**116th Congress Use of Social Media**

Math 287 Fall 2020 Christian Park Gabe Smithline

Dr. Trent Gaugle Marilyn Golden

Data Modeling Project 1 September 29, 2020

**Introduction:**

There has been, in recent years, added scrutiny on social media’s presence in our society. We have seen social media usage growing in our daily lives. We have also seen growth of social media usage in the arena of politics in general and specifically in the area of elections. This has ranged from advertising a certain candidate or policy to slandering another. It is clear that use of social media has become an integral part of politics.

Social media usage by members of Congress (includes both members of the Senate and members of the House of Representatives) changes based on the time, issues, and events occurring each day. The responses online include: re-tweets, shares, reactions and favorites. Beneath this frenzy of activity, there have been many changes in how Congress members in each party use social media and interact with their constituent audiences.

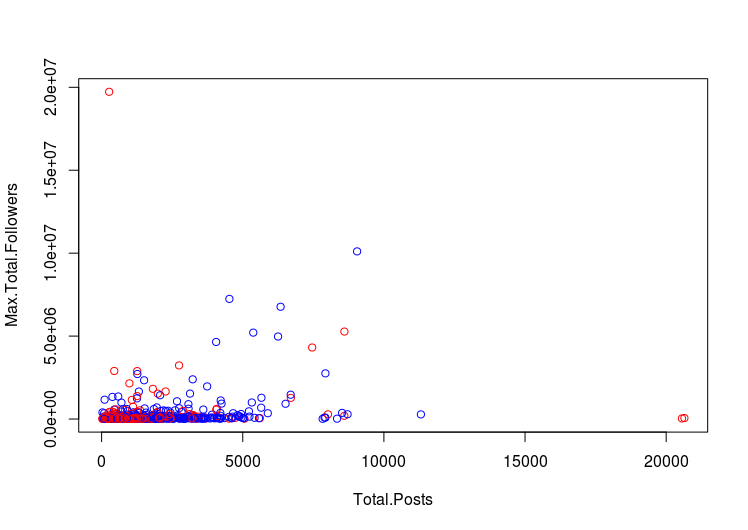
With the upcoming 2020 election and with political tensions at a new high, our project team decided to explore finding a way to analyze the usage of social media by the 535 members of Congress. Congress members use these platforms to influence people and reach constituents. As a result, if a member has a higher number of followers and usage (postings), then they could reach more people who support them and attract those who do not currently support them. Utilizing Pew Research data, we decided to create a model which would help us *predict* a member’s *max total number of followers* as a function of utilizing various social media platforms as well as analyze potential factors/predictors that could have explained the max total number of followers presented in the data. What we were curious to see initially, when thinking about conducting our analysis, was to see if factors like *post activity* and *political party* were good predictors of *max total followers*. (Note: Italicized phrases above are actual column names in the Pew Data we utilized for this analysis).

**Data Description:**

The team acquired the data from the Pew Research Center (*Pew Data 2015 to 2020.csv)* . The columns we analyzed are: *Party, Max Total Followers, Total Posts, Average Number of Reactions to Posts, and, Average Number of Shares.* The *Party* column in the dataframe specified which political party the observations were in as either Republican or Democrat labeled as a ‘R’ or ‘D’. The *Total Posts* column specifies the total number of posts that have been shared by the individual on Twitter and Facebook and the *Max Total Followers* tab shows the highest number of followers attained throughout the use of the individuals social media platforms including Facebook and Twitter. The *Average Number of Reactions* to *Posts* column specifies the average number of likes and other interactions (shares, re-tweets, likes and reactions) that are made from these social media posts. The *Average number of Shares* column specifies the average number of times a post by the individual is shared. The data ranges over the years 2015 to 2020 for the 116th Congress members sworn in January 2019. However, for this analysis, we subsetted the dataframe to include only individuals starting their term in January of 2019 and who are still serving their term. The reason for subsetting the data using this requirement, is to remove the possibility of having multiple entries or giving bias for other factors that could occur in the data from individuals that serve shorter or longer terms as well as eliminating endogeneity with factors like the state of the economy. Utilizing the original data file without this subsetting, might result in the data not being independent which would disallow our using of the model selection methodology.

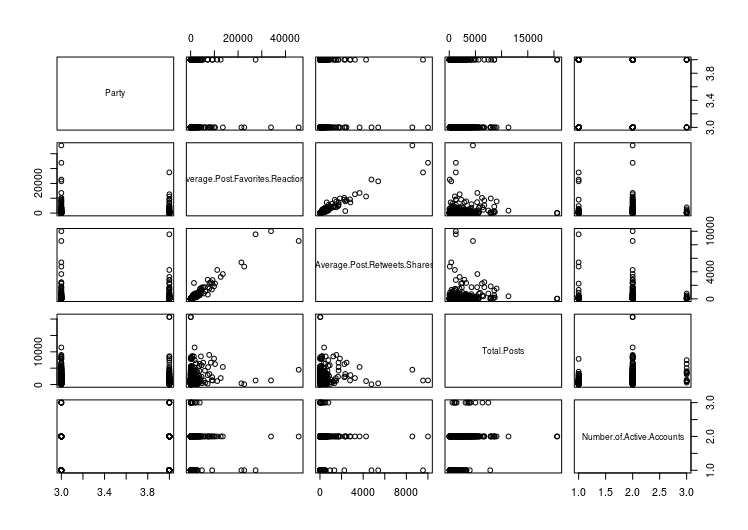
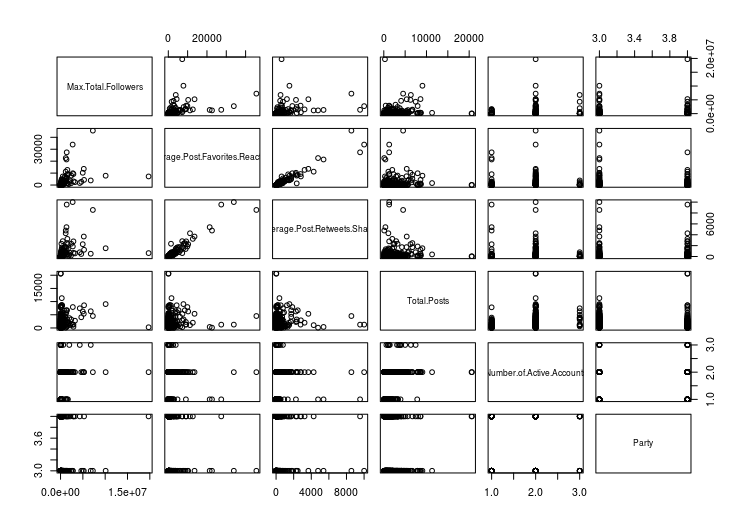
The overall objective in this analysis is to find if any single or multiple of these variables have any significance in helping us create a model to predict the *Max Total Number of Followers* as well as helping explain *the Max Total Number of Followers* present in the data*.*

**Model Selection:**

****

***Figure 1.1***

When initially looking at the data, the first step in our process of building a model was to see if there was a distinct difference between members of a *Party* and their T*otal Number of Followers.* We initially created a scatter plot to illustrate how the *Total Post*s related to the T*otal Number of Followers.* We highlighted this result with the *Party* by using red points for Reublican members and blue points for Democratic members. Looking at ***Figure 1.1,*** it shows that the majority of the Democratic Party points (blue) are scattered at a higher level on the y-axis and have more *Max Total Followers* in comparison to Republicans. This observation led us to hypothesize that *Party* might be a significant variable in the model for predicting and explaining the trends in the data for *Max Total Followers.*

****

***Figure 1.2 Figure 1.3***

We then attempted to find other variables that could be strong predictors of followers as well, as we wanted these to help increase the accuracy and significance in the model for *Max Total Followers.* We utilized the Automated Stepwise Model selection process and looked at the **AIC** results it gave to assist us in our initial choice of predictors.

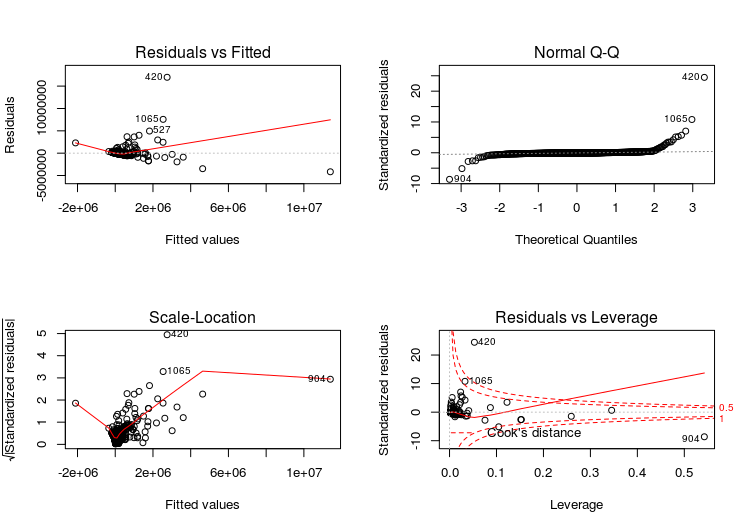
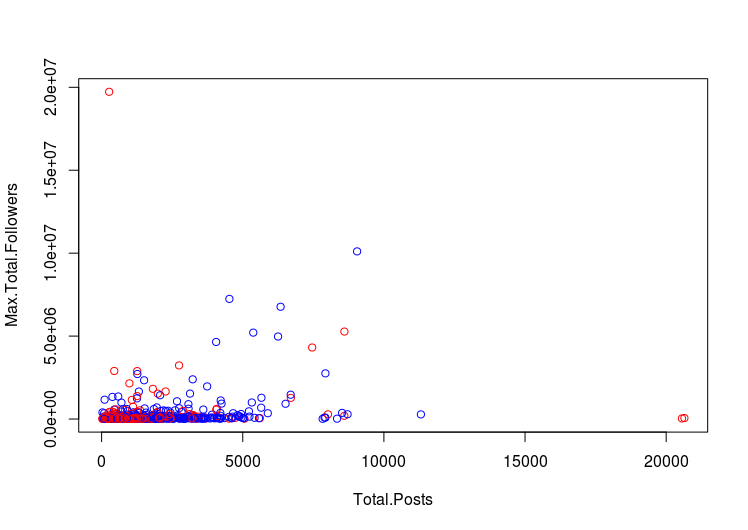
What we found when conducting this method for model selection was that the most significant predictors for *Max Total Followers* were *Average Number of Favorites and Reactions on a Post, Party, Average Number of Retweets and Shares on a Post, Total Number of Posts,* and *Number of Active Accounts.*

After finding these predictors, we wanted to see the relationship between these predictors to help us gain a better understanding of the direction of the relationship as well as if there could be possible multicollinearity among these predictors. As seen in ***Figure 1.2***when initially looking at the matrix there is a very strong positive relationship between Average post favorites and reactions and average posts reshares. This led us to hypothesize that these predictors may pose **issues with multicollinearity** in the model. Along with what we saw in ***Figure 1.2***, when looking at the matrix showing the interaction between *Party* and other predictors, we noticed that Republican’s consistently had much lower y-axis values in comparison to Democrats. After looking at the matrix in ***Figure 1.1*** and hypothesizing the potential predictors that could cause issues with multicollinearity, we created the initial model and ran the summary statistics as well as the **VIF.**

The summary shown in the figure below labeled ***Model 1 Summary*** shows very low *p*-values leading us to reject the null hypothesis. However, we noticed an issue with multicollinearity in the model by finding the **VIF** of all of our predictors. For example, we saw in the matrix an issue with A*verage Favorites and Reactions and Average Number of Shares* as these predictors posted the highest VIF greater than 12*.* It would seem that a member with more shares would have more reactions to their posts. With these high VIF results in the model, we decided to remove the *Average Favorites and Reactions* predictor from the initial model to eliminate the multicollinearity.

After removing this predictor to eliminate the issue of multicollinearity, we decided to plot this new model which included all the same predictors from the previous model except for *average number of favorites and reactions.*  We wanted to look at the **residual plots** as well as the **VIF** and **summary** statistics. The summary statistics showed low *p*-values for all of our predictors and all of the VIF values for these predictors were below 5. This helped us reject the null hypothesis that these predictors were not significant as well as helping us remove multicollinearity among the model after removing the *average favorites and reactions* predictor from the previous model However, as it can be seen in ***Figure 1.4*** below, the **residual plots** did not look very promising in terms of the spread of the variance within the data. The majority of the points were shifted towards the left and spread rather far from each other towards the right.

Having noted the problems in the **residual plots**, we decided to conduct a **power transformation** in an effort to correct these issues seen in the residuals plot which would ultimately help us make the model more linear and make these predictions more viable..



***Figure 1.4***

Residuals:

Min 1Q Median 3Q Max

-4160842 -109515 -57916 22714 16975952

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -339912.71 23774. 2.746 0.00614 \*\*

Average.Post.Favorites.Reactions 485.22 32.50 14.931 < 2e-16 \*\*\*

Average.Post.Retweets.Shares -1288.62 125.97 -10.230 < 2e-16 \*\*\*

Total.Posts 84.52 14.44 5.854 6.47e-09 \*\*\*

Number.of.Active.Accounts 138932.35 3834.16 2.176 0.02975 \*

PartyR 99288.92 46230.58 2.148 0.03198 \*

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 719500 on 1009 degrees of freedom

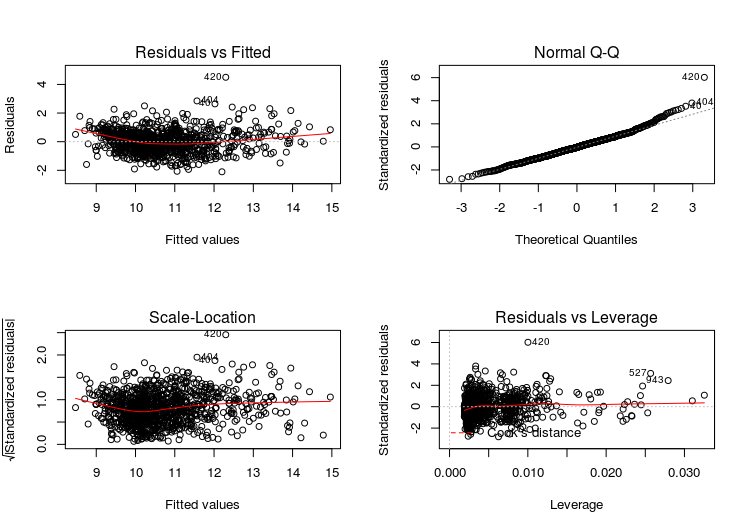
(37 observations deleted due to missingness)

Multiple R-squared: 0.3389, Adjusted R-squared: 0.3357

F-statistic: 103.5 on 5 and 1009 DF, p-value: < 2.2e-16

***Model 1 Summary***

Using the **power-transformation** function in order to transform the predictors to suit the model better, we took the log of the M*ax Total Followers,* the *Average Number of Retweets and Shares* andthe *Total Posts* while squaring the *Number of Active Accounts* and keeping the P*arty* predictor unchanged as it is a categorical variable. When looking at ***Figure 2.1*** below, we saw that this power transformation process resolved all of the issues that were seen in the previous residuals plot. The new residuals plot on the next potential model with the transforms included, resulted in a much more even spread throughout the residual plots as well as a showing a relatively flat line.



***Figure 2.1***

To cover all the bases in the end to help us come to a conclusion for solidifying this model as our final model we performed a best subsets regression. We looked at the BIC values to determine if this 4 predictor model after the power transformations was the optimal model. Looking at ***Figure 2.2*** below, it did show that the four predictor model has the lowest BIC helping reassure us that this was the best model. Finally, after completing all the steps to improve the initial model coming out of the **AIC** process, we ran a **summary** of the transformed model to review the statistics. We felt very comfortable in using this array of variables to predict the *total number of followers*. To our surprise, our expectations were confirmed as *political parties* and *post activity* were two very strong predictors used in our model for predicting the *total number of followers*.

**Model Overview:**

Residuals:

Min 1Q Median 3Q Max

-2.0740 -0.5163 -0.0296 0.4559 4.4846

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 6.76375 0.19853 34.069 < 2e-16 \*\*\*

log(Average.Post.Retweets.Shares) 0.67385 0.01710 39.409 < 2e-16 \*\*\*

log(Total.Posts) 0.22155 0.02917 7.594 7. 04e-14 \*\*\*

I(Number.of.Active.Accounts^2) 0.06212 0.02080 2.986 0.00289 \*\*

PartyR -0.27638 0.04907 -5.632 2.30e-08 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

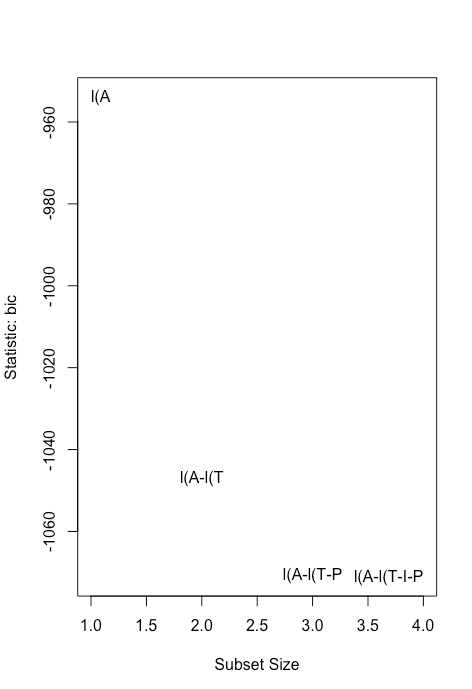
Residual standard error: 0.7502 on 1010 degrees of freedom

(37 observations deleted due to missingness)

Multiple R-squared: 0.6594, Adjusted R-squared: 0.6581

F-statistic: 488.9 on 4 and 1010 DF, p-value: < 2.2e-16

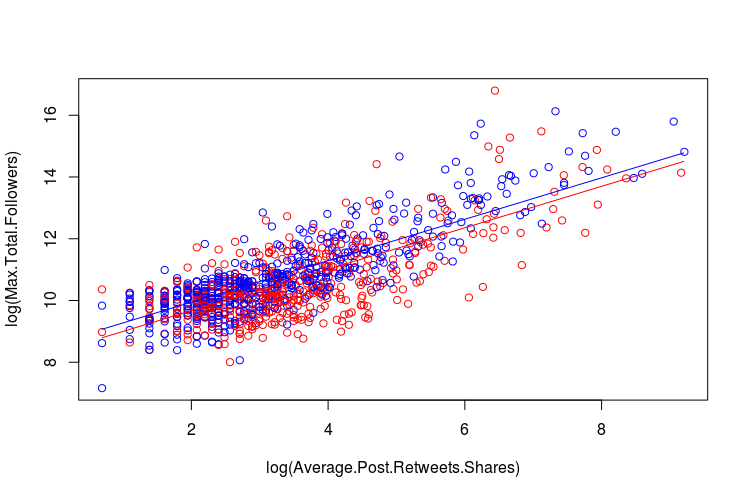
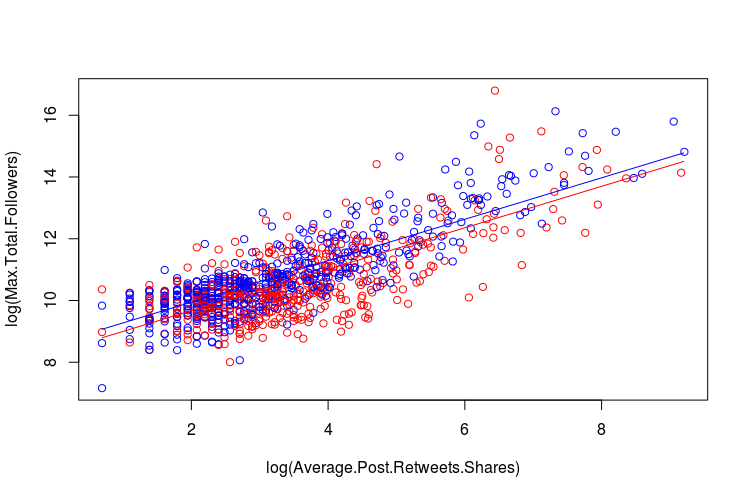
***Model 2 Summary***



***Figure 2.2***

Looking at our final model, when evaluating the summary statistics, it can be seen that all of our predictors we kept in the model are significant. All of the predictors have a *p-*value much less than .05 leading us to reject the null hypothesis that they all have no effect on the *max total followers* with 95% confidence. The R2 of the model was also relatively high at .65 which shows that 65% of the variance in the data was explained by the model. Overall after conducting the **VIF** tests and looking at the **summary** statistics and **residual plots**, we can conclude that these four variables are the best set of predictors, given our data, for finding out the *max total number of followers* for members of the 116th Congress.

**Conclusion:**

****

***Figure 2.3 Figure 2.4 (Interaction/Non Parallel Lines)***

We found that all of the predictors in the initial AIC model were statistically significant. But, when we analyzed the matrix to see the relationship among the predictors and calculated the VIF of our predictors, we found an issue with multicollinearity *among average post shares* and *average likes and reactions.* As a result, we removed the variable with the highest VIF from the initial model. The significance increased showing our new model to be stronger with no issues of multicollinearity present. We had better AIC values, much lower *p*-values and a much higher R2 value compared to our initial model earlier in the model building process. Although these summary statistics and AIC values looked promising, when looking at our residual plots we noticed that the trend of points didn’t have an even spread of variance. With that observation we performed a power transformation on all the predictors which ultimately fixed the residual plots. All of these successive steps led us to conclude that we should reject the null hypothesis that all of our predictors after the power transformation *(Average Amount of Retweets/Shares Per Post, Total Posts, Number of Active Accounts, and Party*) have no effect on *Max Total Follower*s in favor of the alternative hypothesis that they do impact the *log of Max Total Followers*.

With the final model being utilized, we can interpret that being a Republican will lead to having fewer followers compared to a Democrat if all other predictors remain constant. This is also supported when looking at ***Figure 2.3*** as the model and the line of best fit shows that Democrats ultimately have a higher number of max total followers in comparison to Republicans. In general, we can infer that this trend with the relationship between party and max total followers has to do with the demographic that follows these respective parties. Typically it has been seen that Democrats appeal to a younger population which generally will be more active on social media. This could explain the trend in the data we see as to why there is such a distinct difference between the parties. Also the interaction between party and the other predictors seemed to be parallel as when looking at ***Figure 2.4*** it can be seen even when creating interaction between party and the log of the average post shares the lines continue to remain parallel. We can ultimately interpret that *Party* does have an effect in the total *max number of followers* one has in our data, as well as other factors such as  *total posts* and *number of active accounts* having a significant effect on the max total number of followers. Although, we cannot clearly interpret the model in terms of what a one unit increase in a predictor would lead to for the specific coefficients after conducting the transformations on the predictors. But, we can be confident that the final model we chose ultimately had an effect on the M*ax Total Number of Followers* in the data. To use the model and find the *max total number of followers,* we would have to perform a back- transformation to the given y-value in our model as this would then give us the *max total number of followers* in the data at a specific point. We can also be confident in the model’s ability after only including entries from a given specific time period as external factors including the state of the country and policies that were put in place during that time period to have an effect on the max total number of followers. To further our point that this four predictor model was in fact the best, we also performed a best subsets regression and analyzed BIC as well which ultimately showed us that the four predictor model was the best. In conclusion we were able to see that the Party factor had a drastic effect on the max total followers in the data which was very surprising to see and helped support our hypothesis.

**Bibliography**

Kessel, P. van, Widjaya, R., Shah, S., Smith, A., & Hughes, A. (2020, August 28). *How Congress Uses Facebook and Twitter*. Pew Research Center: Internet, Science & Tech. https://www.pewresearch.org/internet/2020/07/16/congress-soars-to-new-heights-on-social-media/.