Research Project

Investigating the Association between Social Media Use and Mental Health

Overview

1) Introduction
2) Data Collection
3) Data Cleaning and Preparation
4) Statistical Modeling
    a) Linear Regression
    b) Logistic Regression
5) Conclusions

Introduction

- Social media is a huge part of the daily lives of billions of people around the globe
- Many scroll through social media for hours a day, without thinking twice about its consequences
- **Question of Interest:** Is increased social media usage related to worse mental health?

Data Collection

- Collected survey data on 20 quantitative and qualitative variables encompassing demographics, social media use habits, and mental health
- Received 310 survey responses

Demographic Variables

Continuous

- Age (in years)
- Estimated gross annual income

Categorical

- Race
  - American Indian or Alaskan Native, Asian, Black or African American, Hispanic or Latino, Native Hawaiian or Pacific Islander, White, Other
- Gender
  - Male, Female, Other
- Highest Level of Education Completed
  - Middle school or less, High School, Bachelors, Masters, PhD/MD/other professional degree
- Employment Status
  - Full-time, part-time, self-employed, student, unemployed, retired

Social Media Use Variables

Continuous

- Estimated time spent on social media per day (in hours)
- Estimated length of social media use (in years)

Discrete

- Number of social media platforms used regularly

Binary

- Do you use social media? (yes/no)
- Do you use social media for work? (yes/no)

Mental Health Variables

Continuous

- Estimated sleep duration per night (in hours)

Categorical

- Cyberbullying (yes/no/unsure)

Discrete

- Stress levels measured from 0 (not at all) to 10 (very severe)
- Anxiety levels measured from 0 to 10
- Depression levels measured from 0 to 10
- Loneliness, low self esteem, inadequacy measured from 0 to 10

Binary

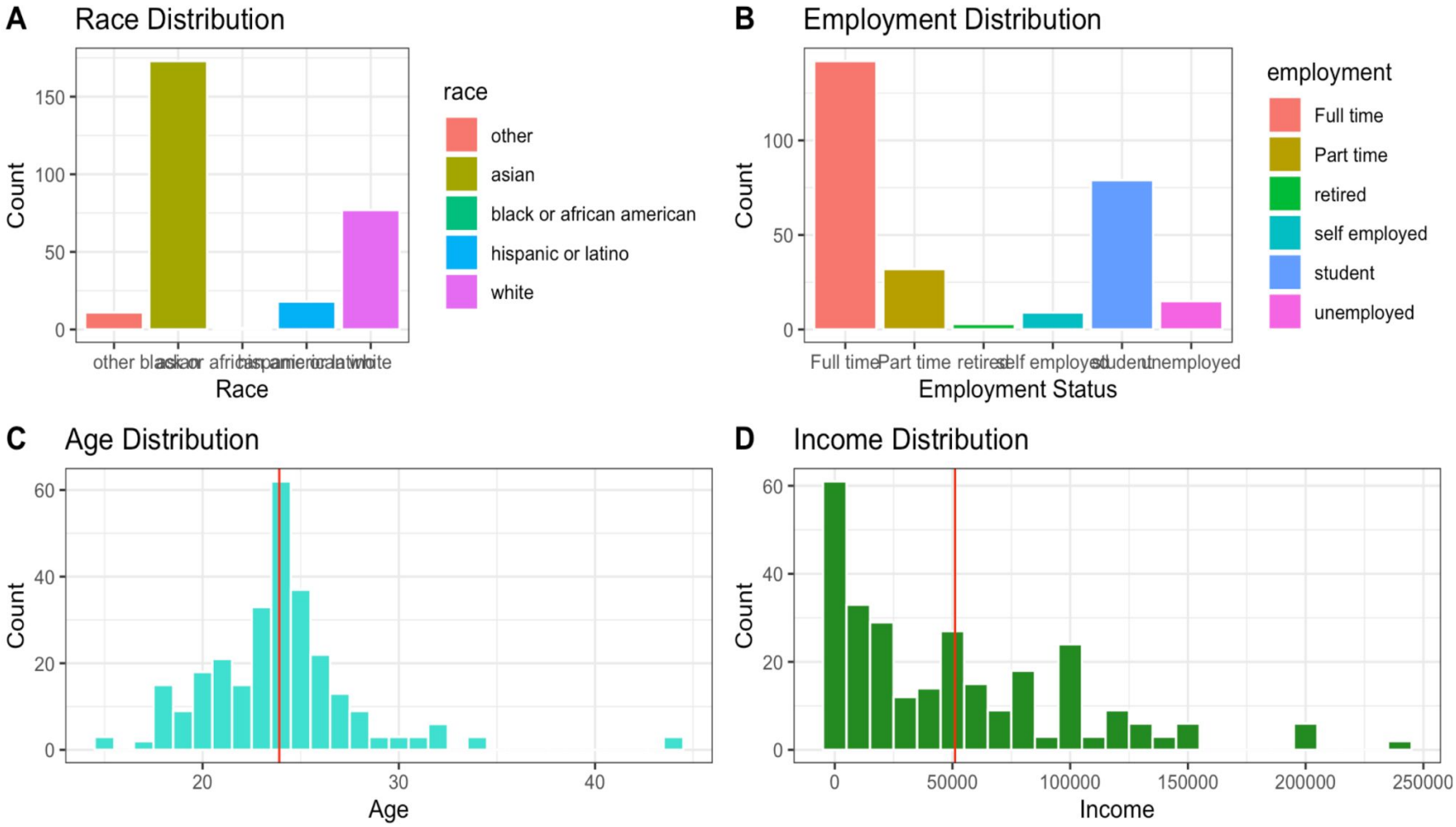
- Mental health diagnosis (yes/no)
- Therapy or counseling (yes/no)
- Currently taking medication for mental health (yes/no)

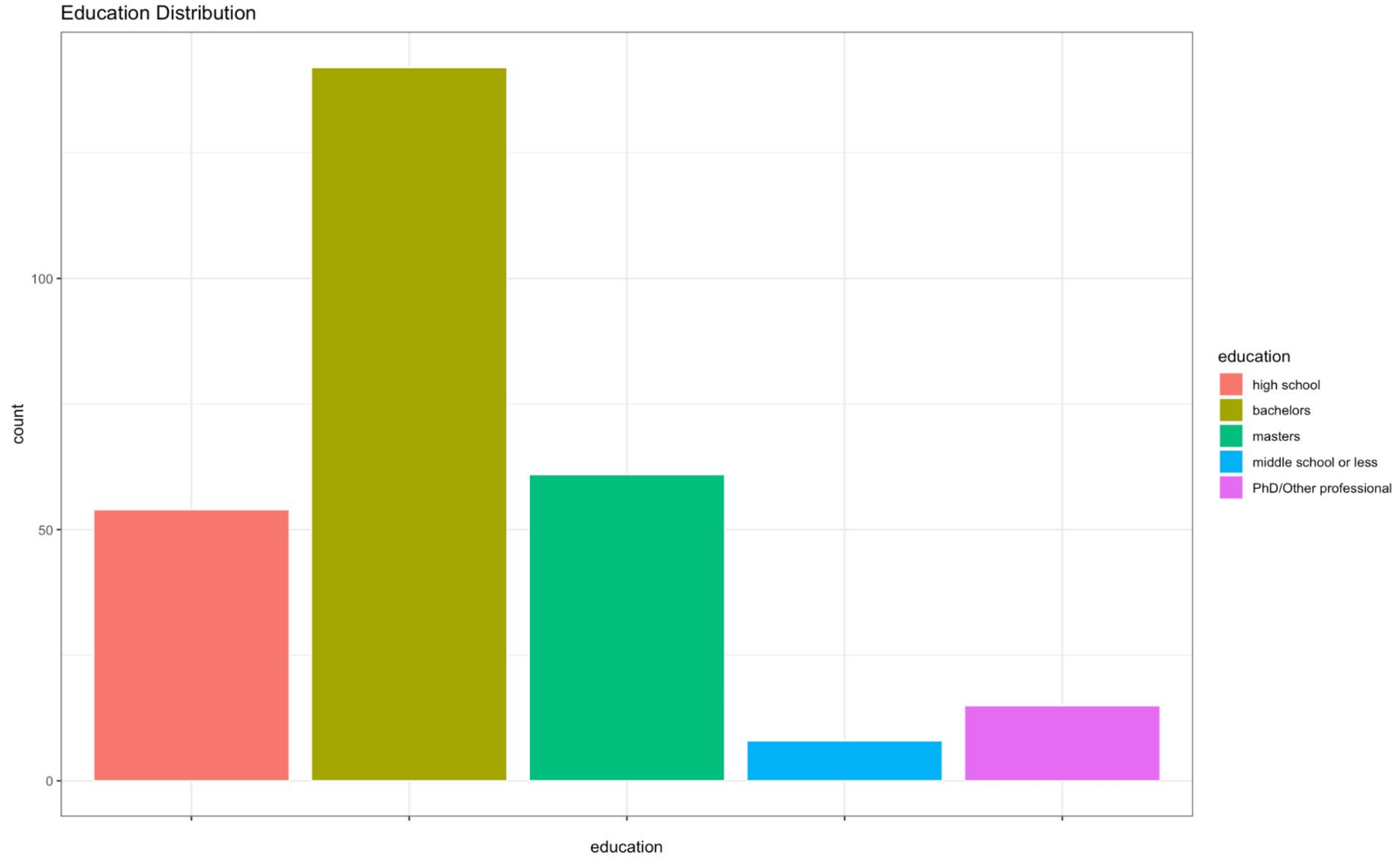
Initial Data Cleaning

- Removed timestamp column
- Removed unneeded commas, dollar signs, and letters from raw survey data
- Out of 310 survey responses, 30 had missing values for one or more variables
- Left with 280 responses after omitting NA’s

Variables

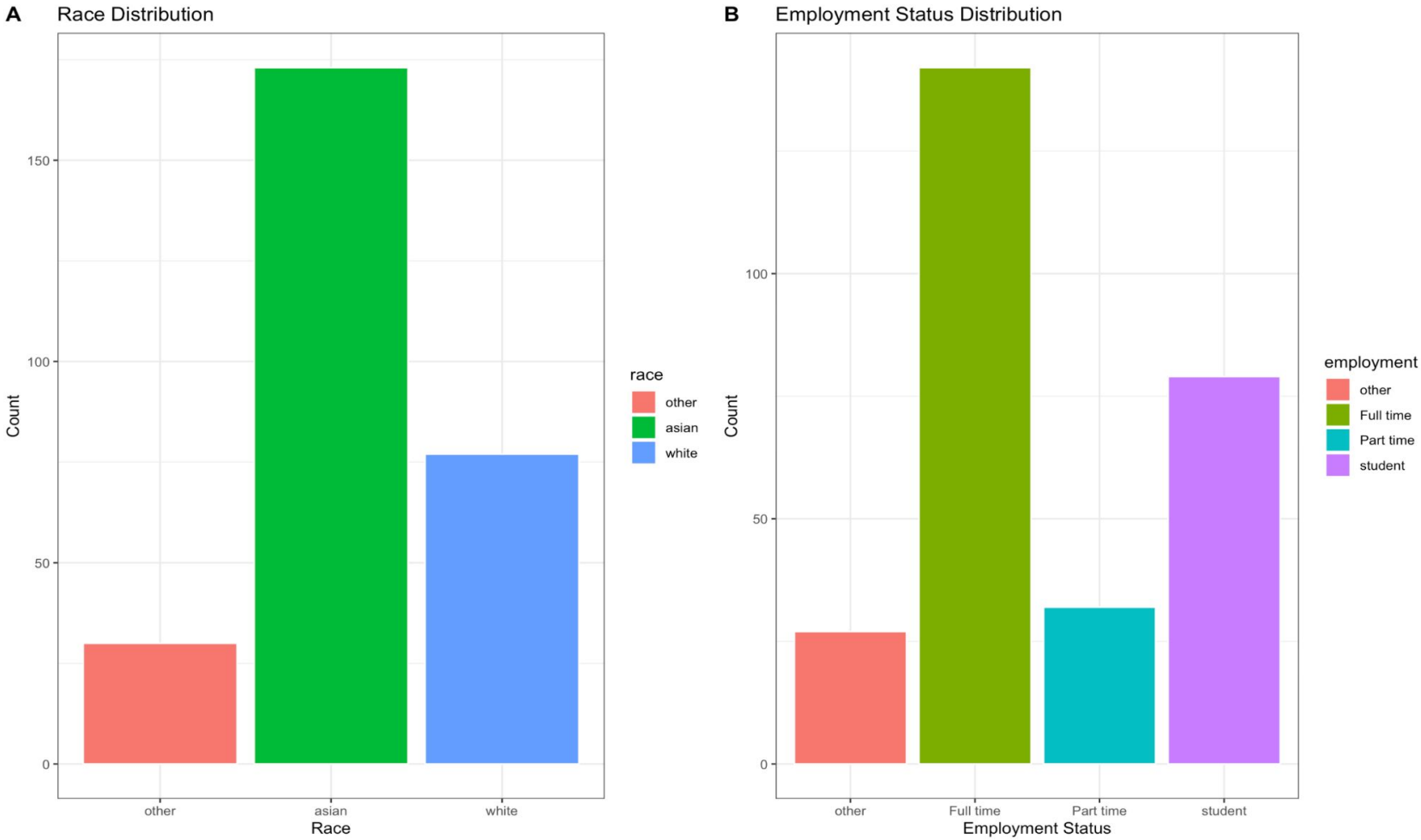
- Creating new binary variables:
  - **stress.bin, anxiety.bin, depression.bin, inadequacy.bin**
  - 1 if >= 5, 0 if otherwise
- Converting yes/no questions to binary variables:
  - 1 for yes, 0 for no
- Total mental health score: Stress + anxiety + depression + feelings of inadequacy
  - **total.score** = stress + anxiety + depression + inadequacy


A
Race Distribution
Count
150
100
50
0
other
asian
black or african american
hispanic or latino
white
Race
race
other
asian
black or african american
hispanic or latino
white
B
Employment Distribution
Count
100
50
0
Full time
Part time
retired
self employed
student
unemployed
Employment Status
employment
Full time
Part time
retired
self employed
student
unemployed
C
Age Distribution
Count
60
40
20
0
20
30
40
Age
D
Income Distribution
Count
60
40
20
0
0
50000
100000
150000
200000
250000
Income

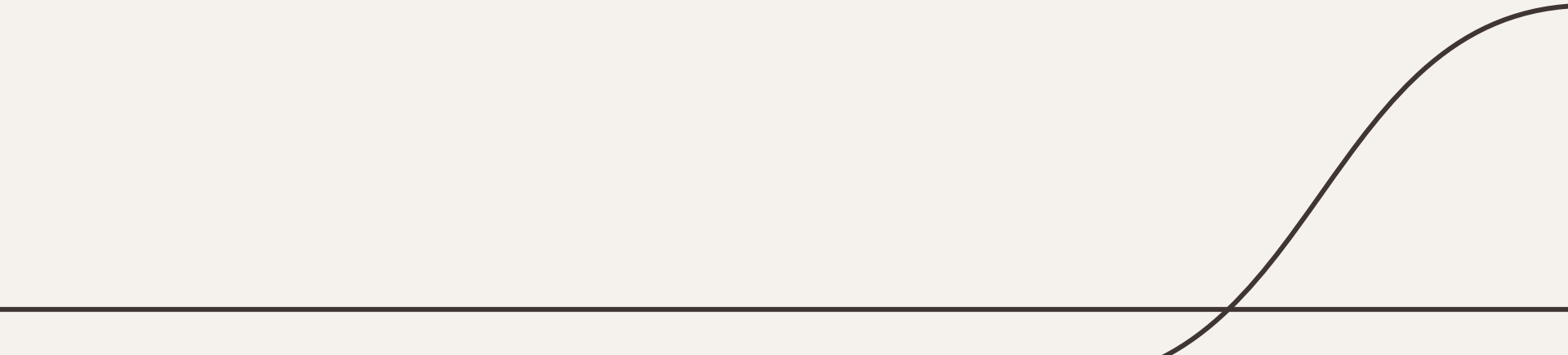

Education Distribution
count
100
50
0
education
education
high school
bachelors
masters
middle school or less
PhD/Other professional

Merging Categories

- Race: White, Asian, Other
- Employment: Full-time, part-time, student, other
- Decided to leave Education as is

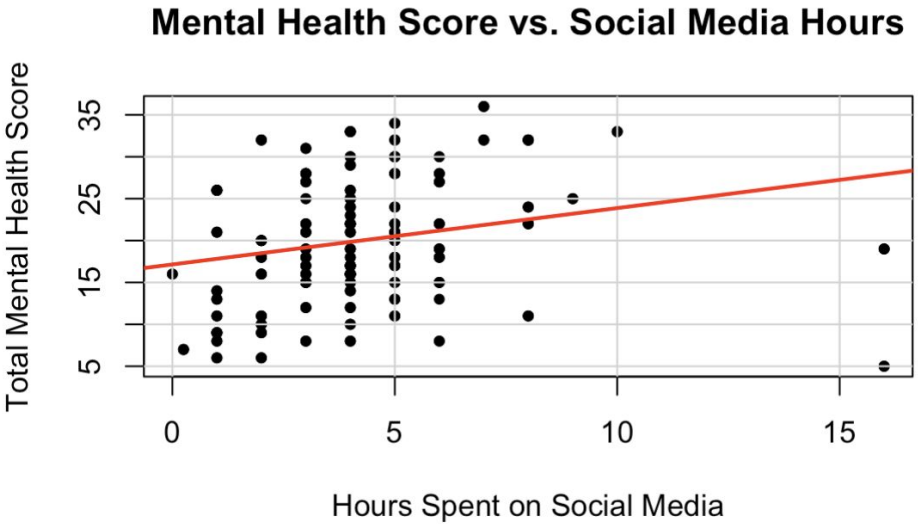

A
Race Distribution
Count
150
100
50
0
other
asian
white
Race
race
other
asian
white
B
Employment Status Distribution
Count
100
50
0
other
Full time
Part time
student
Employment Status
employment
other
Full time
Part time
student

Linear Regression



Initial Analysis

```
Call:
lm(formula = total.score ~ sm_hours, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-22.9143  -5.5124  -0.4937   6.8147  14.1418

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  17.1479     0.8665  19.790  < 2e-16 ***
sm_hours      0.6729     0.1731   3.887 0.000127 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.663 on 278 degrees of freedom
Multiple R-squared:  0.05155,   Adjusted R-squared:  0.04814
```

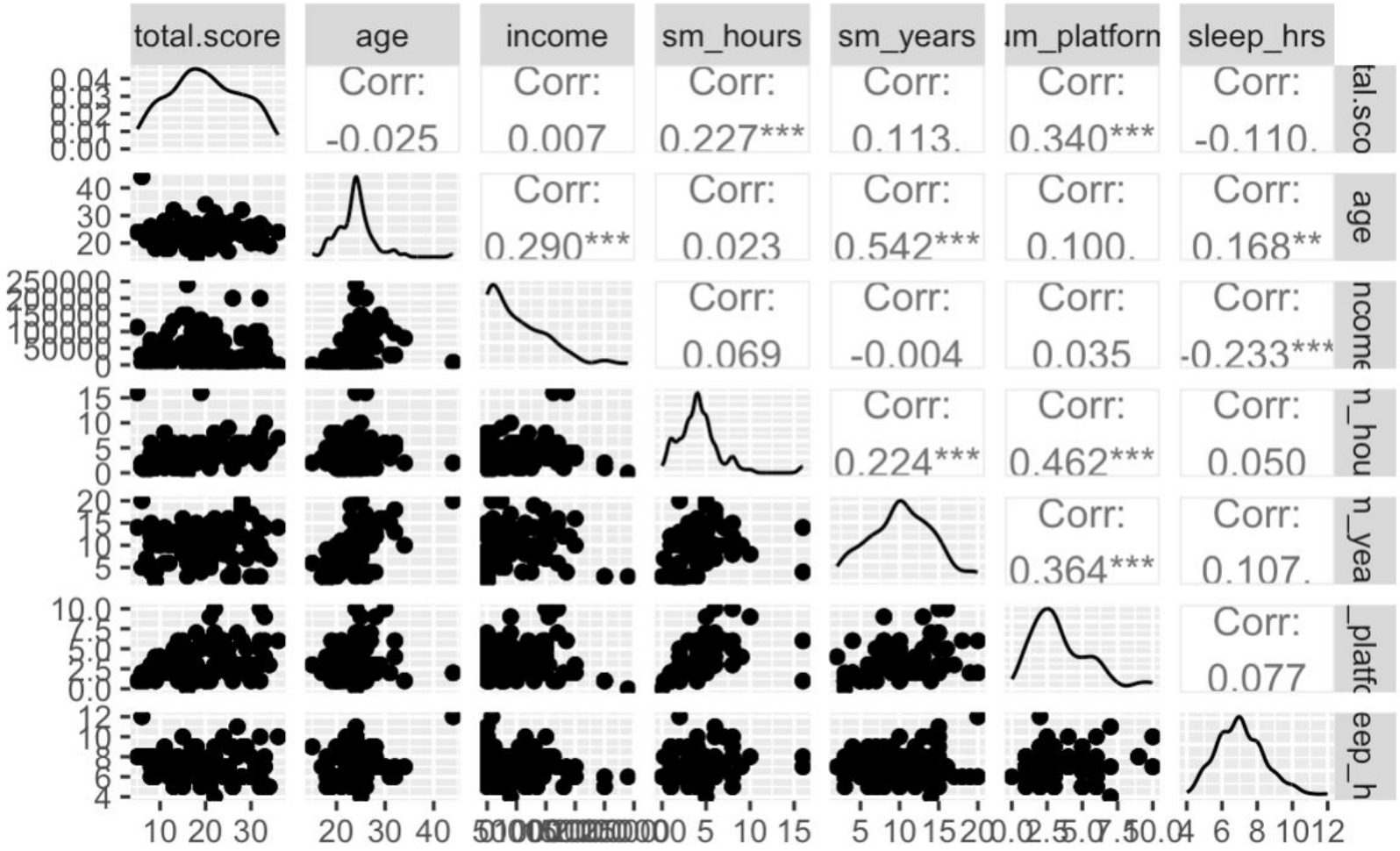


Linear Regression Model Setup

- Outcome: ***total.score*** (total mental health score)
- Main Variable: ***sm_hours*** (estimated hours spent on social media daily)
- Included Variables: age, income, sm_years, num_platforms, sleep_hrs
- Include all variables and perform backwards selection by AIC

Initial Model:

lm(total.score ~ sm_hours + age + income + sm_years + num_platforms + sleep_hrs)

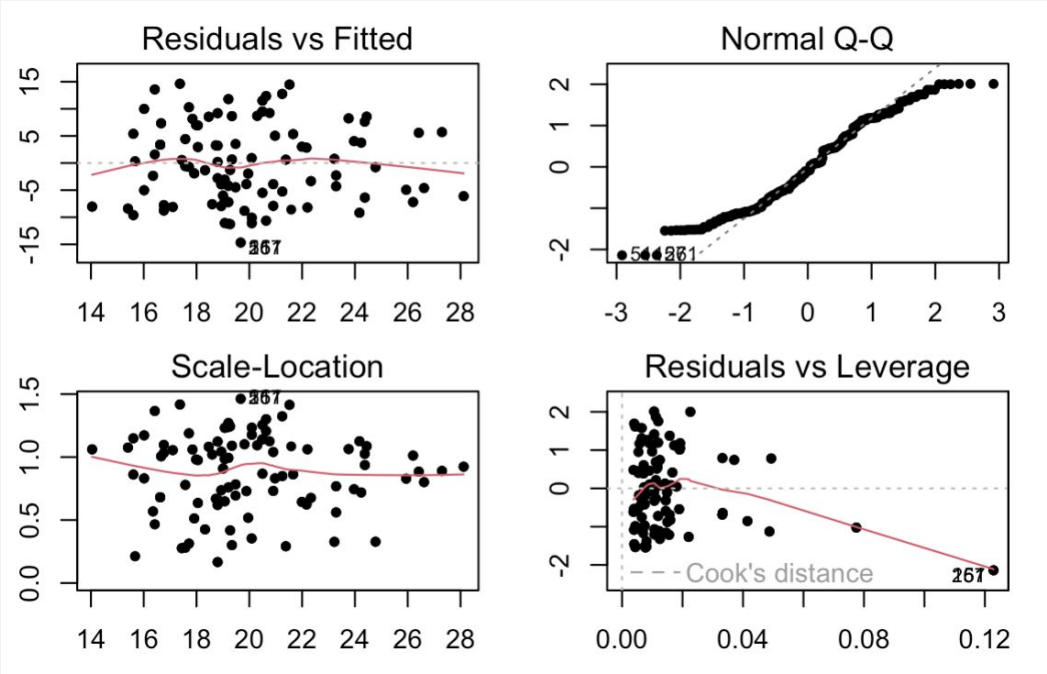

total.score
age
income
sm_hours
sm_years
ım_platform
sleep_hrs
Corr:
-0.025
Corr:
0.007
Corr:
0.227***
Corr:
0.113.
Corr:
0.340***
Corr:
-0.110.
Corr:
0.290***
Corr:
0.023
Corr:
0.542***
Corr:
0.100.
Corr:
0.168**
Corr:
0.069
Corr:
-0.004
Corr:
0.035
Corr:
-0.233***
Corr:
0.224***
Corr:
0.462***
Corr:
0.050
Corr:
0.364***
Corr:
0.107.
Corr:
0.077
tal.sco
age
ncome
n_hou
n_yea
_platfc
eep_h

After Backwards Selection

lm(total.score ~ sm_hours+num_platforms+sleep_hrs)

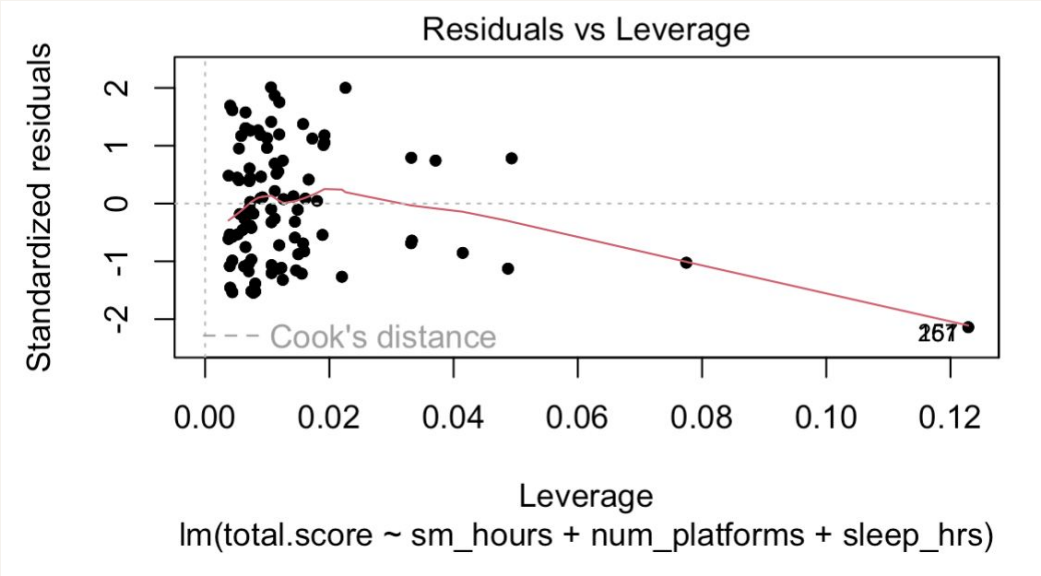
```
Residuals:
     Min       1Q   Median       3Q      Max
-14.6717  -6.1175  -0.7135   5.6120  14.6292

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   20.1554     2.2902   8.801  < 2e-16 ***
sm_hours       0.2711     0.1865   1.454    0.147
num_platforms  1.1566     0.2363   4.894 1.68e-06 ***
sleep_hrs     -0.7472     0.3020  -2.474    0.014 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.32 on 276 degrees of freedom
Multiple R-squared:  0.1409,    Adjusted R-squared:  0.1316
F-statistic: 15.09 on 3 and 276 DF,  p-value: 4.015e-09
```

Model Diagnostics



- Linearity Assumption
- Normality Assumption does not hold too well
- Homoscedasticity
  - Non-constant variance test results suggest constant error variance
- Presence of influential observations

Outliers and Influential Points



- High leverage points: 261 and 157
- Outlier: 51
- Points left as is

Coefficient Interpretation

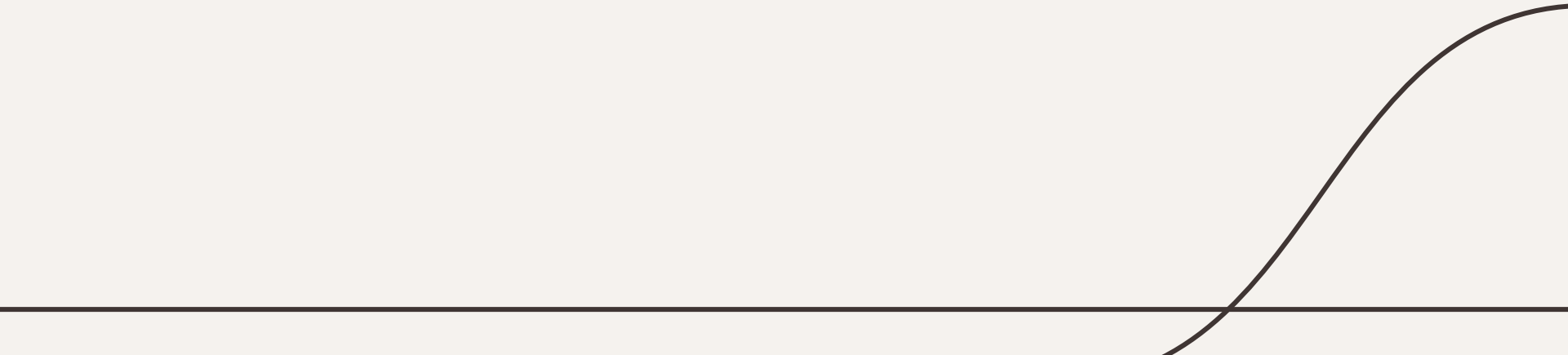
- **sm_hours:**
  - For every 1 hour increase in time spent on social media daily, a person's total mental health score increases by an estimate of 0.27 (95% CI: -0.1, 0.63)
  - 95% CI contains 0 and associated p-value is not significant at 5% level

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	20.1554208	2.2901511	8.800913	1.515284e-16
sm_hours	0.2710963	0.1864561	1.453942	1.470986e-01
num_platforms	1.1566089	0.2363310	4.894021	1.682396e-06
sleep_hrs	-0.7472305	0.3020310	-2.474020	1.396191e-02
	2.5 %	97.5 %		
(Intercept)	15.64703768	24.6638039		
sm_hours	-0.09596046	0.6381530		
num_platforms	0.69136856	1.6218492		
sleep_hrs	-1.34180751	-0.1526535		

Discussion

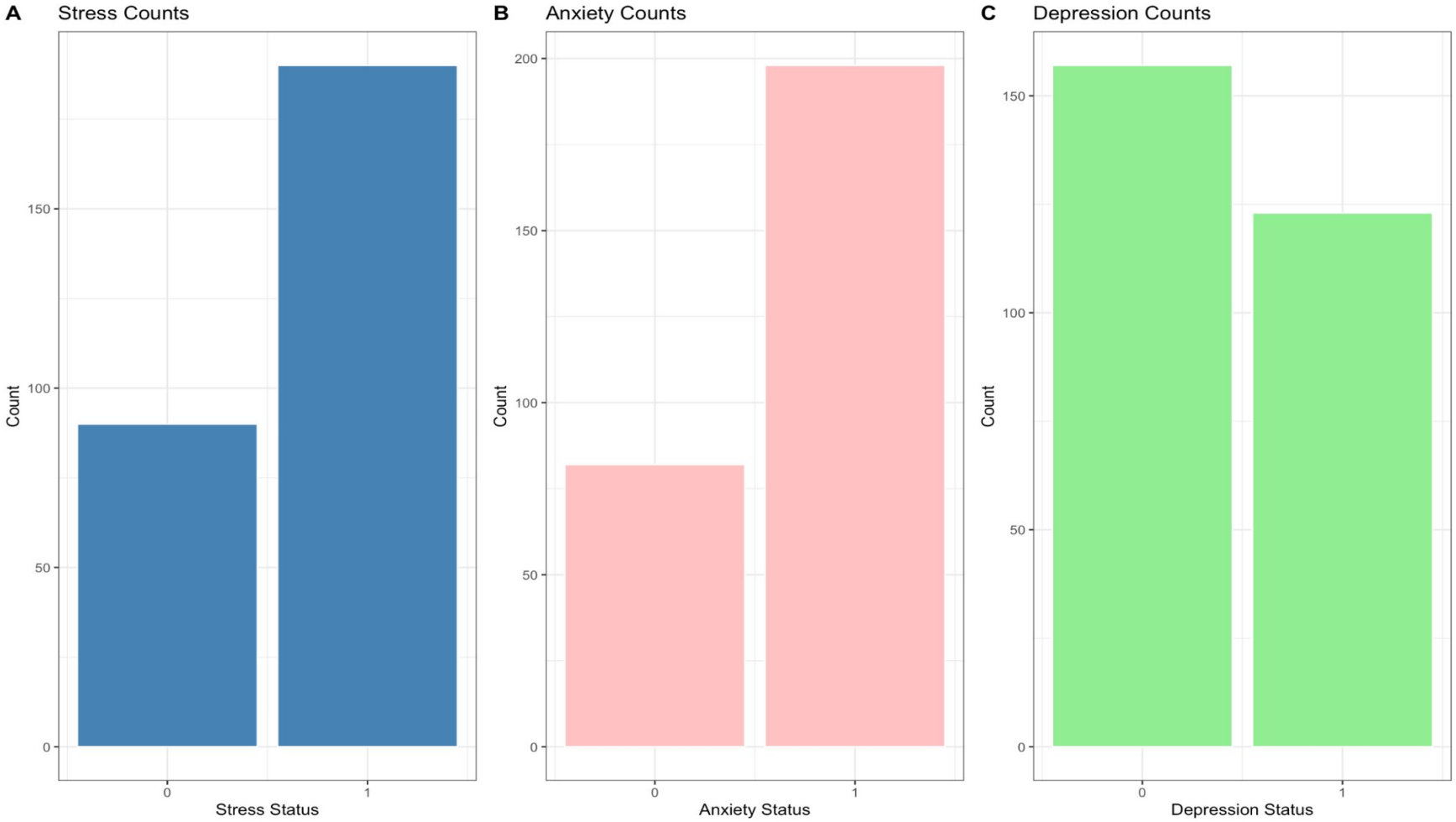
- Linear model provides some preliminary evidence to support a relationship between increased social media usage and worse mental health
- Many factors unaccounted for
- ~10% missing values

Logistic Regression



Initial Model Setup

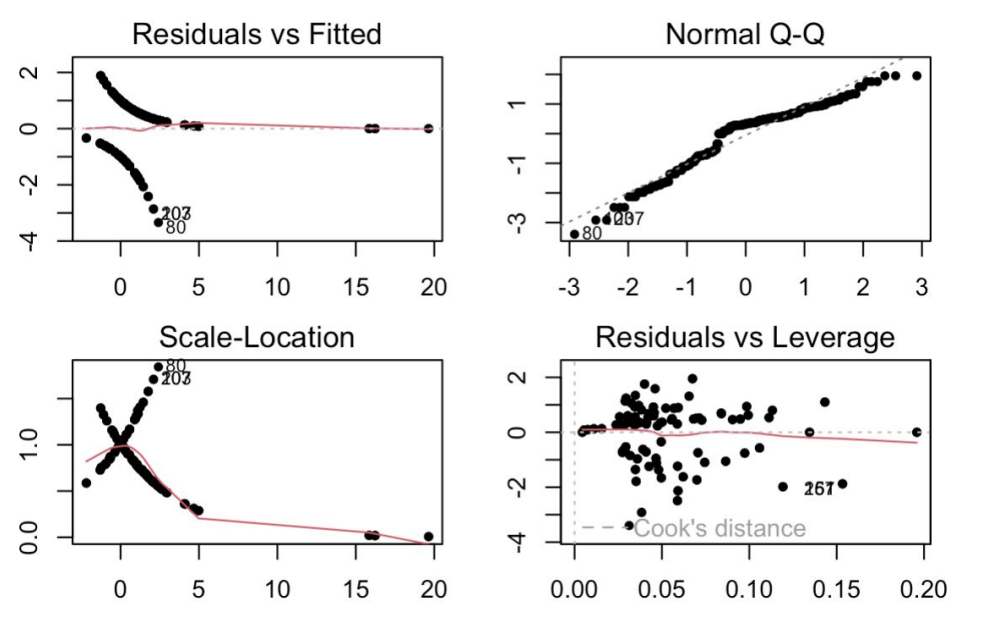
- Model 1: glm(**stress.bin** ~ **sm_hours** + age + race + gender + income + education + employment + num_platforms + sm_years + sm_work + therapy + cyberbullying, family=binomial(link="logit"), data=data)
- Model 2: glm(**anxiety.bin** ~ **sm_hours** +age + race + gender + income + education + employment + num_platforms + sm_years + sm_work + therapy + cyberbullying, family=binomial(link="logit"), data=data)
- Model 3: glm(**depression.bin** ~ **sm_hours** +age + race + gender + income + education + employment + num_platforms + sm_years + sm_work + therapy + cyberbullying, family=binomial(link="logit"), data=data)


A
Stress Counts
Count
150
100
50
0
0
1
Stress Status
B
Anxiety Counts
Count
200
150
100
50
0
0
1
Anxiety Status
C
Depression Counts
Count
150
100
50
0
0
1
Depression Status

Backwards Selection by AIC

- Initially exclude sm_hours, then add back
- New Model: glm(stress.bin ~ sm_hours + race + income + education + employment + num_platforms+cyberbullying, family=binomial(), data=data)
- Interpretation:
  - For every 1 hour increase in time spent on social media daily, odds of experiencing stress increase by an estimated multiplicative factor of 1.06 (95% CI: 0.93, 1.23)
  - 95% CI contains 1; associated p-value is not significant at 5% level
  - Certain education and employment statuses had very large estimates

Model Diagnostics




```
        Hosmer and Lemeshow test (binary model)

data:  data$stress.bin, mod.stress$fitted.values
X-squared = 4.0394, df = 8, p-value = 0.8536
```

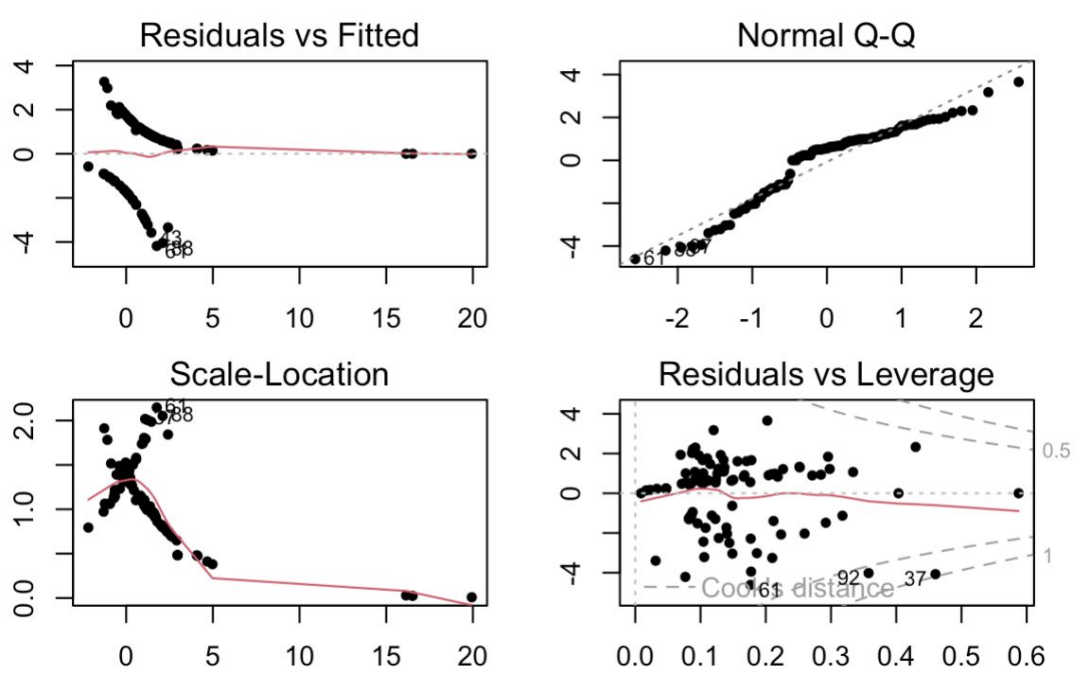
- Diagnostic plots suggest model is not a good fit
- Hosmer Lemeshow test: cannot reject null, suggests model is a good fit
- Overall conclusion: model is not a good fit

Data Aggregation

Aggregated model after backwards selection:

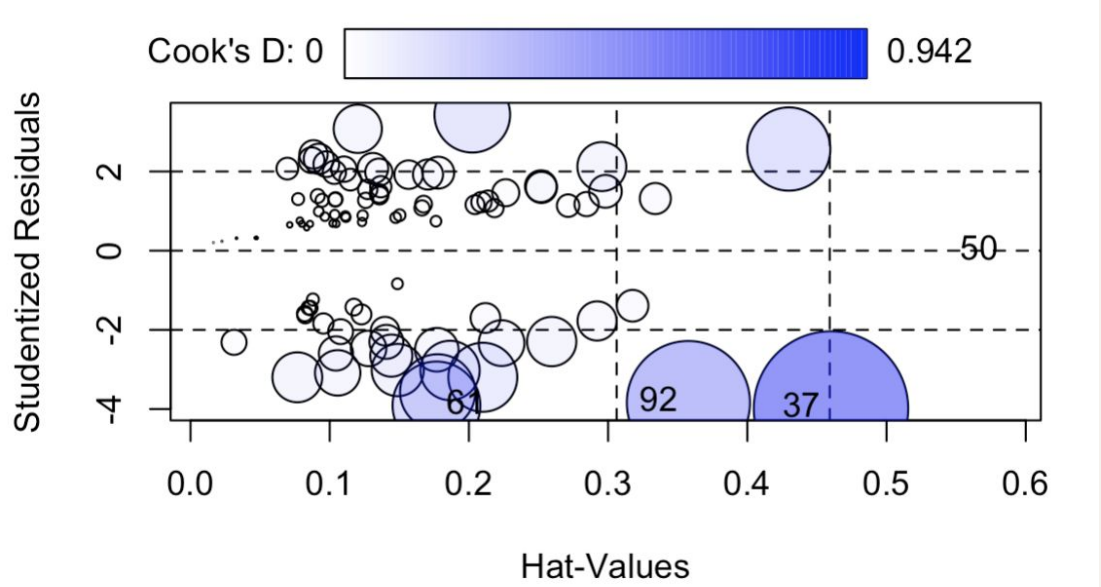
glm(stress.bin/tot ~ sm_hours + race + income + education + employment + num_platforms + cyberbullying, family = binomial(link = "logit"), data = agg_data, weights = tot)

Model Diagnostics



- Diagnostic plots suggest aggregated model is not a good fit
- Deviance GOF test yields p-value of 1.13e-22
  - Suggests aggregated model is NOT a good fit

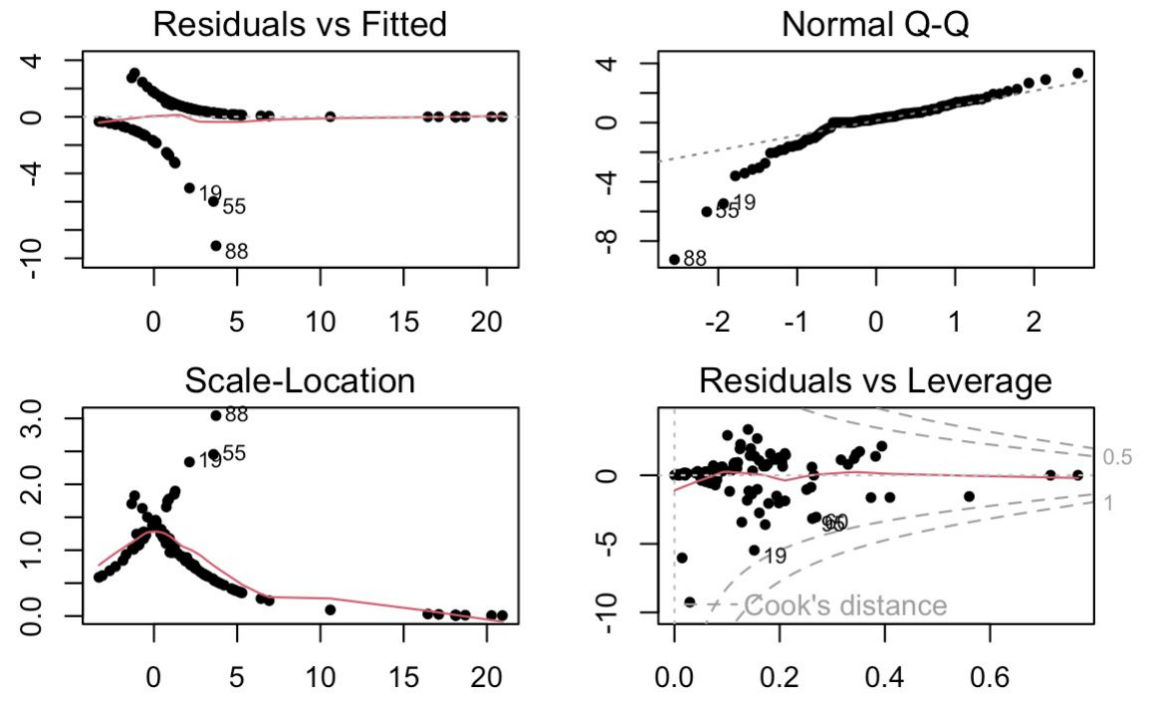
Outliers and Influential Points



	StudRes <dbl>	Hat <dbl>	CookD <dbl>
37	-3.9900258856	0.4600875	9.417926e-01
50	0.0009993508	0.5876731	6.718983e-08
61	-3.9095642164	0.1766712	3.040181e-01
92	-3.8396707698	0.3576578	5.994204e-01

- Removed points 37, 50, 61, 92 and re-fit aggregated model to new dataset

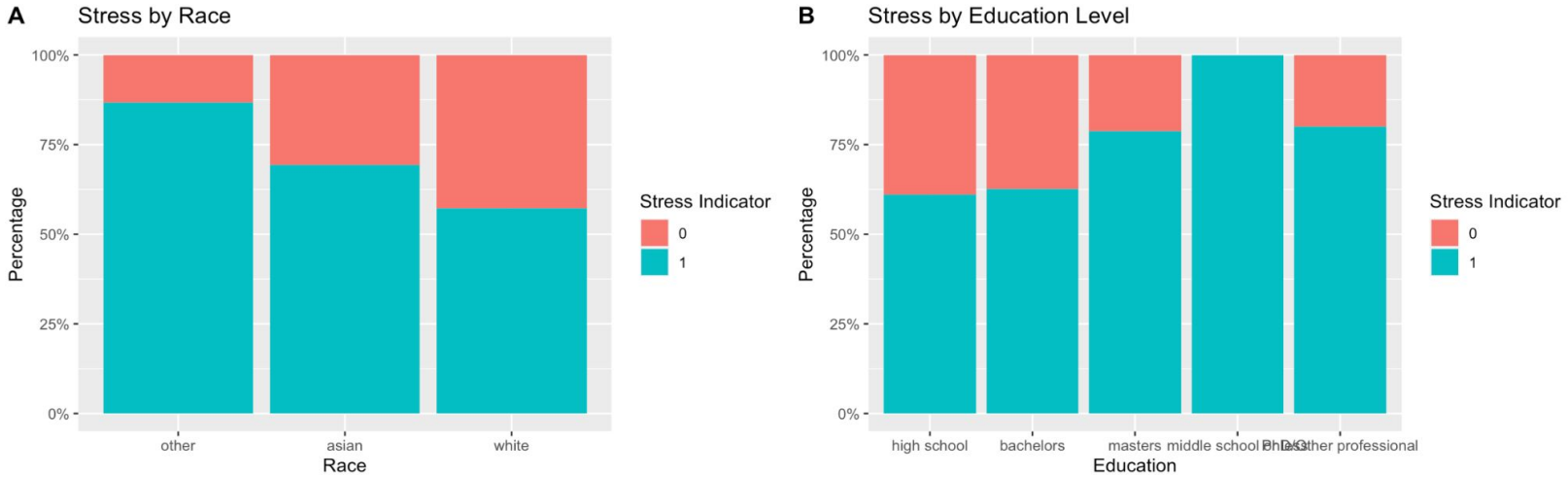
Model Diagnostics



- Diagnostic plots suggest aggregated model is not a good fit
- Deviance GOF test yields p-value of 1.33e-12
  - Less significant than before, but still suggests aggregated model is not a good fit

Aggregated Model Interpretation

- For every 1 hour increase in time spent on social media daily, odds of experiencing stress increase by an estimated multiplicative factor of **1.97 (95% CI: 1.56, 2.54)**
- Other things to note:
  - Race: Asian and White were associated with decreased odds of stress compared to Other.
  - Education: All categories were associated with increased odds of stress compared to High School.
  - Employment: Being employed full-time or part-time were associated with decreased odds of stress, while being a student was associated with increased odds of stress.


A
Stress by Race
100%
75%
50%
25%
0%
Percentage
other
asian
white
Race
Stress Indicator
0
1
B
Stress by Education Level
high school
bachelors
masters
middle school
professional
Education
C
Stress by Employment
other
Full time
Part time
student
Employment

Discussion

- Generalized linear models provided some preliminary evidence to support a relationship between increased social media usage and certain mental health conditions
- Extreme estimates, wide confidence intervals, and NA’s caused by low counts in some categories
- None of the models appeared to fit the data well - additional complexity needed

Overall Conclusions

- Both models provided some preliminary evidence to suggest a relationship between increased social media usage and worse mental health
- Limitations:
  - Accuracy of self-reported survey data
  - Rating scale for mental health conditions is subjective
  - Many factors unaccounted for
  - Missing values
- Further study with more data or a more complex model is needed to reach a firm conclusion