Trending YouTube Videos

By: Brian Grisham, John Chau, Manan Patel, Robert Kaszubski, John DeBrouse

DSC 323 - Data Analysis and Regression
Professor: Nandhini Gulasingam
Final Project Report

Introduction

As one of the world's largest content sharing platforms, YouTube has become ingrained in our culture. YouTube allows anybody to create their own channel and post videos. Each channel can be customized in any way a content creator desires in order to best represent and promote their videos. The goal of any content creator is to build a dedicated viewer base often represented by a channel's subscriber count. These popular channels consist of numerous videos, each with a high viewer count. These highly viewed videos are frequently displayed on YouTube's homepage consisting of a Trending Video section.

We decided to examine what exact factors lead to a YouTube video obtaining a lot of views and being promoted on the Trending page. Glancing at the Trending page at nearly any time of the day, all of the videos posted on there hold close to or over one million views. Occasionally, there is a video that was just recently posted already featured on the page despite a low view count. but we believe that this is the exception, not the rule. Therefore, using our dataset consisting of specific data about videos found on the Trending page, we believe that we would be able to build a model that uses different variables such as the likes, dislikes, category or genre of the video in order to predict the number of views a video will get. Based on our initial background research, nobody knows exactly how the Trending page works or how videos are classified on it. The algorithm utilized by Google, YouTube's parent company, is a mystery. However, according to some experts, the key to a viral or trending video is its shareability. Videos are more likely to be shared amongst people if they are engaging, inspiring, and elicit positive emotions (Pozin). As a result, we predict that the amount of likes a video receives must positively impact the amount of views the video obtains. Our initial hypothesis is that ultimately total like count is the most important factor towards the amount of views and that the genre of video does not matter since it is possible for any type of video to elicit positive emotions.

Many of these videos generate a substantial amount of revenue running advertisements. Google's AdSense is a program that places advertising on videos and websites based on the contents of the video. The more views a video gets, the more revenue the publisher receives. Many people start channels with hopes of building a large viewer base and eventually being able to make a considerable amount of income by uploading videos. This has led to YouTube becoming a full time job for many. As these people are often the ones featured on the Trending video page, our motivation is to deduce exactly what leads to a video going viral and whether viewer count is truly as influential as we believe.

Methodology

Our dataset is Trending YouTube Video Statistics. It was sourced from the website Kaggle.
Link: https://www.kaggle.com/datasnaek/youtube-new

Kaggle provided multiple datasets that contained videos from numerous different regions, we chose to focus only on the dataset with videos trending in the United States. Our data set featured 16 variables and 40,950 observations.

Variable Definitions:

- Video_id - The text/numeric code at the end of a youtube URL
- Trending_date - the date the video was featured on the Trending page
- Title - The title of the video
- Channel_title - The title of the channel that uploaded the video
- Category_id - The category or genre of the video, each number corresponds to an id found in an external json file as tabulated in **M.A**.
- Publish_time - The date and exact time the video was publicly posted
- Tags - The tags associated with the video and used for search engine optimization
- Views - The number of views of the video at the time of trending
- Likes - The number of likes on the video at the time of trending
- Dislikes - The number of dislikes on the video at the time of trending
- Comment_count - The number of comments on the video at the time of trending
- Thumbnail_link - A URL to the thumbnail image
- Comments_disabled - True if comments are disabled on the video, False if comments are enabled on the video
- Ratings_disabled - True if ratings are disabled on the video, False if ratings are enabled on the video
- Video_error_or_removed - True if there was an error or the video was removed, False if there was no video and the video is still published
- Description - Text based description of the video provided by the uploader

Data cleansing:

The data was first inspected to search for missing values or any immediately noticeable discrepancies of which there were none.

Dummy variables were created for any of the categorical or boolean value variables. This included category_id, comments_disabled, ratings_disabled, video_error_or_removed. For the comments, ratings, and video error dummy variables, a true value was represented by a 1 while a

false value was represented by a 0. The dummy variable names for these were com_dis_dummy, rat_dis_dummy, and vid_err_dummy respectively.

For the category_id variable, the dummy variable creation process was more extensive. Due to the plethora of categories provided by the dataset we decided to group certain ones together to broaden the categories and limit the count from 32 specific categories to 6 general categories. The reasoning behind this was that we felt that many of the categories provided overlapped with one another. An example being “Film & Animation” ,“Short Movies", and “Movies”. We would expect similar content to be featured in each and the difference between them to be negligible. The dummy variables are as follows:

- Film - 1 (True) if Category_id is 1, 18, 30-42, or 44, which respectively correspond to categories named “Film & Animation”, “Short Movies”, “Movies”, “Anime/Animation”, “Action/Adventure”, “Classics”, “Comedy”, “Documentary”, “Drama”, “Family”, “Foreign”, “Horror”, “Sci-Fi/Fantasy”, “Thriller”, “Shorts”, and “Trailers”, 0 (False) otherwise
- SciTech - 1 (True) if Category_id is 2, 15, 20, or 28, which respectively correspond to categories named “Autos & Vehicles”, “Pets & Animals”, “Gaming”, and “Science & Technology”, 0 (False) otherwise
- Entertainment - 1 (True) if Category_id is 10, 17, 19, 23, 24, or 43, which respectively correspond to categories named “Music”, “Sports”, “Travel & Events”, “Comedy”, “Entertainment”, and “Shows”, 0 (False) otherwise
- Politics - 1 (True) if Category_id is 25 or 29, which respectively correspond to categories named “News & Politics” and “Nonprofits & Activism”, 0 (False) otherwise
- Blog - 1 (True) if Category_id is 21 or 22, which respectively correspond to categories named “Videoblogging” and “People & Blogs”, 0 (False) otherwise
- Info - 1 (True) if Category_id is 26 or 27, which respectively correspond to categories named “Howto & Style” and “Education”, 0 (False) otherwise

We were also required to transform our dependent variable - views. Upon creating a histogram(**M.B**) of the variable, we realized that the distribution was heavily skewed right and not normal. Around 85% of the dataset was allocated to the first bin of the histogram with the mean view count being around 2.4 million, a big difference from the 648,000 median value. The histogram showed that few videos placed higher than 8.2 million views which was the 95 percentile. In order to transform the variable we used a log transformation and placed the values into a new variable - ln_views. This resulted in an almost symmetric distribution as seen in **M.C**. The mean in this case was almost identical to the median at 13.292 compared to the median at 13.383.

Our next step was removing all of the text variables as they would not be usable in our analysis. These were the video_id, title, channel_title, tags, thumbnail_link, and description.

Additionally we removed all of the variables that were already made into dummy variables including comments_disabled, ratings_disabled, and video_error_or_removed as these were all text based boolean variables and would be unnecessary to keep.

Finally we used simple random sampling in order to narrow down our dataset to 1,000 observations from the original 40,950 as such a high dataset was proving impossible to reliably run in SAS. **Unfortunately, this causes us to get inconsistent results. When we did the analysis, we got the outputs that were shown in the appendix but we get slightly different results when running the code later which we think is caused by the fact that a random sample is chosen every time. Therefore, if we had used the original dataset with 40,950 observations, we may have gotten more consistent results. However, as stated previously, this is impossible to reliably run in SAS.**

Model Approach:

Our approach began with data exploration (detailed below). This consisted of boxplots being generated with each of the categorical or boolean variables as well as scatterplots being generated with each of the numerical variables against ln_views, to try and gather initial patterns and trends. We searched for any association between our variables and ln_views using plots and the correlation matrix (**DE.G)**

Following our data exploration we generated our first full regression model along with the VIF and influential points. We dealt with any outliers that may have popped up and explored whether or not there were any issues with multicollinearity. From there the data was split into a training and testing set (a 75/25 split) and further models were generated using four different selection methods (forward, backward, stepwise, adj-R2).

The top three of these models were determined using their Adj-R2 values. Each of these models were then thoroughly analysed looking at the constant variance, independence, linearity, and normality using the residual plots generated.

From there a top model was picked.

The model was tested using the validation techniques. The ln_views were predicted for the testing dataset and was analysed how accurate the model is. This was determined by examining the rmse value, mae value, R^2 value, yhat, and other statistical values.

Data Exploration

Boxplots were used by plotting each of the categorical variables and boolean dummy variables against ln_views as depicted.

Looking at plot **DE.A**, we see that views between videos with comments enabled or disabled average out to be about the same. The range however is a bit different with many more outliers when comments are enabled with data being expected to fall into the 9-17 range while for comments disabled the range becomes wider from 7-18 resulting in nearly no outliers. The difference between the 1st and 3rd quartile is also much smaller when comments are enabled.

Looking at plot **DE.B**, we see that views between videos with ratings enabled or disabled had about the same median, however the average views for videos with ratings enabled was higher. The first and third quartiles were both higher as well and the overall range was much lower in the comments enabled plot. This suggests that typically views with ratings enabled are expected to receive more views than videos with ratings disabled. The range of the boxplot when ratings are disabled covers nearly all of the values and has less outliers while like in plot **DE.A**, the enabled has a tight range and more outliers.

Looking at plot **DE.C**, we see that views between videos that had errors or were removed and those that were similarly proportioned as the previous two plots. The averages, medians, 1st and 3rd quartiles were almost identical. We can assume that this variable likely does not play a big factor into the views video received.

In plots **DE.D**, we created boxplots for the six category_id dummy variables. Using these boxplots, we can obtain a good idea as to which of these variables are likely significant.

The first chart here shows the distributions for those in the film category and those not in the film category. The mean and median for those in the film category is slightly higher than those not in the film category, which means that a video that fits our film dummy variable is slightly more likely to have more views. However, the difference between the centers of distribution in these two boxplots is not very large so we cannot be certain as to whether or not being in the film category has a significant impact on views.

The second chart shows the distributions for those in the scitech category and those not in the scitech category. Similar to film, the two boxplots have means and median relatively close to one another so no definite claim towards significance can be made. The mean and median for

videos in the scitech category is slightly lower than those not in scitech so our initial inkling is that scitech has a slightly negative effect on the number of views.

The third chart shows the distributions for those in the entertainment category and those not in the entertainment category. Here, the difference in the centers of distribution is greater than the previous two variables. However, the difference is still not very drastic. The mean and median is higher for those in the entertainment category, which indicates that being in the entertainment category likely increases the amount of views, but we cannot be very certain of this claim as the difference is not that large.

The fourth chart shows the distributions for those in the politics category and those not in the politics category. Here, the difference between the distribution centers appears quite large, with the mean and median for those in the politics category being below the 25th percentile of those not in the category. From this observation, we can conjecture that being in the politics category significantly reduces the likelihood of attaining views. However, we cannot be very confident with this claim since there are many outliers present in the non-politics distribution.

The fifth chart shows the distributions for those in the blog category and those not in the blog category. Here, the mean and median for both distributions are nearly equal. Therefore, it appears that whether or not a video is in the blog category has no impact on the amount of views the video obtains. As a result, blog is likely insignificant since it has no impact on the amount of views.

The sixth chart shows the distributions for those in the info category and those not in the info category. The mean and median for both distributions are about the same with those in the info category having a slightly lower mean and median. Therefore, we can conjecture that being in the info category possibly has a slightly negative impact on the number of views a video receives. However, this claim remains uncertain since the difference between the two means and medians is quite small and there are outliers present in both distributions.

Next, scatterplots were used with each of the numeric variables. We plotted likes, dislikes, comment_count against ln_views.

Looking at plots in **DE.E**, we saw how similar each of them came out. The majority of the data hovered closer to the 0 mark than any other number. Only a handful of videos amassed over 1,000,000 likes, or over 50,000 dislikes, or a comment count over 100,000. We do learn that most trending videos have a considerably larger amount of likes than dislikes. However, there is no linear association present.

After removing outliers and running again as seen in plots **DE.F**. The scatterplots remained nearly the same except covering a smaller upper range. However, data still remained clustered near the 0 mark on all of the plots. However it does become clear that videos with the highest amounts of views tend to have the most likes, dislikes, or comments. But at the same time, a video with a low amount of views can have the same amount as one with a much higher amount.

Histograms as explained in the Data Cleansing section above, and pictured in **M.B** and **M.C** were used to look at the distribution of the view count. The view count had to be transformed using a log transformation to achieve a normal and symmetric distribution.

The correlation matrix seen in **DE.G**, shows that there is no issue of multicollinearity since no variables have correlation coefficients greater than 0.9. Also, we see that there are no variables that are strongly correlated with ln_views. In fact, the variable with the strongest correlation to ln_views is likes with a correlation coefficient of 0.49, and this 0.49 can only be considered a moderate correlation. Looking at ln_views versus the other variables, we see that likes, dislikes, comment_count, entertainment, and politics are all significant in relation to ln_views since their p-values are <.0001.

Full Regression Model

After applying a logarithmic transformation and conducting preliminary analysis, a full regression model with all the independent variables was created. The results in **FM.A** allow us to make several key observations.

First, we can see that info, rat_dis_dummy, and vid_err_dummy have parameter estimates of 0 and do not have any of their other statistics calculated. This occurrence happened for rat_dis_dummy and vid_err_dummy since all of the observations in the sample dataset have values of 0 for both rat_dis_dummy and vid_err_dummy. Since every observation in the dataset has a 0 for rat_dis_dummy and vid_err_dummy, there is no variation in these variables, and in effect, these variables act like constants. Regardless of whether a video attains a lot of views or very little views, rat_dis_dummy and vid_err_dummy remain at 0. Therefore, both rat_dis_dummy and vid_err_dummy have no predictive power and are not significant variables. While rat_dis_dummy and vid_err_dummy have parameter estimates of 0 since all observations in the dataset have values of 0 for those two variables, info has a parameter estimate of 0 since info is a linear combination of the other category dummy variables. Above the parameter estimates table, the output says that info has been set to 0 since it is a linear combination of the intercept, film, scitech, entertainment, politics, and blog. Since each observation has a value of 1 for either info, film, scitech, entertainment, politics, or blog, the variables form a linear combination with one another, and essentially, SAS has selected info to be the baseline for the category dummy variables by assigning it a parameter estimate of 0. Thus, although info has a parameter estimate of 0, we cannot discount it and conclude that it is insignificant.

Second, from **FM.A**, we can see that there is no issue with collinearity. Each variable has a variance inflation factor, VIF, less than 10, with the highest value being 7.17. Also, each variable has a tolerance value greater than 0.1, with the lowest value being 0.14. A VIF greater than 10 or a tolerance value less than 0.1 suggests collinearity, and since none of the variables meet either of these criteria, we can conclude that there is no issue with collinearity, which reinforces our findings from the correlation matrix.

Lastly, from **FM.A**, we can see that the full model has an adjusted r-squared of .2935. This value is not great for an adjusted r-squared. However, it is a good starting point as it will improve with the removal of outliers, influential points, and insignificant predictors.

The results in **FM.B** contain the studentized residuals and Cook's distance for the sample dataset, which allow us to identify any outliers and influential points. Any datapoint with a red arrowhead is considered to be an outlier, and any datapoint with a blue arrowhead is considered to be an influential point. Therefore, from the output, we discover that observations 1, 40, 57, 128, 129, 145, 147, 155, 156, 241, 332, 342, 393, 401, 453, 516, 556, 571, 578, 586, 768, 875,

and 994 are either outliers, influential points, or both, and these 23 observations were removed from the dataset for having too much of an impact on the model.

In **FM.C**, we can see how much of an impact the removal of these datapoints has on the model. Here, the adjusted r-squared is .4012, which is a considerable improvement from our previous adjusted r-squared of .2935. By removing these datapoints, our model improved drastically, and it will likely continue to improve with the removal of insignificant variables.

Variable Selection Methods

Before proceeding to the model selection methods, we first separated the dataset into 75% training set and 25% testing set. We then created another data table which contained a new column, new_y. This new variable has the values of ln_views for those selected for the training set. The observations selected for the testing set had an empty new_y value meaning it was left blank. This was done in this manner so that it is easier to run validation techniques on the models.

Now that our data is divided into training and testing, we will now use variable selection methods on the training set. To determine which predictors form the best model, we used four variable selection methods. The tables labeled **VS.A**, **VS.B**, **VS.C**, and **VS.D** respectively show the results for the backward, forward, stepwise, and adjusted r-square selection methods. The following shows the predictors selected from the first three methods as well as their accompanying r-square values.

- Backward: film, entertainment, politics, com_dis_dummy, likes, dislikes. R-square = .4265
- Forward: film, scitech, politics, info, com_dis_dummy, likes, dislikes, comment_count. R-square = .4289
- Stepwise: film, scitech, politics, info, com_dis_dummy, likes, dislikes. R-square = .4282

From these results, we can obtain a good idea as to which variables are likely significant. Film, politics, com_dis_dummy, likes, and dislikes are likely significant since they appear in each selection method. Meanwhile, entertainment, scitech, info, and comment_count may or may not be significant since they appear in at least one, but not all, selection methods.

To provide a better idea as to which model is the best, the adjusted r-square selection method was utilized, and using the results from this selection method as well as the other methods, three top models were selected for further analysis. **VS.D** displays the models with the best adjusted r-square values. Two models tie for the best adjusted r-square value, and both models are selected for further analysis. The first model has an adjusted r-square of .4227 with film, scitech, politics, info, com_dis_dummy, likes, and dislikes as its predictors. This model is the same model selected using the stepwise approach, which is further evidence for this model being one of the top models. The second model also has an adjusted r-square of .4227, but it has film, entertainment, politics, blog, com_dis_dummy, likes, and dislikes as its predictors. This model has many of the same predictors as the previous model with the exception of entertainment and blog in place of scitech and info. Although this model was not selected by any of the other selection methods, it is quite similar to the other top model selected, and even has the same adjusted r-square and r-square as that model, and should be selected for further analysis.

Another two models tie for third in the adjusted r-square selection method. Both have an adjusted r-square of .4226 and r-square of .4289. The first model here has film, entertainment, politics, blog, com_dis_dummy, likes, dislikes, and comment_count as its predictors while the second model has film, scitech, politics, info, com_dis_dummy, likes, dislikes, and comment_count as its predictors. The difference between these two models is that the first one has entertainment and blog whereas the second one has scitech and info. The second model here was chosen for further analysis since scitech and info were a part of the forward and stepwise selections while entertainment only appeared in the backward selection and blog did not appear in any selection method. From the other selection methods, we are more confident that scitech and info are significant compared to entertainment and blog. Also, the second model here is the same model chosen for the forward selection method, which is further evidence for this model being one of the top models.

The three models selected for further analysis are shown below.

M1: film scitech politics info com_dis_dummy likes dislikes
M2: film entertainment politics blog com_dis_dummy likes dislikes
M3: film scitech politics info com_dis_dummy likes dislikes comment_count

Model 1 Regression

As shown by **M1.A** you can see that for most variables' normality will not matter. This is because they are set as true or false. Thus, I looked at the parameter estimate and p value. The p-value when looking at film and dislikes are over the .05 accepted margin. The rest are within the acceptable range. However, there is a clear pattern for likes and dislikes. A majority of videos are clumped up closer to 0 likes/dislikes regardless of views. Thus, this fails the normality test. When looking at the likes and dislikes plot, their data points do not disperse around the zero line, and form a clear pattern. Thus, this fails the constant variance and independence test. However, this could be because of outliers, and would be worth checking after they are removed. I believe it is also worth noting that there are very few videos that have comments disabled. This could lead to exaggerated results from such a small sample of disabled comments. Film, comments disabled, likes, and dislikes all have positive correlations based on parameters estimates, while scitech, politics, info have negative correlations.

Based on image **M1.B** and looking at studentized and cook's D there appears to be a lot of outliers. I will remove data based on Cook's D 0.010 and studentized residuals -3/3. I found 24 different sets of potential outliers that will be removed following my criteria.

Comparing **M1.C to M1.D**, Removing the outliers appeared to have a very small positive influence on our data. Thus, it is not worth keeping the current model, and is better using the original. The change in R2 went from .4281 to .4402. That is only a .0121 difference, which is negligible.

Looking at the original model shown in **M1.C**, it is shown that there are 2 predictors with a p-value above .05. Therefore, it is better to remove it and recreate the model. **M1.E** shows the new model, and all the p-values look to be good. Also it should be noted that the R-Squared dropped from .4402 to .43 from losing the insignificant predictors. The f-value is 106.23 with p-value of <.0001.

To confirm model adequacy, I will write down the test hypotheses.

Ho: Bj = 0

Ha: Bj does not = 0

F-Value = 106.23

P-value <0.0001

Conclusion: P-value is less than alpha (.05), therefore we can reject Ho.

This means there is at least 1 predictor that is significantly associated with Y. There is strong support.

As shown in **M1.F**, likes has the biggest impact by far with it being at .629 correlation. There seems to be no problem of multicollinearity. The only fields that are closely related to views(the y variable) are likes, and the rest are as weakly correlated. No variables are closely correlated to each other outside of views. I would also like to point out that because of how strong likes have an impact on views, it seems like **YouTube's algorithm that determines what's trending or what to recommend might closely depend on the total amount of likes. Thus, the total views would increase.**

M1.F

The final model equation would be ln_views = 13.08929 - .45986scitech - .98624politics - .35861info + .96513com_dis_dummy + .00000664likes This means politics, scitech, info, are less likely to get views. While comments disabled, and likes will likely get more views.

However, because we used log transformation, we must now backtranform them using (e^x)-1 and * 100 to turn it into a percent so we will be able to understand it.

scitech = ((e^.45986)-1) *100 = If true, 58.38% decrease in views

politics = ((e^.98624)-1) *100 = If true, 168.11% decrease in views

info = ((e^.35861)-1) *100 = If true, 43.13% decrease in views

com_dis_dummy = ((e^.96513)-1) *100 = If true, 162.51% increase in views

likes = ((e^.00000664)-1) *100 = Per like, there will be a 0.000664% increase in views

This model ended with a R2 of .43 and adj R2 of .426 which means this model is weak to use, and is still far away from the 1.0 you would strive for. The adj r2 is close to the r2 which means the added variables are significant and worth using. The model's data had clear patterns, so transformations should probably be used.

Model 2 Regression

As shown in **M2.A**, film, entertainment, politics, blog, com_dis_dummy (whether comments are disabled) fail the assumption of constant variance because there is no random spread pattern at all since these are dummy variables. These variables also fail the assumption of independence because there is a clear pattern to the spread (all points fall on either 0 or 1). These variables also fail the assumption of linearity because there is no straight line to the pattern of the spread.

Likes and dislikes also fail the assumption of constant variance because there is no random spread pattern. These variables also fail the assumption of independence because there is a clear pattern to the spread (most points falling on zero and then fanning out down and to the right). These variables also fail the assumption of linearity because there is no straight line to the pattern of the spread.

The studentized vs predicted values residual plot also fails the assumptions of constant variance, and independence because the points are not randomly distributed over the zero (horizontal) line.

The model also fails the normality assumption because as shown in M2.A, the normality plot does not show the points falling on a straight line connecting the lower left to the upper right of the normality plot.

Based on the findings in **M2.B**, the outliers are any data points above 3 or below -3 on the studentized residual; these are data points 338, 320, 306, 284, 204, 163, 55, 26, 1, and they need to be removed, After doing so, there are more outliers that appear: 764, 767, 148, 27, 28. After removing outliers a second time, more outliers appear and need to be removed: 184, 747.

The first picture in **M2.C** shows p-values, f-values MSE, parameter estimates after outliers are removed. There are still insignificant variables to be removed since the parameter estimates table shows that there are some variables that have a p-value greater than 0.05 which means they are insignificant. The second picture in M2.C shows everything after the insignificant predictors (blog, com_dis_dummy) have been removed. As you can see, the F-value went up (from 93.33 to 129.85), while the adj r-sq value went down very slightly (from .4744 to .4736 which shows how insignificant these predictors were) and the Root MSE went up very slightly (from 1.04653 to 1.04733). This means that this model accounts for 45.07% of the variability in new_y caused by film, entertainment, politics, likes, and dislikes.

The final model for model 2 is new_y = 12.75887 + .48042*film + .27209*entertainment -0.50479*politics + 0.00000687*likes + 0.00003712*dislikes. This means that film, & entertainment videos will get more views, while politics videos will get less views. However, we still need to retransform each coefficient with an exponential transformation before we can interpret it properly because there was a log transformation at the beginning of our analysis. If a video is in the film category it will have (e^(.48042)-1)*100=61.67% increase in views, and if a video is in the entertainment category, it will have (e^(.27209)-1)*100=31.27% increase in views. If a video is in the politics category, it will have (e^(0.50479)-1)*100= 65.66% decrease in views. For every like a video gets, it will have (e^(0.00000573)-1)*100=0.00069% increase in views. For every dislike a video gets, it will have (e^(0.00003712)-1)*100=0.00371% increase in views. Film, entertainment, and politics are dummy variables so they either have a value of 0 or 1 (a video is either in the film, entertainment, or politics category or it isn't). If they have a value of 1, their coefficients are taken into account. Otherwise, they are not.

Model 3 Regression

The third model chosen through the help of the forward selection method uses the variables film, scitech, politics, info, com_dis_dummy, likes, dislikes, and comment_count to predict the ln_views. Firstly, a basic regression analysis was done to determine how well the model is. The figure **M3.A** shows us that the model is not the best for predicting the dependent variable and that it can be improved. This is because the R value is only 0.4289 and has an Adj-Re^2 of only 0.4226. This means that the model only accounts about 43% of variability to the dependent variable. As these values are pretty low, it tells us that the model is not the best.

To further inspect the model, I conducted a residual analysis. The figures **M3.B** gives us a lot of insights about this model. As the variables film, scitech, politics, info, and com_di_dummy are dummy variables, the residual analysis does not tell us anything about these variables. However, valuable information can be gained from the residual analysis of the other variables (likes , dislikes, and comment_count). The residual analysis of the variable likes shows us that there is no constant variance and independence as the distribution does not seem to be random. This can also be said for the variables dislikes and comment_count. In all of these graphs, the majority of points are clustered on the left side of the graph with few points toward the right side. This pattern can also be seen in the predicted value graph. The normal probability tells us that the normality assumptions are satisfied though contains few curves. All this could be due to the outliers and influential points that can be seen on **M3.B** plots. This analysis further shows how this model is poor at predicting the dependent variable.

Though this a poor model at the moment, it can be improved for better performance in predicting the dependent variable. The first thing that must be done is to check and find the outliers and influential points seen on the **M3.B** plots**. M3.C** shows how there are many outliers and influential points in this model. Since there were so many of these points, I decided to remove them. After removing these outliers and influential points, I conducted the residual analysis again. As seen in **M3.D,** the second residual plot shows an improvement in the model. The new R^2 is 0.5444 and the Adj-R^2 is 0.5388. Both of this value increased from the previous analysis of this model meaning that the model is better after removal of these outliers and influential points. The residual plots of the variables likes, dislikes, and coument_count and predicted value still do not seem to have constant variance and independence. The normal probability plot also shows an improvement as it is much more linear. It still satisfies the normality assumption. These plots also show us that there are more outliers and influential points.

As seen in the figure **M3.E**, there are still many more outliers and influential points present in the data. Therefore I decide to remove these as well. After removing these points, I conduct the residual analysis again. From **M3.F,** we get the new R^2 is 0.5622 and the Adj-R^2 value of 0.5572. The residual plots, though improved, still show the same pattern of no constant variance and no independence. The normality plot is also more linear and still satisfies normality assumptions. They are still showing that few outliers and influential points exist.

After finding these points from **M3.G**, they were removed and the residual analysis was re-done. The **M3.H** residual plots show the R^2 value of 0.5665 and Adj-R^2 value of 0.5616. The residual plots, especially the comment_count and predicted value, now show some constant variance and independence. The normality plot still satisfies the normality assumption. There are also no more outliers since there are no points above 3 and below -3. The model is now much better at predicting the dependent variable than how it originally was. It now accounts for 56% of variability to the dependent variable.

To further improve the model after removing all of the outliers and influential points, I checked for any insignificant variables. As it can be seen in **M3.I**, the variables scitech, politics, and comment_count are insignificant as their p-values are greater than 0.05 and the com_dis_dummy ended up being 0. Therefore I removed these variables from my final model 3. To do this, I first removed the variable politics as it had the greatest p-value and then re-ran the analysis. Then I saw that the comment_count variable was insignificant and therefore removed it from the model. I repeated the analysis and saw that scitech was still an insignificant variable and so removed it from my model. After doing the regression analyses again I saw that there were no more insignificant variables as they all had p-values greater than 0.05 threshold. However, there was still one variable that just ended up being 0 after I removed all the outliers and influential points. This was the variable com_dis_count which had a value of 0 for each observation. Since it was 0 and had no effect on the dependent variable, I chose to remove it from the model. Now, my final models only include the variables film, info, likes, and dislikes which predict the dependent variable new_y which is the training set for ln_views.

My final model 3 is new_y = 12.37909 + 0.42405(film) - 0.17170(info) + 0.00001429(likes) + 0.00005554(dislikes), where new_y = ln_viewes, film = 1 when “true”, and info = 1 when “true”. This final model, as seen in **M3.J**, the RMSE value is 0.75928, R^2 value is 0.5648 and Adj-R^2 is 0.5620. This means that the model accounts for 56% of the dependent variable. To further check if this model is correctly fit, I conducted a goodness of fit test. For this, first, we assume null hypothesis (Ho) to be $B_j = 0$, meaning that none of the independent variables have an association with the dependent variable. Next we say that the alternate hypothesis is $B_j \neq 0$. In other words, there is at least one independent variable that has a significant association to the dependent variable. Now since, as it can be seen in **M3.J,** the

F-value is 201.17 and p-value < 0.05, we can safely reject the null hypothesis. This shows that the model passes the goodness of fit test.

To interpret the final model, new_y = 12.37909 + 0.42405(film) - 0.17170(info) + 0.00001429(likes) + 0.00005554(dislikes), we must re-transform it since it was originally transformed logarithmically. Therefore, a film video has (e^(0.42405)-1)*100=52.81% increase in views. An info video would have (e^(0.17170)-1)*100=18.73% decrease in the views. A video with 1 like has (e^(0.00001429)-1)*100=0.0014% increase in views. And a video with 1 dislike has (e^(0.00005554)-1)*100=0.00555% increase in views. The independent variable that has the most effect on the dependent variable is the film variable as it has the largest coefficient.

After finding the final model 3, I used the viladiation technique to test my model. I tested my model with the testing dataset that was separated earlier. As it can be seen in **M3.K**, the rmse is 1.55573 and mae is 1.02840 which are pretty low. In practice these two values should be minimized. The R^2 is $yhat^2$ which is $(.57837)^2$=0.33451 as seen in **M3.L.** Also according to **M3.L**, the stats between the original and predicted stats are somewhat close but not the best. The mean for original is 13.26 and the mean for predicted is 13.61. The standard deviation for original is 1.59 whereas the predicted has standard deviation of 1.78. This shows how good this model is at predicting the views from the testing dataset. This model is fairly good at predicting the ln_views for the testing dataset but not as good as I would like to see.

Predictions

The best model is the Model 3. We looked at the F-value, RMSE and adj r-square value to determine which model we would choose to do this. As shown in **M1.D** Model 1 had a F-value of 78.87, RMSE value of 1.17543 and adj r-square value of .4346. As shown in **M2.C** (the second table is the one that shows the correct values after insignificant predictors and outliers are removed)**,** Model 2 had a F-value of 117.35, RMSE value of 1.1, and adj r-sq value of .4469. Model 2 has a higher F-value than Model 1, a SLIGHTLY higher adj r-sq value (0.0123 more to be exact) and a lower RMSE value (by .6). Since the RMSE is lower in Model 2 (it's an error term so it's better to be lower), and the F-value and adj r-sq values are higher in Model 2 than in Model 1, Model 2 is a better model than Model 1. Now, we compare Model 3 to Model 2. As shown in **M3.I**, Model 3 has a F-value of 201.17, a RMSE of 0.75928, and an adj r-sq value of .56. Model 3 is much better than Model 2 because the F-value, and adj r-sq values are significantly higher than Model 2 (83.82, and .1151 respectively), while the RMSE value is much lower than Model 2 as well (by .34). Based on these measures, Model 3 was chosen for model validation techniques.

As shown in **M3.I**, the final equation for model 3, which is our best model, is new_y = 12.73909 + 0.42405*film -0.17170*info + 0.00001429*likes + 0.00005554*dislikes. New_y is ln_views, our dependent variable. Since there was a log transformation at the beginning of our analysis, we have to retransform each coefficient before we can interpret it. For predictors, if a video is in the film category, it is going to get (e^(0.42405)-1)*100 = 52.81% increase in views, while a video in the info category is going to get (e^(-0.17170)-1)*100 =18.732% decrease in views, meaning that it is going to get less views if it is in the info category. For every like and dislike on a video, a video will get (e^(0.00001429)-1)*100= 0.0014% increase in views and (e^(0.00005554)-1)*100=0.005554% increase in views respectively.

Since Model 3 was the best model, we also used it to predict the dependent variable, views, for two different scenarios. The two scenarios are as follows:

- Scenario 1: A video in the info category and not in the film category. This video has 3000 likes and 500 dislikes.
    - As shown by **P.A.**, the prediction interval for a video in the info category, not in the film category, has 3000 likes, and 500 dislikes has a 95% prediction interval of 11.1392 to 14.1368. Since there was a log transformation at the beginning of our analysis, these values must be retransformed with an exponential

transformation to make sense. Therefore, the actual 95% prediction interval is (e^(11.1392)-1)*100=6881558 to (e^(14.1368)-1)*100=137890363.6.

- Scenario 2: A video in the film category and not in the info category. This video has 2000 likes and 600 dislikes.

    - As shown by **P.B.**, the prediction interval for a video in the film category, not in the info category, has 2000 likes and 600 dislikes has a 95% prediction interval of 11.7047 to 14.7454. Since there was a log transformation at the beginning of our analysis, these values must be retransformed with an exponential transformation to make sense. Therefore, the actual 95% prediction interval is (e^(11.7047)-1)*100=121138.74 to (e^(14.7454)-1)*100=253422798.3.

Conclusion

The best model is Model 3. We looked at F-value, RMSE and adj r-square value to determine which model we would choose to do this. As shown in **M1.D** Model 1 had a F-value of 78.87, RMSE value of 1.17543 and adj r-square value of .4346. As shown in **M2.C** (the second table is the one that shows the correct values after insignificant predictors and outliers are removed)**,** Model 2 had a F-value of 117.35, RMSE value of 1.1, and adj r-sq value of .4469. Model 2 has a higher F-value than Model 1, a SLIGHTLY higher adj r-sq value (0.0123 more to be exact) and a lower RMSE value (by .6). Since the RMSE is lower in Model 2 (it's an error term so it's better to be lower), and the F-value and adj r-sq values are higher in Model 2 than in Model 1, Model 2 is a better model than Model 1. Now, we compare Model 3 to Model 2. As shown in **M3.J**, Model 3 has a F-value of 201.17, a RMSE of 0.75928, and an adj r-sq value of .56. Model 3 is much better than Model 2 because the F-value, and adj r-sq values are significantly higher than Model 2 (83.82, and .1151 respectively), while the RMSE value is much lower than Model 2 as well (by .34). Based on these measures, Model 3 was chosen for model validation techniques.

Our best model, model 3, only accounts for 56% of the variability in y as shown by the adj r-sq value in **M3.J**. Though this is not very good for a model, this is the best one from the three models we had. In this model, views are only accounted for by our significant variables. In theory, a good model should have a much higher adj R-sq value which means that it's better at predicting y. This model has low/moderate Adj R-sq value which means that it is not very good at predicting the number of views. This can be seen when the model is validated. In **M3.L,** we saw that the yhat value was only 0.57837 meaning the R-square value is (0.57837)^ 2=0.335. This means that it only accounts for 33% of variance in the dependent variable. This model may be the best of the three but is an okay model in practicality. It is an okay model but it could be a lot better.

In our best model, the model 3, the predictors with the strongest correlation with the dependent variables are likes followed by dislikes, info, and film. This is due to the computed correlation coefficients seen in **DE.G.** Likes has the highest correlation value of 0.489 and dislikes has 0.401. The film and info have much lesser correlation values since their correlation values were -0.0779 and 0.0413 respectively, showing that they have a very weak correlation with views.

Though the film variable changes the dependent variable the most, the variable that is the strongest predictor for this model is the variable likes. Likes has the greatest influence on the dependent variable as it has the greatest standardized estimate of 0.64911 followed by dislikes with standardized estimate of 0.14959 as seen in **M3.J.** Thus, our initial hypothesis that likes has

the greatest influence on a video's view count is correct, which also supports the experts' claims since likes could be considered a good measure of a video's ability to elicit positive emotions and as we discovered, likes is the most important factor towards a video's view count. The other two variables, film and info, do not have a great influence on the dependent variable as their standardized estimates are much lower at 0.07386 and -0.05491 respectively. Therefore, the part of our hypothesis where we predicted that a video's category will not impact its view count turned out to be incorrect. Both film and info are significant variables, which means that a video's category does impact the amount of view the video will obtain. However, the standardized estimates of film and info indicate that these impacts are minimal.

.

As it can be seen, the model 3 was the best of the three models chosen but is not the best model. It is not a model you can reliably use to predict the number of views one might get on their video. Even after removing all of the outliers and influential points and removing the insignificant variables, we were only able to account 56% of variance in the dependent variable with this model. This is not really something you would want in a practical scenario. There can be many reasons for the poor performance of this model. One such reason could be the data itself. During the data exploration stage, we ended up taking a sample of 1000 observations from the huge dataset because the program ran too slowly on the virtual server. The model was created from these 1000. Perhaps if some other set of 1000 observations were selected or even if more observations were selected, then maybe the model would have been better at predicting the views of youtube videos. Another possible reason could be human error. This could have occured in many places. For example, there could have been errors in coding of the project or even in certain interpretations leading to wrong decisions. One final reason for the poor model could be that the dataset itself does not have any predictability. So many people post videos on Youtube everyday and these videos could be very different from each other. There can be so many other variables that factor into the views these videos get that are not included in this dataset. For example, Youtube has their own algorithms for recommending videos and those that get recommended most possibly have a higher chance to get more views. Sometimes, seemingly random videos from many years ago can pop up in a lot of people's Youtube feeds all of a sudden, and a huge influx of people check out the video. It is not uncommon for people to ask in the Youtube comments of old videos on why this video was recommended in their feed all of a sudden, followed by a huge amount of likes/comments on their comment, which means that they are not the only ones that have been randomly recommended this video in their feed. There are so probably so many other things such as this that factor into the amount of views a video gets. Further testing must be conducted to see whether a reliable model can be created from this dataset.

For future work, we would like to continue studying the data with different sample sizes and create more models to see if there is one that predicts the views better than our model 3.

Some of the predictors such as politics, ratings_disabled that were not included in our final models might be stronger predictors with a larger sample size. The reason for this is that politics is actually a highly debated topic in the United States, with all the protests and debates going on ever since the Trump election. Also, ratings_disabled videos can sometimes and often do have a lot of views, but have their ratings disabled BECAUSE the ratings may overwhelmingly be dislikes which would cause the viewer to be negatively biased towards the video rather than focusing on the actual content of the video itself. Another thing that can be worked on is finding a relationship between another dependent variable and the other independent variable. For example, we would choose the variable likes as a dependent variable and the other variable as independent variables and see how they affect/predict the number of likes a certain video would get. Same thing can be done for the dislike variable. It would be interesting to find out what categories of videos are the most liked, as well as the most disliked.

Alternatively, since this analysis was done on videos on the United States, it would be interesting to do a regression analysis on the other datasets that are for youtube videos that are not in the United States. In the original folder where we got our dataset for US youtube video data from, there is data for Canada, Germany, France, Great Britain, India, Japan, South Korea, Mexico, and Russia. With our analysis on 1000 random samples on the original US video dataset, we discovered likes, and dislikes are the strongest predictors for views. If a video has more likes, and dislikes, it will have more views. We removed politics, and scitech as predictors in our model since they were insignificant. However, this was for the United States video set. In a country that is always on the cutting edge of technology such as Japan (based on their adoption of mag-lev trains, real mechs, etc.) it may be the case that scitech will be a strong predictor in the Japan video data set.

Similarly (or differently), we could perform a regression analysis on multiple of these datasets and compare the models for each of the countries' datasets. As mentioned previously, we believe scitech will be a stronger predictor of views in a country such as Japan. But it would be interesting to see if scitech as well as other variables are stronger predictors of views in other countries. Would politics be a predictor of views in the Great Britain dataset or the Russia dataset? If so, would it be a stronger predictor of views in the Great Britain dataset or the Russia dataset? Would music be a predictor of views in the Korea dataset? We know that K-pop is a very big "thing" in Korea, and even so in the United States. Would this mean that music is the strongest predictor of views in Korea since K-pop is such an important part of their culture? This kind of analysis could give us an insight to the culture of these different countries.

If a good model can be generated from this analysis, it can be really helpful to those who want to post videos on Youtube. This model could give them insights to what type of video they should post to get the most amount of views. It can help youtubers make decisions about their

content and also tell them how popular their video has a chance to be. Youtube is a very big community with both content creators and content watchers. In fact there are so many people with their full time job being a youtuber and there are many who struggle to get to a position where they can earn tons of money from it. If they are able to use a successful model that predicts views, these people can have a better chance of getting to popularity and success in the Youtube community, and YouTube becomes a much more viable option as a career, allowing more people to do what they love, which is making content that others will enjoy.

Something important to note is whether or not views are unique or not. It is possible for someone to like a video so much that they view it many times over and over again. The question is whether or not it matters how many views a video has if many of them are from the same person/people. Or maybe these videos have artificially increased views because someone has paid a website to send bots (or people) to view the video. This large artificial increase in views may cause more people to be recommended this video in their YouTube feed because as we all know, popular videos tend to get more views, which could further drive up the number of views.

References

- Pozin, Ilya. “6 Qualities To Make Your Videos Go Viral.” *Forbes*, Forbes Magazine, 7 Aug. 2014, www.forbes.com/sites/ilyapozin/2014/08/07/6-qualities-to-make-your-videos-go-viral/#10434685154e.

Appendix

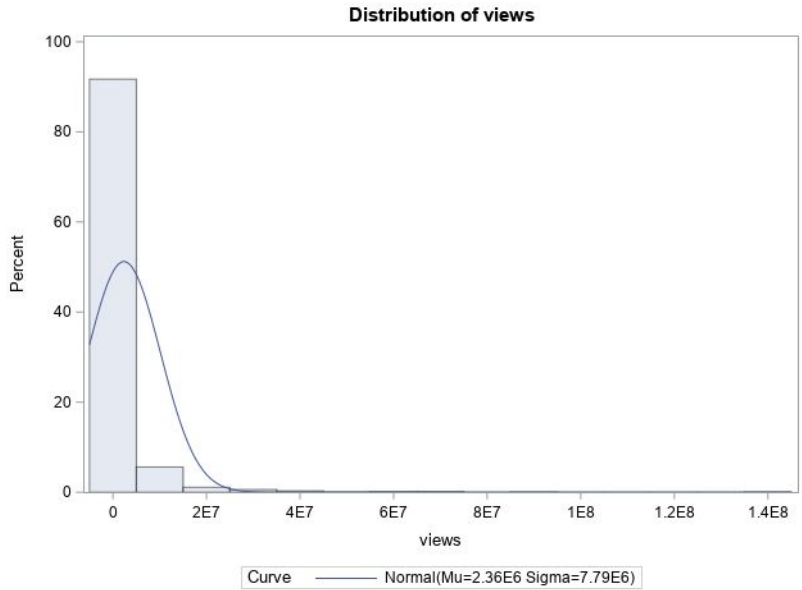
Methodology Appendix

M.A - JSON file for Category_ID variable tabulated

Category	ID Value
Film & Animation	1
Autos & Vehicles	2
Music	10
Pets & Animals	15
Sports	17
Short Movies	18
Travel & Events	19
Gaming	20
Video Blogging	21
People & Blogs	22
Comedy	23
Entertainment	24
News & Politics	25
Howto & Style	26
Education	27
Science & Technology	28
Nonprofits & Activism	29
Movies	30
Anime/Animation	31
Action/Adventure	32

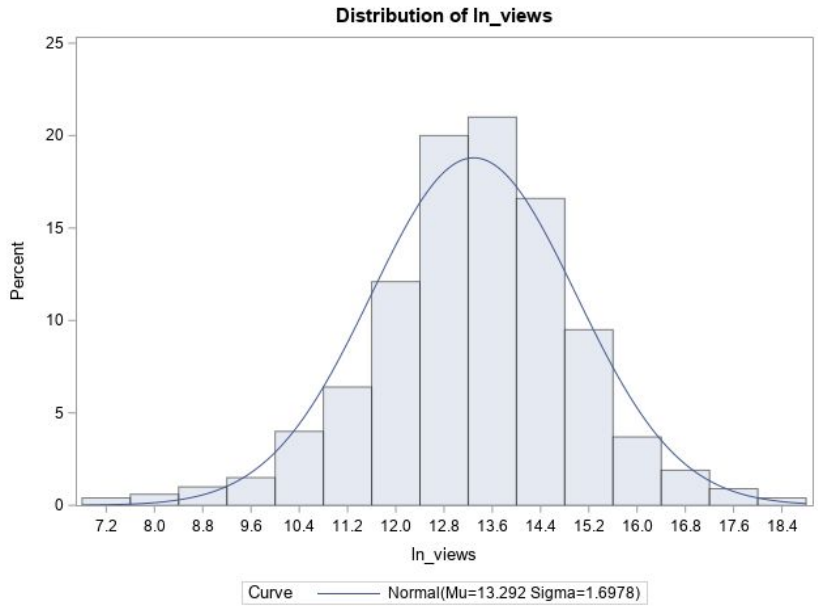
Classics	33
Comedy	34
Documentary	35
Drama	36
Family	37
Foreign	38
Horror	39
Sci-Fi/Fantasy	40
Thriller	41
Shorts	42
Shows	43
Trailers	44

M.B - Original histogram of views



Quantiles for Normal Distribution		
	Quantile	
Percent	Observed	Estimated
1.0	4311.50	-15755037
5.0	30300.50	-10448251
10.0	72141.00	-7619222
25.0	230296.50	-2892040
50.0	648581.50	2360203
75.0	1742846.50	7612446
90.0	4121417.00	12339628
95.0	8170926.50	15168656
99.0	33796925.50	20475443

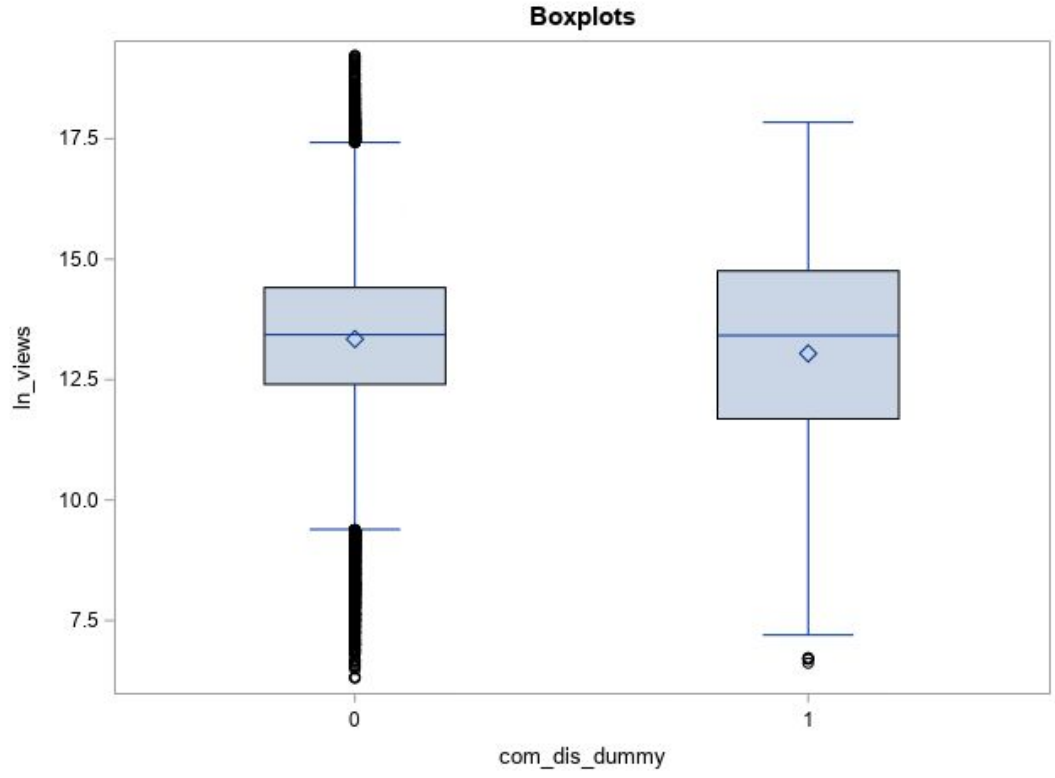
M.C - Histogram of ln_views (transformed views)



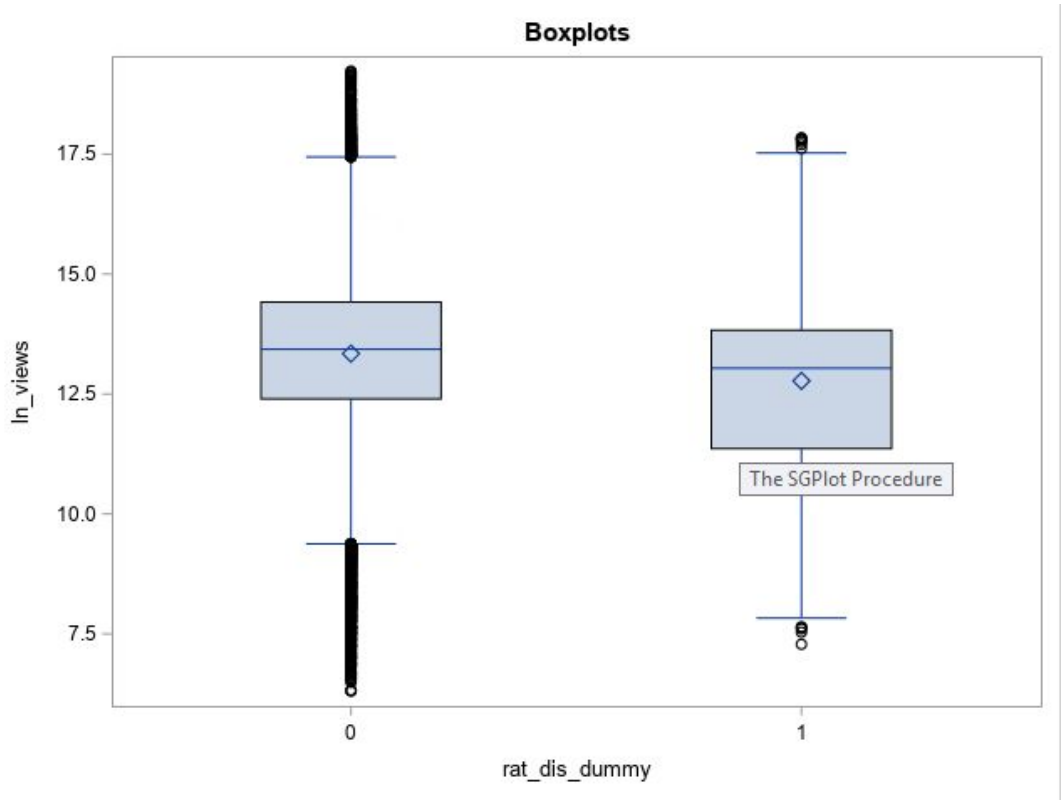
Quantiles for Normal Distribution		
	Quantile	
Percent	Observed	Estimated
1.0	8.36620	9.34210
5.0	10.31888	10.49912
10.0	11.18638	11.11593
25.0	12.34710	12.14658
50.0	13.38254	13.29171
75.0	14.37102	14.43684
90.0	15.23171	15.46749
95.0	15.91605	16.08429
99.0	17.33262	17.24131

Data Exploration Appendix

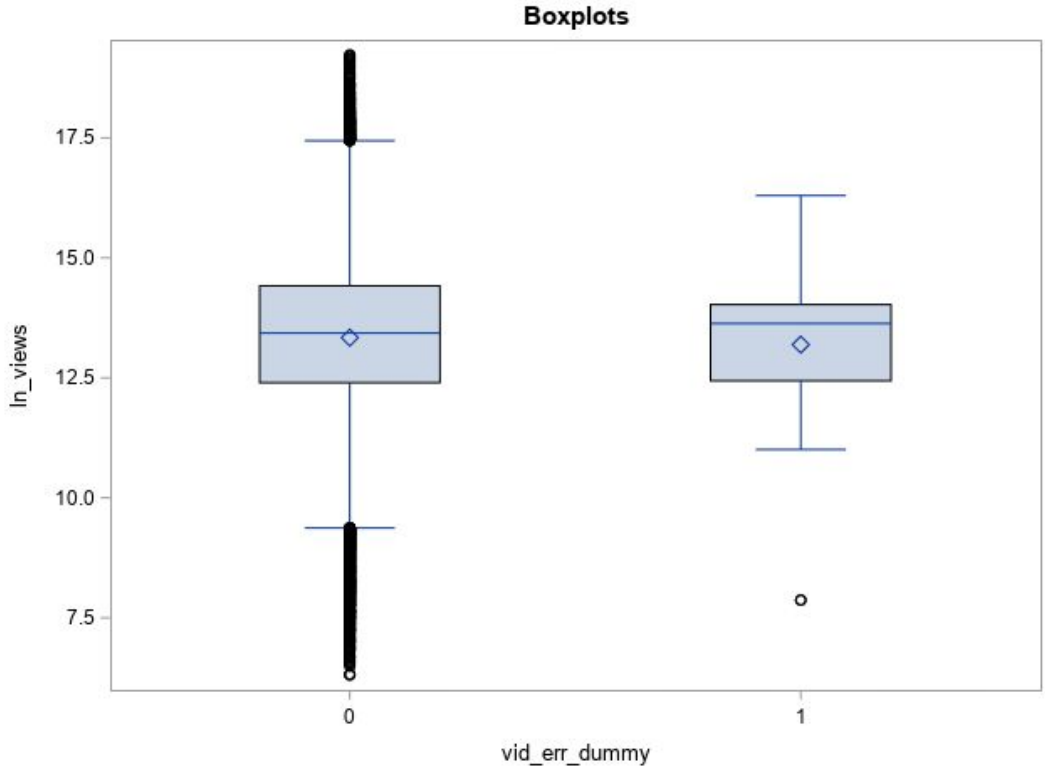
DE.A



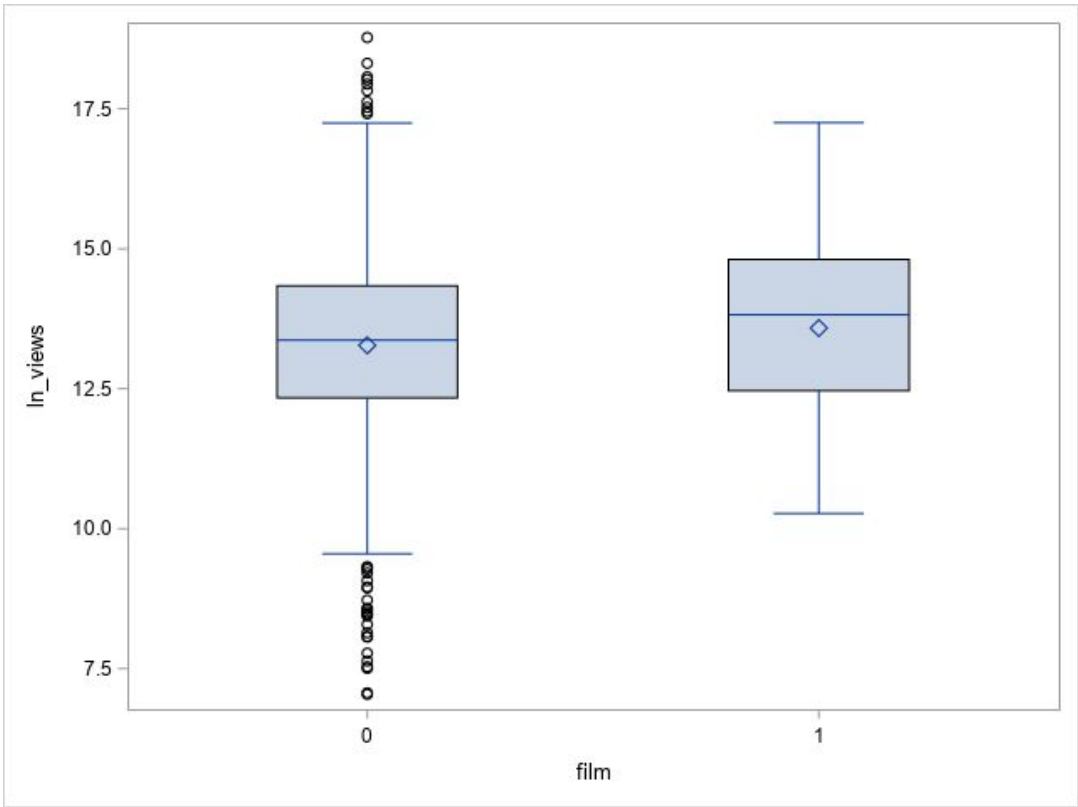
DE.B

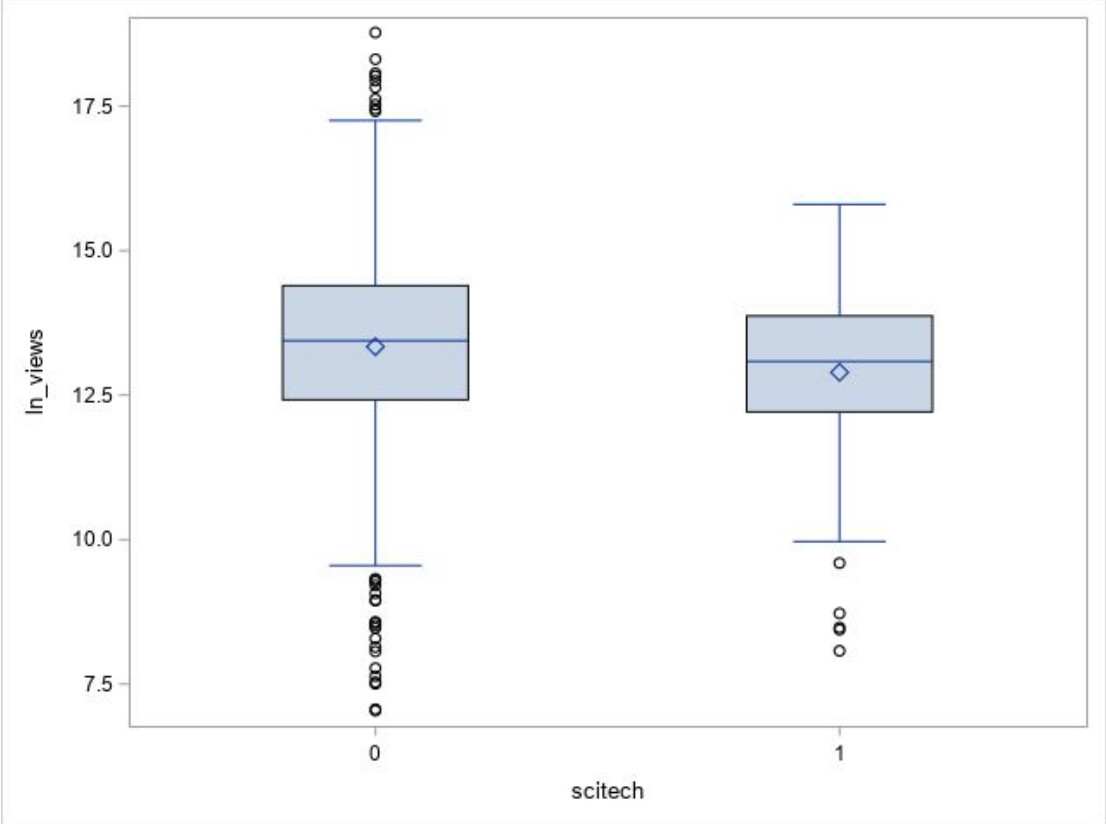

Boxplots
ln_views
17.5
15.0
12.5
10.0
7.5
0
1
rat_dis_dummy
The SGPlot Procedure

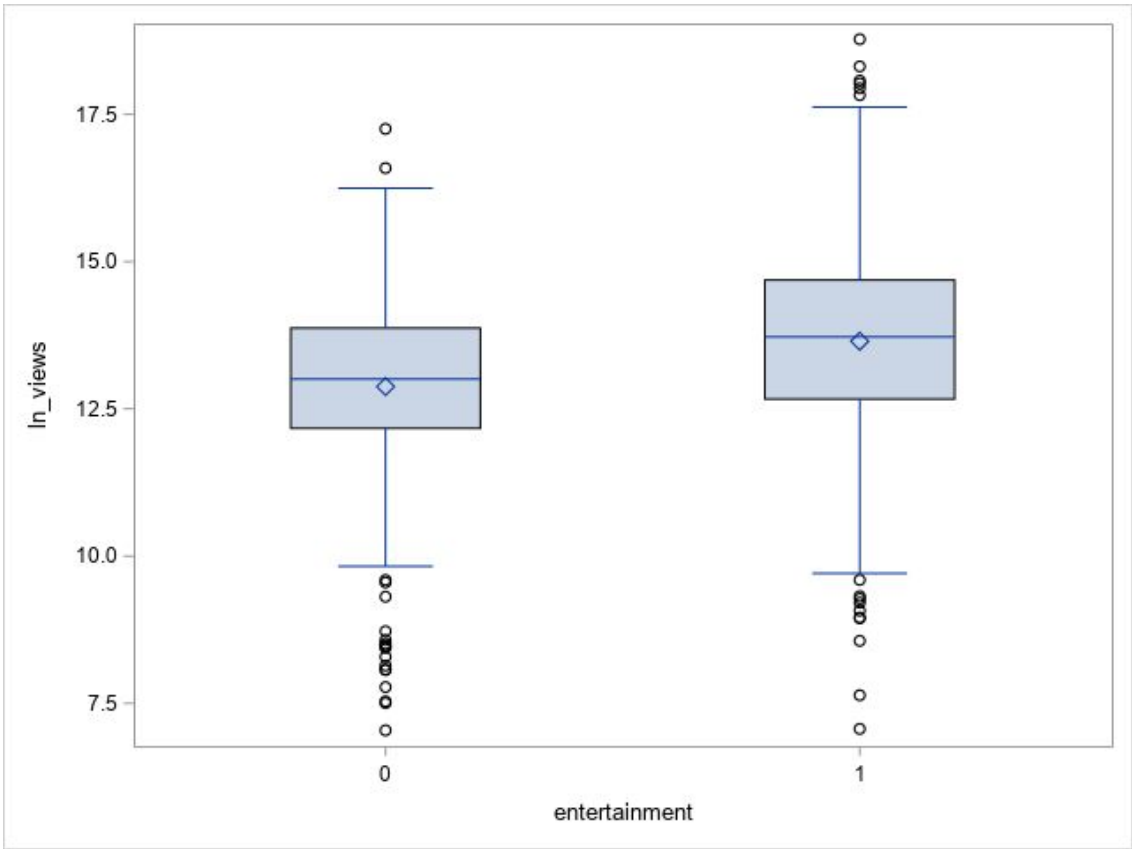
DE.C

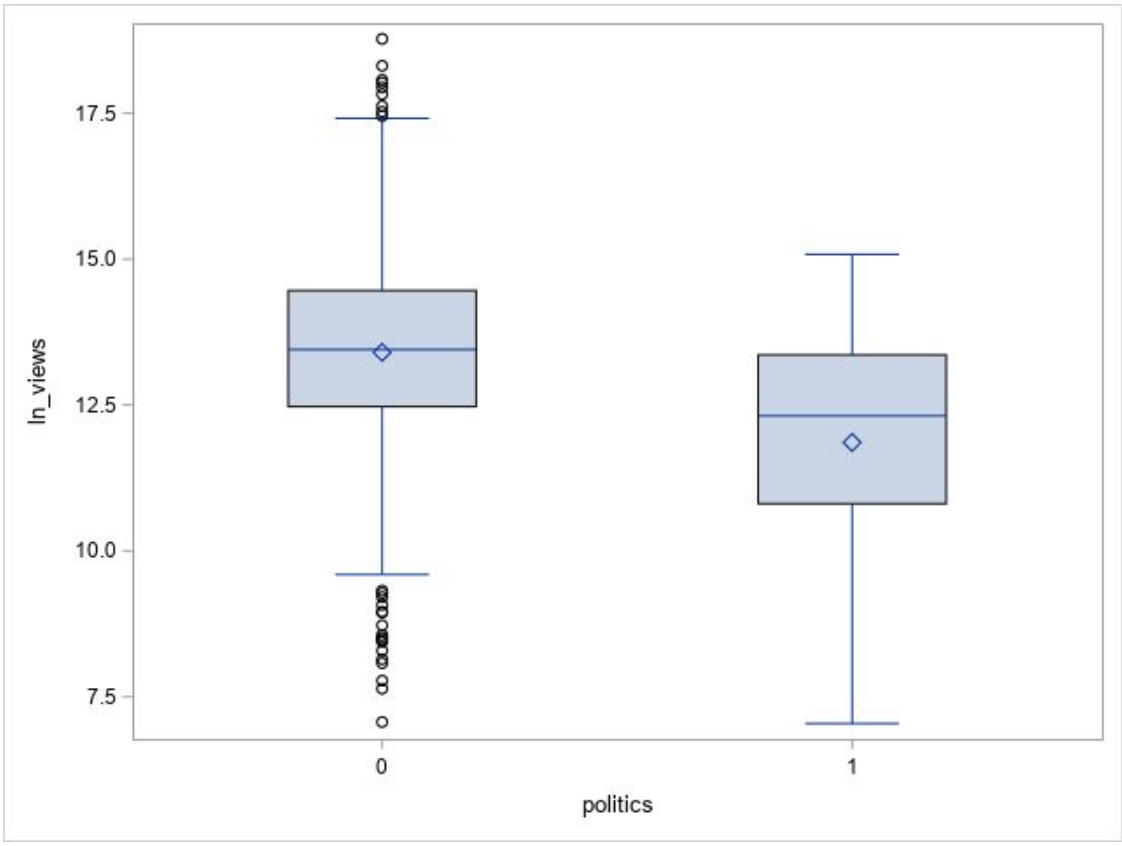

Boxplots
ln_views
17.5
15.0
12.5
10.0
7.5
0
1
vid_err_dummy

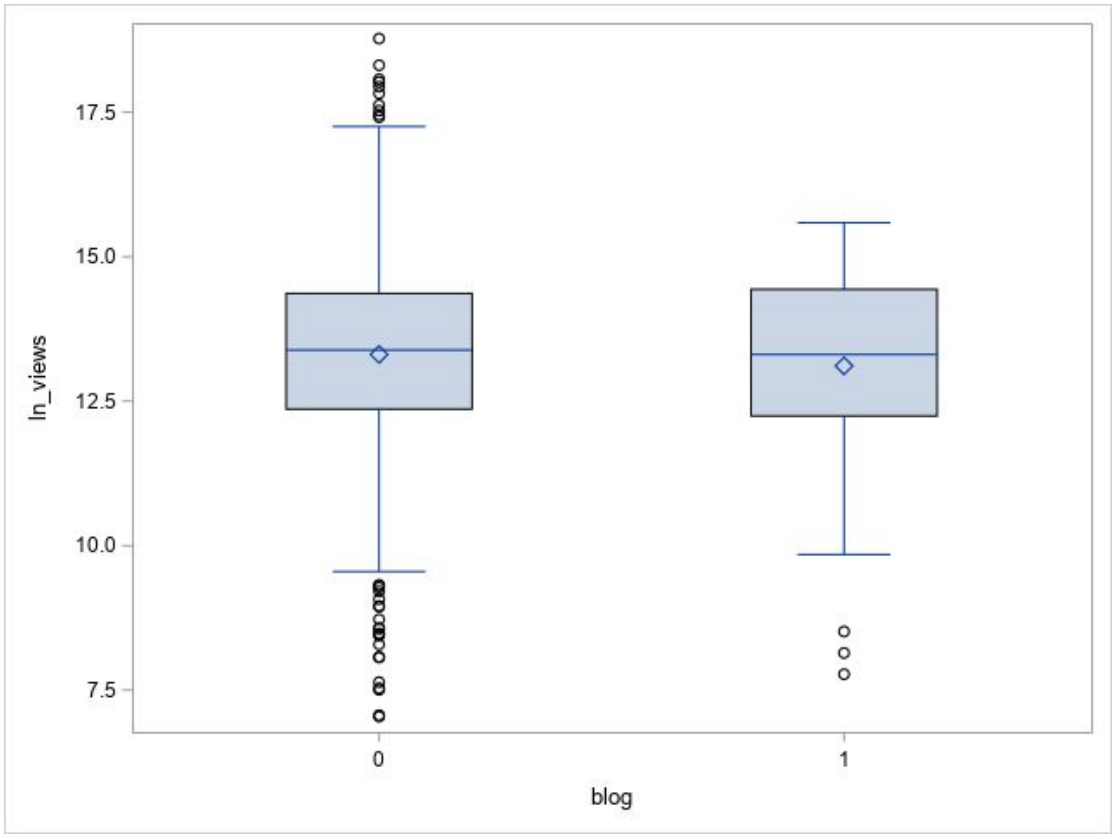
DE.D

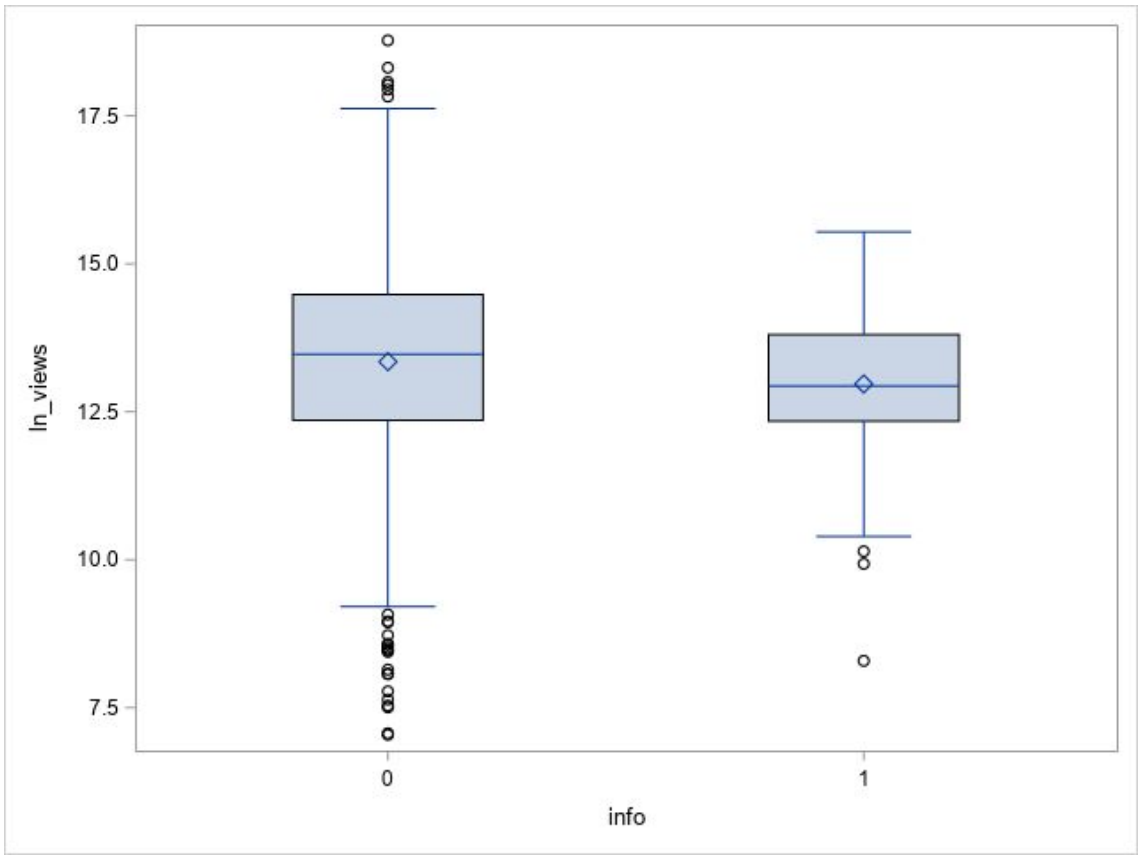

17.5
15.0
12.5
10.0
7.5
ln_views
0
1
film


17.5
15.0
12.5
10.0
7.5
ln_views
0
1
scitech

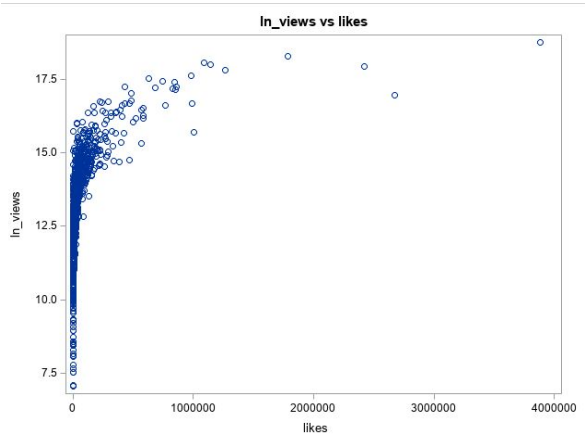

ln_views
17.5
15.0
12.5
10.0
7.5
0
1
entertainment

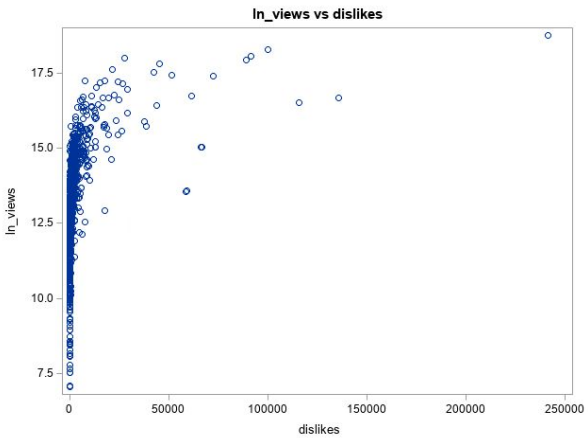

ln_views
17.5
15.0
12.5
10.0
7.5
0
1
politics

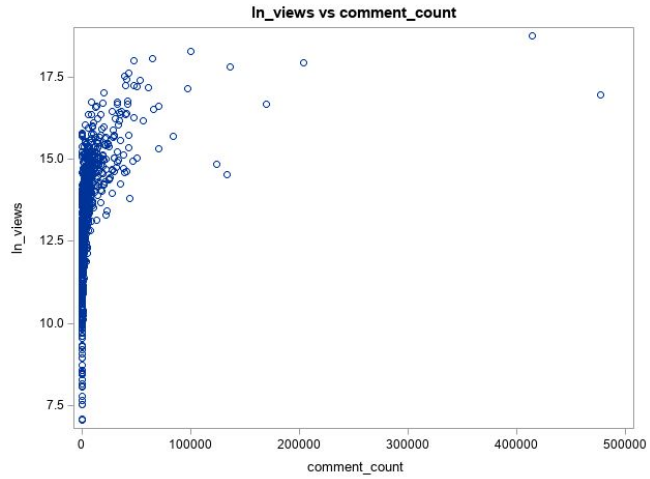

ln_views
17.5
15.0
12.5
10.0
7.5
0
1
blog


ln_views
17.5
15.0
12.5
10.0
7.5
0
1
info

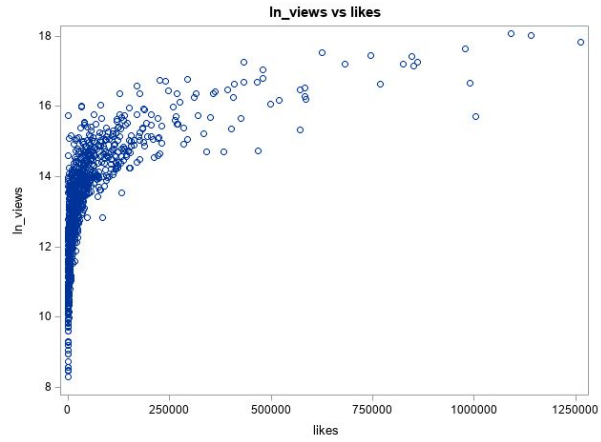
DE.E Initial Scatterplots

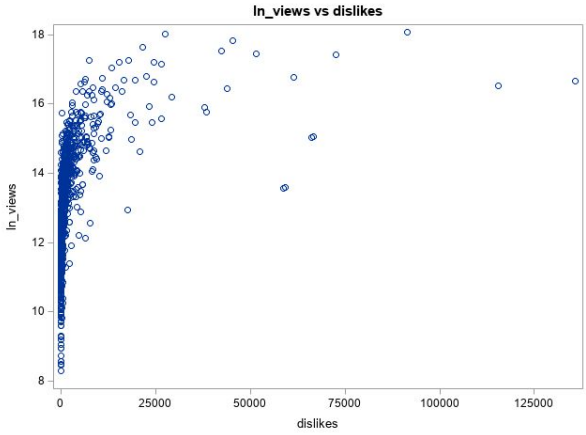


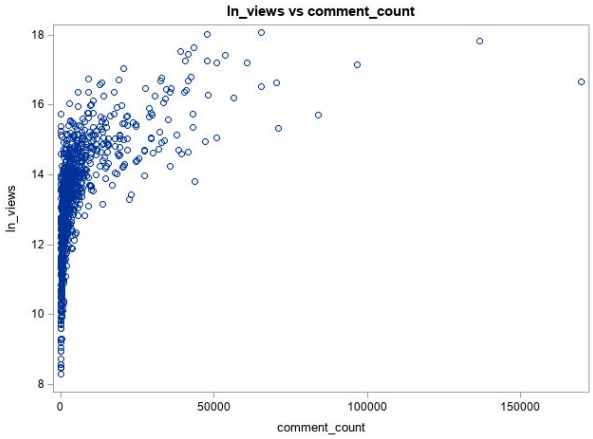




DE.F - Scatterplots post outlier removal







DE.G - Correlation Matrix

Pearson Correlation Coefficients, N = 1000
Prob > |r| under H0: Rho=0

	ln_views	com_dis_dummy	rat_dis_dummy	vid_err_dummy	likes	dislikes	comment_count	film	scitech	entertainment	politics	blog	info
ln_views	1.00000	-0.00583 0.8539	-0.03918 0.2158	.	0.48860 <.0001	0.40114 <.0001	0.40062 <.0001	0.04132 0.1917	-0.08108 0.0103	0.22645 <.0001	-0.23555 <.0001	-0.03259 0.3031	-0.07792 0.0137
com_dis_dummy	-0.00583 0.8539	1.00000	0.25338 <.0001	.	-0.02618 0.4082	0.07355 0.0200	-0.04225 0.1819	-0.03446 0.2762	0.04298 0.1744	-0.03954 0.2115	0.09837 0.0018	0.03330 0.2928	-0.05836 0.0651
rat_dis_dummy	-0.03918 0.2158	0.25338 <.0001	1.00000	.	-0.01789 0.5719	-0.01449 0.6471	-0.01571 0.6197	-0.01323 0.6760	0.04016 0.2044	-0.05919 0.0613	0.05546 0.0796	0.04885 0.1227	-0.02241 0.4791
vid_err_dummy	. .	. .	. .	. .	. .	. .	. .	. .	. .	. .	. .	. .	. .
likes	0.48860 <.0001	-0.02618 0.4082	-0.01789 0.5719	.	1.00000	0.76430 <.0001	0.90320 <.0001	-0.01580 0.6178	-0.06592 0.0372	0.16460 <.0001	-0.08545 0.0069	-0.03329 0.2929	-0.07631 0.0158
dislikes	0.40114 <.0001	0.07355 0.0200	-0.01449 0.6471	.	0.76430 <.0001	1.00000	0.68068 <.0001	-0.01902 0.5481	-0.05411 0.0872	0.10486 0.0009	-0.05060 0.1098	0.02341 0.4597	-0.07046 0.0259
comment_count	0.40062 <.0001	-0.04225 0.1819	-0.01571 0.6197	.	0.90320 <.0001	0.68068 <.0001	1.00000	-0.00350 0.9120	-0.05468 0.0839	0.11643 0.0002	-0.06763 0.0325	-0.02528 0.4246	-0.04518 0.1534
film	0.04132 0.1917	-0.03446 0.2762	-0.01323 0.6760	.	-0.01580 0.6178	-0.01902 0.5481	-0.00350 0.9120	1.00000	-0.08351 0.0082	-0.26034 <.0001	-0.06720 0.0336	-0.07353 0.0200	-0.09855 0.0018
scitech	-0.08108 0.0103	0.04298 0.1744	0.04016 0.2044	.	-0.06592 0.0372	-0.05411 0.0872	-0.05468 0.0839	-0.08351 0.0082	1.00000	-0.37354 <.0001	-0.09642 0.0023	-0.10550 0.0008	-0.14140 <.0001
entertainment	0.22645 <.0001	-0.03954 0.2115	-0.05919 0.0613	.	0.16460 <.0001	0.10486 0.0009	0.11643 0.0002	-0.26034 <.0001	-0.37354 <.0001	1.00000	-0.30058 <.0001	-0.32890 <.0001	-0.44081 <.0001
politics	-0.23555 <.0001	0.09837 0.0018	0.05546 0.0796	.	-0.08545 0.0069	-0.05060 0.1098	-0.06763 0.0325	-0.06720 0.0336	-0.09642 0.0023	-0.30058 <.0001	1.00000	-0.08490 0.0072	-0.11378 0.0003
blog	-0.03259 0.3031	0.03330 0.2928	0.04885 0.1227	.	-0.03329 0.2929	0.02341 0.4597	-0.02528 0.4246	-0.07353 0.0200	-0.10550 0.0008	-0.32890 <.0001	-0.08490 0.0072	1.00000	-0.12450 <.0001
info	-0.07792 0.0137	-0.05836 0.0651	-0.02241 0.4791	.	-0.07631 0.0158	-0.07046 0.0259	-0.04518 0.1534	-0.09855 0.0018	-0.14140 <.0001	-0.44081 <.0001	-0.11378 0.0003	-0.12450 <.0001	1.00000

Full Model

FM.A

First Full Model

The REG Procedure
Model: MODEL1
Dependent Variable: ln_views

Number of Observations Read	1000
Number of Observations Used	1000

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	10	865.56752	86.55675	42.51	<.0001
Error	989	2013.97660	2.03638		
Corrected Total	999	2879.54412			

Root MSE	1.42702	R-Square	0.3006
Dependent Mean	13.29171	Adj R-Sq	0.2935
Coeff Var	10.73614		

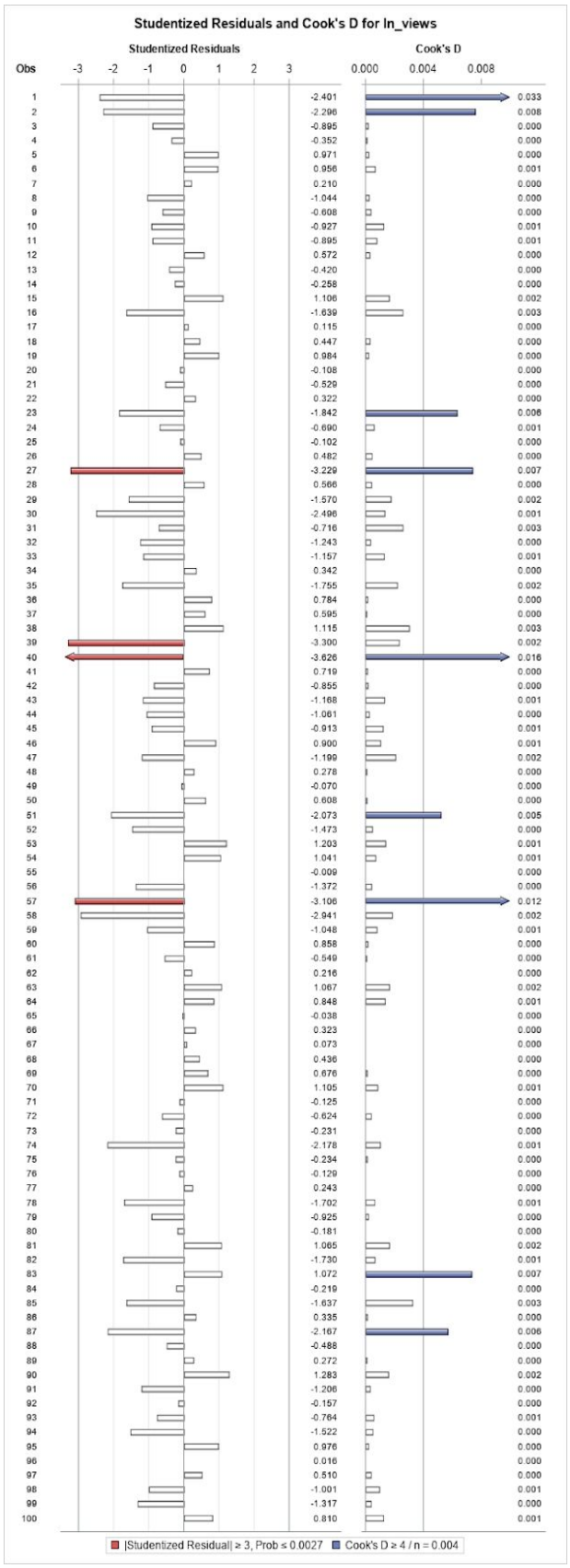
ımeters are not unique. Some statistics will be misleading. A reported DF of 0 or B me

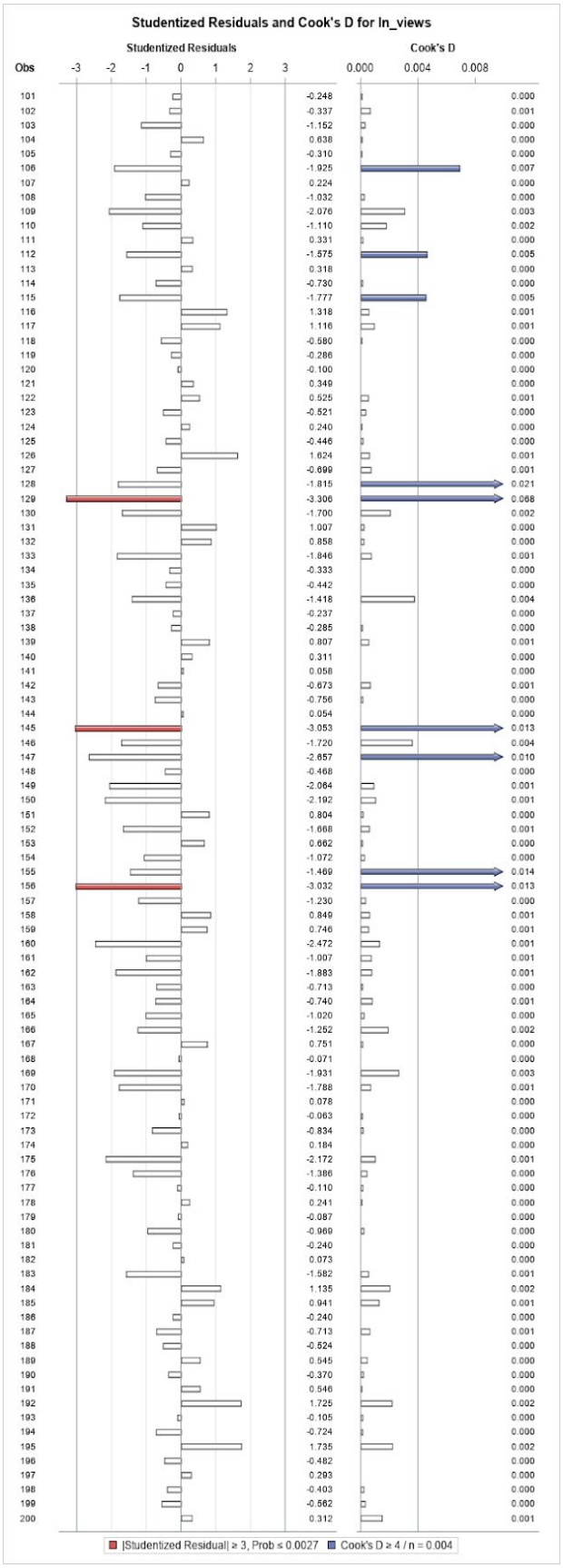
ıables are a linear combination of other variables as shown.

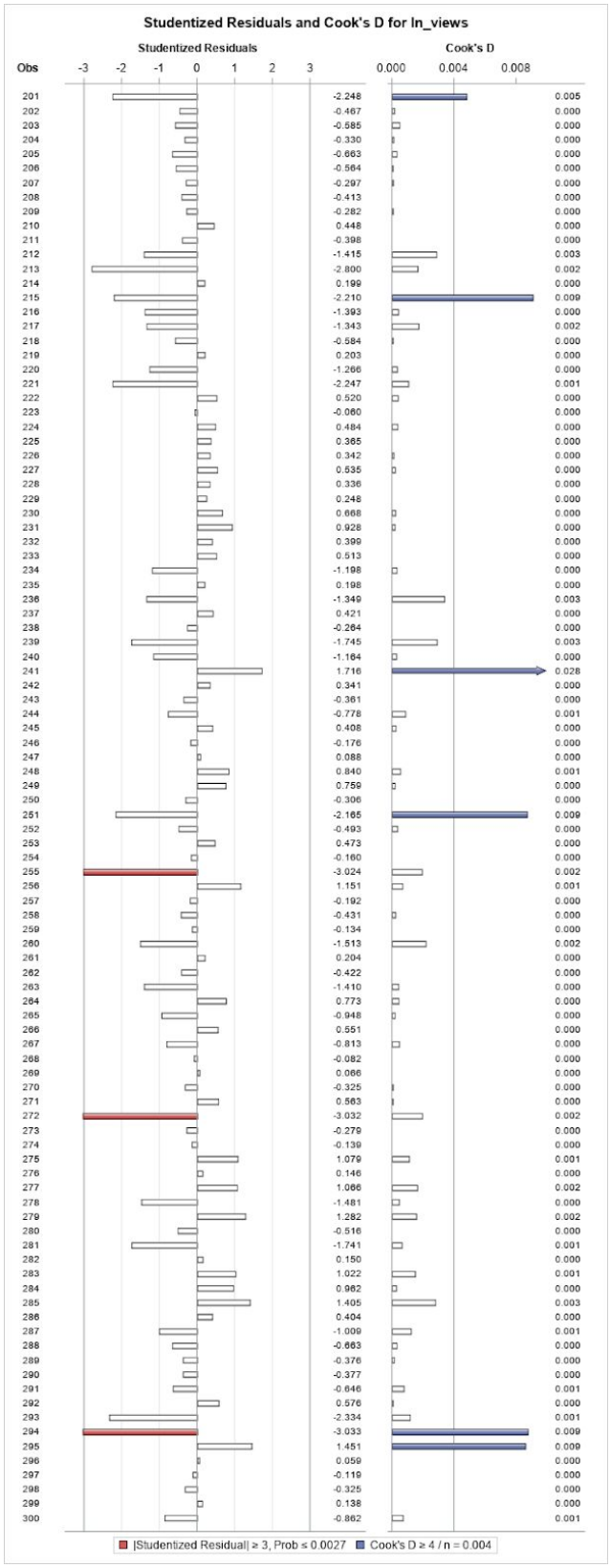
info =	Intercept - film - scitech - entertainment - politics - blog
vid_err_dummy =	0

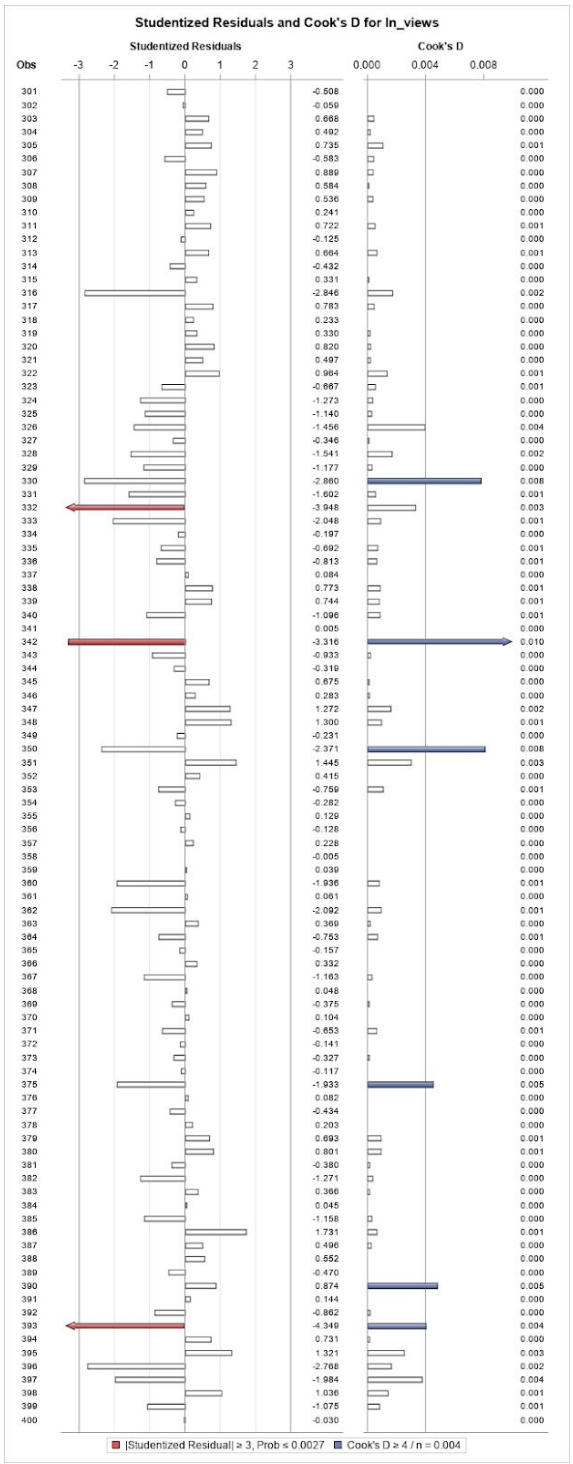
Parameter Estimates								
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate	Tolerance	Variance Inflation
Intercept	B	12.88087	0.11964	107.66	<.0001	0	.	0
film	B	0.51577	0.22644	2.28	0.0230	0.06929	0.76376	1.30931
scitech	B	-0.09626	0.18281	-0.53	0.5986	-0.01753	0.63738	1.56892
entertainment	B	0.38341	0.13554	2.83	0.0048	0.11265	0.44577	2.24330
politics	B	-1.05193	0.20748	-5.07	<.0001	-0.16024	0.70765	1.41312
blog	B	0.03768	0.19616	0.19	0.8477	0.00619	0.68014	1.47029
info	0	0	.	.	.	.	.	.
com_dis_dummy	1	0.23317	0.32850	0.71	0.4780	0.01924	0.96233	1.03914
rat_dis_dummy	0	0	.	.	.	.	.	.
vid_err_dummy	0	0	.	.	.	.	.	.
likes	1	0.00000448	5.536942E-7	8.10	<.0001	0.57636	0.13945	7.17084
dislikes	1	0.00000932	0.00000561	1.66	0.0968	0.06949	0.40451	2.47216
comment_count	1	-0.00001302	0.00000424	-3.07	0.0022	-0.19094	0.18248	5.48000

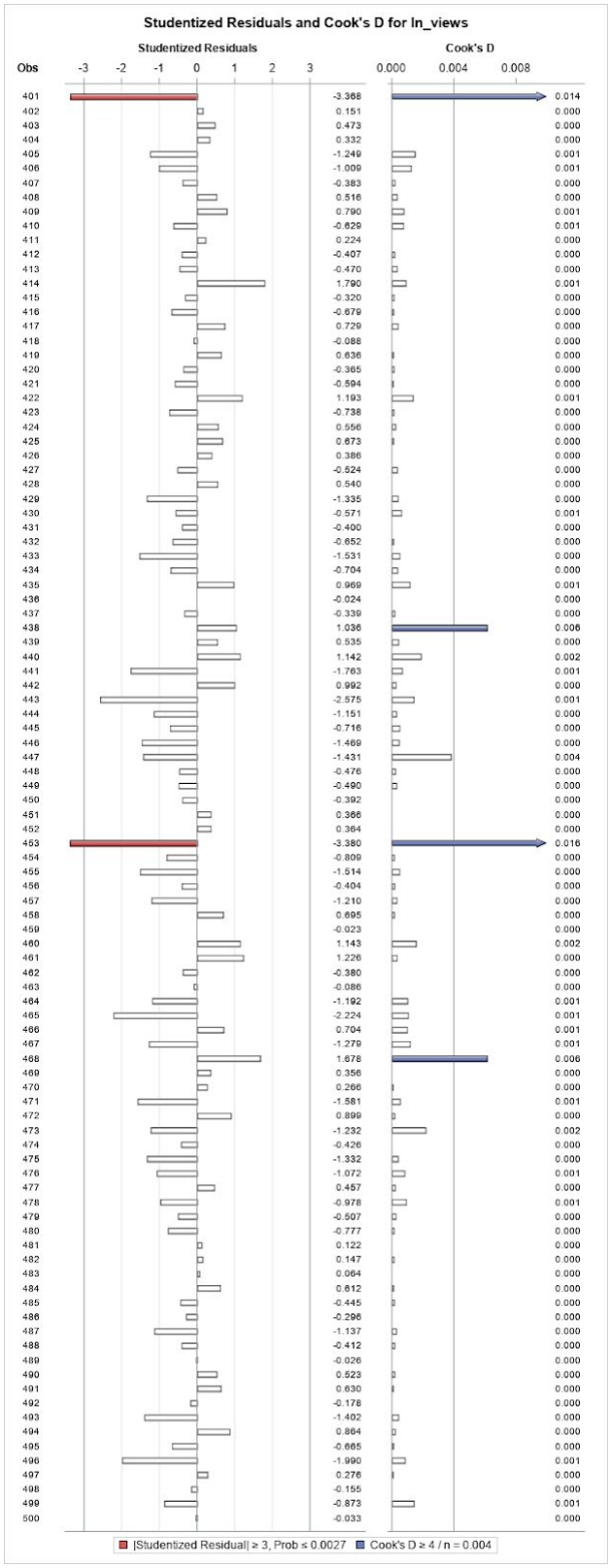
FM.B

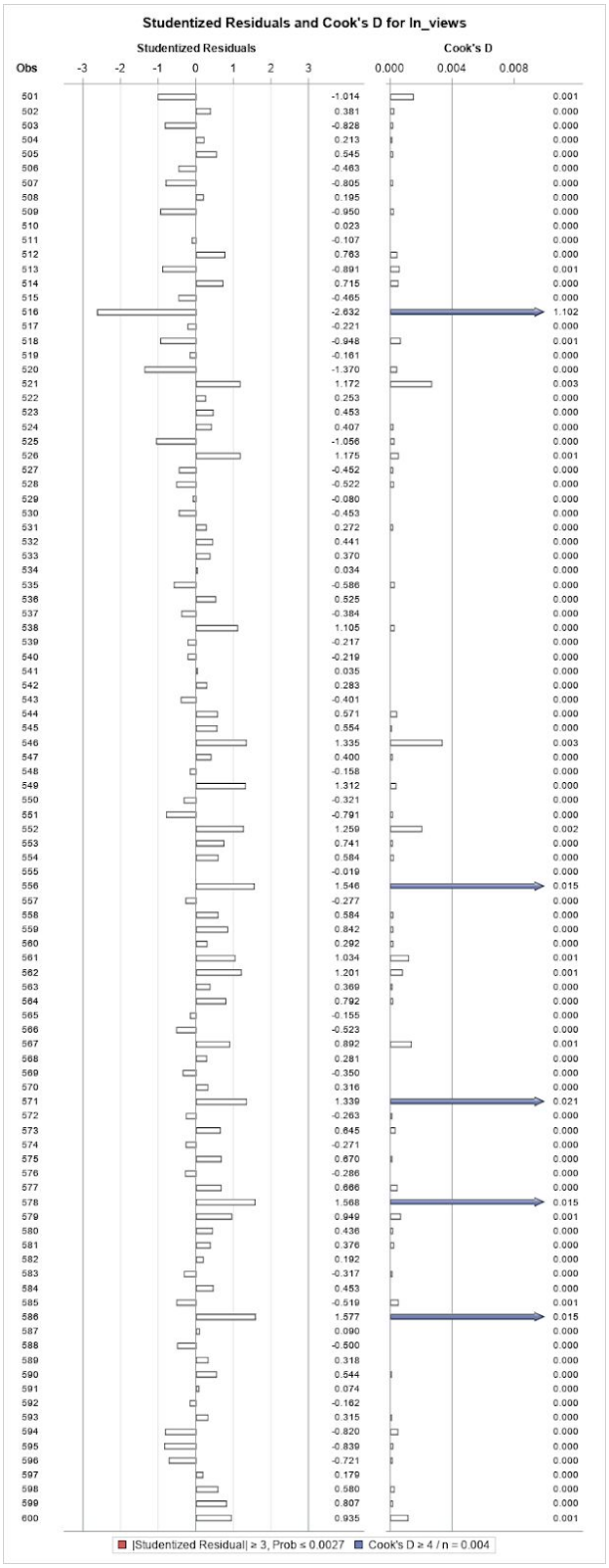

Studentized Residuals and Cook's D for ln_views
Studentized Residuals
Cook's D
Obs
-3 -2 -1 0 1 2 3
0.000 0.004 0.008
1 -2.401 0.033
2 -2.296 0.008
3 -0.895 0.000
4 -0.352 0.000
5 0.971 0.000
6 0.956 0.001
7 0.210 0.000
8 -1.044 0.000
9 -0.608 0.000
10 -0.927 0.001
11 -0.895 0.001
12 0.572 0.000
13 -0.420 0.000
14 -0.258 0.000
15 1.106 0.002
16 -1.639 0.003
17 0.115 0.000
18 0.447 0.000
19 0.984 0.000
20 -0.108 0.000
21 -0.529 0.000
22 0.322 0.000
23 -1.842 0.006
24 -0.690 0.001
25 -0.102 0.000
26 0.482 0.000
27 -3.229 0.007
28 0.566 0.000
29 -1.570 0.002
30 -2.496 0.001
31 -0.716 0.003
32 -1.243 0.000
33 -1.157 0.001
34 0.342 0.000
35 -1.755 0.002
36 0.784 0.000
37 0.595 0.000
38 1.115 0.003
39 -3.300 0.002
40 -3.626 0.016
41 0.719 0.000
42 -0.855 0.000
43 -1.168 0.001
44 -1.061 0.000
45 -0.913 0.001
46 0.900 0.001
47 -1.199 0.002
48 0.278 0.000
49 -0.070 0.000
50 0.608 0.000
51 -2.073 0.005
52 -1.473 0.000
53 1.203 0.001
54 1.041 0.001
55 -0.009 0.000
56 -1.372 0.000
57 -3.106 0.012
58 -2.941 0.002
59 -1.048 0.001
60 0.858 0.000
61 -0.549 0.000
62 0.216 0.000
63 1.067 0.002
64 0.848 0.001
65 -0.038 0.000
66 0.323 0.000
67 0.073 0.000
68 0.436 0.000
69 0.676 0.000
70 1.105 0.001
71 -0.125 0.000
72 -0.624 0.000
73 -0.231 0.000
74 -2.178 0.001
75 -0.234 0.000
76 -0.129 0.000
77 0.243 0.000
78 -1.702 0.001
79 -0.925 0.000
80 -0.181 0.000
81 1.065 0.002
82 -1.730 0.001
83 1.072 0.007
84 -0.219 0.000
85 -1.637 0.003
86 0.335 0.000
87 -2.167 0.006
88 -0.488 0.000
89 0.272 0.000
90 1.283 0.002
91 -1.206 0.000
92 -0.157 0.000
93 -0.764 0.001
94 -1.522 0.000
95 0.976 0.000
96 0.016 0.000
97 0.510 0.000
98 -1.001 0.001
99 -1.317 0.000
100 0.810 0.001
|Studentized Residual| ≥ 3, Prob ≤ 0.0027
Cook's D ≥ 4 / n = 0.004

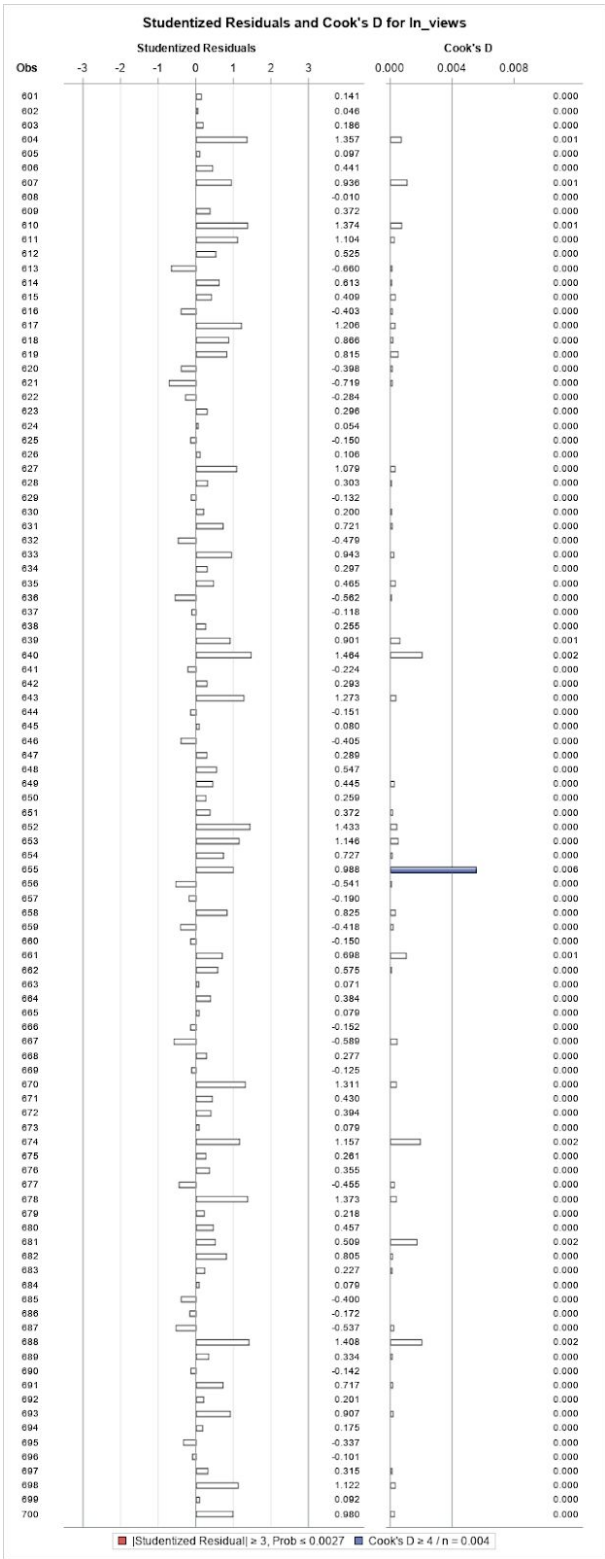

Studentized Residuals and Cook's D for ln_views
Studentized Residuals
Cook's D
Obs
-3 -2 -1 0 1 2 3
0.000 0.004 0.008
101 -0.248 0.000
102 -0.337 0.001
103 -1.152 0.000
104 0.638 0.000
105 -0.310 0.000
106 -1.925 0.007
107 0.224 0.000
108 -1.032 0.000
109 -2.076 0.003
110 -1.110 0.002
111 0.331 0.000
112 -1.575 0.005
113 0.318 0.000
114 -0.730 0.000
115 -1.777 0.005
116 1.318 0.001
117 1.116 0.001
118 -0.580 0.000
119 -0.286 0.000
120 -0.100 0.000
121 0.349 0.000
122 0.525 0.001
123 -0.521 0.000
124 0.240 0.000
125 -0.446 0.000
126 1.624 0.001
127 -0.699 0.001
128 -1.815 0.021
129 -3.306 0.068
130 -1.700 0.002
131 1.007 0.000
132 0.858 0.000
133 -1.846 0.001
134 -0.333 0.000
135 -0.442 0.000
136 -1.418 0.004
137 -0.237 0.000
138 -0.285 0.000
139 0.807 0.001
140 0.311 0.000
141 0.058 0.000
142 -0.673 0.001
143 -0.756 0.000
144 0.054 0.000
145 -3.053 0.013
146 -1.720 0.004
147 -2.657 0.010
148 -0.468 0.000
149 -2.084 0.001
150 -2.192 0.001
151 0.804 0.000
152 -1.668 0.001
153 0.662 0.000
154 -1.072 0.000
155 -1.469 0.014
156 -3.032 0.013
157 -1.230 0.000
158 0.849 0.001
159 0.746 0.001
160 -2.472 0.001
161 -1.007 0.001
162 -1.883 0.001
163 -0.713 0.000
164 -0.740 0.001
165 -1.020 0.000
166 -1.252 0.002
167 0.751 0.000
168 -0.071 0.000
169 -1.931 0.003
170 -1.788 0.001
171 0.078 0.000
172 -0.063 0.000
173 -0.834 0.000
174 0.184 0.000
175 -2.172 0.001
176 -1.386 0.000
177 -0.110 0.000
178 0.241 0.000
179 -0.087 0.000
180 -0.969 0.000
181 -0.240 0.000
182 0.073 0.000
183 -1.582 0.001
184 1.135 0.002
185 0.941 0.001
186 -0.240 0.000
187 -0.713 0.001
188 -0.524 0.000
189 0.545 0.000
190 -0.370 0.000
191 0.546 0.000
192 1.725 0.002
193 -0.105 0.000
194 -0.724 0.000
195 1.735 0.002
196 -0.482 0.000
197 0.293 0.000
198 -0.403 0.000
199 -0.562 0.000
200 0.312 0.001
|Studentized Residual| ≥ 3, Prob ≤ 0.0027
Cook's D ≥ 4 / n = 0.004

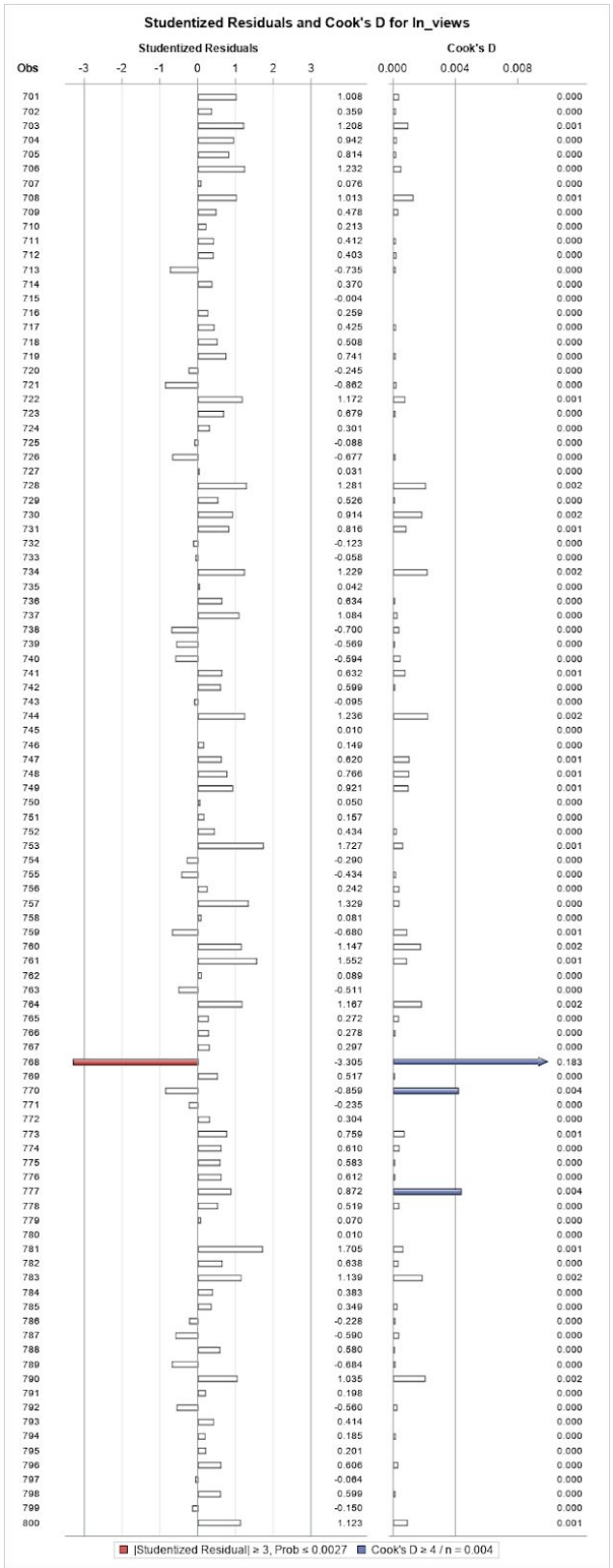

Studentized Residuals and Cook's D for ln_views
Studentized Residuals
Cook's D
Obs
-3 -2 -1 0 1 2 3
0.000 0.004 0.008
|Studentized Residual| ≥ 3, Prob ≤ 0.0027
Cook's D ≥ 4 / n = 0.004

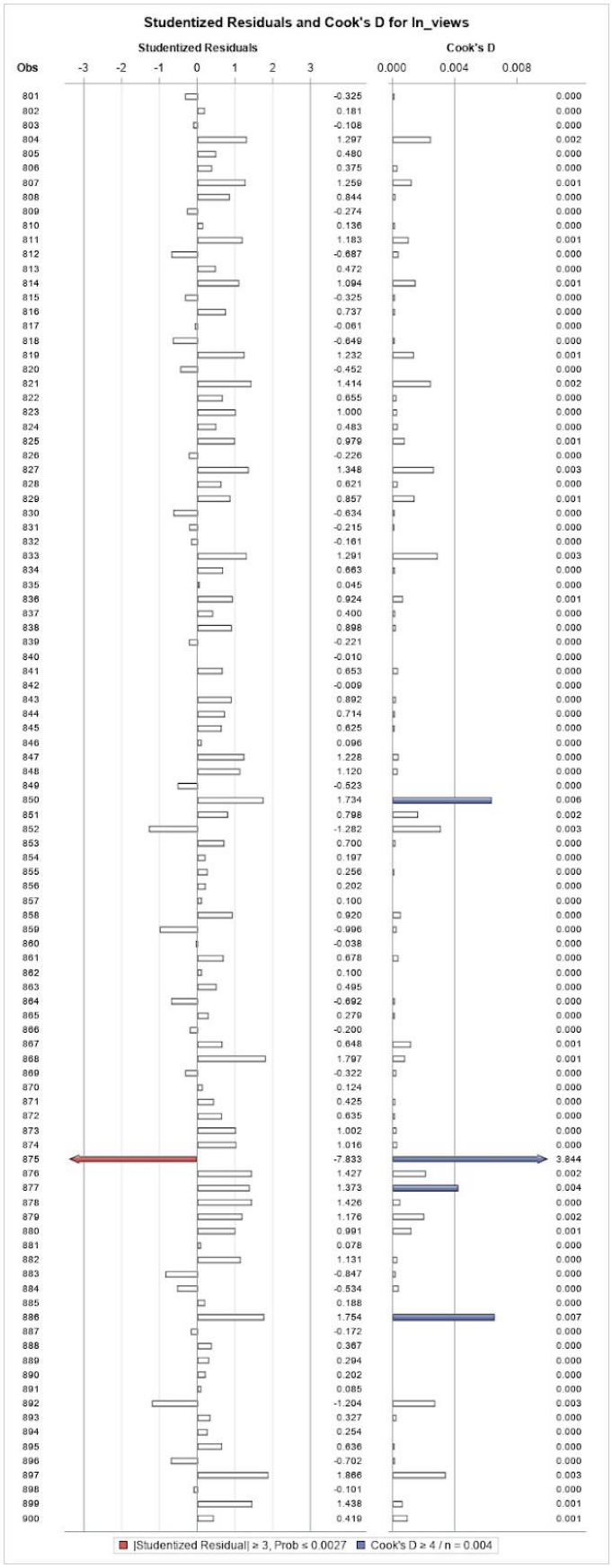

Studentized Residuals and Cook's D for ln_views
Studentized Residuals
Cook's D
Obs
-3 -2 -1 0 1 2 3
0.000 0.004 0.008
|Studentized Residual| ≥ 3, Prob ≤ 0.0027
Cook's D ≥ 4 / n = 0.004

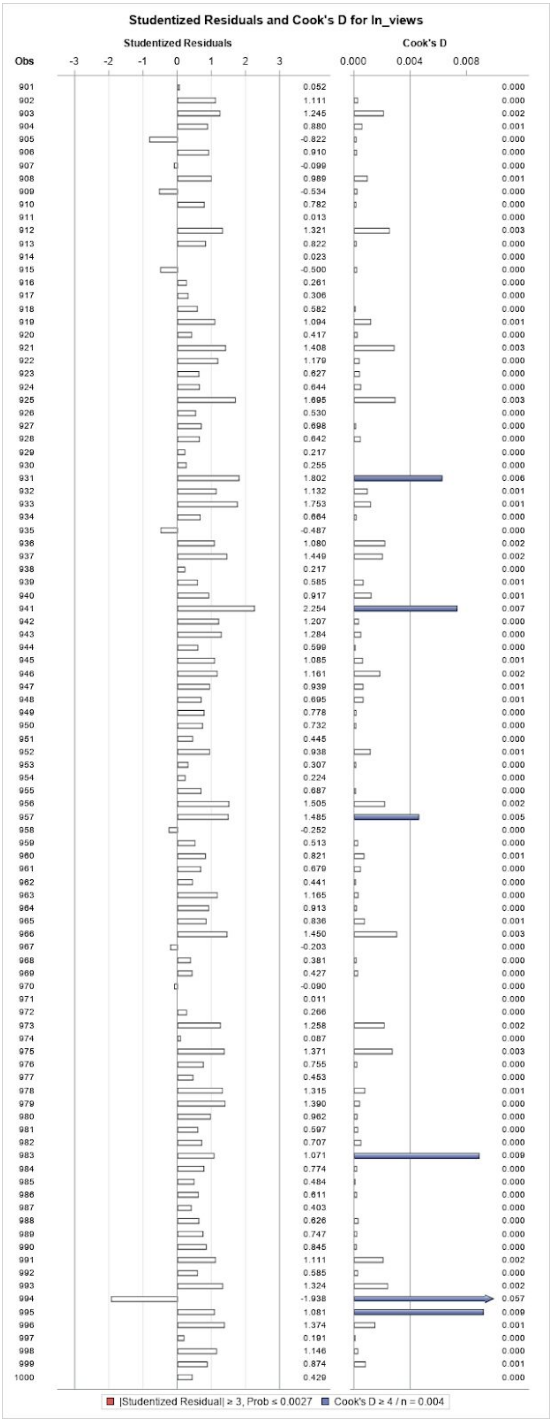

Studentized Residuals and Cook's D for ln_views
Studentized Residuals
Cook's D
Obs
-3 -2 -1 0 1 2 3
0.000 0.004 0.008
401 -3.368 0.014
402 0.151 0.000
403 0.473 0.000
404 0.332 0.000
405 -1.249 0.001
406 -1.009 0.001
407 -0.383 0.000
408 0.516 0.000
409 0.790 0.001
410 -0.829 0.001
411 0.224 0.000
412 -0.407 0.000
413 -0.470 0.000
414 1.790 0.001
415 -0.320 0.000
416 -0.679 0.000
417 0.729 0.000
418 -0.088 0.000
419 0.636 0.000
420 -0.365 0.000
421 -0.594 0.000
422 1.193 0.001
423 -0.738 0.000
424 0.556 0.000
425 0.673 0.000
426 0.386 0.000
427 -0.524 0.000
428 0.540 0.000
429 -1.335 0.000
430 -0.571 0.001
431 -0.400 0.000
432 -0.652 0.000
433 -1.531 0.000
434 -0.704 0.000
435 0.969 0.001
436 -0.024 0.000
437 -0.339 0.000
438 1.036 0.006
439 0.535 0.000
440 1.142 0.002
441 -1.763 0.001
442 0.992 0.000
443 -2.575 0.001
444 -1.151 0.000
445 -0.716 0.000
446 -1.469 0.000
447 -1.431 0.004
448 -0.476 0.000
449 -0.490 0.000
450 -0.392 0.000
451 0.366 0.000
452 0.364 0.000
453 -3.380 0.016
454 -0.809 0.000
455 -1.514 0.000
456 -0.404 0.000
457 -1.210 0.000
458 0.695 0.000
459 -0.023 0.000
460 1.143 0.002
461 1.226 0.000
462 -0.380 0.000
463 -0.086 0.000
464 -1.192 0.001
465 -2.224 0.001
466 0.704 0.001
467 -1.279 0.001
468 1.678 0.006
469 0.356 0.000
470 0.266 0.000
471 -1.581 0.001
472 0.899 0.000
473 -1.232 0.002
474 -0.426 0.000
475 -1.332 0.000
476 -1.072 0.001
477 0.457 0.000
478 -0.978 0.001
479 -0.507 0.000
480 -0.777 0.000
481 0.122 0.000
482 0.147 0.000
483 0.064 0.000
484 0.612 0.000
485 -0.445 0.000
486 -0.296 0.000
487 -1.137 0.000
488 -0.412 0.000
489 -0.026 0.000
490 0.523 0.000
491 0.630 0.000
492 -0.178 0.000
493 -1.402 0.000
494 0.864 0.000
495 -0.665 0.000
496 -1.990 0.001
497 0.276 0.000
498 -0.155 0.000
499 -0.873 0.001
500 -0.033 0.000
|Studentized Residual| ≥ 3, Prob ≤ 0.0027
Cook's D ≥ 4 / n = 0.004


Studentized Residuals and Cook's D for ln_views
Studentized Residuals
Cook's D
Obs
-3 -2 -1 0 1 2 3
0.000 0.004 0.008
501 -1.014 0.001
502 0.381 0.000
503 -0.828 0.000
504 0.213 0.000
505 0.545 0.000
506 -0.463 0.000
507 -0.805 0.000
508 0.195 0.000
509 -0.950 0.000
510 0.023 0.000
511 -0.107 0.000
512 0.763 0.000
513 -0.891 0.001
514 0.715 0.000
515 -0.465 0.000
516 -2.632 1.102
517 -0.221 0.000
518 -0.948 0.001
519 -0.161 0.000
520 -1.370 0.000
521 1.172 0.003
522 0.253 0.000
523 0.453 0.000
524 0.407 0.000
525 -1.056 0.000
526 1.175 0.001
527 -0.452 0.000
528 -0.522 0.000
529 -0.080 0.000
530 -0.453 0.000
531 0.272 0.000
532 0.441 0.000
533 0.370 0.000
534 0.034 0.000
535 -0.586 0.000
536 0.525 0.000
537 -0.384 0.000
538 1.105 0.000
539 -0.217 0.000
540 -0.219 0.000
541 0.035 0.000
542 0.283 0.000
543 -0.401 0.000
544 0.571 0.000
545 0.554 0.000
546 1.335 0.003
547 0.400 0.000
548 -0.158 0.000
549 1.312 0.000
550 -0.321 0.000
551 -0.791 0.000
552 1.259 0.002
553 0.741 0.000
554 0.584 0.000
555 -0.019 0.000
556 1.546 0.015
557 -0.277 0.000
558 0.584 0.000
559 0.842 0.000
560 0.292 0.000
561 1.034 0.001
562 1.201 0.001
563 0.369 0.000
564 0.792 0.000
565 -0.155 0.000
566 -0.523 0.000
567 0.892 0.001
568 0.281 0.000
569 -0.350 0.000
570 0.316 0.000
571 1.339 0.021
572 -0.263 0.000
573 0.645 0.000
574 -0.271 0.000
575 0.670 0.000
576 -0.286 0.000
577 0.666 0.000
578 1.568 0.015
579 0.949 0.001
580 0.436 0.000
581 0.376 0.000
582 0.192 0.000
583 -0.317 0.000
584 0.453 0.000
585 -0.519 0.001
586 1.577 0.015
587 0.090 0.000
588 -0.500 0.000
589 0.318 0.000
590 0.544 0.000
591 0.074 0.000
592 -0.162 0.000
593 0.315 0.000
594 -0.820 0.000
595 -0.839 0.000
596 -0.721 0.000
597 0.179 0.000
598 0.580 0.000
599 0.807 0.000
600 0.935 0.001
|Studentized Residual| ≥ 3, Prob ≤ 0.0027
Cook's D ≥ 4 / n = 0.004


Studentized Residuals and Cook's D for ln_views
Studentized Residuals
Cook's D
Obs
-3 -2 -1 0 1 2 3
0.000 0.004 0.008
601 0.141 0.000
602 0.046 0.000
603 0.186 0.000
604 1.357 0.001
605 0.097 0.000
606 0.441 0.000
607 0.936 0.001
608 -0.010 0.000
609 0.372 0.000
610 1.374 0.001
611 1.104 0.000
612 0.525 0.000
613 -0.660 0.000
614 0.613 0.000
615 0.409 0.000
616 -0.403 0.000
617 1.206 0.000
618 0.866 0.000
619 0.815 0.000
620 -0.398 0.000
621 -0.719 0.000
622 -0.284 0.000
623 0.296 0.000
624 0.054 0.000
625 -0.150 0.000
626 0.106 0.000
627 1.079 0.000
628 0.303 0.000
629 -0.132 0.000
630 0.200 0.000
631 0.721 0.000
632 -0.479 0.000
633 0.943 0.000
634 0.297 0.000
635 0.465 0.000
636 -0.562 0.000
637 -0.118 0.000
638 0.255 0.000
639 0.901 0.001
640 1.464 0.002
641 -0.224 0.000
642 0.293 0.000
643 1.273 0.000
644 -0.151 0.000
645 0.080 0.000
646 -0.405 0.000
647 0.289 0.000
648 0.547 0.000
649 0.445 0.000
650 0.259 0.000
651 0.372 0.000
652 1.433 0.000
653 1.146 0.000
654 0.727 0.000
655 0.988 0.006
656 -0.541 0.000
657 -0.190 0.000
658 0.825 0.000
659 -0.418 0.000
660 -0.150 0.000
661 0.698 0.001
662 0.575 0.000
663 0.071 0.000
664 0.384 0.000
665 0.079 0.000
666 -0.152 0.000
667 -0.589 0.000
668 0.277 0.000
669 -0.125 0.000
670 1.311 0.000
671 0.430 0.000
672 0.394 0.000
673 0.079 0.000
674 1.157 0.002
675 0.261 0.000
676 0.355 0.000
677 -0.455 0.000
678 1.373 0.000
679 0.218 0.000
680 0.457 0.000
681 0.509 0.002
682 0.805 0.000
683 0.227 0.000
684 0.079 0.000
685 -0.400 0.000
686 -0.172 0.000
687 -0.537 0.000
688 1.408 0.002
689 0.334 0.000
690 -0.142 0.000
691 0.717 0.000
692 0.201 0.000
693 0.907 0.000
694 0.175 0.000
695 -0.337 0.000
696 -0.101 0.000
697 0.315 0.000
698 1.122 0.000
699 0.092 0.000
700 0.980 0.000
|Studentized Residual| ≥ 3, Prob ≤ 0.0027
Cook's D ≥ 4 / n = 0.004


Studentized Residuals and Cook's D for ln_views
Studentized Residuals
Cook's D
Obs
-3 -2 -1 0 1 2 3
0.000 0.004 0.008
701 1.008 0.000
702 0.359 0.000
703 1.208 0.001
704 0.942 0.000
705 0.814 0.000
706 1.232 0.000
707 0.076 0.000
708 1.013 0.001
709 0.478 0.000
710 0.213 0.000
711 0.412 0.000
712 0.403 0.000
713 -0.735 0.000
714 0.370 0.000
715 -0.004 0.000
716 0.259 0.000
717 0.425 0.000
718 0.508 0.000
719 0.741 0.000
720 -0.245 0.000
721 -0.862 0.000
722 1.172 0.001
723 0.679 0.000
724 0.301 0.000
725 -0.088 0.000
726 -0.677 0.000
727 0.031 0.000
728 1.281 0.002
729 0.526 0.000
730 0.914 0.002
731 0.816 0.001
732 -0.123 0.000
733 -0.058 0.000
734 1.229 0.002
735 0.042 0.000
736 0.834 0.000
737 1.084 0.000
738 -0.700 0.000
739 -0.569 0.000
740 -0.594 0.000
741 0.632 0.001
742 0.599 0.000
743 -0.095 0.000
744 1.236 0.002
745 0.010 0.000
746 0.149 0.000
747 0.620 0.001
748 0.766 0.001
749 0.921 0.001
750 0.050 0.000
751 0.157 0.000
752 0.434 0.000
753 1.727 0.001
754 -0.290 0.000
755 -0.434 0.000
756 0.242 0.000
757 1.329 0.000
758 0.081 0.000
759 -0.680 0.001
760 1.147 0.002
761 1.552 0.001
762 0.089 0.000
763 -0.511 0.000
764 1.167 0.002
765 0.272 0.000
766 0.278 0.000
767 0.297 0.000
768 -3.305 0.183
769 0.517 0.000
770 -0.859 0.004
771 -0.235 0.000
772 0.304 0.000
773 0.759 0.001
774 0.610 0.000
775 0.583 0.000
776 0.612 0.000
777 0.872 0.004
778 0.519 0.000
779 0.070 0.000
780 0.010 0.000
781 1.705 0.001
782 0.638 0.000
783 1.139 0.002
784 0.383 0.000
785 0.349 0.000
786 -0.228 0.000
787 -0.590 0.000
788 0.580 0.000
789 -0.684 0.000
790 1.035 0.002
791 0.198 0.000
792 -0.560 0.000
793 0.414 0.000
794 0.185 0.000
795 0.201 0.000
796 0.606 0.000
797 -0.064 0.000
798 0.599 0.000
799 -0.150 0.000
800 1.123 0.001
|Studentized Residual| ≥ 3, Prob ≤ 0.0027
Cook's D ≥ 4 / n = 0.004


Studentized Residuals and Cook's D for ln_views
Studentized Residuals
Cook's D
Obs
-3 -2 -1 0 1 2 3
0.000 0.004 0.008
|Studentized Residual| ≥ 3, Prob ≤ 0.0027
Cook's D ≥ 4 / n = 0.004


Studentized Residuals and Cook's D for ln_views
Studentized Residuals
Cook's D
Obs
-3 -2 -1 0 1 2 3
0.000 0.004 0.008
|Studentized Residual| ≥ 3, Prob ≤ 0.0027
Cook's D ≥ 4 / n = 0.004

FM.C

Second Full Model

The REG Procedure
Model: MODEL1
Dependent Variable: ln_views

Number of Observations Read	977
Number of Observations Used	977

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	10	976.29168	97.62917	66.40	<.0001
Error	966	1420.23907	1.47023		
Corrected Total	976	2396.53075			

Root MSE	1.21253	R-Square	0.4074
Dependent Mean	13.33420	Adj R-Sq	0.4012
Coeff Var	9.09338		

Variable Selection

VS.A

Backward Elimination: Step 5

Variable blog Removed: R-Square = 0.4265 and C(p) = 7.4327

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	759.04255	126.50709	89.99	<.0001
Error	726	1020.56754	1.40574		
Corrected Total	732	1779.61008			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	12.78000	0.07733	38395	27313.3	<.0001
film	0.57187	0.20489	10.95060	7.79	0.0054
entertainment	0.29301	0.09822	12.51017	8.90	0.0029
politics	-0.68306	0.18836	18.48575	13.15	0.0003
com_dis_dummy	0.91806	0.42600	6.52865	4.64	0.0315
likes	0.00000603	3.96797E-7	324.50285	230.84	<.0001
dislikes	0.00001135	0.00000556	5.85512	4.17	0.0416

VS.B

Forward Selection: Step 8

Variable comment_count Entered: R-Square = 0.4289 and C(p) = 8.3649

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	8	763.35243	95.41905	67.98	<.0001
Error	724	1016.25765	1.40367		
Corrected Total	732	1779.61008			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	13.05595	0.06145	63358	45137.7	<.0001
film	0.27862	0.19816	2.77507	1.98	0.1601
scitech	-0.41633	0.15129	10.62953	7.57	0.0061
politics	-0.96434	0.18174	39.51982	28.15	<.0001
info	-0.30797	0.12799	8.12733	5.79	0.0164
com_dis_dummy	0.95097	0.42946	6.88251	4.90	0.0271
likes	0.00000561	6.281439E-7	111.76830	79.63	<.0001
dislikes	0.00000773	0.00000623	2.15894	1.54	0.2153
comment_count	0.00000702	0.00000740	1.26179	0.90	0.3434

VS.C

Stepwise Selection: Step 7

Variable film Entered: R-Square = 0.4282 and C(p) = 6.3224

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	7	762.09064	108.87009	77.57	<.0001
Error	725	1017.51944	1.40348		
Corrected Total	732	1779.61008			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	13.06097	0.06122	63882	45516.9	<.0001
film	0.29080	0.19773	3.03576	2.16	0.1418
scitech	-0.41115	0.15118	10.38029	7.40	0.0067
politics	-0.96217	0.18171	39.34836	28.04	<.0001
info	-0.30213	0.12783	7.84032	5.59	0.0184
com_dis_dummy	0.89858	0.42586	6.24851	4.45	0.0352
likes	0.00000607	3.941106E-7	332.79901	237.12	<.0001
dislikes	0.00001040	0.00000556	4.91485	3.50	0.0617

VS.D

AdjR2

The REG Procedure
Model: MODEL1
Dependent Variable: new_y

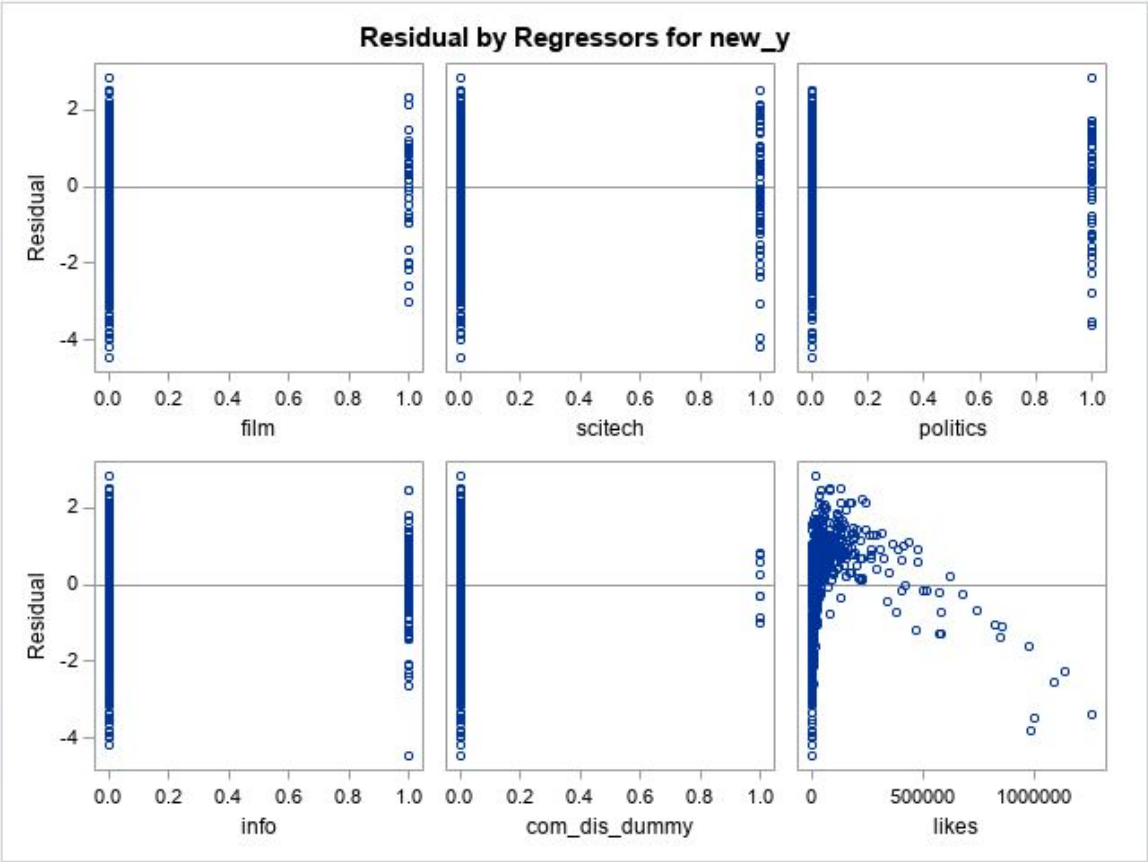
Adjusted R-Square Selection Method

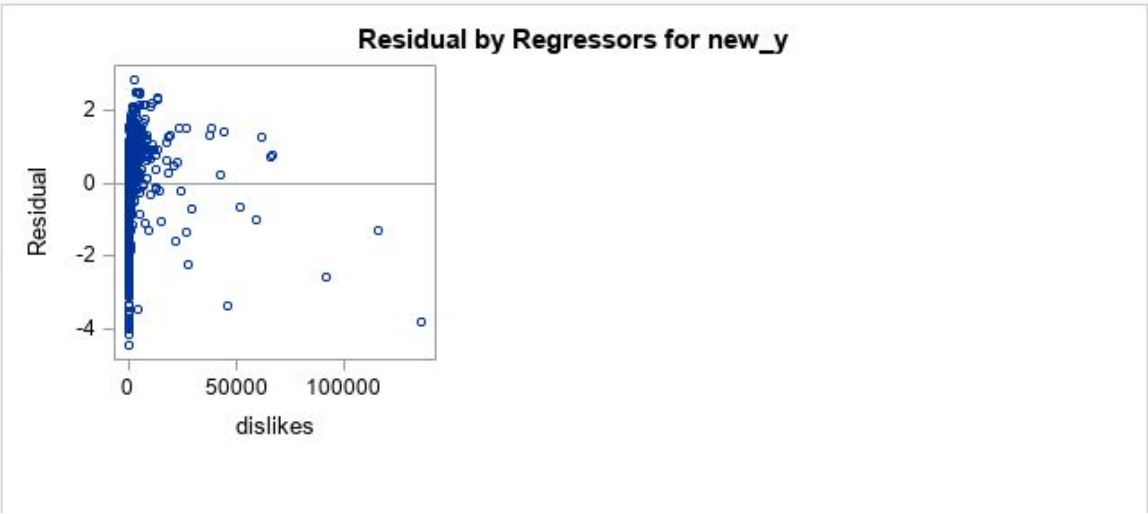
Number of Observations Read	977
Number of Observations Used	733
Number of Observations with Missing Values	244

Number in Model	Adjusted R-Square	R-Square	Variables in Model
7	0.4227	0.4282	film scitech politics info com_dis_dummy likes dislikes
7	0.4227	0.4282	film entertainment politics blog com_dis_dummy likes dislikes
8	0.4226	0.4289	film entertainment politics blog com_dis_dummy likes dislikes comment_count
8	0.4226	0.4289	film scitech politics info com_dis_dummy likes dislikes comment_count

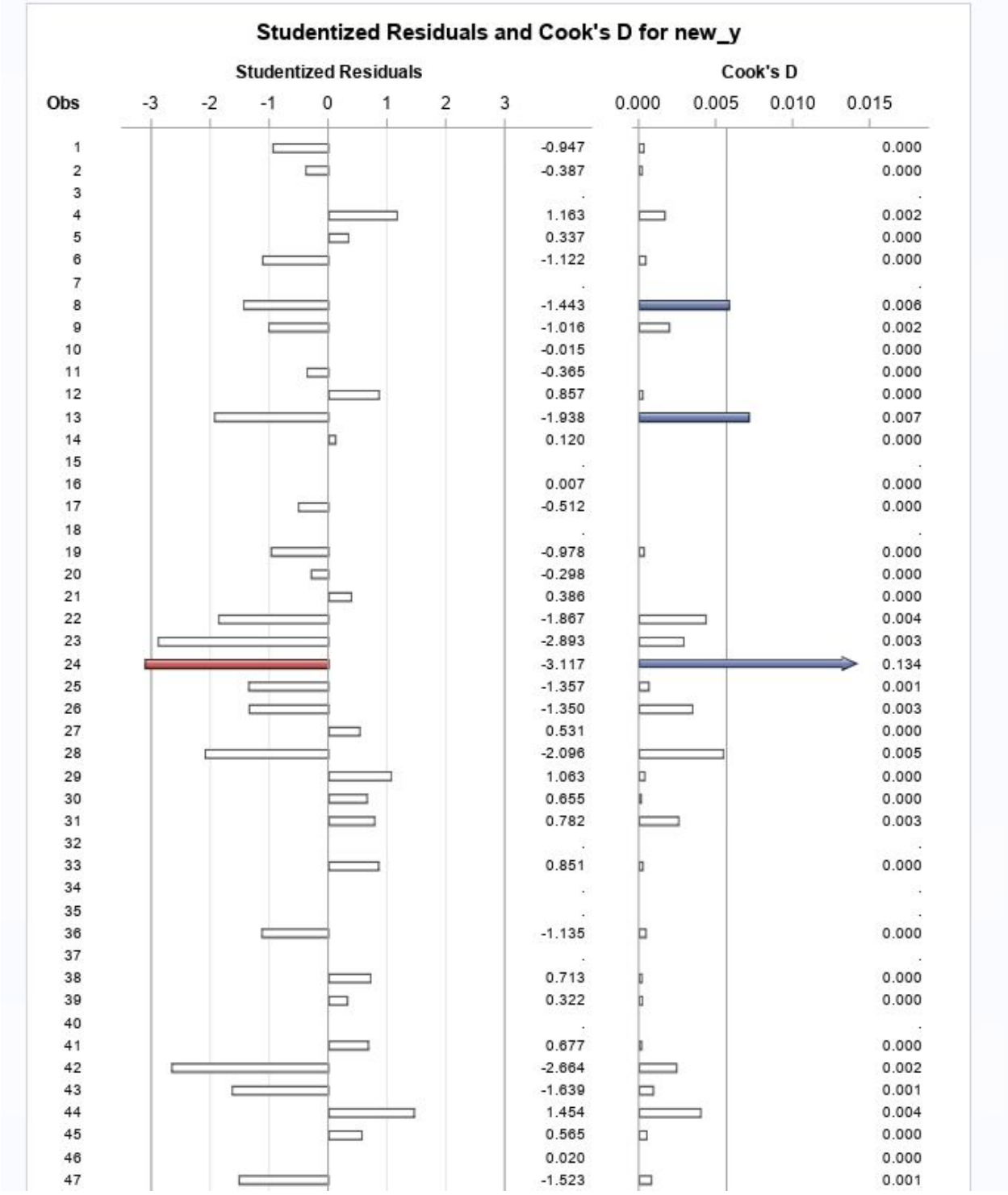
Model 1 Data

M1.A





M1.B


Studentized Residuals and Cook's D for new_y
Studentized Residuals
Cook's D
Obs
-3 -2 -1 0 1 2 3
0.000 0.005 0.010 0.015
1 -0.947 0.000
2 -0.387 0.000
3 . .
4 1.163 0.002
5 0.337 0.000
6 -1.122 0.000
7 . .
8 -1.443 0.006
9 -1.016 0.002
10 -0.015 0.000
11 -0.365 0.000
12 0.857 0.000
13 -1.938 0.007
14 0.120 0.000
15 . .
16 0.007 0.000
17 -0.512 0.000
18 . .
19 -0.978 0.000
20 -0.298 0.000
21 0.386 0.000
22 -1.867 0.004
23 -2.893 0.003
24 -3.117 0.134
25 -1.357 0.001
26 -1.350 0.003
27 0.531 0.000
28 -2.096 0.005
29 1.063 0.000
30 0.655 0.000
31 0.782 0.003
32 . .
33 0.851 0.000
34 . .
35 . .
36 -1.135 0.000
37 . .
38 0.713 0.000
39 0.322 0.000
40 . .
41 0.677 0.000
42 -2.664 0.002
43 -1.639 0.001
44 1.454 0.004
45 0.565 0.000
46 0.020 0.000
47 -1.523 0.001

M1.C

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	7	718.67502	102.66786	75.81	<.0001
Error	709	960.12808	1.35420		
Corrected Total	716	1678.80311			

Root MSE	1.16370	R-Square	0.4281
Dependent Mean	13.36710	Adj R-Sq	0.4224
Coeff Var	8.70571		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	13.06350	0.06046	216.07	<.0001
film	1	0.35741	0.19666	1.82	0.0696
scitech	1	-0.40957	0.15043	-2.72	0.0066
politics	1	-0.89183	0.18205	-4.90	<.0001
info	1	-0.26963	0.12716	-2.12	0.0343
com_dis_dummy	1	0.88374	0.41845	2.11	0.0350
likes	1	0.00000606	3.958538E-7	15.32	<.0001
dislikes	1	0.00000992	0.00000549	1.81	0.0711

M1.D

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	7	762.78172	108.96882	78.87	<.0001
Error	702	969.90544	1.38163		
Corrected Total	709	1732.68715			

Root MSE	1.17543	R-Square	0.4402
Dependent Mean	13.35864	Adj R-Sq	0.4346
Coeff Var	8.79901		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	13.05190	0.06229	209.54	<.0001
film	1	0.28789	0.19887	1.45	0.1482
scitech	1	-0.41821	0.15151	-2.76	0.0059
politics	1	-0.96528	0.18244	-5.29	<.0001
info	1	-0.32244	0.12793	-2.52	0.0119
com_dis_dummy	1	0.80863	0.42333	1.91	0.0565
likes	1	0.00000591	3.930789E-7	15.03	<.0001
dislikes	1	0.00002031	0.00000618	3.29	0.0011

M1.E

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	745.12473	149.02495	106.23	<.0001
Error	704	987.56243	1.40279		
Corrected Total	709	1732.68715			

Root MSE	1.18439	R-Square	0.4300
Dependent Mean	13.35864	Adj R-Sq	0.4260
Coeff Var	8.86612		

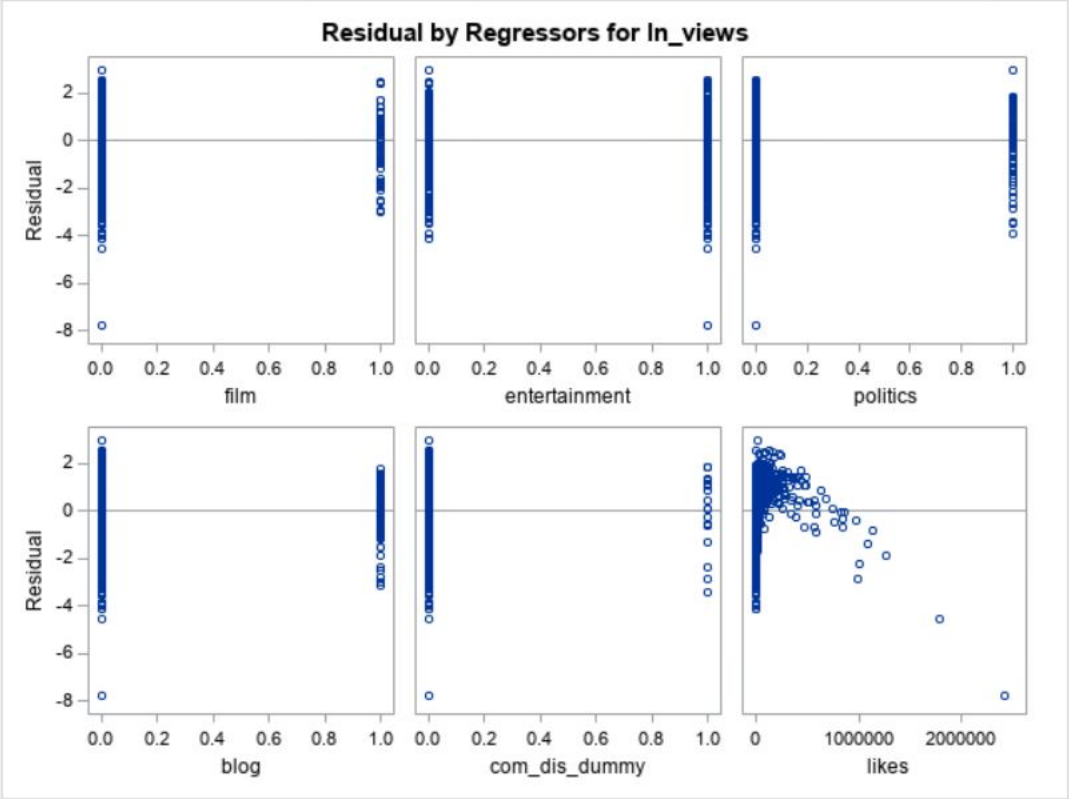
Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	13.08928	0.06030	217.06	<.0001
scitech	1	-0.45986	0.15166	-3.03	0.0025
politics	1	-0.98624	0.18311	-5.39	<.0001
info	1	-0.35861	0.12778	-2.81	0.0051
com_dis_dummy	1	0.96513	0.42305	2.28	0.0228
likes	1	0.00000664	3.232108E-7	20.54	<.0001

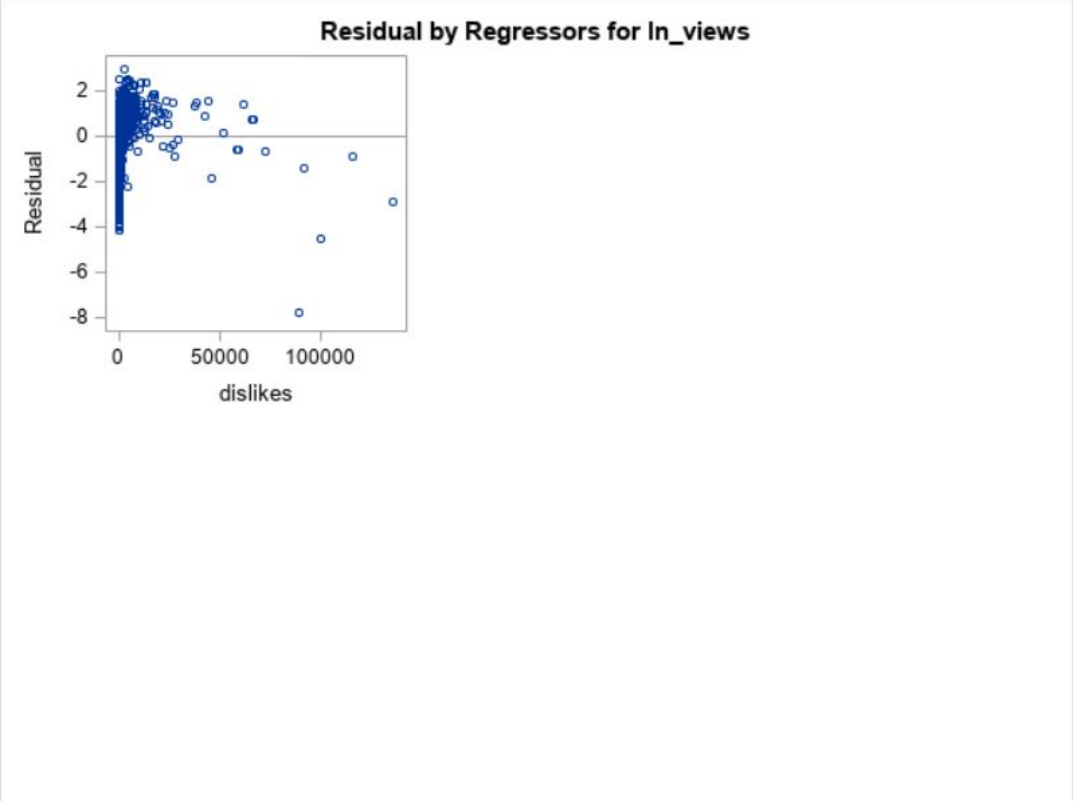
M1.F

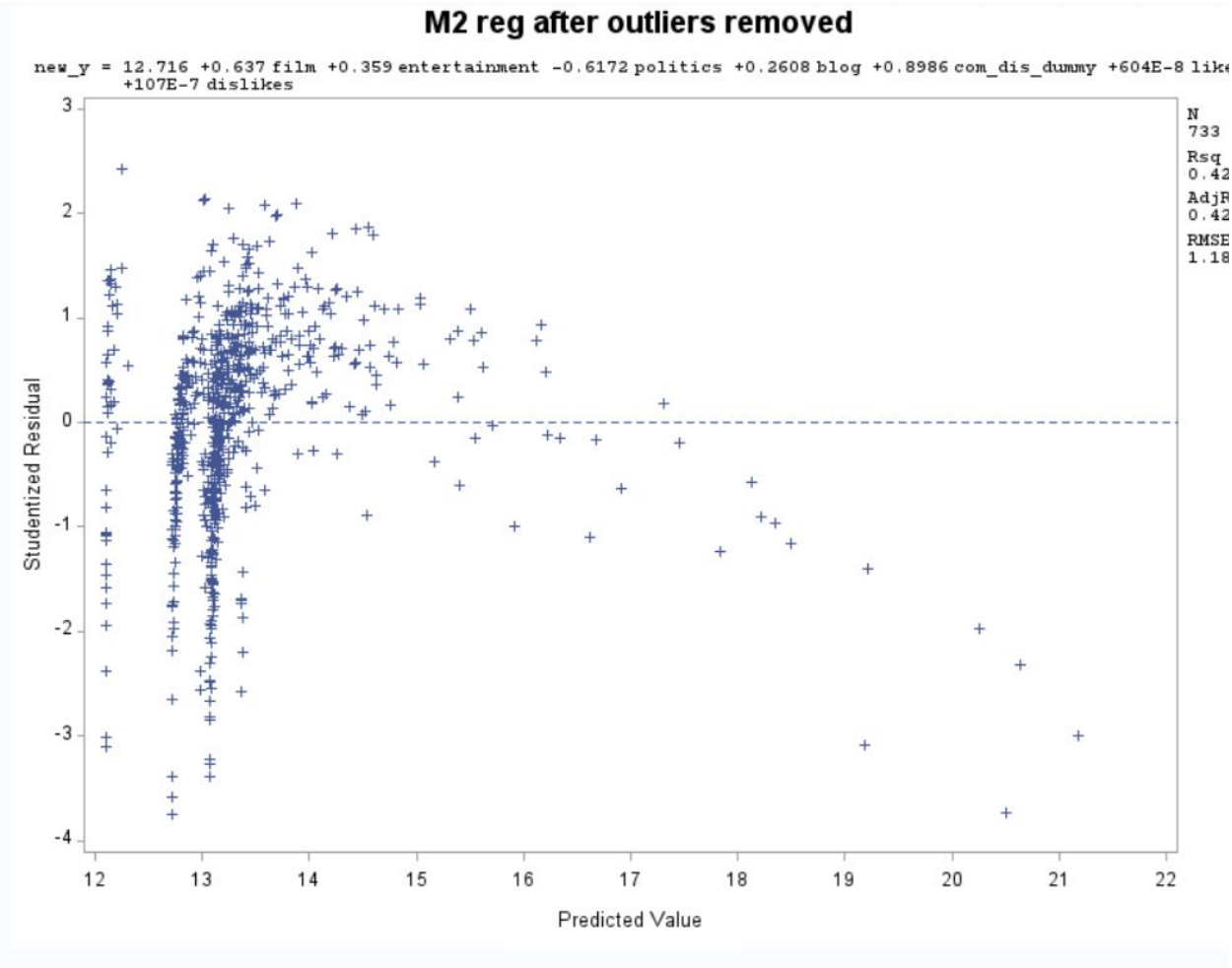
Pearson Correlation Coefficients Prob > |r| under H0: Rho=0 Number of Observations						
	new_y	scitech	politics	info	com_dis_dummy	likes
new_y	1.00000 710	-0.11692 0.0018 710	-0.20201 <.0001 710	-0.11656 0.0019 710	0.04855 0.1963 710	0.62911 <.0001 710
scitech	-0.11692 0.0018 710	1.00000 951	-0.09337 0.0040 951	-0.14382 <.0001 951	-0.04058 0.2112 951	-0.08675 0.0074 951
politics	-0.20201 <.0001 710	-0.09337 0.0040 951	1.00000 951	-0.11301 0.0005 951	0.07577 0.0194 951	-0.11472 0.0004 951
info	-0.11656 0.0019 710	-0.14382 <.0001 951	-0.11301 0.0005 951	1.00000 951	-0.04912 0.1301 951	-0.10130 0.0018 951
com_dis_dummy	0.04855 0.1963 710	-0.04058 0.2112 951	0.07577 0.0194 951	-0.04912 0.1301 951	1.00000 951	-0.03017 0.3527 951
likes	0.62911 <.0001 710	-0.08675 0.0074 951	-0.11472 0.0004 951	-0.10130 0.0018 951	-0.03017 0.3527 951	1.00000 951

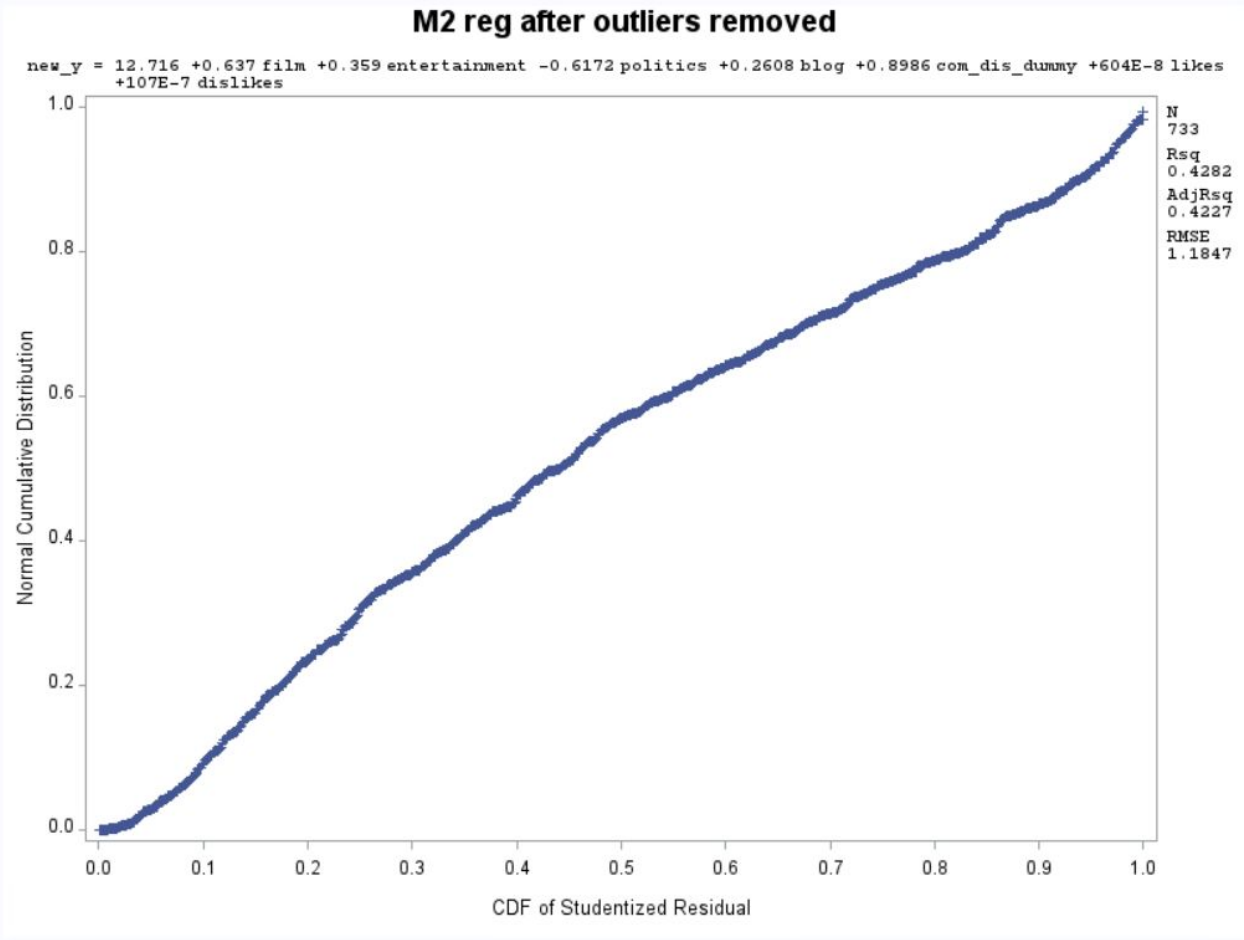
Model 2 Data

M2.A

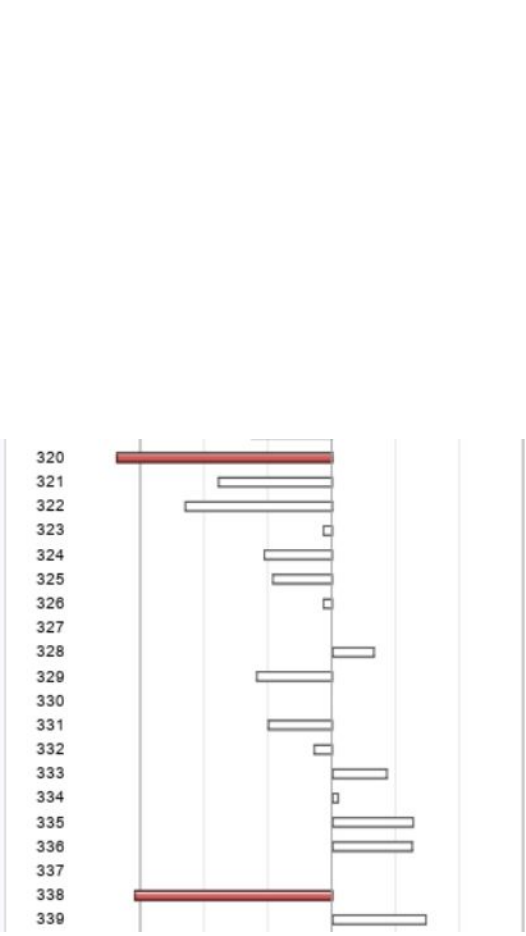

Residual by Regressors for ln_views
Residual
film
entertainment
politics
blog
com_dis_dummy
likes

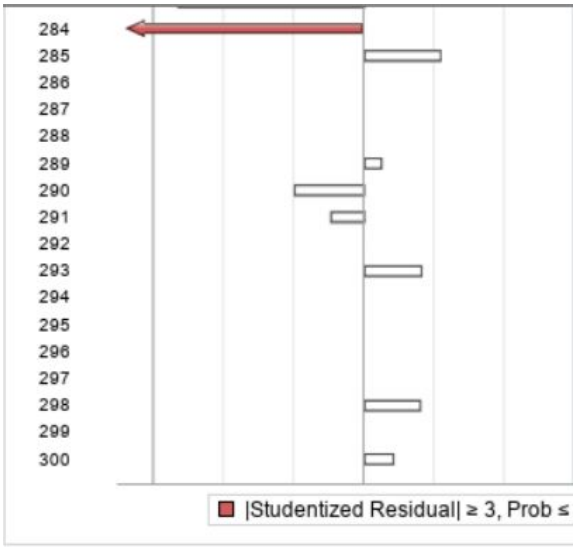

Residual by Regressors for ln_views
Residual
dislikes

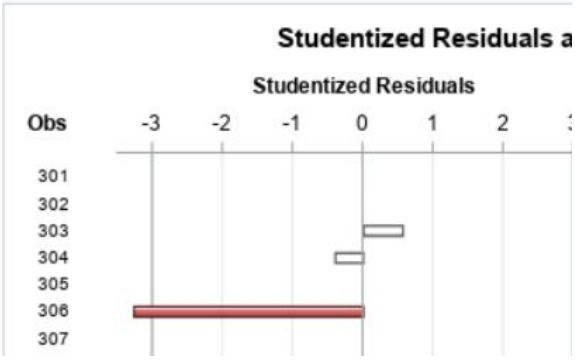

M2 reg after outliers removed
new_y = 12.716 +0.637 film +0.359 entertainment -0.6172 politics +0.2608 blog +0.8986 com_dis_dummy +604E-8 like
+107E-7 dislikes
N
733
Rsq
0.42
AdjR
0.42
RMSE
1.18
Studentized Residual
Predicted Value
3
2
1
0
-1
-2
-3
-4
12
13
14
15
16
17
18
19
20
21
22

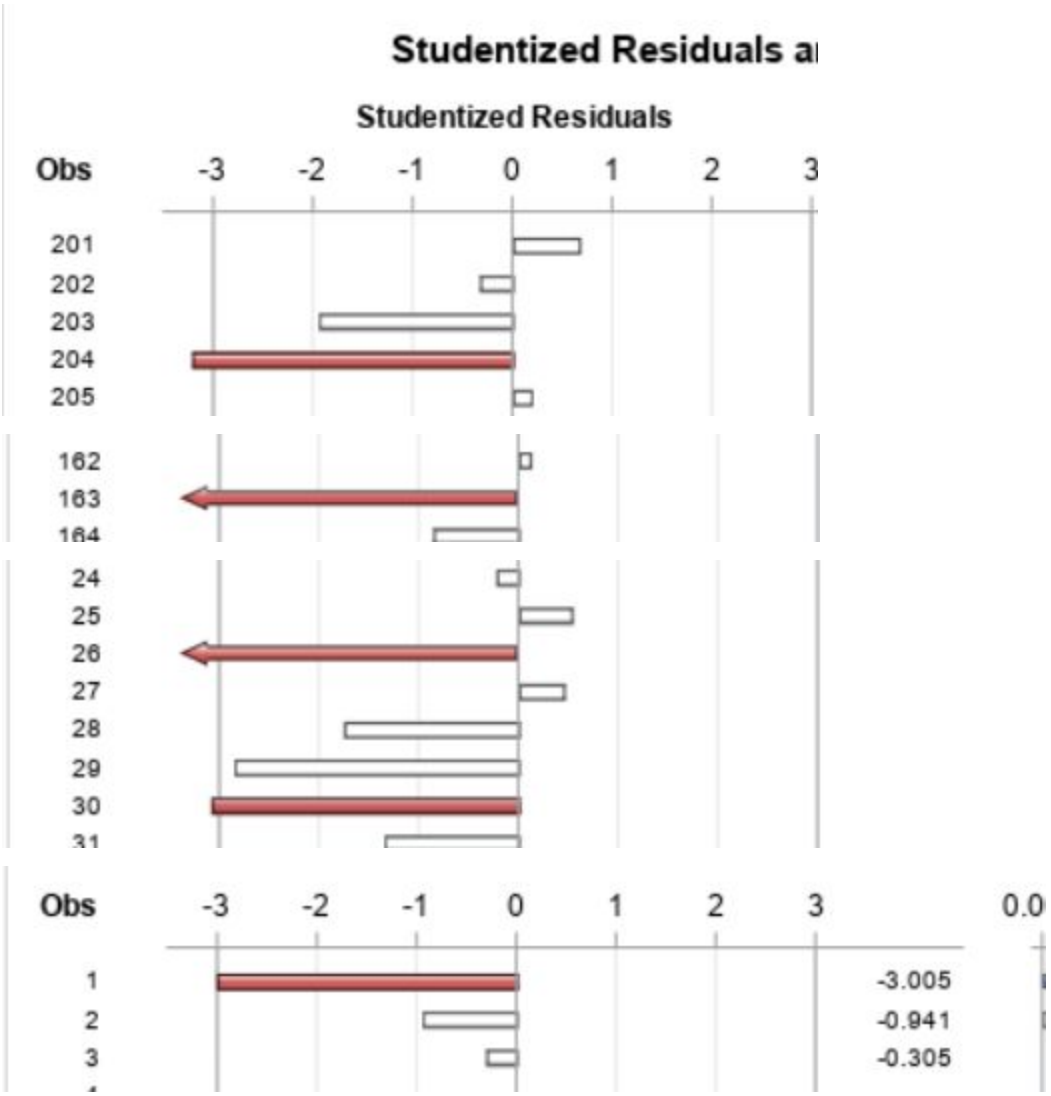

M2 reg after outliers removed
new_y = 12.716 +0.637 film +0.359 entertainment -0.6172 politics +0.2608 blog +0.8986 com_dis_dummy +604E-8 likes
+107E-7 dislikes
N
733
Rsq
0.4282
AdjRsq
0.4227
RMSE
1.1847
Normal Cumulative Distribution
CDF of Studentized Residual
0.0
0.2
0.4
0.6
0.8
1.0
0.1
0.3
0.5
0.7
0.9

M2.B

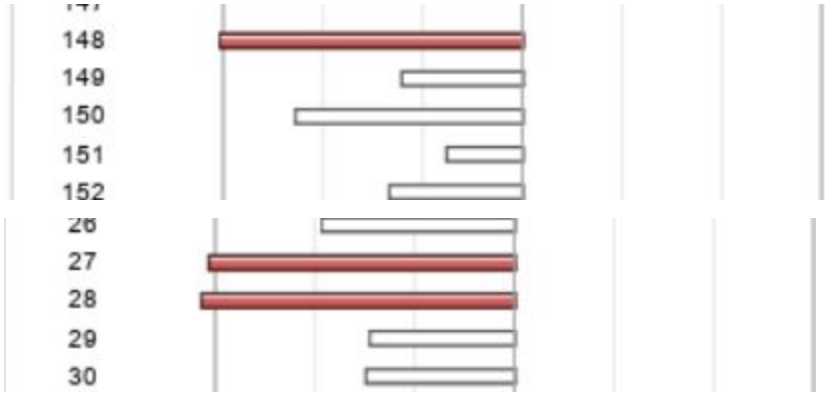

320
321
322
323
324
325
326
327
328
329
330
331
332
333
334
335
336
337
338
339


284
285
286
287
288
289
290
291
292
293
294
295
296
297
298
299
300
|Studentized Residual| ≥ 3, Prob ≤

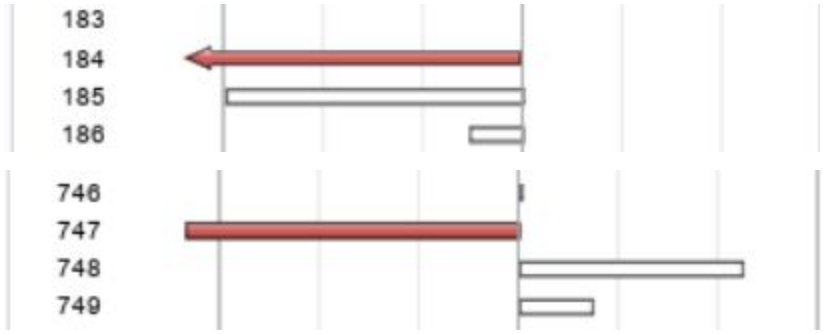

Studentized Residuals a
Studentized Residuals
Obs
-3
-2
-1
0
1
2
301
302
303
304
305
306
307



New outliers that appear after the first round of removing outliers:



More outliers that were removed after second round of outliers:



M2.C

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	7	715.52323	102.21760	93.33	<.0001
Error	709	776.51554	1.09523		
Corrected Total	716	1492.03876			

Root MSE	1.04653	R-Square	0.4796
Dependent Mean	13.38993	Adj R-Sq	0.4744
Coeff Var	7.81581		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	12.73477	0.07937	160.46	<.0001
film	1	0.50579	0.18545	2.73	0.0065
entertainment	1	0.28978	0.09628	3.01	0.0027
politics	1	-0.50500	0.17424	-2.90	0.0039
blog	1	0.09281	0.15778	0.59	0.5566
com_dis_dummy	1	0.62216	0.37860	1.64	0.1008
likes	1	0.00000694	4.149868E-7	16.73	<.0001
dislikes	1	0.00003482	0.00000689	5.05	<.0001

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	712.14148	142.42830	129.85	<.0001
Error	711	779.89728	1.09690		
Corrected Total	716	1492.03876			

Root MSE	1.04733	R-Square	0.4773
Dependent Mean	13.38993	Adj R-Sq	0.4736
Coeff Var	7.82178		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	12.75887	0.06936	183.96	<.0001
film	1	0.48042	0.18122	2.65	0.0082
entertainment	1	0.27209	0.08738	3.11	0.0019
politics	1	-0.50479	0.16909	-2.99	0.0029
likes	1	0.00000687	4.132622E-7	16.62	<.0001
dislikes	1	0.00003712	0.00000676	5.50	<.0001

Model 3 Data

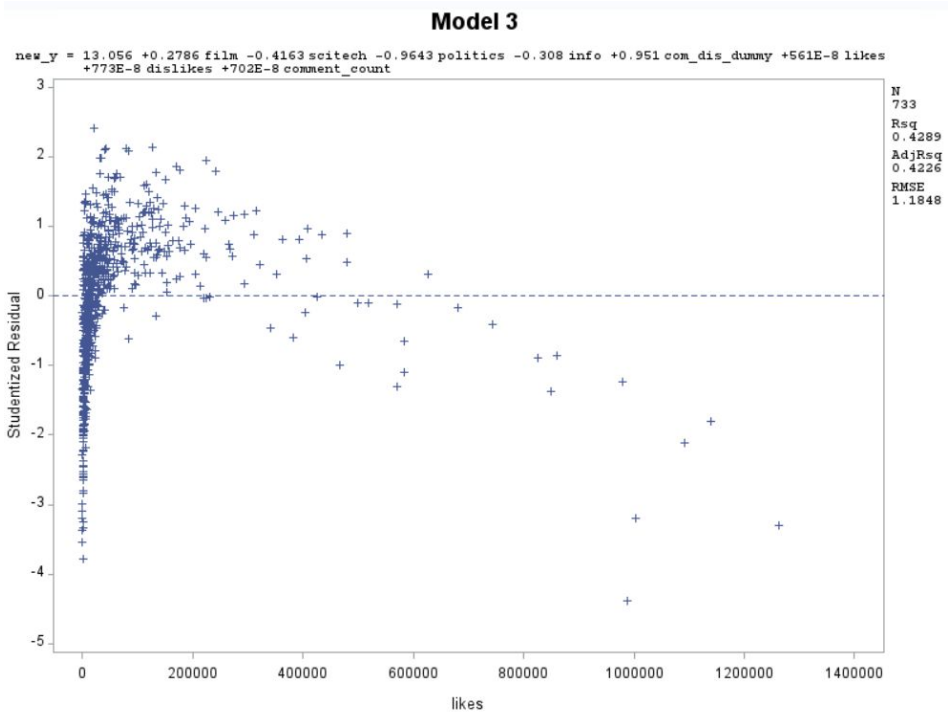
M3.A -

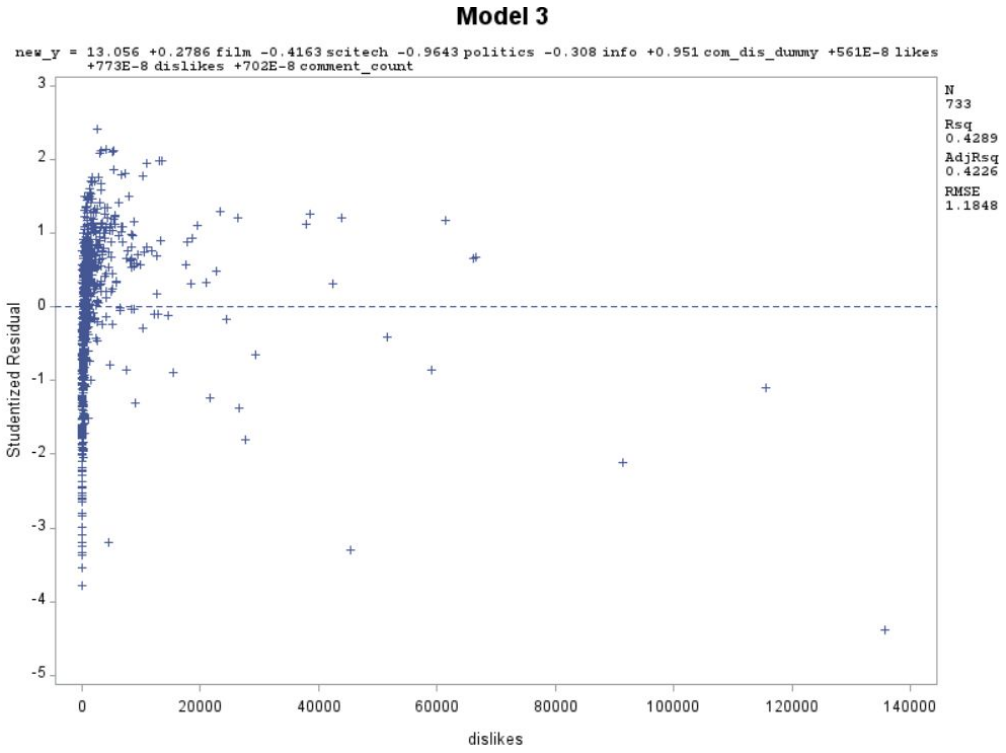
Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	8	763.35243	95.41905	67.98	<.0001
Error	724	1016.25765	1.40367		
Corrected Total	732	1779.61008			

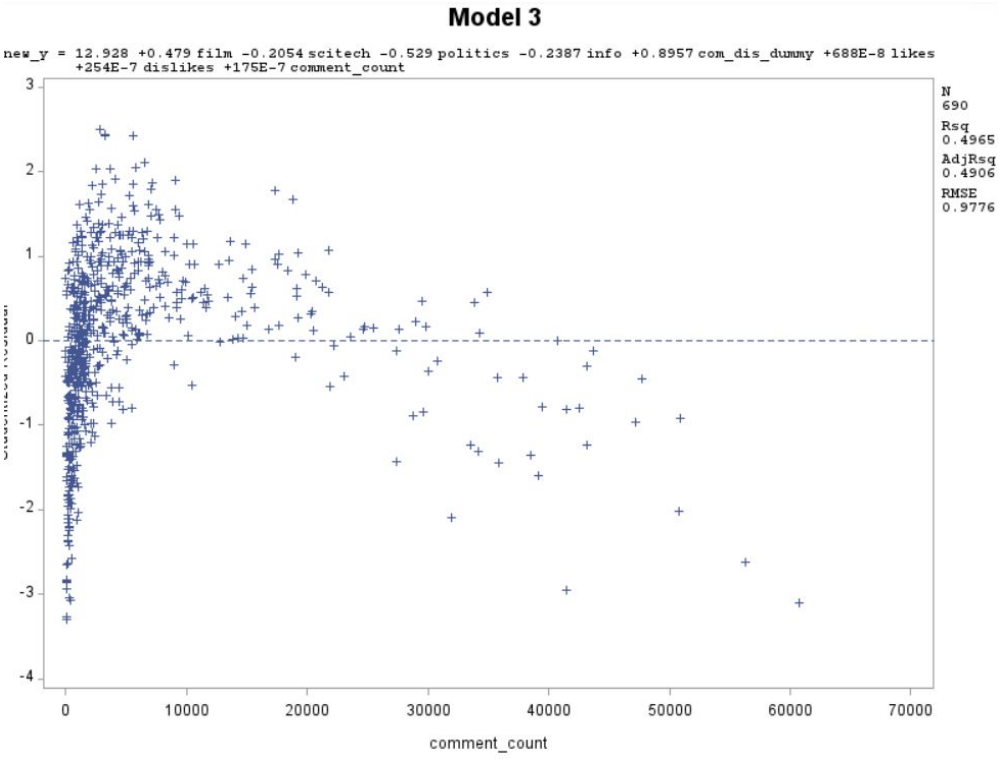
Root MSE	1.18477	R-Square	0.4289
Dependent Mean	13.36009	Adj R-Sq	0.4226
Coeff Var	8.86795		

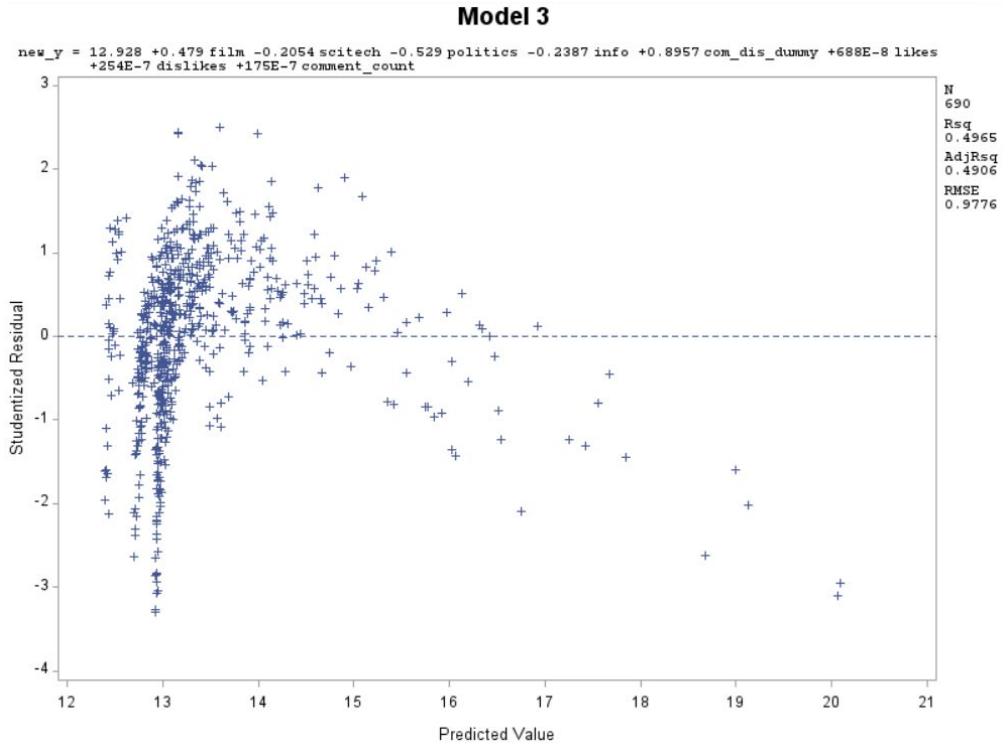
Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	13.05595	0.06145	212.46	<.0001
film	1	0.27862	0.19816	1.41	0.1601
scitech	1	-0.41633	0.15129	-2.75	0.0061
politics	1	-0.96434	0.18174	-5.31	<.0001
info	1	-0.30797	0.12799	-2.41	0.0164
com_dis_dummy	1	0.95097	0.42946	2.21	0.0271
likes	1	0.00000561	6.281439E-7	8.92	<.0001
dislikes	1	0.00000773	0.00000623	1.24	0.2153
comment_count	1	0.00000702	0.00000740	0.95	0.3434

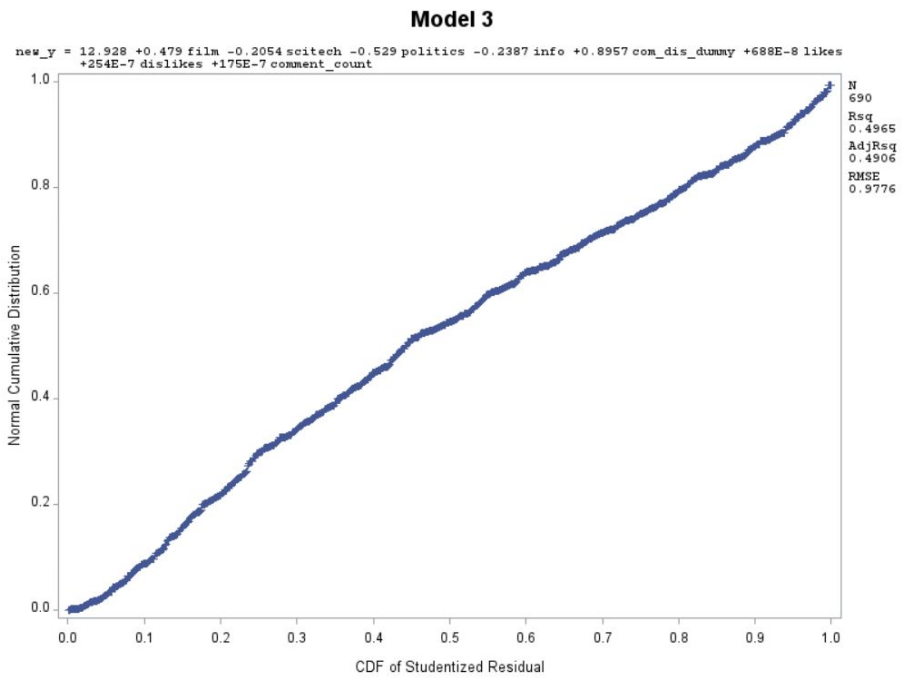
M3.B-


Model 3
new_y = 13.056 +0.2786 film -0.4163 scitech -0.9643 politics -0.308 info +0.951 com_dis_dummy +561E-8 likes
+773E-8 dislikes +702E-8 comment_count
N
733
Rsq
0.4289
AdjRsq
0.4226
RMSE
1.1848
Studentized Residual
likes
0
200000
400000
600000
800000
1000000
1200000
1400000

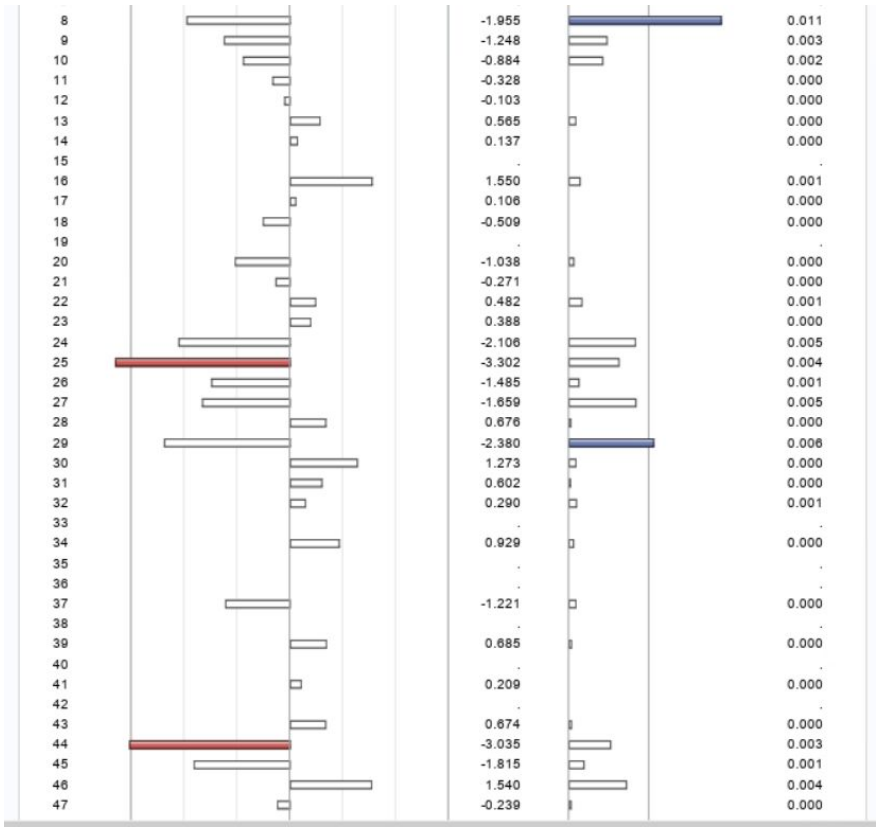

Model 3
new_y = 13.056 +0.2786 film -0.4163 scitech -0.9643 politics -0.308 info +0.951 com_dis_dummy +561E-8 likes
+773E-8 dislikes +702E-8 comment_count
N
733
Rsq
0.4289
AdjRsq
0.4226
RMSE
1.1848
Studentized Residual
dislikes
0
20000
40000
60000
80000
100000
120000
140000


Model 3
new_y = 12.928 +0.479 film -0.2054 scitech -0.529 politics -0.2387 info +0.8957 com_dis_dummy +688E-8 likes
+254E-7 dislikes +175E-7 comment_count
N
690
Rsq
0.4965
AdjRsq
0.4906
RMSE
0.9776
comment_count

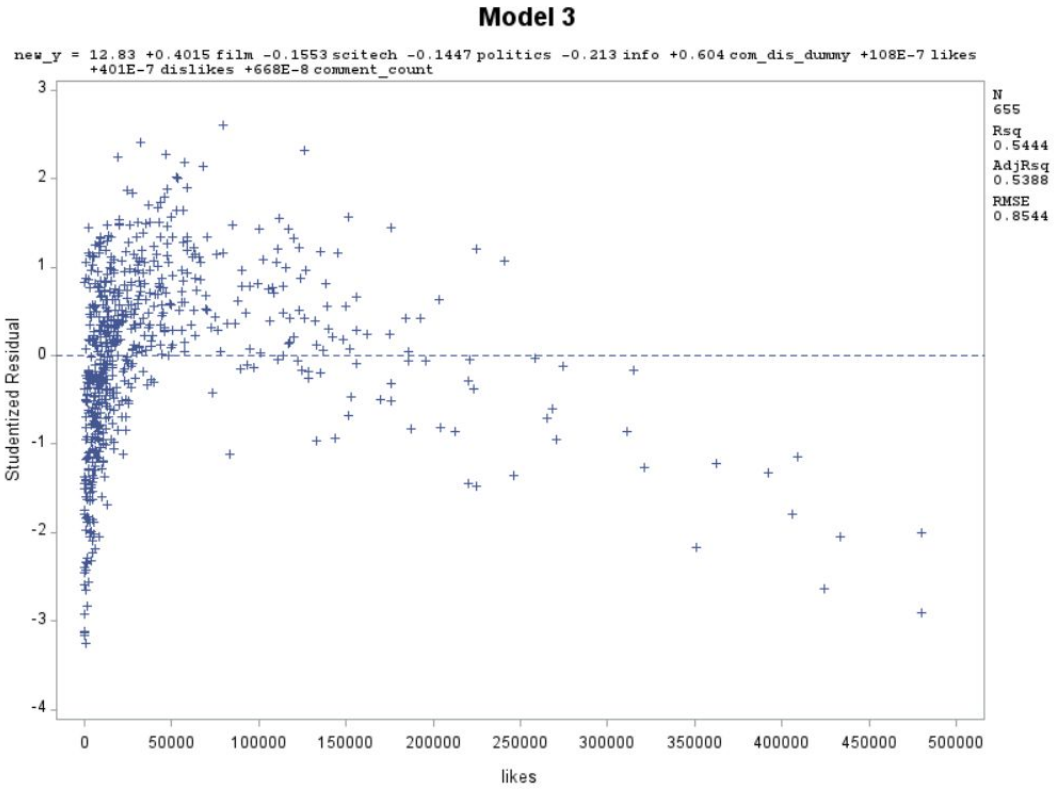

Model 3
new_y = 12.928 +0.479 film -0.2054 scitech -0.529 politics -0.2387 info +0.8957 com_dis_dummy +688E-8 likes
+254E-7 dislikes +175E-7 comment_count
N
690
Rsq
0.4965
AdjRsq
0.4906
RMSE
0.9776
Studentized Residual
Predicted Value

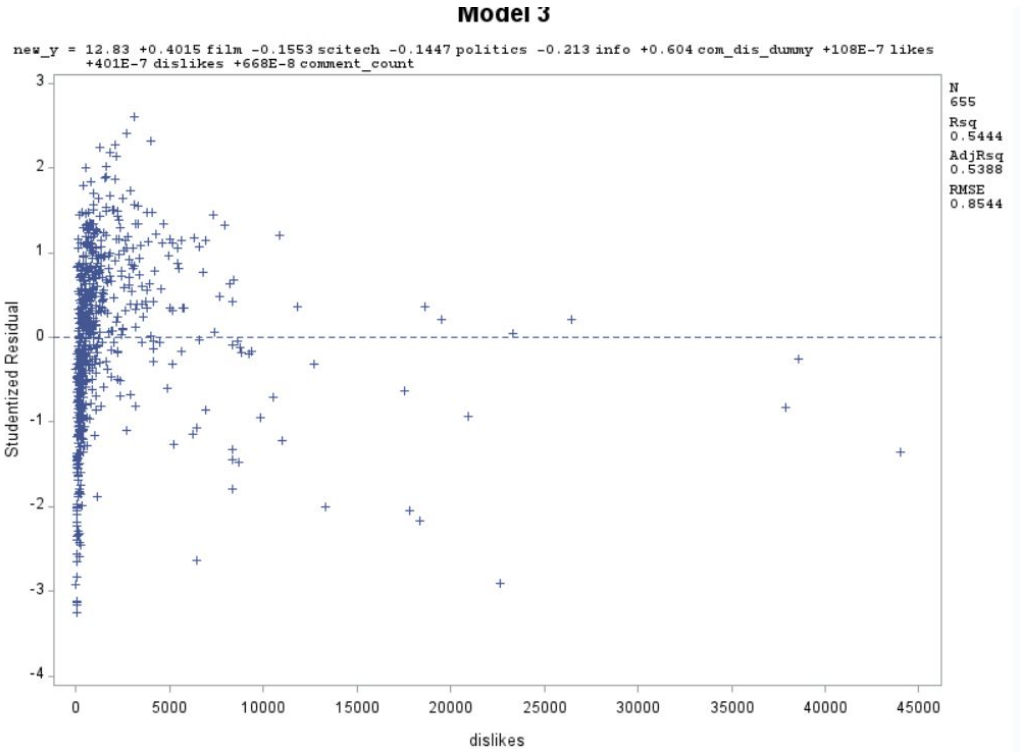

Model 3
new_y = 12.928 +0.479 film -0.2054 scitech -0.529 politics -0.2387 info +0.8957 com_dis_dummy +688E-8 likes +254E-7 dislikes +175E-7 comment_count
N
690
Rsq
0.4965
AdjRsq
0.4906
RMSE
0.9776
Normal Cumulative Distribution
CDF of Studentized Residual
0.0
0.2
0.4
0.6
0.8
1.0
0.1
0.3
0.5
0.7
0.9

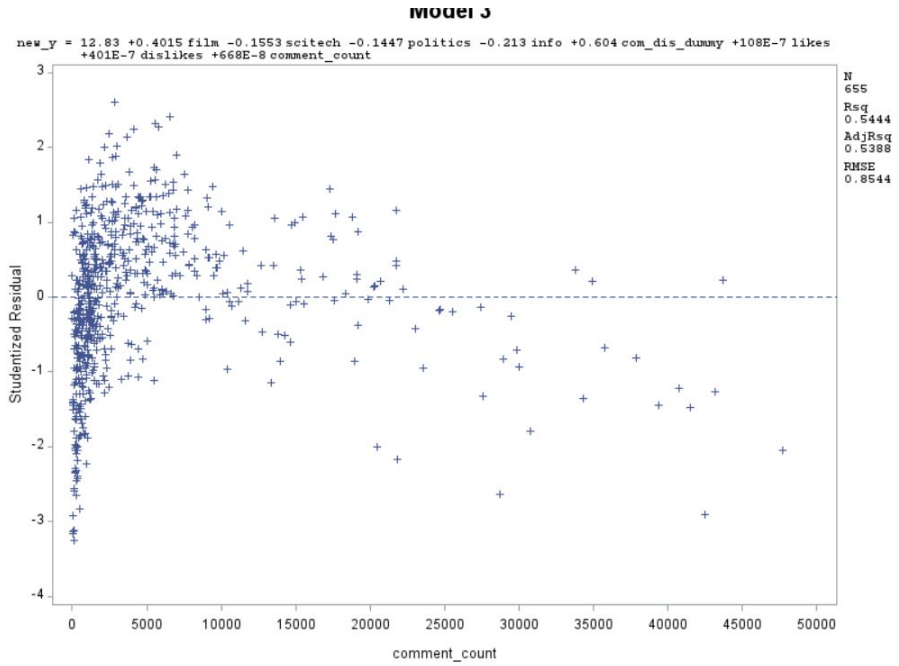
M3.C-

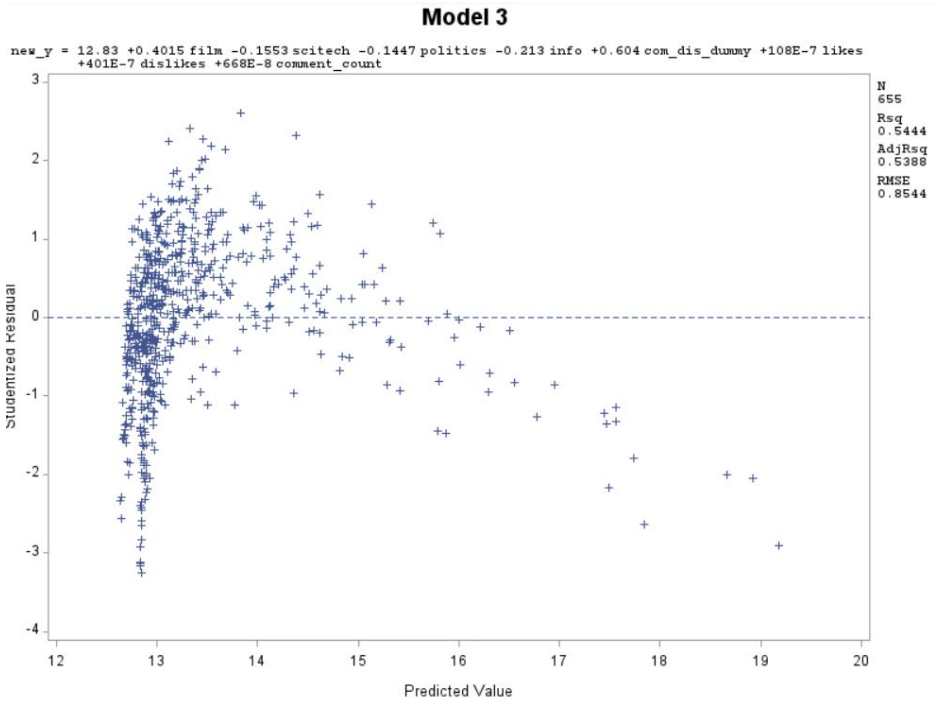

8 -1.955 0.011
9 -1.248 0.003
10 -0.884 0.002
11 -0.328 0.000
12 -0.103 0.000
13 0.565 0.000
14 0.137 0.000
15 . .
16 1.550 0.001
17 0.106 0.000
18 -0.509 0.000
19 . .
20 -1.038 0.000
21 -0.271 0.000
22 0.482 0.001
23 0.388 0.000
24 -2.106 0.005
25 -3.302 0.004
26 -1.485 0.001
27 -1.659 0.005
28 0.676 0.000
29 -2.380 0.006
30 1.273 0.000
31 0.602 0.000
32 0.290 0.001
33 . .
34 0.929 0.000
35 . .
36 . .
37 -1.221 0.000
38 . .
39 0.685 0.000
40 . .
41 0.209 0.000
42 . .
43 0.674 0.000
44 -3.035 0.003
45 -1.815 0.001
46 1.540 0.004
47 -0.239 0.000

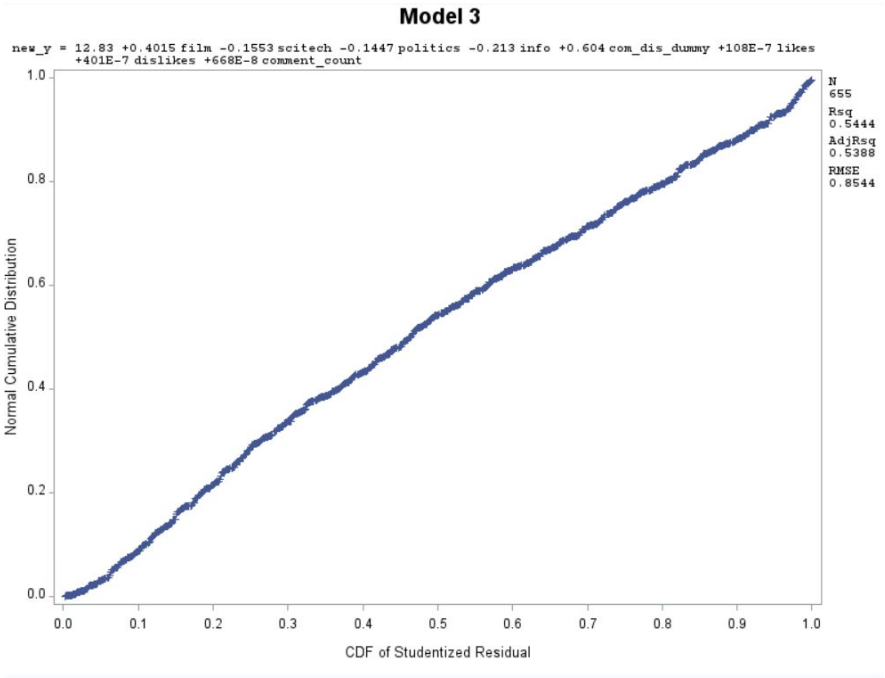
M3.D-


Model 3
new_y = 12.83 +0.4015 film -0.1553 scitech -0.1447 politics -0.213 info +0.604 com_dis_dummy +108E-7 likes +401E-7 dislikes +668E-8 comment_count
N
655
Rsq
0.5444
AdjRsq
0.5388
RMSE
0.8544
Studentized Residual
likes

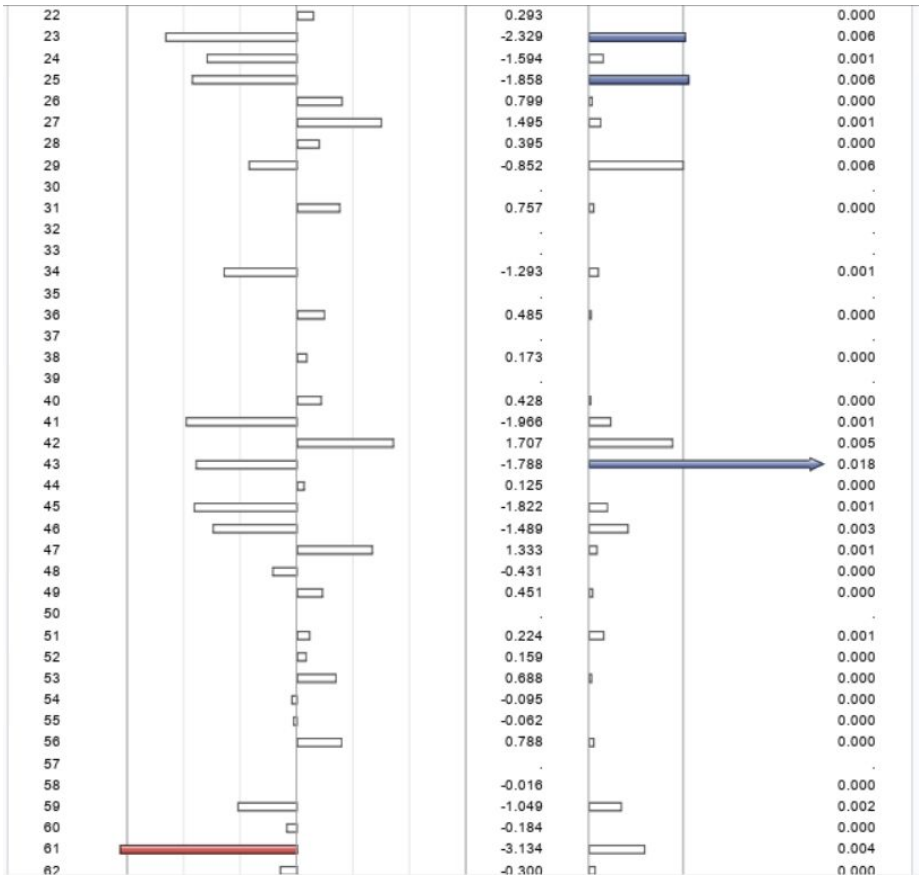

Model 3
new_y = 12.83 +0.4015 film -0.1553 scitech -0.1447 politics -0.213 info +0.604 com_dis_dummy +108E-7 likes +401E-7 dislikes +668E-8 comment_count
N
655
Rsq
0.5444
AdjRsq
0.5388
RMSE
0.8544
Studentized Residual
dislikes


Model 3
new_y = 12.83 +0.4015 film -0.1553 scitech -0.1447 politics -0.213 info +0.604 com_dis_dummy +108E-7 likes
+401E-7 dislikes +668E-8 comment_count
N
655
Rsq
0.5444
AdjRsq
0.5388
RMSE
0.8544
Studentized Residual
comment_count

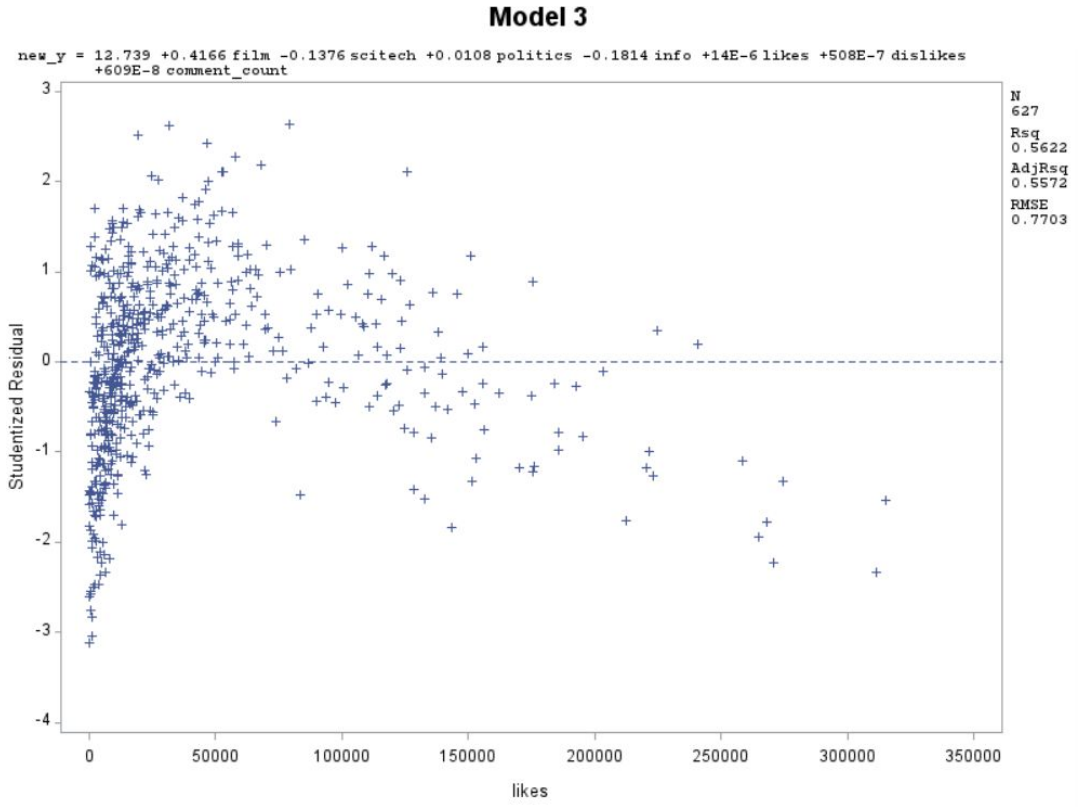

Model 3
new_y = 12.83 +0.4015 film -0.1553 scitech -0.1447 politics -0.213 info +0.604 com_dis_dummy +108E-7 likes
+401E-7 dislikes +668E-8 comment_count
N
655
Rsq
0.5444
AdjRsq
0.5388
RMSE
0.8544
Studentized Residual
Predicted Value

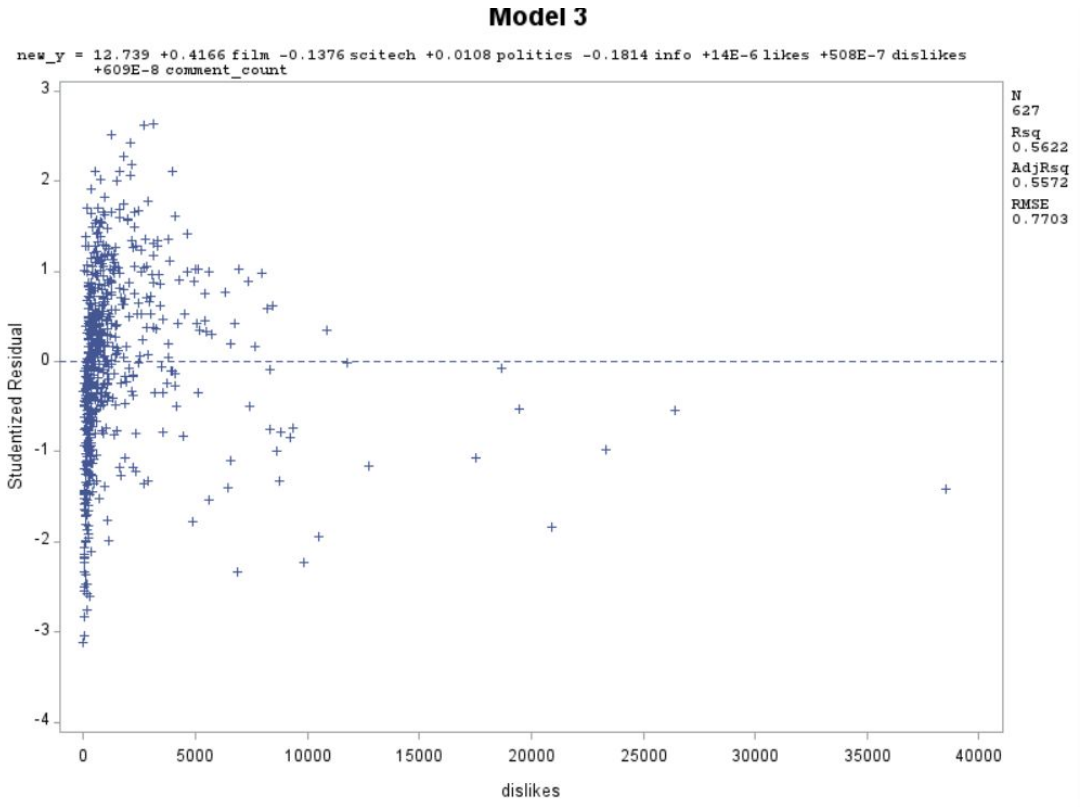

Model 3
new_y = 12.83 +0.4015 film -0.1553 scitech -0.1447 politics -0.213 info +0.604 com_dis_dummy +108E-7 likes
+401E-7 dislikes +668E-8 comment_count
N
655
Rsq
0.5444
AdjRsq
0.5388
RMSE
0.8544
Normal Cumulative Distribution
CDF of Studentized Residual
0.0
0.2
0.4
0.6
0.8
1.0
0.1
0.3
0.5
0.7
0.9

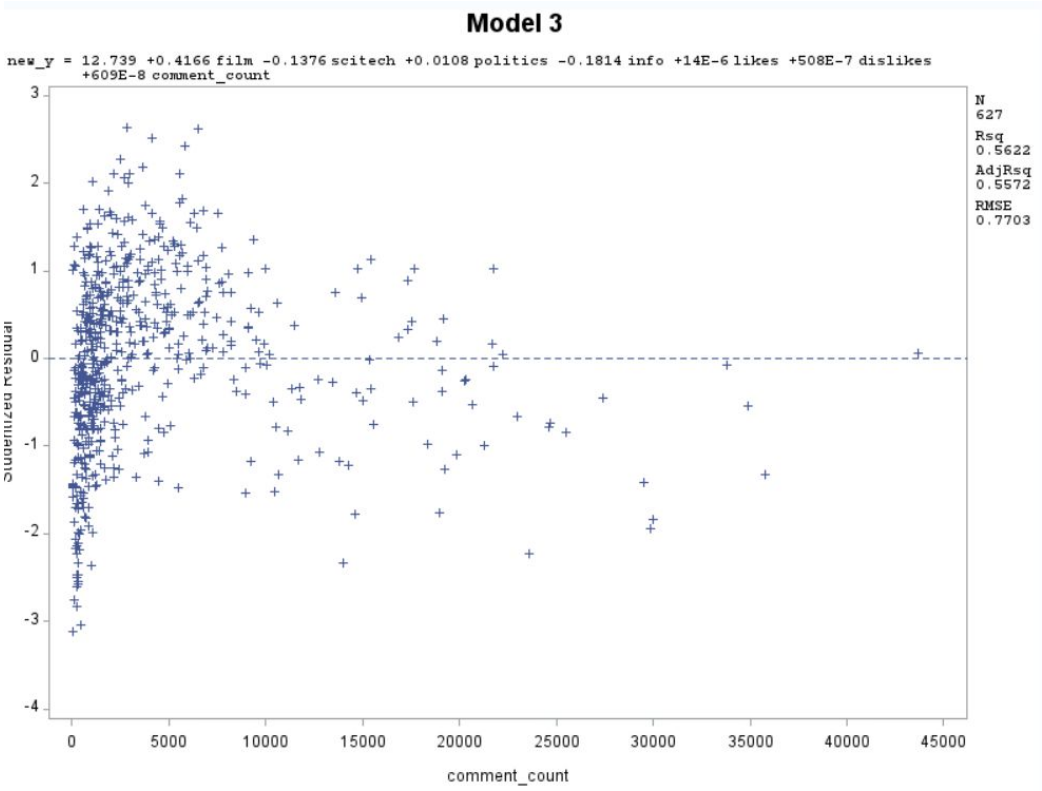
M3.E-

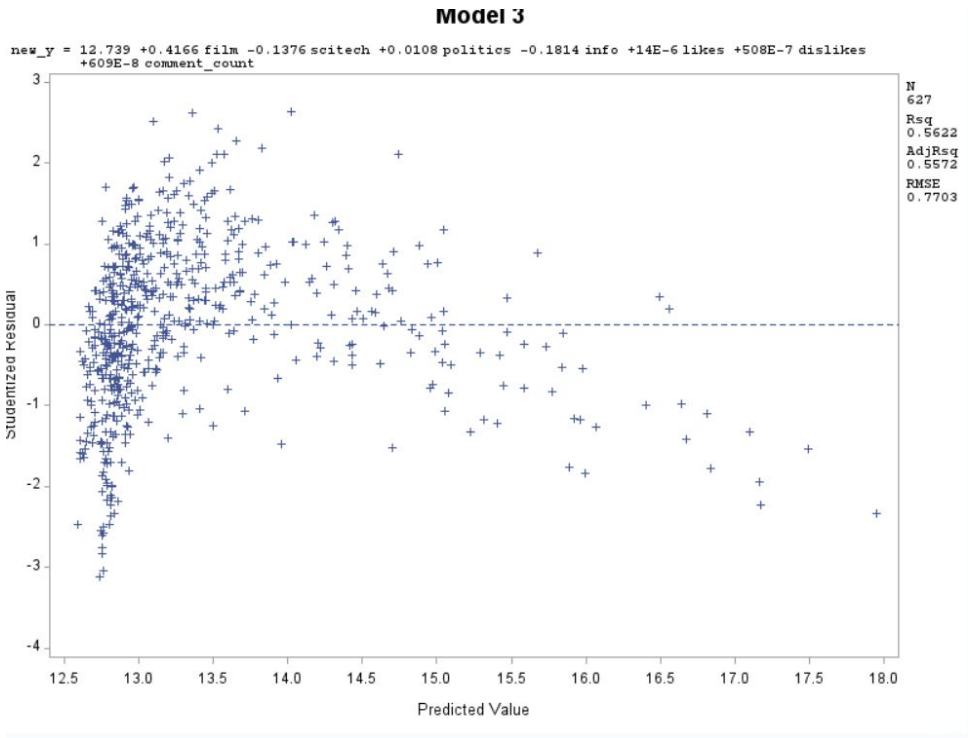

22 0.293 0.000
23 -2.329 0.006
24 -1.594 0.001
25 -1.858 0.006
26 0.799 0.000
27 1.495 0.001
28 0.395 0.000
29 -0.852 0.006
30 . .
31 0.757 0.000
32 . .
33 . .
34 -1.293 0.001
35 . .
36 0.485 0.000
37 . .
38 0.173 0.000
39 . .
40 0.428 0.000
41 -1.966 0.001
42 1.707 0.005
43 -1.788 0.018
44 0.125 0.000
45 -1.822 0.001
46 -1.489 0.003
47 1.333 0.001
48 -0.431 0.000
49 0.451 0.000
50 . .
51 0.224 0.001
52 0.159 0.000
53 0.688 0.000
54 -0.095 0.000
55 -0.062 0.000
56 0.788 0.000
57 . .
58 -0.016 0.000
59 -1.049 0.002
60 -0.184 0.000
61 -3.134 0.004
62 -0.300 0.000

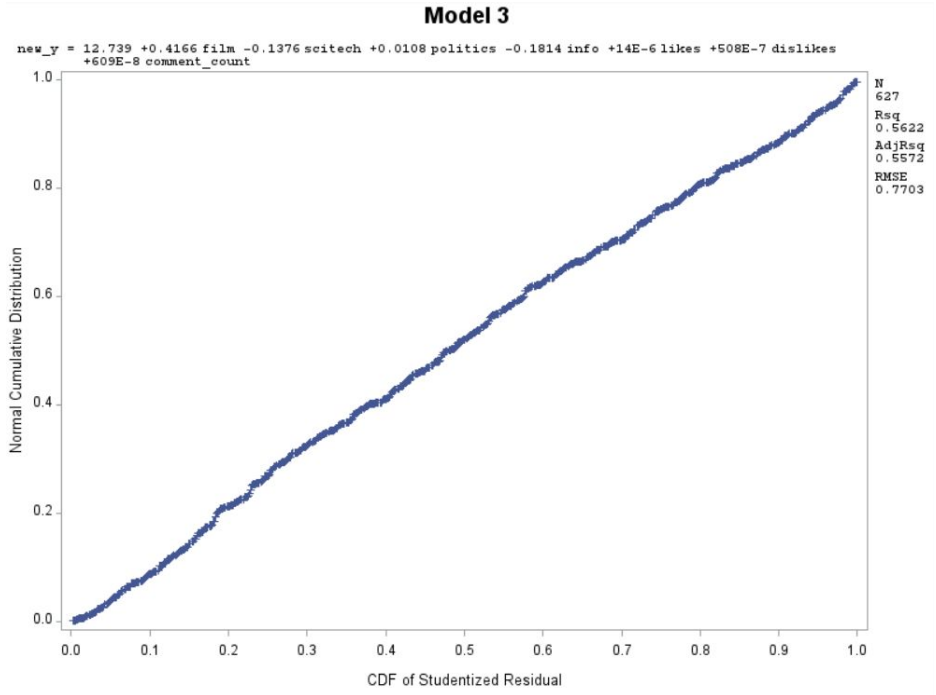
M3.F-


Model 3
new_y = 12.739 +0.4166 film -0.1376 scitech +0.0108 politics -0.1814 info +14E-6 likes +508E-7 dislikes +609E-8 comment_count
N
627
Rsq
0.5622
AdjRsq
0.5572
RMSE
0.7703
Studentized Residual
likes

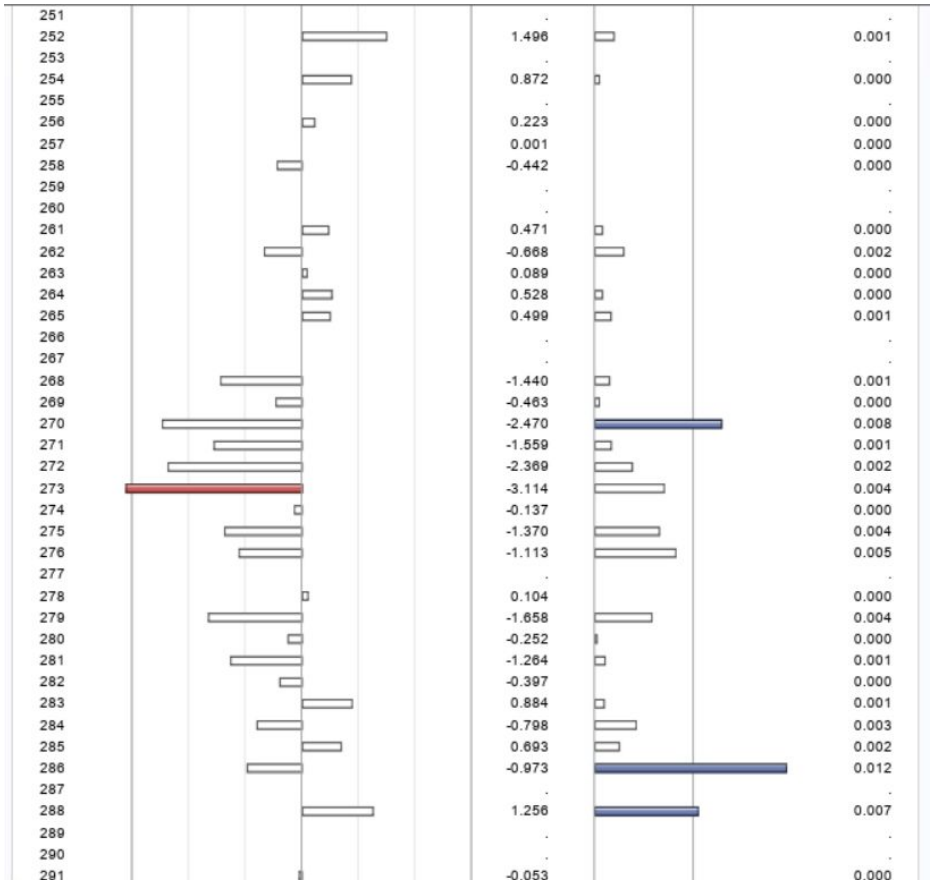

Model 3
new_y = 12.739 +0.4166 film -0.1376 scitech +0.0108 politics -0.1814 info +14E-6 likes +508E-7 dislikes +609E-8 comment_count
N
627
Rsq
0.5622
AdjRsq
0.5572
RMSE
0.7703
Studentized Residual
dislikes


Model 3
new_y = 12.739 +0.4166 film -0.1376 scitech +0.0108 politics -0.1814 info +14E-6 likes +508E-7 dislikes +609E-8 comment_count
N
627
Rsq
0.5622
AdjRsq
0.5572
RMSE
0.7703
Studentized Residual
comment_count
3
2
1
0
-1
-2
-3
-4
0
5000
10000
15000
20000
25000
30000
35000
40000
45000

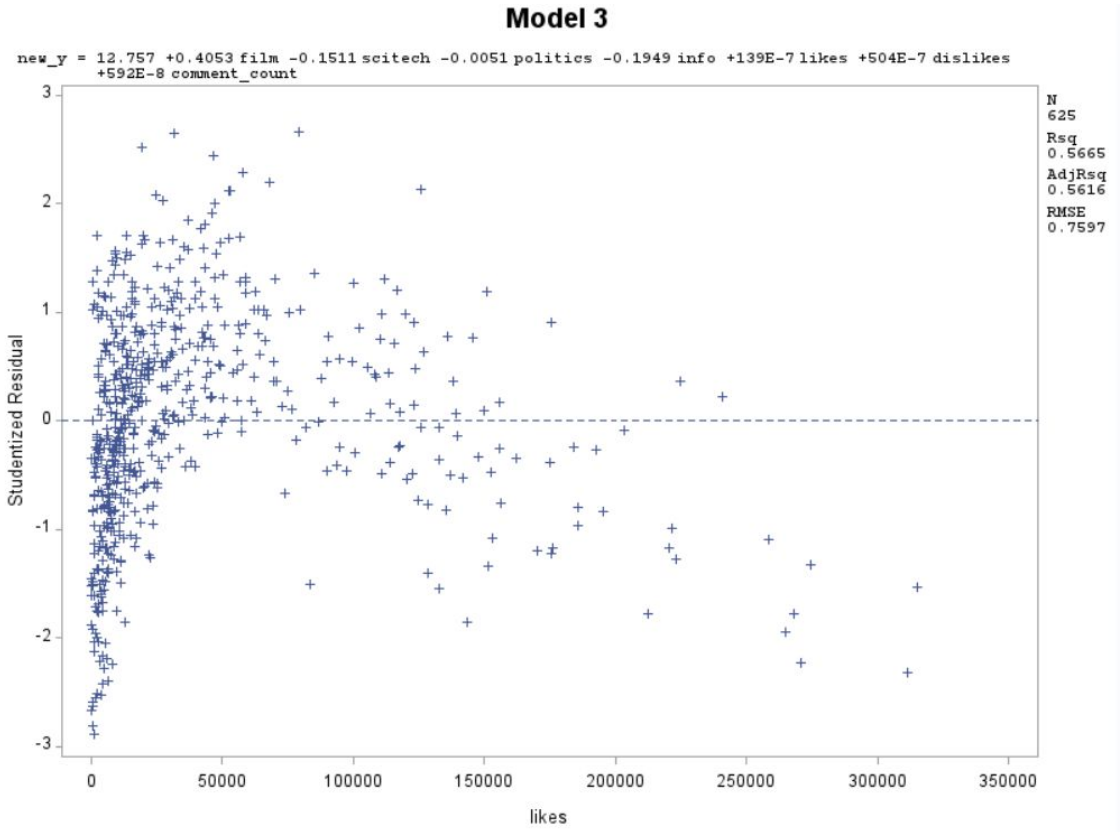

Model 3
new_y = 12.739 +0.4166 film -0.1376 scitech +0.0108 politics -0.1814 info +14E-6 likes +508E-7 dislikes
+609E-8 comment_count
N
627
Rsq
0.5622
AdjRsq
0.5572
RMSE
0.7703
Studentized Residual
Predicted Value

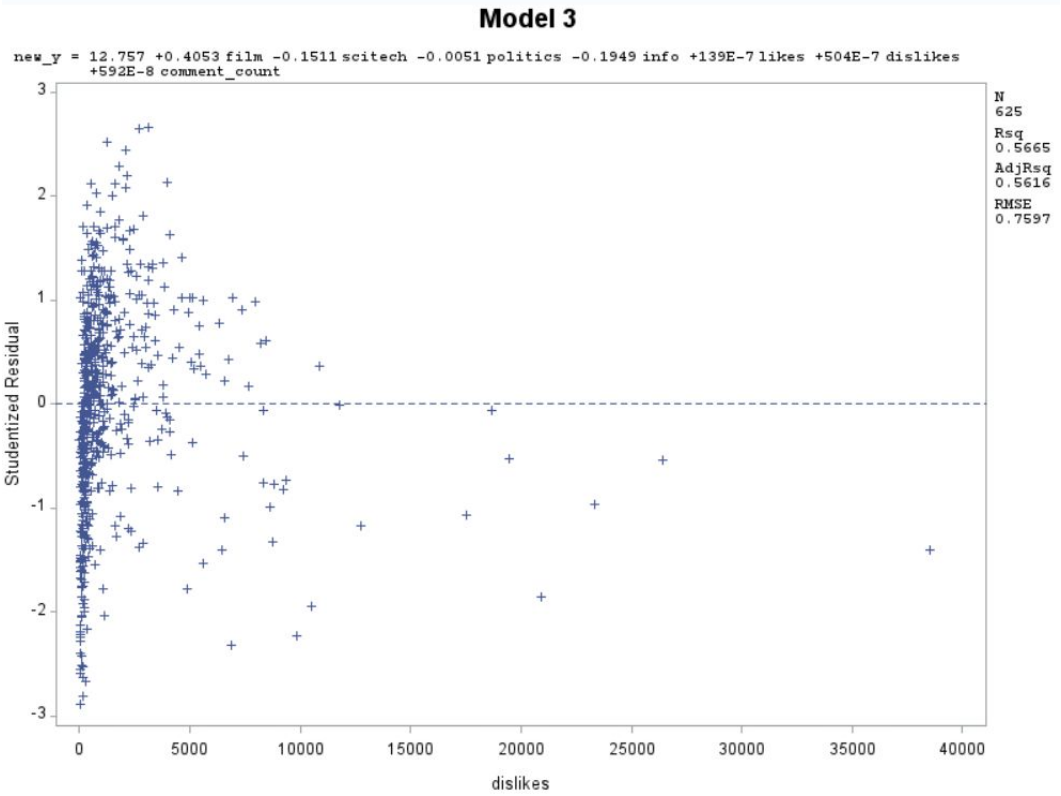

Model 3
new_y = 12.739 +0.4166 film -0.1376 scitech +0.0108 politics -0.1814 info +14E-6 likes +508E-7 dislikes
+609E-8 comment_count
N
627
Rsq
0.5622
AdjRsq
0.5572
RMSE
0.7703
Normal Cumulative Distribution
CDF of Studentized Residual

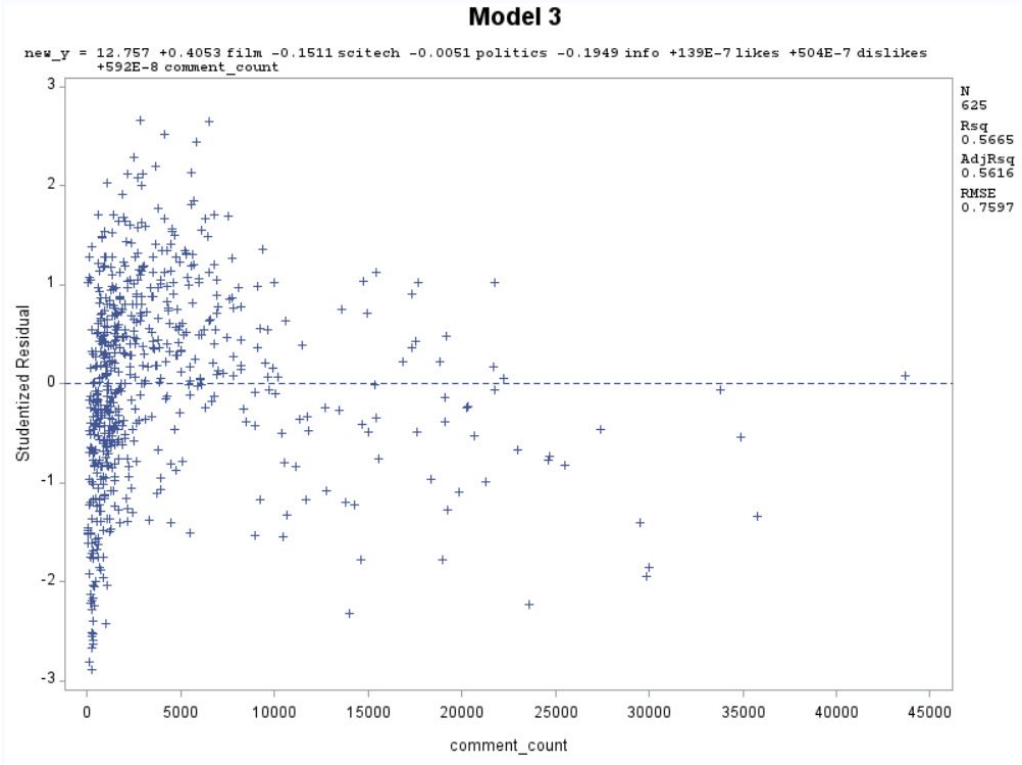
M3.G-

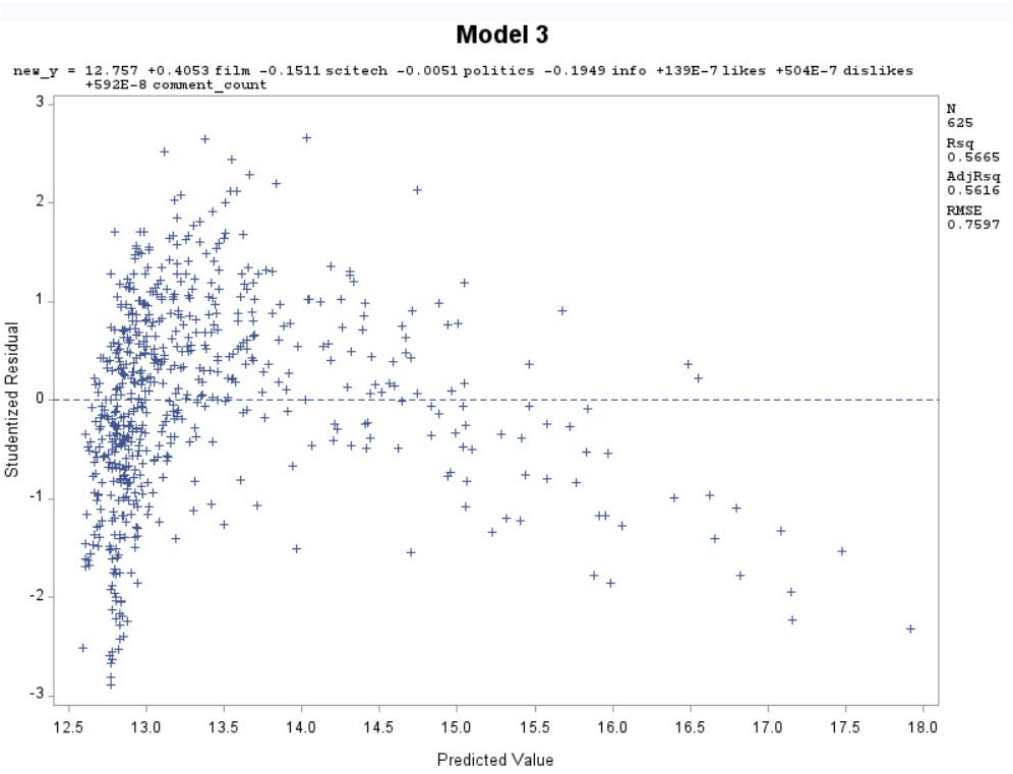

251 . .
252 1.496 0.001
253 . .
254 0.872 0.000
255 . .
256 0.223 0.000
257 0.001 0.000
258 -0.442 0.000
259 . .
260 . .
261 0.471 0.000
262 -0.668 0.002
263 0.089 0.000
264 0.528 0.000
265 0.499 0.001
266 . .
267 . .
268 -1.440 0.001
269 -0.463 0.000
270 -2.470 0.008
271 -1.559 0.001
272 -2.369 0.002
273 -3.114 0.004
274 -0.137 0.000
275 -1.370 0.004
276 -1.113 0.005
277 . .
278 0.104 0.000
279 -1.658 0.004
280 -0.252 0.000
281 -1.264 0.001
282 -0.397 0.000
283 0.884 0.001
284 -0.798 0.003
285 0.693 0.002
286 -0.973 0.012
287 . .
288 1.256 0.007
289 . .
290 . .
291 -0.053 0.000

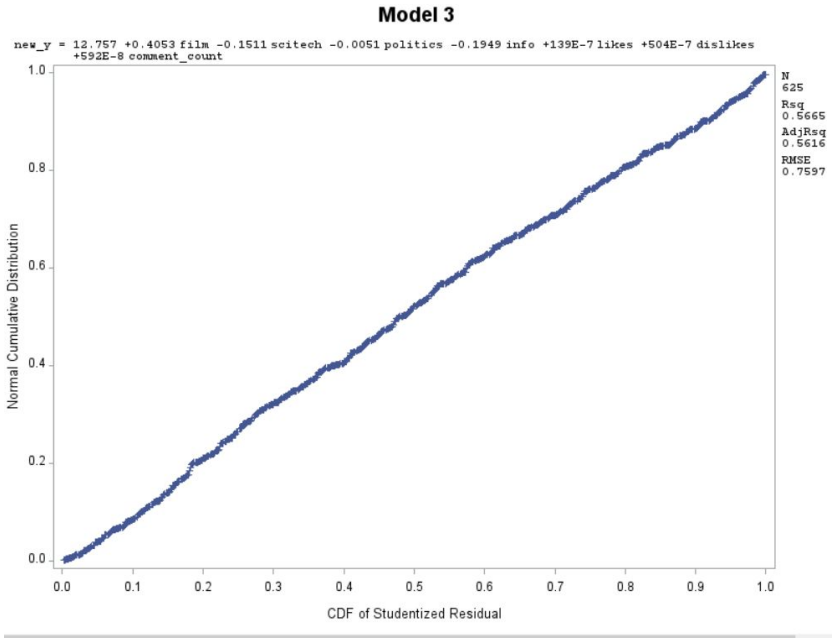
M3.H-


Model 3
new_y = 12.757 +0.4053 film -0.1511 scitech -0.0051 politics -0.1949 info +139E-7 likes +504E-7 dislikes
+592E-8 comment_count
N
625
Rsq
0.5665
AdjRsq
0.5616
RMSE
0.7597
Studentized Residual
likes
3
2
1
0
-1
-2
-3
0
50000
100000
150000
200000
250000
300000
350000


Model 3
new_y = 12.757 +0.4053 film -0.1511 scitech -0.0051 politics -0.1949 info +139E-7 likes +504E-7 dislikes
+592E-8 comment_count
N
625
Rsq
0.5665
AdjRsq
0.5616
RMSE
0.7597
Studentized Residual
dislikes
3
2
1
0
-1
-2
-3
0
5000
10000
15000
20000
25000
30000
35000
40000


Model 3
new_y = 12.757 +0.4053 film -0.1511 scitech -0.0051 politics -0.1949 info +139E-7 likes +504E-7 dislikes
+592E-8 comment_count
N
625
Rsq
0.5665
AdjRsq
0.5616
RMSE
0.7597
Studentized Residual
comment_count
3
2
1
0
-1
-2
-3
0
5000
10000
15000
20000
25000
30000
35000
40000
45000


Model 3
new_y = 12.757 +0.4053 film -0.1511 scitech -0.0051 politics -0.1949 info +139E-7 likes +504E-7 dislikes
+592E-8 comment_count
N
625
Rsq
0.5665
AdjRsq
0.5616
RMSE
0.7597
Studentized Residual
Predicted Value
3
2
1
0
-1
-2
-3
12.5
13.0
13.5
14.0
14.5
15.0
15.5
16.0
16.5
17.0
17.5
18.0



M3.I-

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	12.75657	0.04624	275.86	<.0001
film	1	0.40526	0.15371	2.64	0.0086
scitech	1	-0.15115	0.10666	-1.42	0.1570
politics	1	-0.00514	0.14377	-0.04	0.9715
info	1	-0.19492	0.08562	-2.28	0.0232
com_dis_dummy	0	0	.	.	.
likes	1	0.00001391	8.703841E-7	15.98	<.0001
dislikes	1	0.00005045	0.00001303	3.87	0.0001
comment_count	1	0.00000592	0.00000860	0.69	0.4914

M3-J

Number of Observations Read	869
Number of Observations Used	625
Number of Observations with Missing Values	244

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	463.90727	115.97682	201.17	<.0001
Error	620	357.43439	0.57651		
Corrected Total	624	821.34166			

Root MSE	0.75928	R-Square	0.5648
Dependent Mean	13.36302	Adj R-Sq	0.5620
Coeff Var	5.68195		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate
Intercept	1	12.73909	0.04155	306.58	<.0001	0
film	1	0.42405	0.15283	2.77	0.0057	0.07386
info	1	-0.17170	0.08354	-2.06	0.0403	-0.05491
likes	1	0.00001429	6.855384E-7	20.85	<.0001	0.64911
dislikes	1	0.00005554	0.00001155	4.81	<.0001	0.14959

M3.K-

Validation stats for Model 3

Obs	_TYPE_	_FREQ_	rmse	mae
1	0	244	1.55573	1.02840

M3.L-

Validation stats for Model 3

The CORR Procedure

2 Variables:	ln_views yhat

Simple Statistics							
Variable	N	Mean	Std Dev	Sum	Minimum	Maximum	Label
ln_views	244	13.25641	1.59081	3235	8.56102	17.41341	
yhat	244	13.60802	1.77806	3320	12.57536	28.85160	Predicted Value of new_y

Pearson Correlation Coefficients, N = 244 Prob > |r| under H0: Rho=0		
	ln_views	yhat
ln_views	1.00000	0.57837 <.0001
yhat Predicted Value of new_y	0.57837 <.0001	1.00000

P.A

Output Statistics								
Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	95% CL Mean		95% CL Predict		Residual
1	.	12.6380	0.0774	12.4860	12.7900	11.1392	14.1368	.

P.B

Output Statistics								
Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	95% CL Mean		95% CL Predict		Residual
1	.	13.2250	0.1513	12.9280	13.5221	11.7047	14.7454	.