STA2201 Assignment #1

Timothy Jordan Regis

2023-02-05

Q1

a)

$$Y|\theta \sim Pois(\mu\theta)$$

Thus,

$$\mathbb{E}[Y] = \mathbb{E}[\mathbb{E}[Y|\theta]] = \mathbb{E}[\mu\theta] = \mu\mathbb{E}[\theta] = \mu$$

And,

$$\mathbb{V}[Y] = \mathbb{E}[\mathbb{V}[Y|\theta]] + \mathbb{V}[\mathbb{E}[Y|\theta]] = \mathbb{E}[\mu\theta] + \mathbb{V}[\mu\theta] = \mu + \mu^2\sigma^2 = \mu(1 + \mu\sigma^2)$$

b)

$$\theta \sim Gamma(\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)}\theta^{\alpha-1}e^{-\beta\theta}$$

$$Y|\theta \sim Pois(\mu\theta) = \frac{(\mu\theta)^Y e^{-\mu\theta}}{Y!}$$

$$Y \sim Y|\theta \cdot \theta = \int_0^\infty \frac{(\mu\theta)^Y e^{-\mu\theta}}{Y!} \cdot \frac{\beta^\alpha}{\Gamma(\alpha)}\theta^{\alpha-1}e^{-\beta\theta}d\theta$$

$$= \frac{\mu^Y\beta^\alpha}{Y!\Gamma(\alpha)}\int_0^\infty \theta^{Y+\alpha-1}e^{-\theta(\mu+\beta)}d\theta$$

$$= \frac{\mu^Y\beta^\alpha}{Y!\Gamma(\alpha)}\frac{\Gamma(Y+\alpha)}{(\mu+\beta)^{Y+\alpha}}\int_0^\infty \frac{(\mu+\beta)^{Y+\alpha}}{\Gamma(Y+\alpha)}\theta^{Y+\alpha-1}e^{-\theta(\mu+\beta)}d\theta$$

$$= \frac{\Gamma(Y+\alpha)}{Y!\Gamma(\alpha)}\frac{\mu^Y\beta^\alpha}{(\mu+\beta)^{Y+\alpha}}\int_0^\infty \frac{(\mu+\beta)^{Y+\alpha}}{\Gamma(Y+\alpha)}\theta^{Y+\alpha-1}e^{-\theta(\mu+\beta)}d\theta$$

Note:

$$\frac{(\mu+\beta)^{Y+\alpha}}{\Gamma(Y+\alpha)}\theta^{Y+\alpha-1}e^{-\theta(\mu+\beta)} \sim Gamma(Y+\alpha, \mu+\beta)$$

So,

$$\int_0^\infty \frac{(\mu+\beta)^{Y+\alpha}}{\Gamma(Y+\alpha)}\theta^{Y+\alpha-1}e^{-\theta(\mu+\beta)}d\theta = 1$$

Thus,

$$= \frac{\Gamma(Y+\alpha)}{Y!\Gamma(\alpha)}\frac{\mu^Y\beta^\alpha}{(\mu+\beta)^{Y+\alpha}}$$

$$= \frac{(Y+\alpha-1)!}{Y!(\alpha-1)!}\frac{\mu^Y}{(\mu+\beta)^Y}\frac{\beta^\alpha}{(\mu+\beta)^\alpha}$$

$$= \binom{Y+\alpha-1}{Y}\left(\frac{\mu}{\mu+\beta}\right)^Y\left(\frac{\beta}{\mu+\beta}\right)^\alpha$$

$$\sim NB(\alpha, \frac{\beta}{\mu+\beta})$$

c)

Since the mean of the negative binomial is:

$$= \frac{\alpha\frac{\mu}{\mu+\beta}}{\frac{\beta}{\mu+\beta}}$$
$$= \frac{\alpha\mu}{\beta}$$

And the variance of the negative binomial is:

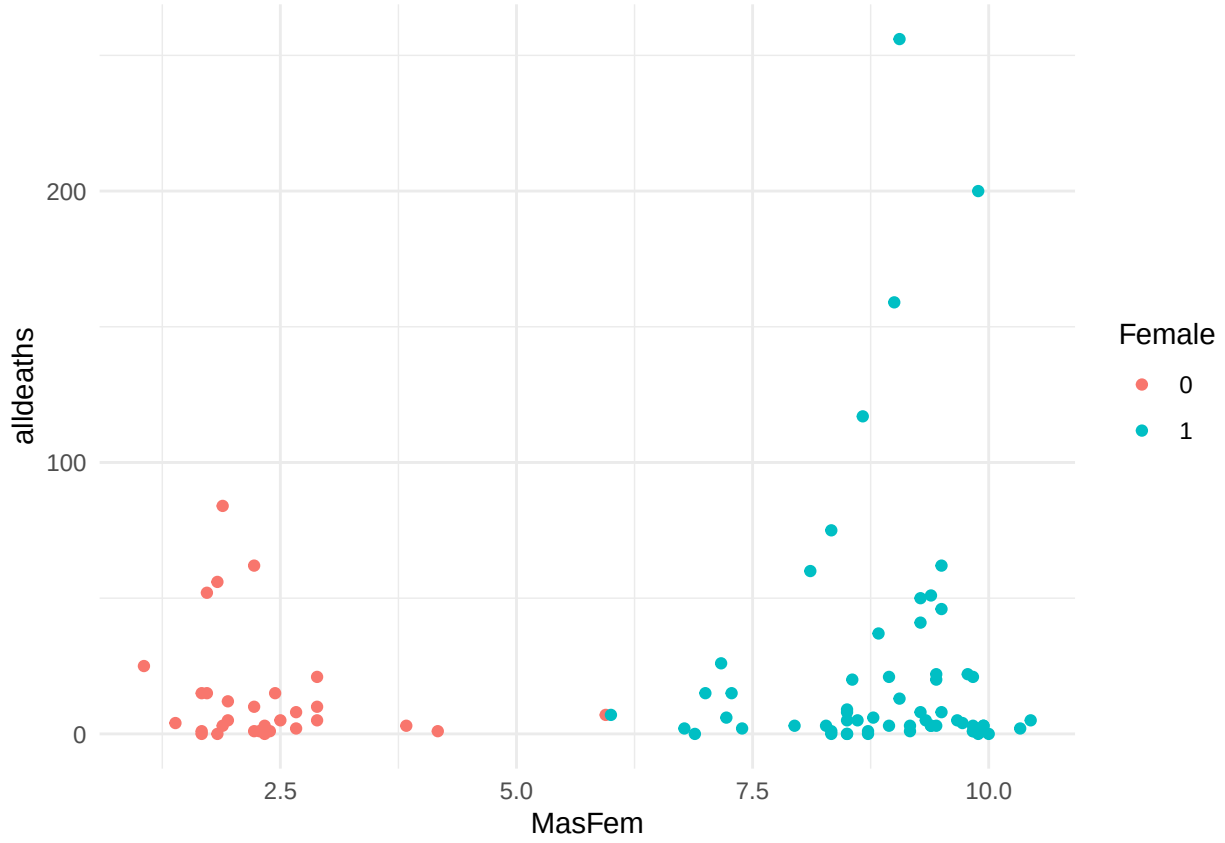
$$\frac{\alpha\frac{\mu}{\mu+\beta}}{(\frac{\beta}{\mu+\beta})^2}$$
$$= \frac{\alpha\mu(\mu+\beta)}{\beta^2}$$
$$\frac{\alpha\mu^2}{\beta^2} + \frac{\alpha\mu}{\beta}$$
$$\sigma^2\mu^2 + \frac{\alpha}{\beta}\mu$$
$$= \mu(\frac{\alpha}{\beta} + \mu\sigma^2)$$

Thus, for the mean to equal μ and variance to equal $\mu(1+\mu\sigma^2)$, we need:

$$\alpha = \beta$$

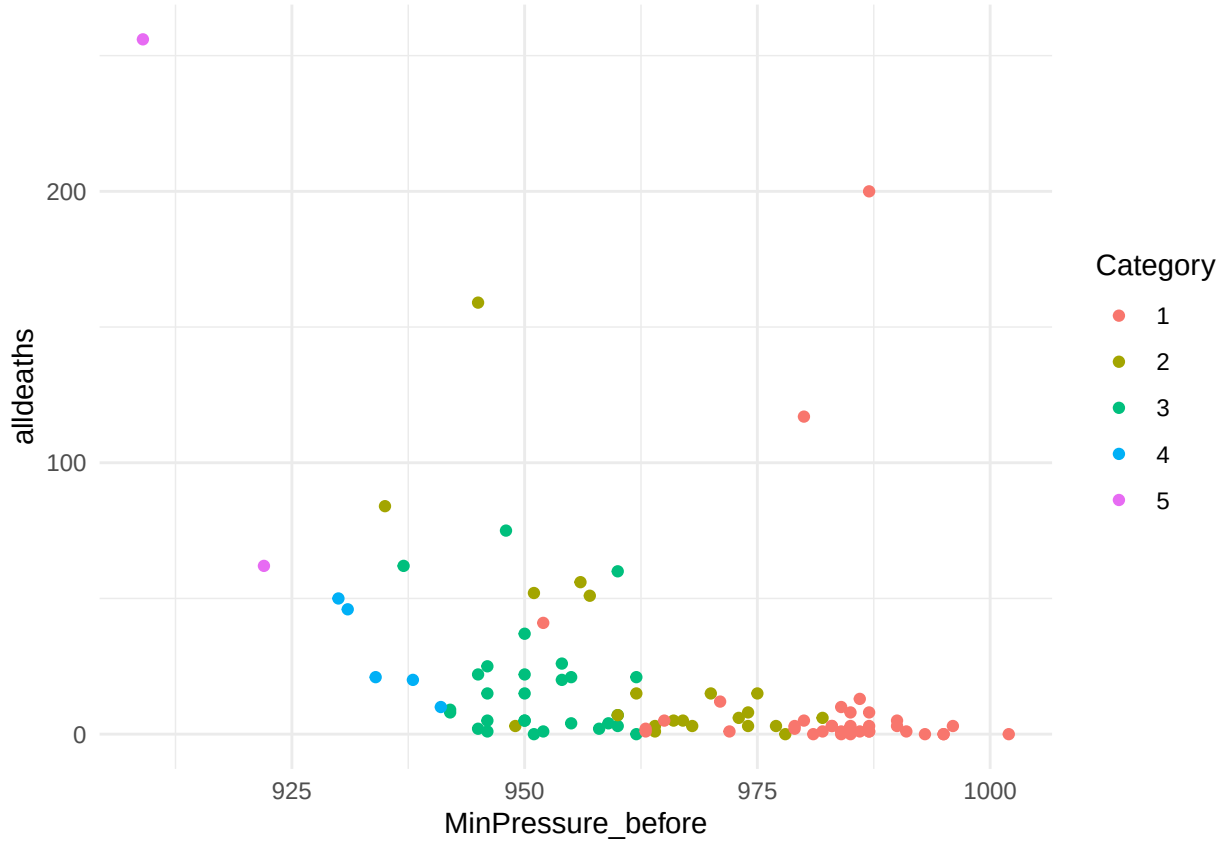
Q2

a)

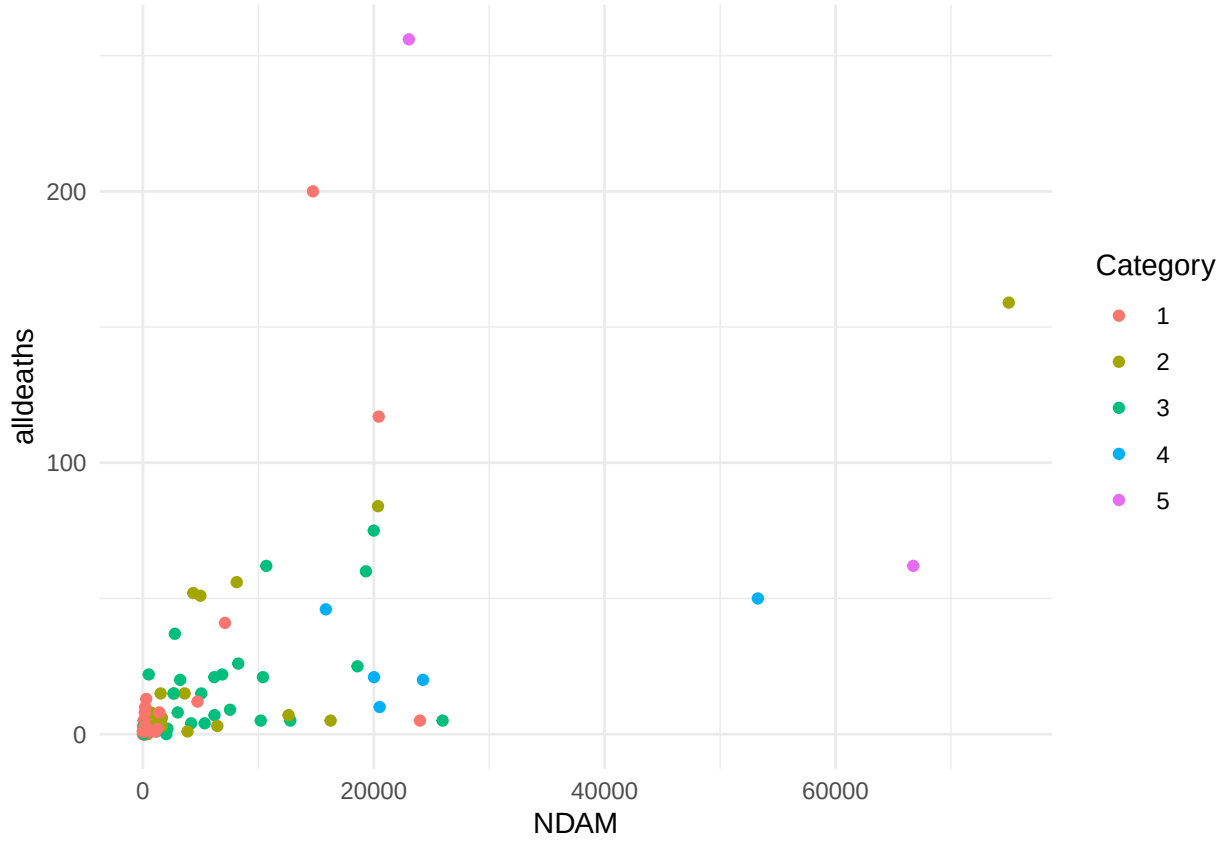


Gender_MF	mean	var
0	14.233	447.840
1	23.758	2253.039

In Figure 1, we have plotted the MFI rating against the number of deaths for all hurricanes, as well as coloring by the actual gender of the storm's name. As we can see, their method of rating the femininity of names was quite successful, with only one 'edge' case in the middle, where the names Ione (Male) and Frances (Female) meet. Otherwise, this did a good job of splitting up male and female names. In its relationship with mortality, we see there is a slight but noticeable change in the left and right halves of the graph. On the left, with male names and low femininity, we see deaths below 100 across all observations, while on the right, with high femininity, we still see a high density of points below 100 again, but there are also some instances, 4 specifically, which extend far beyond 100, up to over 250 deaths. While the averages of the two groups are quite similar, their variances are much different because of this, and we can see that there is a tendency to observe the most deadly storms being those with names of high femininity.



In Figure 2 we plot the minimum pressure against mortality, coloring by the category of the storm. From the graph, we can see there is a very subtle trend in the data, where a majority of the storms with pressure above 970 mb see very few deaths, while those with pressures below 970 display somewhat of a negative relationship between pressure and mortality, with lower pressures gradually increasing the numbers we see. This is to be expected, however, as a hurricane's intensity is inversely related to the pressure within it, with higher pressures being significant of a hurricane losing strength, and we can see this corresponds to the grouping of category 1 storms with higher pressures, and increasing the category rating as we go lower.



Finally, in Figure 3, observing mortality's relationship with damages, we see a naturally high number of storms with very low damages causing very few deaths, which we would expect due to the categorization of hurricanes, where very low ranking storms tend to cause minimal harm but are still considered a hurricane by all important measures. Past this, however, we can see a very loose trend suggesting a positive between damages and deaths, however, we still observe many of the highest damaging hurricanes also leading to very low death counts.

b)

```
##
## Call:
## glm(formula = alldeaths ~ MasFem, family = "poisson", data = q2_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -7.1429  -5.3716  -3.8288  -0.5364  27.4230
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) 2.500370   0.063297  39.502   <2e-16 ***
## MasFem      0.073873   0.007891   9.362   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
```

```
##
##     Null deviance: 4031.9  on 91  degrees of freedom
## Residual deviance: 3937.5  on 90  degrees of freedom
## AIC: 4266.4
##
## Number of Fisher Scoring iterations: 6
```

In the base Poisson method, we find a coefficient estimate for femininity (MFI) of 0.074, which suggests that an increase of 1 point in the MFI of a hurricane's name increases the log count of deaths caused by the hurricane by 0.074. Or, more intuitively, an increase of 1.1 times the deaths.

However, as this is a Poisson model, we cannot guarantee non-spurious results without assessing overdispersion in the data.

```
## [1] 20.65217
```

```
## [1] 1673.152
```

As we can see, with a mean around 20.65, and a variance of 1673.15, it is clear that the deaths are highly overdispersed as the variance is many multiples greater than the mean. Thus, we would be better off fitting a quasipoisson model instead.

```
##
## Call:
## glm(formula = alldeaths ~ MasFem, family = "quasipoisson", data = q2_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -7.1429  -5.3716  -3.8288  -0.5364  27.4230
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.50037    0.54371   4.599 1.38e-05 ***
## MasFem       0.07387    0.06778   1.090    0.279
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasipoisson family taken to be 73.78496)
##
##     Null deviance: 4031.9  on 91  degrees of freedom
## Residual deviance: 3937.5  on 90  degrees of freedom
## AIC: NA
##
## Number of Fisher Scoring iterations: 6
```

From our quasipoisson model, we see an identical estimate for MFI's effect, however, the differences seen in this model are the standard errors, and thus the p-values reported. As we can see, in this model, MFI is no longer significant, with a p-value of 0.279, unlike in the Poisson model, and this is likely due to the addition of the dispersion parameter from the quasipoisson model, in turn telling us we likely saw spurious results previously.

c)

```
##
## Call:
## glm.nb(formula = alldeaths ~ ZMinPressure_A + ZNDAM + ZMasFem +
##     ZMasFem:ZMinPressure_A + ZMasFem:ZNDAM, data = q2_data, init.theta = 0.8112499791,
##     link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5088  -1.0527  -0.4759   0.2903   2.5741
##
## Coefficients:
##                         Estimate Std. Error z value Pr(>|z|)
## (Intercept)              2.4756     0.1222  20.261  < 2e-16 ***
## ZMinPressure_A          -0.5521     0.1503  -3.673 0.000239 ***
## ZNDAM                    0.8635     0.1445   5.976 2.28e-09 ***
## ZMasFem                  0.1723     0.1238   1.392 0.163988
## ZMinPressure_A:ZMasFem   0.3948     0.1521   2.595 0.009453 **
## ZNDAM:ZMasFem            0.7051     0.1501   4.699 2.62e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(0.8112) family taken to be 1)
##
##     Null deviance: 184.86  on 91  degrees of freedom
## Residual deviance: 102.83  on 86  degrees of freedom
## AIC: 658.09
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  0.811
##           Std. Err.:  0.124
##
##  2 x log-likelihood:  -644.091

## [1] -0.1626429
```

In this case, with pressure and damage at the median of the data, the model predicts that a 1 point increase in the MFI of a hurricane leads to a decrease in the log counts of deaths of 0.163. Or, on a natural scale, a decrease to 0.85 times the deaths.

d)

```
##
## Call:
## glm.nb(formula = alldeaths ~ ZMinPressure_A + ZNDAM + ZMasFem +
##     ZMasFem:ZMinPressure_A + ZMasFem:ZNDAM, data = no_sandy,
##     init.theta = 0.8765783305, link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
```

```
## -2.7595  -1.0516  -0.4083   0.2529   2.2112
##
## Coefficients:
##                         Estimate Std. Error z value Pr(>|z|)
## (Intercept)               2.4969     0.1197  20.853  < 2e-16 ***
## ZMinPressure_A           -0.5376     0.1502  -3.579 0.000345 ***
## ZNDAM                     1.0851     0.1781   6.094 1.10e-09 ***
## ZMasFem                   0.1833     0.1205   1.521 0.128289
## ZMinPressure_A:ZMasFem    0.3892     0.1488   2.616 0.008901 **
## ZNDAM:ZMasFem             0.8487     0.1624   5.225 1.74e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(0.8766) family taken to be 1)
##
##     Null deviance: 189.25  on 90  degrees of freedom
## Residual deviance: 101.29  on 85  degrees of freedom
## AIC: 637.84
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  0.877
##           Std. Err.:  0.138
##
##  2 x log-likelihood:  -623.841
```

```
predict.glm(model_nb2, newdata = hurr_sandy %>% dplyr::select(ZMinPressure_A, ZNDAM, ZMasFem), type = "
```

```
##        1
## 113400.6
```

```
hurr_sandy$alldeaths
```

```
## [1] 159
```

As we can see, our model suggests a total of 24796.4 deaths caused by hurricane Sandy, a far shoot from the actual total of 159. This is highly significant of the model failing to generalize to new and unseen data, despite the relatively high confidence seen in the model parameter estimates.

e)

Strengths: I think that the authors had a good idea with their method of constructing the masculinity-femininity index (MFI). In the event of names chosen that are considered unisex, this could prove to be helpful as it would allow for current societal influences to be incorporated into the data, and establish if it plays an effect as well. I also thought that the authors did a very good job of sourcing their data, and were incredibly thorough in their investigation. The use of multiple model types and settings, as well as comparing their results, is beneficial when it comes to making an accurate conclusion.

Weaknesses: The main weakness of the paper is that the data they are collecting is not feasible for answering their questions, the data before 1970s and the data after are distributed in different ways, and furthermore, these names were either all female, or alternating between male and female, with no subjective

decisions based on the hurricane itself. This makes it unsuitable for making inference from. Second, the final models they are able to generate show significantly different results for MFI's connection with mortality, which suggests that there is still information that can explain more of a hurricane's predicted mortality that aren't included in these models.

f)

I'm not incredibly convinced by the results of this paper for multiple reasons. As we can see in Table S2, the coefficient of MFI changes quite a bit as model is changed. From a very low -6.16 with significance, up to 0.172 which lacked significance, this estimate seems rather shaky, and as it is the main question of the paper, this makes it quite hard to be fully confident in their results. I am also uncertain about their collection of deaths data, as they seem to suggest that they have summed both direct and indirect deaths into the 'alldeaths' variable. As it is not entirely clear how 'indirect' these indirect deaths can be, it is difficult to assess whether the data itself is suitable for answering the desired question. As far as additional work goes, it is difficult to suggest further steps which don't require additional data, as the authors have used all the variables provided in their own data, with models that are likely the most reasonable. I think additional work could come from sourcing new variables to include as covariates in the model that could possibly explain more about mortality and paint a clearer picture of MFI's true effects. It's likely very important to have a solid understanding of the areas these hurricanes touch down in, as they can have drastically different effects based on the structures they pass over. Less populous dense fields and farmlands will obviously see lower damages and fewer deaths than more metropolitan or urban cities. The biggest issue, however, is clearly defined at the beginning of the paper. As it is explained, pre-1970s, hurricanes were all named after women, whereas post-1970s, they switched to alternating between male and female names for hurricanes. This is clearly an issue as the pre-1970s and post-1970s data will have much different distributions. Moreover, as the modern method has been to alternate equally between male and female names, the question of a bias here loses value as it is tracking an arbitrary and totally random link to each hurricane. Thus, any future work they could add to this would be futile until the question is re-framed.

Q3

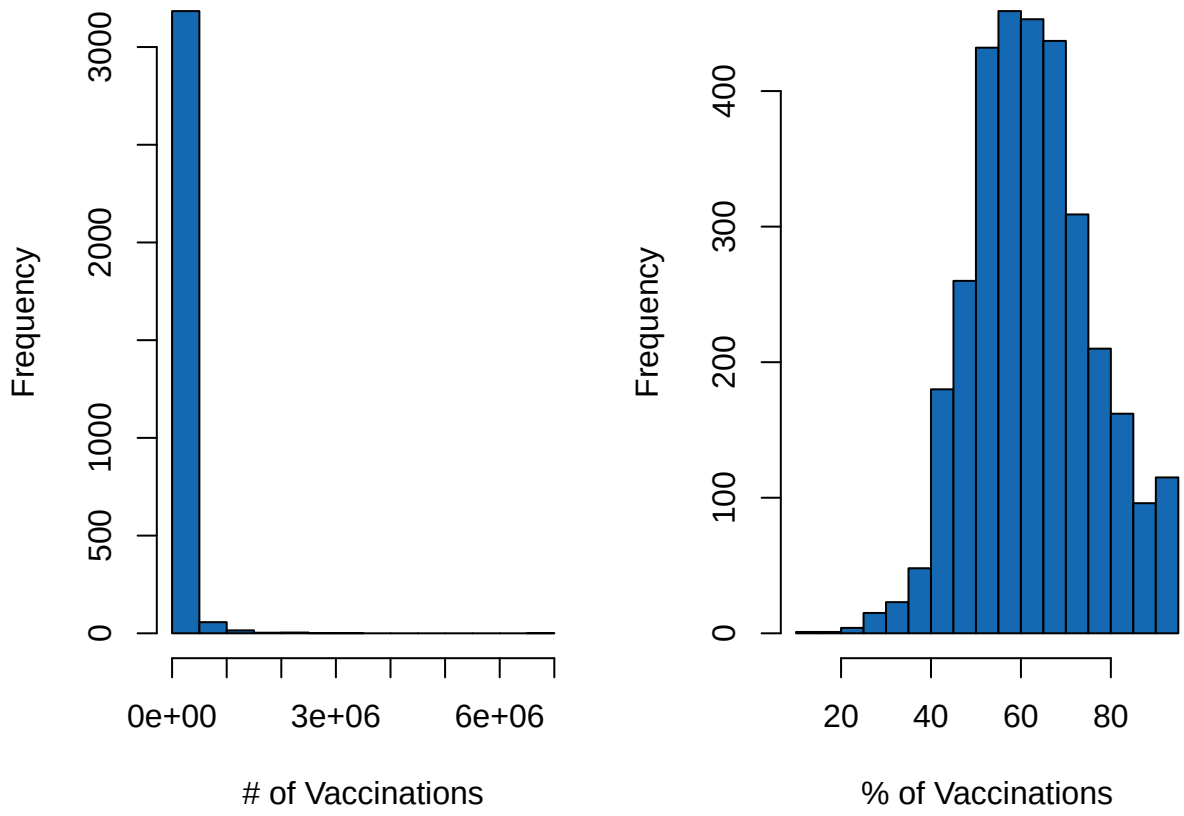
a)

EDA

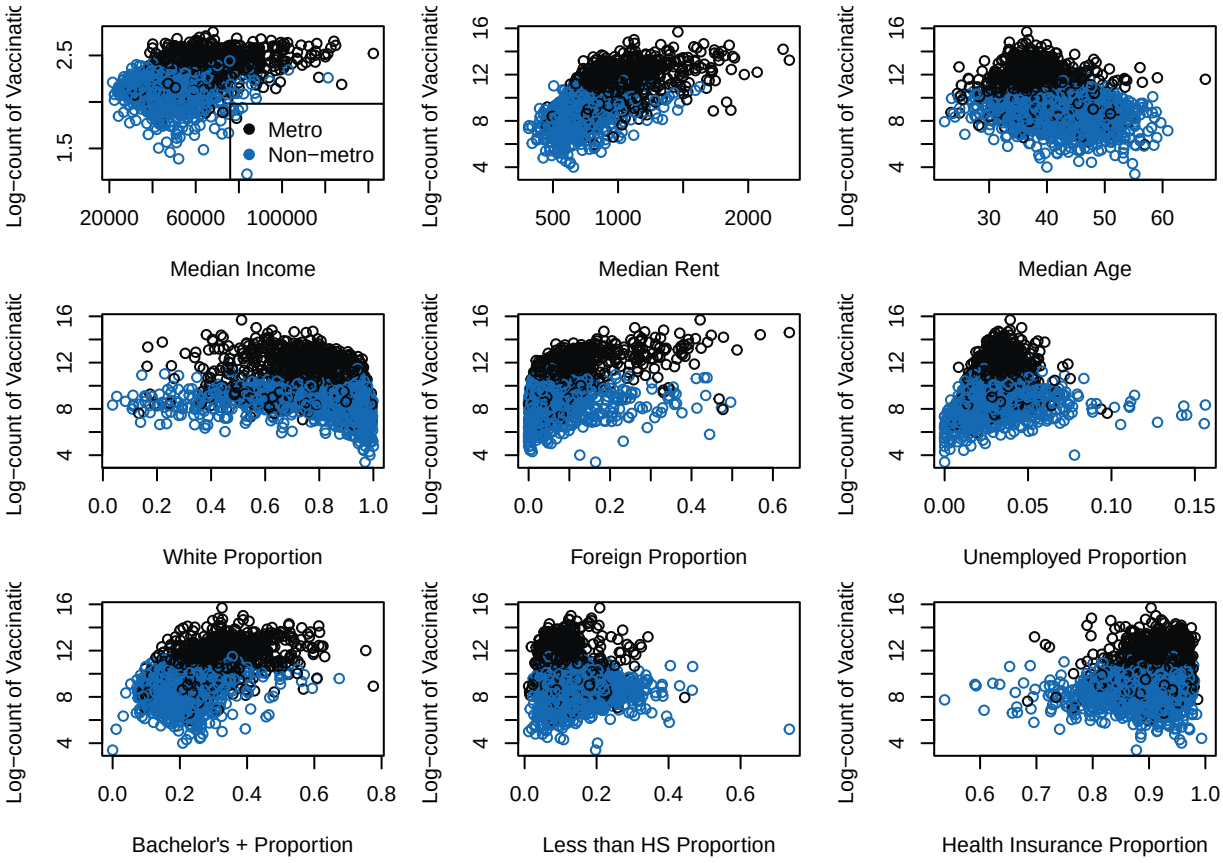
After joining and filtering the data, we performed some basic inspections. First, we ensured that all variables were correctly classified as either a numeric, factor, or string. After this, we checked for NA values in the data. We found quite a few occurrences sparsely populating the data, but there were no discernible patterns we could work out, thus we opted to leave these partially missing observations in the event the missing variables are not included in the final model.

Plots

We start by plotting the distributions of some of the main variables tracking vaccination rates in each county.



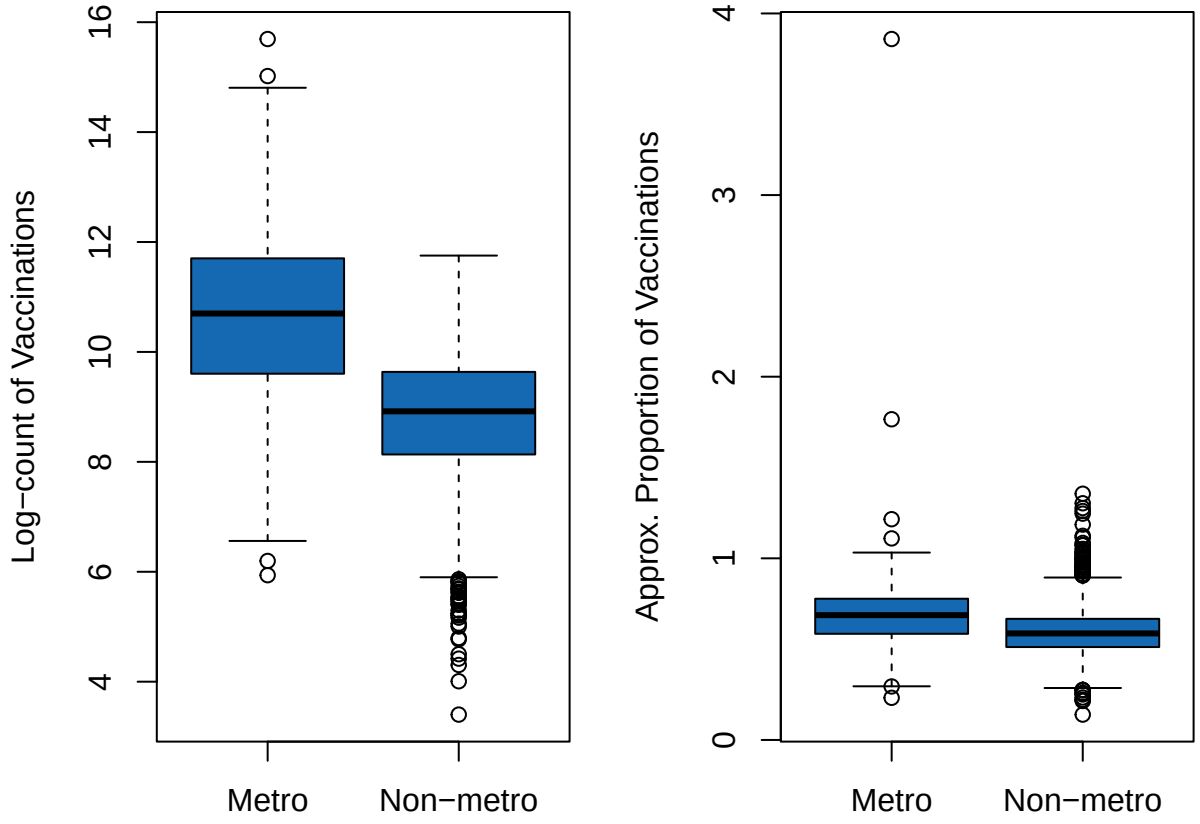
As we can see, the count of vaccinations has an extremely high mass at lower values and a very long tail, this is somewhat expected however, when we consider the sheer number of counties in the US, and compare it with the number of counties with incredibly high populations. In the percent of vaccinated individuals, we see an almost normal distribution, with a high concentration around 50%, but with a sharp cutoff at 100 due to the values it is tracking.



We then moved to plot some of our main predictor variables from the data, this includes; median income, median rent, median age, the proportion of white citizens, the proportion of foreign citizens, the unemployment population proportion, proportion of individuals with a bachelor's degree or greater, proportion of individuals with less than a high school education, and the proportion of individuals with health insurance. We first plotted them against the raw counts of vaccinated individuals in each county, but due to the high mass of low count observations, we decided to plot these variables against the log of the counts of observations. In turn, this allowed for a much clearer picture of the trends in the data and allows us to better investigate variables to include in our model. To also inspect our final variable of interest, we colored the plots by their metro status, to see if there is a difference of effects in metropolitan counties vs. more suburban or rural counties. In doing so we found no significant changes in relationship based on metro status, but we did find some interesting relationships to the full variables.

Firstly, we found that median income and median rent both have subtle positive relationships with vaccination counts, suggesting counties with a higher income and typically high rents and housing costs are expected to see higher rates of vaccination. There was no clear relationship when we measured median ages against the log-counts as we see a relatively even distribution across the line, there is a slight downward trend but this is largely made up for by the variance of the observations. We also found an interesting trend based on the racial proportions of a county, the proportion of white individuals has a very slight negative relationship with vaccination counts, which becomes more pronounced as we pass 80% of the population being white. One reason for this we believe is that the counties with white proportions close to 100% are likely to be more rural and thus see lower populations overall, in turn potentially leading to more citizens not thinking a vaccine is necessary. Conversely, we found a very subtle positive relationship with the proportion of foreign individuals which becomes less strong as the proportion grows. In a similar reasoning, we believe this is more likely a fact that the counties with greater proportions of foreign citizens are likely larger metropolitan areas and thus see more people living closer to each other, thus raising urgency for vaccination. With the unemployment population proportion, we did not see any clear trends except for a change in variance as we increase the proportion, there appeared to be a small positive relationship, but the variance is too high

to fully confirm any existence. For our education metrics, we found a very slight positive relationship with log-counts of vaccinations with the proportion of citizens with a bachelor's degree or greater, while there was almost no clear relationship with the proportion under a high-school education. And finally, with our health insurance proportion, we found no clear relationship to log vaccination counts, which was somewhat surprising to us. However, it is difficult to parse out exactly why this is the case, and we believe it is probably due to the more confusing nature of health insurance that exists in the US and how it varies across the country. After viewing these relationships, we had one final variable to explore in relation to vaccinations.



Lastly, we compared the distribution of log vaccination counts in metropolitan and non-metropolitan cities using a boxplot. As we can see, metropolitan counties tend to have slightly higher vaccination counts than non-metropolitan cities, however, this result was quite expected as we are studying count data, and metropolitan counties have significantly different population sizes than non-metropolitan cities, which would ultimately influence our results. Thus, we also divided the raw number of vaccinations by the census population size we have in the plot on the right. This shrinks the distance between the two groups by a fair amount, and suggests that the 'metropolitan-ness' of a county plays less of a significant role in the vaccination rates than we initially were led to believe. We also note here that there are a number of proportions above 1, which we believe is due to a slight mismatch in the data between the actual census numbers, but we still see the majority below 1 as expected.

b)

To begin with model creation, we must first decide on a dependent variable. From our earlier plots, we can see that the counts of vaccinations has a very high mass at lower values, with a long tail, encouraging the use of a Poisson or closely related count model. On the other hand, taking the log of this distribution leaves us with a new distribution fairly close to normal, suggesting the use of a log transformed linear regression. Finally, we also saw that the proportion measure has somewhat of a normal distribution, but with a hard

limit at 100, and fairly asymmetric tails, making us more confident in sticking with the count data. Thus, as a first step, we created 3 matching regression models using the counts of vaccinated individuals as a response. Here, we included variables for; median age, median income, median rent, metro status, proportion of white and foreign born citizens, proportion of those with less than high school education and a bachelor's degree or above, the proportion of unemployed citizens, and the proportion of citizens with health insurance. Our first model was a base Poisson model, however, after checking for overdispersion, we found that there is significant overdispersion, thus leading us to use a quasipoisson model. In turn, we also included a negative binomial model as another way of dealing with overdispersion. Another method we also considered was using a simple log-transformed regression model, as we notice the distribution of log counts of 18+ vaccinated individuals is quite close to normal.

```
##
## Call:
## glm(formula = Series_Complete_18Plus ~ median_age + median_income +
##     median_rent + Metro_status + prop_white + prop_foreign_born +
##     prop_less_than_hs + prop_bachelor_above + prop_unemployed +
##     prop_health_insurance, family = "poisson", data = q3_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1414.6   -112.1    -38.8     37.3   3502.7
##
## Coefficients:
##                           Estimate Std. Error z value Pr(>|z|)
## (Intercept)            8.950e+00  2.095e-03  4271.3   <2e-16 ***
## median_age            -7.551e-03  1.736e-05  -434.9   <2e-16 ***
## median_income         -2.548e-05  8.198e-09 -3108.1   <2e-16 ***
## median_rent            1.219e-03  5.628e-07  2166.7   <2e-16 ***
## Metro_statusNon-metro -1.662e+00  2.505e-04 -6635.1   <2e-16 ***
## prop_white            -2.293e-01  6.054e-04  -378.8   <2e-16 ***
## prop_foreign_born      5.791e+00  9.174e-04  6312.6   <2e-16 ***
## prop_less_than_hs     -3.358e+00  2.362e-03 -1421.8   <2e-16 ***
## prop_bachelor_above    2.688e+00  1.114e-03  2413.6   <2e-16 ***
## prop_unemployed        1.644e+01  8.403e-03  1955.9   <2e-16 ***
## prop_health_insurance  2.282e+00  2.060e-03  1107.6   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 584341797  on 3120  degrees of freedom
## Residual deviance: 157042685  on 3110  degrees of freedom
##   (162 observations deleted due to missingness)
## AIC: 157078210
##
## Number of Fisher Scoring iterations: 6
```

```
##
## Call:
## glm(formula = Series_Complete_18Plus ~ median_age + median_income +
##     median_rent + Metro_status + prop_white + prop_foreign_born +
##     prop_less_than_hs + prop_bachelor_above + prop_unemployed +
##     prop_health_insurance, family = "quasipoisson", data = q3_data)
##
```

```
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1414.6   -112.1    -38.8     37.3   3502.7
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)            8.950e+00  5.521e-01  16.212  < 2e-16 ***
## median_age            -7.551e-03  4.575e-03  -1.651   0.0989 .
## median_income         -2.548e-05  2.160e-06 -11.797  < 2e-16 ***
## median_rent            1.219e-03  1.483e-04   8.224 2.86e-16 ***
## Metro_statusNon-metro -1.662e+00  6.601e-02 -25.184  < 2e-16 ***
## prop_white            -2.293e-01  1.595e-01  -1.438   0.1506
## prop_foreign_born      5.791e+00  2.417e-01  23.960  < 2e-16 ***
## prop_less_than_hs     -3.358e+00  6.223e-01  -5.397 7.30e-08 ***
## prop_bachelor_above    2.688e+00  2.934e-01   9.161  < 2e-16 ***
## prop_unemployed        1.644e+01  2.214e+00   7.424 1.46e-13 ***
## prop_health_insurance  2.282e+00  5.428e-01   4.204 2.69e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasipoisson family taken to be 69412.99)
##
##     Null deviance: 584341797  on 3120  degrees of freedom
## Residual deviance: 157042685  on 3110  degrees of freedom
##   (162 observations deleted due to missingness)
## AIC: NA
##
## Number of Fisher Scoring iterations: 6


##
## Call:
## glm.nb(formula = Series_Complete_18Plus ~ median_age + median_income +
##     median_rent + Metro_status + prop_white + prop_foreign_born +
##     prop_less_than_hs + prop_bachelor_above + prop_unemployed +
##     prop_health_insurance, data = q3_data, init.theta = 1.357541626,
##     link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.8610  -0.9693  -0.3292   0.3264   5.4719
##
## Coefficients:
##                           Estimate Std. Error z value Pr(>|z|)
## (Intercept)            5.585e+00  3.976e-01   14.05  < 2e-16 ***
## median_age            -4.208e-02  3.358e-03  -12.53  < 2e-16 ***
## median_income         -3.607e-05  2.145e-06  -16.82  < 2e-16 ***
## median_rent            3.568e-03  1.437e-04   24.82  < 2e-16 ***
## Metro_statusNon-metro -9.837e-01  3.828e-02  -25.70  < 2e-16 ***
## prop_white             3.762e-01  1.198e-01    3.14  0.00169 **
## prop_foreign_born      3.480e+00  3.284e-01   10.60  < 2e-16 ***
## prop_less_than_hs     -9.234e-01  4.294e-01   -2.15  0.03152 *
## prop_bachelor_above    3.407e+00  2.900e-01   11.75  < 2e-16 ***
## prop_unemployed        1.621e+01  1.391e+00   11.65  < 2e-16 ***
## prop_health_insurance  4.725e+00  3.814e-01   12.39  < 2e-16 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(1.3575) family taken to be 1)
##
##     Null deviance: 12554.9  on 3120  degrees of freedom
## Residual deviance:  3486.9  on 3110  degrees of freedom
##   (162 observations deleted due to missingness)
## AIC: 68206
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  1.3575
##           Std. Err.:  0.0311
##
##  2 x log-likelihood:  -68181.7470

##
## Call:
## lm(formula = log(Series_Complete_18Plus) ~ median_age + median_income +
##     median_rent + Metro_status + prop_white + prop_foreign_born +
##     prop_less_than_hs + prop_bachelor_above + prop_unemployed +
##     prop_health_insurance, data = q3_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.0934 -0.5581  0.0907  0.6666  3.2378
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)            3.602e+00  4.607e-01   7.818 7.29e-15 ***
## median_age            -4.554e-02  3.891e-03 -11.703  < 2e-16 ***
## median_income         -3.359e-05  2.486e-06 -13.513  < 2e-16 ***
## median_rent            3.586e-03  1.666e-04  21.531  < 2e-16 ***
## Metro_statusNon-metro -9.966e-01  4.437e-02 -22.464  < 2e-16 ***
## prop_white             2.769e-01  1.388e-01   1.994   0.0462 *
## prop_foreign_born      2.654e+00  3.806e-01   6.974 3.74e-12 ***
## prop_less_than_hs     -2.516e-01  4.976e-01  -0.506   0.6132
## prop_bachelor_above    2.138e+00  3.360e-01   6.362 2.28e-10 ***
## prop_unemployed        1.416e+01  1.612e+00   8.786  < 2e-16 ***
## prop_health_insurance  6.892e+00  4.420e-01  15.595  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9946 on 3110 degrees of freedom
##   (162 observations deleted due to missingness)
## Multiple R-squared:  0.6138, Adjusted R-squared:  0.6126
## F-statistic: 494.4 on 10 and 3110 DF,  p-value: < 2.2e-16
```

Checking these model coefficients, we found high significance across almost all included variables in the Poisson model, but lost significance on the proportion of white citizens and the median age in the county in the quasipoisson. In the negative binomial we found significance on all predictors. And in the linear model,

we find strong significance across all predictors except the proportion of those with less than a high school education, and the proportion of white citizens which just barely passes our tests.

Despite this, however, we are also tempted to explore the correlation matrix of these variables. In doing so, we find many correlations with magnitudes greater than 0.4, leading us to run multiple models dropping variables in many different combinations, and exploring their impact on AIC and overall significance.

We found high correlations between median income, rent, and the bachelor's+ proportion, which were both to be expected, as well as many others. After running multiple iterations of removing these highly correlated variables, we converged on the following covariates, used in the quasipoisson, negative binomial, and log-normal models; median age, metro status, bachelor's+ proportion, and the unemployed proportion.

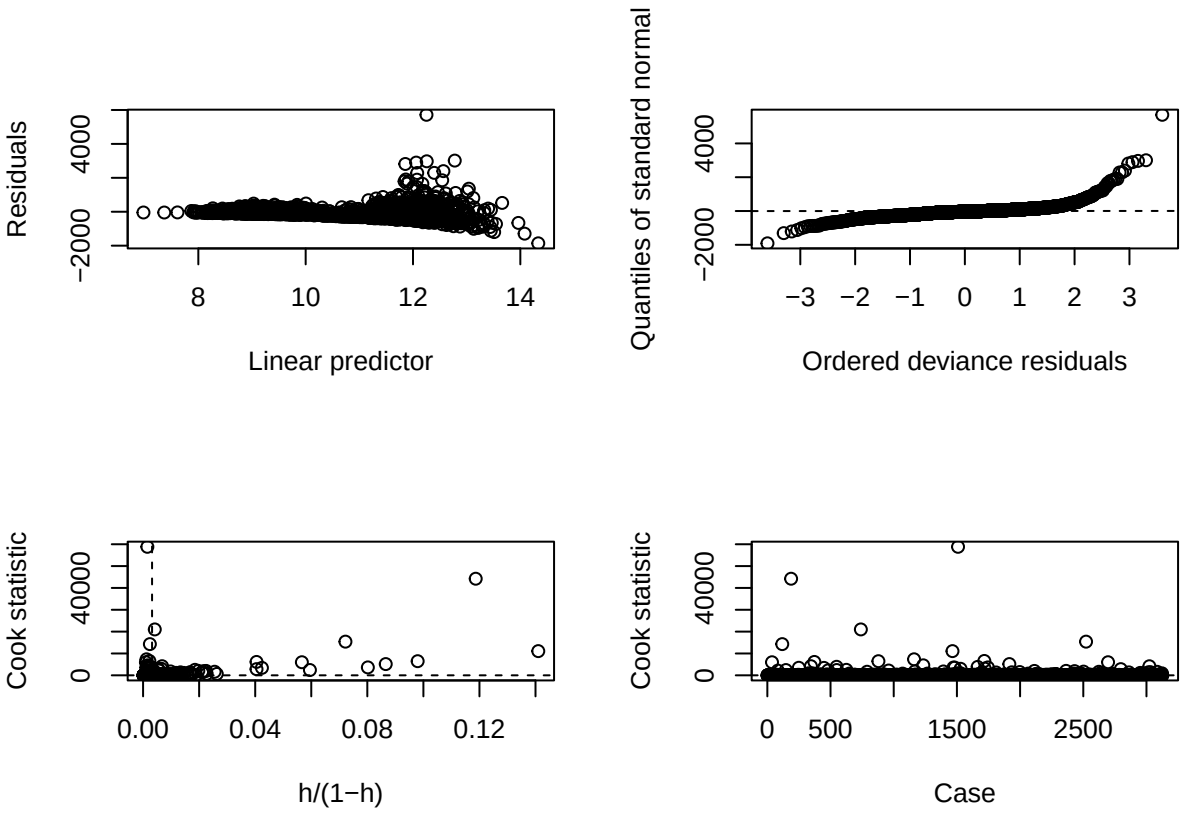
```
##
## Call:
## glm(formula = Series_Complete_18Plus ~ median_age + Metro_status +
##     prop_bachelor_above + prop_unemployed, family = "quasipoisson",
##     data = q3_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1806.8   -144.5    -39.5     41.3   5688.9
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)            10.155898   0.334268  30.382  < 2e-16 ***
## median_age             -0.021352   0.006782  -3.148  0.00166 **
## Metro_statusNon-metro -1.999431   0.103200 -19.374  < 2e-16 ***
## prop_bachelor_above     5.404370   0.243879  22.160  < 2e-16 ***
## prop_unemployed        28.384045   2.265247  12.530  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for quasipoisson family taken to be 182880.1)
##
##     Null deviance: 584816067  on 3124  degrees of freedom
## Residual deviance: 261649631  on 3120  degrees of freedom
##   (158 observations deleted due to missingness)
## AIC: NA
##
## Number of Fisher Scoring iterations: 8
```

```
##
## Call:
## glm.nb(formula = Series_Complete_18Plus ~ median_age + prop_bachelor_above +
##     Metro_status + prop_unemployed, data = q3_data, init.theta = 1.053647375,
##     link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.5915  -1.0338  -0.4004   0.2578   7.6604
##
## Coefficients:
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)             9.816534   0.178095   55.12   <2e-16 ***
## median_age             -0.035886   0.003498  -10.26   <2e-16 ***
```

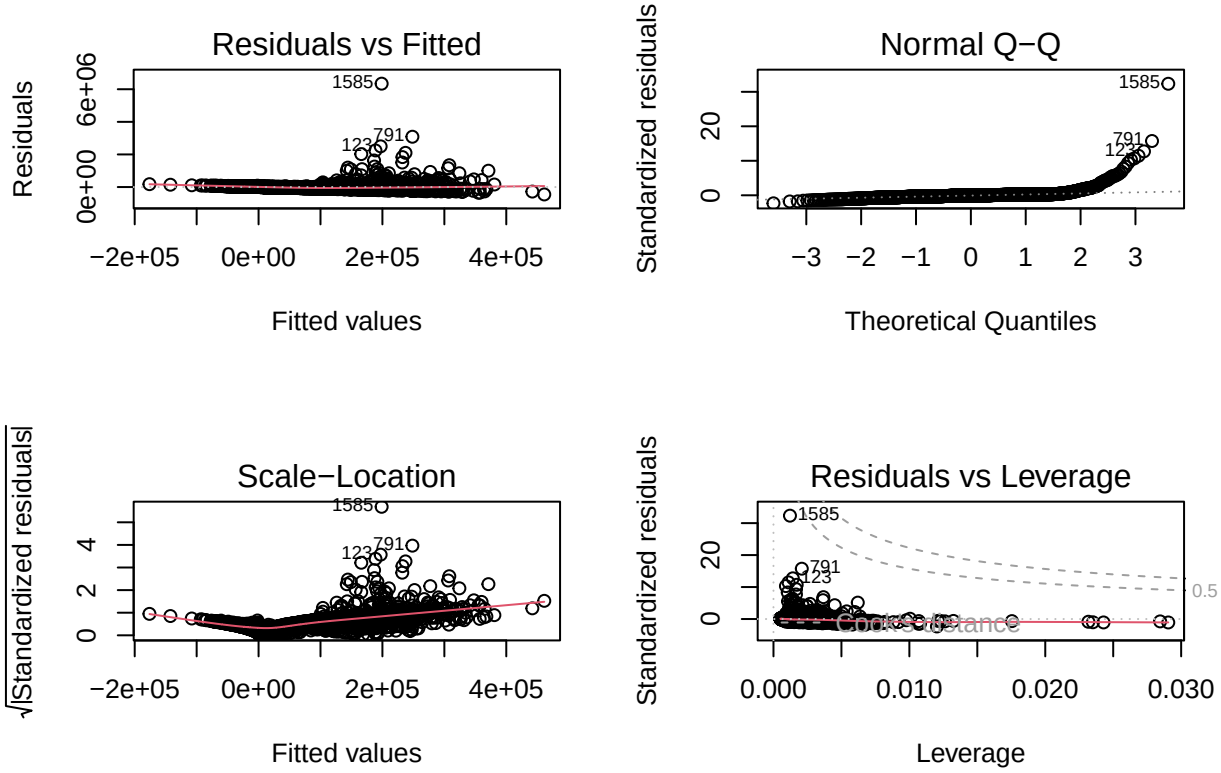
```
## prop_bachelor_above    7.376639   0.205747   35.85   <2e-16 ***
## Metro_statusNon-metro -1.379286   0.040373  -34.16   <2e-16 ***
## prop_unemployed       33.330088   1.365390   24.41   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(1.0536) family taken to be 1)
##
##     Null deviance: 9785.1  on 3124  degrees of freedom
## Residual deviance: 3585.7  on 3120  degrees of freedom
##   (158 observations deleted due to missingness)
## AIC: 69237
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  1.0536
##           Std. Err.:  0.0236
##
##  2 x log-likelihood:  -69225.0370


##
## Call:
## lm(formula = Series_Complete_18Plus ~ median_age + prop_bachelor_above +
##     Metro_status + prop_unemployed, data = q3_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -453866  -57319   -7695   24458 6341536
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)              -3500.3    35879.4  -0.098   0.9223
## median_age               -1756.6      704.7  -2.493   0.0127 *
## prop_bachelor_above     629094.2    41450.9  15.177  < 2e-16 ***
## Metro_statusNon-metro   -75714.1     8133.7  -9.309  < 2e-16 ***
## prop_unemployed        1568081.6   275074.6   5.701 1.31e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 196300 on 3120 degrees of freedom
##   (158 observations deleted due to missingness)
## Multiple R-squared:  0.1625, Adjusted R-squared:  0.1614
## F-statistic: 151.3 on 4 and 3120 DF,  p-value: < 2.2e-16
```

After filtering these variables in each model, we must come to a final model. To do so, we examine some diagnostics of each candidate (Order: Quasipoisson, Linear Model, Negative Binomial):

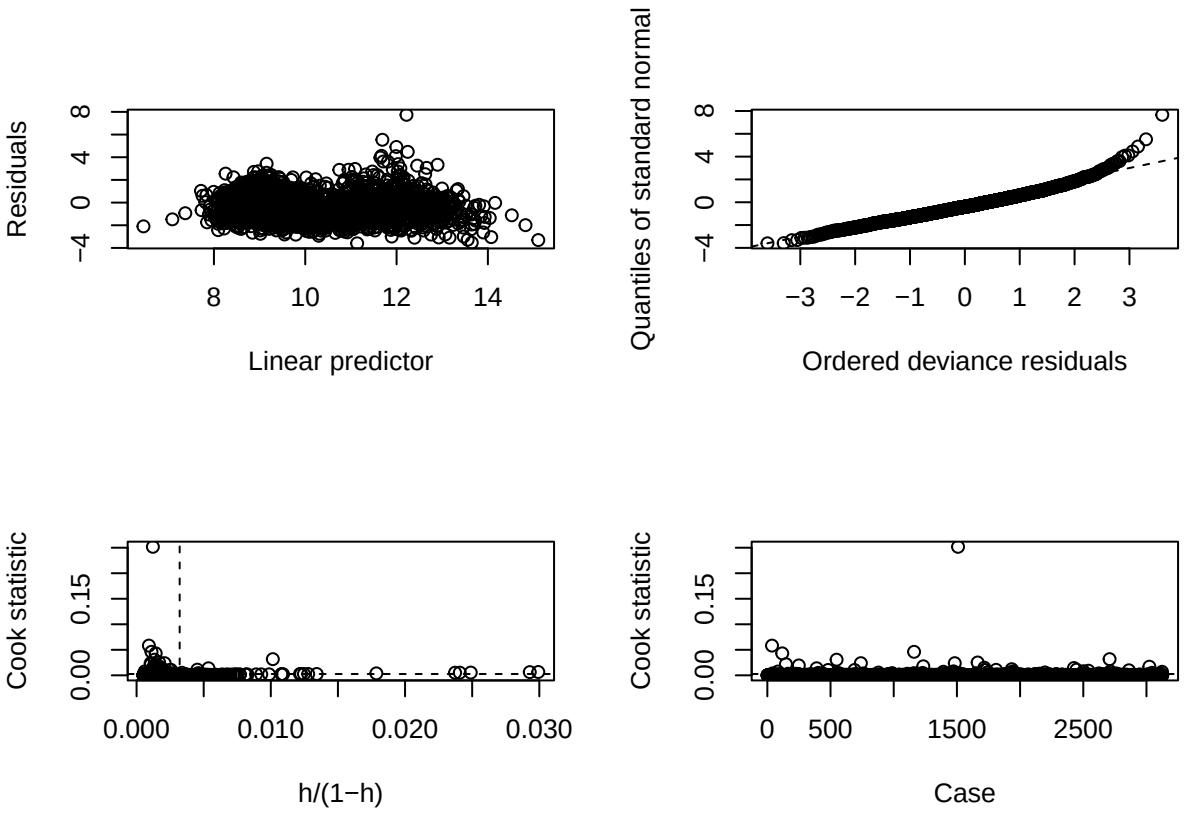
```
glm.diag.plots(q3_model5) #  Quasipoisson
```




```
par(mfrow = c(2, 2))
plot(q3_model7)           #  Log-Transformed
```




```
glm.diag.plots(q3_model6) #  Negative-Binomial
```



As we can see, in the quasipoisson model, there are some concerns in the residual vs. fitted plot, where we see significant signs of heteroskedasticity of the errors. Moreover, while Poisson regression doesn't require normality, there is a much more pronounced deviation than in the negative binomial model as seen in the qq-plot. Furthermore, the cook's statistics are much greater in the quasipoisson model as well, bringing up some concerns with influential outliers that can potentially impact our results. Second, in the log transformed model, we found major issue with many of the diagnostic plots including significant heteroskedasticity of errors as seen in the residuals vs and Scale-Location plots. Here we can also see a significant divergence from normality in the qq-plot, but overall there does seem to be very few highly influential points in the data. However, these results were still not satisfactory given the multiple places they appeared to fail.

Finally, in the negative binomial plots on the other hand, we see much better results. Residuals show almost no sign of heteroskedasticity, the qq-line is well followed except for a slightly heavy tail, and the Cook's distance statistics are much smaller, with little to no evidence of any points acting as influential outliers. Thus, we selected the negative binomial model as our final model.

In the final model, we find the following coefficient estimates:

```
##
## Call:
## glm.nb(formula = Series_Complete_18Plus ~ median_age + prop_bachelor_above +
##     Metro_status + prop_unemployed, data = q3_data, init.theta = 1.053647375,
##     link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.5915  -1.0338  -0.4004   0.2578   7.6604
##
```

```
## Coefficients:
##                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)             9.816534   0.178095   55.12   <2e-16 ***
## median_age             -0.035886   0.003498  -10.26   <2e-16 ***
## prop_bachelor_above     7.376639   0.205747   35.85   <2e-16 ***
## Metro_statusNon-metro -1.379286   0.040373  -34.16   <2e-16 ***
## prop_unemployed       33.330088   1.365390   24.41   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(1.0536) family taken to be 1)
##
##     Null deviance: 9785.1  on 3124  degrees of freedom
## Residual deviance: 3585.7  on 3120  degrees of freedom
##   (158 observations deleted due to missingness)
## AIC: 69237
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  1.0536
##           Std. Err.:  0.0236
##
##  2 x log-likelihood:  -69225.0370
```

To interpret each coefficient, we first make the assumption that all other variables are held fixed, and recall the use of a log link for this model. For a 1 unit increase in the median age of the county, the log count of vaccinated individuals is expected to decrease by 0.036, which after converting becomes a decrease of about 3.5%. Increasing the proportion of individuals with a bachelor's degree or greater by 1 unit is expected to lead to an increase of almost 7.38 in the log count of vaccinations, or an increase 1600 times larger. For metropolitan vs. non-metropolitan counties, we found that non-metropolitan counties saw a decrease in log counts of vaccinated individuals of about 1.38, or a decrease 4 times smaller. And finally, with a coefficient of 33.33, for the unemployed population, we found that a 1 unit increase in their proportion leads to an increase that is astronomically large. While this poses a potential problem, we believe this is likely due to there being additional uncovered data we have not accessed yet which could influence vaccination rates and paint a better picture, as in all other models the coefficient on unemployment was significantly higher than all other covariates in the model.

Remembering the log-link using in each model, these can be interpreted as the change in the log counts of the number of vaccinated individuals in a county, or, after exponentiation, the multiplicative change in the counts of vaccinated individuals. Thus, interpreting these as multiplications on the natural scale is as follows (Taking the exp() of all values in the 'Estimate' Column):

```
## [1] "Age"               "0.964750268745051"

## [1] "Bachelors"       "1598.2091498165"

## [1] "Metro Status"      "0.251758244288776"

## [1] "Unemployed"      "298588653106955"
```

c)

```
predict.glm(q3_model5_pred, newdata = q3_data_ada, type = "response")
```

```
##        1
## 195855.6
```

```
q3_data_ada$Series_Complete_18Plus
```

```
## [1] 284600
```

```
predict.glm(q3_model5_pred, newdata = q3_data_ada, type = "response")/q3_data_ada$Census2019_18PlusPop
```

```
##         1
## 0.5295441
```

```
q3_data_ada$Series_Complete_18Plus/q3_data_ada$Census2019_18PlusPop
```

```
## [1] 0.7694866
```

```
predict.glm(q3_model5_pred, newdata = q3_data_ada, type = "response")/q3_data_ada$total_pop_18plus
```

```
##         1
## 0.5643407
```

```
q3_data_ada$Series_Complete_18Plus/q3_data_ada$total_pop_18plus
```

```
## [1] 0.82005
```

As we can see, our model is still considerably far from the true value of vaccinated proportions in Ada County, with a margin of almost 100,000 individuals, and over a 20% difference in proportions. Despite this, however, we believe it is likely due to the widely different ways in which individual counties can live very different lives, and the fact that we are likely missing data which can help explain these differences to a better degree. Ultimately, this prediction is not incredibly good, but the model's significance as well as strong performance under diagnostic testing leads us to believe we still have a fairly strong candidate, and that additional predictions should be carried out to fully understand the model's capabilities.

d)

Our analysis showed that the vaccination rates of a county are impacted by many factors, in many different ways. Importantly, we saw that vaccination rates were positively influenced by the proportion of individuals with a bachelor's degree or higher, and the proportion of individuals that are currently unemployed, whereas we saw that vaccination rates are negatively impacted by the median age in a county, and that more rural and non-metropolitan counties also see lower vaccination rates. These results are somewhat expected when we consider their relevence in the real world. Those with a bachelor's degree could be more forward thinking and proactive with their health than those of lower education, thus leading to higher vaccination rates. The unemployed population is also something heavily impacted by Covid itself, and could mean that during

covid, when these unemployed people are looking for work, they will almost certainly need to be vaccinated before being allowed to work in many places. With non-metroplitan counties, we would expect these areas to see lower vaccination rates as we believe the less population dense areas will feel less of an urgency to become vaccinated if they are unlikely to interact with many people. Moreover, these areas are less likely to have travellers from other states or countries, making them less likely to see the same levels of transmission. Finally, we were somewhat surprised by the effect of age on vaccinations, as we would believe older populations will feel a stronger need to be vaccinated due to the greater health risks. However, we considered that this outcome could be the result of the fact that areas with lower aged citizens likely have a greater active workforce, and thus have a greater population of individuals that need to get vaccinated to continue working. But these judgements are difficult to be certain of without further exploration.

In the future we think additional variables of interest would be those that track more personal factors, such as political beliefs, medical history, or marital status, but these becomes difficult to include without massively increasing the granularity to an individual basis. Other factors could also include more descriptions on population density, and poverty or homelessness rates. We also believe additional work could be done on model creation, through the inclusion of offsets in the Poisson model, and interactions between certain terms. While our own modelling path did not show any significance in interaction terms, there were are still many areas in which interactions could take on significant effects once new variables are included. In doing so, we could answer many more questions about vaccine hesitancy, and possibly work out the issue we saw with the unemployment rates.

e)

Option 1 provides the lowest granularity, as it provides data only at the state level, with little information on individual county level data. In turn, this would leave us with a maximum of 50 observations, which also runs the risk becoming too small for us to generate reliable results from which we could be confident in. Option 2 provides a similar level of granularity, but the outcome variable averages over county level data, rather than total population data. This can help differentiate between states with high populations only in one area, versus those with evenly spread populations, and paint a finer picture of each individual state's vaccination rate. Option 3, on the other hand, provides very good granularity, as we would have data on every county in the country, however, its use of state fixed effects is somewhat hindering to the question they wish to answer. These state effects could provide some interesting information about the general feelings regarding the vaccine in the state, but it would be difficult to track vaccination rates at the state level here without combining the results of each county. Moreover, state fixed effects means that we get 50 more variables in our model, which can also hurt predictive performance if we have too few data points. The most appropriate analysis is likely dependent on the specific question asked, however, as they all measure similar outcomes, but with slightly different contents.