1. (i)

Multiple Linear Regression Model is:

$$Y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \varepsilon_i \ ,$$

And if written in the form of matrix,

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_N \end{pmatrix}, X = \begin{pmatrix} 1 & X_{11} & \cdots & X_{1p} \\ 1 & X_{21} & \cdots & X_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & \cdots & X_{np} \end{pmatrix}, \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}, \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

It can be represented as:

$$Y = X\beta + \varepsilon$$

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) β 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 ε are independent.

(3)Each ε has normal distribution.

(4)Each ε 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

$$\frac{\partial \log L_n(\theta; x)}{\partial \theta} = 0$$

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

$$I(\theta) = -\frac{\partial^2 \log L_n(\theta; x)}{\partial \theta \partial \theta^T}|_{\theta} = \hat{\theta}_n$$

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

1. (iv)

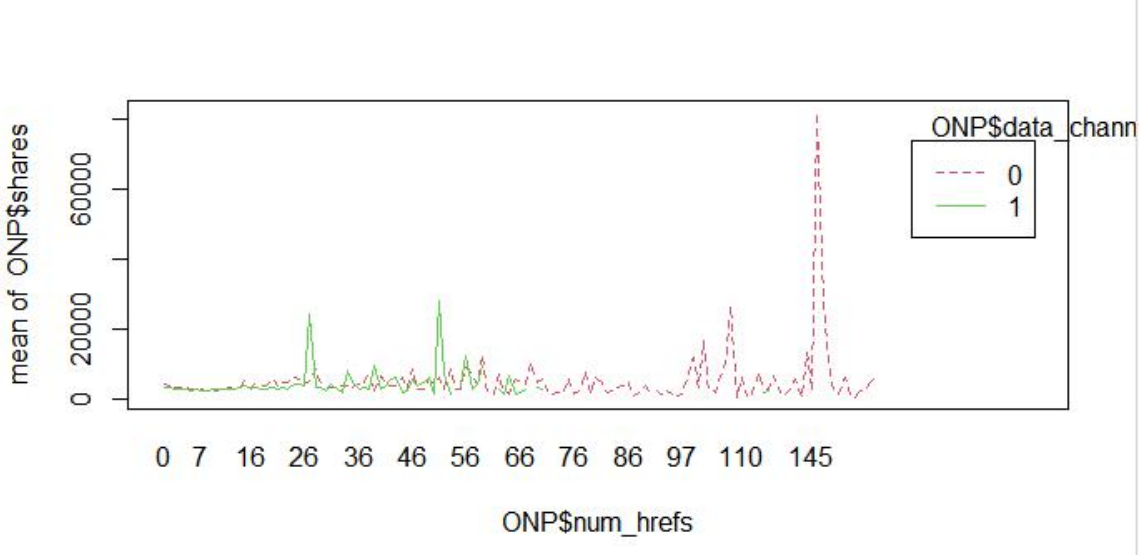
In this dataset, I just use a continuous variable - num_hrefs and a categorical variable – data_channel_is_tech to explain the interaction between a continuous variable and a binary categorical variable.

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

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

```
                                          Df    Sum Sq   Mean Sq F value
ONP$num_hrefs                              1 1.105e+10 1.105e+10  81.909
ONP$data_channel_is_tech                   1 5.875e+08 5.875e+08   4.355
ONP$num_hrefs:ONP$data_channel_is_tech     1 7.975e+08 7.975e+08   5.913
Residuals                              39640 5.347e+12 1.349e+08
                                      Pr(>F)
ONP$num_hrefs                         <2e-16 ***
ONP$data_channel_is_tech              0.0369 *
ONP$num_hrefs:ONP$data_channel_is_tech 0.0150 *
Residuals
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

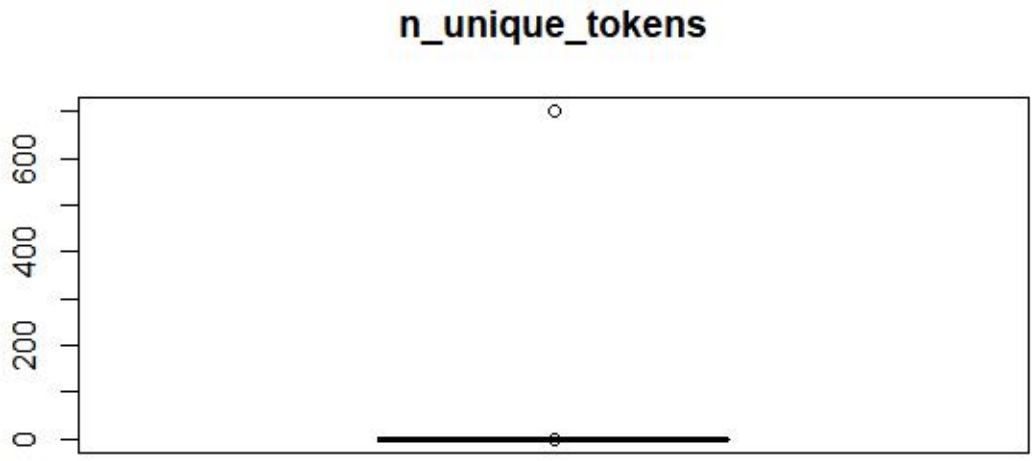
In the line chart shown below, we could found that the dot line is not parallel to the solid one, which means they do interact with each other.

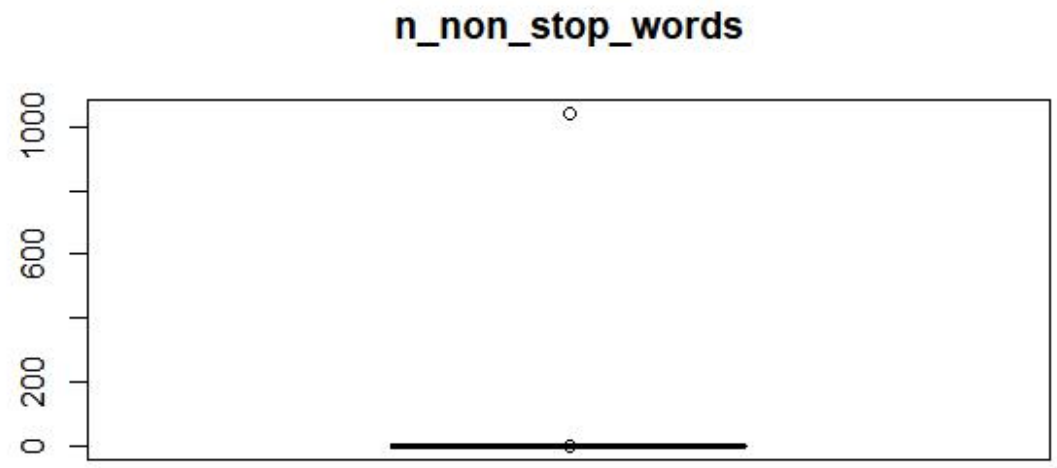


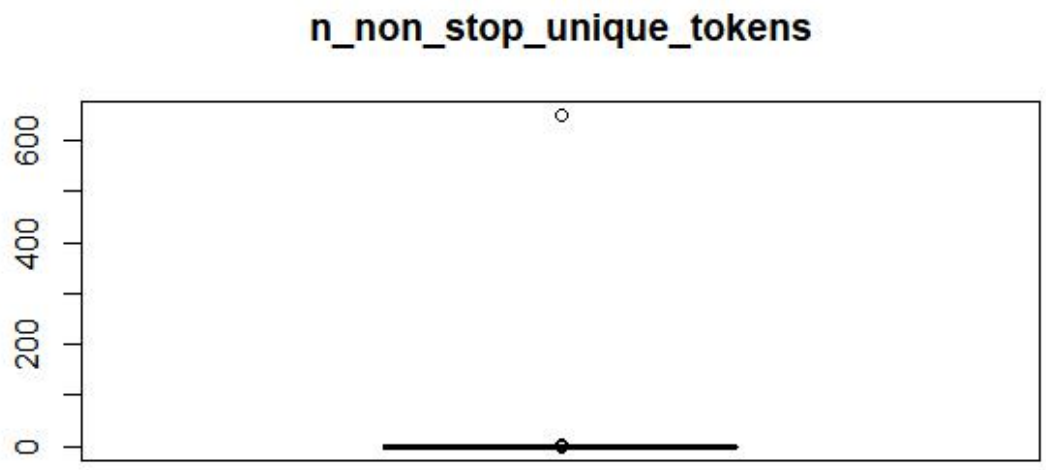
2. (i)
(1)There are numerical data,character data and categorical data in this dataset.
(2)Interaction exists between some variables.
(3)There are some errors and outliers in this dataset.
After using R to summary this dataset, I found some abnormal data. Thus, I draw boxplots to observe them better.

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.

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-1.092e+02	8.221e+02	-0.133	0.894303	
n_tokens_title	9.019e+01	2.928e+01	3.080	0.002072	**
n_tokens_content	6.118e-01	2.270e-01	2.694	0.007053	**
n_unique_tokens	3.875e+03	1.950e+03	1.987	0.046925	*
n_non_stop_words	-1.761e+03	6.841e+03	-0.257	0.796890	
n_non_stop_unique_tokens	-1.512e+03	1.657e+03	-0.913	0.361395	
num_hrefs	2.530e+01	6.814e+00	3.713	0.000205	***
num_self_hrefs	-5.832e+01	1.805e+01	-3.230	0.001238	**
num_imgs	1.259e+01	9.083e+00	1.386	0.165874	
num_videos	6.720e+00	1.597e+01	0.421	0.673846	
average_token_length	-5.526e+02	2.471e+02	-2.236	0.025365	*
num_keywords	4.832e+01	3.791e+01	1.275	0.202406	
data_channel_is_lifestyle	-1.024e+03	4.009e+02	-2.553	0.010672	*
data_channel_is_entertainment	-1.181e+03	2.586e+02	-4.564	5.02e-06	***
data_channel_is_bus	-7.951e+02	3.881e+02	-2.049	0.040498	*
data_channel_is_socmed	-5.976e+02	3.784e+02	-1.579	0.114276	
data_channel_is_tech	-5.343e+02	3.766e+02	-1.419	0.155943	
data_channel_is_world	-4.673e+02	3.822e+02	-1.223	0.221402	
kw_max_min	8.468e-02	5.052e-02	1.676	0.093688	.
kw_avg_min	-3.323e-01	3.103e-01	-1.071	0.284261	
kw_min_max	-2.797e-03	1.377e-03	-2.032	0.042193	*
kw_max_max	-1.120e-03	3.917e-04	-2.859	0.004253	**
kw_avg_max	-7.812e-04	8.472e-04	-0.922	0.356485	
kw_min_avg	-3.607e-01	7.701e-02	-4.683	2.84e-06	***
kw_max_avg	-2.054e-01	2.562e-02	-8.018	1.11e-15	***
kw_avg_avg	1.683e+00	1.460e-01	11.526	< 2e-16	***
self_reference_min_shares	2.624e-02	7.595e-03	3.455	0.000551	***
self_reference_max_shares	5.767e-03	4.120e-03	1.400	0.161541	
self_reference_avg_sharess	-5.909e-03	1.053e-02	-0.561	0.574815	
weekday_is_monday	2.700e+02	2.680e+02	1.007	0.313704	
weekday_is_tuesday	-2.830e+02	2.641e+02	-1.071	0.284013	
weekday_is_wednesday	-1.115e+02	2.640e+02	-0.422	0.672735	
weekday_is_thursday	-2.949e+02	2.645e+02	-1.115	0.265015	
weekday_is_friday	-2.360e+02	2.743e+02	-0.860	0.389646	

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.

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	1.624e+03	7.486e+02	2.170	0.030027	*
n_tokens_title	8.222e+01	2.918e+01	2.818	0.004838	**
n_tokens_content	2.467e-01	1.848e-01	1.335	0.181912	
num_hrefs	2.761e+01	6.536e+00	4.225	2.40e-05	***
num_self_hrefs	-7.046e+01	1.775e+01	-3.969	7.24e-05	***
num_imgs	1.557e+01	8.435e+00	1.846	0.064835	.
num_videos	8.240e+00	1.584e+01	0.520	0.603029	
num_keywords	1.111e+02	3.727e+01	2.981	0.002874	**
data_channel_is_lifestyle	-1.255e+03	3.990e+02	-3.144	0.001665	**
data_channel_is_entertainment	-1.960e+03	2.497e+02	-7.850	4.27e-15	***
data_channel_is_bus	-1.762e+03	3.789e+02	-4.651	3.31e-06	***
data_channel_is_socmed	-9.553e+02	3.765e+02	-2.537	0.011180	*
data_channel_is_tech	-1.235e+03	3.711e+02	-3.329	0.000873	***
data_channel_is_world	-1.364e+03	3.743e+02	-3.644	0.000269	***
kw_min_max	-3.365e-03	1.378e-03	-2.442	0.014599	*
kw_max_max	-7.532e-04	3.865e-04	-1.949	0.051351	.
kw_avg_max	2.050e-03	7.946e-04	2.580	0.009876	**
kw_min_avg	1.966e-01	6.055e-02	3.247	0.001167	**
kw_max_avg	6.201e-02	1.016e-02	6.103	1.05e-09	***
self_reference_min_shares	2.299e-02	3.411e-03	6.740	1.61e-11	***
self_reference_max_shares	3.811e-03	1.682e-03	2.265	0.023498	*
weekday_is_monday	2.208e+02	2.682e+02	0.823	0.410494	
weekday_is_tuesday	-2.989e+02	2.644e+02	-1.130	0.258299	
weekday_is_wednesday	-1.498e+02	2.643e+02	-0.567	0.570909	
weekday_is_thursday	-3.262e+02	2.649e+02	-1.231	0.218164	
weekday_is_friday	-2.441e+02	2.747e+02	-0.889	0.374119	
weekday_is_saturday	4.214e+02	3.277e+02	1.286	0.198482	
LDA_00	6.178e+02	4.651e+02	1.328	0.184154	
LDA_01	1.492e+02	5.129e+02	0.291	0.771119	
LDA_02	-1.340e+03	4.625e+02	-2.897	0.003769	**
LDA_03	9.438e+02	4.820e+02	1.958	0.050255	.
global_subjectivity	2.089e+03	7.793e+02	2.681	0.007353	**
global_sentiment_polarity	1.248e+03	1.604e+03	0.778	0.436442	
global_rate_positive_words	-1.282e+04	5.950e+03	-2.154	0.031241	*
global_rate_negative_words	3.328e+03	9.816e+03	0.339	0.734584	
avg_positive_polarity	-2.068e+03	1.343e+03	-1.541	0.123436	

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.

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	1.624e+03	7.486e+02	2.170	0.030027	*
n_tokens_title	8.222e+01	2.918e+01	2.818	0.004838	**
n_tokens_content	2.467e-01	1.848e-01	1.335	0.181912	
num_hrefs	2.761e+01	6.536e+00	4.225	2.40e-05	***
num_self_hrefs	-7.046e+01	1.775e+01	-3.969	7.24e-05	***
num_imgs	1.557e+01	8.435e+00	1.846	0.064835	.
num_videos	8.240e+00	1.584e+01	0.520	0.603029	
num_keywords	1.111e+02	3.727e+01	2.981	0.002874	**
data_channel_is_lifestyle	-1.255e+03	3.990e+02	-3.144	0.001665	**
data_channel_is_entertainment	-1.960e+03	2.497e+02	-7.850	4.27e-15	***
data_channel_is_bus	-1.762e+03	3.789e+02	-4.651	3.31e-06	***
data_channel_is_socmed	-9.553e+02	3.765e+02	-2.537	0.011180	*
data_channel_is_tech	-1.235e+03	3.711e+02	-3.329	0.000873	***
data_channel_is_world	-1.364e+03	3.743e+02	-3.644	0.000269	***
kw_min_max	-3.365e-03	1.378e-03	-2.442	0.014599	*
kw_max_max	-7.532e-04	3.865e-04	-1.949	0.051351	.
kw_avg_max	2.050e-03	7.946e-04	2.580	0.009876	**
kw_min_avg	1.966e-01	6.055e-02	3.247	0.001167	**
kw_max_avg	6.201e-02	1.016e-02	6.103	1.05e-09	***
self_reference_min_shares	2.299e-02	3.411e-03	6.740	1.61e-11	***
self_reference_max_shares	3.811e-03	1.682e-03	2.265	0.023498	*
weekday_is_monday	2.208e+02	2.682e+02	0.823	0.410494	
weekday_is_tuesday	-2.989e+02	2.644e+02	-1.130	0.258299	
weekday_is_wednesday	-1.498e+02	2.643e+02	-0.567	0.570909	
weekday_is_thursday	-3.262e+02	2.649e+02	-1.231	0.218164	
weekday_is_friday	-2.441e+02	2.747e+02	-0.889	0.374119	

n_tokens_title	1.089206	n_tokens_content	2.187474
num_hrefs	1.589743	num_self_hrefs	1.360603
num_imgs	1.425050	num_videos	1.228195
num_keywords	1.411477	data_channel_is_lifestyle	2.289876
data_channel_is_entertainment	2.635174	data_channel_is_bus	5.475841
data_channel_is_socmed	2.257740	data_channel_is_tech	6.038189
data_channel_is_world	6.603736	kw_min_max	1.319244
kw_max_max	2.010585	kw_avg_max	3.189683
kw_min_avg	1.353987	kw_max_avg	1.125379
self_reference_min_shares	1.325504	self_reference_max_shares	1.395734
weekday_is_monday	2.897081	weekday_is_tuesday	3.045725
weekday_is_wednesday	3.061405	weekday_is_thursday	3.025341
weekday_is_friday	2.658228	weekday_is_saturday	1.780191
LDA_00	4.307292	LDA_01	3.668160
LDA_02	4.799763	LDA_03	5.858726
global_subjectivity	2.360444	global_sentiment_polarity	6.946051
global_rate_positive_words	3.081476	global_rate_negative_words	3.259806
avg_positive_polarity	5.628448	min_positive_polarity	1.913673

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

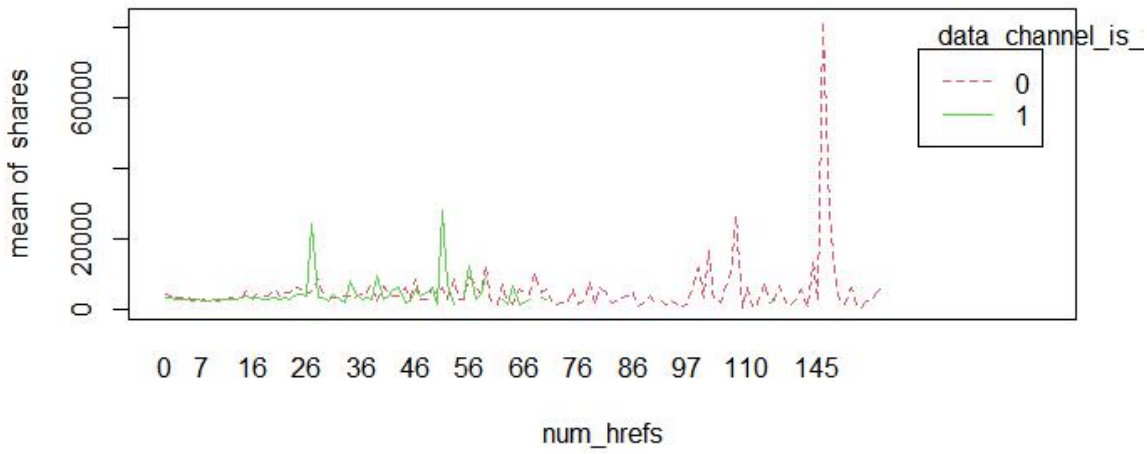
$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2$$

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
num_hrefs	1	1.067e+10	1.067e+10	77.747	<2e-16	***
data_channel_is_tech	1	7.177e+08	7.177e+08	5.232	0.0222	*
num_hrefs:data_channel_is_tech	1	8.525e+08	8.525e+08	6.215	0.0127	*
Residuals	38800	5.323e+12	1.372e+08			

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

As the picture listed above, the p-value of the interaction term $X_1 X_2$ 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.

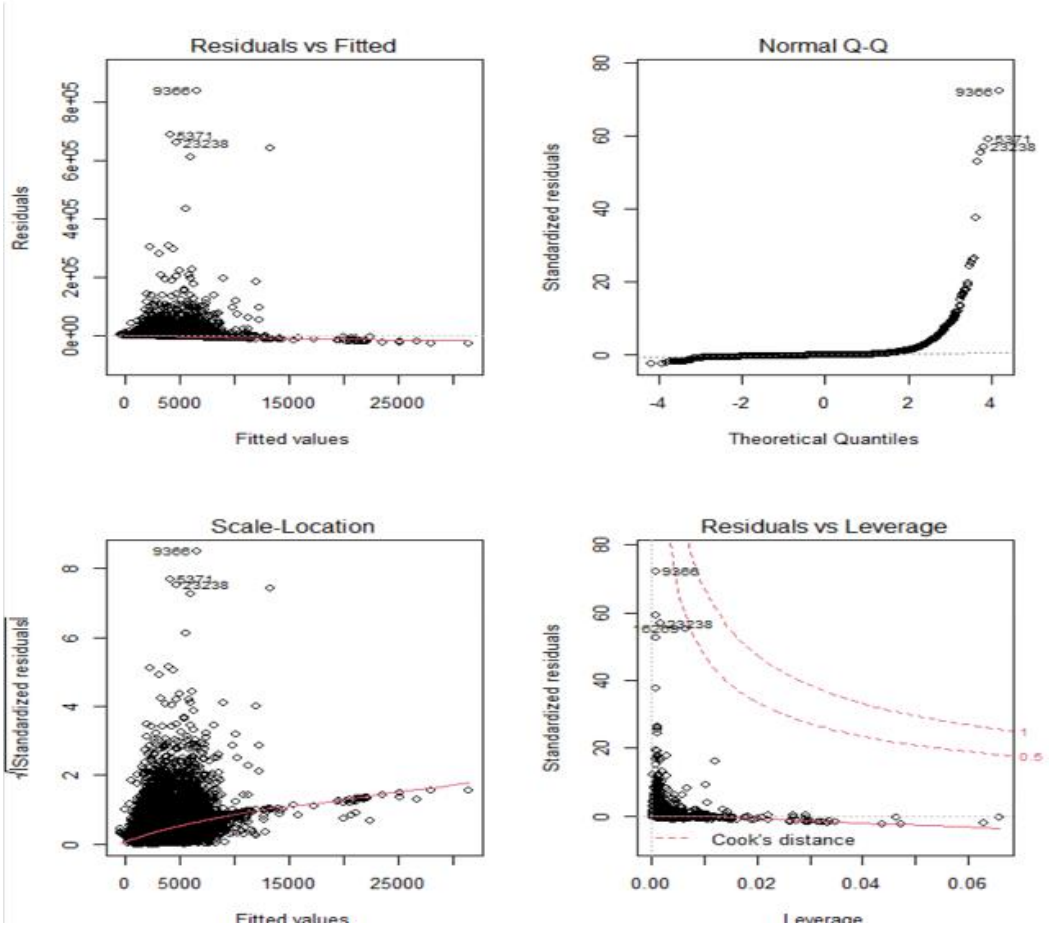


Use this interaction term to fit a new linear model with other exploratory variables against the response variable shares. From the picture of summary of current fitted model, we could found that the p-value of

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.

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	1.707e+03	7.488e+02	2.279	0.022645	*
n_tokens_title	8.728e+01	2.920e+01	2.989	0.002805	**
n_tokens_content	1.874e-01	1.854e-01	1.011	0.312074	
num_hrefs	2.306e+01	6.638e+00	3.475	0.000512	***
num_self_hrefs	-8.583e+01	1.818e+01	-4.721	2.36e-06	***
num_imgs	1.659e+01	8.438e+00	1.967	0.049241	*
num_videos	1.104e+01	1.586e+01	0.696	0.486436	
num_keywords	1.058e+02	3.729e+01	2.836	0.004571	**
data_channel_is_lifestyle	-1.248e+03	3.990e+02	-3.129	0.001754	**
data_channel_is_entertainment	-1.956e+03	2.496e+02	-7.837	4.73e-15	***
data_channel_is_bus	-1.781e+03	3.789e+02	-4.702	2.59e-06	***
data_channel_is_socmed	-9.228e+02	3.765e+02	-2.451	0.014254	*
data_channel_is_tech	-1.874e+03	4.054e+02	-4.623	3.79e-06	***
data_channel_is_world	-1.387e+03	3.743e+02	-3.706	0.000211	***
kw_min_max	-3.451e-03	1.378e-03	-2.505	0.012253	*
kw_max_max	-7.471e-04	3.865e-04	-1.933	0.053225	.
kw_avg_max	2.040e-03	7.944e-04	2.567	0.010256	*
kw_min_avg	1.949e-01	6.054e-02	3.220	0.001284	**
kw_max_avg	6.156e-02	1.016e-02	6.059	1.38e-09	***
self_reference_min_shares	2.277e-02	3.411e-03	6.674	2.52e-11	***
self_reference_max_shares	4.076e-03	1.683e-03	2.422	0.015456	*

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.

(2) 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.

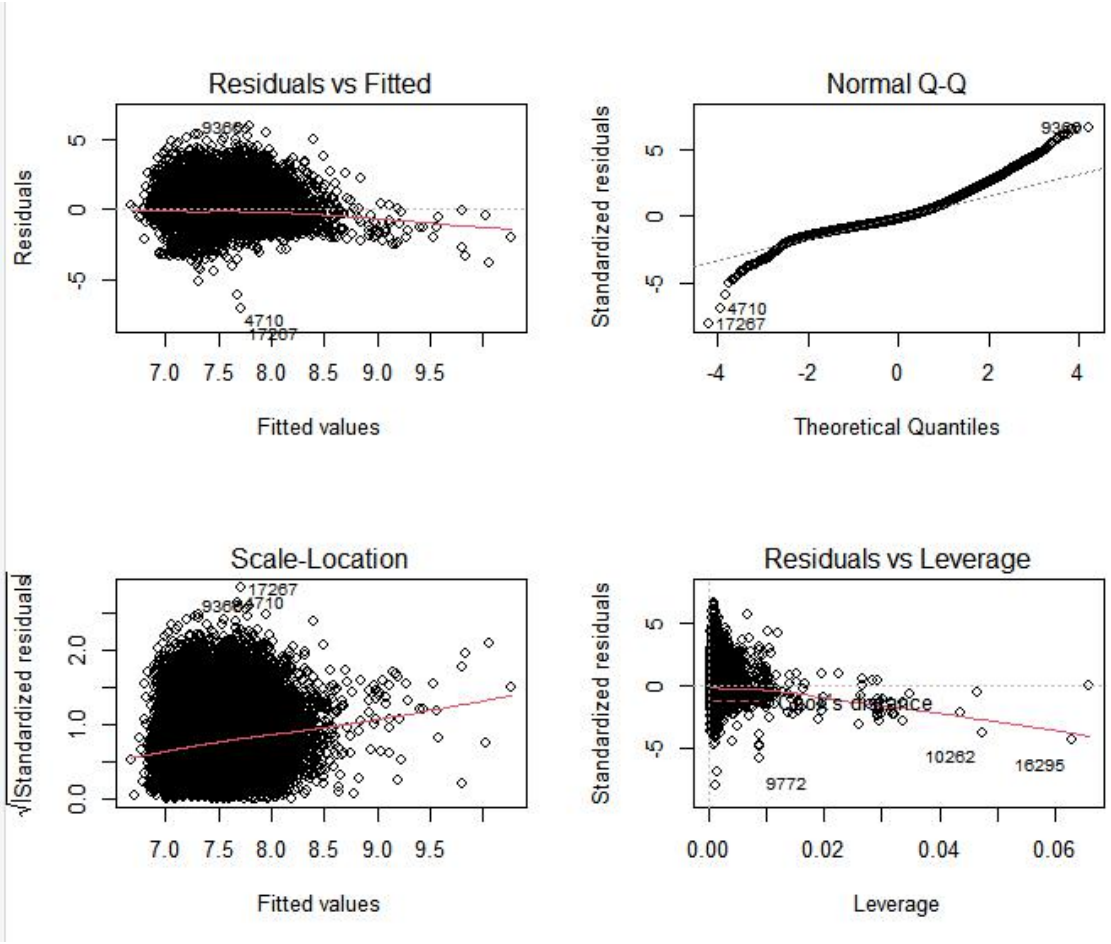
(3) Homoscedasticity

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.

(4) 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.

3. (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:

```
Coefficients:
                               Estimate Std. Error t value Pr(>|t|)
(Intercept)                   7.154e+00  7.784e-02  91.902  < 2e-16 ***
n_tokens_title                5.621e-02  2.463e-02   2.282 0.022465 *
n_tokens_content             -3.611e-02  6.720e-03  -5.373 7.77e-08 ***
num_hrefs                     5.610e-03  4.953e-04  11.326  < 2e-16 ***
num_self_hrefs               -1.024e-02  1.341e-03  -7.635 2.32e-14 ***
num_imgs                      5.680e-03  6.226e-04   9.123  < 2e-16 ***
num_videos                    5.316e-02  7.471e-03   7.116 1.13e-12 ***
num_keywords                  2.502e-02  2.850e-03   8.780  < 2e-16 ***
data_channel_is_lifestyle    -1.068e-01  2.994e-02  -3.567 0.000361 ***
data_channel_is_entertainment -3.382e-01 1.919e-02 -17.630  < 2e-16 ***
data_channel_is_bus          -3.320e-01  2.879e-02 -11.530  < 2e-16 ***
data_channel_is_socmed        1.137e-01  2.849e-02   3.992 6.56e-05 ***
data_channel_is_tech         -6.202e-03  2.773e-02  -0.224 0.823033
data_channel_is_world        -1.865e-01  2.848e-02  -6.547 5.93e-11 ***
kw_min_min                   -3.785e-07  1.051e-07  -3.600 0.000319 ***
kw_max_min                   -7.695e-08  2.953e-08  -2.606 0.009155 **
kw_avg_min                    1.405e-07  6.124e-08   2.294 0.021780 *
kw_min_avg                    6.719e-05  4.613e-06  14.564  < 2e-16 ***
kw_max_avg                    8.701e-06  7.631e-07  11.401  < 2e-16 ***
LDA_00                        2.668e-01  3.523e-02   7.574 3.69e-14 ***
LDA_01                       -5.435e-02  4.961e-02  -1.096 0.273210
LDA_02                       -3.347e-01  3.495e-02  -9.576  < 2e-16 ***
LDA_03                        2.186e-02  4.865e-02   0.449 0.653189
global_subjectivity           4.581e-01  6.269e-02   7.307 2.79e-13 ***
global_sentiment_polarity    -5.677e-03  1.352e-01  -0.042 0.966508
global_rate_positive_words   -8.861e-01  4.764e-01  -1.860 0.062879 .
global_rate_negative_words   -8.274e-01  7.753e-01  -1.067 0.285895
avg_positive_polarity        -1.173e-01  1.405e-01  -0.835 0.403544
min_positive_polarity        -3.790e-01  8.645e-02  -4.384 1.17e-05 ***
max_positive_polarity         2.298e-02  5.877e-02   0.391 0.695844
avg_negative_polarity        -7.768e-02  9.699e-02  -0.801 0.423204
min_negative_polarity        -7.955e-02  3.420e-02  -2.326 0.020006 *
max_negative_polarity         1.111e-01  7.973e-02   1.394 0.163316
title_subjectivity            7.079e-02  2.112e-02   3.352 0.000802 ***
title_sentiment_polarity      8.582e-02  1.938e-02   4.427 9.57e-06 ***
abs_title_subjectivity        1.627e-01  2.815e-02   5.779 7.56e-09 ***
abs_title_sentiment_polarity  4.725e-02  3.057e-02   1.546 0.122224
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8803 on 38759 degrees of freedom
Multiple R-squared:  0.102,     Adjusted R-squared:  0.101
F-statistic: 100.1 on 44 and 38759 DF,  p-value: < 2.2e-16
```

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.

n_tokens_title	n_tokens_content
1.089484	3.480953
num_hrefs	num_self_hrefs
1.570075	1.334773
num_imgs	num_videos
1.335343	1.284012
num_keywords	data_channel_is_lifestyle
1.419085	2.216548
data_channel_is_entertainment	data_channel_is_bus
2.676391	5.438064
data_channel_is_socmed	data_channel_is_tech
2.222916	5.799601
data_channel_is_world	kw_min_max
6.577377	1.320688
kw_max_max	kw_avg_max
2.017593	3.258305
kw_min_avg	kw_max_avg
1.352002	1.091655
LDA_00	LDA_01
4.248783	3.328283
LDA_02	LDA_03
4.714078	5.499265
global_subjectivity	global_sentiment_polarity
2.627509	6.919966
global_rate_positive_words	global_rate_negative_words
3.128151	3.333664
avg_positive_polarity	min_positive_polarity
6.316191	1.886986
max_positive_polarity	avg_negative_polarity
4.142578	7.583152
min_negative_polarity	max_negative_polarity
4.857064	2.889795
title_subjectivity	title_sentiment_polarity
2.460250	1.319727
abs_title_subjectivity	abs_title_sentiment_polarity
1.431016	2.636172

3.(iii)

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

data_channel_is_entertainment	data_channel_is_bus
-0.1397226151	-0.1302498416
LDA_02	data_channel_is_world
-0.1007252572	-0.0813457164
n_tokens_content	num_self_hrefs
-0.0485664672	-0.0427299089
min_positive_polarity	data_channel_is_lifestyle
-0.0291756622	-0.0257274243
min_negative_polarity	kw_min_max
-0.0248369391	-0.0200420194
kw_max_max	global_rate_positive_words
-0.0179345047	-0.0159375019
avg_negative_polarity	avg_positive_polarity
-0.0106840244	-0.0101699662
LDA_01	global_rate_negative_words
-0.0096838669	-0.0094393197
data_channel_is_tech	global_sentiment_polarity
-0.0026091740	-0.0005350936
(Intercept)	max_positive_polarity
0.0000000000	0.0038546562
LDA_03	abs_title_sentiment_polarity
0.0051046653	0.0093314791
max_negative_polarity	n_tokens_title
0.0101433113	0.0115414719
kw_avg_max	title_subjectivity
0.0200623030	0.0212603863
title_sentiment_polarity	data_channel_is_socmed
0.0271147652	0.0288338639
abs_title_subjectivity	num_videos
0.0290107613	0.0390636807
num_keywords	num_imgs
0.0506701467	0.0510687027
global_subjectivity	kw_max_avg
0.0573775795	0.0577087543
num_hrefs	LDA_00
0.0687506478	0.0756342375

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

4. (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 log(x/(1-x)).We can assume that all observations Y with the given the X are independent and the model structure is appropriate.

(2) 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:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-1.809e+02	1.262e+03	-0.143	0.886048	
timedelta	-3.252e-04	7.873e-05	-4.131	3.62e-05	***
n_tokens_title	-2.489e-03	5.962e-03	-0.417	0.676343	
n_tokens_content	1.391e-04	5.053e-05	2.752	0.005915	**
n_unique_tokens	-9.238e-02	4.011e-01	-0.230	0.817858	
n_non_stop_words	6.886e-01	1.215e+00	0.567	0.571012	
n_non_stop_unique_tokens	-7.098e-01	3.399e-01	-2.088	0.036783	*
num_hrefs	1.216e-01	2.212e-02	5.495	3.90e-08	***
num_self_hrefs	-1.881e-02	4.047e-03	-4.648	3.36e-06	***
num_imgs	1.129e-03	1.909e-03	0.591	0.554358	
num_videos	2.763e-03	3.330e-03	0.830	0.406778	
average_token_length	-1.100e-01	5.084e-02	-2.164	0.030446	*
num_keywords	3.300e-02	7.485e-03	4.409	1.04e-05	***
data_channel_is_lifestyle	-1.334e-02	8.758e-02	-0.152	0.878950	
data_channel_is_entertainment	-2.802e-01	5.193e-02	-5.397	6.79e-08	***
data_channel_is_bus	-9.078e-02	8.178e-02	-1.110	0.266974	
data_channel_is_socmed	1.233e+00	9.435e-02	13.063	< 2e-16	***
data_channel_is_tech	5.812e-01	7.996e-02	7.269	3.62e-13	***
data_channel_is_world	1.612e-02	7.784e-02	0.207	0.835957	
kw_max_min	2.941e-06	4.997e-06	0.589	0.556182	
kw_min_max	-6.221e-07	2.332e-07	-2.668	0.007641	**
kw_max_max	-8.409e-07	8.732e-08	-9.630	< 2e-16	***
kw_avg_max	-6.338e-07	1.756e-07	-3.609	0.000307	***
kw_max_avg	-8.244e-05	4.686e-06	-17.592	< 2e-16	***
kw_avg_avg	6.438e-04	2.475e-05	26.009	< 2e-16	***
self_reference_min_shares	6.511e-06	2.726e-06	2.389	0.016908	*
self_reference_max_shares	1.042e-06	1.225e-06	0.851	0.394955	
self_reference_avg_sharess	2.579e-06	3.217e-06	0.802	0.422751	
weekday_is_monday	-1.182e+00	6.679e-02	-17.701	< 2e-16	***
weekday_is_tuesday	-1.303e+00	6.615e-02	-19.701	< 2e-16	***
weekday_is_wednesday	-1.333e+00	6.603e-02	-20.182	< 2e-16	***
weekday_is_thursday	-1.281e+00	6.622e-02	-19.346	< 2e-16	***
weekday_is_friday	-9.840e-01	6.828e-02	-14.411	< 2e-16	***
weekday_is_saturday	1.348e-01	9.184e-02	1.468	0.142234	

```
weekday_is_sunday                    NA         NA      NA       NA
is_weekend                           NA         NA      NA       NA
LDA_00                         1.826e+02  1.262e+03   0.145 0.884993
LDA_01                         1.817e+02  1.262e+03   0.144 0.885567
LDA_02                         1.816e+02  1.262e+03   0.144 0.885623
LDA_03                         1.816e+02  1.262e+03   0.144 0.885593
LDA_04                         1.823e+02  1.262e+03   0.144 0.885183
global_subjectivity            8.549e-01  1.738e-01   4.918 8.75e-07 ***
global_sentiment_polarity     -1.685e-01  3.459e-01  -0.487 0.626272
global_rate_positive_words    -2.186e+00  1.501e+00  -1.456 0.145304
global_rate_negative_words     4.789e+00  2.839e+00   1.687 0.091610 .
rate_positive_words           -2.869e-03  1.187e+00  -0.002 0.998071
rate_negative_words           -6.340e-01  1.196e+00  -0.530 0.595998
avg_positive_polarity          1.399e-01  2.779e-01   0.503 0.614774
min_positive_polarity         -5.705e-01  2.338e-01  -2.440 0.014684 *
max_positive_polarity         -1.929e-01  8.819e-02  -2.187 0.028730 *
avg_negative_polarity         -2.283e-01  2.585e-01  -0.883 0.377126
min_negative_polarity          6.750e-02  9.419e-02   0.717 0.473592
max_negative_polarity          1.582e-01  2.160e-01   0.733 0.463852
title_subjectivity             9.262e-02  5.868e-02   1.578 0.114508
title_sentiment_polarity       1.312e-01  5.142e-02   2.552 0.010710 *
abs_title_subjectivity         2.895e-01  7.603e-02   3.808 0.000140 ***
abs_title_sentiment_polarity   2.173e-02  1.117e-01   0.195 0.845720
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 47288  on 39643  degrees of freedom
Residual deviance: 42542  on 39589  degrees of freedom
AIC: 42652

Number of Fisher Scoring iterations: 8
```

4.(iv)

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

	term	estimate	std.error	statistic	p.value	conf.low	conf.high
1	(Intercept)	-1.809092e+02	1.262390e+03	-0.143306900	8.860478e-01	-3.377101e+03	*NA*
2	timedelta	-3.252312e-04	7.873438e-05	-4.130738852	3.615992e-05	-4.795944e-04	-1.709487e-04
3	n_tokens_title	-2.488753e-03	5.961675e-03	-0.417458723	6.763429e-01	-1.417226e-02	9.198163e-03
4	n_tokens_content	1.390745e-04	5.052788e-05	2.752430282	5.915474e-03	4.091826e-05	2.390173e-04
5	n_unique_tokens	-9.238369e-02	4.011444e-01	-0.230300326	8.178584e-01	-8.777810e-01	6.947670e-01
6	n_non_stop_words	6.885819e-01	1.215368e+00	0.566562309	5.710116e-01	-1.501915e+00	3.761358e+00
7	n_non_stop_unique_tokens	-7.098428e-01	3.399355e-01	-2.088169262	3.678257e-02	-1.376482e+00	-4.389310e-02
8	num_hrefs	1.215710e-01	2.212209e-02	5.495456762	3.897004e-08	7.820926e-02	1.649310e-01
9	num_self_hrefs	-1.881093e-02	4.047358e-03	-4.647706763	3.356456e-06	-2.668436e-02	-1.081152e-02
10	num_imgs	1.128619e-03	1.908895e-03	0.591242233	5.543581e-01	-2.595193e-03	4.889114e-03
11	num_videos	2.762773e-03	3.330338e-03	0.829577449	4.067777e-01	-3.705094e-03	9.353468e-03
12	average_token_length	-1.100375e-01	5.084362e-02	-2.164234242	3.044638e-02	-2.096244e-01	-1.030650e-02
13	num_keywords	3.300475e-02	7.485181e-03	4.409345894	1.036833e-05	1.833821e-02	4.766083e-02
14	data_channel_is_lifestyle	-1.333867e-02	8.758109e-02	-0.152300809	8.789497e-01	-1.844796e-01	1.588658e-01
15	data_channel_is_entertainment	-2.802497e-01	5.193058e-02	-5.396620904	6.790766e-08	-3.820866e-01	-1.785089e-01
16	data_channel_is_bus	-9.077767e-02	8.177734e-02	-1.110058994	2.669736e-01	-2.510149e-01	6.956598e-02
17	data_channel_is_socmed	1.232542e+00	9.435327e-02	13.063054203	5.353604e-39	1.049205e+00	1.419167e+00
18	data_channel_is_tech	5.812143e-01	7.995599e-02	7.269177626	3.616826e-13	4.246235e-01	7.380618e-01
19	data_channel_is_world	1.611773e-02	7.783794e-02	0.207067780	8.359569e-01	-1.363610e-01	1.687775e-01
20	kw_max_min	2.941046e-06	4.997345e-06	0.588521712	5.561822e-01	-6.789770e-06	1.267108e-05
21	kw_min_max	-6.221242e-07	2.332193e-07	-2.667550359	7.640644e-03	-1.074490e-06	-1.593716e-07

22	kw_max_max	-8.408944e-07	8.732323e-08	-9.629674847	5.991878e-22	-1.012413e-06	-6.700901e-07
23	kw_avg_max	-6.338294e-07	1.756281e-07	-3.608928792	3.074640e-04	-9.778322e-07	-2.893486e-07
24	kw_max_avg	-8.243580e-05	4.686047e-06	-17.591757333	2.849037e-69	-9.160852e-05	-7.320467e-05
25	kw_avg_avg	6.437802e-04	2.475220e-05	26.009010189	3.916374e-149	5.953820e-04	6.924121e-04
26	self_reference_min_shares	6.511229e-06	2.725837e-06	2.388707857	1.690774e-02	1.105040e-06	1.182025e-05
27	self_reference_max_shares	1.042297e-06	1.225272e-06	0.850665165	3.949554e-01	-1.202593e-06	3.655155e-06
28	self_reference_avg_sharess	2.579122e-06	3.217237e-06	0.801657604	4.227510e-01	-3.845139e-06	8.870603e-06
29	weekday_is_monday	-1.182305e+00	6.679463e-02	-17.700604268	4.148322e-70	-1.314548e+00	-1.052643e+00
30	weekday_is_tuesday	-1.303092e+00	6.614510e-02	-19.700501279	2.134958e-86	-1.434099e+00	-1.174740e+00
31	weekday_is_wednesday	-1.332690e+00	6.603439e-02	-20.181752780	1.416339e-90	-1.463486e+00	-1.204561e+00
32	weekday_is_thursday	-1.281118e+00	6.622089e-02	-19.346134421	2.197161e-83	-1.412269e+00	-1.152613e+00
33	weekday_is_friday	-9.840074e-01	6.828077e-02	-14.411193257	4.400415e-47	-1.119081e+00	-8.513500e-01
34	weekday_is_saturday	1.347723e-01	9.183667e-02	1.467521359	1.422343e-01	-4.474419e-02	3.154065e-01
35	weekday_is_sunday	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*
36	is_weekend	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*
37	LDA_00	1.825945e+02	1.262387e+03	0.144642226	8.849934e-01	*NA*	3.378783e+03
38	LDA_01	1.816780e+02	1.262386e+03	0.143916323	8.855665e-01	*NA*	3.377865e+03
39	LDA_02	1.815871e+02	1.262387e+03	0.143844239	8.856235e-01	*NA*	3.377775e+03
40	LDA_03	1.816352e+02	1.262385e+03	0.143882529	8.855932e-01	*NA*	3.377821e+03
41	LDA_04	1.822908e+02	1.262386e+03	0.144401755	8.851832e-01	*NA*	3.378477e+03
42	global_subjectivity	8.548866e-01	1.738365e-01	4.917762895	8.753885e-07	5.143790e-01	1.195843e+00
43	global_sentiment_polarity	-1.684692e-01	3.459463e-01	-0.486980738	6.262720e-01	-8.469366e-01	5.092432e-01
44	global_rate_positive_words	-2.185697e+00	1.500836e+00	-1.456319666	1.453043e-01	-5.125076e+00	7.584833e-01
45	global_rate_negative_words	4.788829e+00	2.838728e+00	1.686962863	9.161049e-02	-7.614883e-01	1.036574e+01
46	rate_positive_words	-2.869308e-03	1.187034e+00	-0.002417208	9.980713e-01	-3.043406e+00	2.126692e+00
47	rate_negative_words	-6.339789e-01	1.195814e+00	-0.530164960	5.959976e-01	-3.684432e+00	1.514646e+00
48	avg_positive_polarity	1.398682e-01	2.779187e-01	0.503270153	6.147743e-01	-4.045752e-01	6.849086e-01
49	min_positive_polarity	-5.704654e-01	2.337903e-01	-2.440073494	1.468427e-02	-1.028119e+00	-1.115918e-01
50	max_positive_polarity	-1.928927e-01	8.819288e-02	-2.187168911	2.873020e-02	-3.657488e-01	-2.002326e-02
51	avg_negative_polarity	-2.283267e-01	2.585207e-01	-0.883204531	3.771258e-01	-7.353692e-01	2.780740e-01
52	min_negative_polarity	6.749892e-02	9.418721e-02	0.716646353	4.735923e-01	-1.172208e-01	2.520074e-01
53	max_negative_polarity	1.581992e-01	2.159659e-01	0.732519476	4.638516e-01	-2.653756e-01	5.812623e-01
54	title_subjectivity	9.261941e-02	5.868492e-02	1.578248888	1.145084e-01	-2.213294e-02	2.079291e-01
55	title_sentiment_polarity	1.312332e-01	5.142348e-02	2.552009935	1.071035e-02	3.025226e-02	2.318539e-01
56	abs_title_subjectivity	2.895352e-01	7.603284e-02	3.808027589	1.400796e-04	1.405776e-01	4.386375e-01
57	abs_title_sentiment_polarity	2.173065e-02	1.116787e-01	0.194581904	8.457203e-01	-1.969871e-01	2.408200e-01

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:

```
lm.pred <- fitted(lm6)
lm.pred
lm.pred[lm.pred <= 0.5] <- 0
lm.pred[lm.pred > 0.5] <- 1
lm.rmse <- rmse(lm.pred, ds$shares)
lm.rmse
```

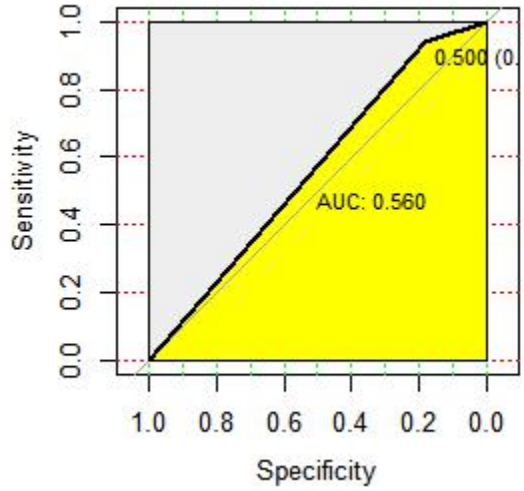
```
[1] 6.548008
```

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.

(2) Logisitic Regression

```
Confusion Matrix and Statistics

          logit.target
lo.pred     0     1
      0  2040  1748
      1  9207 26649

               Accuracy : 0.7237
                 95% CI : (0.7192, 0.7281)
    No Information Rate : 0.7163
    P-Value [Acc > NIR] : 0.000566

                  Kappa : 0.1498

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.18138
            Specificity : 0.93844
         Pos Pred Value : 0.53854
         Neg Pred Value : 0.74322
             Prevalence : 0.28370
         Detection Rate : 0.05146
   Detection Prevalence : 0.09555
      Balanced Accuracy : 0.55991

       'Positive' Class : 0
```

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.
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 C

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.