COVID-19 Vaccination Rates in the United States

Kristine Villaluna

21/01/2022

Vaccinations

This analysis relates to COVID-19 vaccination rates in the United States. We are interested in exploring factors that are associated with differences in vaccine coverage by US county.

Data Description

The latest data on vaccination coverage (and the data dictionary) can be downloaded here: https://data.cdc.gov/Vaccinations/COVID-19-Vaccinations-in-the-United-States-County/8xkx-amqh/data.

For this analysis, we will consider data from the 15th of January. Note that we will be interested in people who have had at least two vaccinations, which refers to columns that have the Series_Complete prefix.

We also have a dataset acs that contains a range of different demographic, socioeconomic, and health variables by county. These were obtained from the American Community Survey (ACS) via the R package tidycensus.

EDA

After reading in the vaccination and ACS data, the two data sets were merged by `fips` and EDA performed.

First, a quick check was done to see the minimum and maximum counts of fully vaccinated individuals 18+ by each state was done to see the differences in vaccinations between counties in the same state. The table is as follows:

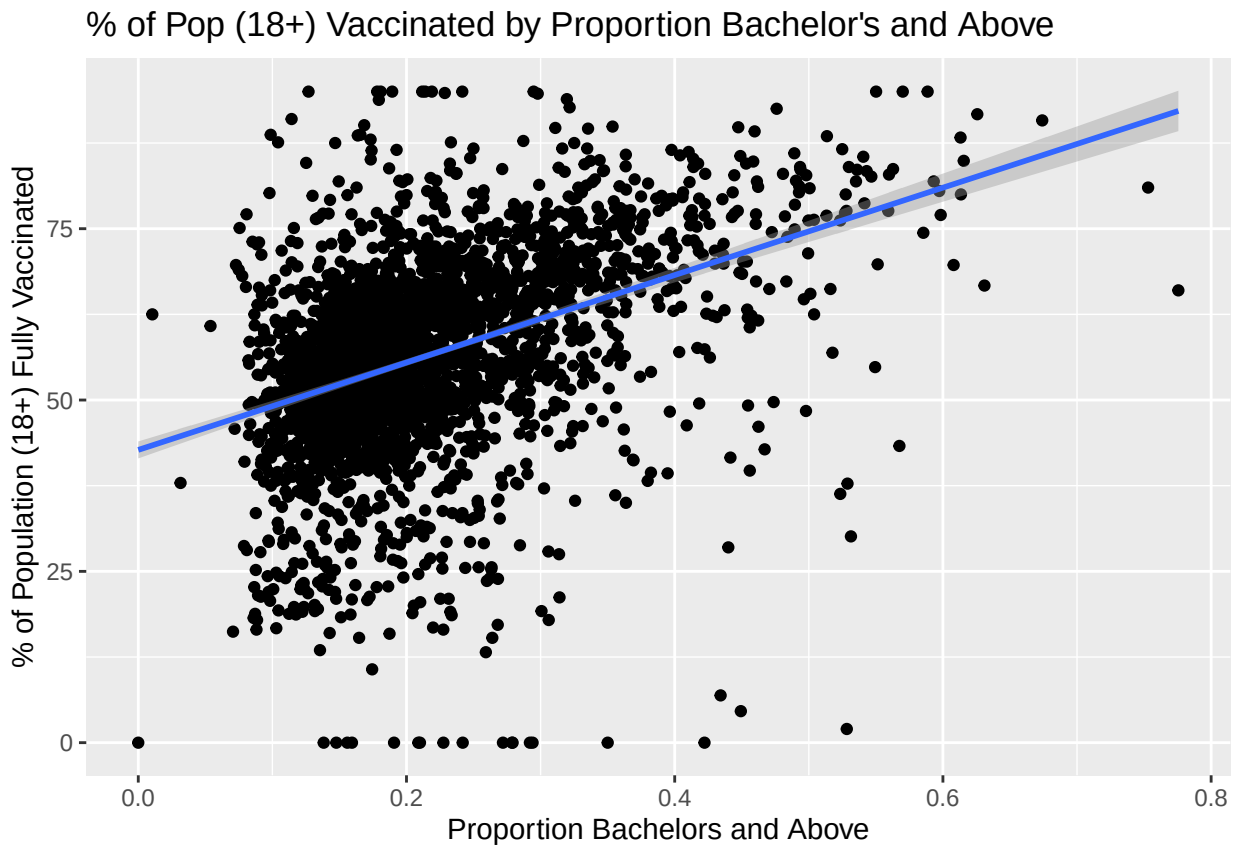
recip_state	Max18Plus	Min18Plus
AK	158115	394
AL	336970	3690
AR	195216	1751
AS	26925	26925
AZ	2219225	3712
CA	6143055	0
CO	480468	479
CT	646279	42879
DC	429556	21658
DE	328419	22773
FL	2052183	2693
FM	42264	42264
GA	2111788	302
GU	91142	22332
HI	0	0
IA	282910	1870
ID	263542	304
IL	3130832	1301

recip_state	Max18Plus	Min18Plus
IN	471631	3098
KS	379432	529
KY	427391	744
LA	251501	1185
MA	1085205	180
MD	770994	11716
ME	223038	9467
MH	21646	21646
MI	837092	1175
MN	800364	1862
MO	542944	780
MP	33176	33176
MS	116368	449
MT	74437	143
NC	679086	1793
ND	96160	79
NE	316599	68
NH	234731	19188
NJ	618456	30274
NM	408347	338
NV	1128542	336
NY	1671155	3436
OH	729964	4714
OK	443036	646
OR	562274	686
PA	933700	2284
PR	203691	1559
PW	14970	14970
RI	390461	32337
SC	253255	4595
SD	110553	233
TN	450322	1331
TX	2557110	0
UT	664316	542
VA	1007196	0
VI	21386	2650
VT	114400	2213
WA	1559027	784
WI	521926	2237
WV	104921	2220
WY	47474	700

Here we can see some states like Hawaii (HI) with 0 for their counts. Upon further inspection, we find that this is because there is no county level data available for the state of Hawaii. Moreover, we can see that there are some states like Texas (TX) and Virginia (VA) which have counties with 0 counts.

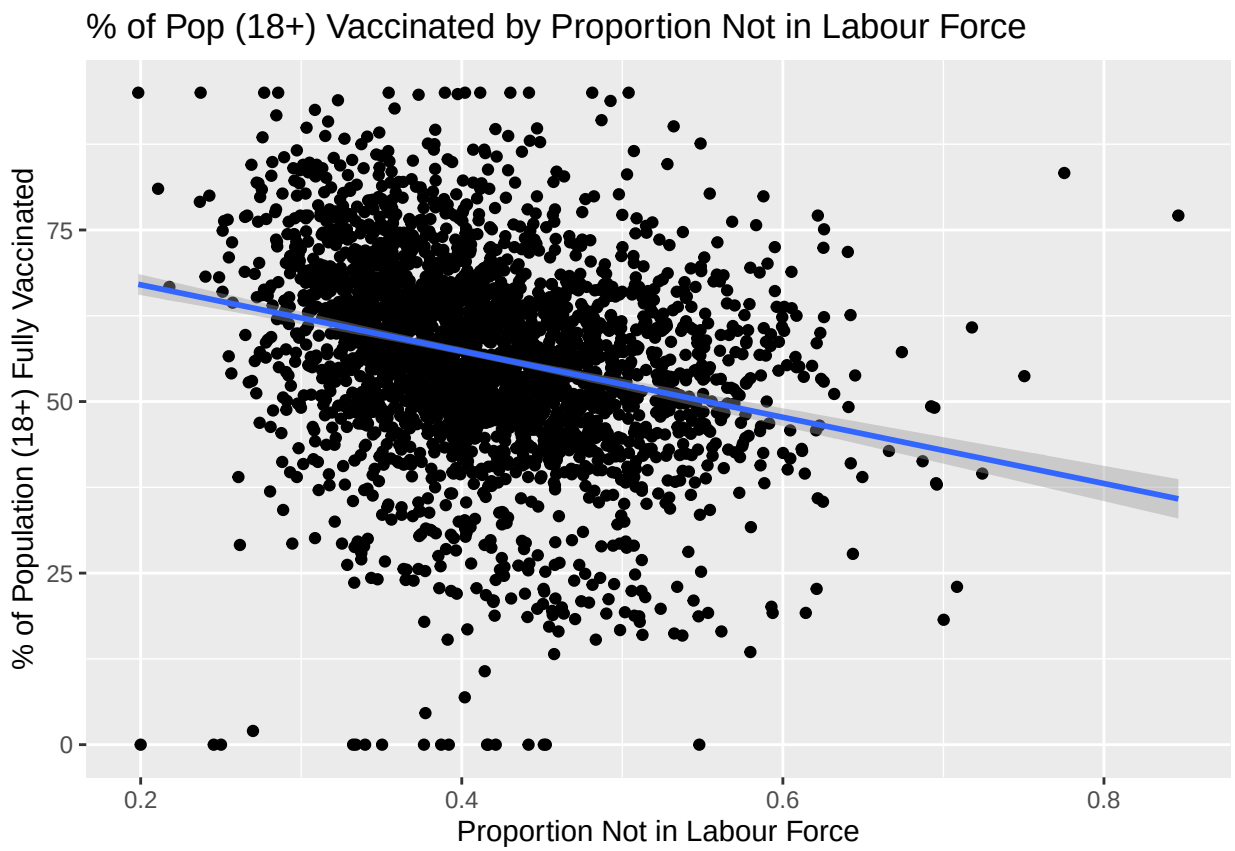
Next, we plotted some potential explanatory variables to see the distributions and potentially motivate our model building.

Education - Bachelor's and Above



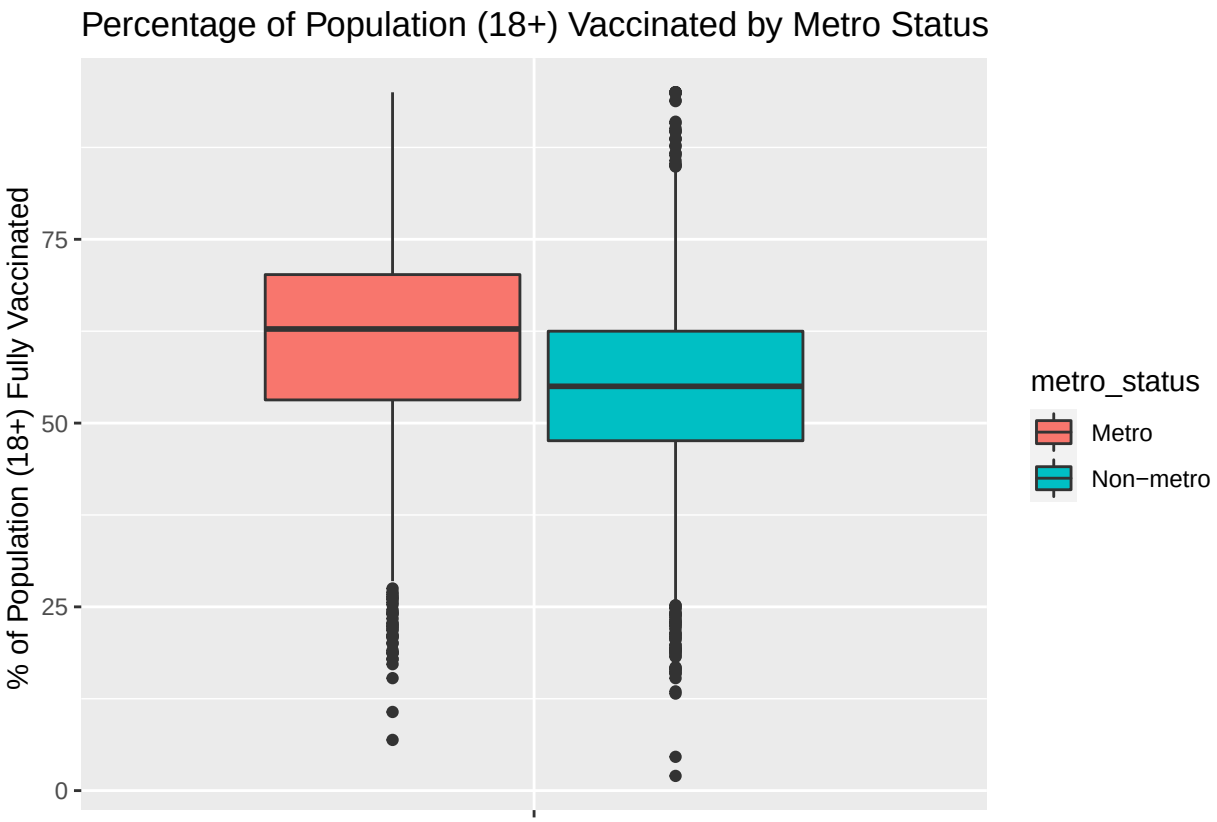
Here we can see a positive relationship, indicating that the higher the proportion of the population with a Bachelor's degree or higher, the higher the percent of the population fully vaccinated. This is not surprising as this is something you may intuitively suspect.

Not in Labour Force



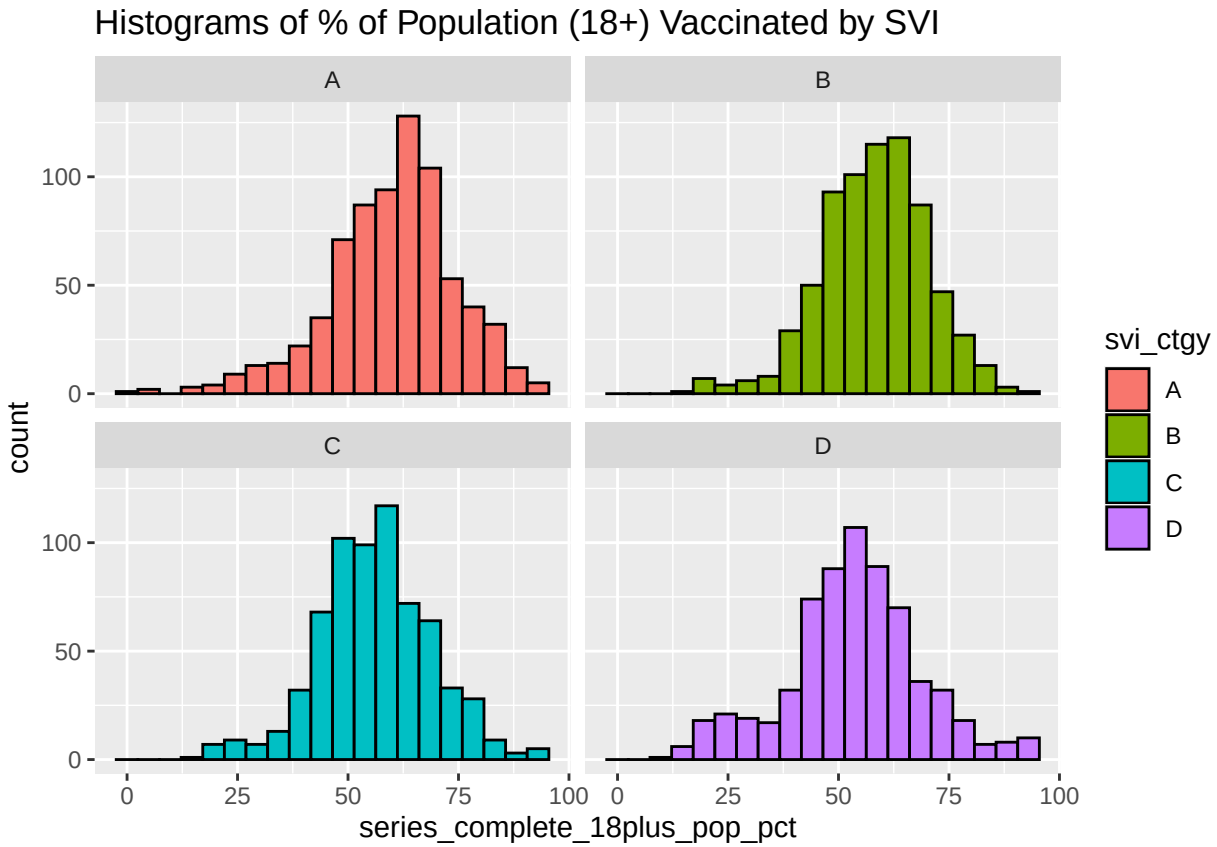
In this plot we see a negative relationship between the proportion of the population not in the labour force and the percentage of the 18+ population that is fully vaccinated. This means that the higher the proportion of the population that is not in the labour force, the lower the percentage of the 18+ population that is fully vaccinated.

Metro Status



In this plot we can see that the percentage of the 18+ population that is fully vaccinated is a bit higher in metro areas versus non-metro areas. The mean appears to be approximately 60% in the metro areas, and approximately 55% in non-metro areas.

Social Vulnerability Index



Finally, here we see a plot including four histograms faceted by the CDC's Social Vulnerability Index (SVI). The CDC defines this as the potential negative effects on communities caused by external stresses on human health. Such stresses include natural or human-caused disasters, or disease outbreaks. Here we can see that the categories A and B have slightly higher percentage of 18+ people vaccinated as we can see from the peaks.

For reference, the rank categories are: A = 0– 0.25 SVI rank, B = 0.2501–0.50 SVI rank, C = 0.5001–0.75 SVI rank, and D = 0.7501–1.0 SVI rank.

Additional EDA plots are available in the .Rmd file.

Analysis

Next, we would like to build a regression model at the county level to help investigate patterns in the full vaccination rate for the population aged 18+ (that is, people aged 18+ who have received at least two vaccines).

For this analysis, I will consider modeling the outcome measure as a proportion of the population aged 18+ who are fully vaccinated by county using a binomial model, where n is equal to the total population count in the county, y is the number of people 18+ who are fully vaccinated, and p is the proportion of the population aged 18+ who are fully vaccinated by county.

To calculate the proportion of the population aged 18+ who are fully vaccinated by county, I took the number of people aged 18+ who have received two doses, and divided it by the total population 18+. The summary of this new variable is as follows:

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##  0.0000  0.4895  0.5728  0.5717  0.6604  4.6954     457
```

Here we can see that there is something strange going on with the maximum, as proportions should be bounded by 0 and 1. Upon further inspection, there were 8 observations where there was a higher number of people vaccinated than there was total population.

For example, Chattahoochee County, GA had a value of 39042 for the variable `series_complete_18plus`, but only had a total population of 8315 from the ACS data. For the purpose of this analysis, these 8 observations were filtered out.

Next, we did a quick check of a correlation matrix to see how correlated the potential explanatory variables are. This is done as we would like to avoid adding correlated variables into our model which can produce multicollinearity. The code for this correlation matrix is available in the .Rmd file.

As expected, we see some high correlations (0.6+) with some variables such as proportion of bachelor above and median income, and proportion with less than high school and low income-poverty ratio.

Model Building

Model 1 - Binomial

For our first model, we can try and look at a larger range of covariates. Here, we will first try proportion foreign born, median age, proportion bachelor above, proportion not in labour force, proportion with health insurance, SVI, metro status, and proportion low IP ratio. This is just so we can first get a sense of the data.

Please note that the output has been suppressed for the PDF, please see .Rmd file for model summary.

Looking at the model fitted with all of the mentioned covariates, we can see that all of the variables are significant at the 0.05 level. This is a bit suspicious so we can potentially remove some variables to see how this affects the model.

Model 2 - Binomial

Based on our EDA, we can keep only the ones we suspect to be important, namely, proportion foreign born, proportion with bachelor or above, proportion not in labour force, proportion with health insurance, metro status, and SVI index. The resulting binomial model is available in the .Rmd file.

Again, we are seeing very small p-values, and all predictors are statistically significant. This can be indicative of over-dispersion. We can also see that the residual deviance is very high compared to the degrees of freedom. One way to try and combat this is to use a **quasi-binomial** model instead.

Model 3 - Quasi-Binomial

We then fit a quasi-binomial model using the same covariates from the previous binomial model. The summary is available in the .Rmd file.

In this model we can see that proportion not in labour force and one level of SVI are not as significant as the other variables, as they are only significant at the 0.1 level. It is important to note that when we are working with categorical variables such as SVI, that we must treat the variable as a whole, and should not remove one level from the model and not the others.

We can use the `drop1()` and `anova()` functions with the likelihood ratio test (LRT) to see which variables we can potentially remove from our model.

Using the `drop1()` function:

```
## Single term deletions
##
## Model:
## cbind(series_complete_18plus, total_pop_18plus - series_complete_18plus) ~
##     prop_foreign_born + prop_bachelor_above + prop_nilf + prop_health_insurance +
##         metro_status + svi_ctgy
##                         Df Deviance scaled dev.  Pr(>Chi)
## <none>                     11558911
## prop_foreign_born       1 13989673      560.14 < 2.2e-16 ***
## prop_bachelor_above     1 11994035      100.27 < 2.2e-16 ***
## prop_nilf               1 11574823        3.67 0.0555096 .
## prop_health_insurance   1 11988896       99.09 < 2.2e-16 ***
## metro_status            1 11681569       28.27 1.058e-07 ***
## svi_ctgy                3 11633340       17.15 0.0006578 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Using the `anova()` function:

```
## Analysis of Deviance Table
##
## Model: quasibinomial, link: logit
##
## Response: cbind(series_complete_18plus, total_pop_18plus - series_complete_18plus)
##
## Terms added sequentially (first to last)
##
##
##                        Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                                    2816   20128860
## prop_foreign_born       1  5657881      2815   14470979 < 2.2e-16 ***
## prop_bachelor_above     1  2199862      2814   12271118 < 2.2e-16 ***
## prop_nilf               1      190      2813   12270928 0.8342376
## prop_health_insurance   1   534183      2812   11736744 < 2.2e-16 ***
## metro_status            1   103404      2811   11633340 1.053e-06 ***
## svi_ctgy                3    74429      2808   11558911 0.0006578 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Looking at the output of both the drop1 and ANOVA functions, we can see that we can remove the variable `prop_nilf`.

Model 4 - Quasi-Binomial

The resulting model with `prop_nilf` removed is as follows:

```
##
## Call:
## glm(formula = cbind(series_complete_18plus, total_pop_18plus -
##     series_complete_18plus) ~ prop_foreign_born + prop_bachelor_above +
##     prop_health_insurance + metro_status + svi_ctgy, family = "quasibinomial",
##     data = vax)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1545.31    -18.00     -1.61     16.77    385.75
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)            -2.47826    0.26165  -9.472  < 2e-16 ***
## prop_foreign_born       2.75705    0.12274  22.462  < 2e-16 ***
## prop_bachelor_above     1.39302    0.13942   9.991  < 2e-16 ***
## prop_health_insurance   2.83101    0.28092  10.077  < 2e-16 ***
## metro_statusNon-metro -0.15280    0.03041  -5.025 5.36e-07 ***
## svi_ctgyB             -0.10375    0.03125  -3.320  0.00091 ***
## svi_ctgyC             -0.10630    0.03299  -3.222  0.00129 **
## svi_ctgyD             -0.05738    0.04150  -1.383  0.16682
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasibinomial family taken to be 4342.381)
##
##     Null deviance: 20128860  on 2816  degrees of freedom
## Residual deviance: 11574823  on 2809  degrees of freedom
##   (1 observation deleted due to missingness)
## AIC: NA
##
## Number of Fisher Scoring iterations: 4
```

Here we can see everything is significant except for category D of SVI.

After checking the `drop1()` and `anova()` functions, we now see that all the variables left in the model are significant using the LRT tests. Now that we have sufficiently chosen our main effects, let us try some interactions.

Model 5 - Quasi-Binomial with Interactions

Now that we have the main effects chosen, we can consider adding interaction terms into the model. For this analysis, I will only consider two-way interaction terms. The summary is available in the .Rmd file.

After examining all of the second-order interaction terms using the `drop1()` function, we can keep interactions between the proportion foreign born and proportion health insurance, metro status and SVI, as well as interactions between SVI and proportion of bachelor above and health_insurance.

Model 6 - Final Model

Therefore, the final model including interactions is as follows:

```
##
## Call:
## glm(formula = cbind(series_complete_18plus, total_pop_18plus -
##     series_complete_18plus) ~ prop_foreign_born + prop_bachelor_above +
##     prop_health_insurance + metro_status + svi_ctgy + prop_foreign_born:prop_health_insurance +
##     prop_foreign_born:metro_status + prop_foreign_born:svi_ctgy +
##     prop_bachelor_above:svi_ctgy + prop_health_insurance:svi_ctgy,
##     family = "quasibinomial", data = vax)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1381.94    -18.89     -1.55     16.82    373.64
##
## Coefficients:
##                                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)                              -7.84484    0.91715  -8.553  < 2e-16
## prop_foreign_born                        13.03301    1.96517   6.632 3.96e-11
## prop_bachelor_above                       2.08029    0.29718   7.000 3.19e-12
## prop_health_insurance                     8.32002    0.98325   8.462  < 2e-16
## metro_statusNon-metro                    -0.04890    0.03615  -1.353 0.176162
## svi_ctgyB                                 3.83139    1.02884   3.724 0.000200
## svi_ctgyC                                 2.19575    0.98074   2.239 0.025243
## svi_ctgyD                                 4.98841    0.99085   5.034 5.10e-07
## prop_foreign_born:prop_health_insurance -11.70419    2.05517  -5.695 1.36e-08
## prop_foreign_born:metro_statusNon-metro  -1.91278    0.46301  -4.131 3.72e-05
## prop_foreign_born:svi_ctgyB              -2.39399    0.49370  -4.849 1.31e-06
## prop_foreign_born:svi_ctgyC               0.65542    0.45161   1.451 0.146808
## prop_foreign_born:svi_ctgyD               0.78073    0.46639   1.674 0.094244
## prop_bachelor_above:svi_ctgyB             1.39248    0.38467   3.620 0.000300
## prop_bachelor_above:svi_ctgyC            -1.43348    0.36832  -3.892 0.000102
## prop_bachelor_above:svi_ctgyD            -1.34686    0.47172  -2.855 0.004332
## prop_health_insurance:svi_ctgyB          -4.34310    1.10886  -3.917 9.19e-05
## prop_health_insurance:svi_ctgyC          -1.92556    1.06312  -1.811 0.070212
## prop_health_insurance:svi_ctgyD          -5.11149    1.08491  -4.711 2.58e-06
##
## (Intercept)                              ***
## prop_foreign_born                        ***
## prop_bachelor_above                      ***
## prop_health_insurance                    ***
## metro_statusNon-metro
## svi_ctgyB                                ***
## svi_ctgyC                                *
## svi_ctgyD                                ***
## prop_foreign_born:prop_health_insurance ***
## prop_foreign_born:metro_statusNon-metro ***
## prop_foreign_born:svi_ctgyB              ***
## prop_foreign_born:svi_ctgyC
## prop_foreign_born:svi_ctgyD              .
## prop_bachelor_above:svi_ctgyB            ***
## prop_bachelor_above:svi_ctgyC            ***
## prop_bachelor_above:svi_ctgyD            **
```

```
## prop_health_insurance:svi_ctgyB         ***
## prop_health_insurance:svi_ctgyC         .
## prop_health_insurance:svi_ctgyD         ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasibinomial family taken to be 3797.125)
##
##     Null deviance: 20128860  on 2816  degrees of freedom
## Residual deviance: 10593771  on 2798  degrees of freedom
##   (1 observation deleted due to missingness)
## AIC: NA
##
## Number of Fisher Scoring iterations: 5
```

The way to interpret the coefficients of a quasi-binomial model is similar to the way you would for a normal logistic regression. We must exponentiate the coefficients to produce an odds ratio for success (odds of being fully vaccinated).

Take for example, metro status. Taking, the exponential of the coefficient, we get:

```
exp(-0.04890)
```

```
## [1] 0.9522764
```

```
1-exp( -0.04890)
```

```
## [1] 0.04772365
```

Which means that the odds of being fully vaccinated in non-metro areas is 0.95 times that of the odds in metro areas, i.e., approximately 5% less in non-metro areas than metro areas.

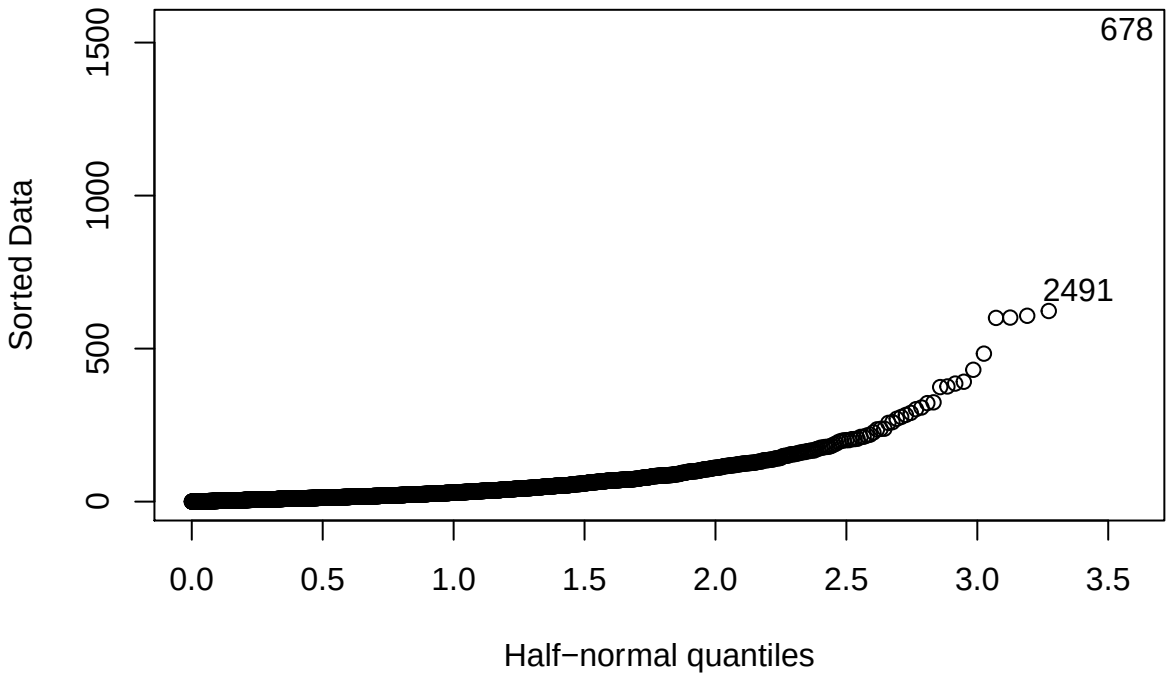
Finally, we can check that all variables are significant using the LRT test one last time.

```
## Single term deletions
##
## Model:
## cbind(series_complete_18plus, total_pop_18plus - series_complete_18plus) ~
##     prop_foreign_born + prop_bachelor_above + prop_health_insurance +
##         metro_status + svi_ctgy + prop_foreign_born:prop_health_insurance +
##         prop_foreign_born:metro_status + prop_foreign_born:svi_ctgy +
##         prop_bachelor_above:svi_ctgy + prop_health_insurance:svi_ctgy
##                                        Df Deviance scaled dev.  Pr(>Chi)
## <none>                                    10593771
## prop_foreign_born:prop_health_insurance  1 10718059      32.732 1.058e-08 ***
## prop_foreign_born:metro_status           1 10656913      16.629 4.545e-05 ***
## prop_foreign_born:svi_ctgy               3 10926087      87.518 < 2.2e-16 ***
## prop_bachelor_above:svi_ctgy             3 10912163      83.851 < 2.2e-16 ***
## prop_health_insurance:svi_ctgy           3 10740854      38.735 1.975e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Here we can see that all variables are significant and we can now move on to model checking.

Model Checking

First, a half-norm plot can help us look for outliers and influential points.

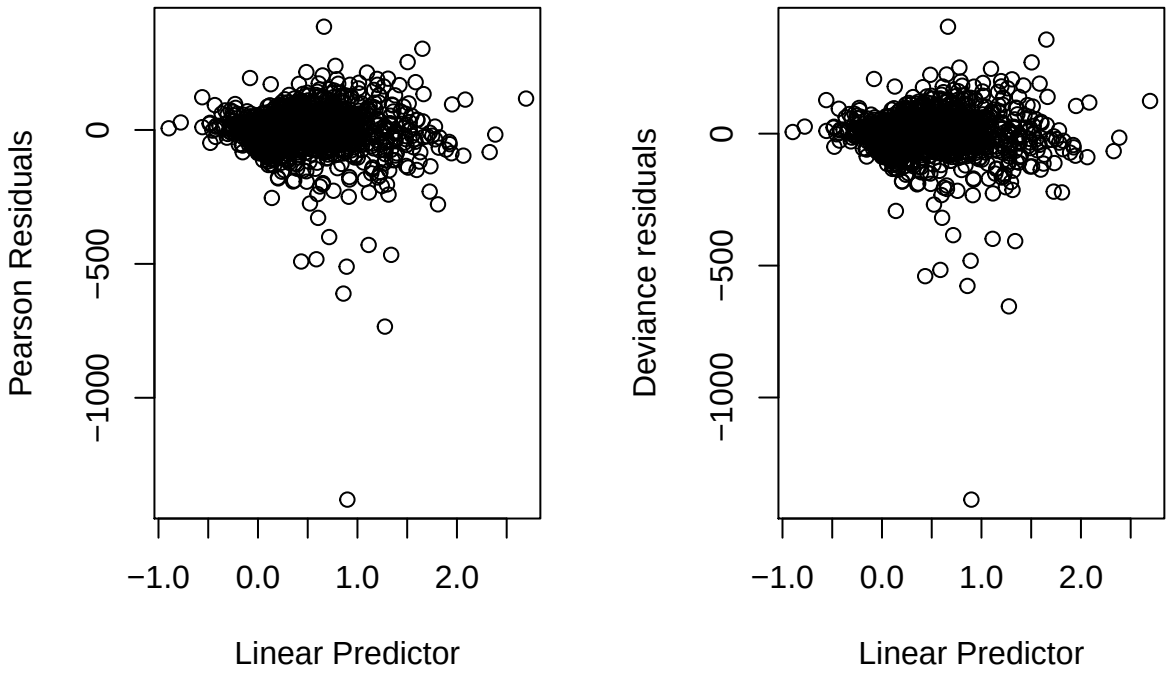


Here we see that obs 678 is quite influential, we can consider a fit without this case.

After removing that influential point, the model is refit as previously specified.

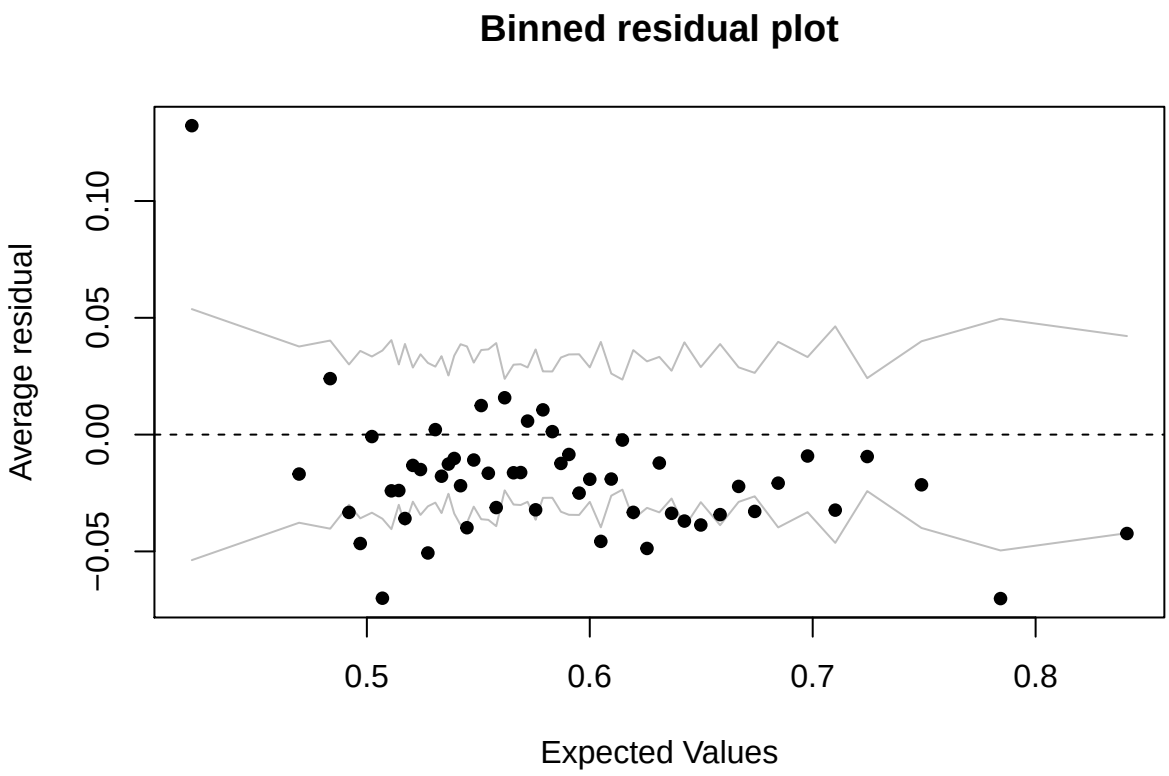
Residual Plots

Next, we can take a look at some residual plots.



On the left, we can see the linear predictor plotted against the Pearson residuals, and on the right the linear predictor plotted against the deviance residuals. Both plots are similar and we can see that the residuals tend to scatter around 0 with a number of outliers.

Furthermore, we can look at the binned residual plot:



The grey lines represent +- 2 SE bands, which we would expect to contain about 95% of the observations. We say that a model is reasonable if the majority of the fitted values seem to fall within the SE bands.

The model does not look reasonable, as there are more outliers among the residuals than we would expect from chance alone (alpha = .05). It looks like the model does not do well when the fitted value is between 0.5 and 0.6. This is indicative of bad model fit.

Summary

For this analysis, I tried to consider modeling the outcome measure as a proportion of the population aged 18+ who are fully vaccinated by county using a binomial model, where n was equal to the total population count in the county and y was the number of people 18+ who are fully vaccinated. Assuming that there is independence in the population, we try and get the probability, which is then an unbiased estimate of the proportion of 18+ vaccinated.

First, I tried fitting a binomial model. I then found there to be over-dispersion in the data. To remedy this, I then tried a quasi-binomial model. Using the LRT test in the `drop1()` and `anova()` functions, I then reduced my model down to main effects. After that, I tested two way interaction terms which I could include into the model.

The final model included the following variables: proportion foreign born, proportion bachelor above, proportion health insurance, metro status, SVI and interactions between the proportion foreign born and proportion health insurance, metro status and SVI, as well as interactions between SVI and proportion of bachelor above and health_insurance.

After looking at the diagnostic and residual plots, it does not appear that my model was very good. In future work, some other variables that may be of interest to investigate are different indicators of health status such as information about whether an individual is immuno-compromised or not, and social demographic information such as household composition, martial status, and ethnic/diversity index.