DATA 7202 Assignment 1 Report

Name: Yupeng Wu
No: 45960600

1. (i)

Generalized Linear Models:

The linear regression models assume that:

- Linearity of the data. The relationship between the predictor (x) and the outcome (y) is assumed to be linear.
- Normality of residuals. The residual errors are assumed to be normally distributed.
- Homogeneity of residuals variance. The residuals are assumed to have a constant variance (homoscedasticity).
- Independence of residuals error terms.

$Y = Y_1, \dots, Y_n$ are independent random variables.
$X = X_1, \dots, X_p$ are a corresponding set of covariate vectors.

1. Response variables $Y_1, \dots, Y_n$ have distributions from the same exponential family:
$$f_{Y_i}(y; \theta_i) = exp\{\frac{a(y)b(\theta_i) + c(\theta_i)}{e(\varphi)} + d(y; \varphi)\}$$
2. A set of covariates $X_1, \dots, X_p$ (each X_i is a $(d+1)$ vector) and associated set of parameters $\beta_0, \dots, \beta_d$ forming linear predictors $X_i^T\beta$. Which refers to
$$Y_i = \beta_1 X_{i1} + \cdots + \beta_p X_{ip} + \varepsilon_i$$
3. A link function $g(\cdot)$ such that
$$g(\mu_i) = X_i^T\beta,$$
Where $\mu_i = E(Y_i)$.

Logistic Regression Models:

The logistic regression models assume that:

- The outcome is a binary or dichotomous variable like yes vs no, positive vs negative, 1 vs 0.
- There is a linear relationship between the logit of the outcome and each predictor variables. Recall that the logit function is
$$logit(p) = log(\frac{p}{1-p})$$
where p is the probabilities of the outcome.
- There are no influential values (extreme values or outliers) in the continuous predictors
- There is no high intercorrelations among the predictors.

When $Y_1, \dots, Y_n$ is binary, with $P(Y_i = 1|X_i) = p_i$, we use
$$Y_i|X_i \sim Bernoulli(p_i)$$
Where
$$X_i^T\beta = log(\frac{p_i}{1-p_i})$$

As $\mu_i = p_i$, the link function $g(\cdot)$ is

$$g(\mu_i) = log\left(\frac{\mu_i}{1-\mu_i}\right) = logit(\mu_i)$$

1. (ii)

Maximum likelihood estimation (MLE)

Using a Newton-Raphson procedure with iteratively reweighted least squares to find the MLE in the GLM.

Let $X_1, \dots, X_n$ be an iid sample from a population with pdf or pmf $f(x;\ \theta = (\theta_1, \dots, \theta_k)$. Then the likelihood function is:

$$L_n(\theta; x) = L_n(\theta_1, \dots, \theta_k; x_1, \dots, x_n) = \prod_{i=1}^{n} f(x_i;\ \theta_1, \dots, \theta_k)$$

And the log-likelihood by:

$$log\ L_n(\theta; x) = \sum_{i=1}^{n} log(f(x_i; \theta_1, \dots, \theta_k))$$

The estimator $\hat{\theta}$ is a value of θ that maximized $L_n(\theta; x)$.

We compute the MLE by solving

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

We compute the negative of the second order derivative:

$$I(\theta) = -\left(\frac{\partial^2\ log\ L_n(\theta; x)}{\partial\theta\partial\theta^T}\right)_{\theta=\widehat{\theta}}$$

Ordinary least squares (OLS)

OLS minimizes the sum of squares of the difference between Y_i and their expected values. Minimize the expression:

$$\sum_{i=1}^{n}\left(Y_i - X_i^T\beta\right)^2 = (Y - X\beta)^T(Y - X\beta)$$

The optimal solution is:

$$\hat{\beta} = (X^TX)^{-1}X^TY$$

This is the same solution as the maximum likelihood estimator assuming normal errors with constant variance.

1. (iii)

$Y = Y_1, \dots, Y_n$ are independent random variables.

$X = X_1, \dots, X_p$ are a corresponding set of covariate vectors.

The p.d.f of Y_i is

$$f_i(x) = \frac{1}{\sqrt{2\pi}\sigma} exp(-\frac{1}{2\sigma^2}\left(x - \sum_{i=1}^{p}\sum_{j=1}^{p}\beta_j X_{ij}\right)^2)$$

The likelihood function is

$$L(\sigma) = \prod_{i=1}^{n} f_i(Y_i) = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(Y_i - \sum_{i=1}^{p}\sum_{j=1}^{p}\beta_j X_{ij}\right)^2\right)$$

$$= \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n exp\left(-\frac{1}{2\sigma^2}|Y - X\beta|^2\right)$$

In order to minimize $|Y - X\beta|^2$

$$|Y - X\beta|^2 = Y^T Y - 2X^T Y\beta + X^T X\beta^2$$

Then setting the derivatives in each β_i equal to zero

$$-2X^T Y + X^T X\beta = 0$$

$\mathrm{X^T}X$ is a $p \times p$ matrix which is invertible. We get

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

$$\hat{\sigma}^2 = \frac{1}{n}|Y - X(X^T X)^{-1} X^T Y|^2$$

1. (iv)

Dataset: “beaver2” in R

The “activ” variable is categorical variable and the “time” is continuous variable. Using “aov” to compute the interaction between these two variables, we can find out the p-value of the “time:activ” is 9.62e-05, which is lower than 0.05. Since the p-value is very low, the interaction between two variables will have an infect on the regression model.

2.(i)

From the given dataset, we can find out that there are two types of variables: continuous variables and categorical variables. The categorical variables are listed in the following table.

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	weekday_is_tuesday	weekday_is_wednesday
weekday_is_thursday	weekday_is_friday	weekday_is_saturday
weekday_is_sunday	is_weekend	

As we know, in order to reduce the interference of category variables on the model, we should delete a category variable in a category. So that my model ignores “weekday_is_sunday” and “is_weekend”. Besides, the total sum of “LDA_00”, “LDA_01”, “LDA_02”, “LDA_03” and “LDA_04” is supposed to be one. For similar reason, my model ignores “LDA_04”.

As we can see from the summary statistics, some variable has extremely big maximum value according to its mean and has a large standard deviation. This shows that we should detect the outliers in such variables. I performed some boxplot to find out which variable has an outlier. It turns that “n_tokens_content”, “n_non_stop_words” and “n_non_stop_unique_token” have some

outliers. These outliers might have a great influence on the model.

Also, we can find out there is some negative values in “kw_min_min”, “kw_avg_min” and “kw_min_avg”. All the value should be positive so these are error values. There are 22979 rows contain negative “kw_min_min” and I decided to delete this variable from the model. And there are only 833 rows contain negative “kw_avg_min” and 6 rows contain negative “kw_min_avg”. So, I deleted these rows from the dataset.

Finally, perform VIF test. Delete all the variables which their variance inflation factor is larger than 10.

2.(ii)

I am interested in the interaction between “num_videos” and “num_imgs”. Because both images and videos are multimedia data so I want to find out whether they have some interactions.

Set the “num_videos*num_imgs” as a variable for a new linear regression model. From the summary of the “aov” function, the p-value of its estimate coefficients is 0.0641. With this low p-value, I have the confident of adding this to my model for a better performance.

Before adding this “num_videos*num_imgs” to my model, the model summary is:

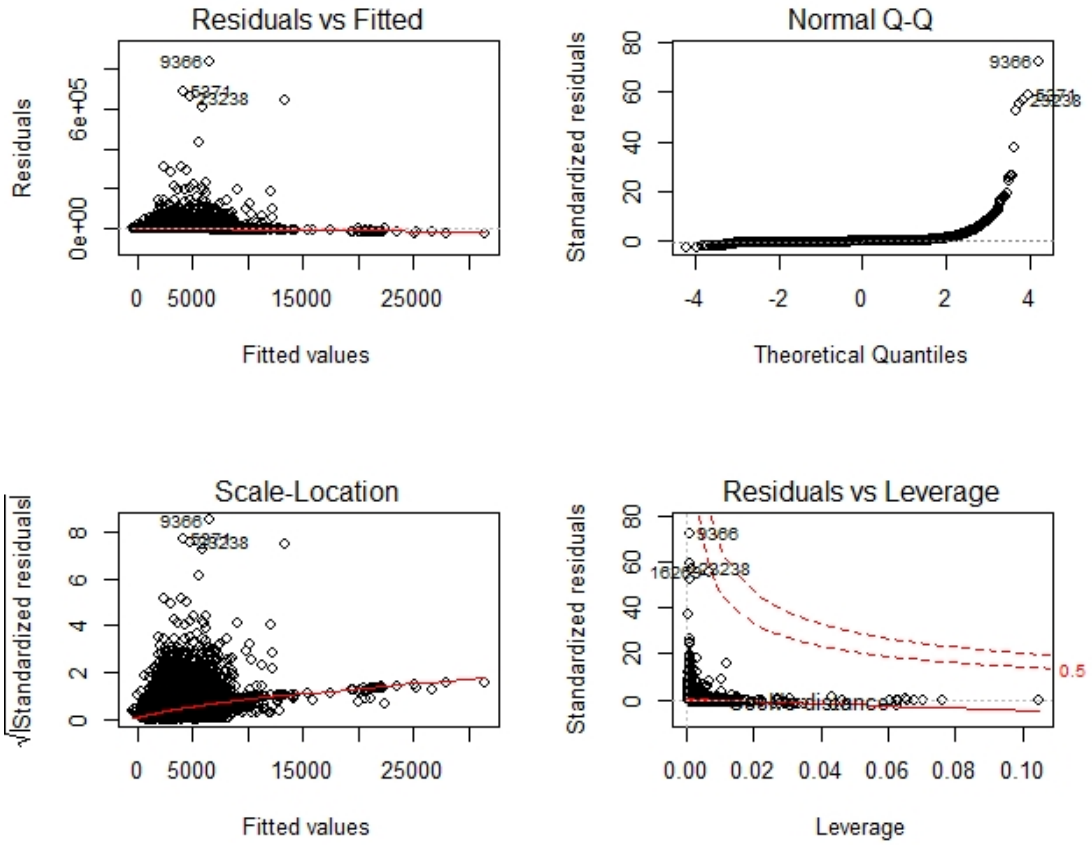
Multiple R-squared	Adjusted R-squared	F-statistic	p-value (<)
0.019	0.01789	17.06	2.2e-16

After adding this new variable to the model, the summary is:

Multiple R-squared	Adjusted R-squared	F-statistic	p-value (<)
0.01903	0.01789	16.7	2.2e-16

There is an increase in R-squared and a decrease in F-statistic, which means this new variable helps improve the model.

2. (iii)



Linearity:

From the Residuals vs Fitted plot, we can find that most of the points match the red line and the red line is approximately horizontal at zero, which shows the data fit the linearity assumption.

Normal distribution:

From the Q-Q plot, the points are not normally distributed because points are not fitted on the straight line.

Homoscedasticity:

From the Scale-Location plot, we can find out the points are located equally along the red line, which can prove the homoscedasticity assumption.

High leverage points:

From the Residual vs Leverage plot, the upper right corner doesn't have any point which means there are no extreme values.

2. (iv)

The variables' variance inflation factor and transportation method are listed in the following table.

Variable name (x)	**VIF**	**Transformation Method** $y(x)$
n_tokens_title	1.089206	x
n_tokens_content	2.187474	$log(x+1)$

num_hrefs	1.589743	x
num_self_hrefs	1.360603	$log(x+1)$
num_imgs	1.425050	x
num_videos	1.228195	x
num_keywords	1.411477	x
data_channel_is_lifestyle	2.289876	x
data_channel_is_entertainment	2.635174	x
data_channel_is_bus	5.475841	x
data_channel_is_socmed	2.257740	x
data_channel_is_tech	6.038189	x
data_channel_is_world	6.603736	x
kw_min_max	1.319244	$log(x+1)$
kw_max_max	2.010585	$log(x+1)$
kw_avg_max	3.189683	$log(x+1)$
kw_min_avg	1.353987	x
kw_max_avg	1.125379	$log(x+1)$
self_reference_min_shares	1.325504	x
self_reference_max_shares	1.395734	x
weekday_is_monday	2.897081	x
weekday_is_tuesday	3.045725	x
weekday_is_wednesday	3.061405	x
weekday_is_thursday	3.025341	x
weekday_is_friday	2.658228	x
weekday_is_saturday	1.780191	x
LDA_00	4.307292	$log(x+1)$
LDA_01	3.668160	$log(x+1)$
LDA_02	4.799763	$log(x+1)$
LDA_03	5.858726	$log(x+1)$
global_subjectivity	2.360444	$log(x+1)$
global_sentiment_polarity	6.946051	x^3
global_rate_positive_words	3.081476	x
global_rate_negative_words	3.259806	x
avg_positive_polarity	5.628448	x
min_positive_polarity	1.913673	x
max_positive_polarity	3.239946	x
avg_negative_polarity	7.708756	x
min_negative_polarity	5.201475	x
max_negative_polarity	2.902986	x
title_subjectivity	2.350436	x
title_sentiment_polarity	1.326015	x
abs_title_subjectivity	1.413176	x
abs_title_sentiment_polarity	2.399035	x

The reason why I use $log(x+1)$ as the transformation form is that this doesn't change the range. For example, the "num_self_hrefs" is ranged from 0 to 116 and $log(0)$ makes no sense. But $log(x+1)$ makes sure that the minimum value for this variable after the transformation is still zero.

While I was attempting to choose the transformations of the variables, my aim is to minimize the p-value and to maximize the R-square.

After I transform all the variables, I also transform "shares" with a log function. The final linear regression model's R-square becomes 0.1187, which is higher than the previous model. With the better R-squared in this linear regression model, these transformations help to improve the model.

3. (i)

The time of publication is unknown before publication, so the "weekday_is_monday", "weekday_is_tuesday", "weekday_is_wednesday","weekday_is_Thursday", "weekday_is_Friday", "weekday_is_Saturday", "weekday_is_Sunday" and "is_weekend" will not be available.

Besides, there will be no share if the article hasn't been published. So, "self_reference_min_shares", "self_reference_max_shares" and "self_reference_avg_sharess" will not be available.

3. (ii)

Coefficients:

	Estimate	**Error**	**t value**	**Pr(>|t|)**	**Confint 2.5%**	**Confint 97.5%**
(Intercept)	6.602e+00	9.211e-02	71.678	2e-16	6.421388e+00	6.782446e+00
n_tokens_title	9.027e-03	2.189e-03	4.124	3.72e-05	4.737441e-03	1.331747e-02
n_tokens_content	-3.912e-02	6.680e-03	-5.856	4.78e-09	-5.221385e-02	-2.602646e-02
num_hrefs	4.103e-03	4.766e-04	8.608	2e-16	3.168496e-03	5.036768e-03
num_self_hrefs	-5.434e-03	7.498e-03	-0.725	0.468615	-2.013161e-02	9.262728e-03
num_imgs	4.791e-03	6.353e-04	7.542	4.71e-14	3.546310e-03	6.036573e-03
num_videos	3.449e-03	1.228e-03	2.808	0.004985	1.041598e-03	5.855912e-03
num_keywords	1.087e-02	2.882e-03	3.772	0.000162	5.221070e-03	1.651864e-02
data_channel_is_lifestyle	-1.233e-01	2.988e-02	-4.127	3.69e-05	-1.818705e-01	-6.473958e-02

data_channel_is_entertainment	-2.545e-01	1.890e-02	-13.462	2e-16	-2.915121e-01	-2.174147e-01
data_channel_is_bus	-2.488e-01	2.868e-02	-8.675	2e-16	-3.049893e-01	-1.925750e-01
data_channel_is_socmed	1.038e-01	2.830e-02	3.666	0.000246	4.829239e-02	1.592386e-01
data_channel_is_tech	2.358e-02	2.775e-02	0.850	0.395412	-3.080546e-02	7.796954e-02
data_channel_is_world	-1.530e-01	2.812e-02	-5.439	5.38e-08	-2.080783e-01	-9.784239e-02
kw_min_max	-5.974e-02	2.786e-03	-21.446	2e-16	-6.520055e-02	-5.428088e-02
kw_max_max	-6.881e-02	1.502e-02	-4.580	4.65e-06	-9.824851e-02	-3.936285e-02
kw_avg_max	2.033e-02	1.593e-02	1.276	0.201917	-1.089526e-02	5.155593e-02
kw_min_avg	2.743e-04	1.126e-05	24.360	2e-16	2.522321e-04	2.963726e-04
kw_max_avg	1.962e-01	1.044e-02	18.787	2e-16	1.757231e-01	2.166608e-01
self_reference_min_shares	1.726e-06	2.555e-07	6.753	1.47e-11	1.224679e-06	2.226384e-06
self_reference_max_shares	4.686e-07	1.247e-07	3.758	0.000171	2.242373e-07	7.130576e-07
weekday_is_monday	-2.170e-01	2.014e-02	-10.771	2e-16	-2.564331e-01	-1.774723e-01
weekday_is_tuesday	-2.888e-01	1.984e-02	-14.555	2e-16	-3.276763e-01	-2.499001e-01
weekday_is_wednesday	-2.903e-01	1.983e-02	-14.637	2e-16	-3.291908e-01	-2.514379e-01
weekday_is_thursday	-2.824e-01	1.987e-02	-14.212	2e-16	-3.213598e-01	-2.434629e-01
weekday_is_friday	-2.099e-01	2.061e-02	-10.186	2e-16	-2.502985e-01	-1.695134e-01
weekday_is_saturday	1.110e-02	2.459e-02	0.451	0.651846	-3.710585e-02	5.929833e-02
LDA_00	2.815e-01	4.704e-02	5.985	2.18e-09	1.893318e-01	3.737220e-01
LDA_01	-1.381e-01	5.039e-02	-2.741	0.006130	-2.368773e-01	-3.934709e-02
LDA_02	-3.287e-01	4.672e-02	-7.036	2.01e-12	-4.202902e-01	-2.371448e-01

LDA_03	-5.504e-02	3.601e-02	-1.529	0.126391	-1.256181e-01	1.553791e-02
global_subjectivity	5.524e-01	8.736e-02	6.323	2.60e-10	3.811262e-01	7.235750e-01
global_sentiment_polarity	-8.137e-01	4.951e-01	-1.644	0.100254	-1.784073e+00	1.566221e-01
global_rate_positive_words	-6.025e-01	3.495e-01	-1.724	0.084707	-1.287509e+00	8.246927e-02
global_rate_negative_words	-1.000e+00	5.270e-01	-1.898	0.057757	-2.032983e+00	3.289969e-02
avg_positive_polarity	-2.204e-02	8.834e-02	-0.249	0.803023	-1.951934e-01	1.511205e-01
min_positive_polarity	-3.782e-01	8.429e-02	-4.487	7.25e-06	-5.434099e-01	-2.129910e-01
max_positive_polarity	1.265e-02	3.303e-02	0.383	0.701747	-5.208315e-02	7.737849e-02
avg_negative_polarity	-9.706e-02	9.090e-02	-1.068	0.285642	-2.752306e-01	8.110993e-02
min_negative_polarity	-6.051e-02	3.373e-02	-1.794	0.072854	-1.266314e-01	5.607660e-03
max_negative_polarity	1.342e-01	7.807e-02	1.719	0.085583	-1.880092e-02	2.872517e-01
title_subjectivity	5.934e-02	2.094e-02	2.835	0.004592	1.830783e-02	1.003753e-01
title_sentiment_polarity	8.127e-02	1.915e-02	4.244	2.20e-05	4.373766e-02	1.188093e-01
abs_title_subjectivity	1.456e-01	2.788e-02	5.223	1.77e-07	9.097476e-02	2.002636e-01
abs_title_sentiment_polarity	4.532e-02	3.035e-02	1.493	0.135463	-1.417851e-02	1.048142e-01
num_imgs:num_videos	-2.297e-04	1.923e-04	-1.195	0.232215	-6.065141e-04	1.471404e-04

Residuals:

Min	1Q	Median	3Q	Max
-7.0629	-0.5521	-0.1633	0.3920	6.0071

Multiple R-squared	Adjusted R-squared	F-statistic	p-value (<)
0.1187	0.1177	116	2.2e-16

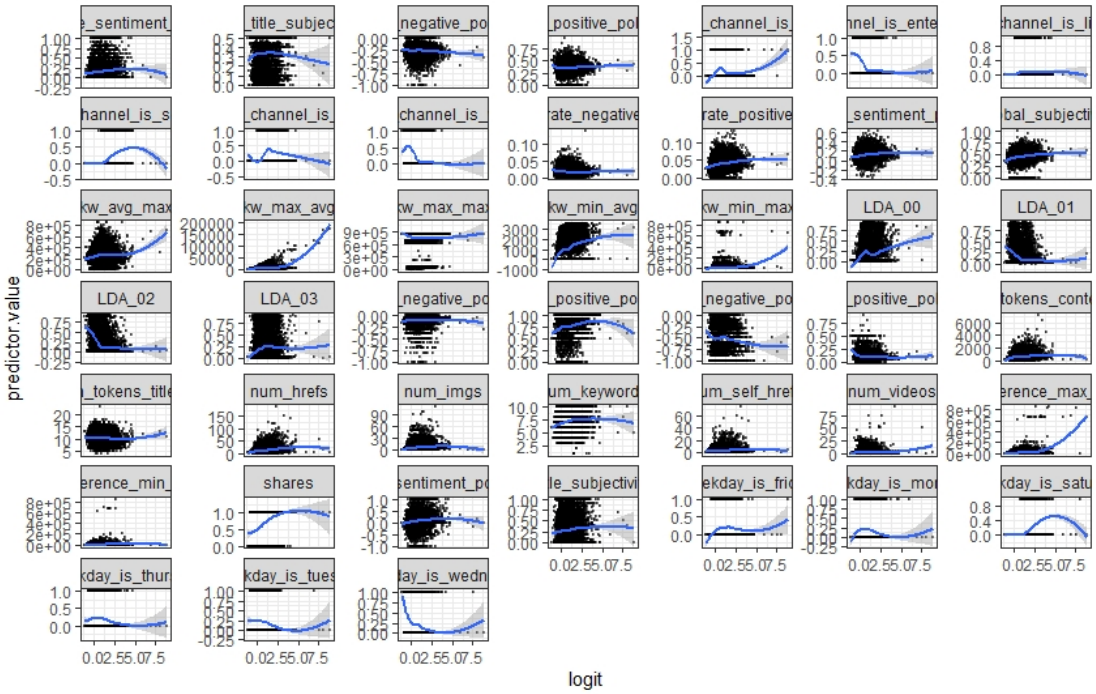
3. (iii)

From the standardized coefficients of the linear regression model. The two most significant slop parameters are “kw_min_avg” and “kw_min_max”. Their estimates are 0.334807154 and -0.288703067.

4. (i)

(In this question, all the plots are made by a sample of 10000 from the original dataset)

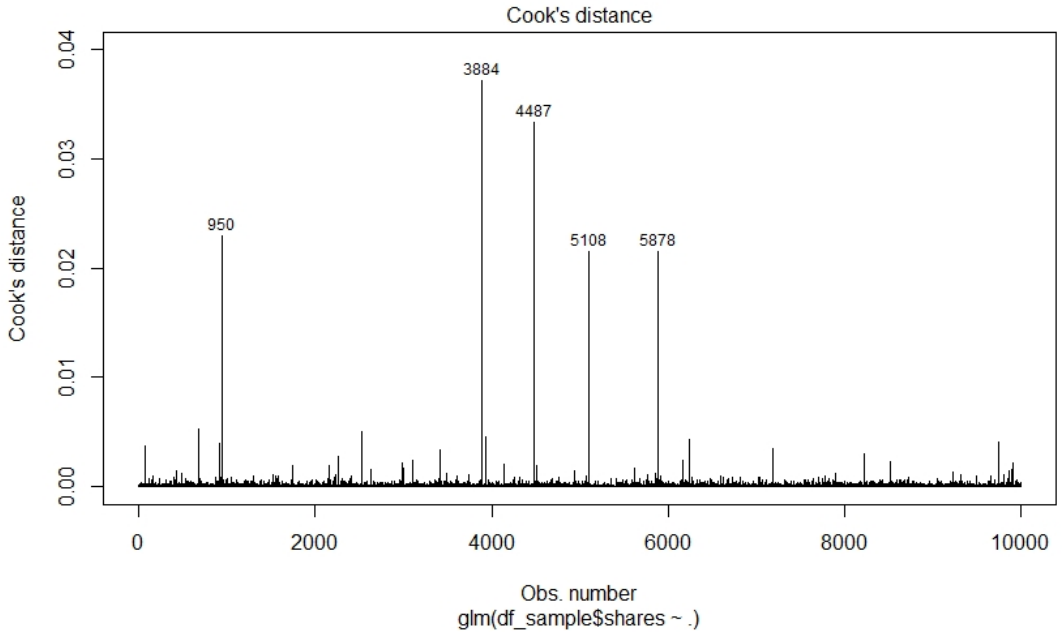
Linearity assumption:



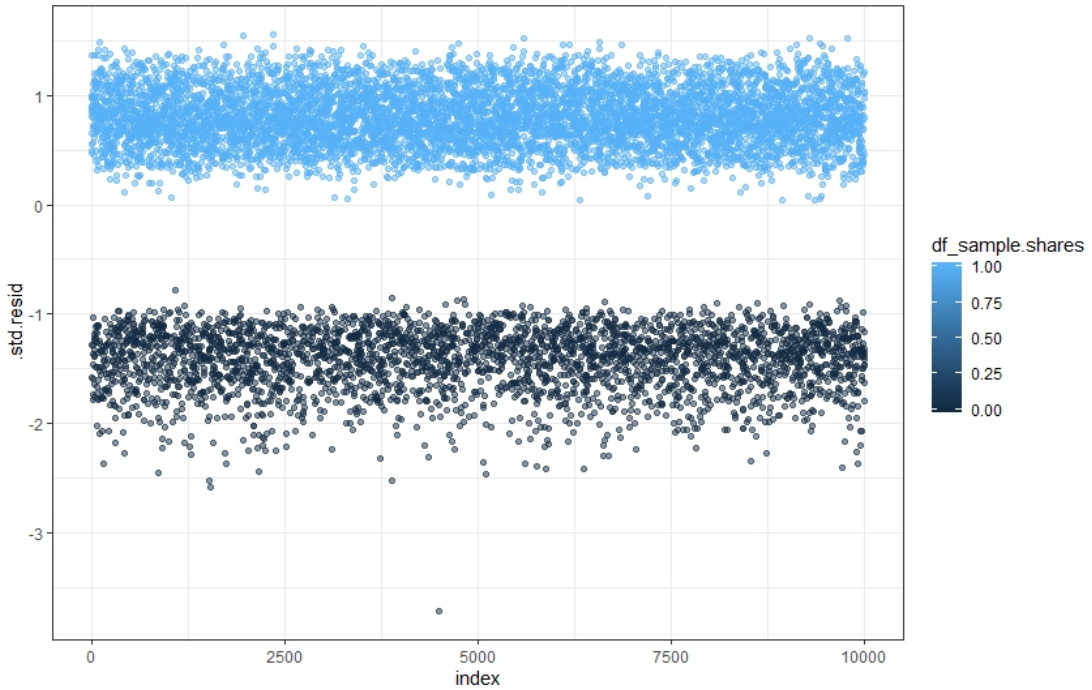
The smoothed scatter plots show that most of the variables are not so linearly associated with the “shares” outcome in logit scale. This model still needs some transformations.

Influential values:

By visualizing the Cook's distance values, I labeled the top 5 largest values in this sampled data:



Not all outliers are influential observations. So, we need to check the standardized residual error:



Filter potential influential data points with $abs(.std.res) > 3$. We can find out there is one influential point in this sample dataset. We need delete this row from the dataset.

Multicollinearity:

Multicollinearity corresponds to a situation where the data contain highly correlated pr edictor variables. Using R function $vif()$ to compute the variance inflation factors:

Variables	VIF
n_tokens_title	1.082689
n_tokens_content	2.274245
num_hrefs	1.606108
num_self_hrefs	1.444862
num_imgs	1.397468
num_videos	1.238083
num_keywords	1.414845
data_channel_is_lifestyle	2.408923
data_channel_is_entertainment	2.968919
data_channel_is_bus	6.335027
data_channel_is_socmed	1.759353
data_channel_is_tech	5.673780
data_channel_is_world	7.661875
kw_min_max	1.329530
kw_max_max	2.135288
kw_avg_max	3.296159
kw_min_avg	1.372637
kw_max_avg	1.151887
self_reference_min_shares	2.157653
self_reference_max_shares	2.243924
weekday_is_monday	3.761741
weekday_is_tuesday	4.141289
weekday_is_wednesday	4.208495
weekday_is_thursday	4.140024
weekday_is_friday	3.383684
weekday_is_saturday	1.652677
LDA_00	4.419670
LDA_01	4.158484
LDA_02	5.250756
LDA_03	6.357658
global_subjectivity	2.505051

global_sentiment_polarity	6.889925
global_rate_positive_words	3.244281
global_rate_negative_words	3.192553
avg_positive_polarity	5.853665
min_positive_polarity	1.926483
max_positive_polarity	3.451019
avg_negative_polarity	8.124184
min_negative_polarity	5.349909
max_negative_polarity	3.124473
title_subjectivity	2.433952
title_sentiment_polarity	1.229667
abs_title_subjectivity	1.465249
abs_title_sentiment_polarity	2.297644

As we can see from the table, all the variance inflation factors are smaller than 10.

4. (ii)

Here I chose to use the same transportation method as I did in linear regression model. Because after using the formal transformation, from the model summary I found out all the p-value are less than 0.05, which I think it is doing a great job. So, I did no further transformations.

4. (iii)

	Estimate	**Std. Error**	**z value**	**Pr(>|z|)**	**Confint 2.5%**	**Confint 97.5%**
(Intercept)	1.169e-01	2.908e-01	0.402	0.687709	-4.561318e-01	6.848247e-01
n_tokens_title	-6.512e-04	6.881e-03	-0.095	0.924603	-1.413386e-02	1.283912e-02
n_tokens_content	-4.469e-02	2.095e-02	-2.133	0.032946	-8.577869e-02	-3.642235e-03
num_hrefs	1.030e-02	1.702e-03	6.050	1.45e-09	6.990903e-03	1.366362e-02
num_self_hrefs	-1.375e-02	2.380e-02	-0.578	0.563438	-6.043033e-02	3.288071e-02
num_imgs	5.387e-03	2.015e-03	2.674	0.007504	1.463042e-03	9.363539e-03

num_videos	-2.366e-03	3.823e-03	-0.619	0.535925	-9.808717e-03	5.188593e-03
num_keywords	3.567e-02	9.151e-03	3.897	9.72e-05	1.773681e-02	5.361049e-02
data_channel_is_lifestyle	1.895e-02	1.001e-01	0.189	0.849838	-1.765147e-01	2.158828e-01
data_channel_is_entertainment	-3.937e-01	5.966e-02	-6.599	4.15e-11	-5.107312e-01	-2.768378e-01
data_channel_is_bus	-2.413e-01	9.381e-02	-2.573	0.010090	-4.251907e-01	-5.745009e-02
data_channel_is_socmed	1.214e+00	1.093e-01	11.110	< 2e-16	1.002025e+00	1.430507e+00
data_channel_is_tech	4.312e-01	9.037e-02	4.772	1.82e-06	2.543155e-01	6.085811e-01
data_channel_is_world	-1.293e-01	8.829e-02	-1.465	0.142970	-3.022885e-01	4.380676e-02
kw_min_max	-1.396e-01	8.858e-03	-15.756	< 2e-16	-1.569364e-01	-1.222138e-01
kw_max_max	-2.026e-01	4.818e-02	-4.206	2.60e-05	-2.971512e-01	-1.082774e-01
kw_avg_max	8.031e-02	5.091e-02	1.578	0.114667	-1.943492e-02	1.801202e-01
kw_min_avg	6.966e-04	3.787e-05	18.394	< 2e-16	6.225417e-04	7.709932e-04
kw_max_avg	3.823e-01	3.486e-02	10.966	< 2e-16	3.143390e-01	4.510397e-01
self_reference_min_shares	6.079e-06	1.888e-06	3.220	0.001280	2.513055e-06	9.887555e-06
self_reference_max_shares	2.078e-06	6.339e-07	3.277	0.001048	9.221469e-07	3.419478e-06
weekday_is_monday	-1.198e+00	7.750e-02	-15.453	< 2e-16	-1.351179e+00	-1.047289e+00
weekday_is_tuesday	-1.247e+00	7.671e-02	-16.261	< 2e-16	-1.399510e+00	-1.098696e+00
weekday_is_wednesday	-1.292e+00	7.657e-02	-16.875	< 2e-16	-1.443956e+00	-1.143705e+00
weekday_is_thursday	-1.285e+00	7.667e-02	-16.762	< 2e-16	-1.437067e+00	-1.136434e+00
weekday_is_friday	-9.228e-01	7.927e-02	-11.640	< 2e-16	-1.079711e+00	-7.688604e-01
weekday_is_saturday	1.222e-01	1.051e-01	1.162	0.245154	-8.336852e-02	3.289777e-01
LDA_00	3.713e-01	1.543e-01	2.406	0.016122	6.935045e-02	6.743020e-01

LDA_01	-5.893e-01	1.570e-01	-3.753	0.00017 5	-8.969151 e-01	-2.81350 1e-01
LDA_02	-9.972e-01	1.464e-01	-6.809	9.80e-12	-1.284258 e+00	-7.10167 9e-01
LDA_03	-6.417e-01	1.565e-01	-4.100	4.13e-05	-9.483481 e-01	-3.34815 7e-01
global_subjectivity	8.573e-01	2.727e-01	3.144	0.00166 7	3.231424e -01	1.39211 4e+00
global_sentiment_po larity	-2.941e+00	1.546e+00	-1.902	0.05714 8	-5.943140 e+00	1.30334 4e-01
global_rate_positive _words	-1.083e-01	1.101e+00	-0.098	0.92167 4	-2.264437 e+00	2.05170 1e+00
global_rate_negative _words	-3.624e+00	1.661e+00	-2.181	0.02916 1	-6.873954 e+00	-3.61266 1e-01
avg_positive_polarit y	2.155e-01	2.765e-01	0.779	0.43586 8	-3.263528 e-01	7.57629 9e-01
min_positive_polarit y	-8.453e-01	2.616e-01	-3.232	0.00123 0	-1.357422 e+00	-3.32028 5e-01
max_positive_polari ty	-1.334e-02	1.037e-01	-0.129	0.89763 3	-2.166105 e-01	1.89939 5e-01
avg_negative_polari ty	-8.304e-02	2.872e-01	-0.289	0.77245 9	-6.463722 e-01	4.79377 9e-01
min_negative_polari ty	-7.045e-02	1.067e-01	-0.660	0.50906 1	-2.797182 e-01	1.38535 7e-01
max_negative_polar ity	4.422e-01	2.458e-01	1.799	0.07200 6	-3.978676 e-02	9.23830 9e-01
title_subjectivity	9.072e-02	6.758e-02	1.342	0.17948 3	-4.144511 e-02	2.23505 7e-01
title_sentiment_pola rity	1.637e-01	6.157e-02	2.658	0.00786 0	4.268737e -02	2.84073 4e-01
abs_title_subjectivit y	3.264e-01	8.936e-02	3.652	0.00026 0	1.513284e -01	5.01627 1e-01
abs_title_sentiment_ polarity	1.264e-01	9.814e-02	1.288	0.19772 0	-6.555420 e-02	3.19195 9e-01

After the transformation, the model AIC is 30583. And before the transportation, the model AIC is 30760. This shows that the transportation has a very good effect on the mo del.

4. (iv)

From the standardized coefficients of the linear regression model. The two most significant slop parameters are “kw_min_avg” and “kw_min_max”. Their estimates are 1.566687929 and -1.239621939.

5.

GLM:

RSE	Multiple R-squared	Adjusted R-squared	F-statistic	p-value (<)
0.8721	0.1147	0.1132	78.07	2.2e-16

1. Residual standard error

The RSE corresponding to the prediction error, represents roughly the average difference between the observed outcome values and the predicted values by the model. Here the $RSE = 0.8724$ means that the observed shares deviate from the predicted values by approximately 0.8721 units in average. The formula is $\frac{RSE}{mean(train\$shares)} \approx 11.7\%$, which is low.

2. R-squared and Adjusted R-squared:

In this model, the adjusted R-squared is 0.1132, which is not close to 1. This seems that the regression model did not explain much of the variability in the outcome.

3. F-Statistic

A large F-statistic will correspond to a statistically significant p-value ($p < 0.05$). In our example, the F-statistic equal 78.07 producing a p-value of 2.2e-16, which is highly significant.

Advantages:

The adjusted R-squared represents the proportion of the variance for a dependent variable. And the model has an ability to determine the relative influence of predictor variables to the criterion value.

Disadvantages:

If the data has a very pool linear relationships between its variables, then the linear regression will performed bad.

Logistic Regression Model (with sampled data):

Since the output is binary, we generate a confusion matrix to compute its accuracy, precision, recall and F1.

True\ Predict	0	1
0	914	677
1	2750	7301

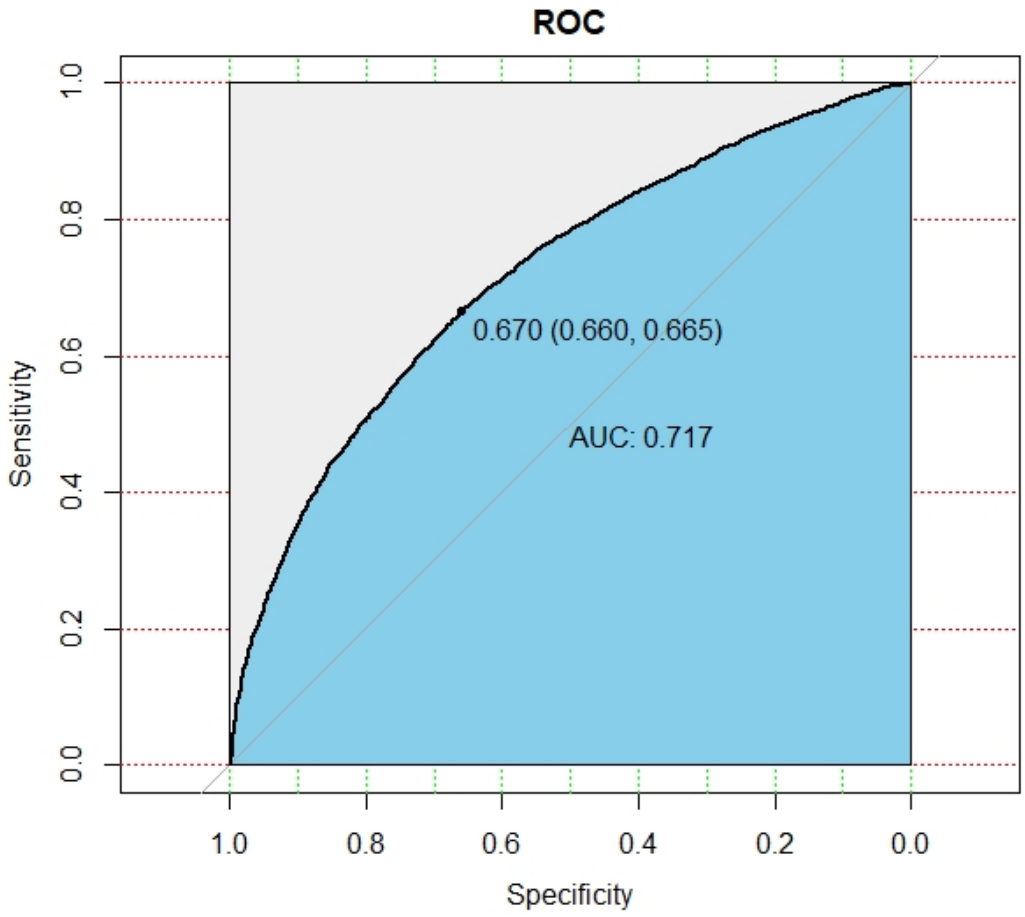


Fig. ROC curves

From this matrix and the ROC plot, we can compare with the results of Fernandes *et al.* with random forest. Then we have:

	Accuracy	Precision	Recall	F1	AUC
R.F.	0.67	0.67	0.71	0.69	0.73
Logistic	0.71	0.73	0.92	0.80	0.72

From the table above, we can clearly see that the logistic model has a higher accuracy with higher precision, higher recall, higher F1 value and almost the same AUC value.

Advantages:

The ROC curve can visualize the model performance. ROC is a probability curve and AUC represent a degree of measure of separability.

Disadvantages:

If the data has too much extreme values, the ROC curve cannot truly reflect the predicted value.

6.

In my opinion, although the multiple linear regression model can predict the “shares” as a number, I still prefer the logistic regression. First reason is that the logistic regression is 71% accurate and the F1 is 0.80, which is very high and makes the model more convince. Second reason is that the adjusted R-square value of the linear regression model is 0.1147, which is pretty low and shows this model doesn’t measure many points.

7. (i)

GLM:

For the multiple linear regression model, you can find the most influential variable by sorting the estimate coefficient. For example, a very large absolute estimate coefficient in my model is “data_channel_is_entertainment”, which is -2.433e-01. Since this estimate is negative, it shows that this variable has a negative effect on “shares”. This means if you want to get more shares, please avoid having a data from entertainment channel.

Logistic Regression Model:

Similar analysis in the logistic regression model.

For example, the “global_rate_negative_words” has the largest absolute estimate coefficient, which is -3.624. This means more negative words in your content will decrease the probability of your article to have more than 1000 shares. By the way, in my logistic regression model the “data_channel_is_entertainment” still has a large negative estimate, which is a same result as multiple linear regression model.

7. (ii)

Highest predicted popularity article (GLM):

Viral Dove Campaign Becomes Most Watched Ad Ever

http://mashable.com/2013/05/20/dove-ad-most-watche

Highest predicted probability of being popular (LRM):

GamerX: World’s First LGBT Gamer Convention Comes to San Francisco

http://mashable.com/2013/07/21/gaymerx

7. (iii)

R command: summary(dataset)

Take the min value if the estimate coefficient is negative, Take the max value if the estimate coefficient is positive.

Note: Some variable’s value has been transformed.

Variable Name	Value in Linear Regression	Value in Logistic Regression
n_tokens_title	23	2
n_tokens_content	9.045	0
num_hrefs	304	304
num_self_hrefs	0	0
num_imgs	128	128
num_videos	91	0
num_keywords	10	10
data_channel_is_lifestyle	0	1
data_channel_is_entertainment	0	0
data_channel_is_bus	0	0
data_channel_is_socmed	1	1
data_channel_is_tech	1	1

data_channel_is_world	0	0
kw_min_max	0	0
kw_max_max	0	0
kw_avg_max	13.65	13.65
kw_min_avg	3613	3613
kw_max_avg	12.606	12.606
self_reference_min_shares	843300	843300
self_reference_max_shares	843300	843300
weekday_is_monday	0	0
weekday_is_tuesday	0	0
weekday_is_wednesday	0	0
weekday_is_thursday	0	0
weekday_is_friday	0	0
weekday_is_saturday	1	1
LDA_00	0.65596	0.65596
LDA_01	0.01802	0.01802
LDA_02	0.01802	0.01802
LDA_03	0.01802	0.01802
global_subjectivity	0.6931	0.6931
global_sentiment_polarity	-0.0610466	-0.0610466
global_rate_positive_words	0	0
global_rate_negative_words	0	0
avg_positive_polarity	0	1
min_positive_polarity	0	0
max_positive_polarity	1	0
avg_negative_polarity	-1	-1
min_negative_polarity	-1	-1
max_negative_polarity	-1	0
title_subjectivity	1	9.072e-02
title_sentiment_polarity	1	1.637e-01
abs_title_subjectivity	0.5	3.264e-01
abs_title_sentiment_polarity	1	1.264e-01

7. (iv)

For the hypothetical article in linear regression model. The article's data channel could not be both "Social Media" and "Tech". So, this article cannot be produced.

For the hypothetical article in logistic regression model. The number of words in the content cannot be zero and its data channel cannot be "lifestyle", "Social Media" and "Tech". This article cannot be produced.