Number of Hours Spent on Social Media

Ada Niu[1] and Golnazsadat Abrishami Osgoui[2] and Rashi Saxena[3]

***Abstract*— In this research, we are investigating a linear relationship between different factors and spending hours on social media. We used convenience sampling to create our dataset. We have four different approaches to the hypothesis. We investigate the relation between the independent variables and the dependent variable in the dataset in three ways, by using the coefficient of determination and p values, using a scatter plot, and checking the linearity assumptions. In the end, we find that age and gender have an impact on the spending hours on social media, but we do not have any linear relationship is not a linear relationship between the predictor variables and the response variable.**

I. INTRODUCTION

The earliest method of communicating great distances used written correspondence written and delivered by hand over an elongated periods of time. Today, we can connect to anyone in the world within seconds. How is this possible? The answer is technology, especially, social media. Social media has single-handedly changed the way we communicate. Social media has provided humans with a platform to share raw, personal moments with their loved ones living thousands of miles away. It has helped businesses by providing a platform for digital advertising- it is extremely rare to find a business that does not have an online presence. It has brought news back into the millennial life. With the perks of social media, we have also been informed about the cons of it. During a light conversation regarding this project, we began to discuss how much time we personally spend on social media, which brought a question to our attention- How much time to people on average spend on social media? According to a study conducted by Common Sense Media 60 percent of teens- on average spend 9 hours a day on social media.1 After reading a little bit about how time on social media by users is growing, we were curious to research the implications of this behavior- Research has suggested that young people who spend more than two hours a day on social networking sites are more likely to report poor mental health.2 This has become such a growing problem that companies, such as Google and Apple, are looking into adding features to help parents restrict the time their kids spent on social media. Our research on this topic inspired us to investigate if we could predict the number of hours spent on social media by a person. Our goal is to have accurate predictions to help develop features in phones for specific users to restrict their time spent on social media.

[1]A. Niu is with Student of Data Science, University of the Pacific

[2]G. Abrishami is with with Student of Data Science, University of the Pacific

[3]R. Saxena is with with Student of Data Science, University of the Pacific

II. DATASET

We decided to collect our own dataset for a couple reasons-1) we were not able to find the dataset, with the features of our interest, we were interested in online 2) we wanted to work with a dataset as clean as possible. To conduct our research, we began with creating a Google survey form with questions of interest to us- age, gender, marital status, occupation, education level, total number of active social media accounts, and hours spent on social media daily. Our predictor variables are a persons gender, age, occupation, marital status, number of social media accounts she/he has, and his/her education, and our response variable is her/him spending hours on social media. The scale of measurement for age, number of social media accounts, and spending hours on social media is the ratio, and it is nominal for gender, occupation, marital status. The educations scale is ordinal. We used the convenience sampling method to collect the data. Although we would have ideally wanted to use simple random sampling, which is recommended for survey research, due to time restrictions we were not able to randomly pass out our survey, so we used Facebook posts and emails to friends to encourage them to answer our questions, and also asked to pass it on to whomever they know too. We were hoping for a large data set because according to central limit theorem, in a large sample size, at least larger than thirty, the x has a relatively normal distribution. Fortunately, we were able to obtain a relatively large dataset with one hundred sixty-nine observations. Therefore, the mean of our sample is normally distributed.

III. HYPOTHESIS

From our research, we saw most studies using the age of the group spending time on social media, so for our research, we were mainly interested in the relationship between two predictor variables- age, and number of social media account and the response variable- hours spent on social media.

Null Hypothesis: Taken together, Age and Number of Social Media account (NSM) are not useful for predicting Hours spent on.The coefficients of Age and NSM are equal to 0.

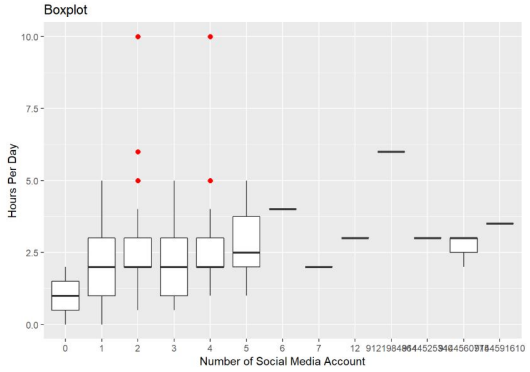
$$\beta_1 = \beta_2 = 0$$

Alternative Hypothesis: Taken together, age and number of social media accounts are useful in predicting the time that a person spend in social medias per day. Atleast one of the coefficients is not equal to 0.

$$\beta_1 \neq \beta_2$$

A. Data Cleaning

Before cleaning, it is best to gain some understanding of the dataset- we used the boxplot and the scatter plot.The boxplot, besides the outliers, reveals some anomalies in the number of social media accounts and age.



So, we checked our dataset and figured out that this anomaly derives from incorrect data-a few people have written their phone numbers in the number of social media accounts column. We also notice a boxplot for 0 social media accounts, which could be puzzling. The reason for this is because when we sent our survey out, we had only accounted for Facebook, Instagram, Pinterest, and Snapchat as our social media platforms, but to some of our responders count WhatsApp, WeChat, and Reddit as social media accounts so they entered 0 for the the number of social media accounts we asked for, and entered the number of hours they spend on social media using other platforms.

IV. Models

Now, we create our multilinear regression models. We will be using the the adjusted Rsquared value and p value to help us decide whether we can reject our the hypothesis. One of the most significant parameters of a linear regression model is the coefficient of determination or R2, which shows how many percentages of the variation in the response variable is explained by the predictor variables or the regression. The other important parameter is the Pvalue.

In the first model, we use only age and number of social media accounts as predictors which shows pretty low R2. In this model, only two percent of the variation of the response variable is explained by predictors. But, our p value is not low at .08, so at the .05 significance level, the data does not provide sufficient evidence to conclude that age and number of social media accounts are together useful for predicting hours of social media per day. We were not satisfied with this answer, so we explored the other variables in our data.

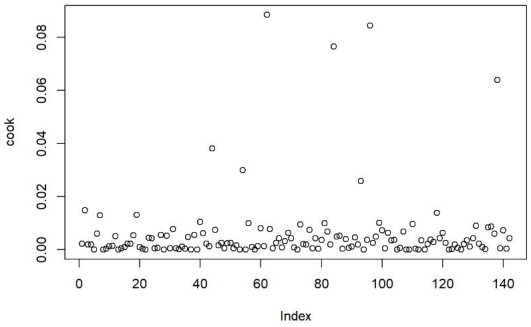
In the second model, we plug in all the explanatory variables into the model. In this model, we see a slight improvement in our R2, but also the pvalue increases, which shows that our current predictors do not have a relationship with the number of hours on social media per day. But looking at the pvalues of each variable, we can see that age, number of social media per hours, and the gender(male specifically) shows a pvalue of less than .05.

So in the third model, age, gender, and a number of social media accounts are the explanatory variables. We see a slight jump in R2 but also a very low Pvalue. We can also notice that the number of social media accounts, has a high Pvalue, so it might be interesting to look at a model with just age and gender as the predictor variables.

Finally, in the last model, we just use gender and age as explanatory variables. Our R2 is around the same as the third model-this explains to us that adding number of social media accounts as a predictor variable would not have helped our mode. We have a pretty low p value of .003, from which can conclude that at .05 significance level, the data provides sufficient evidence that age and gender taken together are useful predictors of number of hours spent on social media. Although our p value is pretty low, we have a low Rsquared, so we want to investigate the linear relationship between our predictor variables and response variable, and see if we can identify any influential outliers or multicollinearity to improve our model.

V. Result and Discussion

We used Cooks Distance to identify any influential outliers, and as you see from our plot, there are none. According to plot we do not have any observation with coock distance greater than one. So, none of the outliers are influential and we can keep them. Let's see if we have a way of finding some correlation between coefficients, meaning that two predictor variables are correlated.

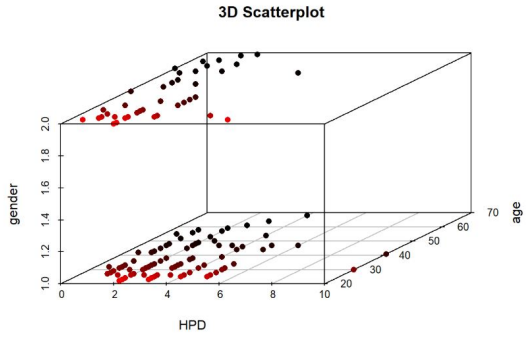


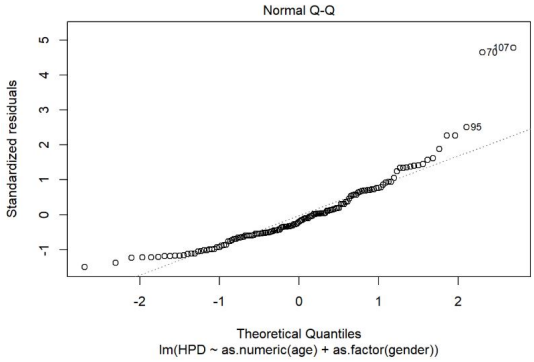
Next, we check the variance inflation factors for this two predictors. We notice that our variance influence factor values are close to 1, which indicates that the standard deviation of the coefficients will remain stable with the inclusion in the regression equation of the other predictor variables. Therefore, we do not have any correlation between our predictor variables.

We also use 3D scatterplot, to see a relationship between variables, and we do not see any linear relationship between our predictor variables and the response variable. Let's see if our model violates linear assumptions.

A. Linear Assumptions

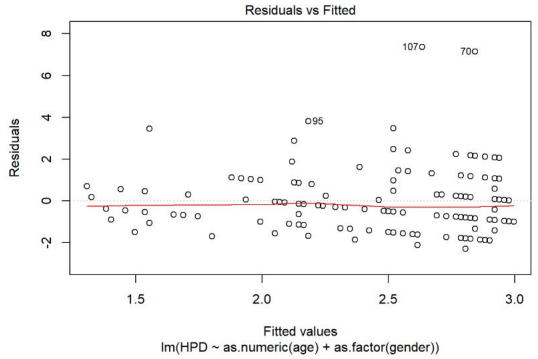
There are four assumptions(conditions) for regression inferences -1) Population regression line 2) Equal standard





QQ.JPG

deviation 3) Normal population 4) Independent observations. The fourth assumption (Independent observations) is valid because the observations of the response variable are independent of one other, in which one persons response will not affect anothers response. The first three assumptions can be examined from the residual plot and the normal Q-Q plot. Our residual plot shows that the conditional standard deviations of the response variable are not the same for all values of the predictor variable since the dots are clumped near the zero line and they do not form an exact horizontal band. We can try to eliminate the inequality of variance by implementing log or inverse transformation. However, due to the nature of our dataset, this method does not help us to increase R2 that much. We can also conclude that our model is slightly right-skewed, which is not normally distributed from the normal Q-Q plot. From our examination on assumptions for regression inferences, we can conclude that linear regression model may not be the best model for our dataset.



VI. CONCLUSION

Given our p-value in our final model, we can say that taken together, age and gender indeed have impacts on hours spend on social media. But we can also say that given our significantly low adjusted R squared, there is no linear relationship. Though regression model is not the most ideal model for our dataset, through our regression analysis, we learned what variable matters the most, and how certain are we about all these variables. Our last model shows partial disapprove of our original hypothesis, in which number of social media accounts is in fact not a significant influence factor, but shows us that gender is a better variable to look into when predicting hours spent on social media.

A. Research Limitations

There are some limitations and further improvements worth discussing for this research topic. One concern can be related to the sample size. One hundred sixty-nine observations are not sufficient enough to draw accurate conclusions for the whole population. Another concern is that the conception of Social Media may vary from one person to the other, which may affect our research result. It would be better if we give out definition of social media in our survey. Moreover, without time restrictions, we could use simple random sampling, which is the recommended approach for survey research, such as by giving out surveys on the street to random people to have an accurate depiction of the population. In real world example and datasets, our research method can be a good way to examine potential causal relationship or build a linear relationship between predictor and response variables.

REFERENCES

[1] https://qz.com/1367506/pew-research-teens-worried-they-spend-too-much-time-on-phones/
[2] http://www.bbc.com/future/story/20180118-how-much-is-too-much-time-on-social-media
[3] http://circaedu.com/hemj/how-social-media-changed-the-way-we-communicate/
[4] https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/f-statistic-value-test/