ib), odds of dividends payout multiply by 1.0005666.
* This might seem small, but when ib increases by 1000 units, odds of dividend payouts is multiplied by 1.76 times. When ib increases by 10000 units, odds of dividend payouts is multiplied by 288 times. The odds increase exponentially.

1. **Decision Tree (CART)**
   * For decision trees, I did not perform train-test split since rpart will do that for us.
   * I used the following variables for Decision tree:
     + Fyear, at, ib, idit, spi, xint, exchg and costat
     + dvc.bracket for categorical Y. Hence method for rpart is “class”
   * Since the dataset is pretty big (18893 rows), I increased the minsplit size to 500 so that the minimum number of observations in a node is set to 500. It would not be realistic to have too big a tree. I used cp = 0 initially to let the tree grow to its maximum.

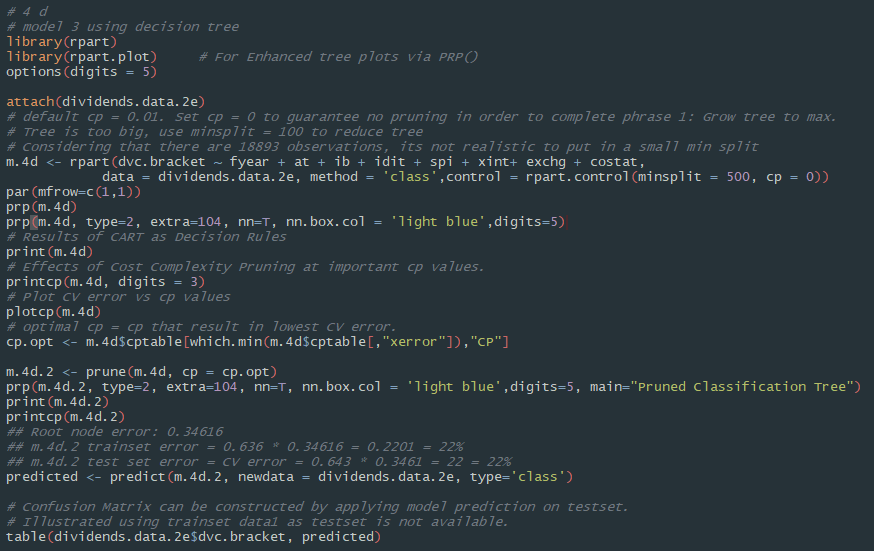


Fig 48. R-code for decision tree (CART)

* I obtained the following tree and cpplot:

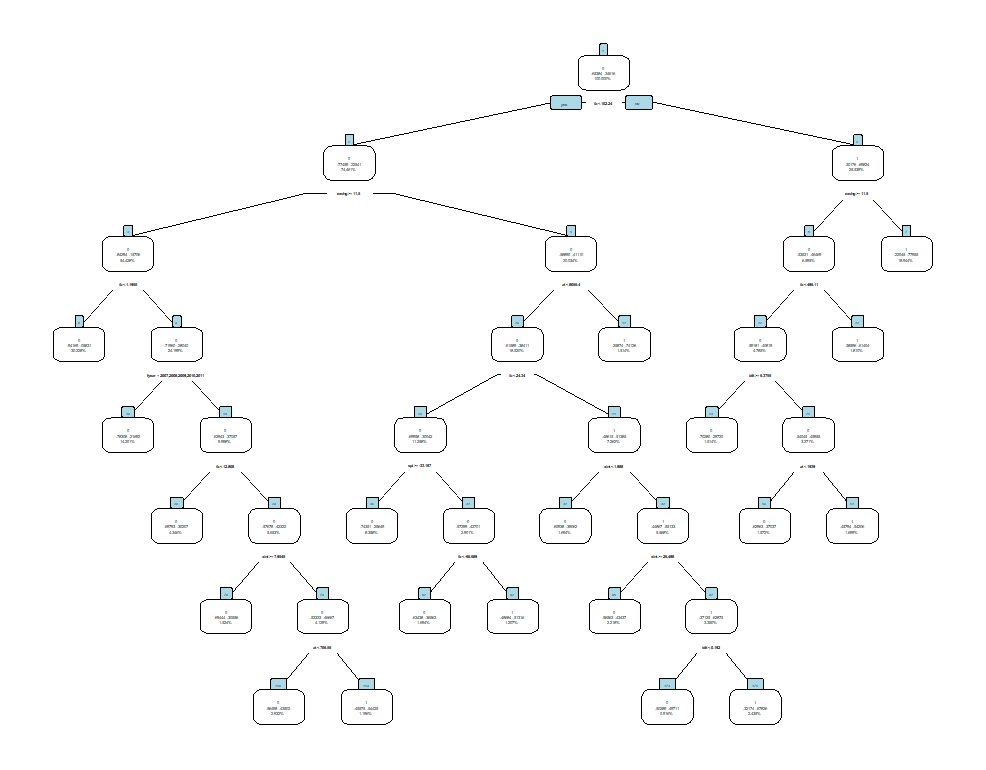
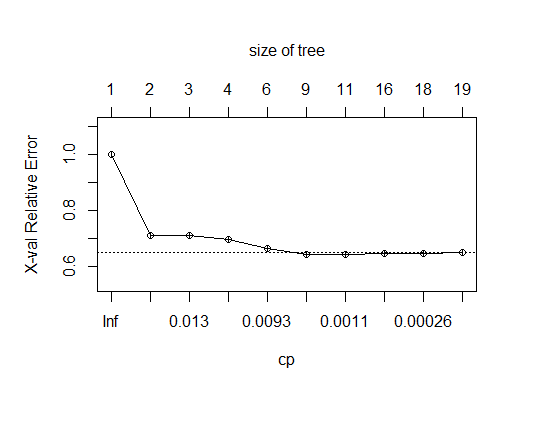
 

Fig 49. Decision tree before pruning

* Using printcp() I observed the cp, rel error and xerror values:

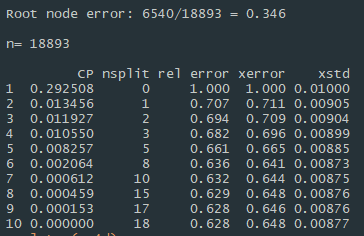


Fig 50. printcp() output before pruning

* Next, I performed pruning to reduce the size of the tree. I pruned the tree based one the optimal cp with lowest cross val. error. The optimal cp is 0.0020642.
* I obtained the following tree with 8 splits.

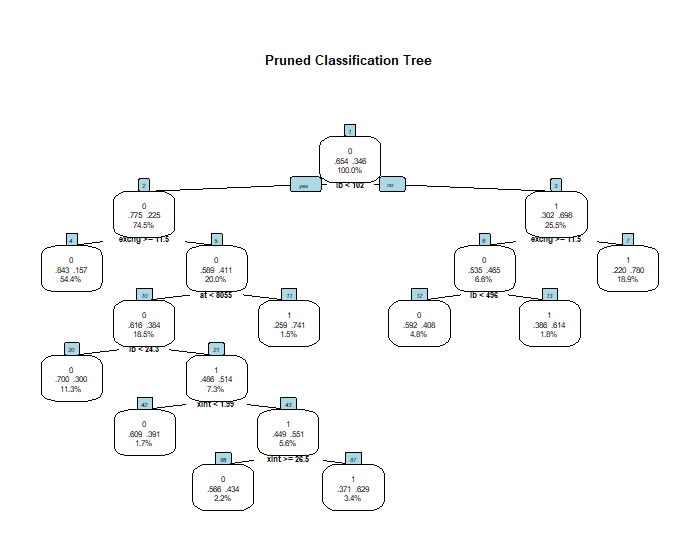


Fig 51. Pruned Classification Tree

* Root node error: 0.34616
* Trainset error: 0.636\*0.346161 = 22%
* Testset error: 0.643\*0.346161 = 22%
* Confusion Matrix:
  + False positives: 1230
  + False negative: 2930
  + Misclassification rate: 22%

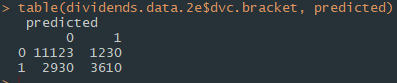


Fig. 52. Confusion matrix for CART

* Back to the question on whether dividends is affected by earnings (ib)
  + In the first node, 65.4% of the observations do not have dividends.
  + The first split condition is if ib < 102.
  + For observations with ib < 102, the percentage of those without dividends increases to 77.5%. For those with ib > 102, the percentage of those without dividends fall greatly to 30.2%.
  + A similar pattern is seen in splitting at node 10, depending on whether ib is < 24.3.
  + For these splits that occur, splitting into a lower ib value increases the percentage of observations with no dividend payouts, and splitting into a higher ib value decreases this percentage.
  + Thus, **the higher the earnings, the more likely for dividends** to be given out in general.

1. **Segregating data into 3 timeframes and repeating 4b-4d**

* I ran the following R-code to segregate the data into timeframe1, timeframe2 and timeframe3.
* I performed train-test split on all 3 timeframes with split ratio 0.7.

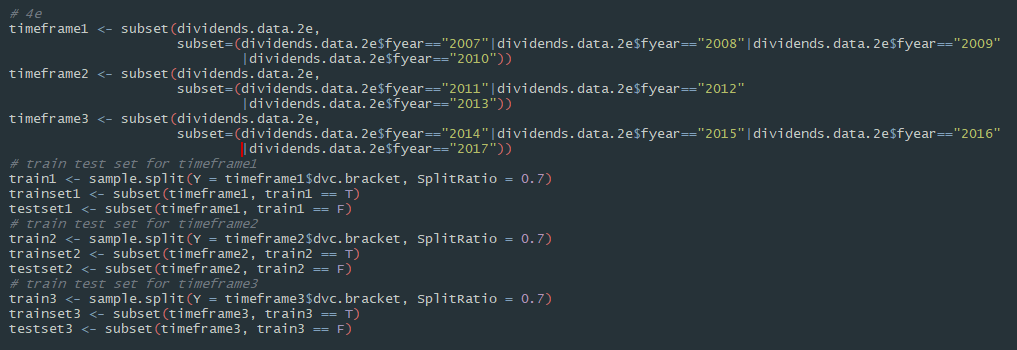


Fig 53. Segregating data and train-test splits

* Then I repeated 4b-4d for each of the 3 splits. I will not be cluttering the report with the repeated R-code. However, I will be analysing the following.
  + Coefficient for ib in linear regression
  + Odds ratio for ib in logistic regression
  + Splits in the decision trees
* For each timeframe I obtained the following coefficients for linear regression:
  + I observed that the coefficient for time frame 2 is the highest at 0.287. This means that for every unit increase in ib during that time period will increase the amount of dividends the most.
  + The effect of ib on dvc is the least in timeframe 3 since the coefficient is lowest, at 0.194.



Fig 54. Coef. for timeframe 1



Fig 55. Coef. for timeframe 2



Fig 56. Coef. for timeframe 3

* For each timeframe I obtained the following Odds ratio for logistic regression:
  + Odds ratio for ib is the highest for time frame 1. For every unit increase in ib, the odds of dividends being paid out is increase by 1.00094. Odds ratio for ib is the lowest at timeframe 3 at 1.00032.



Fig 57. Odds ratio for timeframe 1



Fig 58. Odds ratio for timeframe 2



Fig 59. Odds ratio for timeframe 3

* For each time frame I obtained the following Decision Trees in Fig 60 – Fig 62.
  + I observed the following that I marked in red:
    - For splits that involve ib, similar to that in 4d, splits that result in lower ib tend to have a higher percentage of observations with no dividend payouts (split left) and splits that result in higher ib have a significantly lower percentage of observations with no dividend payouts.
    - This is consistent throughout all 3 timeframes.
  + Thus, we can draw the same conclusion as that from question 4d. Thus, **the higher the earnings, the more likely for dividends** to be given out in general.

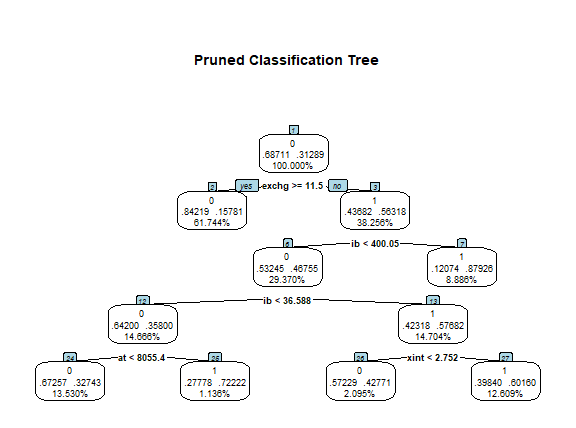


Fig 60. Pruned Tree for timeframe 1

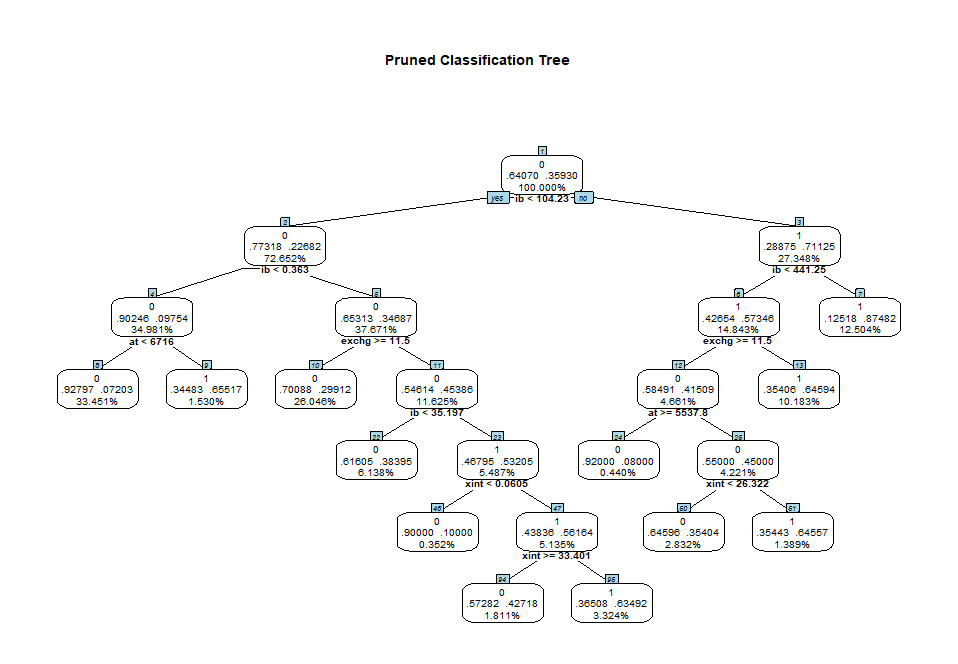


Fig. 61. Pruned Tree for timeframe 2

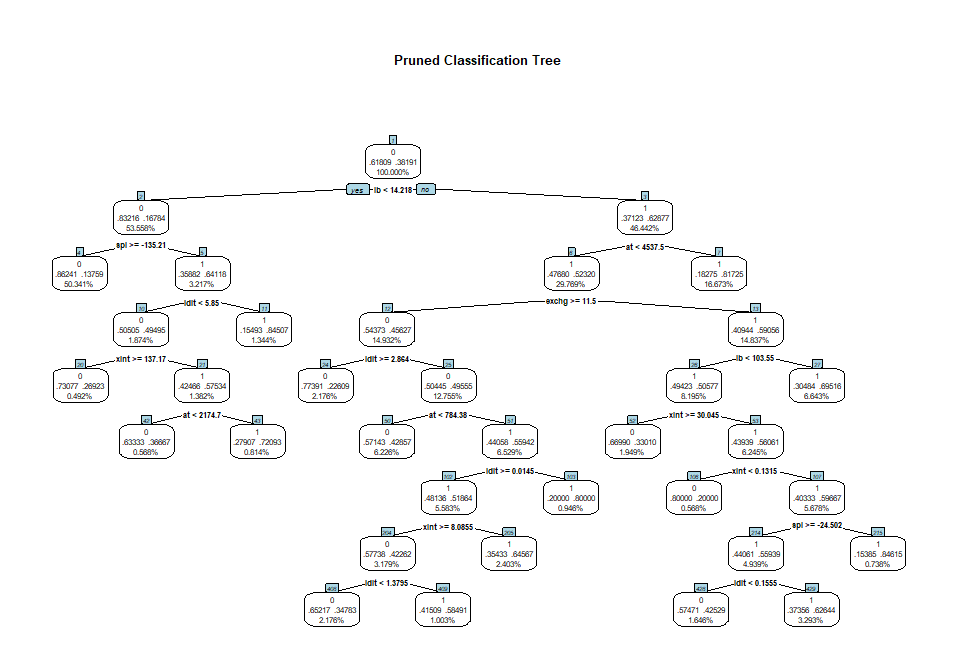


Fig. 62. Pruned Tree for timeframe 3

1. **Question 5**

**Business Hypothesis**

Companies are reluctant to cut dividends if losses are temporary. That is, the relationship between reporting losses and paying dividends might be less strong if the losses are attributable to special items.

**Assumption:** As I lack understanding in what special items really are, for the purpose of testing the hypothesis, I assume that losses are attributable to special items when both ib and spi are less than 0. However, this might not be true in an actual scenario.

Before testing the hypothesis, I wanted to explore the following:

* No. of rows from data that has ib<0
* No. of rows from data that has spi<0
* No. of rows from data with both ib and spi <0
* Percentage of observations that has 0 dividends if ib and spi < 0
* Percentage of observations with 0 dividends if ib <0.
* A higher percentage of observations have no dividend payouts if income is less than 0 (losses). When special items (spi) is negative the percentage of companies not giving payouts decreases from 88.02% to 82.9%.
* I will perform linear regression to further explore this phenomenon.

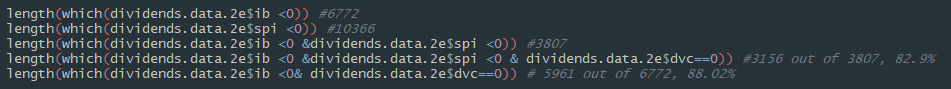


Fig 63. Exploration of data

For this question, I further cleaned the data since we are only concerned about observations with losses.

* For the first model, I cleaned the data to have only observations with income less than 0.
* For the second model, I further cleaned the data to also only include observations with spi less than 0.

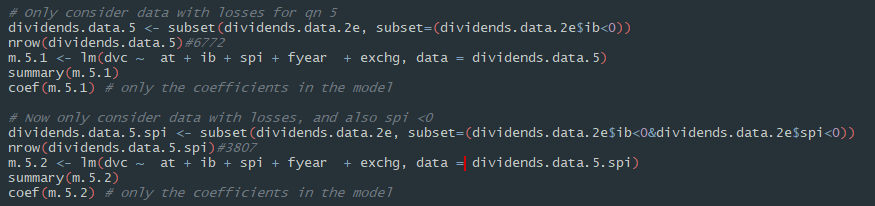


Fig 64. Linear Regression for Question 5

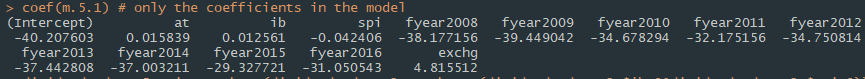
I obtained the following coefficients for the above models. 

Fig 65. Coef. for model with ib < 0

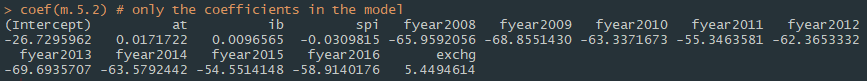


Fig 65. Coef. for model with ib and spi < 0

* The model with only losses has coefficient of 0.012561 for ib, which is higher than the model with negative spi, which is 0.0096565.
* For every unit of loss, dividends would be cut by 0.012561 in the 1st model.
* For every unit of loss, dividends would be cut by 0.0096565 in the 2nd model.
* Thus, based on the assumption that observations with ib and spi < 0 means that losses are attributable to special items, I can conclude that the above business **hypothesis is true.**
* ***If we only consider losses that are attributable to special items, a unit decrease in ib (loss) will result in a smaller cut in dividends.***

1. **Question 6**

**Does Q4 fully address the main question: Does paying dividends still provide information about earnings prospect? Explain. If it does not fully address the main question, explain what needs to be done.**

In Question 4, we used linear regression, logistic regression and CART to find out if dividends pay-out policy is affected by earnings. In those methods, we are using earnings as one of the independent variables X, and find out its effect on the dependent variable Y (dividend pay-out).

The main question asks if dividends still provide information about earnings prospects, and it is phrased in such a way that implies that dividends is the dependent variable X used to predict earnings prospects variable Y. This is different from what we have done in question 4. Thus question 4 does not address the main question.

The introduction also mentions that stock repurchase is becoming a popular pay-out method to shareholders. **Thus, to fully address the main question, we can use both dividends and stock repurchase as dependent variables to find out their effect on earnings prospect (income) of the firm.**

1. **Question 7**

**Recommendation:** Selecting the simplest tree that is within 1 standard error of the lowest testset cross-validation error. 1SE rule.

* In question 4d, I performed pruning to reduce the size of the tree. I pruned the tree based one the optimal cp with lowest cross val. error. The optimal cp is 0.0020642.
* Using the above recommendation, the lowest testset cross validation error is 0.641, and the standard error (xstd) for that row is 0.00873. I would select the simplest tree within 0.641 + 0.0083.

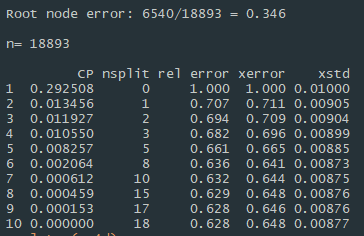


Fig 66. printcp() from question

**Any difference?**

* In this case, the resulting tree would be the same. And the resulting tree would have 8 splits.
* However, it is noted that this using the 1SE rule, it is not always the case where the row with the lowest cross-validation error (xerror) is selected. In the presence of a higher standard error (xstd), there is a possibility that smaller trees will be pruned.

**Value of such a recommendation?**

* This value of this recommendation would be the ability to obtain the simplest possible tree. This could be very useful for cases when the tree is still too large after pruning.