2020

Assignment - 1

DATA 7202

KIRTI KHADE || SID - 45733130

TABLE OF CONTENT

Question - 1

1(i)

Give relevant equations and assumptions to describe the general linear model and the logistic regression model for multivariate data (assume X has dimension p and Y is a single variable in each case). Define any notation used.

General Linear Model(Multivariate data):

- Equations:

$$Y_i = \beta_o + \sum_{j=1}^{p} \beta_j * x_{ij} + \in_i$$

$$or$$

$$Y = X * \beta + \in$$

Where, Y_i is the dependent or the response variable ; $x_{ij}$is the explanatory variable ; β_j is coefficient ; β_o is the intercept ;and $\in_i$ is the random additive error term.
In matrix format we can write this as:

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ . \\ . \\ Y_n \end{pmatrix} \quad X = \begin{pmatrix} 1 \dots x_1^T \\ 1 \dots x_2^T \\ . \\ . \\ 1 \dots x_n^T \end{pmatrix} \quad \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ . \\ . \\ \beta_n \end{pmatrix} \quad \text{and} \quad \in = \begin{pmatrix} \in_1 \\ \in_2 \\ . \\ . \\ \in_n \end{pmatrix}$$

- Assumptions:
  - The relationship between each explanatory variable and the response variable is approximately a straight line
  - No highly influential outliers
  - Residuals are approximately normally distributed
  - Residuals have approximately constant variance for any combination of x
  - Residuals are independent of each other
  - Observations are independent of each other
  - Number of explanatory variables are at least n/3 times of total sample size

Logistic Regression:

- Equations:

$$\log\left(\frac{p_i}{1-p_i}\right) = \sum_{j=1}^{p} \beta_j * x_{ij} + \in_i$$

Where, Y_i is binary with $p_i = P(Y_i = 1 \mid X_i)$ and $1 - p_i = P(Y_i = 0 \mid X_i)$

$\log\left(\frac{p_i}{1-p_i}\right)$ is known as log odds ; $X_i\ is\ the\ vetor\ of\ all\ p\ variables$;rest of the variables are as above.

- Logistic Regression Assumptions:
  - Dependent variable should be binary or ordinal classes
  - Independent variable independent of each other
  - Independent variables should not be multicollinear
  - The relationship between each explanatory variable and the log odd of the response variable is approximately a straight line
  - No highly influential outliers
  - Number of explanatory variables are at least n/3 times of total sample size

1(ii)

Briefly describe the algorithms used to fit each of these models and their mathematical basis.

General Linear Model(Multivariate data):

1) The equation for General linear regression is represented above
2) For this model, we need to find $\in_i$ and β_i
3) $\in_i$ is normally distributed and has a constant variance . Therefore, we can say that $\in_i \sim N(0, \sigma^2 * I_n)$, where I_n is an identity matrix of n * n dimension.
   Hence, $Y = X * \beta + \in$ can be written as $Y \sim N(X\beta, \sigma^2 * I_n)$
4) For the equation : $Y = \beta * X + \in$, we need to estimate the β, which can be done by using (a) maximum likelihood estimation or (b) ordinary least squares
5) Maximum Likelihood estimations: This method is used for estimating the probability for a sample of observation.
   We have,
   MLE of $\hat{\beta}$ = $(X^T * X)^{-1} (X^T * Y)$
   MLE of $\widehat{\sigma^2}$ = 1/ n * $(Y - X * \hat{\beta})^T (Y - X * \hat{\beta})$
6) Ordinary least square: This method minimises the sum of squared of the differences between Y_i and their expected value. This method, minimises,
   $$\sum_{i=1}^{n} (Y_i - X_i^T * \beta)^2 = (Y - X * \beta)^T (Y - X * \beta)$$
   To get the optimal solution of
   $$\hat{\beta} = (X^T * X)^{-1} (X^T * Y)$$
7) Both of the methods assume, normal error and constance variance.
8) The interpretation of linear regression is as follows: With an increase of X_i by 1, the Y will approximately increase by β_i

Logistic Regression:

1) The equation for logistic regression is represented above. Logistic Regression has binary variables
2) The logistic function can be seen as a linear model, if we replace Y in linear regression with the log odds defined above.

3) β , $\widehat{\sigma^2}$ can be calculated in the same manner as above. Considering Y = $\log\left(\frac{p_i}{1-p_i}\right)$
4) The interpretation of logistic regression is as follows: With an increase of X_i by 1, the value of log odd will approximately increase by β_i. And the value of probability will approximately increase by exp(β_i)/ (1 + exp(β_i))

1(iii)

Derive the maximum likelihood estimate of σ^2 for multiple linear regression (= the general linear model). List any assumptions.

We have log of MLE as,

$$\ln(L(\mu,\sigma^2)) = -\frac{n}{2} * \ln(2\Pi) - \frac{n}{2} * \ln(2 * \sigma^2) - \frac{1}{2 * \sigma^2} * \sum_{i=1}^{n}(X_i - \mu)^2$$

Differentiation w.r.t σ^2 , we get;

$$\frac{\partial}{\partial \sigma^2}\left(-\frac{n}{2} * \ln(2\Pi)\right) = 0 \ (As, it\ is\ a\ constant\ term)$$

$$\frac{\partial}{\partial \sigma^2}\left(-\frac{n}{2} * \ln(2 * \sigma^2)\right) = -\frac{n}{2} * \frac{1}{2 * \sigma^2} * 2 = -\frac{n}{2} * \frac{1}{\sigma^2}$$

$$\frac{\partial}{\partial \sigma^2}\left(\frac{1}{2 * \sigma^2} * \sum_{i=1}^{n}(X_i - \mu)^2\right) = \frac{1}{2} * \frac{1}{2 * \sigma^4} * \sum_{i=1}^{n}(X_i - \mu)^2$$

Combining the three equations above we have,

$$\frac{\partial}{\partial \sigma^2} L = -\frac{n}{2} * \frac{1}{\sigma^2} + \frac{1}{4 * \sigma^4} * \sum_{i=1}^{n}(X_i - \mu)^2$$

Assumptions:

- The Log maximum likelihood functions ($\ln L(\mu,\sigma^2)$) should be continuous/consistent
- $\widehat{\sigma^2}$ is biased, the unbiased can be obtained via

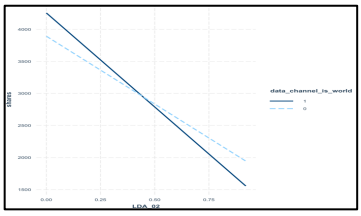
$$S^2 = \sigma^2 = \frac{1}{n-p} * \left(Y - X * \hat{\beta}\right) * \left(Y - X * \hat{\beta}\right)^T$$

1(iv)

In lectures, interactions have been described between two continuous variables and between two categorical variables. Explain the interaction between a continuous variable and a binary categorical variable. Give an example, including equations and a plot. Also attempt to find evidence of one such interaction in this dataset - report your evidence and conclusions

Interaction between Continuous and Binary Categorical Variable

- An interaction variable exists between two explanatory variables. If effect of one explanatory variable on response variable changes due to the value of one or more explanatory variable, this effect is called interaction variable. Adding interaction variable in the regression model, would help in understanding the relationship of variables and will also help in increasing the accuracy.
- For a binary categorical variable, having value 0 and 1, an interaction between the line graphs of the explanatory variable would mean that the continuous explanatory variable is different for different levels of categorical variable.
- One example of the continuous and binary categorical variable can be seen as the followings: Continuous variable - LDA_02 ; Categorical variable - data_channel_is_world. Since the lines are "interacting" at a point, we can say that the variables are interaction variable.

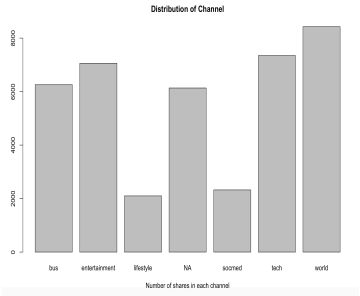


Question - 2

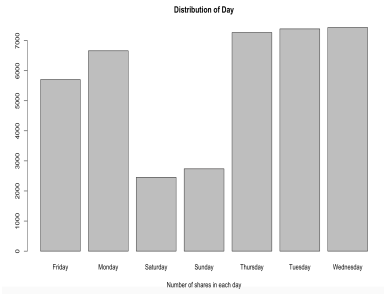
2(i)

Perform exploratory data analysis as relevant to the construction of the regression models. Investigate and highlight any apparent structure in the data.

- **EDA on Categorical Variables:**
  - **Distribution of channels**: Highest number in "world", followed by "tech" and "entertainment". Some observations have no channel. None of the items had multiple channels



  - **Distribution of days**: Highest number of shares is on Thursday, Tuesday and Wednesday. Lowest share is in Sunday and Saturday.

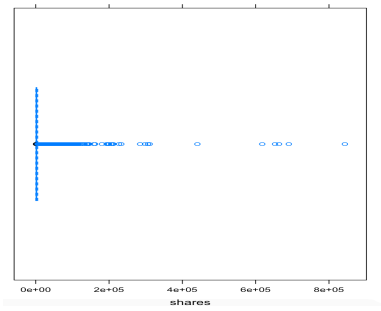


- Among all the categorical variables, we can **delete** the following before analysis because they are redundant:
  - is_weekend: As, we have variables for both the weekends (Saturday and Sunday)
  - weekday_is_sunday: As, we have variables for all weekdays, and there is no data point with no day. So, we will be removing this for analysis.
  - All the channel will be considered as is, because there are observations that are not on any channel.

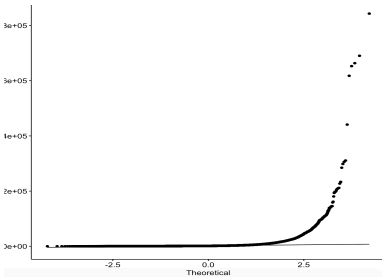
- **The most correlated(absolute) columns are as follows:**

Feature - 1	Feature - 2	Absolute Correlation
n_non_stop_words	n_non_stop_unique_tokens	1
n_unique_tokens	n_non_stop_words	1
n_unique_tokens	n_non_stop_unique_tokens	1
kw_avg_min	kw_max_min	0.941
kw_max_max	kw_min_min	0.857
self_reference_avg_sharess	self_reference_max_shares	0.853

- **EDA on "shares":**
  - Box plot : There are many outliers in "shares" column



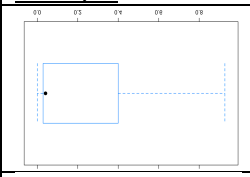
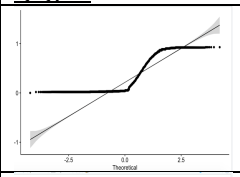
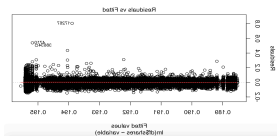
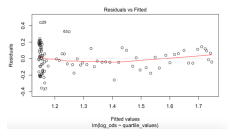
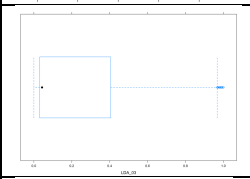
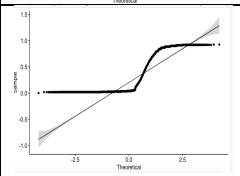
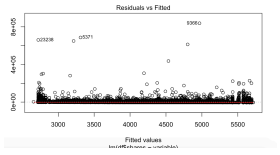
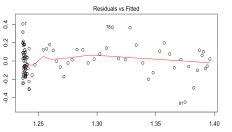
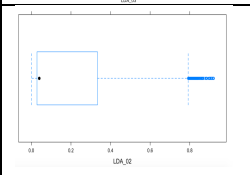
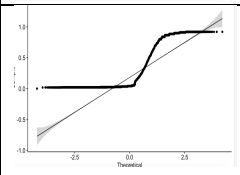
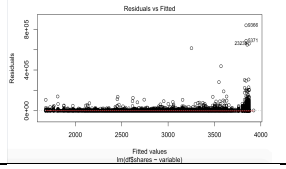
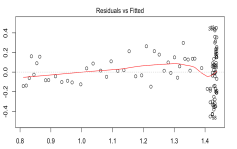
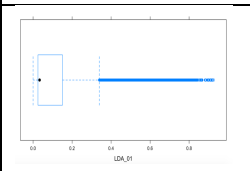
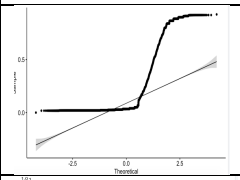
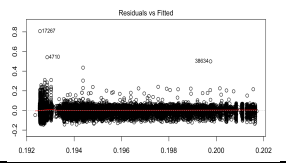
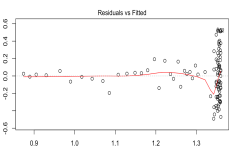
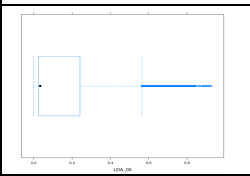
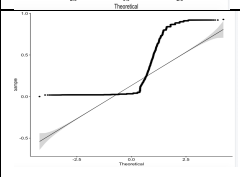
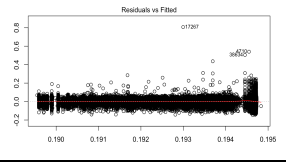
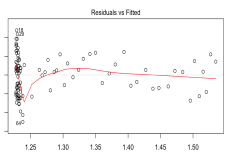
  - Q- Q plot on "shares": The Q-Q plot does not form a straight line, hence "shares" is not normally distributed



- **<u>EDA on the continues variable:</u>**
  - Outlier analysis:
    - We will use **Box plot** to see if the variable has many outliers
    - However, Box plot will not show us if the outliers are "*influential*". We will be analysing these outliers during linear regression, to make sure that there are no influential outliers.
    - I have also analysed each variable using the "describe", "max", "min" functions. I have transformed the variable if :
      - The was values less than 0, and if it logically does not make sense for the variable, I have caped the value at 0
      - One of two values were extremely high – I have caped the value with 95% of the variable.
    - Additionally, I have also checked for influential outliers using "Residual vs Leverage" plot from linear regression between the response and the explanatory variable – None of the variables had any " influential" outliers.
  - Normal Distribution:
    - We will use Q-Q plot value to see if the variable is normally distributed
  - Linearity analysis:
    - After some pre-liminary analysis, I deducted that share to the power of "-0.22" is transformation required to get a linear -relationship between the dependent and the independent variables
    - **Residuals vs Fitted :**
      - This plot shows if residuals have non-linear patterns. If the pattern is non-linear, it would mean that the explanatory and the response variable has non-linear relationship
      - There could be a non-linear relationship between predictor variables and an outcome variable, and the pattern could show up in this plot if the model doesn't capture the non-linear relationship.
      - If the residuals are spread equally, both side of residual plot then it indicates that there is no non-linear relationship between the two
  - Logistic analysis:
    - For logistic analysis, we need to make sure that "log odds" are linearly dependent with every variable.
      - For that, firstly we converted the response variable into binomial (if share > 1000 then 1 or else 0)
      - We divide the data into 100 or 1000 equal parts using the variable in study
      - We compare the linearly dependency(as above) between mean of response and mean of variables in the 100 or 1000 parts of data
  - Any transformation needed: In the process, we will check if there is any transformation that may be needed for the variables understudy

- **<u>EDA on LDA variables</u>**
  - LDA columns represents the "closeness" with the particular topic. The values are between 0 – 1

- Analysing variable based on their outlier plot, Q-Q plot and scatter plot below:

Variable	Outlier plot	Q-Q plot	Linearity Test	Log Odd Quartile Test
LDA_04				
LDA_03				
LDA_02				
LDA_01				
LDA_00				

- Outliers: LDA_01 has the greatest number of outliers as seen by box-plot method.
- The Q-Q plot of the variable shows that these variables are not normal.
- The linear dependence of thee variable can be seen “straight line” in the residual vs fitted curve.
- The log odd of the variables is approximately “straight line” in the residual vs fitted curve. It’s not a perfect straight line because there are only 100 or 1000 data points that are analysed and so, the linear regression may be highly influenced by one or two points
    - I also tried getting the exact transformation using box-cox transform. Most of the plots were still not linear, and were generating large number of nan/ infinite numbers

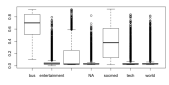
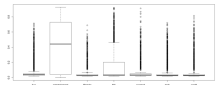
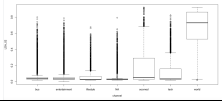
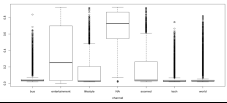
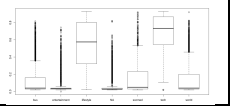
EDA on rest of the variables

- Performed the same analysis as above in the rest of the keywords, however, won’t be including here due to limit in the size of the file. Almost all the variables had huge number of outliers as seen from box plot and most of the Q-Q plots were not normal.
- I am summarising the results from the analysis on each variable here, as including the plots was increasing the size of the file more than 10MB
- To get the linear relationship, for log odd quartile test and linearity test of a variable, I made the following transformation:

Variable	Linearity Test	Log Odd Quartile Test
kw_max_min	log(kw_max_min + 1)	log(kw_max_min + 1)
kw_avg_min	log(kw_avg_min + 2)	log(kw_avg_min + 2)
kw_min_max	log(kw_min_max + 1)	log(kw_min_max + 1)
kw_max_max	log(kw_max_max + 1)	kw_max_max

kw_avg_max	log(kw_avg_max + 1)	kw_avg_max
kw_min_avg[kw_min_avg <0] = 0 kw_min_avg	log(kw_min_avg + 1)	kw_min_avg
kw_max_avg	log(kw_max_avg + 1)	kw_max_avg
kw_avg_avg +	log(kw_avg_avg + 1)	log(kw_avg_avg + 1)
global_sentiment_polarity	global_sentiment_polarity	global_sentiment_polarity ** -2
min_positive_polarity	min_positive_polarity	min_positive_polarity ** 0.06
n_tokens_content	log(n_tokens_content + 1)	log(n_tokens_content + 1)
n_unique_tokens[(n_unique_tokens > 1)] = 1 n_unique_tokens	n_unique_tokens	n_unique_tokens
n_non_stop_words[(n_non_stop_words > 1)] = 1 n_non_stop_words	n_non_stop_words	n_non_stop_words ** -0.10
num_hrefs	log(num_hrefs+1)	log(num_hrefs+1)
num_imgs	log(num_imgs + 1)	log(num_imgs + 1)

- **EDA on continuous vs categorical variables**
  - Channel variables:
    - LDA is a topic modelling method, and the variables (LDA 01, 02 , 03 , 04) that indicates closeness to a topic has a relationship with channel variable

LDA_00	LDA_01	LDA_02	LDA_03	LDA_04
				

    - LDA_00 is a dominated by “bus” and “scomed” channel
    - LDA_01 is a dominated by “entertainment” channel
    - LDA_02 is a dominated by “world” channel
    - LDA_03 is a dominated by “entertainment” and “scomed” channel
    - LDA_04 is a dominated by “lifestyle” and “tech” channel

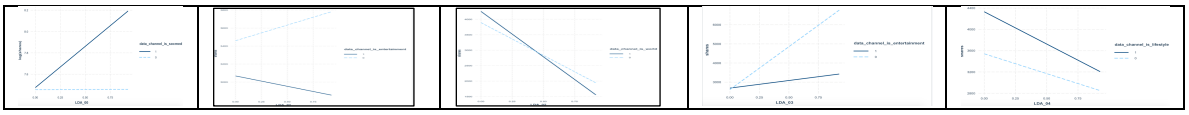
  - No other relevant relationship between the channels/days could be found with any continuous variable

2(ii)

Search for at least one reasonable use of an interaction term between two explanatory variables and include it in your model. Note that the variables can be of any type, but the interaction must be in addition to the one considered in 1(iv) and you must justify why you considered this interaction (relevant graphs or tables) and explain the effect of including it in the model.

- Interactions with respect to a Continuous and a Categorical variable(Note – only LDA_02 was mentioned in 1(iv)):

LDA_00	LDA_01	LDA_02	LDA_03	LDA_04



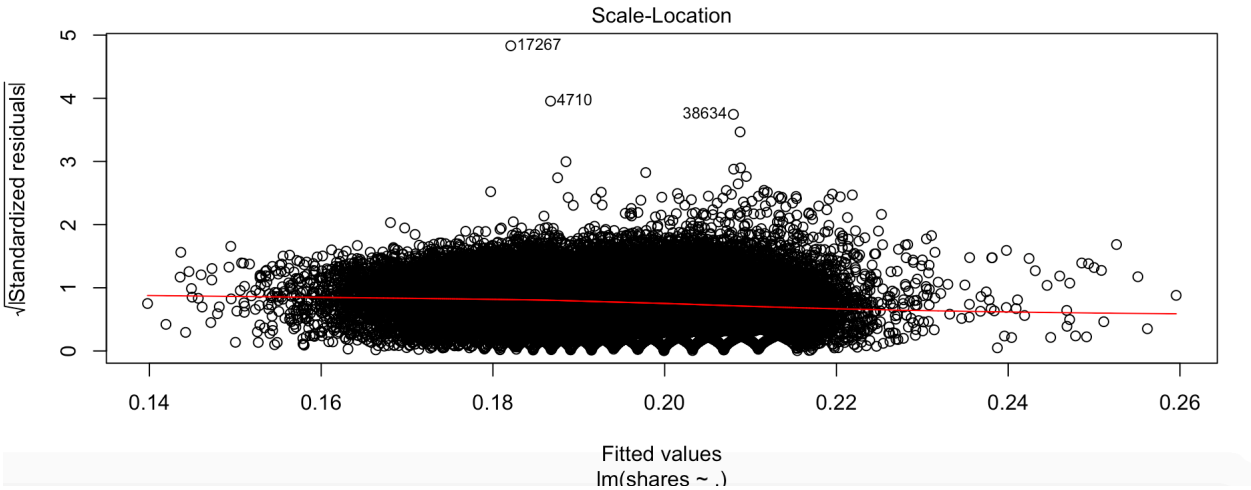
- We can see that the interaction term matches with the results of EDA on continuous vs categorical variable above. This could be because some channel(s) is approximately represented by each LDA as it is a topic modelling algorithm. Because of this there is interaction between these variables
- The general R-squared error without interactions is 16.30% while with interactions (adding all together)is 16.41%

2(iii)

Carefully evaluate whether or not the assumptions of a multiple regression model hold. Consider, provide evidence, act to remedy (if reasonable) and discuss standard assumptions of multiple linear regression, collinearity, any outliers or influential observations. Note: you may have to consider the conditions many times as you adjust the model (including for transformations, as discussed below).

- **NOTE** – For this analysis I am removing all the variables that calculate itself "share" in any way. Firstly, because they are procured after "sharing" or the online news. Secondly, because, we do not need these variables in the further questions.
- **Assumptions**:
  - The relationship between each explanatory variable and the response variable is approximately a straight line – Using the findings in EDA in 2(i), we will be modifying the explanatory variables, in order to get a linear relationship with the response variable. However, we will be verifying linearity w.r.t the model, using the Partial-residual plot
  - Residuals have approximately constant variance for any combination of x – Using Homoscedasticity analysis, we will be analysing this
  - Residuals are approximately normally distributed – Using Normal Q-Q plot we will be analysing this
  - Residuals are independent of each other – We will be using ACF or autocorrelation factor to analyse this
  - Observations are independent of each other – VIF analysis will be used to verify this
  - No highly influential outliers – We will be using cook's distance analysis to understand any influential outliers
  - Number of explanatory variables are at least n/3 times of total sample size – The number of explanatory variables is n/3 times of total sample size

- **Homoscedasticity**. The residuals should have constant variance.
  - Upon doing "Box-cox" transformation, the best value of lambda came out to be "-0.22". Hence, we applied that transformation for the rest of the analysis.
  - **Scale-Location plot**: This plot shows if residuals are spread equally along the ranges of predictors. This helps us in checking the assumption of equal variance.

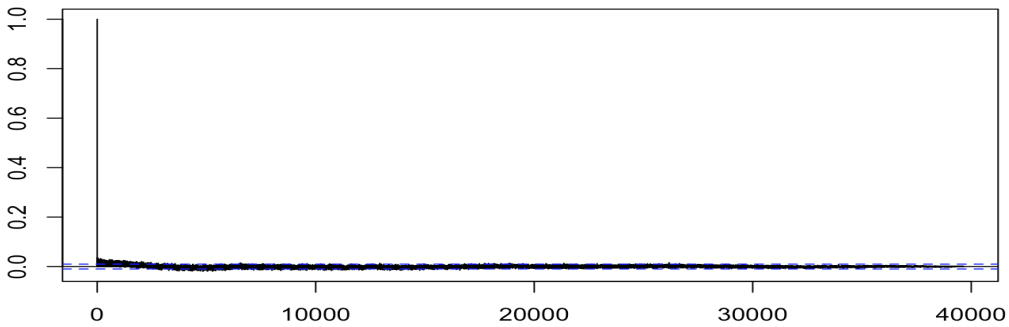


- On analysing the constant variance using the "ncvtest"
  - **BP tests the null assumes homoscedasticity**. So, if p_value > 0.05 we **reject** the null and conclude there may not be heteroscedasticity.
  - The results for ncvtest are shown below:

```
> ncvTest(fit)
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 1.524437, Df = 1, p = 0.21695
```

  - Since p value > 0.05, we can say that there is homoscedasticity in the dataset

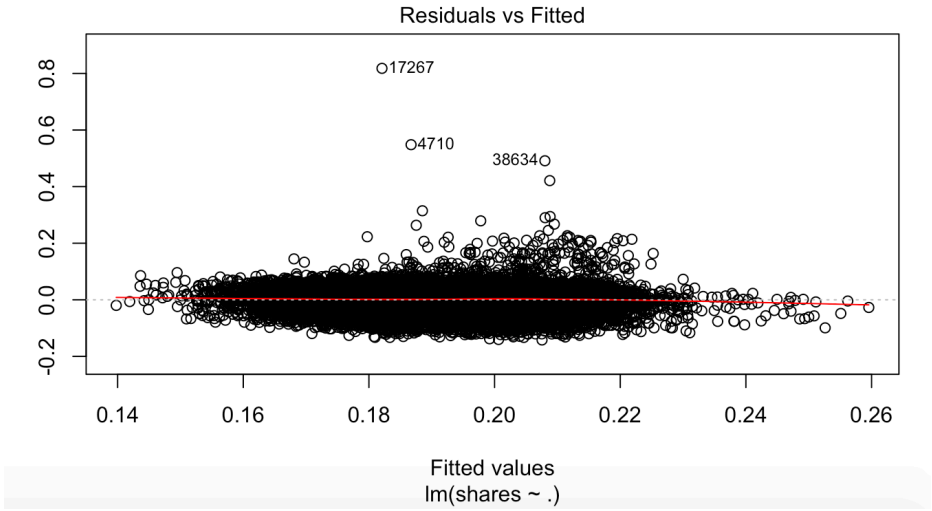
- **Independent Residual.** The residuals should be independent of each other.
  - To check this, we will be using ACF – Autocorrelation function, as it plots the sample autocorrelation up to the lag. From the below figure we can see that the correlation is always below 0.1, hence it's safe to say that no correlation exists between the residulas



- **Independence of variable**. The explanatory variables should be independent of each other
  - Using Variance Inflation Factor, I was able to remove the variables with high VIF, one by one. I stopped when the highest VIF was less than 10. The following variables were removed using this process:
    - LDA_03
    - rate_positive_words
    - n_non_stop_words
    - kw_min_max

- kw_avg_max
- kw_max_min
- kw_avg_avg

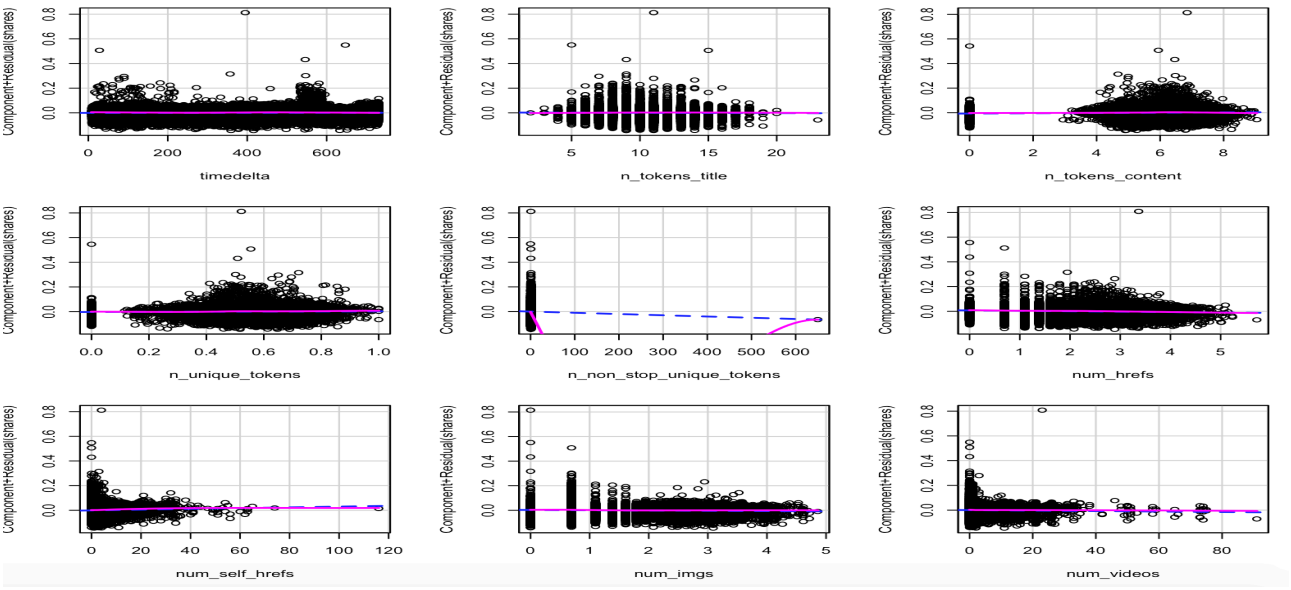
- **Linearity**. In other words, the model should capture all the systematic variance present in the data, leaving nothing but random noise.

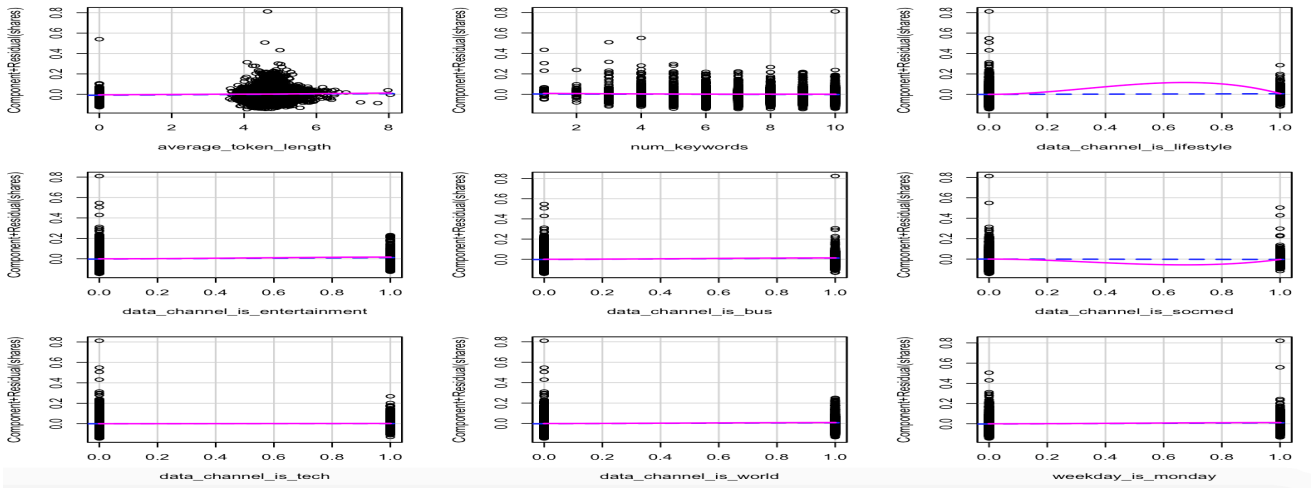


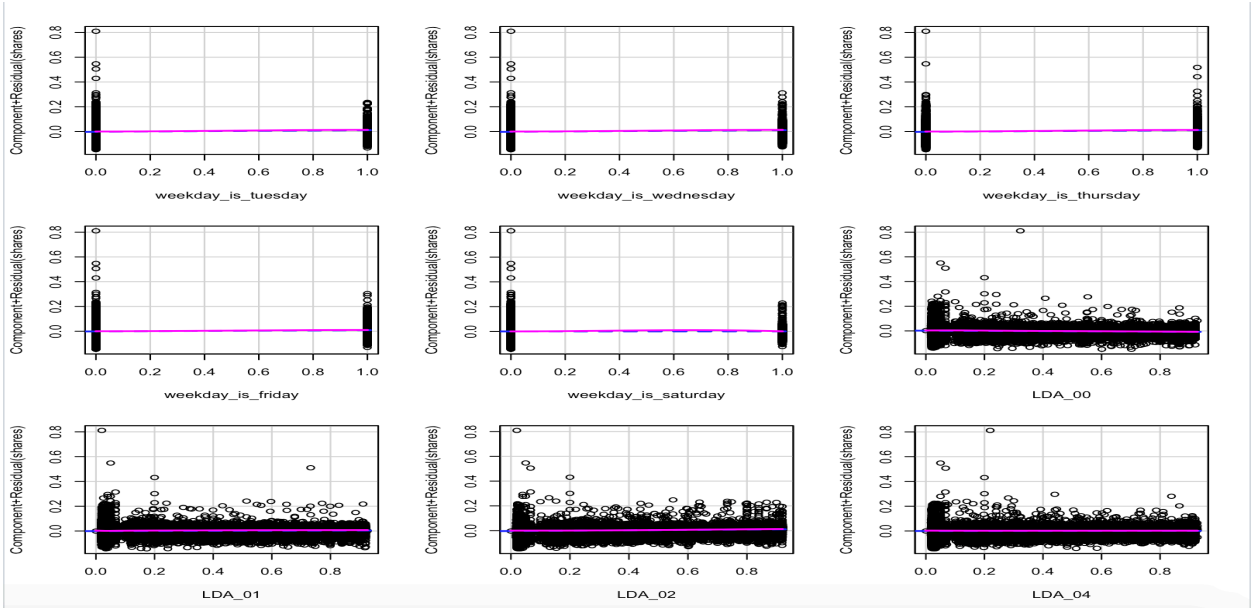
- "Residual vs fitted" graph is a straight line, with equal number of variance both side. Hence, we can say that the relationship between the independent and dependent variables is linear
- We will also analyse the "Partial-residual plots" or the "Component + residual plots"
  - These plots show the relationship between given independent variable and the response variable, given that all the other independent variable are in the model. These plots are formed as (wiki):

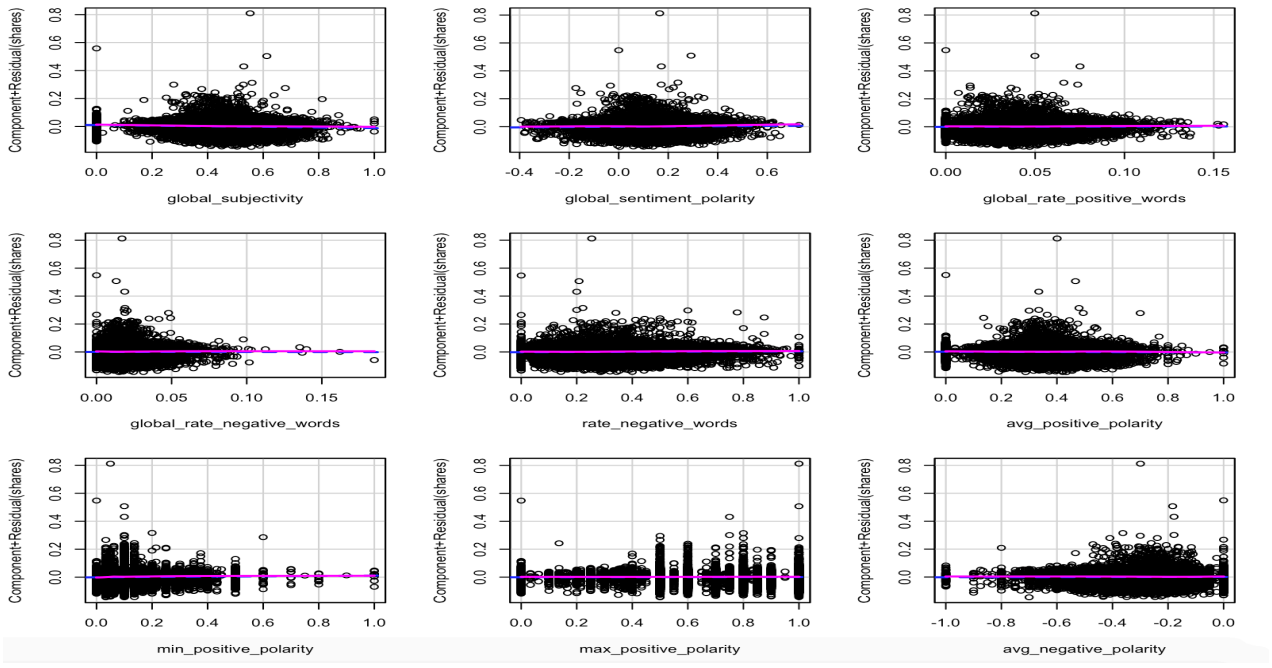
$$Residual + X * \beta_i \quad versus\ X$$

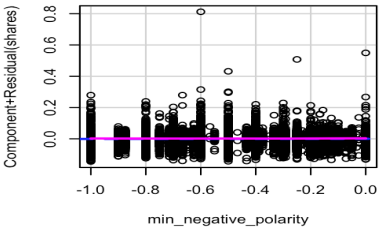
  - The lose line is residuals against the independent variable. While, the bold line represents the best fit. If both of the line seems to be linear, then the relationship is linear. If there is a curved lines there is likely a linearity problem

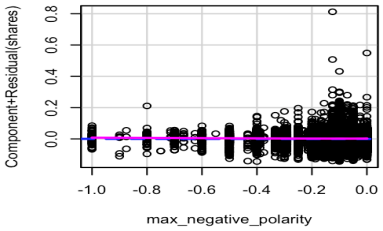

Component+Residual(shares)
timedelta
n_tokens_title
n_tokens_content
n_unique_tokens
n_non_stop_unique_tokens
num_hrefs
num_self_hrefs
num_imgs
num_videos

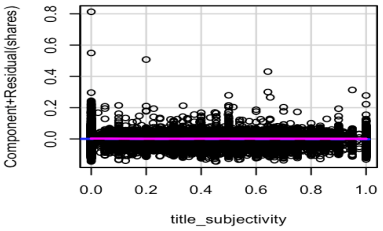

Component+Residual(shares)
average_token_length
num_keywords
data_channel_is_lifestyle
data_channel_is_entertainment
data_channel_is_bus
data_channel_is_socmed
data_channel_is_tech
data_channel_is_world
weekday_is_monday

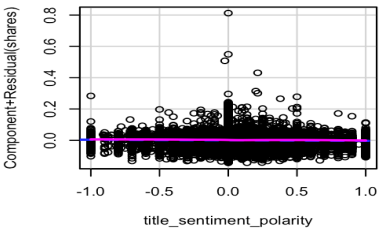

Component+Residual(shares)
weekday_is_tuesday
weekday_is_wednesday
weekday_is_thursday
weekday_is_friday
weekday_is_saturday
LDA_00
LDA_01
LDA_02
LDA_04

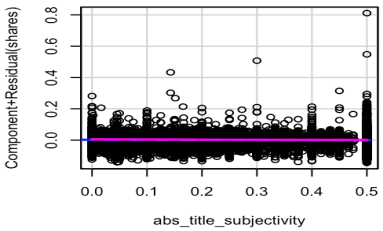

Component+Residual(shares)
global_subjectivity
global_sentiment_polarity
global_rate_positive_words
global_rate_negative_words
rate_negative_words
avg_positive_polarity
min_positive_polarity
max_positive_polarity
avg_negative_polarity

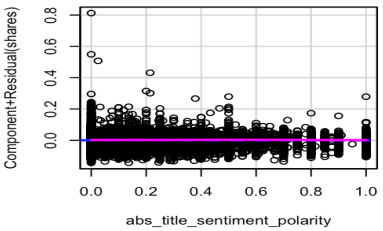


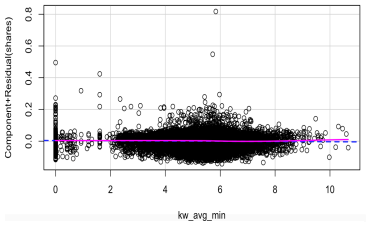


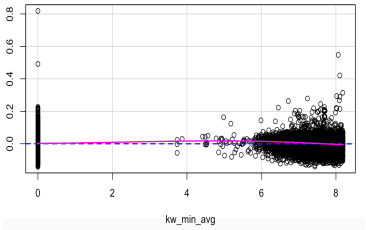


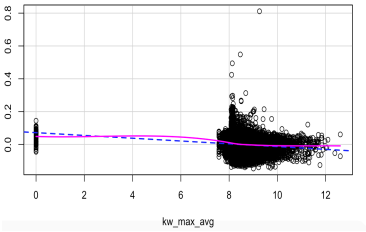


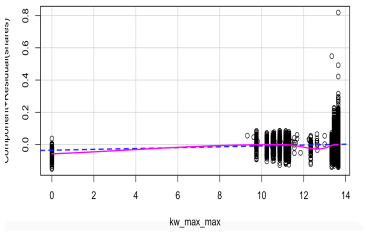




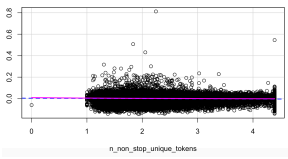




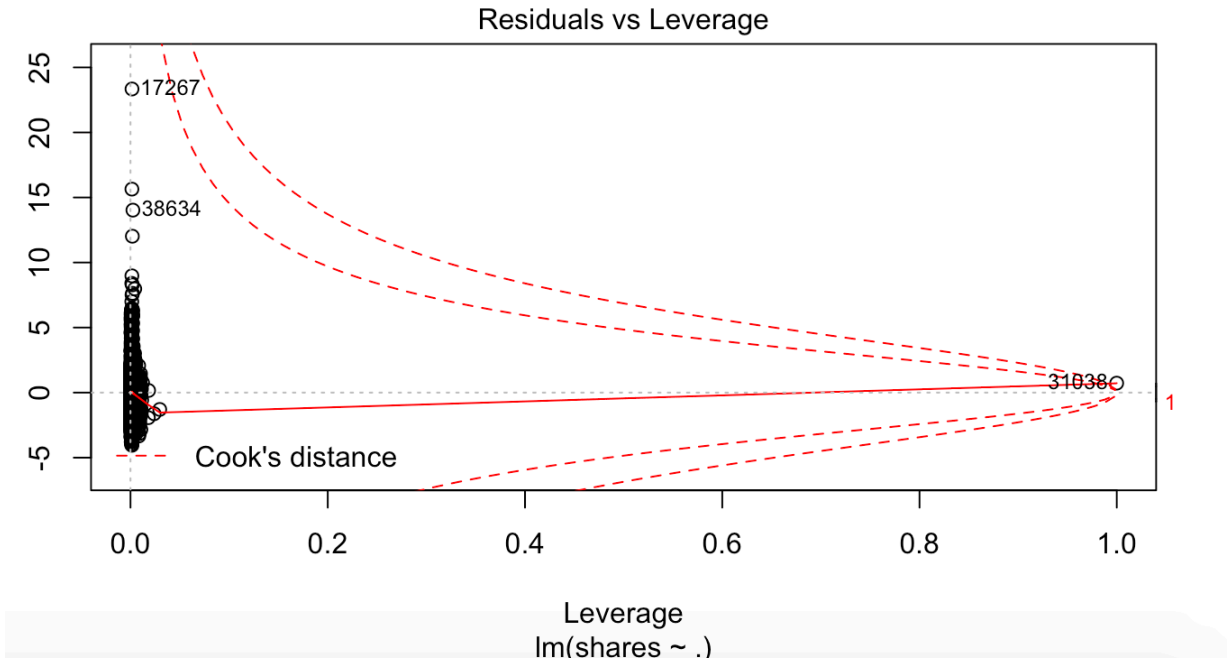




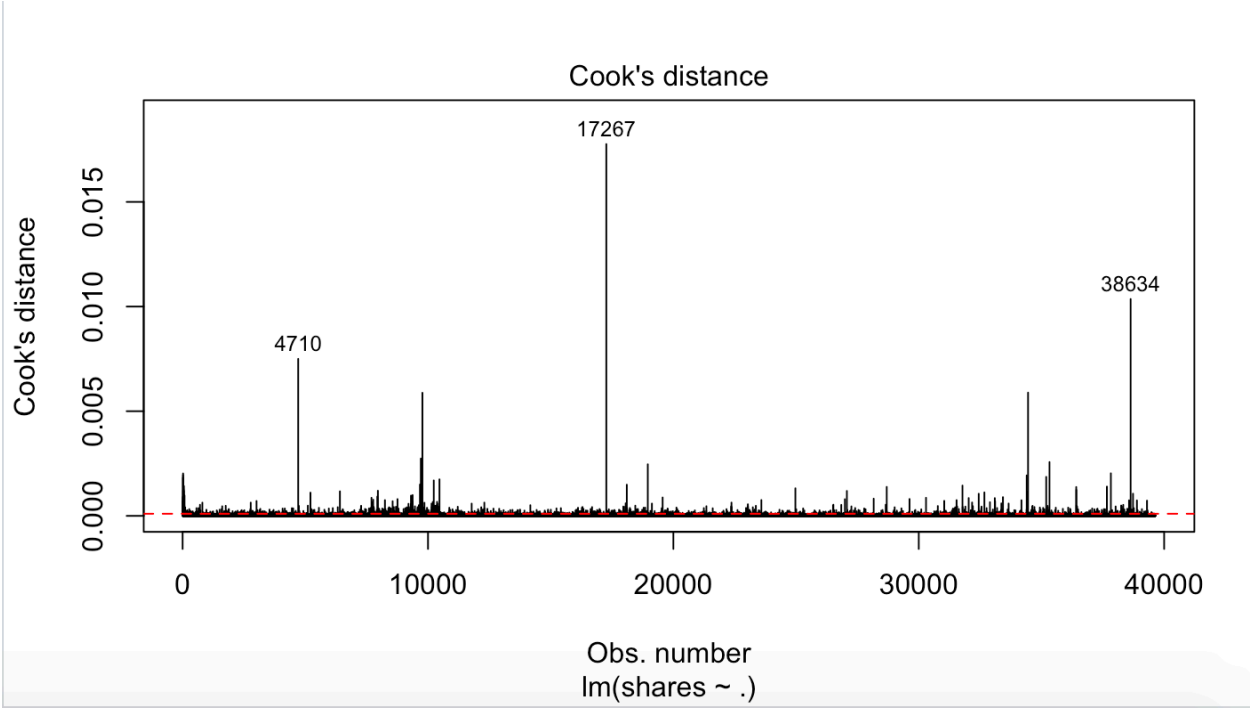
- We can see that following variables have a non-linear graph:
  - n_non_stop_unique_tokens
  - data_channel_is_lifestyle
  - data_channel_is_socmed
- Since the latter two are categorical variable, we won't be able to make much changes to it, but we need to make sure that n_non_stop_unique_tokens has a linear relationship. After applying the following transformation, we got the linear plot for n_non_stop_unique_tokens:
  - df$n_non_stop_unique_tokens <- df$n_non_stop_unique_tokens ** -2
  - df$n_non_stop_unique_tokens[df$n_non_stop_unique_tokens <0] = 0
  - df$n_non_stop_unique_tokens[df$n_non_stop_unique_tokens > 4.390] = 4.390
  - Plotting again after the above transformations, we get a linear plot:



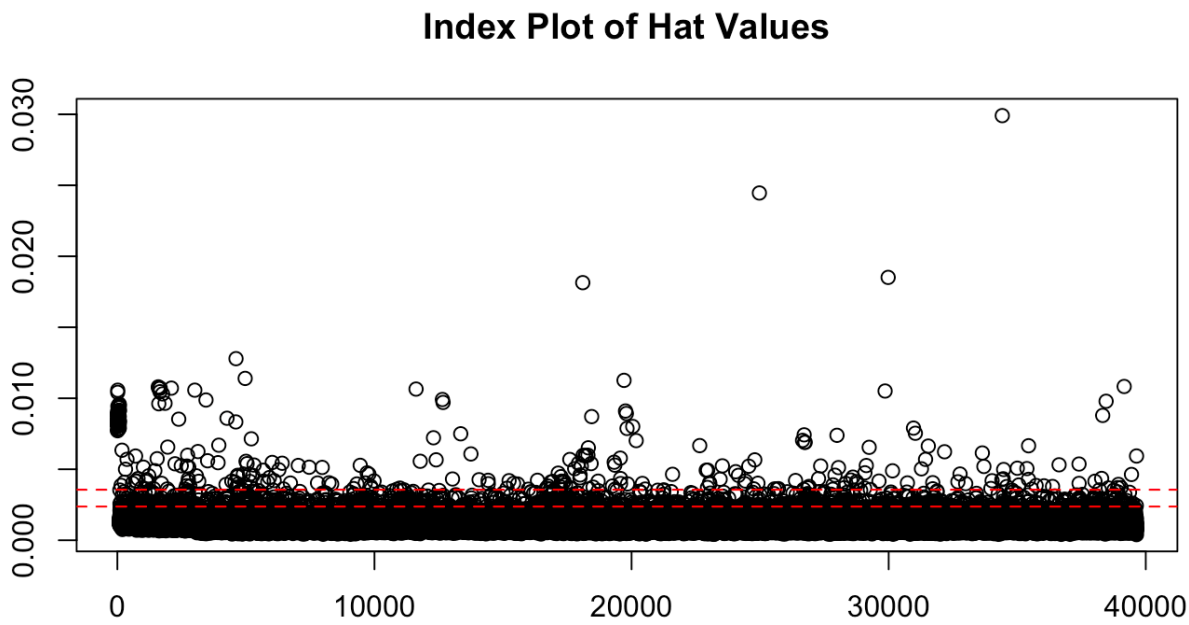
- **Outlier Analysis** :
  - All the observations in "Residual vs Leverage" graph are below the cook's distance, but we will need to dig a little deeper



- Plotting Influential observations:



- Plotting the high leverage plot (2/3 times of cook's distance):



- There are high number of outliers even when we plot 3 times of the cook's distance. As the standard suggests, we will be removing the rows yielding a residual that is 4 times of the mean of the cooks distance
    - Removing 1839 rows from the data frame. Accuracy increased to 16.3% from 11.7%

- **Normality**: The residuals should have a distribution which is approximately normal (with zero mean and constant variance and be independent).
    - We can see that towards the higher and lower t-quantiles , the q-q plot deviates a little. But it is still approximately in a straight line.

- **GVLMA**: In, Global validation of linear model assumption we can see that data does not meet all the statistical assumptions for regression model. But we won't be dwelling much into this.

	Value	p-value	Decision
Global Stat	1388.431	0.000000	Assumptions NOT satisfied!
Skewness	1269.610	0.000000	Assumptions NOT satisfied!
Kurtosis	7.392	0.006553	Assumptions NOT satisfied!
Link Function	2.629	0.104909	Assumptions acceptable.
Heteroscedasticity	108.800	0.000000	Assumptions NOT satisfied!

2(iv)

Attempt to choose and use at least some transformations of some of the variables to improve model fit and compliance with assumptions. Justify these transformations and discuss the results, with further discussion of the extent to which the assumptions are met after this

In the above table, I have made multiple transformations of the variables.
The transformations is done so that there is a Linear relationship between the dependent and the independent variables. This transformation is done after checking the variable is linearly dependent or not on the prediction variable, I experimented on the various transforms and choose the one best suited.

For example:

- log(kw_avg_avg + 1)
- log(n_tokens_content + 1)
- log(num_imgs + 1)
- Please refer to 2(ii) for detailed analysis on variables

Question - 3

Use a general linear model (multiple regression) to attempt to predict the number of times an article will be shared based on the other variables which are available before publication.

3(i)

List and explain which of the variables you think would and would not be available before publication.

The following tokens will not be available before publication, as these variable counts the number of times an article gets shared with respect to various other parameters:

- shares: Number of shares (target)
- self_reference_min_shares: Min. shares of referenced articles in Mashable
- self_reference_max_shares: Max. shares of referenced articles in Mashable
- self_reference_avg_shares: Avg. shares of referenced articles in Mashable
- **NOTE**: kw_min/max/avg_min/max/avg variables do seem like they fall in this category, however dwelling a bit into the reference research paper, I came to conclusion that they do not belong to this category because: This metrics is calculated by articles previously published in the same by Mashable.
    - To calculate the metrics; for each of the article keywords, they extract the minimum, average and maximum number of shares. In data channel categories (Fernandase, 2015)

3(ii)

Give a table including all the estimated model parameters (including error variance), confidence intervals, test statistics and p-values.

Column1	Estimate	Error_Variance	T_value	P_value	2.50%	97.50%
(Intercept)	0.220	0.004	60.692	0.000	0.213	0.227
n_unique_tokens	0.016	0.007	2.242	0.025	0.002	0.029
data_channel_is_entertainment	0.015	0.001	16.725	0.000	0.013	0.017
weekday_is_wednesday	0.014	0.001	17.633	0.000	0.012	0.015
data_channel_is_world:LDA_02	0.014	0.002	5.673	0.000	0.009	0.018
weekday_is_tuesday	0.014	0.001	17.294	0.000	0.012	0.015
weekday_is_thursday	0.013	0.001	16.989	0.000	0.012	0.015
LDA_01	0.011	0.002	6.137	0.000	0.008	0.015
weekday_is_monday	0.011	0.001	13.741	0.000	0.009	0.013
weekday_is_friday	0.010	0.001	12.563	0.000	0.009	0.012
global_sentiment_polarity	0.009	0.005	1.818	0.069	-0.001	0.019
min_positive_polarity	0.009	0.003	2.583	0.010	0.002	0.016
rate_negative_words	0.007	0.003	2.048	0.041	0.000	0.013
data_channel_is_lifestyle	0.007	0.002	3.753	0.000	0.003	0.010
data_channel_is_bus	0.007	0.001	5.446	0.000	0.004	0.009
LDA_02	0.006	0.002	3.658	0.000	0.003	0.010

data_channel_is_socmed	0.003	0.001	2.146	0.032	0.000	0.006
LDA_04	0.003	0.002	1.828	0.068	0.000	0.006
kw_max_max	0.003	0.000	10.692	0.000	0.002	0.003
average_token_length	0.002	0.001	3.091	0.002	0.001	0.003
min_negative_polarity	0.002	0.001	1.476	0.140	-0.001	0.005
n_tokens_content	0.002	0.000	3.240	0.001	0.001	0.003
max_negative_polarity	0.001	0.003	0.220	0.826	-0.005	0.007
num_self_hrefs	0.001	0.000	9.830	0.000	0.000	0.001
max_positive_polarity	0.001	0.001	0.392	0.695	-0.002	0.003
weekday_is_saturday	0.001	0.001	0.534	0.593	-0.001	0.002
data_channel_is_world	0.000	0.002	0.287	0.774	-0.003	0.003
kw_min_avg	0.000	0.000	-2.375	0.018	0.000	0.000
num_keywords	0.000	0.000	-1.320	0.187	0.000	0.000
num_videos	0.000	0.000	-3.157	0.002	0.000	0.000
n_tokens_title	0.000	0.000	-2.480	0.013	0.000	0.000
kw_avg_min	-0.001	0.000	-4.303	0.000	-0.001	0.000
avg_positive_polarity	-0.001	0.004	-0.228	0.820	-0.009	0.007
self_reference_avg_sharess	-0.001	0.000	-19.596	0.000	-0.001	-0.001
abs_title_sentiment_polarity	-0.001	0.001	-1.038	0.299	-0.004	0.001
avg_negative_polarity	-0.001	0.004	-0.370	0.711	-0.009	0.006
num_imgs	-0.002	0.000	-6.505	0.000	-0.002	-0.001
title_subjectivity	-0.002	0.001	-2.780	0.005	-0.004	-0.001
title_sentiment_polarity	-0.003	0.001	-4.317	0.000	-0.005	-0.002
n_non_stop_unique_tokens	-0.003	0.005	-0.636	0.525	-0.014	0.007
num_hrefs	-0.004	0.000	-10.853	0.000	-0.004	-0.003
data_channel_is_tech	-0.004	0.001	-3.131	0.002	-0.006	-0.001
LDA_00	-0.004	0.002	-2.526	0.012	-0.008	-0.001
abs_title_subjectivity	-0.005	0.001	-4.817	0.000	-0.007	-0.003
data_channel_is_lifestyle:LDA_04	-0.007	0.003	-2.435	0.015	-0.013	-0.001
global_rate_positive_words	-0.008	0.021	-0.362	0.717	-0.050	0.034
kw_max_avg	-0.008	0.000	-19.270	0.000	-0.009	-0.007
data_channel_is_entertainment:LDA_01	-0.013	0.002	-5.928	0.000	-0.017	-0.009
global_subjectivity	-0.019	0.003	-7.553	0.000	-0.024	-0.014
global_rate_negative_words	-0.020	0.041	-0.480	0.631	-0.101	0.061
data_channel_is_socmed:LDA_00	-0.026	0.003	-8.438	0.000	-0.031	-0.020

3(iii)

Interpret the two most significant slope parameters:

(1) Polarity: High rate of negative words or minimum positive polarity yields higher response variable
(2) Channel of sharing: Entertainment, Bus, World and Lifestyle gives good number of shares. 'Tech' and 'Socmed' does not yields higher response variable
(3) Day of sharing: An article shared on Saturday/Sunday, usually does not yields higher response variable. Any other day does yields higher response variable
(4) **Note**: The relationship between "shares" and the response variable is inverse, so the inverse is true for "shares" variable. As,

$$Y = shares^{-0.22}$$

Question - 4

Use a logistic regression model to attempt to predict whether or not an article will be popular (defined here as ≥1000 shares)

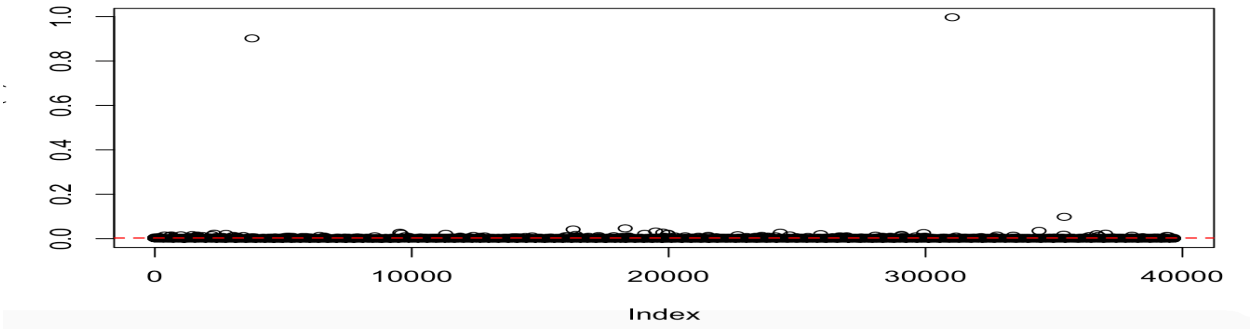
4(i)

You should attempt to check all assumptions of the model and report on this.

- **Dependent variable should be binary** – Using the condition, '1' if shares >= 1000 and '0' otherwise, I have made the dependent variable binary
- **Linearity with log odds** - The relationship between each explanatory variable and the log odd of the response variable is approximately a straight line -
    - For each variable we followed the following steps:
    - We divide the data into 100 or 1000 equal part using the variable in study
    - We compare the linearly dependency between mean of response and mean of variables in the 100 or 1000 parts of data
    - For linear dependency, I used the "Residue vs Fitted" graph to analyse the linear dependency between the variables
    - Note – Most of the graphs where approximately linear, I experimented with various transforms, including the inverse of box-cox transform but the relationship because very complex with lots of NaN and Infinite values. Which is why, I choose a transformation that is approximately linear instead of searching for a transform that gives a linear relationship
    - Please refer to 2(ii) for detailed analysis on variables
    - Other than this, I also experimented plotting the ggplot to understanding the linearity of the model. However, I was not able to get any information from the graphs. This was because of high number of independent variable present
- **Independent variable independent of each other**
    - Using Variance Inflation Factor, I was able to remove the variables with high VIF, one by one. I stopped when the highest VIF was less than 10. The following variables were removed using this process:
        - LDA_02

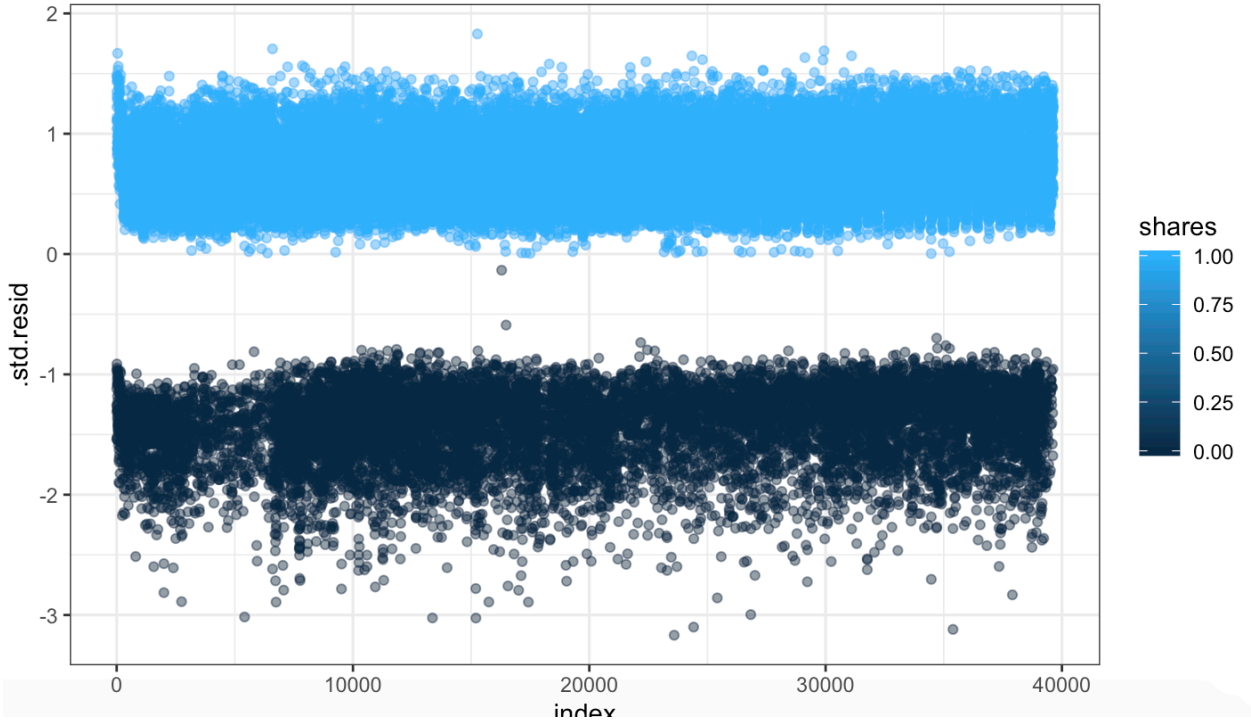
- rate_positive_words
- kw_avg_min
- n_tokens_content
- n_unique_tokens

- **No highly influential outliers** – Using cook's distance analysis to understand any influential outliers
  - Examine the high leverage points



  - Removed 7 most influential outliers, using the same method used for linear regression as explained in 2(iii)

- Residuals have clear distinction between the two classes



- Number of explanatory variables are at least n/3 times of total sample size – The number of explanatory variables is n/3 times of total sample size

4(ii)

By default, re-use any transformations or other processing of the X variables attempted with multiple linear regression. Comment on whether any further transformations or representations of the data may be useful. Use them if you think they help.

The following transformations were applied to the variables (More detail in 2(ii) and 4(i)) :

Variable	Log Odd Quartile Test
kw_max_min	log(kw_max_min + 1)
kw_avg_min	log(kw_avg_min + 2)
kw_min_max	log(kw_min_max + 1)
kw_min_avg[kw_min_avg <0] = 0 kw_min_avg	kw_min_avg
kw_avg_avg	log(kw_avg_avg + 1)
global_sentiment_polarity	global_sentiment_polarity ** -2
min_positive_polarity	min_positive_polarity ** 0.06
n_tokens_content	log(n_tokens_content + 1)
n_unique_tokens[(n_unique_tokens > 1)] = 1 n_unique_tokens	n_unique_tokens
n_non_stop_words[(n_non_stop_words > 1)] = 1 n_non_stop_words	n_non_stop_words ** -0.10
num_hrefs	log(num_hrefs+1)
num_imgs	log(num_imgs + 1)

4(iii)

Give a table including all the estimated model parameters, confidence intervals, test statistics and p-values.

Column1	Estimate	Error Variance	T value	P value	2.50%	97.50%
(Intercept)	-1.80	0.33	-5.46	0.00	-2.46	-1.17
global_rate_negative_words	2.93	2.63	1.11	0.27	-2.22	8.10
data_channel_is_socmed	1.16	0.09	12.46	0.00	0.98	1.35
LDA_00	1.01	0.11	9.55	0.00	0.80	1.22
global_subjectivity	0.93	0.17	5.51	0.00	0.60	1.26
LDA_04	0.71	0.09	7.67	0.00	0.53	0.90
data_channel_is_tech	0.50	0.08	6.34	0.00	0.35	0.66
kw_avg_avg	0.42	0.04	10.48	0.00	0.35	0.51

abs_title_subjectivity	0.28	0.08	3.66	0.00	0.13	0.42
max_negative_polarity	0.21	0.21	1.01	0.31	-0.20	0.63
LDA_01	0.18	0.10	1.79	0.07	-0.02	0.39
LDA_03	0.17	0.10	1.73	0.08	-0.02	0.36
num_hrefs	0.13	0.02	6.07	0.00	0.09	0.17
weekday_is_saturday	0.13	0.09	1.38	0.17	-0.05	0.31
title_sentiment_polarity	0.11	0.05	2.10	0.04	0.01	0.21
num_imgs	0.09	0.02	5.24	0.00	0.05	0.12
title_subjectivity	0.08	0.06	1.39	0.16	-0.03	0.19
abs_title_sentiment_polarity	0.05	0.08	0.59	0.56	-0.11	0.21
kw_max_min	0.03	0.01	3.31	0.00	0.01	0.05
num_keywords	0.03	0.01	4.42	0.00	0.02	0.05
num_videos	0.01	0.00	1.68	0.09	0.00	0.01
n_tokens_title	0.00	0.01	0.74	0.46	-0.01	0.02
kw_min_avg	0.00	0.00	18.78	0.00	0.00	0.00
self_reference_avg_sharess	0.00	0.00	9.01	0.00	0.00	0.00
kw_avg_max	0.00	0.00	2.56	0.01	0.00	0.00
global_sentiment_polarity	0.00	0.00	0.30	0.76	0.00	0.00
kw_max_max	0.00	0.00	-8.73	0.00	0.00	0.00
kw_max_avg	0.00	0.00	-2.28	0.02	0.00	0.00
data_channel_is_lifestyle	0.00	0.09	-0.03	0.98	-0.17	0.17
average_token_length	0.00	0.04	-0.08	0.94	-0.08	0.08
num_self_hrefs	-0.01	0.00	-3.44	0.00	-0.02	-0.01
min_negative_polarity	-0.04	0.09	-0.47	0.64	-0.22	0.13
data_channel_is_world	-0.05	0.08	-0.69	0.49	-0.20	0.10
max_positive_polarity	-0.08	0.08	-0.97	0.33	-0.24	0.08
kw_min_max	-0.13	0.01	-17.62	0.00	-0.15	-0.12
avg_positive_polarity	-0.13	0.20	-0.67	0.50	-0.52	0.26
avg_negative_polarity	-0.19	0.24	-0.78	0.43	-0.66	0.28
data_channel_is_bus	-0.19	0.08	-2.39	0.02	-0.35	-0.03
data_channel_is_entertainment	-0.34	0.05	-6.85	0.00	-0.44	-0.25
rate_negative_words	-0.39	0.21	-1.83	0.07	-0.81	0.03
min_positive_polarity	-0.43	0.24	-1.77	0.08	-0.92	0.04
global_rate_positive_words	-0.45	1.33	-0.34	0.73	-3.06	2.16
n_non_stop_unique_tokens	-0.62	0.15	-3.99	0.00	-0.92	-0.32
weekday_is_friday	-0.99	0.07	-14.52	0.00	-1.13	-0.86
weekday_is_monday	-1.17	0.07	-17.51	0.00	-1.30	-1.04
weekday_is_thursday	-1.30	0.07	-19.63	0.00	-1.43	-1.17
weekday_is_tuesday	-1.32	0.07	-19.97	0.00	-1.45	-1.19

weekday_is_wednesday	-1.36	0.07	-20.54	0.00	-1.49	-1.23

4(iv)

Interpret the two most significant slope parameters.

(1) **global_rate_negative_words** : The slope for global_rate_negative_words is 2.93, this means that on increasing global_rate_negative_words by 1, the log odd probability increases by 2.93.

a. On the other hand, on increasing global_rate_negative_words by 1 the probability increases by a factor of 0.94 (exp(estimation) / (1 + exp(estimation)))

(2) **data_channel_is_socmed**: The slope for data_channel_is_socmed is 1.16, this means that if the article is published in the channel "socmed", the log odd probability increase by 1.16.

a. On the other hand, on increasing global_rate_negative_words by 1 the probability increases by a factor of 0.76 (exp(estimation) / (1 + exp(estimation))

5.

Evaluate the predictive performance of each of the above two models using two appropriate metrics. Include details of each metric and its advantages and disadvantages. Compare your results with logistic regression to the reported results of Fernandes et al. with random forests

Metrics	Linear Regression	Logistic Regression
R - square	16.41%	-
Pseudo R -squarte	-	11.45%
RMSE	0.028	-
Accuracy	-	72.67%

- **Multiple Linear Regression:**
  - **R-square**: R-squared measures how close the data is to fitted regression line. Statistically, it compared the proportion of variance explained by the model to the overall variance of the response variable.

$$R^2 = 1 - \sum_{i=1}^{n} \frac{(y_i - \bar{y})^2}{(y_i - \hat{y}_i)^2}$$

Where, y_i is the response variable;

$\hat{y}_i$ is the predicted value or response variable;

$\bar{y}$ is the size of residual from a null model

- Accuracy: The accuracy of the model using R-square metrics is 16.41%
- Advantages:
  - Scale of R-squared value is intuitive, it ranges from 0 to 1. The more the model is close to 1, the more accurate it is.

        - This statistic measures the percentage of dependent variable explained. And usually, the large the R-square value – the better regression model fits the observation if the residual plots are acceptable.
    - Disadvantages:
        - R-square can give a low R-squared with for linearly model as well and high R-squared error for non-linear dependent model as well. We, need to analyse the residual plots to make sure that the predictions are not biased.
        - It always increases as more predictors are added to the model, ergo we need to consider metrics such as Adjusted R – square is used

- **RMSE**: RMSE measures the square root of variance of residual.
    - Accuracy: The accuracy of the model using R-square metrics is 0.028
    - Advantages:
        - In RMSE, we exactly know how much the prediction deviates from the actual value
        - Same unit as that of response variable
    - Disadvantages: Although, lower value of RMSE indicates better fit, there is no limit to the value of RMSE. So, we are not sure how low enough for a good model

- **Logistic Regression:**
    - **Confusion metrics:** It is used to summarize performance of a classification algorithm. The confusion metrics helps in understanding the type of error made by the prediction model. This metrics measures statistics in various forms such as accuracy, recall, precision and F-measure. I will be using "Accuracy" for this model

      $$\text{Accuracy} = (TP + TN)/ (TP + TN + FP + FN)$$

      Where,

      TP = Number of observations that are positive and the prediction is positive as well
      TN = Number of observations that are negative and the prediction is negative as well
      FP = Number of observations that are positive and the prediction is negative
      TP = Number of observations that are negative and the prediction is positive
        - Accuracy: The accuracy of the model is 72.67% .
        - Advantages: Helps to understand in which class the algorithm is failing
        - Disadvantages:
            - If the data does not have uniform distribution of classes, the information could be miss leading
            - If one class dominates, the data, evaluation will be dominated by that class

    - **Pseudo R-square:** Pseudo R-square, is seen as an equivalent of R-square.

$$R^2 = 1 - \sum_{i=1}^{n} \frac{(y_i - \hat{y}_i)^2}{(y_i - \bar{y})^2}$$

Where, y_i is the response variable;
$\bar{y}$ is the mean of , y_i
$\hat{y}_i$ is the predicted probabilities or response variable;

- Accuracy: The accuracy of the model is 11.45%
- Advantages:
  - Measures the goodness of fit of Logistic Model
  - they look like R-squared i.e. they have similar scale, ranging from 0 to 1
- Disadvantages:
  - While calculating the Pseudo – R square, we do not minimize variance, hence they cannot be interpreted as one will interpret R square in OLS
  - R-squared and different pseudo R-squared can arrive at very different values.

- **Comparison of Evaluation between Logistic Regression and Fernandes et al Random Forest :**
  - We cannot directly compare the two metrics directly, as our Logistic Regression is based on shares >= 1000; while for Fernandes et al Random Forest is based on shares>= 1400.
  - On doing a rough analysis(without checking all the assumptions, suing the same as done for shares >= 1000) at shares >= 1400, we get the following results

	Accuracy	Precision	Recall	F1
Logistic Regression	65.7%	66.90%	70.68%	68.74%
Random forest	67%	67%	71%	69%

6.

For the aims of the study, as first outlined by Fernandez, and any related aims which you might reasonably ascribe to Mashable, which of the two models (multiple linear regression or logistic regression) do you think is more useful and why? This is mainly an argument about whether or not it is worthwhile to discretise the shares variable

The aim of the study is to prediction of the news popularity.

For **linear regression model**, I was able to get a max accuracy of 16.41% only. Which means that the model is able to explain 16.41% of the response variable and 83.59% of the response variable remains unexplained. So, even if Linear regression is able to predict the exact number of 'shares' in an article. The prediction is not satisfactory , probably due to the fact that explanatory variables are not enough to explain the dependent variables.

On the other hand, the **Logistic model** (shares > 1000) does gives a better accuracy of 72.6% w.r.t the model. However, there are two issues with the logistic regression here. Firstly, it only tells if the article will be shared greater than 1000 times or not. Secondly, number of articles less than 1000 shares is just 28% of the total data, so, it does not cover a satisfactory distribution of data. In the paper by Fernandes et al., the division was made on shares >= 1400 which covers 50% of the data in both the classes giving a much statistically useful model

If I have to choose from the two models, I will definitely choose the **Logistic model.** Because of the following reason:

(1) Since we are predicting the number of shares for a given article, 16.41% R^2 in linear regression model, is very less to make any reliable predictions.
(2) The logistic regression model gives a better accuracy, not only at shares >1000. But as seen in Question – 5 above, at shares > 1400 as well the logistic regression model gives an accuracy which is at par with the accuracy using Random Forest model
(3) Using the logistic regression, we are still able to meet the aim of the study i.e. prediction of news popularity

7.

The model used by Fernandes et al. was a random forest. You may not know this model, but it is essentially a large ensemble of decision tree models, with generally strong predictive capabilities, but weak interpretability. It is hoped that the regression models you used will be more interpretable.

7(i)

Explain how one can determine how to improve the predicted popularity of an article with either of the regression models considered here.

Let's say, we want to improve the predicted popularity of an article, using the Linear Regression model. In the current model, the equation that can be seen as

$$Shares^{-0.23} = \sum(\beta * Variable) + \in$$

Where, shares is the response variable, ; β is the coefficient estimation ; Variable is all the variables considered in linear regression and ∈ is the error term

This equation can be written as:

$$Shares = \frac{1}{(\sum(\beta * Variable) + \epsilon)^{1/0.23}}$$

Or,

$$Shares = \frac{1}{(\sum(\beta * Variable) + \epsilon)^{4.35}}$$

So, to increase the popularity of an article, we need to decrease parameters that have a positive coefficient(or estimation) with the dependent variable and increase the ones with negative coefficient(or estimation).
For instance, we need to increase the value of the following coefficients:

Variables	Estimate	Error Variance	T value	P value	2.50%	97.50%
abs_title_subjectivity	-0.01	0.00	-5.19	0.00	-0.01	0.00
LDA_00	-0.01	0.00	-6.79	0.00	-0.01	-0.01
global_subjectivity	-0.02	0.00	-7.82	0.00	-0.03	-0.02

7(ii)

Using your two fitted regression models, identify the article with the highest predicted popularity and predicted probability of being popular.

Linear Regression most popular articles	Logistic Regression most popular articles
http://mashable.com/2013/04/20/top-comments-8/	http://mashable.com/2013/05/05/earned-media/
http://mashable.com/2013/01/26/facebook-pranks/	http://mashable.com/2014/07/20/apollo-11-45th-anniversary/

7(iii)

For each of your two fitted regression models, list the attributes of two hypothetical articles (fake news?) which would give the highest possible predicted popularity and predicted probability of being popular, respectively. You should give values for every attribute, but for each variable, keep them within the range seen in the dataset.

For Linear Regression, the dependent and independent variable has inverse relationship. I have created the hypothetical article with "highest popularity" in the following manner:

- if the estimate < 0, then I kept the maximum value of the variable.
- If estimate >0, then I kept the minimum value of the variable
- Maximum and Minimum is chosen so that the range of the variable is maintained. At the same time the variable meets the condition required for maximum shares

- This will help to get the lowest value of the prediction thus increasing the shares due to the inverse relationship

For Logistic Regression, the dependent and independent variable has a direct relationship. I have created the hypothetical article with "highest popularity" in the following manner:

- if the estimate > 0, then I kept the maximum value of the variable.
- If estimate < 0, then I kept the minimum value of the variable
- Maximum and Minimum is chosen so that the range of the variable is maintained. At the same time the variable meets the condition required for maximum shares

For the both of the regression model, a hypothetical article with the maximum popularity will have the following attribute:

Variable	Estimate_Linear_Regression	Variable_Linear_Regression	Estimate_Logistic_Regression	Variable_Logistic_Regression
global_rate_negative_words	-0.043	0.185	2.869	0.185
global_subjectivity	-0.017	1.000	0.938	1.000
LDA_00	-0.014	0.927	1.015	0.927
kw_max_avg	-0.010	12.606	0.000	0.000
data_channel_is_socmed	-0.006	1.000	1.165	1.000
LDA_04	-0.005	0.927	0.716	0.927
abs_title_subjectivity	-0.005	0.500	0.275	0.500
title_sentiment_polarity	-0.004	1.000	0.108	1.000
num_hrefs	-0.003	5.720	0.132	304.000
max_negative_polarity	-0.003	0.000	0.215	0.000
title_subjectivity	-0.002	1.000	0.081	1.000
data_channel_is_tech	-0.002	1.000	0.497	1.000
avg_negative_polarity	-0.001	0.000	-0.187	-1.000
num_imgs	-0.001	4.860	0.088	128.000
global_rate_positive_words	-0.001	0.155	-0.463	0.000
self_reference_avg_sharess	-0.001	13.645	0.000	843300.000

kw_avg_min	-0.001	10.665	0.000	0.000
weekday_is_saturday	-0.001	1.000	0.127	1.000
kw_min_avg	0.000	8.193	0.001	3613.040
num_videos	0.000	91.000	0.006	91.000
n_non_stop_unique_tokens	0.000	650.000	-0.606	0.000
n_non_stop_words	0.000	0.000	0.000	0.000
shares	0.000	1.000	0.000	843300.000
kw_avg_avg	0.000	0.000	0.426	43567.660
kw_avg_max	0.000	0.000	0.000	843300.000
kw_max_min	0.000	0.000	0.034	298400.000
kw_min_max	0.000	0.000	-0.132	0.000
LDA_03	0.000	0.000	0.174	0.927
n_tokens_title	0.000	2.000	0.004	23.000
num_keywords	0.000	1.000	0.034	10.000
n_tokens_content	0.001	0.000	0.000	0.000
num_self_hrefs	0.001	0.000	-0.013	0.000
LDA_01	0.001	0.000	0.187	0.926
avg_positive_polarity	0.001	0.000	-0.132	0.000
max_positive_polarity	0.001	0.000	-0.078	0.000
abs_title_sentiment_polarity	0.002	0.000	0.048	1.000
min_negative_polarity	0.002	-1.000	-0.042	-1.000
average_token_length	0.002	0.000	-0.005	0.000
kw_max_max	0.002	0.000	0.000	0.000
LDA_02	0.006	0.000	0.000	0.000
data_channel_is_lifestyle	0.006	0.000	-0.004	0.000
global_sentiment_polarity	0.007	-0.394	0.000	0.728
rate_negative_words	0.008	0.000	-0.386	0.000
data_channel_is_world	0.008	0.000	-0.055	0.000

weekday_is_friday	0.009	0.000	-0.989	0.000
weekday_is_monday	0.011	0.000	-1.170	0.000
data_channel_is_bus	0.012	0.000	-0.196	0.000
n_unique_tokens	0.012	0.000	0.000	0.000
min_positive_polarity	0.013	0.000	-0.426	0.000
data_channel_is_entertainment	0.013	0.000	-0.348	0.000
weekday_is_thursday	0.013	0.000	-1.299	0.000
weekday_is_tuesday	0.013	0.000	-1.320	0.000
weekday_is_wednesday	0.013	0.000	-1.355	0.000
(Intercept)	0.224	0.000	-1.825	0.000

7(iv)

Based on your analysis of dependence between variables in the dataset, comment on whether or not these hypothetical articles could be produced.

Linear Regression:

(1) The hypothetical article that I have produced is not possible, as the "predicated" share value of this article is coming to be negative

(2) Another problem is two data channels is being considered, while in the data, we saw that every article had one data cannel

Logistic Regression:

(1) The hypothetical article seems to be valid, with the log odd of 0.244 and the probability of 0.56, indicating that the number of shares is greater than 1000

Resources:

https://www.theanalysisfactor.com/assessing-the-fit-of-regression-models/
https://www.geeksforgeeks.org/confusion-matrix-machine-learning/
https://stats.idre.ucla.edu/other/mult-pkg/faq/general/faq-what-are-pseudo-r-squareds/
https://www.lexjansen.com/wuss/2018/130_Final_Paper_PDF.pdf
https://www.lexjansen.com/wuss/2018/130_Final_Paper_PDF.pdf
https://stats.idre.ucla.edu/stata/faq/how-can-i-understand-a-categorical-by-continuous-interaction-stata-12/
https://www.theanalysisfactor.com/interpreting-interactions-in-regression/

https://stattrek.com/multiple-regression/interaction.aspx
https://data.library.virginia.edu/diagnostic-plots/
https://stats.stackexchange.com/questions/58941/ncvtest-from-r-and-interpretation
http://r-statistics.co/Logistic-Regression-With-R.html