US Attitudes Towards the Covid-19 Vaccine

Salomé Garnier, Simone Lewis, Neha Gupta, Corbin Duncan, Connor Brown

5/03/2021

Introduction

The novel coronavirus leading to COVID-19 infection is regarded in terms of mortality, economic ruin and transmissibility, to be the most severe pandemic in over a century. Dr. Anthony Fauci, the United States' leading infectious disease expert and White House advisor, has expressed that vaccine uptake and acceptance will be critical to any return to 'normal' pre-pandemic life. At the time of writing, there have been 140 million recorded cases of the infectious disease, 80 million recovered cases and 3.01 million fatalities.

The public health response to the COVID-19 pandemic across the United States has been uncoordinated across state lines. Public cooperation with public health guidelines and restrictions has also varied greatly across the country and appears to be correlated with partisanship. Early scholarship suggests the infection has a disproportionately high infection rate among poor Americans. Manely and Shrestha (2021) find supportive evidence for counties with higher rates of poverty being less responsive to stay-at-home orders and having a higher infection rate, even in low-density locations. Some experts also expect there will be discernible variations in vaccine uptake along ethnic and wealth cleavages. What we can deduce from the aforementioned data is that Americans' experiences of the pandemic and their relationship to health authorities is highly heterogeneous and variable. We can further intuit that health outcomes and vaccine uptake may correlate with group identity or individual characteristics.

In this project, we seek to predict and assign a score to the willingness of individuals to be vaccinated based on their demographic characteristics. We will test and train our model using a dataset of factors associated with acceptance of the vaccine. We will report the efficiency statistic for this model. Our group will then seek to make cursory comparisons between predicted vaccine acceptance and compliance with public health restrictions in regions in the United States, by determining the accuracy of our vaccine hesitancy model in also predicting non-compliance with public health guidelines. Finally, we will consider the limits of our model, horizons for future research, and policy implications.

Data

To answer the question of whether sociodemographic factors can be determinative in predicting vaccine acceptance, we manipulated the dataset 'The factors determining the preference for COVID-19 Vaccine,' collected by Mondal and Sinharoy (2021) of Penn State College of Medicine Department of Pediatrics and published by Mendeley Data. The survey was conducted over the period of May 2020 to January 2021, with a binary indicator variable distinguishing between responses before and after the first successful vaccine announcement on November 1st, 2020. The number of eligible participants, defined by the researchers as participants who completed all survey sections and resided within the U.S, totaled to 2,978.

In addition to demographic data and data about the willingness to accept a covid-19 vaccine, two continuous variables were constructed by the authors. Firstly, the researchers calculated a measure of perceived threat for each respondent based, not solely on participant input, but on the age and chances of severe infection for a respondent, whether they are a healthcare worker and the severity of outbreak in a participant's area.

Secondly, the researchers consolidated responses gauging first-hand and secondary knowledge of the virus into a single measure. Respondents were asked 71 questions, seeking to record an individual's pandemic-related stress, efforts to prevent disease spread, support for public health measures, factual knowledge of the infection, trust in local hospital efficacy and hope for the future. These responses were recorded alongside individual characteristics such as age, race, education level, family income, gender, U.S region, healthcare worker status, and healthcare access.

With 2,978 respondents, our dataset achieves a diversity in responses which will improve the predictive power of our model. Of these respondents, 28.7% report having at least one family member infected by COVID-19, 18.3% report being a healthcare worker, and 81.1% report a willingness to accept a coronavirus vaccine. In terms of race, white respondents are over represented, but we record a value above 4% for each racial identity included.

Modeling Methodology & Results

To build the model of vaccine hesitancy, we first begun by splitting the cleaned survey data into training and testing data. Splitting the dataset will prevent overfitting of the model by preventing the prediction value from becoming too aligned with the entire set of observations. Using random sampling, we decided to allocate 75% of the data (N = 2233) towards the training data for model building, and the remaining 25% (N = 745) for testing the accuracy and efficiency statistics of our model's predictions of vaccine hesitancy. We decided to use a generalized multivariable logistic regression to predict the outcome variable (Y) of vaccine willingness or hesitancy. The outcome variable, "covid_vaccine," is coded as 1 for someone who answered that they would receive the vaccine when offered, and 0 for refusal. We fitted the first model of COVID-19 vaccine willingness with the predictor variables of gender ("gender"), age group ("age_group"), education level ("education"), race ("race"), and household income level ("financial_status"). The second model also included whether the participants had received the flu shot within the last 3 years from taking this survey. We then reported the results from both models 1 and 2.

```
## [1] "Dimensions of training dataset"
```

```
## [1] 2233   11
```

```
## [1] "Dimensions of testing dataset"
```

```
## [1] 745  11
```

Starting with model 1, from the sampled training data, every predictor variable was significant from at least the $p < 0.05$ level. The strongest predictors were education level (p = 7.23e-6) and being Black (p = 0.000135). Education had a positive relationship, where level of education rose with willingness to receive the COVID-19 vaccine, being Black had a negative relationship where they were less likely to accept the vaccine of all racial groups. All other races exhibited some negative willingness to receive the vaccine, but none near the magnitude and significance of Black respondents. Financial status was also a strong predictor (p = 0.002418) and had a positive coefficient showing that as income levels increased, so did willingness to receive the vaccine. Gender also showed that being male had a positive association with wanting to receive the vaccine, although not statistically significant.

Model 1: Excluding flu shot

```
##
## Call:
## glm(formula = fm1, family = "binomial", data = n_train, na.action = na.exclude)
```

```
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5030   0.3885   0.5337   0.6576   1.3128
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)       0.70419    0.46500   1.514 0.129930
## genderMale        0.38876    0.15209   2.556 0.010587 *
## age_group         0.13070    0.05422   2.410 0.015932 *
## education         0.22514    0.05018   4.487 7.23e-06 ***
## raceBlack        -1.66140    0.43522  -3.817 0.000135 ***
## raceHispanic     -1.18410    0.44273  -2.675 0.007483 **
## raceOther        -1.22936    0.47058  -2.612 0.008990 **
## raceWhite        -0.79680    0.40442  -1.970 0.048810 *
## financial_status  0.28862    0.09515   3.033 0.002418 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1993.1  on 2117  degrees of freedom
## Residual deviance: 1881.3  on 2109  degrees of freedom
##   (115 observations deleted due to missingness)
## AIC: 1899.3
##
## Number of Fisher Scoring iterations: 5
```

Model 2: Including flu shot

```
##
## Call:
## glm(formula = fm2, family = "binomial", data = n_train, na.action = na.exclude)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5743   0.3602   0.4694   0.5872   1.6862
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -0.17200    0.48347  -0.356 0.722013
## genderMale        0.40901    0.15782   2.592 0.009553 **
## age_group         0.07706    0.05741   1.342 0.179522
## education         0.18554    0.05251   3.533 0.000410 ***
## raceBlack        -1.61114    0.44690  -3.605 0.000312 ***
## raceHispanic     -1.24920    0.45333  -2.756 0.005858 **
## raceOther        -1.10759    0.48267  -2.295 0.021748 *
## raceWhite        -0.79682    0.41250  -1.932 0.053401 .
## financial_status  0.22091    0.09849   2.243 0.024891 *
## flu_shot          1.62677    0.13908  11.696  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

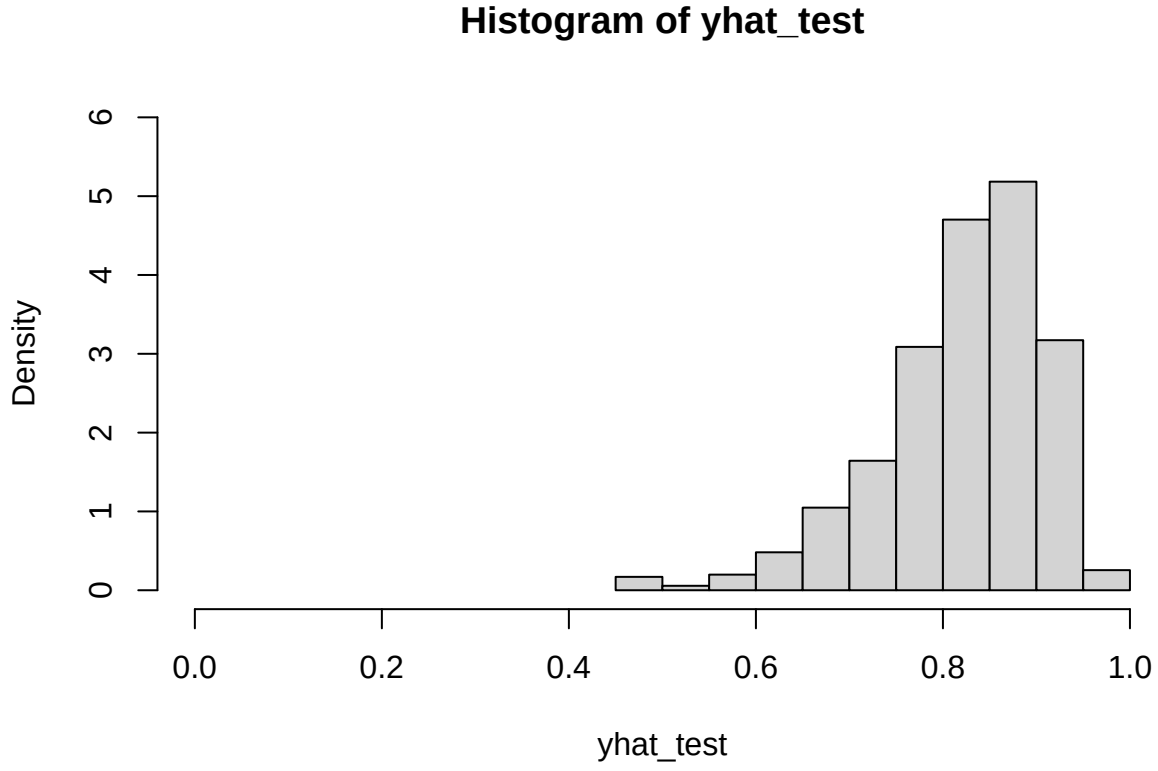
```
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1992.7  on 2116  degrees of freedom
## Residual deviance: 1749.4  on 2107  degrees of freedom
##   (116 observations deleted due to missingness)
## AIC: 1769.4
##
## Number of Fisher Scoring iterations: 5
```

Collinearity of flu shot and Covid-19 vaccine:

```
## [1] 0.3261726
```

An interesting observation we noted was found when running the second model including flu shot uptake. It had by far the most significant p-value of all indicators in both models (p = 2.25 x 10-7) and the strongest coefficient estimate. Flu shot uptake also had strong collinearity (r = 0.326) with COVID vaccine willingness, as a result, all other demographic predictors lost or had reduced significance relative to model 1. This will be discussed further in the policy proposal, but it likely indicates that COVID vaccine willingness is heavily predicted by overall individual sentiment to the pharmaceutical and vaccine industry in general, rather than the COVID vaccine's lack of field research and development relative to other inoculations with greater mainstream prevalence (Hopkins, 2020). As the flu shot skewed the significance of demographic variables predictive power, we decided to proceed with testing the efficiency of model 1 to isolate what demographics need greater targeting to increase their trust in the vaccine.

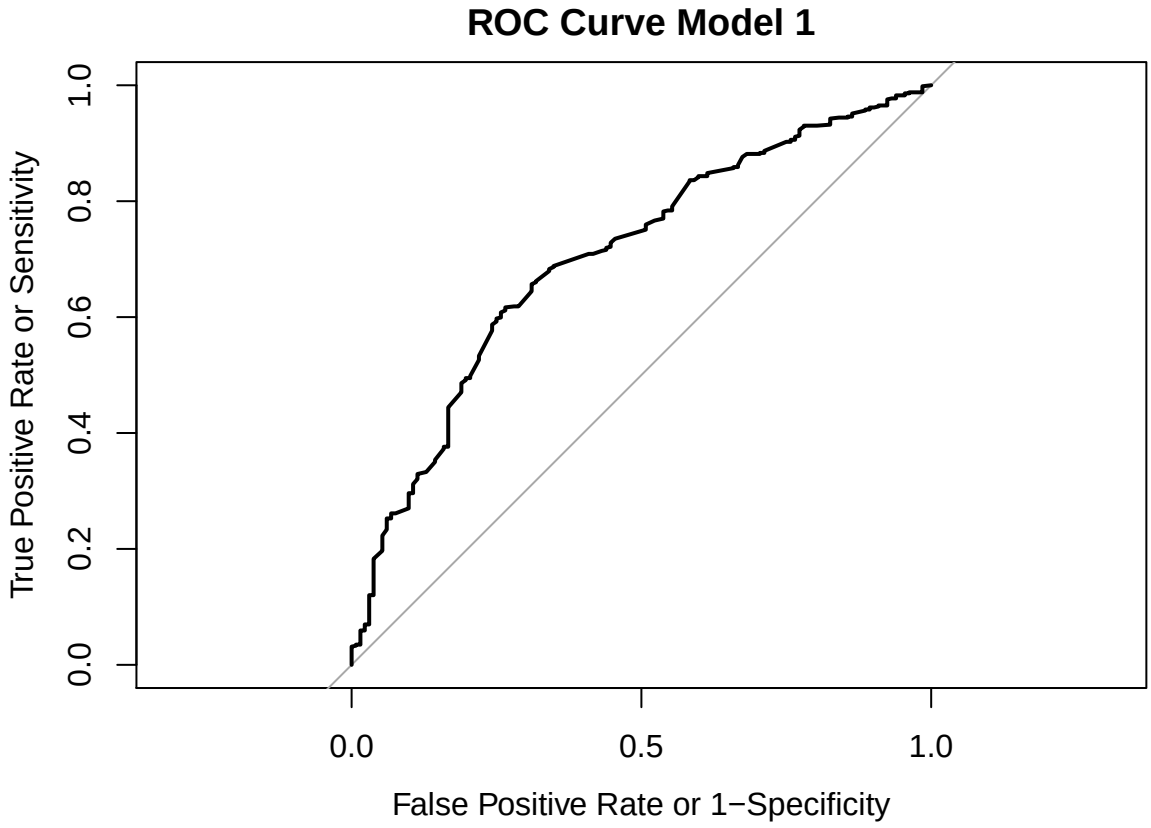
We then decided to create a prediction of vaccine willingness using model 1 and the remaining test data, "yhat_test," the density plot outlined below shows that the average willingness prediction trended to a 0.82 chance to receive the vaccine.



Mean of yhat:

```
## [1] 0.8205602
```

After creating our predictor variable from the test data, we then plotted a receiver operating characteristic (ROC) curve for vaccine willingness model outlined below.




```
## Area under the curve: 0.703
```

The ROC curve visually represents the probability that a true positive case (sensitivity) outranks a false positive error (1 – specificity), we computed an area under curve (AUC) value of 0.703, indicating an above 70% chance that a true positive case will occur relative to a false positive. This would position our model in the acceptable discrimination range (Hosmer & Lemeshow, 2013), where any value below 0.5 would suggest no discrimination between positive and negative cases, no better than random chance. Using the results from the ROC curve, we then calculated the threshold to classify an observation inserted into model 1 as a positive case of wanting to receive the COVID-19 vaccine. Utilizing best weights, we computed a threshold value of 0.827 to classify an observation as a positive case.

threshold	accuracy	precision	recall
0.827	0.639	0.91	0.617

f-score:

```
## [1] 0.7353897
```

Model Efficiency

```
##
##                              Deny Vaccine (True) Seek Vaccine (True)
##   Deny Vaccine (Predicted)           0.13739377          0.31161473
##   Seek Vaccine (Predicted)           0.04957507          0.50141643
```

```
## [1] "False negative rate is 0.312 (top right)"
```

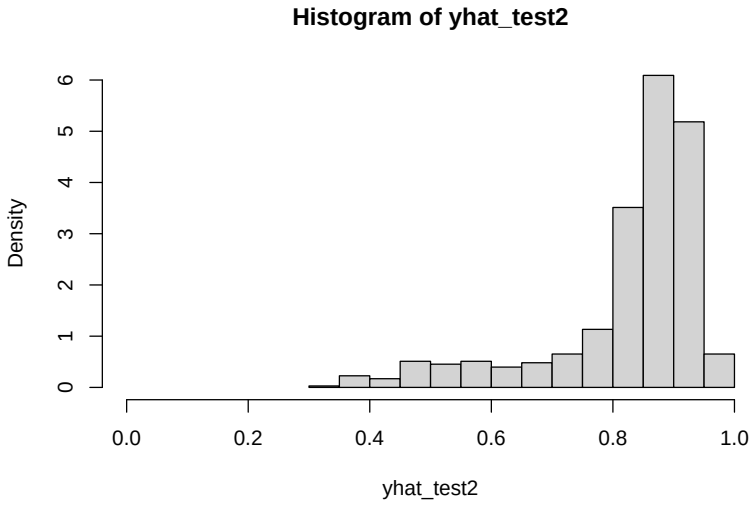
```
## [1] "False positive rate is 0.0496 (bottom left)"
```

An efficiency table of this calculation showed that our model had an accuracy, the proportion of correct predictions, at 0.639, its precision, the percentage of relevant results, at 0.91, and recall, the percentage relevant results that are correctly classified by the model at 0.617. We computed the F1-score, a truer measure of accuracy accounting for precision and recall, giving more weight to false negatives and false positives while not letting large numbers of true negatives influence the score, this result came to 0.735. Finally, we produced a 2x2 table to summarise the true positivity and negativity rate, as well as the false positivity and negativity rate which aligned with the efficiency statistics above.

From a public policy perspective, the true aim of our model is to accurately predict which demographics are hesitant to receive the COVID vaccine, and then employ effective targeting of these groups to allay risks of seeking inoculation to hasten the end of the pandemic and reach herd immunity. This is especially true for minority groups and those from lower socio-economic backgrounds, who are more likely to experience more severe consequences from contracting the virus due to greater prevalence of comorbid conditions relative to white, and wealthier groups (Obermeyer et. al, 2019). In this case, the F1-score is especially important as it weighs false positives, an observation that we predict will seek the vaccine, but will actually deny the vaccine in reality, more heavily. These are cases that our model must necessarily avoid at all costs, as it would result in groups not receiving sufficient targeting to receive the vaccine due to predicting they would in reality this is not the case. Our false positivity rate was below 5% (0.0496) showing that these vulnerable groups are in most cases being captured by the model, and is reflected in the robustness of the model's F1-score.

Discussion of Model 2

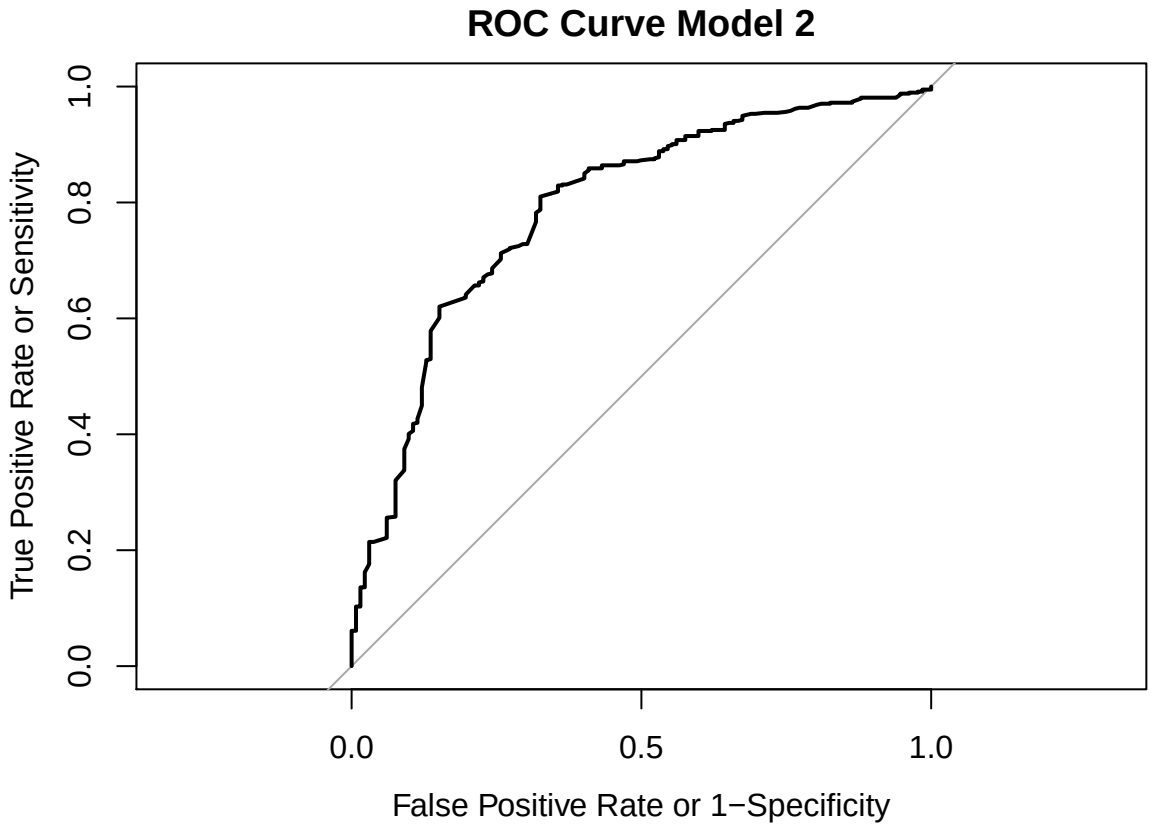
We briefly decided to test the efficiency statistics of Model 2 in comparison to Model 1 to test its relative efficacy. After plotting its predicted value density plot, we found the mean predicted value of observations was almost identical to Model 1 at 0.8218831.



Mean of yhat 2:

```
## [1] 0.8218831
```

Plotting the ROC curve for Model 2, we found that its discriminatory power between true positive and false positive cases was 0.7902 for its area under curve value. This shows that the probability of a positive case outranking was 12.4% better at discriminating between cases than Model 1 showing greater predictive power of detecting accurate cases when including flu shot uptake.




```
## Area under the curve: 0.7902
```

Analysing the efficiency of Model 2, we computed the model's accuracy to be 0.785, its precision to be 0.915, and the recall at 0.81; these were all improvements over the figures produced by Model 1 discounting flu shot uptake. This was reflected in the improved F1-score of Model 2, which we computed as 0.859, a 17% relative improvement over Model 1. However, as discussed previously, the primary goal of our model is to minimise false positives, an observation that we predict will seek the vaccine, but will actually deny the vaccine in reality. A high false positive rate would result in groups not receiving sufficient targeting to receive the vaccine as our model would predict they would when in reality they refuse uptake. For Model 2, our false positivity rate was 0.0609, this was 12% higher than Model 1 showing that vulnerable groups would receive less accurate targeting in the former. As a result, our further analysis in this memo will incorporate Model 1 into our additional calculations and analyses.

threshold	accuracy	precision	recall
0.82	0.785	0.915	0.81

f-score:

```
## [1] 0.8593043
```

Model Efficiency:

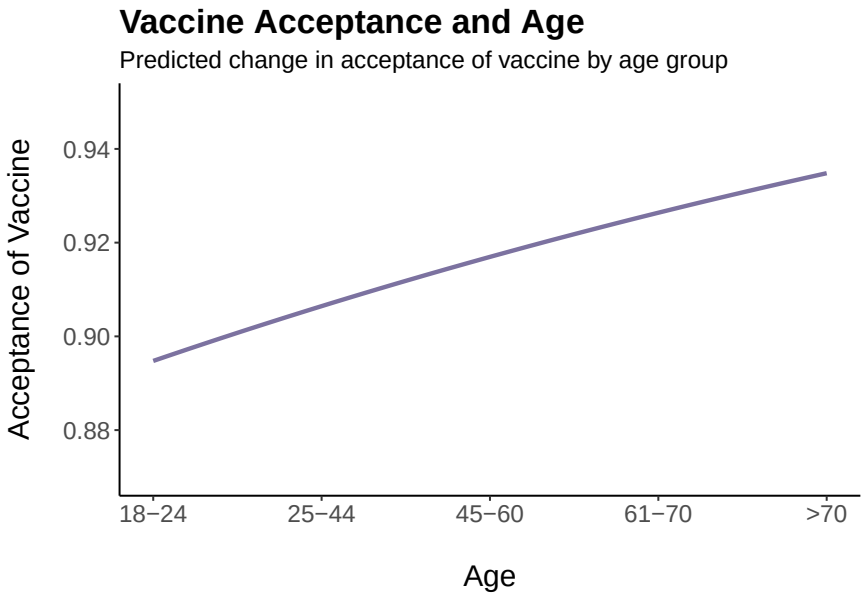
```
##
##                             Deny Vaccine (True) Seek Vaccine (True)
##   Deny Vaccine (Predicted)          0.12606232          0.15439093
##   Seek Vaccine (Predicted)          0.06090652          0.65864023
```

```
## [1] "False negative rate is 0.154 (top right)"
```

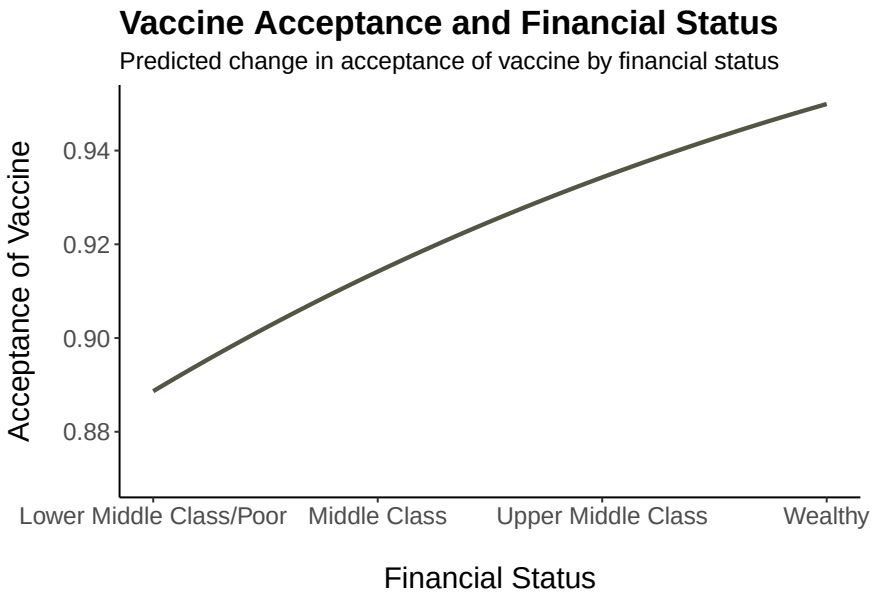
```
## [1] "False positive rate is 0.0609 (bottom left)"
```

Discussing and Visualizing the Results

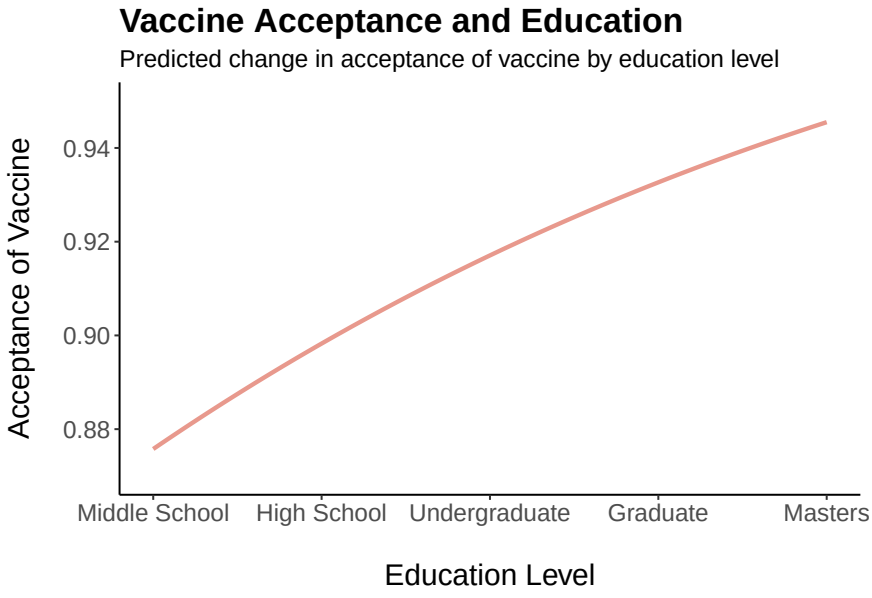
Overall, being white, older, richer, and more educated were strong predictors of acceptance of the vaccine.



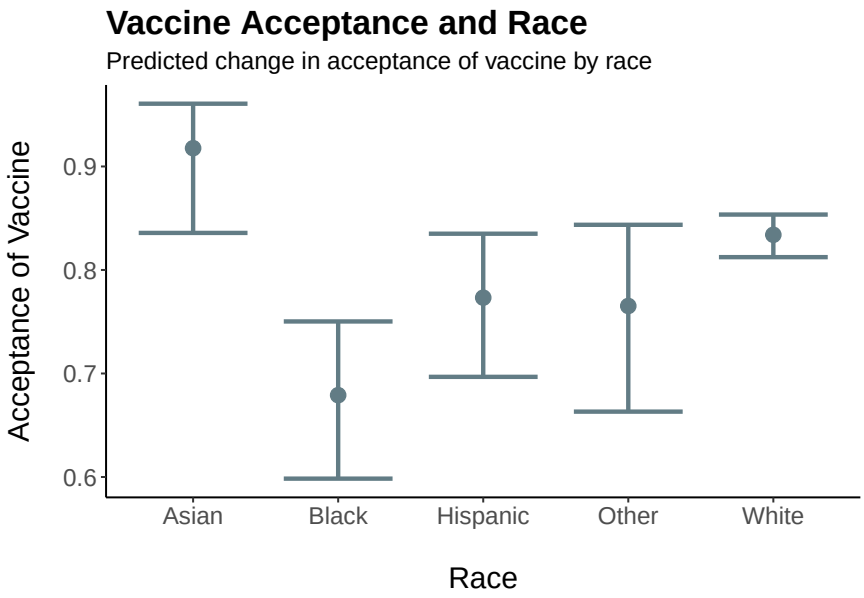
We observe a positive and strong relationship between age and greater COVID-19 vaccine acceptance. The most obvious reasoning behind this observation is the parallel relationship between age and the risk of severe illness from COVID-19. Older individuals face a greater risk of hospitalization and death from the virus and 8 out of 10 COVID-19 deaths reported in the U.S. have been among those 65 and older (CDC 2021). Differences in the level of risk for COVID-19 reflects the variability of vaccine acceptance among age groups with older adults having greater incentives to take the vaccine.



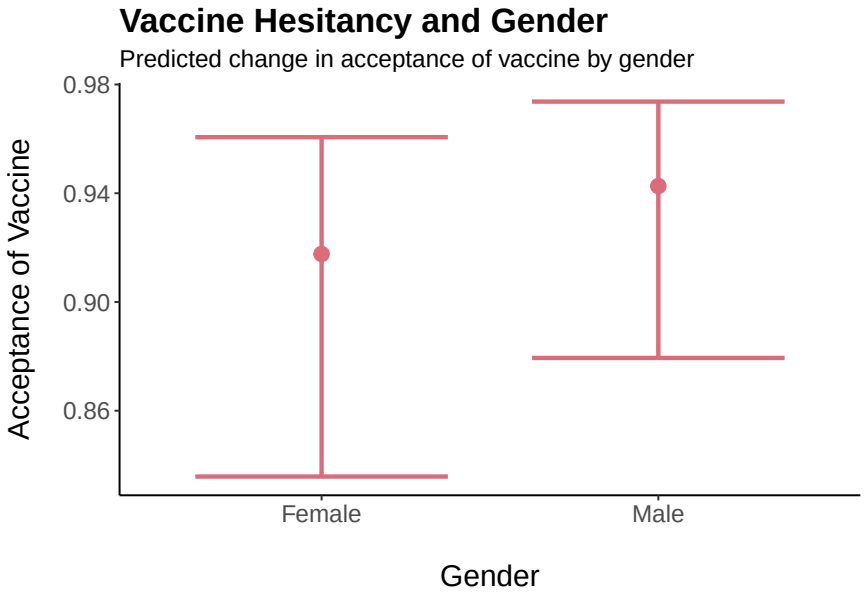
We also observe a strong and positive relationship between an individual's reported financial status and their likelihood of vaccine acceptance. In the United States, higher income typically allows for better access to healthcare services and greater interaction with healthcare professionals while the opposite is true for lower income. This connection is a potential explanation for the greater mistrust towards healthcare among those with a low income. Cost related concerns are also more prominent among low-income individuals and an additional factor that can promote vaccine hesitancy. Notably, a recent Census survey revealed that nearly 7 million Americans were unaware that the COVID vaccine was free (Fortune 2021).



Given the strong correlation between education and financial status, there is a lot of overlap in the mechanisms that help explain why COVID-19 vaccine acceptance varies along these dimensions. For instance, lower education is also associated with lower interaction with healthcare professionals and a related lower level of trust. Additionally, education is closely related to awareness about the details of vaccines and general health literacy (Khubchandani et al. 2021).



Results show that acceptance of the vaccine varies significantly by race, with Asian Americans trusting the vaccine overwhelmingly more and African Americans trusting it significantly less. There are many potential explanations for the low levels of acceptance among Black people, including a history of unethical medical experimentation on Black Americans in the United States and more generalized distrust of the government and its associated institutions. But African Americans also suffer from discrimination and biases in the healthcare system that could explain lower levels of trust in the health system.For example, health professionals tend to hold differential perceptions of Black patients in terms of intelligence and pain tolerance, which reduces the rates at which non-Black doctors admit Black patients for healthcare interventions. Also, when Black patients are assigned to a physician, there is a higher uptake of recommended care when the provider is Black. Overall, these exhibited racial biases on the part of human error and prejudice can explain the significantly reduced trust Black individuals have in the healthcare system (Obermeyer et. al, 2019).



Although the predicted vaccine acceptance among men is slightly higher than for women, the confidence intervals for these estimates overlap and we cannot say for certain that there is a relationship between

gender and COVID vaccine acceptance.

Vaccine acceptance and public health responsiveness

After designing a predictive model of vaccine acceptance, our team sought to exploit other observations within the database. We hypothesized that there may be a relationship between vaccine acceptance and resistance to public health restrictions. In terms of mechanisms, such a result would perhaps stem from a broader distrust of health authorities. Alternatively, we considered that those who were most opposed to public health restrictions may also be those most inclined to get the vaccine, so that restrictions would disappear more rapidly.

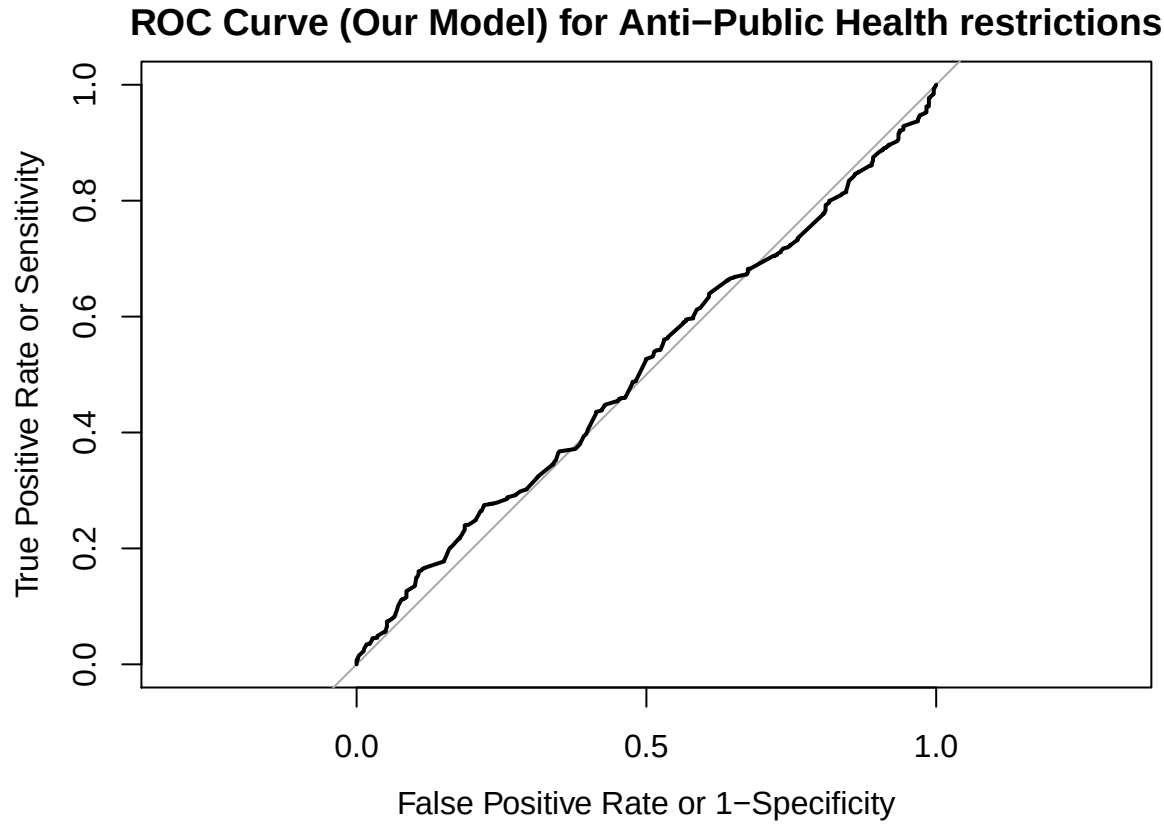
Mean yhat for larger dataset:

```
## [1] 0.8205602
```

To consider this question, we first generated a new variable for resistance to public health advice, based upon views on mandatory face masks, lockdowns and whether re-opening was a good or necessary action.

Mean resistance to public health guidelines:

```
## [1] 0.8230536
```




```
## Area under the curve: 0.5085
```

Efficacy of Predictive Power We computed an area under curve (AUC) value of 0.5085, indicating an approximate 50% chance that a true positive case will occur relative to a false positive. This would position our model below the acceptable discrimination range (Hosmer & Lemeshow, 2013), where essentially our model could not discriminate better than random chance.

As we undertook for our predictive model for vaccine hesitancy, using the results from the ROC curve, we calculated a threshold to classify observations as a positive case of being resistant to public health restrictions. Utilizing best weights, we computed a threshold value of 0.885 to classify an observation as a positive case.

threshold	accuracy	precision	recall
0.885	0.36	0.859	0.275

Analysis of Vaccine Model on Public Health Opinions

With an accuracy statistic of just 36% and a initial predictive power essentially as good as random, we find that our model's ability to predict vaccine acceptance does not extend to an ability to predict resistance to public health policies such as face masks and lockdowns.

Of course, if we wished to analyze this issue critically, we could similarly generate a model based on our dataset. We cannot rule out a relationship between rejecting public health advice and vaccine acceptance, but we can say that predictors of such opinions are not necessarily the same.

Conclusion

Policy Implications

Those who are less educated, financially disadvantaged, and black have the lowest predicted vaccine acceptance. These groups are also some of the most likely to work in essential sectors involving more covid-exposure and experience the greatest burden of COVID-19 (KFF, 2020). Thus, targeting these groups and their concerns in additional outreach is critical to minimize the spread and burden of infection.

Community-driven interventions have been recommended to improve vaccine uptake among hesitant groups. Both the American Psychological Association and government of Connecticut recommend community-driven strategies which involve the following steps:

Step 1: Use data to identify and prioritize racial/ethnic minority communities that may be less likely to receive a COVID-19 vaccine.

Step 2: For each community of focus, identify relevant government officials and community partners to form a "community partner network."

Step 3: Work with the community partner network to understand barriers in the community and create an implementation plan for vaccination messaging, outreach, and administration.

Step 4: Help community partner networks implement plans, providing funding and support as needed.

Step 5: Conduct continuous program evaluation through data collection and analysis to inform possible changes to the ongoing strategies. This data analysis assists with Step 1 of the community engagement process by identifying the groups that would be most hesitant to receive the vaccine. Further steps involve understanding barriers in the community and creating targeted messaging, outreach, and administration to address these issues. The Kaiser Family Foundation's survey found that among those who are hesitant to get a COVID-19 vaccine, the main reasons are: worries about possible side effects (59%) lack of trust in the government to ensure the vaccines' safety and effectiveness (55%), concerns that the vaccine is too new (53%) concerns over the role of politics in the development process (51%).

Additionally, the survey found that about half of black adults who say they probably or definitely won't get vaccinated cite that they don't trust vaccines in general (47%) or that they are worried they may get COVID-19 from the vaccine (50%).

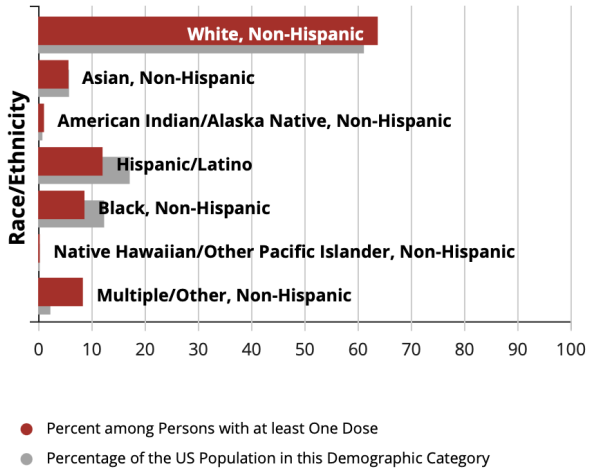
Thus, outreach addressing these specific worries would be particularly important for raising turnout in these hesitant populations. Mitesh Patel in Nature magazine recommends greater use of behavioral nudges such as framing vaccination as the norm, providing peer-comparision feedback to increase social incentive, and making choice activities time-bound as these strategies have been proven to increase vaccine use.

Further Research

Further research could test how predicted vaccine acceptance (answering "Yes" to the question, "Would you like to get COVID-19 vaccine, if available?") compares to actual vaccine uptake. This comparison could give insight into which groups have a high desire for the vaccine but low use pointing to issues of access and indicating which groups have a high marginal benefit from improving access. Additionally, it would be useful for future surveys to ask why individuals feel hesitant to take the vaccine to understand which barriers policymakers should target to most effectively increase vaccine uptake.

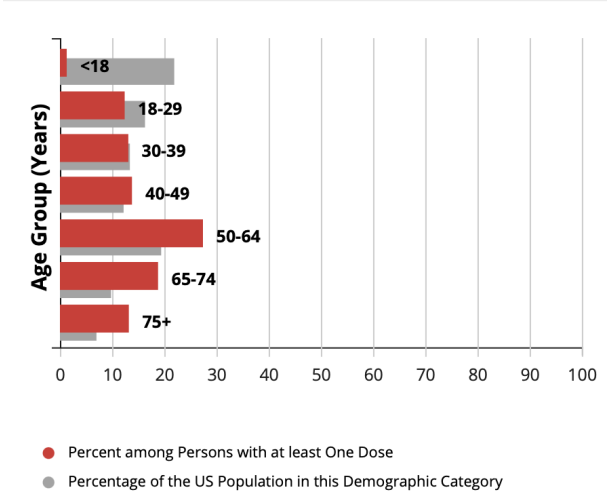
Preliminary data about Covid-19 vaccine uptake in the United States somewhat confirms the predictions of our model. Data from the CDC shows that white people are overrepresented among those vaccinated in the US as of April 2021, while Black and Hispanic populations are largely underrepresented.

Figure 1: Vaccine Uptake and Race (CDC)



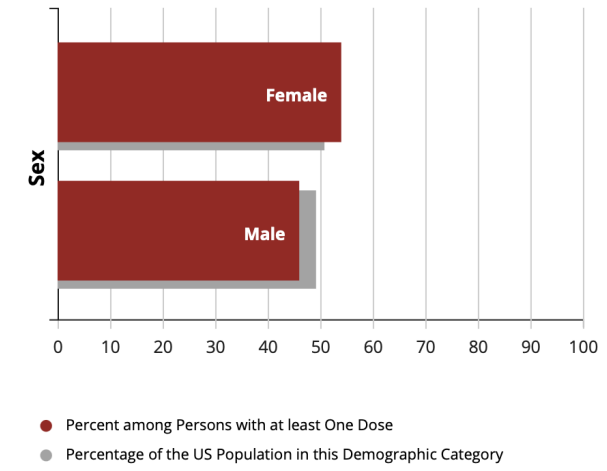
Similarly, older individuals are more likely to be vaccinated on average, although this is likely a function of the US vaccination timeline that gave the elderly priority.

Figure 2: Vaccine Uptake and Age (CDC)



Finally, while our model suggested that women would be less likely to accept the Covid-19 vaccine, the coefficients were not statistically significant, and the CDC data shows that the opposite is true. Unfortunately, the CDC provides no data on financial status and education, which were two of our strongest predictors. Future research can potentially focus on vaccine uptake as a function of those two demographic predictors.

Figure 3: Vaccine Uptake and Gender (CDC)



Unfortunately, the CDC provides no data on financial status and education, which were two of our strongest predictors. Future research can potentially focus on vaccine uptake as a function of those two demographic predictors.

Limitations

A major limitation of the logistic regression is the assumption of linearity between the log-odds of the dependent variable (vaccine acceptance) and independent variables (gender, age group, education, race, financial status, flu shot). Some of these variables, such as age, may not be linearly but exponentially related to vaccine acceptance but our logistic regression model would not capture this relationship.

Additionally, attitudes toward the vaccine fluctuate as new information is released so these survey responses are a reflection of attitudes 05/2020 - 01/2021. The model would likely lose predictive power when applied to other time periods as information around the vaccine evolves.

Works Cited:

Center for Disease Control. "COVID-19 and Your Health." Centers for Disease Control and Prevention, 11 Feb. 2020, https://www.cdc.gov/coronavirus/2019-ncov/need-extra-precautions/older-adults.html

Center for Disease Control, "Demographic Characteristics of People Receiving COVID-19 Vaccinations in the United States," April 27, 2021. https://covid.cdc.gov/covid-data-tracker/#vaccination-demographic

CNBC.com staff. (2021, April 15). Covid live updates: Fauci says universal vaccine is the endgame to combat new variants. CNBC.

COVID Live Update: 141,407,599 Cases and 3,026,333 Deaths from the Coronavirus - Worldometer. (2021). Worldometre, Coronavirus Pandemic by the Numbers. https://www.worldometers.info/coronavirus/

Fortune, Nearly 7 Million Americans Might Not Get a COVID-19 Vaccine Because They Don't Know It's Free." https://fortune.com/2021/03/10/covid-vaccine-free-people-not-getting-coronavirus-vaccines-cost-price/. Accessed 29 Apr. 2021.

Gates, B. (2020). Responding to Covid-19 — A Once-in-a-Century Pandemic? New England Journal of Medicine, 382(18), 1677–1679. https://doi.org/10.1056/nejmp2003762

Hamel et al (2020), "KFF COVID-19 Vaccine Monitor: December 2020." https://www.kff.org/coronavirus-covid-19/report/kff-covid-19-vaccine-monitor-december-2020/

Hopkins, J. (2020). A Covid-19 Vaccine Problem: People Who Are Afraid to Get One. Wall Street Journal.

Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). Applied logistic regression (Vol. 398). John Wiley & Sons.

Jung, J., Manley, J., & Shrestha, V. (2021). Coronavirus infections and deaths by poverty status: The effects of social distancing. Journal of Economic Behavior & Organization, 182, 311–330. https://doi.org/10.1016/j.jebo.2020.12.019

Khubchandani, Jagdish et al. "COVID-19 Vaccination Hesitancy in the United States: A Rapid National Assessment." Journal of community health vol. 46,2 (2021): 270-277

Mondal, Pritish; Sinharoy, Ankita (2021), "The factors determining the preference for COVID-19 Vaccine", Mendeley Data, V1, doi: 10.17632/kbzskd37zy.1

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. Science, 366(6464), 447-453.

Royles, D. (2020, December 15). Years of medical abuse make Black Americans less likely to trust the coronavirus vaccine. Washington Post. https://www.washingtonpost.com/outlook/2020/12/15/years-medical-abuse-make-black-americans-less-likely-trust-covid-vaccine/

Code

```
knitr::opts_chunk$set(echo = FALSE)

#load packages
library(knitr)
library(tidyverse)
library(pROC)
library(glmnet)
library(readr)
library(ggplot2)
library(jtools)

#load dataset
demographics <- read_csv("demographics_id.csv")
vaccine_data <- read.csv("vaccine_data.csv")


summarystat <-do.call(data.frame,
list("Family member infected" = mean(vaccine_data$family_member_covid,
                                     na.rm=TRUE),
"Is healthcare worker" = mean(vaccine_data$healthcare_worker,
                              na.rm=TRUE),
"Vaccine acceptance" =  mean(vaccine_data$covid_vaccine,
                             na.rm=TRUE)
    ))

knitr::kable(summarystat,
             col.names = gsub("[.]", " ",
                              names(summarystat)),
             digits = 3)

knitr::kable(prop.table(table(vaccine_data$your_race)),
             col.names = c("Race","Proportion"),
             digits = 3)

knitr::kable(prop.table(table(vaccine_data$your_education)),
             col.names = c("Education","Proportion"),
             digits = 3)


#split demographics voter data into training and test set using sampling
#Set seed for replication
set.seed(02138)

## For Regression analysis
#allocated 75% of data to training and 25% of data to testing
row.number <- sample(1:nrow(demographics),
                     0.75 * nrow(demographics))
n_train = demographics[row.number,]
n_test = demographics[-row.number,]

#summarise
```

```r
paste("Dimensions of training dataset")
dim(n_train)
paste("Dimensions of testing dataset")
dim(n_test)
## USA Model
#fit the demographic variables onto the outcome variable of vaccine acceptance

#exclude flu_shot
fm1 <- covid_vaccine ~ gender + age_group +
  education + race + financial_status

#estimate model 1 (logistic regression)
model1 <- glm(fm1, data = n_train,
              family = "binomial",
              na.action = na.exclude)
summary(model1)


#include flu shot
fm2 <- covid_vaccine ~ gender + age_group +
  education + race + financial_status + flu_shot

#estimate model 2 (logistic regression)
model2 <- glm(fm2, data = n_train,
              family = "binomial",
              na.action = na.exclude)
summary(model2)

#report collinearity of vaccines
cor(demographics$covid_vaccine,
    demographics$flu_shot,
    use = "complete.obs")
##Calculate prediction value on the test set
yhat_test<- predict(model1, newdata = n_test, "response")

#Plot prediction value density of test set
hist(yhat_test, freq = FALSE,
     xlim = c(0, 1), ylim = c(0, 6))

#Create yhat column on the country level model
demographics$yhat <- as.numeric(predict(model1,
                                        newdata = demographics,
                                        "response"))

mean(demographics$yhat, na.rm = TRUE)
#Compute ROC curve
dem.roc1 <- roc(
  response  = n_test$covid_vaccine,
  predictor = yhat_test
)

#Plot ROC curve
plot(dem.roc1,
```

```
     xlab = "False Positive Rate or 1-Specificity",
     ylab = "True Positive Rate or Sensitivity",
     main = "ROC Curve Model 1",
     legacy.axes = TRUE)

#Calculate AUC = 0.703, chance a true case outranks a false case.
auc(dem.roc1)

#Calculate the threshold for positive case from test set
threshold <- coords(dem.roc1, x = "best", transpose = TRUE,
                      ret = c("threshold","accuracy","precision","recall"))

#Report efficiency table of model
knitr::kable(rbind(threshold), row.names = FALSE, digits = 3)

n_test$threshold <- threshold[1]

#Calculate f1-score
f1 <- 2*((0.91*0.617)/(0.91+0.617))
f1
#Use boolean logic to determine model v. actual prediction value
n_test$yhat_actual <- as.numeric(yhat_test >= n_test$threshold)

#Create prop table of model efficiency
prop_table_covid <- prop.table(table(n_test$yhat_actual,
                                     n_test$covid_vaccine))
rownames(prop_table_covid) <- c("Deny Vaccine (Predicted)",
                                "Seek Vaccine (Predicted)")
colnames(prop_table_covid) <- c("Deny Vaccine (True)",
                                "Seek Vaccine (True)")
prop_table_covid

paste("False negative rate is 0.312 (top right)")
paste("False positive rate is 0.0496 (bottom left)")

##Model 2 calculate prediction value on the test set
yhat_test2<- predict(model2,
                     newdata = n_test,
                     "response")

#Plot prediction value density of test set
hist(yhat_test2, freq = FALSE,
     xlim = c(0, 1), ylim = c(0, 6))

#Create yhat column on the country level model
demographics$yhat2 <- predict(model2,
                              newdata = demographics,
                              "response")

mean(demographics$yhat2, na.rm = TRUE)
##Compute ROC curve Model 2
dem.roc2 <- roc(
  response  = n_test$covid_vaccine,
```

```
  predictor = yhat_test2
)

#Plot ROC curve
plot(dem.roc2,
     xlab = "False Positive Rate or 1-Specificity",
     ylab = "True Positive Rate or Sensitivity",
     main = "ROC Curve Model 2",
     legacy.axes = TRUE)

#Calculate AUC = 0.7902, chance a true case outranks a false case.
auc(dem.roc2)

##Model 2 calculate the threshold for positive case from test set
threshold_mod2 <- coords(dem.roc2, x = "best",
                         transpose = TRUE,
                       ret = c("threshold","accuracy",
                               "precision","recall"))

#Report efficiency table of model
knitr::kable(rbind(threshold_mod2),
             row.names = FALSE, digits = 3)

n_test$threshold_mod2 <- threshold_mod2[1]

#Calculate f1-score
f1_mod2 <- 2*((0.915*0.81)/(0.915+0.81))
f1_mod2


#Use boolean logic to determine model v. actual prediction value
n_test$yhat_actual_mod2 <- as.numeric(yhat_test2 >= n_test$threshold_mod2)

#Create prop table of model efficiency
prop_table_covid_mod2 <- prop.table(table(n_test$yhat_actual_mod2,
                                          n_test$covid_vaccine))
rownames(prop_table_covid_mod2) <- c("Deny Vaccine (Predicted)",
                                     "Seek Vaccine (Predicted)")
colnames(prop_table_covid_mod2) <- c("Deny Vaccine (True)",
                                     "Seek Vaccine (True)")
prop_table_covid_mod2

paste("False negative rate is 0.154 (top right)")
paste("False positive rate is 0.0609 (bottom left)")

# Effect of age
effect_plot(model1, pred = age_group, colors = "#7C72A0") +
         theme_classic() +
         labs(title = "Vaccine Acceptance and Age",
              subtitle = "Predicted change in acceptance of vaccine by age group",
              x = "Age",
              y = "Acceptance of Vaccine") +
           theme(plot.title = element_text(face = "bold",
```

```
                                        size = 18),
              plot.subtitle = element_text(size = 13),
              axis.text.x = element_text(size = 13),
              axis.text.y = element_text(size = 13),
              axis.title.x = element_text(size = 16,
                                          margin = margin(t = 20)),
              axis.title.y = element_text(size = 16,
                                          margin = margin(r = 15))) +
         scale_x_continuous(breaks = c(1, 2, 3, 4, 5),
                            labels = c("18-24", "25-44", "45-60", "61-70", ">70"))+
  ylim(0.87, 0.95)
#Effect of financial status
effect_plot(model1, pred = financial_status, colors = "#545643") +
         theme_classic() +
         labs(title = "Vaccine Acceptance and Financial Status",
              subtitle = "Predicted change in acceptance of vaccine by financial status",
              x = "Financial Status",
              y = "Acceptance of Vaccine") +
           theme(plot.title = element_text(face = "bold",
                                        size = 18),
              plot.subtitle = element_text(size = 13),
              axis.text.x = element_text(size = 13),
              axis.text.y = element_text(size = 13),
              axis.title.x = element_text(size = 16,
                                          margin = margin(t = 20)),
              axis.title.y = element_text(size = 16,
                                          margin = margin(r = 15))) +
         scale_x_continuous(breaks = c(1, 2, 3, 4),
                            labels = c("Lower Middle Class/Poor",
                                       "Middle Class", "Upper Middle Class",
                                       "Wealthy")) +
 ylim(0.87, 0.95)
# Effect of education
effect_plot(model1, pred = education, colors = "#E8998D") +
         theme_classic() +
         labs(title = "Vaccine Acceptance and Education",
              subtitle = "Predicted change in acceptance of vaccine by education level",
              x = "Education Level",
              y = "Acceptance of Vaccine") +
           theme(plot.title = element_text(face = "bold",
                                        size = 18),
              plot.subtitle = element_text(size = 13),
              axis.text.x = element_text(size = 13),
              axis.text.y = element_text(size = 13),
              axis.title.x = element_text(size = 16,
                                          margin = margin(t = 20)),
              axis.title.y = element_text(size = 16,
                                          margin = margin(r = 15))) +
         scale_x_continuous(breaks = c(1, 2, 3, 4, 5, 6, 7),
                            labels = c("Middle School", "High School",
                                       "Undergraduate", "Graduate",
                                       "Masters", "Doctorate", "Professional")) +
ylim(0.87, 0.95)
```

```r
# Effect of race
effect_plot(model1, pred = race, colors = "#627C85") +
         theme_classic() +
         labs(title = "Vaccine Acceptance and Race",
              subtitle = "Predicted change in acceptance of vaccine by race",
              x = "Race",
              y = "Acceptance of Vaccine") +
           theme(plot.title = element_text(face = "bold",
                                           size = 18),
                 plot.subtitle = element_text(size = 13),
                 axis.text.x = element_text(size = 13),
                 axis.text.y = element_text(size = 13),
                 axis.title.x = element_text(size = 16,
                                             margin = margin(t = 20)),
                 axis.title.y = element_text(size = 16,
                                             margin = margin(r = 15)))
# Effect of gender
effect_plot(model1, pred = gender, colors = "#DB6C79") +
         theme_classic() +
         labs(title = "Vaccine Hesitancy and Gender",
              subtitle = "Predicted change in acceptance of vaccine by gender",
              x = "Gender",
              y = "Acceptance of Vaccine") +
           theme(plot.title = element_text(face = "bold",
                                           size = 18),
                 plot.subtitle = element_text(size = 13),
                 axis.text.x = element_text(size = 13),
                 axis.text.y = element_text(size = 13),
                 axis.title.x = element_text(size = 16,
                                             margin = margin(t = 20)),
                 axis.title.y = element_text(size = 16,
                                             margin = margin(r = 15)))

# standardising new dataset var names
names(vaccine_data)[names(vaccine_data) == "Gender_string"] <- "gender"
names(vaccine_data)[names(vaccine_data) == "your_age"] <- "age_group"
names(vaccine_data)[names(vaccine_data) == "your_education"] <- "education"
names(vaccine_data)[names(vaccine_data) == "your_race"] <- "race"

yhat_vd<- predict(model1,
                  newdata = vaccine_data, "response")

#generating y_hat on larger dataset
vaccine_data$yhat <- as.numeric(predict(model1,
                                        newdata = vaccine_data,
                                        "response"))

mean(vaccine_data$yhat, na.rm = TRUE)
#Establish measure of public health responsiveness
vaccine_data$resistph <- as.numeric(vaccine_data$ld >= 4 &
                                      vaccine_data$fm >= 4 |
                                      vaccine_data$ld >= 4 &
                                      vaccine_data$reopening <= 2 |
```

```
                                    vaccine_data$reopening <= 2 &
                                    vaccine_data$fm >= 4)

mean(vaccine_data$resistph,na.rm=TRUE)

#Compute ROC curve
dem.roc3 <- roc(
  response  = vaccine_data$resistph,
  predictor = yhat_vd
)

#Plot ROC curve
plot(dem.roc3,
     xlab = "False Positive Rate or 1-Specificity",
     ylab = "True Positive Rate or Sensitivity",
     main = "ROC Curve (Our Model) for Anti-Public Health restrictions",
     legacy.axes = TRUE)

#Calculate AUC
auc(dem.roc3)

#Calculate the threshold for positive case from test set
threshold_rph <- coords(dem.roc3, x = "best", transpose = TRUE,
                        ret = c("threshold","accuracy","precision","recall"))

#Report efficiency table of model
knitr::kable(rbind(threshold_rph), row.names = FALSE, digits = 3)
```