Youtube Trending Video Analysis

Math 261A Regression Theory

December 2018

Jonathan Schwartz
Jung-a Kim
Mengqi Yin
Sina Sadeh

Introduction

Youtube was invented by Jawed Karim, Steve Chen, and Chad Hurley in 2005 as one platform to share videos around the world. After HD videos became available, it has been used for different purposes such as video blogs, national events, or learning tools. It maintains a list of the top trending videos on the platform. The dataset is a daily record of the top trending Youtube videos between November 2017 and March 2018 collected by Mitchell J. using the Youtube API.

Our team was curious which factors contribute to the popularity of the trending videos. Specifically, do the factors that the author can control such as the published time, the number of tags, or the length of the title affect the number of views on a trending video? What are the attributes that contribute to the number of views?

Variables of Interest

views	The number of times video was viewed by Youtube users(the response variable)
publish_hour	The time(in 24-hour clock) when the video was uploaded or published on Youtube site
likes	The number of likes by users
dislikes	The number of dislikes by users
comments	The number of comments made by users or the publisher
tag_appeared_in_title_count	The number of times that a Youtube video title contains one if its own tags
tag_appeared_in_title**	TRUE if a Youtube video title contains one of its tags. This attribute is derived from “tag_appeared_in_title_count” attribute.
trend_day_count	The number of days the video appeared on the Youtube trending video list
trend_publish_diff	The difference in days between published date and the first trending date
trend_tag_highest	Maximum number of times all trending videos used one of the video’s tags
trend_tag_total	Total number of times all trending videos used any of the video’s tags
tags_count	The number of tags attached to the video
subscriber	The number of Youtube users subscribed to the video channel to get notifications about any channel updates
like_ratio	The ratio of likes to all ratings: likes/(likes+dislikes)
total_engagement	Total number of times a viewer engaged in some way: likes+dislikes+comments
engagement_ratio	The percentage of viewers who engaged in some way: total.engagement/views
title_length	The number of characters in the video title

description_length	The number of characters in the video description
weekday*	The day of the week the video was published on
caps**	TRUE if the number of uppercase letters is greater than the number of lower case letter in the title.
exclamation**	TRUE if the title contains at least one "!".
question**	TRUE if the title contains at least one "?".

Total: 22 variables, 4525 observations

* 1 categorical variable: weekday (6 dummy variables from ‘Monday’ to ‘Saturday’; ‘Sunday’ is the baseline.)

** 4 binary variables: tag_appeared_in_title, caps, exclamation, question

Testing the significance of the factors controlled by the author

Our first question was whether the factors controlled by the author affect the number of views on his/her videos. The controllable factors that we defined were the published time, publish weekday, the number of tags, if any tags are included in the title, the number of tags included in the title, the length of the title, the length of the description of the video, inclusion of capital letters, exclamation marks, and question marks.

Using the partial F-test for testing the group of factors, the null hypothesis was that none of those 10 factors had any influence on the number of views of the video given that all the other 11 uncontrolled factors are included in the model. *total_engagement* was not included in the model since it had high multicollinearity with *likes,dislikes,comments*. The F-test statistic for this hypothesis testing was 3.433299 with p-value = 0.0001686354 which is much less than the significance level of 0.01. Thus, we concluded that at least one of those 8 factors had significant influence on the number of views given that all the other 11 variables were included in the regression model. This led us to narrow down the number of significant factors and investigate whether the controlled factors remain in the final model.

Model Building

The response variable of our model was *views*. Including all 21 variables in the first model, we examined the residual plots to identify the non-constant variance and non-linear relationship with *views*.

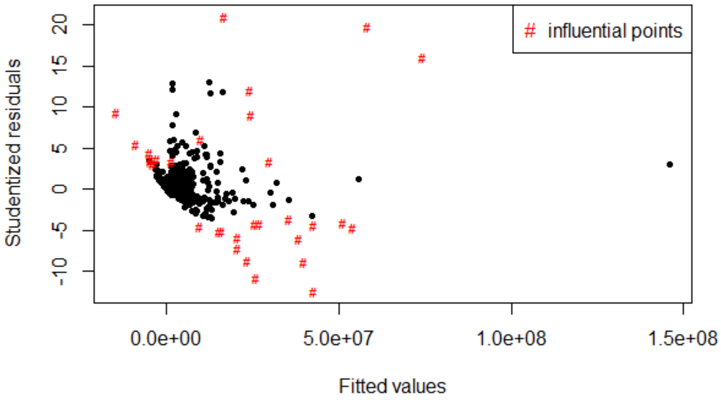


Fig 1. Student residuals of the untransformed dataset

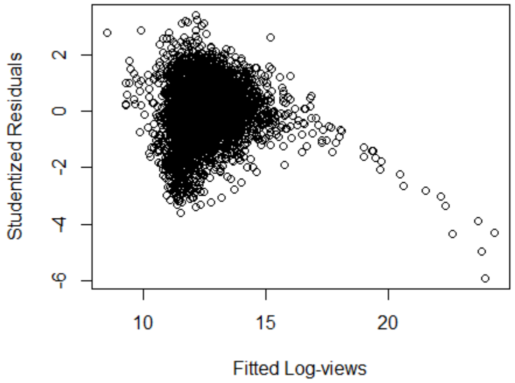


Fig 2. Studentized residual plot after log-transformation on response

The studentized residual plot showed non-constant residual variance with a left-pointing triangular pattern (Fig.1). Using Box-Cox method, we found the optimal power transformation on the response. The optimal power that maximizes the log-likelihood of our data was close to ‘0’. Thus we applied log-transformation on the response variable *views* which stabilized the residual variance except for a few outliers(Fig.2).

We identified 31 influential points[1] from the dataset. We selected the observations whose Cook’s distance was greater than 1 or COVRATIO less than the size-adjusted threshold = 1 - 3*(k+1)/n with h_{ii} greater than 2*(k+1)/n and standardized residual greater than 3. COVRATIO was chosen as the criteria since we wanted to test whether the outliers increase the overall variance of the parameter estimates, which would decrease the accuracy of the relationship between the variables and the response. The influential points were recorded at this step and later used for testing their influence on variable selection.

With the log-transformed response, we examined the residual plots, qq-plots, and scatterplots of the response against the 17 numerical variables. We transformed 10 numerical variables to linearize the functional relationship or stabilize the residual variance. The 4 variables *likes*, *dislikes*, *comments*, and *total_engagement* showed a logarithmic relationship with the response. *subscribers* showed a square root relationship with the response. The 5 variables *trend_tag_total*, *tag_appeared_in_title_count*, *description_length*, *engagement_ratio*, and *like_ratio* showed non-constant residual variance against themselves with the right-pointing triangular pattern. Thus, we applied different types of transformations on each variable to stabilize the residual variance. Each variable was transformed to *log(trend_tag_total+1), exp(tag_appeared_in_title_count), sqrt(description_length), sqrt(engagement_ratio),* and *cube(like_ratio).* ‘1’ was added inside the logarithmic function due to zero values in some of the observations.

The full model fitted with 21 partly transformed predictors had no interaction terms based on the assumption that the effect of the 17 numerical predictors on the response was independent of

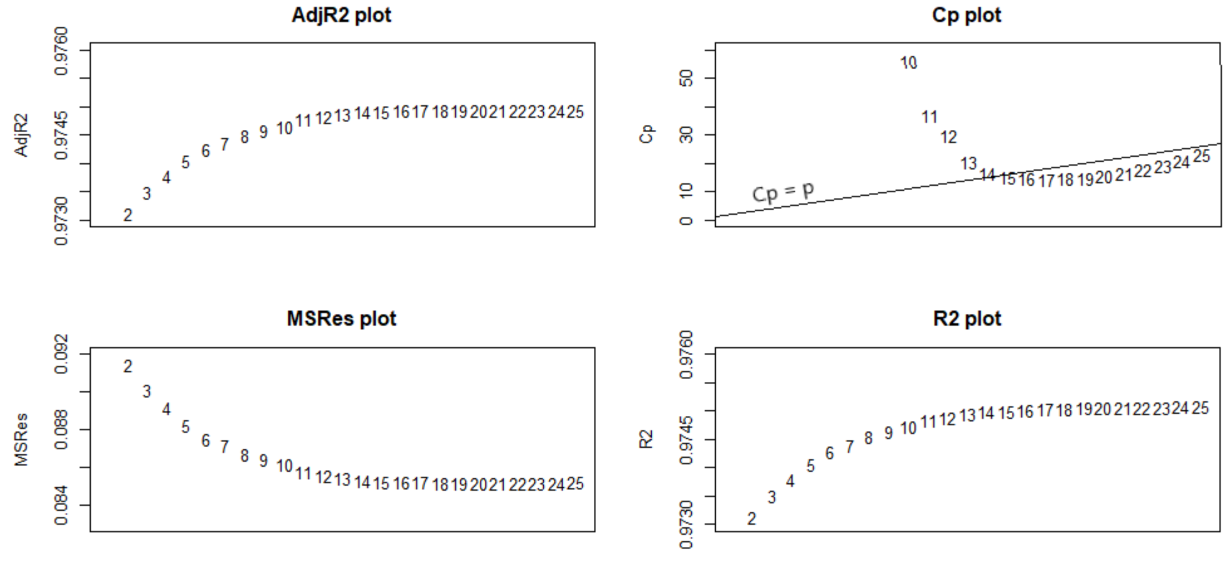


Fig 3. Adjusted Rsquared, Cp, MSRes, Rsquared plot

[1] Influential points by row number: 71, 150, 328, 332, 850, 949, 957, 1026, 1275, 1353, 1355, 1429, 1537, 1741, 1746, 1803, 1849, 2001, 2027, 2086, 2193, 2323, 2351, 2484, 2553, 3578, 3652, 4058, 4062, 4366, 4462

weekday. In order to find the optimal number of predictors, we used R function ‘regsubsets’ to find the best models for each different number of predictors. Using each model’s Adjusted Rsquared, Mallow’s Cp, Mean Squared Residuals, and Rsquared, we plotted them against the number of predictors to find the simplest model with the maximum Rsquared, minimum MSRes, and low Cp.

Fig 3 shows the plot with the models labeled as its number of predictors. Based on the four criteria, the optimal numbers of predictors are 12, 13, and 14. Thus, we selected the three best models with 12, 13, and 14 predictors based on the four criteria. Then we implemented Forward variable selection using 0.05 as the p-value threshold to add a new predictor. Finally, we implemented Backward variable selection using the threshold = 0.1 to drop each predictor starting with the full model.

Predictors	Method
subscribers, likes, dislikes, like_ratio, comments, total_engagement, engagement_ratio, title_length, trend_pub_diff, caps, trend_tag_highest, trend_tag_total (12 predictors)	Exhaustive Search
subscribers, likes, dislikes, like_ratio, comments, total_engagement, engagement_ratio, title_length, trend_pub_diff, caps, trend_tag_highest, trend_tag_total, description_length (13 predictors)	Exhaustive Search
subscribers, likes, dislikes, like_ratio, comments, total_engagement, engagement_ratio, title_length, trend_pub_diff, caps, trend_tag_highest, trend_tag_total, description_length, weekday (14 predictors)	Exhaustive Search
subscribers, likes, dislikes, like_ratio, comments, total_engagement, engagement_ratio, title_length, trend_pub_diff, caps, trend_tag_highest, trend_tag_total, description_length (13 predictors)	Forward Selection
subscribers, likes, dislikes, like_ratio, comments, total_engagement, engagement_ratio, title_length, trend_pub_diff, caps, trend_tag_highest, trend_tag_total, description_length, publish_hour (14 predictors)	Backward Selection

Table 1. Five candidate models based on exhaustive search, Forward variable selection, and Backward variable selection

In Table 1, the second model from the exhaustive search and the Forward selection model are the same. The predictor *total_engagement* in all these four models had extremely high Variation Inflation factor at 893, since it was the sum of *comments*, *likes*, and *dislikes*. However, removing *total_engagement* inflated Mallows’s Cp approximately from 32 to 163 for all the models. It was not surprising since *total_engagement* had the highest correlation with the response variable *views* with r = 0.9064, thus removing *total_engagement* from the model would lower the accuracy of the prediction for *views.* Therefore, we discarded the next highest variation inflator *likes* with VIF = 706.230706 which increased Cp to only 58. After removing *likes*, the VIF of *total_engagement* was still high, but again, to prevent the extreme rise in Cp, we removed *comments* with VIF = 11. This indicated 91% of *comments* could be explained by regressing on the other predictors in the model. Then we removed *dislikes* which had the second-highest VIF around 9.

Predictors	Adj Rsquared	PRESS	MSRes	Cp
“subscribers”, “total_engagement”, “like_ratio”, “engagement_ratio”, “title_length”, “trend_pub_diff”, “caps”, “trend_tag_highest”, “trend_tag_total” (9 predictors)	0.9739	404.8821	0.0896	236.491
"subscribers", "total_engagement", "like_ratio", "engagement_ratio", "title_length", "trend_pub_diff", "caps", "trend_tag_highest", "trend_tag_total", "description_length" (10 predictors)	0.974	404.2452	0.0894	229.0986

Table 2. The models after multicollinearity treatment and test of significance

After the variation inflator check, we deleted the variables with high p-values greater than 0.1 in the the models. This process resulted in making the second,third, fourth, and the last model the same. Table 2 shows the two models treated with multicollinearity and contains only significant predictors. The residual plots for both models looked very similar. The VIF's of all the predictors in each model were under 2.4. The second model had the highest adjusted Rsquared, the lowest PRESS, the lowest MSRes, and the lowest Cp. Thus, we chose the second model as the final model. As the last step of building the model, we re-selected variables with the same process after excluding the 31 influential points. Treating the multicollinearity and excluding the insignificant factors resulted in the same models as those with the influential points.

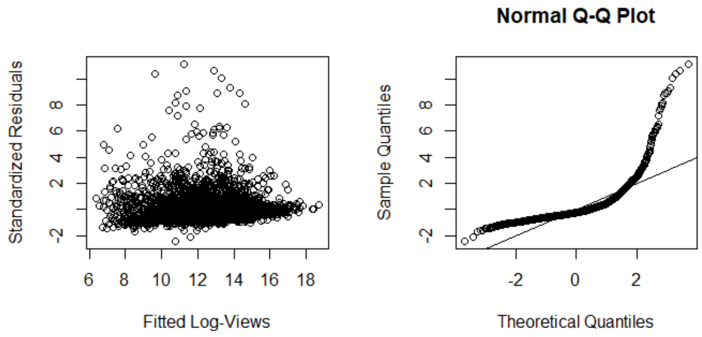


Fig 4. Residual plots for the final model

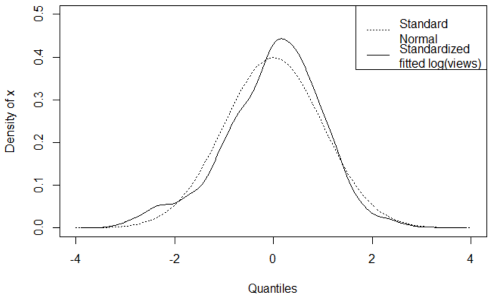
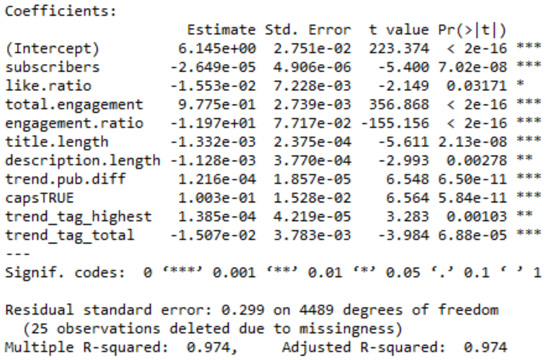


Fig 5. Density of Standard Normal Distribution and Standardized fitted log(views) distribution

Fig 4 shows the residual plot and Q-Q plot of our final model. Except for a few outliers, residuals have a constant variance along the fitted values. There were 77 points whose absolute residual was greater than 3. The Q-Q plot of the standardized residuals showed the u-shaped pattern which indicated the residuals were positively skewed. Thus, we implemented Box-Cox method to find the power transformation parameter for the response variable, but the maximum likelihood was at $\lambda = 0.95$ which indicates no response variable transformation is necessary. The transformation on predictors was already done using the scatter plots and residual plots. We tried double-log transformation on the response and the Q-Q plot looked slightly better with s-shape which means that the residuals follow a t-distribution. However, this transformation raised p-values for some of the predictors in this model and excluding them increased MSRes. Thus, we researched the non-normality of residuals and found an article that for large sample sizes, sample statistics such as estimated coefficients approximately follow a normal distribution due to Central Limit Theorem(J. Bartlett 2013). Thus we could assume that the fitted values which are linear combinations of the coefficients approximately follow a normal distribution. In Fig 5, we can verify that the fitted values are approximately normally distributed. This meant the true response and the true coefficients could be estimated with the t-based confidence intervals without the additional transformation.



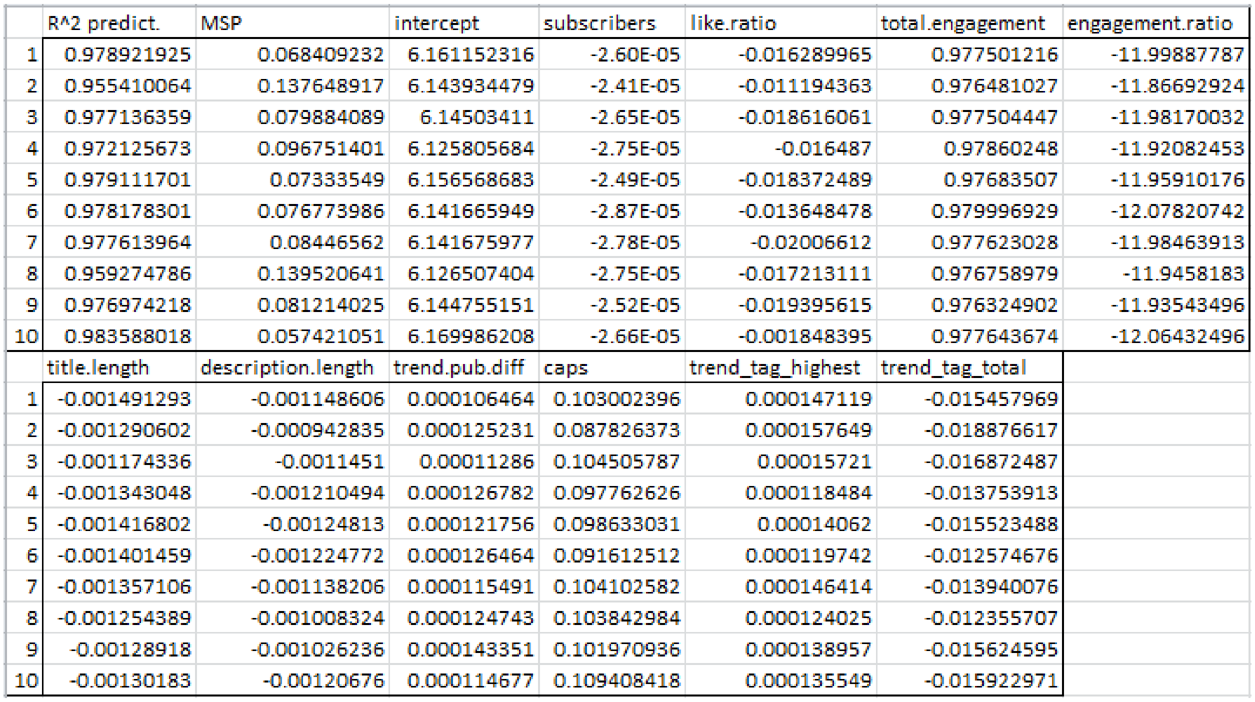
```
Coefficients:
                    Estimate Std. Error  t value Pr(>|t|)    
(Intercept)        6.145e+00  2.751e-02  223.374  < 2e-16 ***
subscribers       -2.649e-05  4.906e-06   -5.400 7.02e-08 ***
like.ratio        -1.553e-02  7.228e-03   -2.149  0.03171 *  
total.engagement   9.775e-01  2.739e-03  356.868  < 2e-16 ***
engagement.ratio  -1.197e+01  7.717e-02 -155.156  < 2e-16 ***
title.length      -1.332e-03  2.375e-04   -5.611 2.13e-08 ***
description.length -1.128e-03  3.770e-04   -2.993  0.00278 ** 
trend.pub.diff     1.216e-04  1.857e-05    6.548 6.50e-11 ***
capsTRUE           1.003e-01  1.528e-02    6.564 5.84e-11 ***
trend_tag_highest  1.385e-04  4.219e-05    3.283  0.00103 ** 
trend_tag_total   -1.507e-02  3.783e-03   -3.984 6.88e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.299 on 4489 degrees of freedom
  (25 observations deleted due to missingness)
Multiple R-squared:  0.974,     Adjusted R-squared:  0.974
```

Fig 6. Coefficients table of the final model

Model Validation

After building the model, we validated the model's predictability and checked the stability of the estimated coefficients in order to describe the true relationship between the regressors and the response. Since collecting "fresh" data was not possible, we created a validation set of data by splitting our data into two parts, estimation data and validation data. In order to avoid the potential problems that may occur for random split of the data, we implemented k-fold cross validation algorithm to maximize accuracy and to validate the model.

In k-fold cross validation, we randomly partition our data set into k equal-sized subsets called folds. Afterwards, each one of the subsets will be retained as validation data once and the remaining k-1 subsets will be used as our estimation data. Therefore, this process will repeat a total of k times. In our case, we used k=10. we divided our dataset into 10 equal-sized subsets with each subset having approximately 452 observations.



	R^2 predict.	MSP	intercept	subscribers	like.ratio	total.engagement	engagement.ratio
1	0.978921925	0.068409232	6.161152316	-2.60E-05	-0.016289965	0.977501216	-11.99887787
2	0.955410064	0.137648917	6.143934479	-2.41E-05	-0.011194363	0.976481027	-11.86692924
3	0.977136359	0.079884089	6.14503411	-2.65E-05	-0.018616061	0.977504447	-11.98170032
4	0.972125673	0.096751401	6.125805684	-2.75E-05	-0.016487	0.97860248	-11.92082453
5	0.979111701	0.07333549	6.156568683	-2.49E-05	-0.018372489	0.97683507	-11.95910176
6	0.978178301	0.076773986	6.141665949	-2.87E-05	-0.013648478	0.979996929	-12.07820742
7	0.977613964	0.08446562	6.141675977	-2.78E-05	-0.02006612	0.977623028	-11.98463913
8	0.959274786	0.139520641	6.126507404	-2.75E-05	-0.017213111	0.976758979	-11.9458183
9	0.976974218	0.081214025	6.144755151	-2.52E-05	-0.019395615	0.976324902	-11.93543496
10	0.983588018	0.057421051	6.169986208	-2.66E-05	-0.001848395	0.977643674	-12.06432496

	title.length	description.length	trend.pub.diff	caps	trend_tag_highest	trend_tag_total
1	-0.001491293	-0.001148606	0.000106464	0.103002396	0.000147119	-0.015457969
2	-0.001290602	-0.000942835	0.000125231	0.087826373	0.000157649	-0.018876617
3	-0.001174336	-0.0011451	0.00011286	0.104505787	0.00015721	-0.016872487
4	-0.001343048	-0.001210494	0.000126782	0.097762626	0.000118484	-0.013753913
5	-0.001416802	-0.00124813	0.000121756	0.098633031	0.00014062	-0.015523488
6	-0.001401459	-0.001224772	0.000126464	0.091612512	0.000119742	-0.012574676
7	-0.001357106	-0.001138206	0.000115491	0.104102582	0.000146414	-0.013940076
8	-0.001254389	-0.001008324	0.000124743	0.103842984	0.000124025	-0.012355707
9	-0.00128918	-0.001026236	0.000143351	0.101970936	0.000138957	-0.015624595
10	-0.00130183	-0.00120676	0.000114677	0.109408418	0.000135549	-0.015922971

Fig 7. Cross validation table with Rsquared prediction, MSP, and regressor coefficients

Regressor	intercept	subscribers	like_ratio	total_engagement	engagement_ratio	title_length	description_length	trend_publish_diff	caps	trend_tag_highest	trend_tag_total
Our model	6.145475	-0.00002648816	-0.01553134	0.9775194	-11.97265	-0.001332407	-0.001128420	0.0001216057	0.1002788	0.0001385073	-0.01507096
Mean of estimates	6.145708596	-0.0000265	-0.01531316	0.977527175	-11.97358585	-0.001332005	-0.001129946	0.000121782	0.100266765	0.000138577	-0.01509025
Difference ratio	0.000038	0.000447	0.0142	0.000008	0.00007816	0.000302	0.00135	0.0014	0.00012	0.000503	0.00128

Table 3. Comparison between the original coefficients and mean estimates

Fig 7 shows the 10 training sets' Rsquared for predictions, Mean Squared Prediction errors, and the estimated parameters from the model. The estimated coefficients seemed stable. The mean of the 10 Rsquared predictions was 0.9738 which was the same as the Rsquared for prediction of our model. Thus, we could say the predictive power of our model does not depend on which subset models are selected. The mean of the 10 MSP's is 0.0895 which was almost the same as MSRes of our model 0.0894. Table 3 shows the difference ratio of our model to the mean coefficient estimates from the 10 training datasets. The maximum difference ratio of all the estimated parameters was 1.4% for *like_ratio*.

Conclusion

Our final model is

log(views) = 6.145475
-0.00002648816*sqrt(subscribers) -0.01553134*(like_ratio^3)
+0.9775194*log(total_engagement + 1) -11.97265*sqrt(engagement_ratio)
-0.001332407*(title_length) -0.001128420*sqrt(description_length)
+0.0001216057*(trend_publish_diff) +0.1002788*(caps)
+0.0001385073*(trend_tag_highest) -0.01507096*log(trend_tag_total + 1)

From the model, we could verify that the factors that are controlled by the author; title length, description length, and including capital letters more than lowercase letters remained significant in the final model.

subscribers	like_ratio	total_engagement	engagement_ratio	title_length	description_length	trend_publish_diff	including more capital letters than lowercase letters in the title	trend_tag_highest	trend_tag_total
20,563,106	0.9845	176,384	0.0626	24	636	7	TRUE	488	1,007

Table 4. The data of the observation 5

Table 4 shows the values of the observation 5 in the dataset. The 95% prediction interval for the average number of views for the data of the video 1 is between 2,681,744 and 2,913,608 with the predicted average views as 2,795,273. The actual number of views in observation 5 is 2,819,118 which is in the prediction interval of the true mean number of views.

If the author reduced the title length of the video by 10 characters given all the other variables fixed, the difference in the log(views) would be -0.001332407*-10 = 0.01332407 which means that the average views would increase by exp(0.01332407)-1 = 0.01341323 = about 1.3%.

According to our model coefficients, a video with shorter description length tend to have a higher views. For instance, a video with 100 characters less than the 5th observed video would have lower log(views) by -0.001128420*(sqrt(536) - sqrt(636)) = 0.002332858 which means the average views is exp(0.002332858) -1 = 0.002335581 = about 0.2% lower for a video with shorter description length given all the other variables are the same.

If the total number of likes, dislikes, and comments on the same video increased by 1000 given that all the other variables are fixed, the difference in the log(views) would be 0.9775194*log[(177,384 + 1) / (176,384+1)] = 0.005526314 which means that the average views would increase by exp(0.005526314) -1 = 0.005541612 = about 0.5%.

Lastly, if the author of the video included less capital letters than the lower case letters, given all the other variables fixed, the difference in the log(views) would be -0.1002788 which means the average views would decrease by 1 - exp(-0.1002788) = 0.09541482 = about 9.5%. The 95% prediction interval for the true coefficient of *caps* is between 0.07032247 and 0.13023513. This indicates that the author can increase the average views of a video by 7~14% by including more capital letters than lower case letters. This strategy is more effective than writing shorter title or description, or having 100 more likes or dislikes or comments, given that all the other variables are fixed.

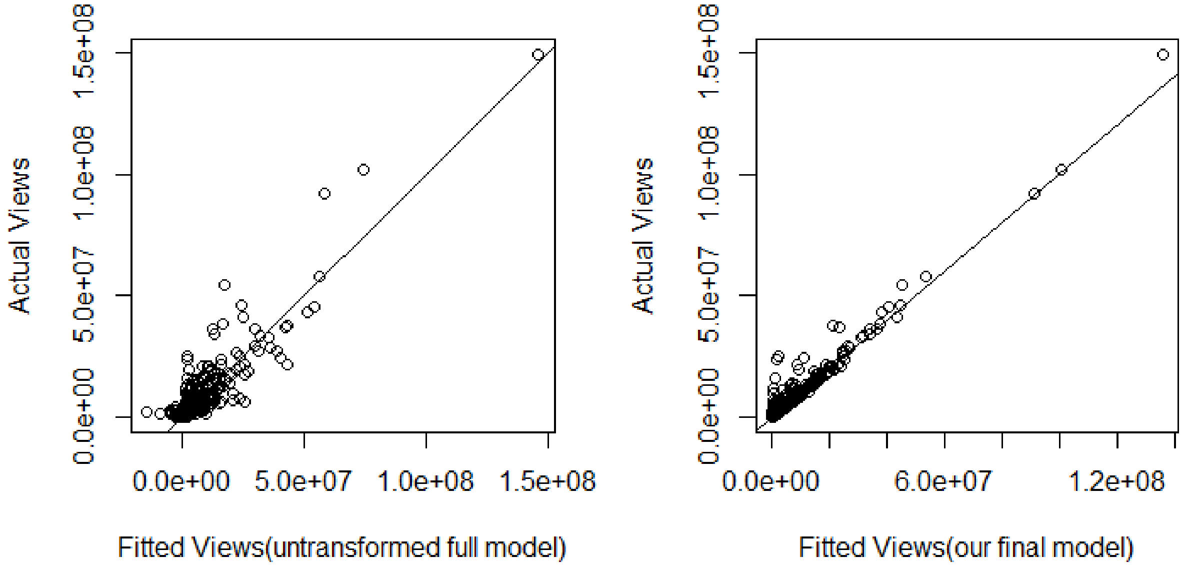


Fig 8. Comparison between the raw full model and the final fitted model

Fig 8 compares the fitness of the untransformed full model and the final model. The final fitted model does not seem to have extreme influential points whereas the full model have a couple of influential points. The mean of actual *views* in the dataset was 1,270,512. The root MSRes of the full model was 1,882,863 in *views*, but the final fitted model decreased the root MSRes to 889,123 in *views* which was 47.2% of the full model. The correlation coefficient between the predicted views and the actual views has also rose from 0.908 to 0.981.

Bibliography

Montgomery, D. C., Peck, E. A., Vining, G. G. (2013), *Introduction to Linear Regression Analysis*, Hoboken, NJ: John Wiley & Sons, Inc.

RStudio Team (2016). RStudio: Integrated Development for R. RStudio, Inc., Boston, MA URL http://www.rstudio.com/.

Mitchell J (2018), “Trending YouTube Video Statistics: Daily statistics for trending YouTube videos”, Kaggle, Available at https://www.kaggle.com/datasnaek/youtube-new/home

Mr.SG (2018), “[Youtube Trends] About the Dataset (full details)”, Kaggle, Available at https://www.kaggle.com/sgonkaggle/youtube-trend-with-subscriber/discussion/57391

Brownlee, J. (2018), “A Gentle Introduction to k-fold Cross-Validation”, Machine Learning Mastery, Available at https://machinelearningmastery.com/k-fold-cross-validation/

Bartlett, J. (2013), “Assumptions for linear regression”, The Stats Geek, Available at http://thestatsgeek.com/2013/08/07/assumptions-for-linear-regression/

Ace X, (2016), “The History of YouTube”, Engadget, Available at https://www.engadget.com/2016/11/10/the-history-of-youtube/