Predicting Covid-19 Preventative Behavior

Section 5
Ella Morrow

Introduction

Introduction to Topic

Ever since COVID-19 began to impact lives all around the world earlier this year, people have reacted in very different ways. Looking at the United States specifically it is clear that people around the country have had extremely varied reactions, and some issues such as wearing a mask have become quickly polarized and politicized. What specific factors could be leading to such differing attitudes? Age in particular has been called out multiple times, younger generations especially receiving large amounts of media coverage for “displaying careless behavior” and “disregarding the pandemic.” But how much of a difference does something like age actually play? And what other factors can we use to successfully predict someone’s behaviors in relation to COVID-19?

Research Questions

What is the relationship between someone’s age and gender and their preventative behaviors towards COVID-19? Is there even a strong enough relationship between the variables that would let us confidently predict a preventative score from just age and gender? How might this relationship have changed overtime as the pandemic became more/less serious? These are all questions I would like to explore in the following report. First we will look into the relationship between age in years and binary gender and a quantitative preventative score value that represents how often people display preventative behavior such as handwashing and mask wearing, while also taking into account the month the survey was completed and how scared participants feel about COVID-19. Then we will continue to use these explanatory variables to look at how well we can predict whether someone has a “high” preventative score.

Introduction to Data

The data set includes survey responses related to various covid related behaviors of 14,031 adults living in the United States. There are 165 total variables, some of which cover basic demographic information like employment status and household size, but most of the variables deal with behaviors related to covid like how often they wash their hands or how many people they come into contact with on a regular basis. This data was collected from March 31st 2020 up until about mid-August 2020 by YouGov and the Institute of Global Health Innovation at Imperial College London. YouGov used active sampling, which involves randomly sampling individuals from a panel of loyal survey respondents, and only those specifically selected by YouGov were able to participate in the survey. The data was collected in order to show how people across the United States are responding to

the global pandemic, and because it is publicly available public health officials and academic organizations will be able to analyze it and learn more about people's reactions to covid.

Multiple Linear Regression Modeling

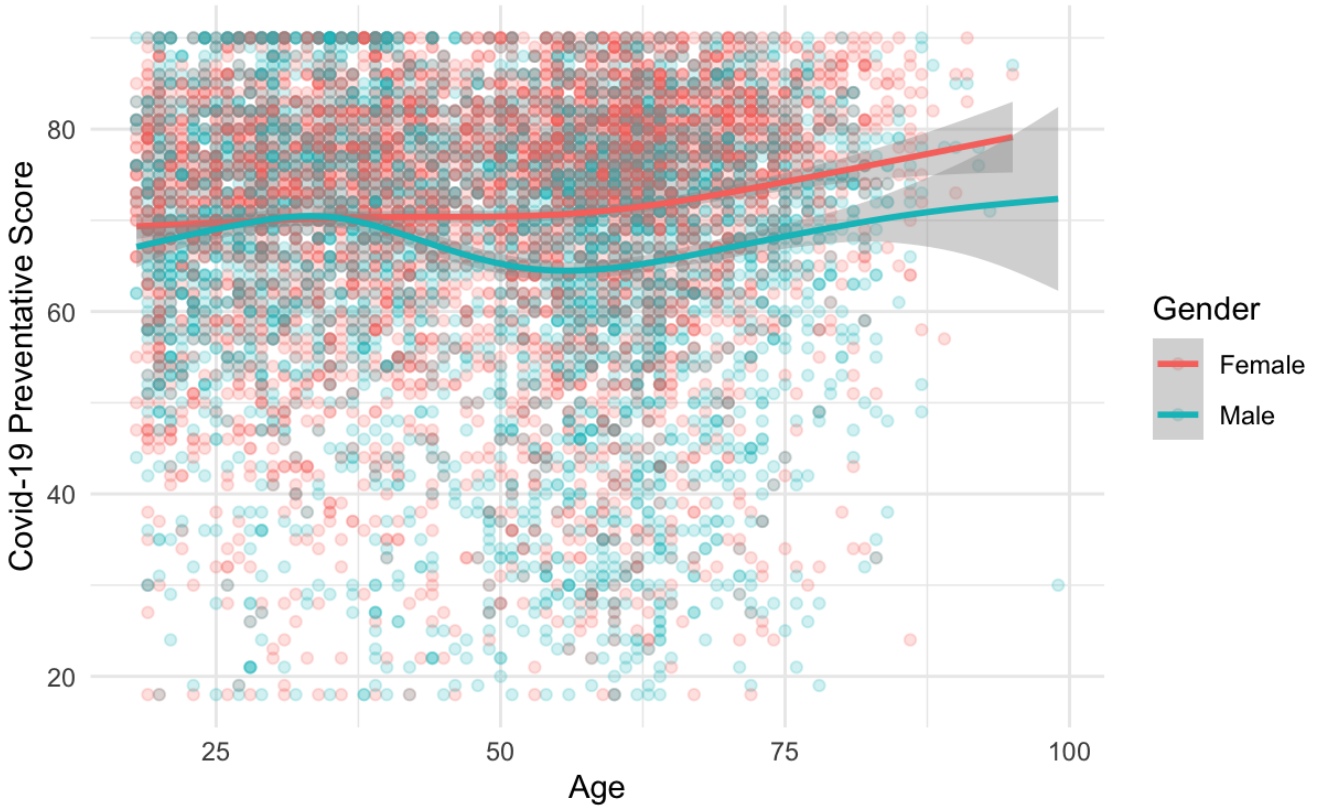
Variable Descriptions

This model predicts a COVID-19 preventative score which was compiled by adding together a range of variables that measured the frequency of precautionary and preventive behaviors like hand washing and mask wearing. The original variables each had five categories ranging from Never to Always which I quantified to represent 1(Never) to 5(Always). The scores range from 18, little to no preventive behavior displayed, to 90, regular display of a high number of preventive behaviors. The explanatory variables used in this model to predict the score are age, gender, month survey was completed, and level of fear towards COVID-19. Age is in years and ranges from 18 to 99. Gender is binary with categories of male or female. There are five month categories that cover every month the study was conducted ranging from April 2020 to August 2020. The last variable measures how scared the participant is about contracting COVID-19 and has four categories: not at all scared, not very scared, fairly scared, and very scared. Another variable that was used in model selection but did not end up in the final model is the region of the US participants live in with the categories being North East, South, North Central, and West.

Sample Description

The sample I will be using is of individuals living in the United States, however this sample does not include the whole dataset because I had to filter the participants who did not answer the question of how scared they are of contracting COVID-19. Among those who did have data for this variable, I filtered out those who answered "Don't know" and those who are "Not applicable/have already contracted COVID-19," leaving me with a sample size of 9102 cases out of the total 14031 in the data set. So keep in mind that only those who answered whether they were scared of covid and have not yet contracted it are included in my subset.

Visualization



Looking at this scatter plot we can see that the relationship between age and preventative score is not very strong or linear. The spread for younger participants is large but more heavily concentrated in having scores above 60, then there is sort of a dip in concentration of preventive score around ages 50 to 75 where we see a higher amount of scores below 50. The spread then seems to narrow back down towards a higher preventative score for older participants above 75 so it looks like what little correlation there is between age and score is positive. Looking at gender's effect on this it seems like there is a fairly even distribution of gender for both age and score. We can also see that, on average, there is a mostly positive linear relationship between age and score for females. For males the relationship is positive at first, then negative, then evens out again and from age 50 on the relationship between score and age for males seems to mirror that of females but with a lower average preventative score. There are also one or two data points that could be considered outliers in the bottom right that represent much older people who have lower preventative scores.

Model Selection

To select the best linear regression model I fitted a variety of models that predicted preventative score with different variations of age, gender, month, how scared, as well as region of the United States. I then looked at each of these model's adjusted R-squared and standard deviation of residuals values to narrow my selection down. I found that models that incorporated how scared someone feels had the highest R-squared values, meaning that these model's could account for higher percentages of the data. I also saw that including the region of the United States where a person lives does not

improve the model as much as month of survey and level of fear towards covid-19 do. After checking each model, I saw that the models with the highest adjusted R-squared values and lowest standard deviation of residuals were ones that, in addition to age, incorporated gender, month, and level of fear. Two of the models I fitted had these explanatory variables, the only difference was one also had an interaction term between age and gender. So, to find whether the model with or without interaction was best I used hypothesis testing and compared the models using a nested hypothesis test. My null hypothesis for this test was that the smaller model without interaction was the correct model. I found that if my null hypothesis was in fact true and the true slope of the interaction term was zero, there was a probability of 1.41e-8 of getting slopes as large or larger than what was observed in the model with interaction. This low p-value gave me enough evidence to reject the null hypothesis that the smaller model without interaction is correct in favor of keeping the interaction term.

Final Model Statement

$$E[preventative\ score \mid age * gender + month + scared] = \beta_0 + \beta_1 age + \beta_2 genderMale + \beta_3 month5 + \beta_4$$
$$+ \beta_5 month7 + \beta_6 month8 + \beta_7 scaredNotVery + \beta_8 scaredFairly + \beta_9 scaredVery + \beta_{10} age * genderMale$$

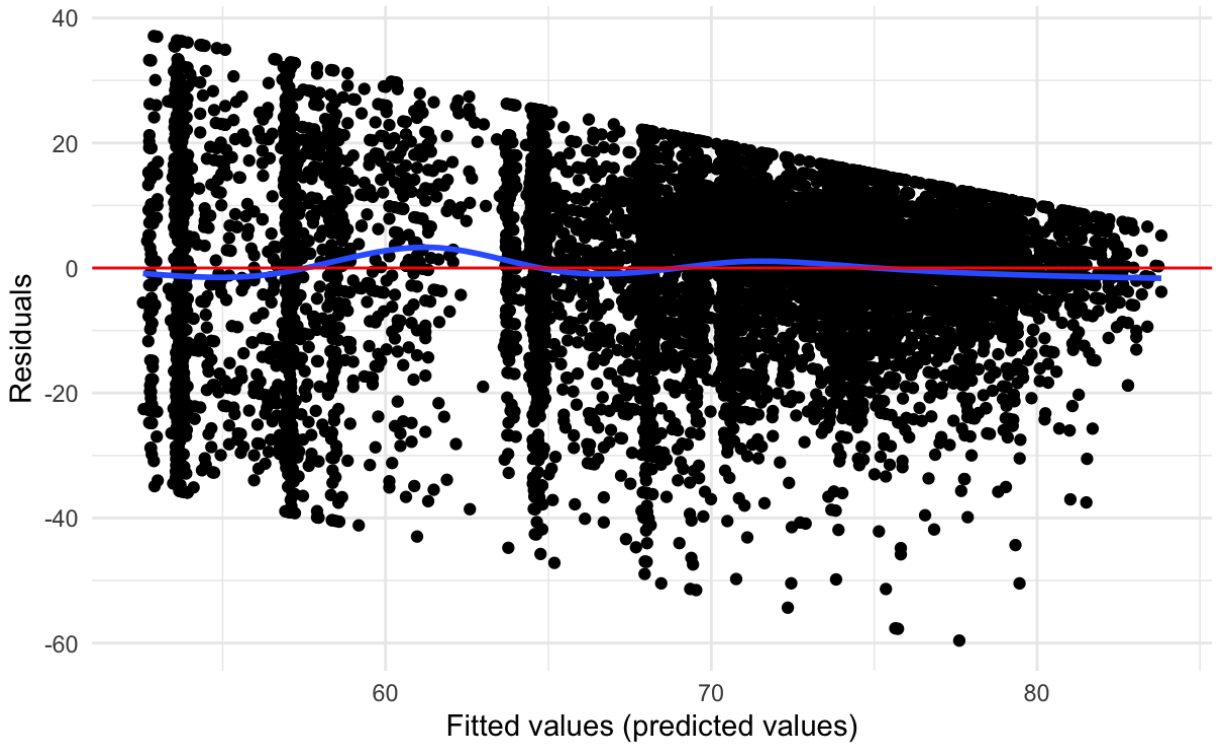
Fitted Model

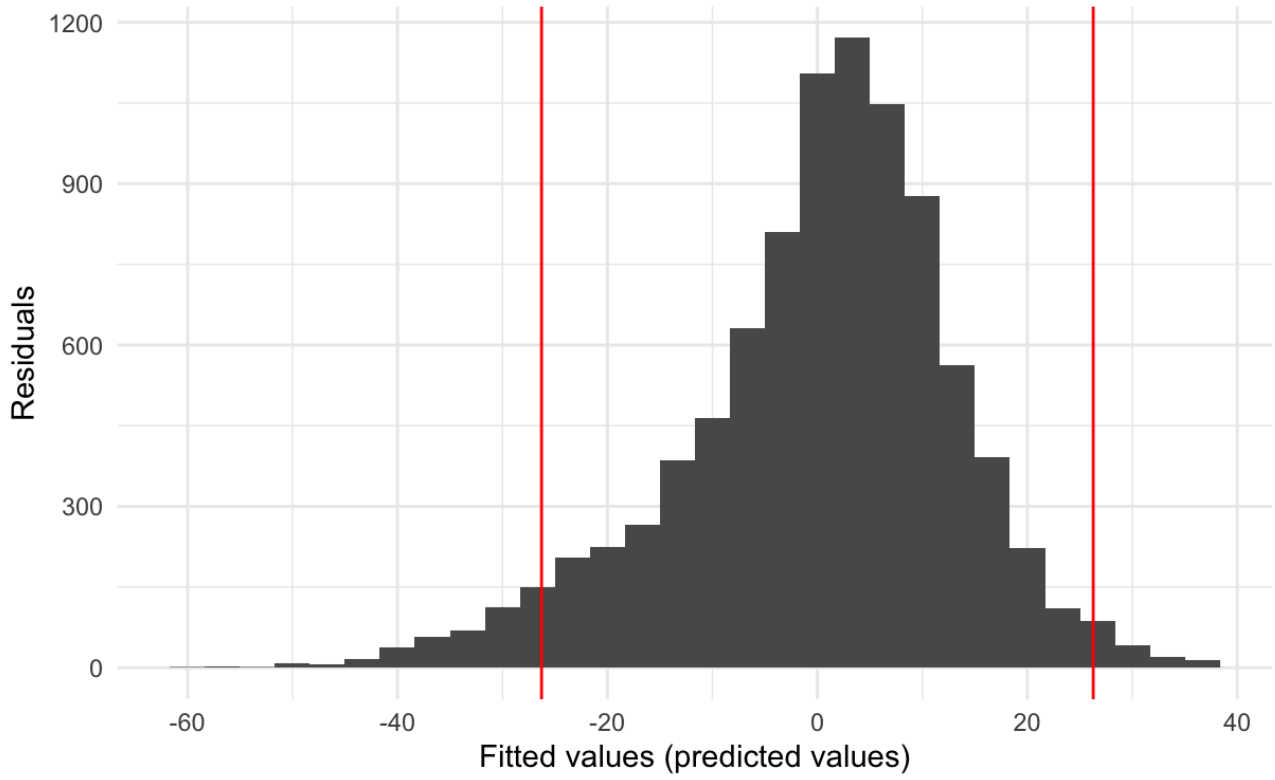
Slope Estimates	Estimates	Standard Errors	95% Confidence Intervals
Intercept	56.59	0.75	(55.12 , 58.072)
age	0.085	0.010	(0.063 , 0.10)
genderMale	2.11	0.83	(0.48 , 3.74)
month5	-1.34	0.47	(-2.27 , -0.41)
month6	-4.79	0.52	(-5.81 , -3.76)
month7	-4.67	0.51	(-5.68 , -3.66)
month8	-5.57	0.58	(-6.71 , -4.43)
scaredNotVery	10.98	0.40	(10.18 , 11.78)
scaredFairly	16.80	0.40	(16.022 , 17.59)
scaredVery	20.73	0.46	(19.83 , 21.63)
age:genderMale	-0.090	0.016	(-0.12 , -0.059)

Interpretations

Taking into account the month the survey was taken and how scared the participant feels about contracting covid-19, this model estimates that the average change in COVID-19 preventative score with every additional year of age for females living in the United States is 0.085. Looking at the confidence interval, we can see that 95% of the data we have has a slope estimate that falls into the interval (0.06, 0.10). And because 95% of samples from the greater target population are expected to generate confidence intervals that contain the true population parameter value, it is likely that there is a true relationship between age and preventative score for females because 0 is not in the interval. This likelihood of a relationship is also strengthened by having a percentage of 4.71e-15 that, should there be no true relationship between age and preventative score for females, we would get a slope equal to or greater than our estimate. So, this model estimates that there may be a very small positive relationship between age and increased preventative COVID-19 behavior for females however we cannot say for certain. Now looking at males, and continuing to hold month of survey and level of fear constant, this model estimates a difference of -0.09 between the average change in preventative score for females and the average change in preventative score for males for every additional year of age someone has. Considering the standard deviation of this estimate, a range of plausible values for this difference slope for males and females could also be (-0.12, -0.05). And because there is a low probability of getting an estimate equal to or greater than this if there was no difference in the relationship between age and preventative score between genders and 0 is not present in the interval of plausible values, it is likely that the relationship is in fact different for males and females. So, it is likely that, on average, males have an even smaller relationship between age and preventative behavior that could even be negative. This means it is also possible that there is no true relationship between age and preventative score for males.

Model Evaluation





Looking at the scatter plot of residuals, based on the line of best fit it seems like the relationship between the residuals and the fitted values for this model is fairly straight, but the spread is not quite equal as it starts off wider with lower scores then with higher scores we see more precision in the predictions. At the spread's widest point there are cases being predicted easily as much as 35 points off from their true value, and even for higher scores when the model gets more accurate it would not be hard to find residuals greater than 10. There are also a few outliers in the bottom right, so although predictions narrow in as the score increases, there are still a handful of cases getting over predicted. Moving on, the model has an R-squared value of 0.24 which means that age, gender, month of survey, and how scared someone is can explain about a quarter of the total variation of COVID-19 preventative scores. Now, looking at a histogram of residuals we can see that it is unimodal and mostly symmetric which means that about 95% of the residuals are within 2 standard deviations of 0. Knowing this, we can now interpret the residual standard error for the model, which is 13.14. So, the model can successfully predict preventative score for 95% of the total data within about 26 points. Based on all of this, this model is certainly not a "great" model and I would not place a huge amount of confidence in its ability to accurately predict preventative score as it can easily be as far as 26 points off. That being said, when considering the model can explain about 25% of variation of preventative score and the score is based on human behavior which is something that is not at all linear and often unpredictable, I do think the model does a decent job considering the context of the outcome variable.

Multiple Logistic Regression Modeling

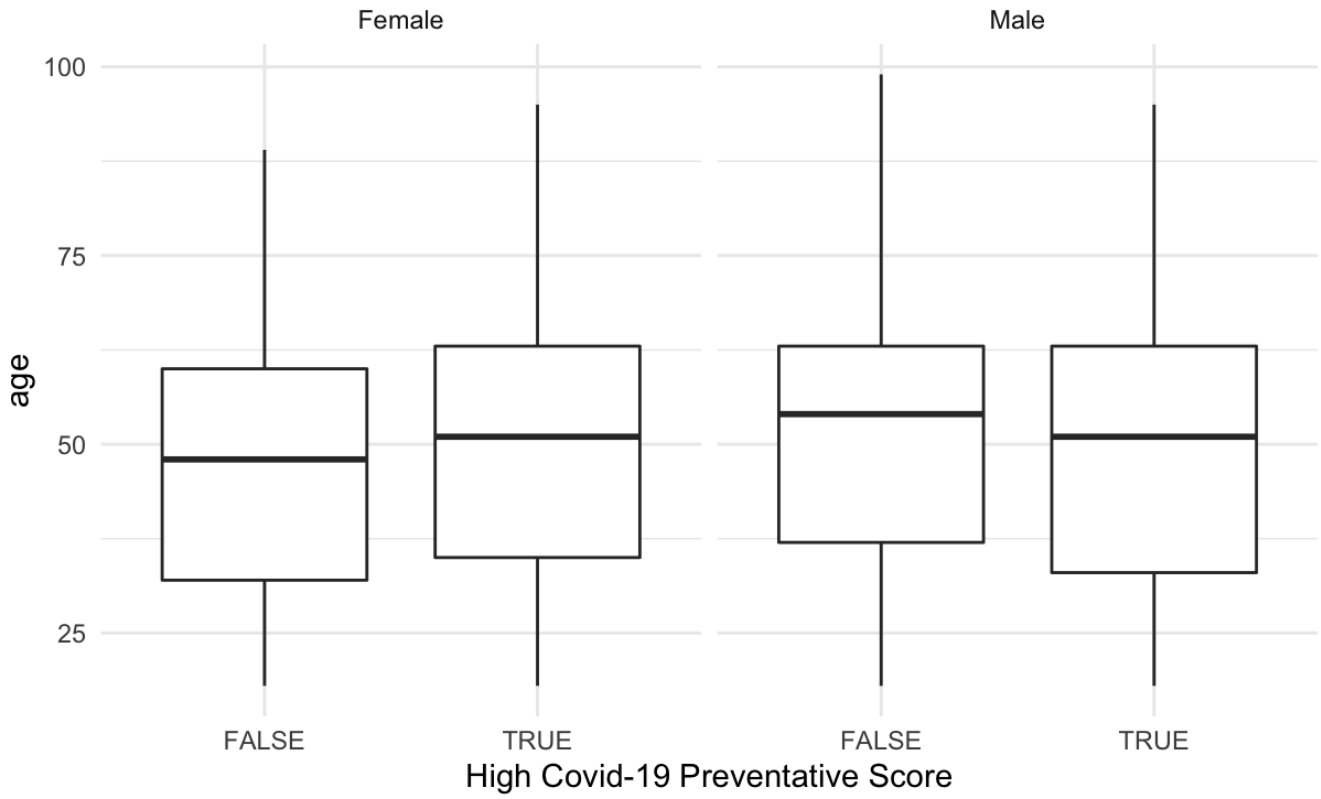
Variable Descriptions

The binary outcome variable for this logistic regression model is whether or not someone has a high preventative score. Out of a total of 90, a high score is classified as a COVID-19 preventative score of greater than or equal to 60 and means that a person is displaying most of these preventative behaviors most of the time. The explanatory variables used to predict this are age, gender, month survey was completed, and level of fear towards COVID-19. Age is in years and ranges from 18 to 99. Gender is binary with categories of male or female. There are five month categories that cover every month the study was conducted ranging from April 2020 to August 2020. The last variable measures how scared the participant is about contracting COVID-19 and has four categories: not at all scared, not very scared, fairly scared, and very scared. Another variable that was used in model selection but did not end up in the final model is the region of the US participants live in with the categories being North East, South, North Central, and West.

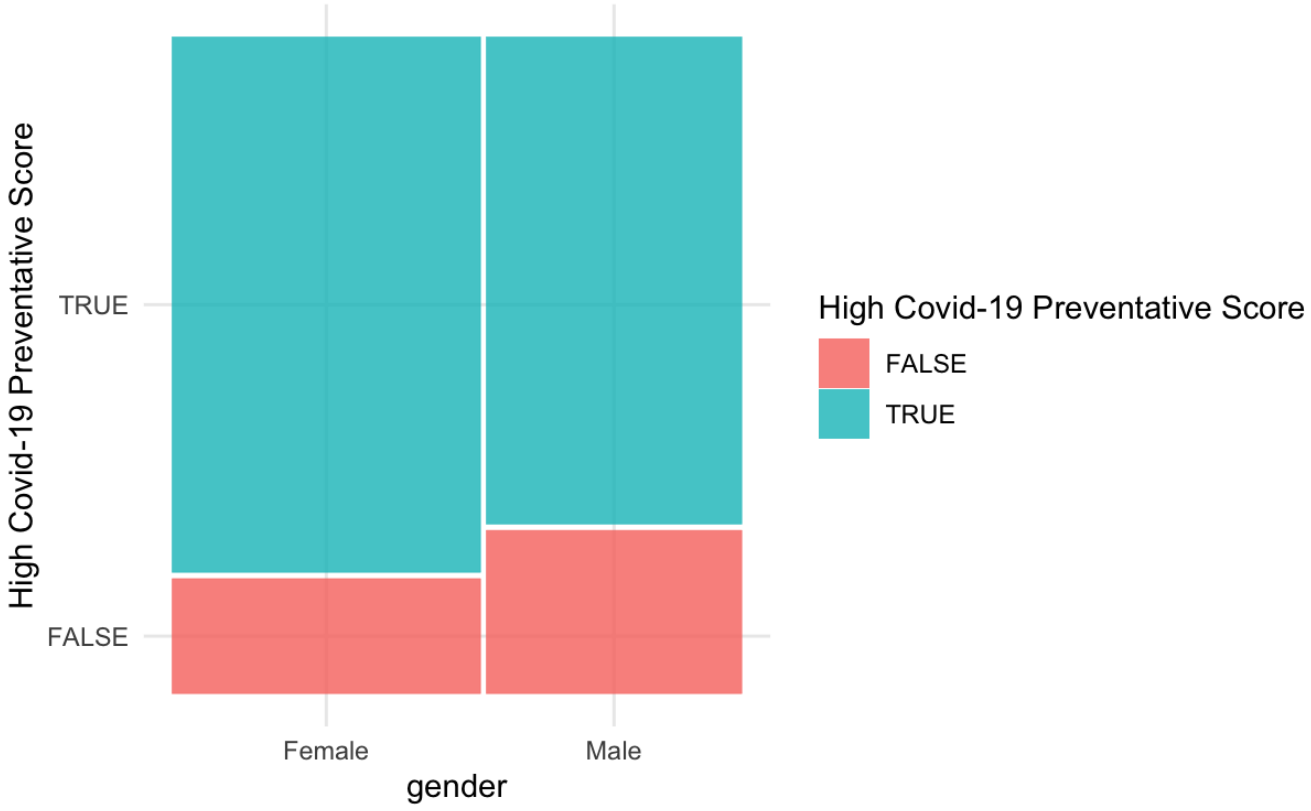
Sample Description

The sample used in this model is the same as the one used above for the linear regression model with 9102 individuals who answered whether they were scared and had not yet contracted COVID-19.

Visualization



These boxplots show that the median age for those who had high COVID-19 preventative scores (60 or greater) is roughly the same for both genders and looks to be just over 50. For individuals who do not have high preventative scores (below 60) the median age is different between genders, seeming to be just under 50 for females and around 55 for males. This is interesting because it means that when comparing low to high scores, the average age for females increases but decreases for males. I'll also note that the interquartile ranges are similar throughout and there aren't any extreme outliers, so both genders have similar spreads and concentrations of age for both high and low scores.



From this mosaic plot we can see that both genders had larger proportions of high preventative scores than low preventative scores. We also see that there are slightly more female participants in this dataset than males. The plot also shows that females have a noticeably larger proportion of high COVID-19 preventative scores than males do.

Model Selection

First to select my final logistic regression model I did a hypothesis test comparing a model with age, gender, month of survey, and how scared of COVID-19 a person is to a nested model without month. The p-value I received from this test was less than 0.001 which gave me enough evidence to reject the null hypothesis that the smaller model (without month) is correct. Then I wanted to see if adding a different variable instead of month would make a better model so I created predicted probability boxplots for the model with age, gender, month, and how scared, and a new one with age,

gender, how scared, and region of the United States. However, the boxplots both looked very similar for both models, the only clear difference being that one for the model with region had slightly higher median predicted probabilities. So, I decided to calculate the accuracy of each model and found that my original model with month predicted 71% of total outcomes correctly while the model with region predicted 68% of total outcomes correctly. This higher accuracy led me to confirm the model with age, gender, how scared, and month as my final logistic regression model.

Final Model Statement

$$log(Odds[highPreventativeScore \mid age + gender + month + scared]) = \beta_0 + \beta_1 age + \beta_2 genderMale$$
$$\beta_3 month5 + \beta_4 month6 + \beta_5 month7 + \beta_6 month8 + \beta_7 scaredNotVery + \beta_8 scaredFairly + \beta_9 scaredVery$$

Fitted Model

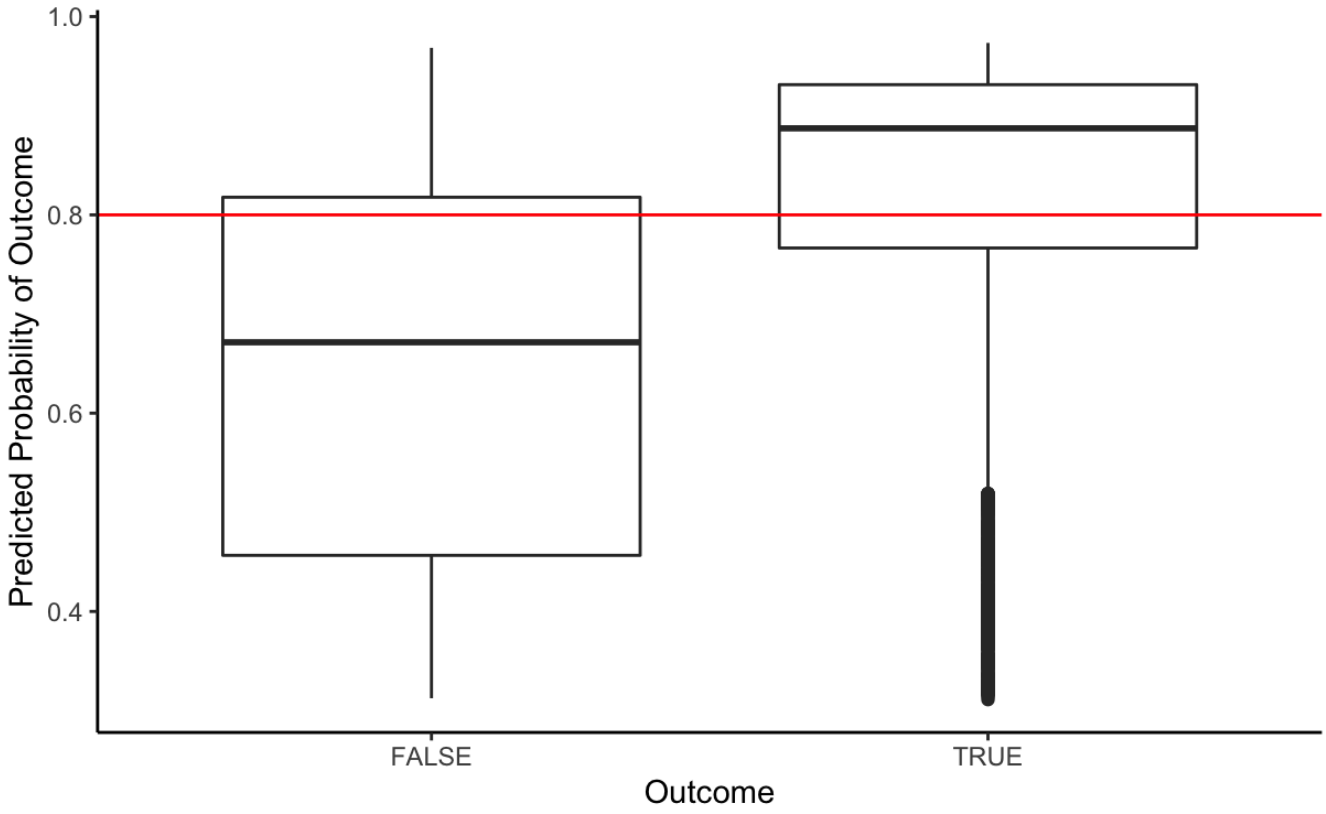
Odds Ratios	Estimates	95% Confidence Intervals
age	1.0065	(1.0033 , 1.0098)
genderMale	0.73	(0.66 , 0.82)
month5	0.73	(0.59 , 0.91)
month6	0.41	(0.32 , 0.51)
month7	0.40	(0.32 , 0.50)
month8	0.38	(0.30 , 0.49)
scaredNotVery	3.60	(3.15 , 4.11)
scaredFairly	11.70	(9.99 , 13.75)
scaredVery	16.38	(13.22 , 20.47)

Interpretations

Holding gender, month of survey, and level of fear towards COVID-19 constant, this model estimates a multiplicative change of 1.006 in the odds of having a high preventative score for each additional year of age. A larger range of plausible values for this multiplicative change in odds is (1.003, 1.009), which should include 95% of the variance in estimates. Although this is a small change from year to year, because 1 is not within the confidence interval, it is more likely than not that there is some sort of relationship between age and increase in odds of having a high COVID-19 preventative score. In addition, the p-value for this estimate is 7.75e-05 which means that there is less than a 5% chance of getting an estimate equal to or higher than the one we have if there was in fact no relationship between high score and age. Now looking at gender, if we hold age, month, and how scared constant, this model

estimates the multiplicative change in odds of having a high COVID-19 preventative score from females to males is 0.73. The range of plausible values for this change in odds between genders is (0.66, 0.82) and the p-value is 1.25e-07. These both point to there being a clear difference in odds of getting a high preventative score between genders, with males being less likely than females to display higher amounts of preventive behavior.

Model Evaluation



After experimenting with different thresholds, for evaluation of this model I have chosen a threshold of 0.8 because it falls between the two median predicted probabilities and balances the sensitivity and specificity of the model. So using this threshold of 0.8 and assuming that every person that's predicted to have an 80% chance or higher of having a high preventative score does have a high preventative score, the accuracy of this model is 0.71. This is fairly high and means that the model correctly predicts 71% of the total number of outcomes. Like I mentioned before, choosing a threshold of 0.8 also balances both the sensitivity and specificity of this model to also both be a little over 0.71 leaving the false positive and false negative rates just over 0.28. This means that the model falsely predicts someone as having both a preventative score of higher than 60 when it was actually lower and lower than 60 when it was actually higher 28% of the time. So, basically this model is right roughly 70% of the time. Something else that could be important to point out is that although the sensitivity is fairly high, there are a number of outliers that truly did have a high preventative score but had predicted probabilities under 50%. Although this certainly is not a perfect model, it is certainly not horrible either; I think this model is good enough for what it does, especially when considering that falsely predicting whether someone has a high COVID-19 preventative score is unlikely to lead to any disastrous real life consequences.

Conclusions

General Takeaways

Reflecting on the linear regression model, it is likely that, taking into account month of survey and level of fear towards COVID-19, there is a very small positive relationship between age and increase in COVID-19 preventative behaviors for females, but it is unlikely that there is much of a relationship between the two for males. So I would caution assuming that there is a definite relationship between age and increase in COVID-19 preventative behaviors in the greater population of the United States. The logistic model for predicting high preventative score backs this up because although it estimated that there could be a relationship between age and odds of having a high preventative score, the estimate itself for the multiplicative change in odds each year of having a high score is very small as well so it is possible that there isn't actually any real relationship between the two in the greater population. I believe the only concrete relationship these models predict is that males have lower odds than females of displaying a high amount and frequency of COVID-19 preventative behavior.

All this being said, these models are certainly not perfect and I would not recommend generalizing their estimates to the broader population of the United States, but if we do take what they say to be true for the sake of providing a general takeaway there are a few things to be touched on. First of all, I think the fact that we've found there to be no definite relationship between age and preventative behavior is really interesting because like I mentioned in my introduction, the media likes to point fingers at younger generations especially for not doing enough to combat the spread of COVID-19. This study would disprove these claims by the media and show that it is unlikely there is a real relationship between age and increase in COVID-19 preventative behavior. It is also interesting that there does seem to be a real relationship between gender and preventative behavior, one would imagine that both genders would have similar relationships with odds of having a high preventative score, especially when taking into account both age and feelings of fear towards COVID-19. But, because we do see a discrepancy between genders here, perhaps this points to the need for health organizations to increase their efforts on targeting males specifically when providing information about safe COVID-19 behavior.

Limitations

In addition to the fact that both our linear and logistic regression models are not necessarily trustworthy or applicable to the general population, moving outwards there are also many potential biases that may be present in this study in terms of data collection and representativeness of our sample. Looking at the data there are 165 different variables so I would assume the survey was fairly time consuming, this means that it is likely that certain people who did not have much free time in their schedules were unable to participate. This skews the data especially when looking at the variables that relate to a person's employment status because those who are unemployed or just working part time would have been more able to participate. In addition to this potential bias is the fact that there are certain groups of people who may be struggling with serious health issues or may be hospitalized who would have been less able to complete the survey. This means that people who may be suffering from serious illnesses, other than COVID-19, and who may be displaying stricter preventative behaviors because of it, are potentially under-represented in this sample.

There are also several information biases that may be present. For example, there are several questions in the survey, and that ended up counting towards my preventative score variable, that ask

participants to remember things like the number of people they came in close contact with in the past week, or the number of times they washed their hands in a day. These variables are liable to recall bias because it may be difficult for certain people to remember exactly how many times they did these things especially if they are trying to recall something that has happened in the past 7 days. Social desirability bias may also be present in this data set due to the fact that there are certain questions people may feel ashamed to answer truthfully especially if it relates to doing something not encouraged by their local government like how often they are attending social gatherings.

When considering limitations of these models as well as potential harmful consequences should this research be released, something important to address is that the gender variable used is binary and therefore excludes the identities of those who don't fall into the gender binary. Also, another limitation of these models is that the sample I used does not include those who have already contracted COVID-19, so the models are not applicable to the population of the whole United States but just those who have not yet contracted COVID-19.