# 1. (i)

Multiple Linear Regression Model is:

，

And if written in the form of matrix,

,,,

It can be represented as:



The definition of notation is:

(1)Y is a column vector of n dependent of response variable.

(2)X is an n by (p+1) matrix of n observations on each of the p explanatory variables.

(3) is a (p + 1) column vector of unknown parameters .

(4)ε is an n column vector of random additive “errors” .

assumptions for Multiple Linear Regression Model:

(1)Relationship between response and explanatory variables is linear.

(2)All  are independent.

(3)Each  has normal distribution.

(4)Each  has constant variance.

(5)n should be larger than p.

1.(ii)

Maximum likelihood estimations-MLE

It is a statistical method based on the maximum likelihood principle. The intuitive idea of themaximum likelihood principle is that if a randomized trial has possible outcomes like A, B, C..., if the result A appears in an experiment, it can be considered that the experimental conditions are

favorable to the appearance of A.

The probability of event A occurring is associated with an unknown parameter, the value is different, the probability of the event A occurs is different also, when we are in A test event A happens, argues that this value should be to maximize in all possible values of t that A maximum likelihood estimation method is to select the t value as A parameter to estimate of t, make the selected samples in the selected overall possibility for most.

1. (iii)

In most cases, we can get MLE by calculating



Afterwards, we should better compute the negative of the second order derivative



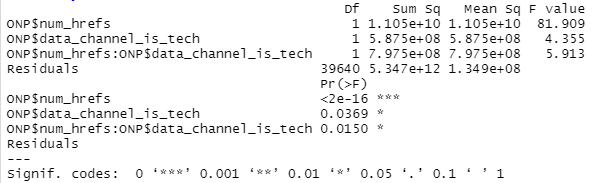
If all eigenvalues are positive, it will give the global maximum of the likelihood.

1. (iv)

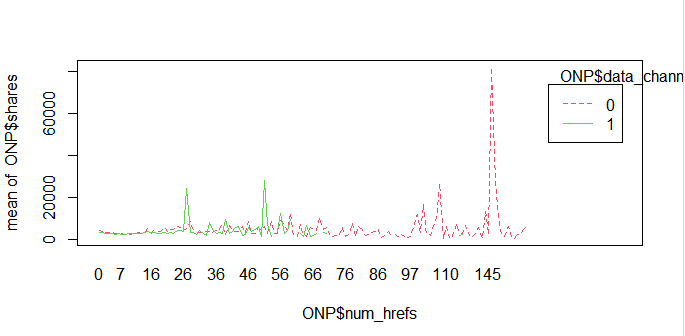
In this dataset, I just use a continuous variable - num\_hrefs and a categorical variable – data\_channel\_is\_tech to explain the interaction between a continuous variable and a binary categorical variable.

As it can be seen from above, the linear model can be written as





In the line chart shown below, we could found that the dot line is not parallel to the solid one, which means they do interact with each other.



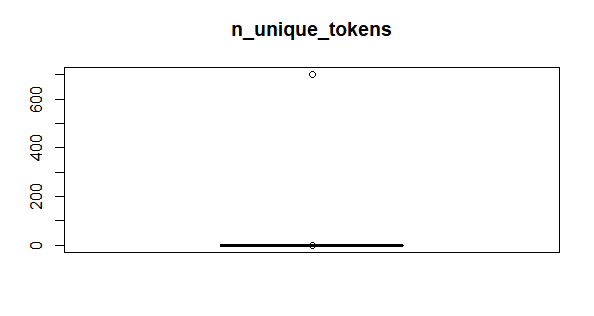
1. (i)
2. There are numerical data,character data and categorical data in this dataset.
3. Interaction exists between some variables.
4. There are some errors and outliers in this dataset.

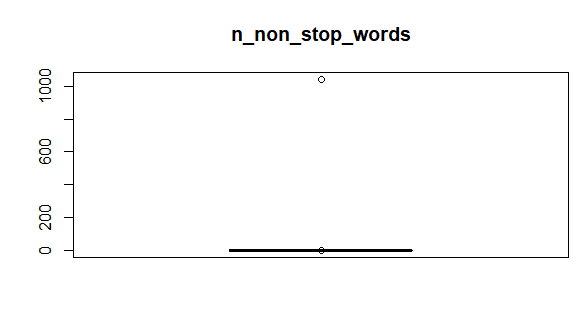
After using R to summary this dataset, I found some abnormal data. Thus, I draw boxplots to observe them better.

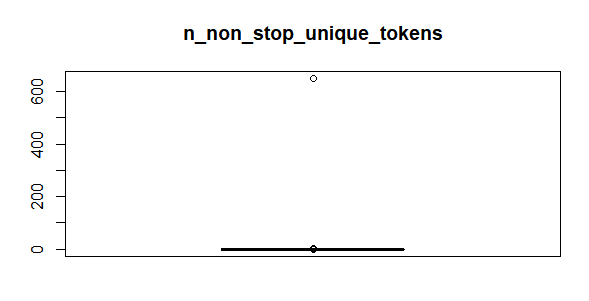
From the picture of boxplot listed below, we could find that the rate of unique words are almost between 0 an 1, and it also should in this range. But there is a outlier which is much higher than others, which value is 701.

Therefore, I think it might be a recording error.

Besides, the same problem is also existed on n\_non\_stop\_unique\_tokens and n\_non\_stop\_words.

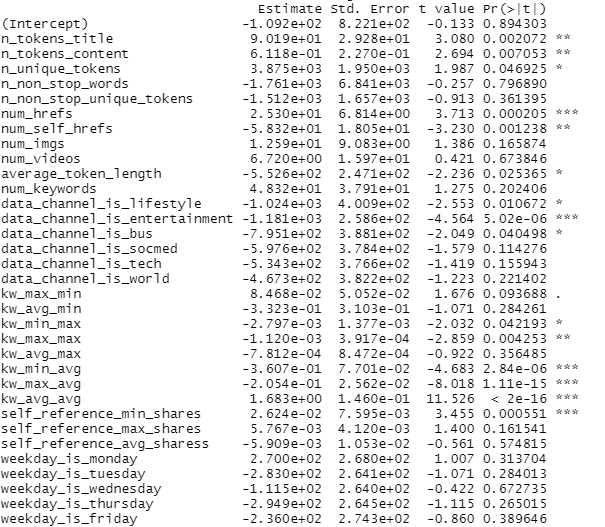




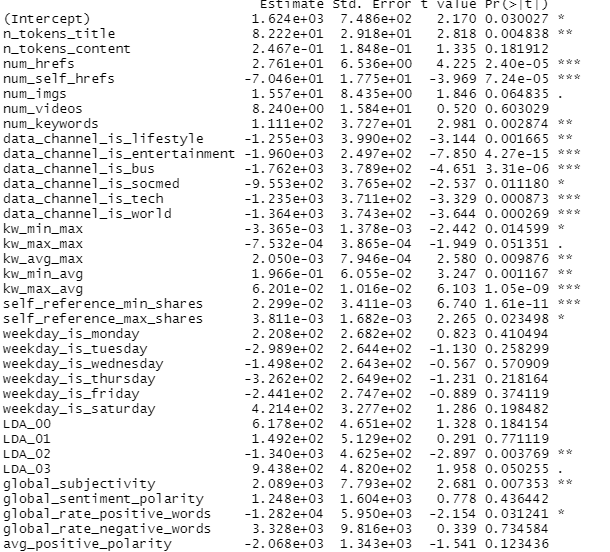


Therefore, I delete the outlier in n\_non\_unique\_tokens. After deleting it, the value of the other two attributes have been changed into (0,1), which means that this outlier is an abnormal value in these three attributes.

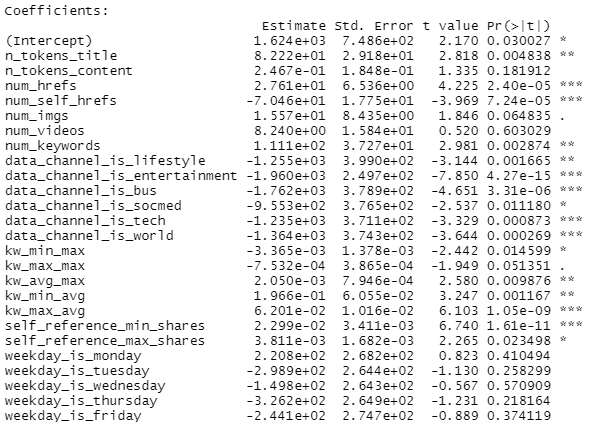
And then I also delete the unusable data and delete negative values. And then I try the first construction. But I found that there are lots of p-values are very large.

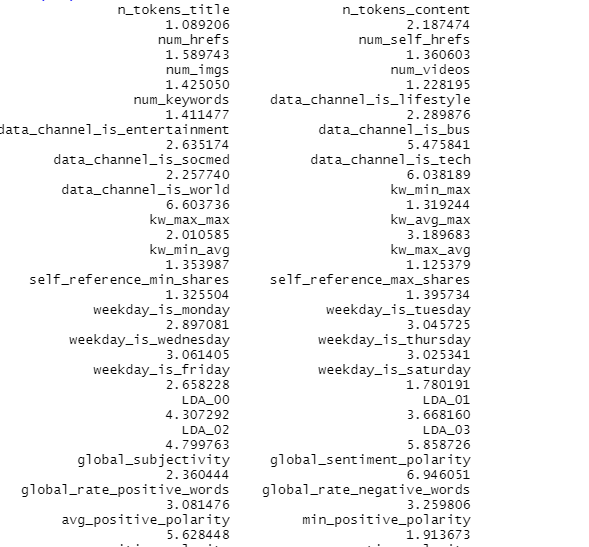


Therefore, I delete variables of "weekday\_is\_sunday","is\_weekend" and "LDA\_04" and try the second construction. But there are still some very large p-values.



And then I use VIF test to check the variance inflation factor and delete all the variables which their variance inflation factor is larger than 10 and reconstruct the model. Now there are lots of small p-value which means almost all variables are significantly related to respond variables VIF of and all of them are smaller than 10.



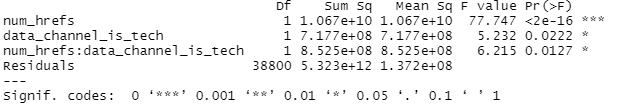


2.(ii)

One reasonable use of an interaction pair is (num\_hrefs,data\_channel\_is\_tech), where data\_channel\_is\_tech is a binary variable indicates whether the post belongs to the ‘Tech’ channel and num\_hrefs means Number of links.

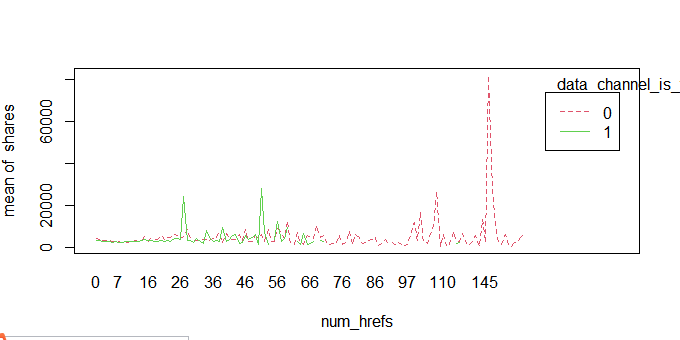
As it can be seen from above, the linear model can be written as



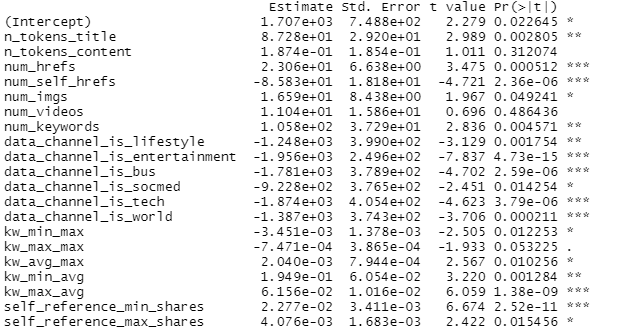


As the picture listed above, the p-value of the interaction term  is significant which equals to 0.0127. Therefore, we should reject the null hypothesis that there the interaction term has no effect to the linear model. According to the lecture PowerPoint on interaction terms, when the interaction effect is significant, then the main effect may not be meaningful. Therefore, we should have the interaction term in the linear model.

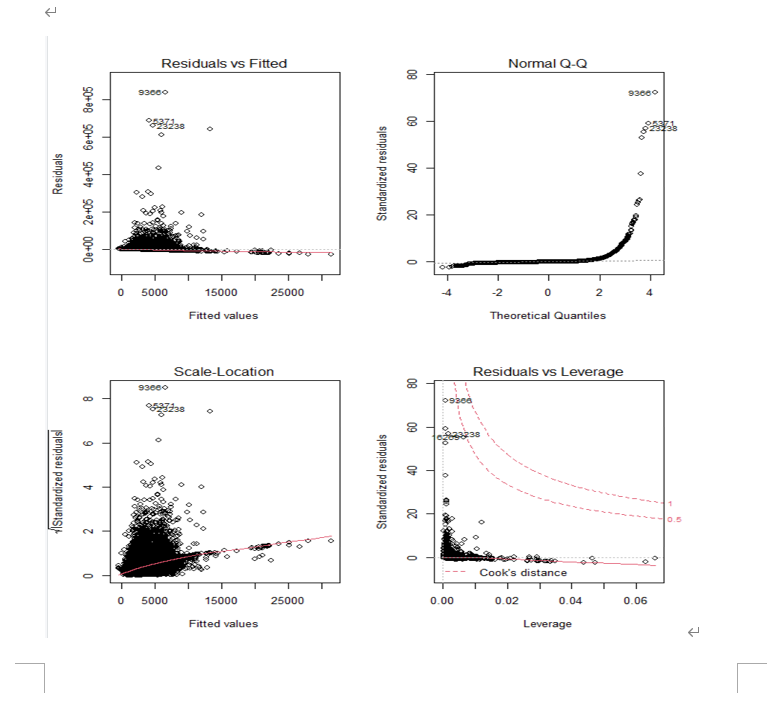
I also use aov( ) function in r to test the interaction which made me found that they are really interact with each other.



Use this interaction term to fit a new linear model with other exploratory variables against the response variable shares. From the picture of summary of current fitted model, we could found that the p-value of interaction term equals to 9.25e-05 which is less than 0.05. This is a significant evidence that we should reject the null hypothesis that the interaction has no effect to the linear model.



2.(iii)



1. Linearity

From the Residuals vs Fitted plot, we could found that the red line is almost flat at zero value of residuals and a large number of points of fitted values are nearly close and some even match the red line, which means the data fit the linearity assumption.

1. Normal distribution

From the Q-Q plot and Histogram of residuals, we could found that the points are not fitted to the straight line and the density of residuals is skewed to the right. Therefore, these data are not normally distribution.

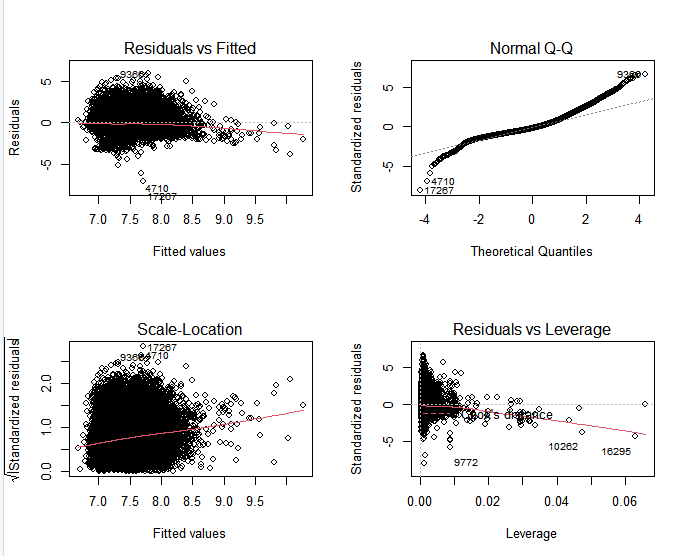
1. Homoscedasticity

From the Scale-Location plot, it could be found out that the red line is positive, which means that the value of the square root of standardized residuals become larger when fitted values become larger. Therefore, these data are not satisfied the homoscedasticity assumption.

1. High leverage points

From the Residual vs Leverage plot, the top right corner have no points. This means that there are no extreme values.

2.(iv)



(a) Because the series of data are very different, I take the log of them to make them become exponential. Specifically, for the data with the attribute: kw\_min\_min, kw\_max\_min, kw\_avg\_min, kw\_min\_max, kw\_max\_max, kw\_avg\_max, kw\_min\_avg, kw\_max\_avg, kw\_avg\_avg. Because the smallest value of these data is -1, which would generate error if taking log of them directly, I add 1 to all the values of these attributes to make them not become negative value and then I take the log of these adjusted values.

(b)From the pictures listed above, from the Q-Q plot and Histogram of residuals, we could found that the points are almost fitted to the straight line. This means that these data are normally distribution after the adjustment.

1. (i)

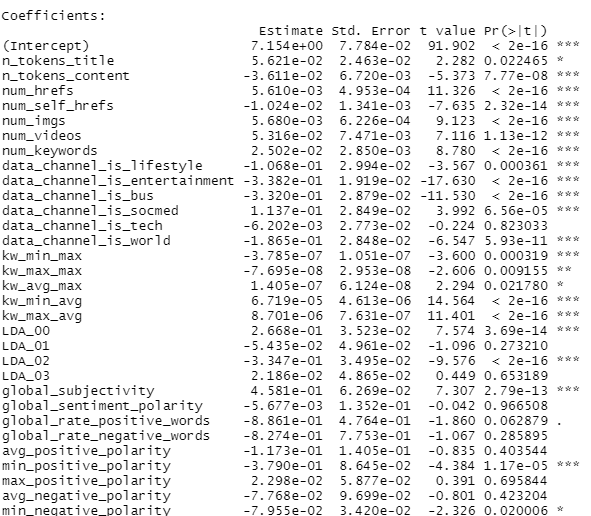
|  |  |  |  |
| --- | --- | --- | --- |
| **Variables can be available before publication** | | | |
| n\_tokens\_title | n\_tokens\_content | n\_unique\_tokens | n\_non\_stop\_words |
| n\_non\_stop\_unique\_tokens | num\_hrefs | num\_self\_hrefs | num\_imgsnum\_videos |
| average\_token\_length | num\_keywords | data\_channel\_is\_lifestyle | data\_channel\_is\_entertainment |
| data\_channel\_is\_world | data\_channel\_is\_bus | kw\_min\_min | data\_channel\_is\_socmed |
| kw\_max\_min | data\_channel\_is\_tech | kw\_avg\_min | kw\_min\_max |
| kw\_max\_max | kw\_avg\_max | kw\_min\_avg | kw\_max\_avg |
| kw\_avg\_avg | LDA\_00 | LDA\_01 | LDA\_02 |
| LDA\_03 | LDA\_04 | global\_subjectivity | global\_sentiment\_polarity |
| global\_rate\_positive\_words | global\_rate\_negative\_words | rate\_positive\_words | rate\_negative\_words |
| avg\_positive\_polarity | min\_positive\_polarity | max\_positive\_polarity | avg\_negative\_polarity |
| min\_negative\_polarity | max\_negative\_polarity | title\_subjectivity | title\_sentiment\_polarity |
| abs\_title\_subjectivity | abs\_title\_sentiment\_polarity |  |  |

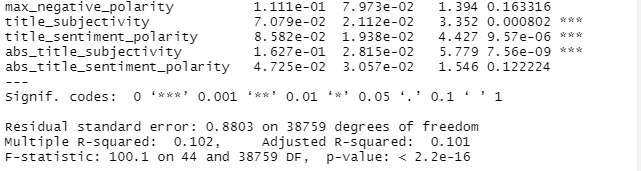
|  |  |  |  |
| --- | --- | --- | --- |
| **Variables cannot be available before publication** | | | |
| weekday\_is\_monday | weekday\_is\_tuesday | weekday\_is\_wednesday | weekday\_is\_thursday |
| weekday\_is\_friday | weekday\_is\_saturday | weekday\_is\_sunday | is\_weekend |
| self\_reference\_min\_shares | self\_reference\_max\_shares | self\_reference\_avg\_sharess | timedelta |

3.(ii)

After calculating the VIF of these variables, some of them whose value of VIF is greater than 10 have been removed to make sure there are not strong colinearity between explanatory variable. And also removing variables with very large p-value to get rid of variables which means that they are not significantly related to the dependent variable, which could also solve the problem of multicollinearity to some extent.

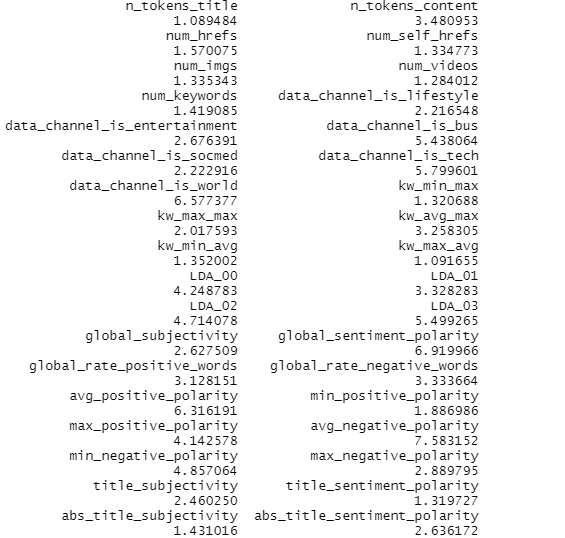
As a result, the remaining estimated model parameters are as follows:





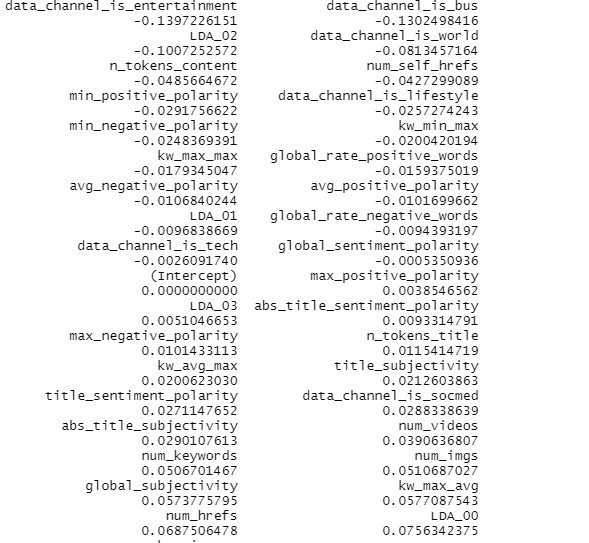
From the table listed above, we could found that the p-value of most explanatory variables are smaller than 0.05. This means that almost all parameters are significantly related to the dependent variable.

Moreover, from the snapshot of the value of VIF of interaction terms listed below, they all smaller than 10,which means that this adjusted model has no problem of multicollinearity.



3.(iii)

By using lm.beta() function in R, we could get a table of the standardized coefficients interval of remaining variables listed below:



We could find that the two most significant slope parameters are data\_channel\_is\_entertainment and data\_channel\_is\_bus,because the abs of their standardized coefficients are larger than others which means they are significantly related to respond variables.

1. (i)

Use the same set of variables as the linear regression model including the interaction term.

Check assumptions of a generalise linear model:

1. Independency

Check duplicates in the dataset and get rid of those repeat data.Manually transform the response variable from numeric to binary through the logit transform Iog(x/(1-x)).We can assume that all observations Y with the given the X are independent and the model structure is appropriate.

1. Multicollinearity

It is corresponding to the situation that the data contain highly correlated predictor variables.Use the vif() function to calculate the variance inflation factor of all exploratory variables in the glm and delete variables with VIF greater than 10.

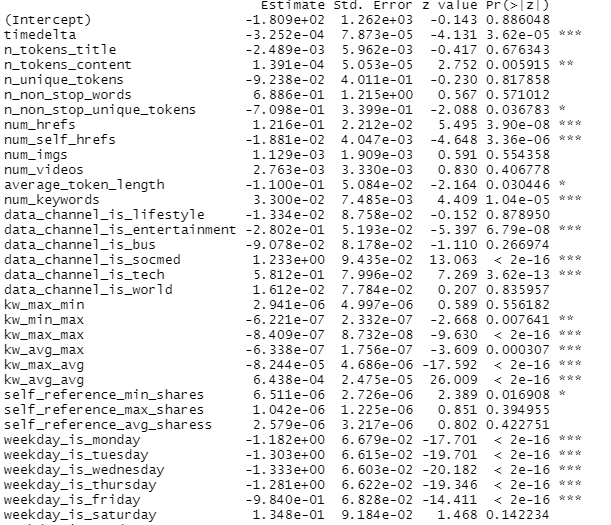
The prediction of the model might not be normally distributed since the plot is not linear.We might need to transform some variables to make it normal.

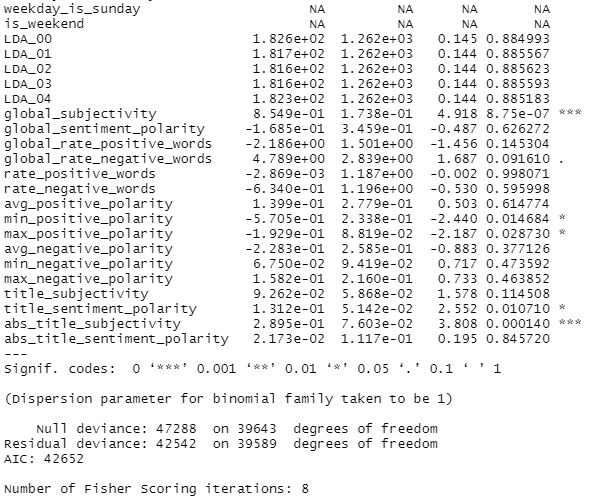
4.(ii)

I still deleted the unusable data (e.g url). And I also delete negative value (-1) of kw\_min\_min, kw\_avg\_min, kw\_min\_avg, because these three variables are discrete variables and negative value of discrete variables are always considered as missing data because they have no negative value at all. Besides, I also take log of some data as well to calculate them better.

4.(iii)

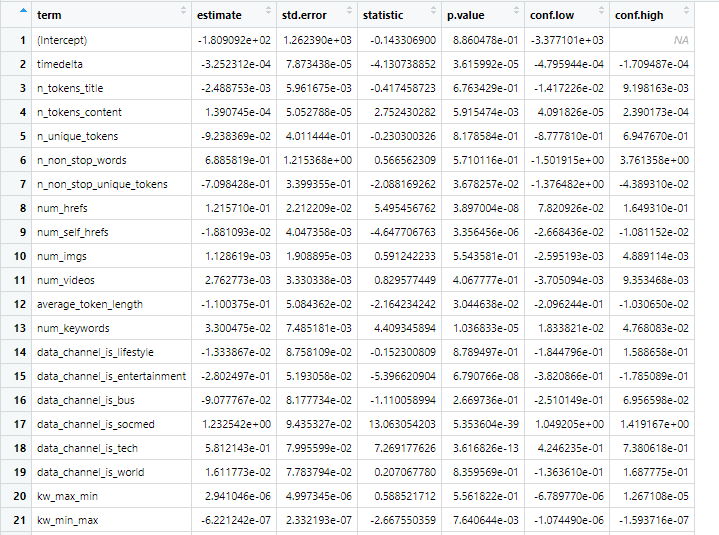
The table of all the estimated model parameters is listed below:

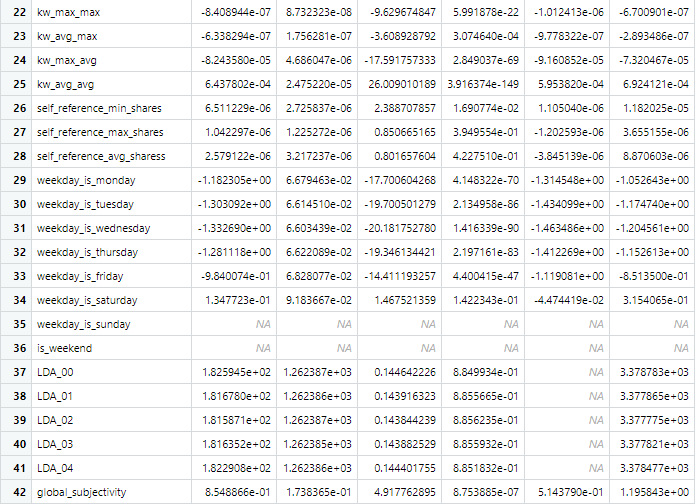


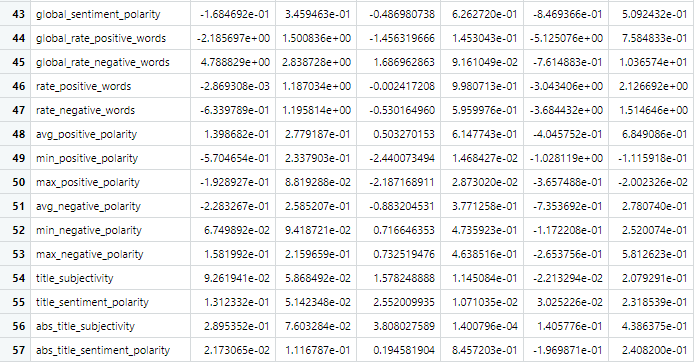


4.(iv)

By using tidy() function in R, I got the table listed below:



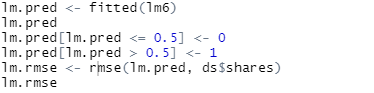




We could find that the two most significant slope parameters are num\_keywords and self\_reference\_min\_share,because the p-value of them are much smaller than others which means they are significantly related to respond variables.

5.

1. Multiple Linear Regression Model:

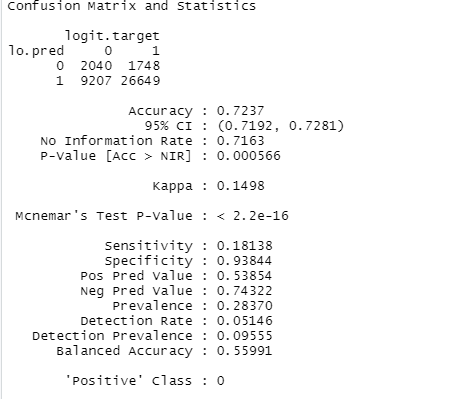




The RMSE is the square root of the square root of the deviation between the observed value and the truth value and the ratio of the observed number m. Thus, it’s always used to measure the deviation between the observed value and the truth value. It is more sensitive to outliers.

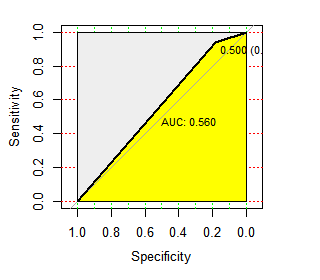
From the snapshot listed above, we could find that the value of RMSE is 6.548, which means that the deviation between the observed value and the truth value is very large in this model. It might be caused by some predicted values which are significantly different from the true value.

1. Logisitic Regression



A confusion matrix is a table that is often used to describe the performance of a classification model (or "classifier") on a set of test data for which the true values are known.

From the picture of confusion matrix listed above, we could find that the accuracy of logistic model is 0.7237, which means the accuracy of this model for prediction is 72.37%.



ROC curve, is a graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied. In general the minimum value of 0.5 we think the AUC (of course there are contrary to the extreme situation, AUC is less than 0.5, this situation is equivalent to the classifier is always the right as wrong, wrong is right, so as long as take the predicted class, then got a classifier of AUC is greater than 0.5) and we hope to achieve the effect of the classifier is: for real category 1 sample, classifier to predict the probability of 1 (namely TPRate), than real category of 0 and predict the probability of category 1 (namely FPRate), namely y>x.

From the picture of ROC curve listed above, we could find that the value of AUC is 0.56 which is larger than standard value 0.5. This means that the accuracy of prediction which is predicted by this model is high.

6.

As far as I am concerned, Logisitic Regression model is better. Because the prediction accuracy is very high (more than 0.7). It is great because we use the prediction model, the probability of it to get the wrong prediction is less than 30%. Besides, the the value of AUC is 0.56, which means that the accuracy of logistic regression model for prediction is 56%. Therefore, from reasons listed above, the accuracy of logistic regression model for prediction is above 50%.

However, the value of RMSE of multiple Linear Regression Model is very large. This means that it always get wrong prediction because this value reflect the deviation between the observed value and the truth value is very large.