# Social Media Usage and Mental Health

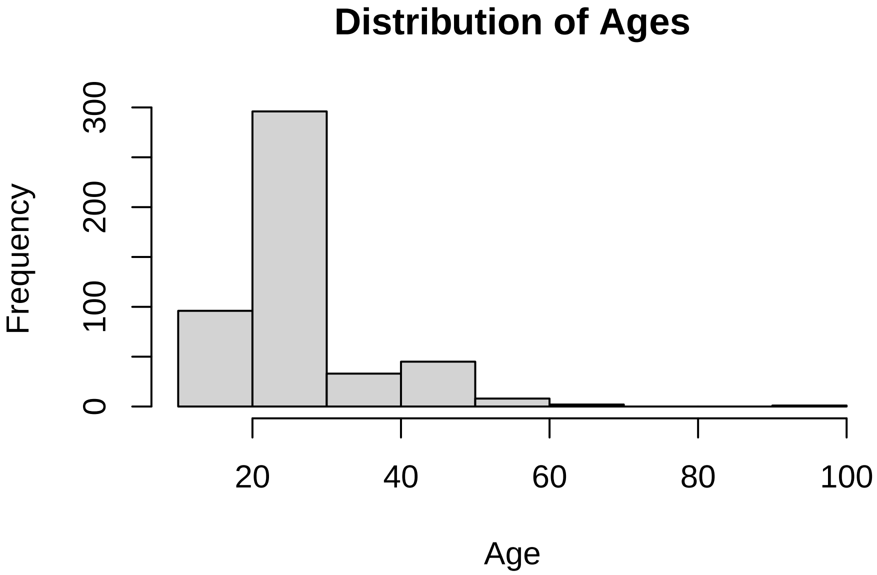
By: Aidan Ouckama

## Introduction

With the expansion of technology within our modern society, comes a multitude of questions on the pros and cons of said technology. Efficiency, of course, is a clear benefit to technological advancements and developments, however, with efficiency comes laziness and comfortability. Because of technology, everything is accessible with a single click, or a single swipe. What we should be asking is, how is this impacting us? Social media, one of the products of this technological revolution, is one of the key contributors to society's worries of technology. Social media provides constant stimulation: likes, follows, updates. It was created to hook people into it, make them become addicted to the application. How is this product of technology impacting our mental health? Our attention span? Our overall well-being? As we embark on this statistical exploration, our primary objective is to dissect the correlation between social media usage patterns and various facets of mental well-being. Are individuals who spend more time on social media platforms prone to higher levels of stress, anxiety, or depression? Does the pursuit of virtual validation through likes and comments contribute to a distorted sense of self-worth? Moreover, how does the constant influx of information affect our attention spans, cognitive abilities, and overall psychological resilience? By employing statistical methods, such as fitting linear models and cross validation, we will find the answers to these questions. The analysis on our data collected will provide valuable insights into the potential risks and benefits associated with our digital interactions, fostering a nuanced understanding of the impact technology has on the human psyche.

### Data Description

The data collected and analyzed is survey data on different individuals' social media usage and questions about different aspects of their life, which will be addressed later. The survey consisted of about five hundred participants, who were asked 20 different questions about their social media usage and a multitude of questions on their mental health, attention span, and other aspects of their life that may or may not have been impacted by their social media usage. Although there is a large range of questions that were asked to the participants, they can be grouped into three categories, mental health / self-image, addiction, and procrastination. The answers to these questions were then given on a scale from 1 to 5. Evaluating the data, most of the participants who answered the survey are between 20 and 30 years old, which makes sense for our analysis, as this age group has high diversity with social media usage, allowing for more variation within our data.



Cleaning the data involved manipulating certain qualitative columns. Instead of qualitative data (e.g. one of the columns involved naming the different social media platforms), the data was transformed into quantitative data (e.g. an integer representation of the different amounts of platforms). Irrelevant columns such as “How much are you bothered by worries?”, and other columns which are unrelated to the input and the three categories of output (mental health / self-image, addiction, and procrastination), were removed from our data as well.

The columns which will be used as input variables for our regression analysis will be the amount of hours a user has indulged in social media, and the manipulated data on the amount of different platforms the user used. Along with these input variables, gender and age will be taken into account as well, to see if these are also factors that impact our model. With the data collected, we will be searching for three different trends with the three categories of output mentioned above. Our null hypothesis will be that there is no correlation between social media usage and the listed output categories. Our alternative hypothesis is that social media does statistically have some sort of positive or negative correlation with the three output categories. Along with this data, real world data collected from friends and family in the same format as the original data was gathered, which was used with generalized linear models as a "test" dataset.

**Methods**

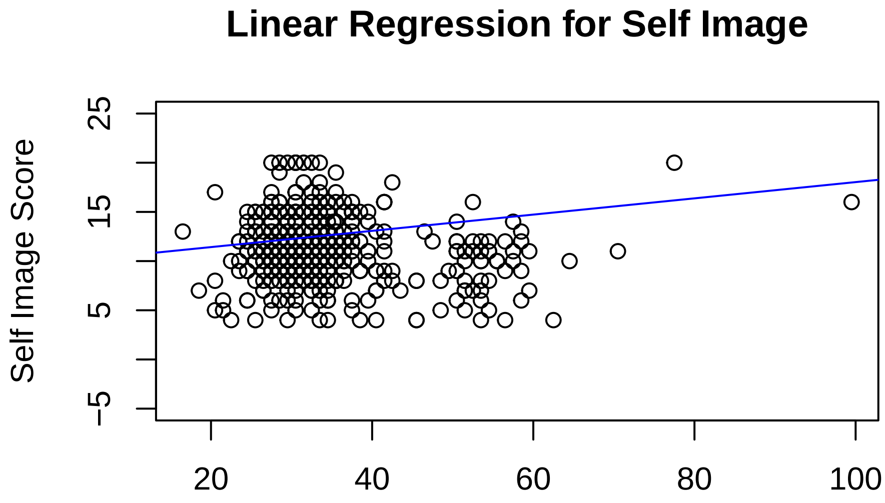
The methods used to evaluate our data include a blend of linear regression, LASSO, and k-fold cross validation. For each of the three output categories, a linear regression model was fitted with the four input variables, hours of social media, amount of different social media platforms, age, and gender. The reasoning behind using linear regression to fit these models is to find if there is any correlation between the input variables and the output, in this case, discover if there is a relationship between social media usage and a person's mental health, procrastination habits, and addiction symptoms. If we do find that the input of social media usage has statistical significance to any of the output categories, then we can reject the null hypothesis and conclude there is in fact some sort of correlation between social media and our psyche.

LASSO was used for variable selection, and to pick and choose which input variables were relevant to our outputs. This stage of the project was not only relevant to the main hypothesis, but it was also used to find a better fitting model for predictive uses. LASSO was initially used to double-down on figuring which input variables were relevant to our desired outputs, as irrelevant variables were zeroed out when fed through the LASSO regression. The difference between LASSO and regular linear regression is there is a cost to each of the variables being added to the model, which allows for more generalized models and prevents overfitting. These generalized models were then used with the hand collected data as a test, to have a real world aspect to the project.

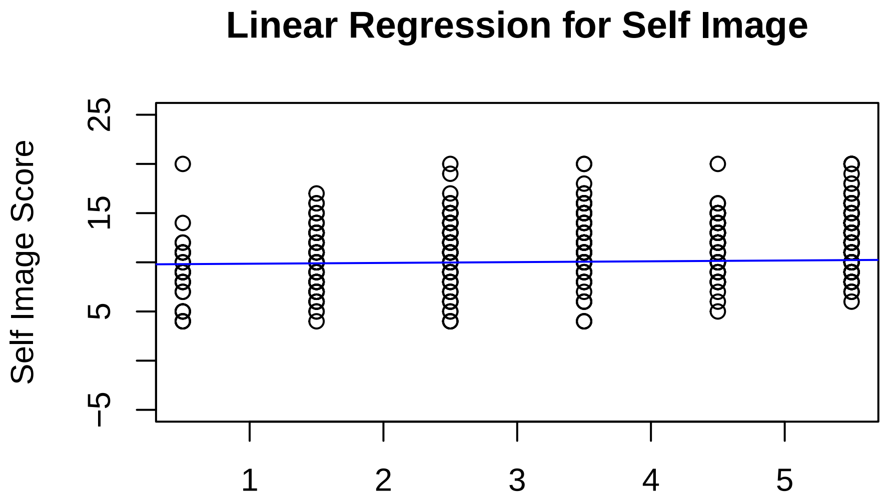
Finally, k-fold cross validation was used to find the cost value for the LASSO regression model. The cost value is important to the model because a value too high can produce an under-fitted model, and a cost too little can produce an over-fitted model. So to find the sweet spot for this c value, k-fold cross validation was used to find coefficients that are well fitted for the training data and test data, resulting in a model that isn't overfitted, but still is generally accurate when given input.

**Results**

With the first stage of the project, the linear model fitting, it was clear there was definitely a correlation between a user's social media usage and the three output categories: mental health issues, addiction, and procrastination. However, with the other input variables, it was not as clear cut.



Above is a graph displaying the linear regression line on one of our models. Due to the congestion created by the variables that do not have as much correlation to the output as social media usage, the graph presents data that does not seem to have any correlation. However, below is a graph with only social media usage as the input variable.



This is the case with all three models, where social media usage had significant correlation with the output, and the other variables had no correlation (and a p-value of $> 0.05$). There was, however, some correlation (p-value $< 0.001$) between age and procrastination, with there being an inverse relationship. However, this is most likely due to the maturity and time management skills an adult develops as they get older, rather than it being to do with social media. But, it was an interesting observation. This analysis showed that no matter the age or gender of a social media user, there is correlation between usage amount and negative mental issues, such as negative self-image, addiction, and procrastination. Now the fun part, fitting a model and testing it on real world data! However, our models are not generalized to handle real world data, therefore we utilized LASSO to do variable selection and regularization on our current linear models. Below are the coefficients and the cost values of each of our LASSO regression models for self-image, addiction, and procrastination respectively.

```
## [1] "Self Image Cost"
```

```
## [1] 0.05556545
```

```
## [1] "Self Image Coefficients"
```

```
## 5 x 1 sparse Matrix of class "dgCMatrix"
##                        s0
## (Intercept)     9.89918449
## platform_count  0.06011540
## avg_time        0.46068712
## age            -0.01932608
## gender          0.16873825
## [1] "Addiction Cost"
```

```
## [1] 0.04647298
```

```
## [1] "Addiction Coefficients"
```

```
## 5 x 1 sparse Matrix of class "dgCMatrix"
##                         s0
## (Intercept)     4.406620748
## platform_count  0.045432304
## avg_time        0.415422245
## age            -0.008406133
## gender         -0.011777943
```

```
## [1] "Procrastination Cost"
```

```
## [1] 0.05667058
```

```
## [1] "Procrastination Coefficients"
```

```
## 5 x 1 sparse Matrix of class "dgCMatrix"
##                        s0
## (Intercept)    11.36685856
## platform_count  0.07933053
## avg_time        0.93269960
## age            -0.05420791
## gender          .
```

Through these coefficients we can see that gender was zeroed out for two out of the three models, and in the self-image model, it has very low impact. This is the case for every input variable except for average time on social media. We saw this was the case within our original linear models, so LASSO solidified the information we already knew. The cost values for each LASSO regression was found using k-fold cross validation. Models with different cost values were evaluated and cross referenced until the best model determined which cost value was the most effective.

These generalized models were then given real world data from my myself, my friends, and my family, to see how accurate they were. These were the results:

```
## [1] "Predictions"
```

```
##             s0
##  [1,] 12.40713
##  [2,] 12.30623
##  [3,] 12.28477
##  [4,] 10.92417
##  [5,] 12.92282
##  [6,] 12.59520
##  [7,] 11.34834
##  [8,] 12.61239
##  [9,] 11.61371
## [10,] 11.61157
## [11,] 10.42941
## [12,] 10.78889
## [13,] 11.02507
## [14,] 10.90698
```

```
## [1] "Y"
```

```
##       [,1]
##  [1,]    8
##  [2,]   10
##  [3,]   17
##  [4,]   12
##  [5,]    8
##  [6,]   15
##  [7,]   14
##  [8,]   15
##  [9,]    8
## [10,]   13
## [11,]    6
## [12,]   12
## [13,]   11
## [14,]   12
```

### Conclusion

Overall, this project was a really fun experience. I had the chance to use data to fit a linear model and “get my hands dirty”. Unfortunately, the results I found were that social media usage had correlations to a multitude of negative mental impacts, such as: self-image, addiction, and procrastination. The time in which I am doing this project alone is proof for these findings. I chose this topic because it was important to me, as social media is greatly embedded within my life. I was only going to fit the linear model and call it a day, but this project was calling for me to challenge myself. I decided later on to fit a LASSO model, and use that model to predict values on real world “test” data. This is where I got my hands really dirty, and where I had the most fun. I had a large knowledge gap when it came to LASSO and anything a step further regular linear regression, so this part of the project definitely challenged me. But, it was fun collecting data from my friends and family, and although the model was not the most accurate, I was proud of myself for playing around with real data and creating my own model. In terms of reflection, I definitely want to work with bigger and more specific data, rather than a survey with linear 1-5 answers. I also saw the kinks of real world data being given to a fitted model, as one of my test subjects brought up, her answers to a lot of the self image and distraction questions were due to other things besides social media, which is definitely something to keep in mind on the next project. All in all I loved this project, along with the class, and I’m super happy I took this course.