STA2201H Winter 2023 Assignment 1

Due: 11:59pm, 3 February 2023

What to hand in: .qmd or .Rmd file and the compiled pdf

How to hand in: Submit files via Quercus

1 Overdispersion

Suppose that the conditional distribution of outcome Y given an unobserved variable θ is Poisson, with a mean and variance $\mu\theta$, so

$$Y|\theta \sim \text{Poisson}(\mu\theta)$$

a) Assume $E(\theta) = 1$ and $Var(\theta) = \sigma^2$. Using the laws of total expectation and total variance, show $E(Y) = \mu$ and $Var(Y) = \mu(1 + \mu\sigma^2)$.
b) Assume θ is Gamma distributed with α and β as shape and scale parameters, respectively. Show the unconditional distribution of Y is Negative Binomial.
c) In order for $E(Y) = \mu$ and $Var(Y) = \mu(1 + \mu\sigma^2)$, what must α and β equal?

a.

$$\begin{aligned}E(Y) &= E(E(Y|\theta))\\ &= E(\mu\theta)\\ &= \mu E(\theta)\\ &= \mu\end{aligned}$$

$$\begin{aligned}Var(Y) &= Var(E(Y|\theta)) + E(Var(Y|\theta))\\ &= Var(\mu\theta) + E(\mu\theta)\\ &= \mu^2\sigma^2 + \mu\\ &= \mu(1+\mu\sigma^2)\end{aligned}$$

b.

$$\begin{aligned}P(Y=y) &= \int_0^\infty P(y|\theta)f(\theta)d\theta\\ &= \int_0^\infty \frac{e^{-(\mu\theta)}(\mu\theta)^y}{y!}\frac{1}{\Gamma(\alpha)}\theta^{(\alpha-1)}(1/\beta)^\alpha e^{-\theta/\beta}\ d\theta\\ &= \mu^y\int_0^\infty \frac{(1/\beta)^\alpha}{y!\Gamma(\alpha)}\theta^{y+\alpha-1}e^{-((1/\beta)+\mu)\theta}\ d\theta\\ &= \mu^y\frac{(1/\beta)^\alpha}{y!\Gamma(\alpha)}\frac{\Gamma(y+\alpha)}{((1/\beta)+\mu)^{y+\alpha}}\int_0^\infty \frac{((1/\beta)+\mu)^{y+\alpha}}{\Gamma(y+\alpha)}\theta^{y+\alpha-1}e^{-((1/\beta)+\mu)\theta}\ d\theta\\ &= \mu^y\frac{(1/\beta)^\alpha}{y!\Gamma(\alpha)}\frac{\Gamma(y+\alpha)}{((1/\beta)+\mu)^{y+\alpha}}(1)\\ &= \frac{\Gamma(y+\alpha)}{\Gamma(\alpha)\Gamma(y+1)}(\frac{(1/\beta)}{(1/\beta)+\mu})^\alpha(\frac{\mu}{(1/\beta)+\mu})^y\\ &= \binom{y+\alpha-1}{y}(\frac{(1/\beta)}{(1/\beta)+\mu})^\alpha(\frac{\mu}{(1/\beta)+\mu})^y\\ &\sim NB(\alpha,\frac{1/\beta}{1/\beta+\mu})\end{aligned}$$

c.

$$\begin{aligned}\mu &= E(Y)\\ &= \frac{\alpha(1-\frac{1/\beta}{1/\beta+\mu})}{\frac{1/\beta}{1/\beta+\mu}}\\ \mu(1+\mu\sigma^2) &= Var(Y)\\ &= \frac{\alpha(1-\frac{1/\beta}{1/\beta+\mu})}{(\frac{1/\beta}{1/\beta+\mu})^2}\end{aligned}$$

- Two equations two unknowns, solve:

$$\begin{aligned}\alpha &= 1/\sigma^2 \\ \beta &= \sigma^2\end{aligned}$$

2 Hurricanes

In 2014 the following paper was published in PNAS:

> Jung, K., Shavitt, S., Viswanathan, M., & Hilbe, J. M. (2014). Female hurricanes are deadlier than male hurricanes. Proceedings of the National Academy of Sciences, 111(24), 8782-8787.

As the title suggests, the paper claimed that hurricanes with female names have caused a greater loss of life. In this question you will be investigating the data set used for the regression part of their analysis.

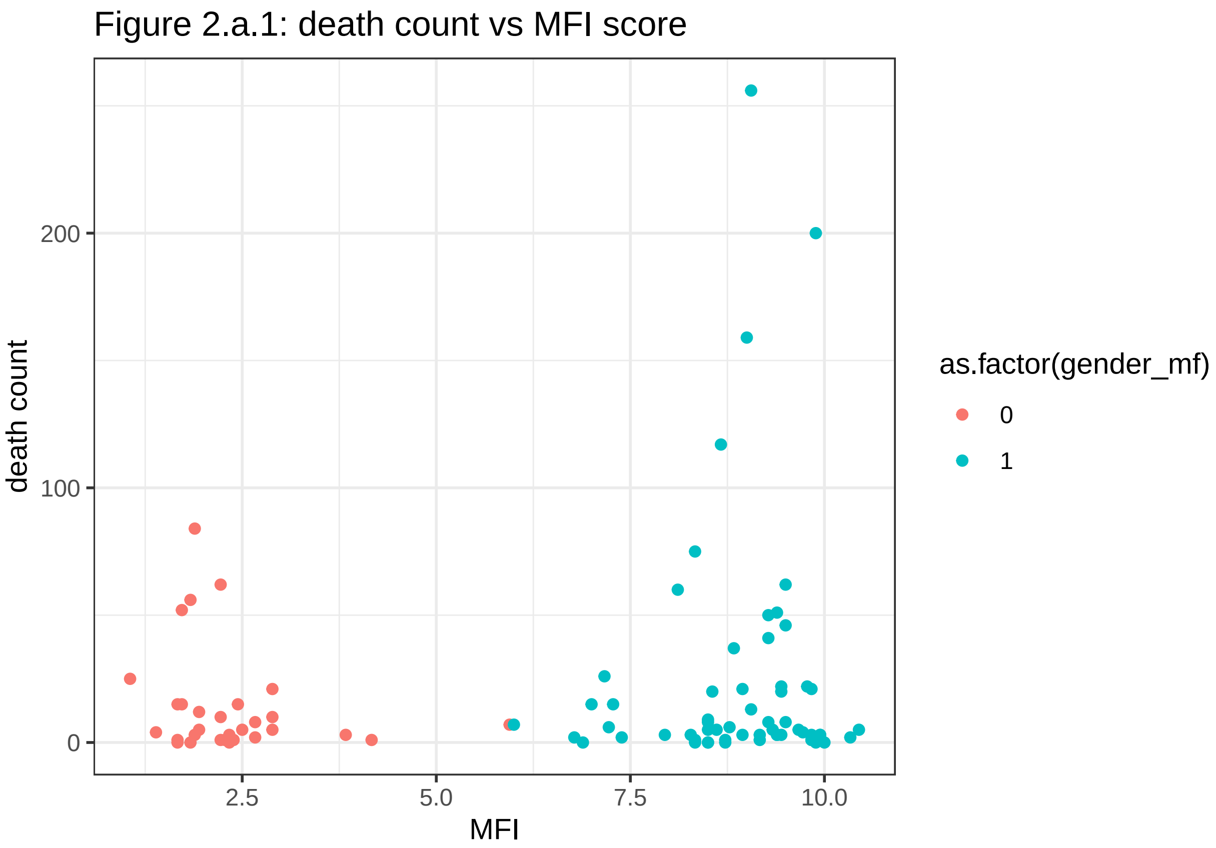
You can download the data from the paper's supporting information here: https://www.pnas.org/doi/10.1073/pnas.1402786111#supplementary-materials

You should skim the whole paper but you will probably find it useful to read the sections on the Archival Study in the most depth (both in the main text and 'Materials and Methods' section).

a) Create three graphs in ggplot that help to visualize patterns in deaths by femininity, minimum pressure, and damage. Discuss what you observe based on your visualizations.
b) Run a Poisson regression with `deaths` as the outcome and `femininity` as the explanatory variable. Interpret the resulting coefficient estimate. Check for overdispersion. If it is an issue, run a quasi-Poisson regression with the same variables. Interpret your results.
c) Reproduce Model 4 (as described in the text and shown in Table S2).[1] Report the estimated effect of femininity on deaths assuming a hurricane with median pressure and damage ratings.
d) Using Model 4, predict the number of deaths caused by Hurricane Sandy. Interpret your results.
e) Describe at least two strengths and two weaknesses of this paper, focusing on the archival analysis. What was done well? What needed improvement?
f) Are you convinced by the results? If you are, explain why. If you're not, describe what additional data and/or analyses you would like to see to further test the author's hypothesis.

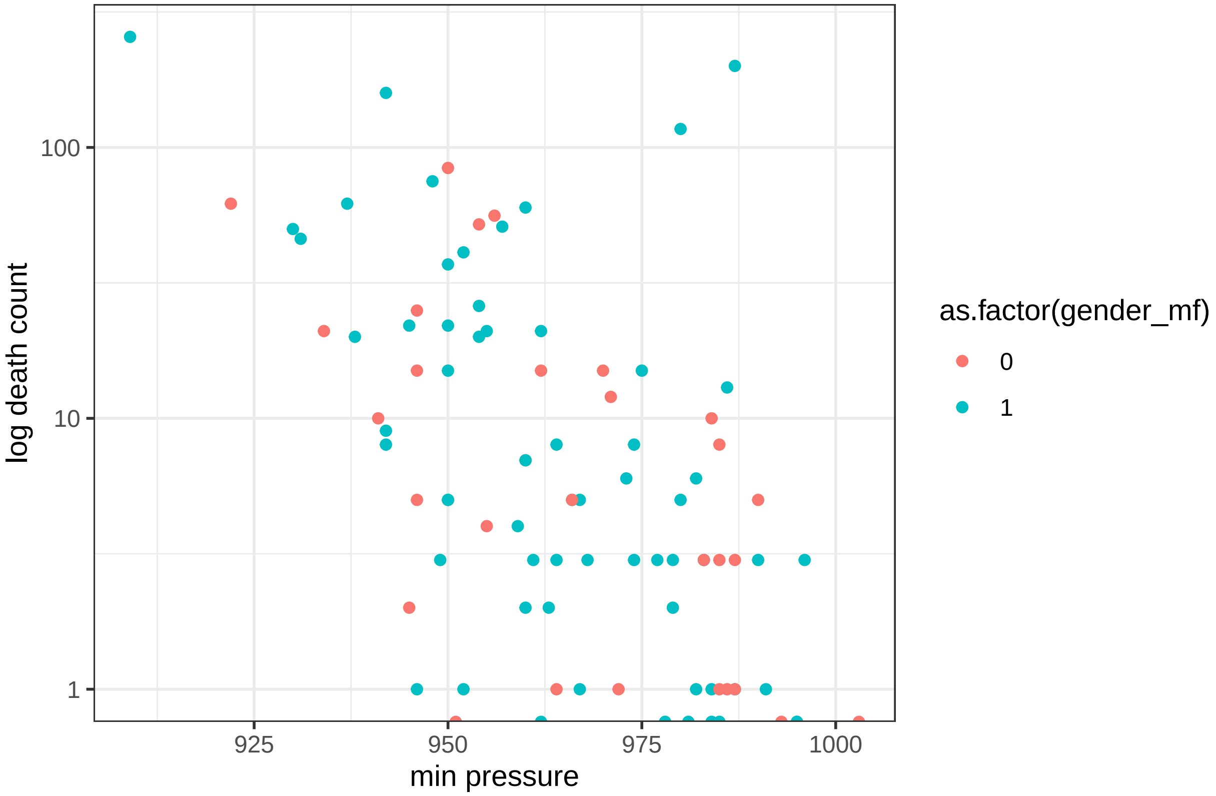
[1] I was able to reproduce the coefficient estimates using the data available but the standard errors were slightly different, so don't worry if that is what you find.

a.

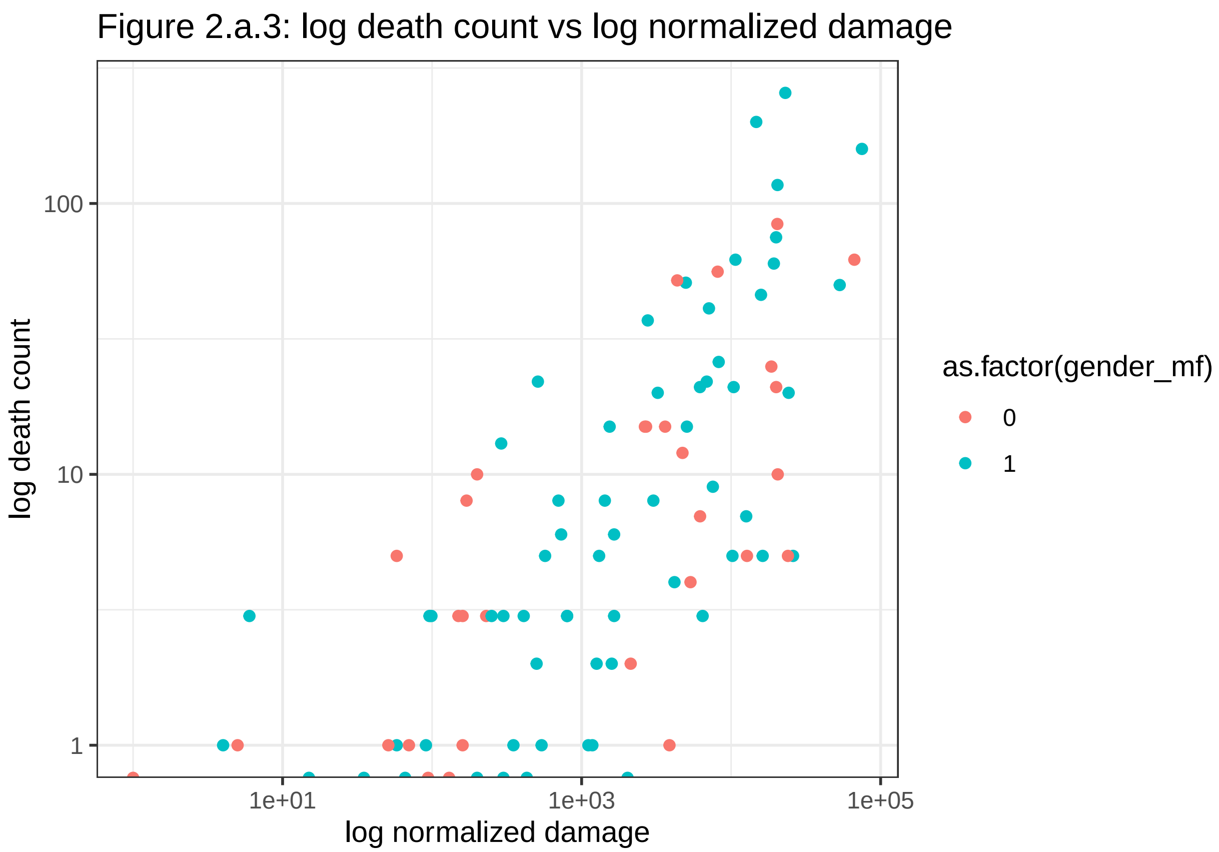


- Hurricanes with higher femininity has some extreme cases with lots of deaths.
- For cases under 100 deaths, both gender hurricanes are similar.
- Few hurricanes have a MFI score between around 4.5 to 6, making a somewhat clear boundary between male and female named hurricanes in terms of MFI.

Figure 2.a.2: log death count vs min pressure



- Death count is negatively correlated with minimum pressure. Gender of the hurricane does not seem to have an effect on this.



- Death count is positively correlated with normalized damage. Gender of the hurricane does not seem to have an effect on this.

b.

```
##
## Call:
## glm(formula = alldeaths ~ mas_fem, family = poisson, data = hurr)
##
## 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 ***
## mas_fem     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
```

- For every 1 point increase in the femininity score if the hurricane, the death count is expected to multiply by roughly exp(0.0738) = 1.0767 times or roughly + 7.7%.
- Comparing the standardized residual sum square with chi-square distribution at 91 df yells a p-value of almost 0, indicating over dispersion exist.

```
##
## Call:
## glm(formula = alldeaths ~ mas_fem, family = quasipoisson, data = hurr)
##
## 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 ***
## mas_fem      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
```

- The p-value for femininity score is no longer significant, indicating we should not draw a conclusion of the relationship between the two variables. If we would trust this model, the result is similar to above: for every 1 point increase in the femininity score if the hurricane, the death count is expected to multiply by roughly exp(0.0738) = 1.0767 times, but with a way higher variance on the beta.

c.

```
##
## Call:
## MASS::glm.nb(formula = alldeaths ~ z_min_pressure_a + zndam +
##     z_mas_fem + z_mas_fem:z_min_pressure_a + z_mas_fem:zndam,
##     data = hurr, 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 ***
## z_min_pressure_a            -0.5521     0.1503  -3.673 0.000239 ***
## zndam                        0.8635     0.1445   5.976 2.28e-09 ***
## z_mas_fem                    0.1723     0.1238   1.392 0.163988
## z_min_pressure_a:z_mas_fem   0.3948     0.1521   2.595 0.009453 **
## zndam:z_mas_fem              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] 0.9508484
```

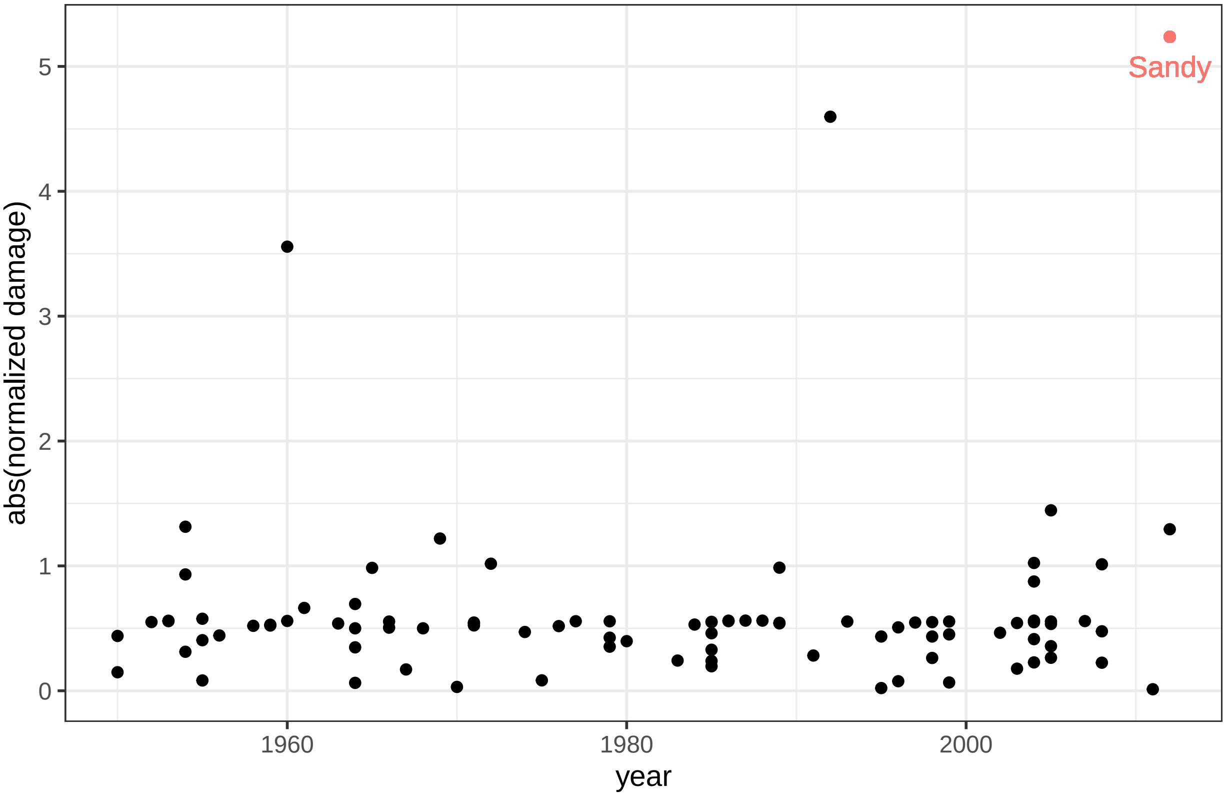
- Every 1 point increase in standardized MFI is expect to multiply the death count by roughly exp(-0.1626) = 0.850 times or roughly - 15.0%.
- Since MFI compare with standardized MFI has a scale of 3.227 and some shifting, every 1 point increase in MFI is expect to multiply the death count by roughly exp(-0.1626/3.227) = 0.951 times or roughly - 4.9%.

d.

```
## $fit
##        1
## 20806.74
##
## $se.fit
##        1
## 16776.47
##
## $residual.scale
## [1] 1
```

- The predicted outcome is 20806 with se.fit being 16776, which is really not looking good, especially when comparing to the true value 159 count. The model itself looks ok with decent p-values. A quick look at the data (there are only 92 rows) showed that "Sandy" has a standardized normalized damage of 5.24, which is way higher than (roughly 4,5 times of) most other hurricanes. Likely the model has its best prediction power within the normal range of standardized normalized damage (0 to 1.5), and this 5.24 is causing the model to be overly sensitive with the prediction. This combines with the exponential likely created this very large prediction and standard error.

Figure 2.d: abs(normalized damage) vs year --- Sandy



e.

Strength:

- Using a continuous variable MFI as the focused predictor instead of just a binary variable male/female. This makes it possible to look at each individual name and predict its influence base on its own femininity score, not just rigid male/female standard.

- Although the data only has 92 rows due to its nature, multiple experiments are conducted to provide more information and make up for the data.

Weakness:

- Although predicted fatality counts for high normalized damage hurricanes increases as MFI increases, for low normalized damage hurricanes the predicted fatality counts actually decreases as MFI increases. The paper states MFI has no effect for less severe storms, but figure. 1 showed a somewhat clear decreasing pattern on fatality counts. I do not know if low damage hurricanes are more often, but I think it may be worth looking deeper into.
- There are 62 female named hurricanes and 30 male named hurricanes. Since the gender of hurricanes are alternated between male and female since late 1970s, this indicates that roughly a third of the hurricanes are from earlier than 1970s where less protective technologies exist, and all these hurricanes are female named. This may lead to a biased sample. And it does appear from the graphs below that the majority of the high death count/ high damage female hurricanes are before 1970s. In fact, I repeated their model 4 but only with hurricanes appeared at 1978 or later, and none of the MFI terms were significant (note only 50+ rows of data after filter for 1978 or later).
- "The practice of naming hurricanes solely after women came to an end in 1978 when men's and women's names were included in the Eastern North Pacific storm lists." https://www.nhc.noaa.gov/aboutnames_history.shtml

Figure 2.e.1: death count vs gender

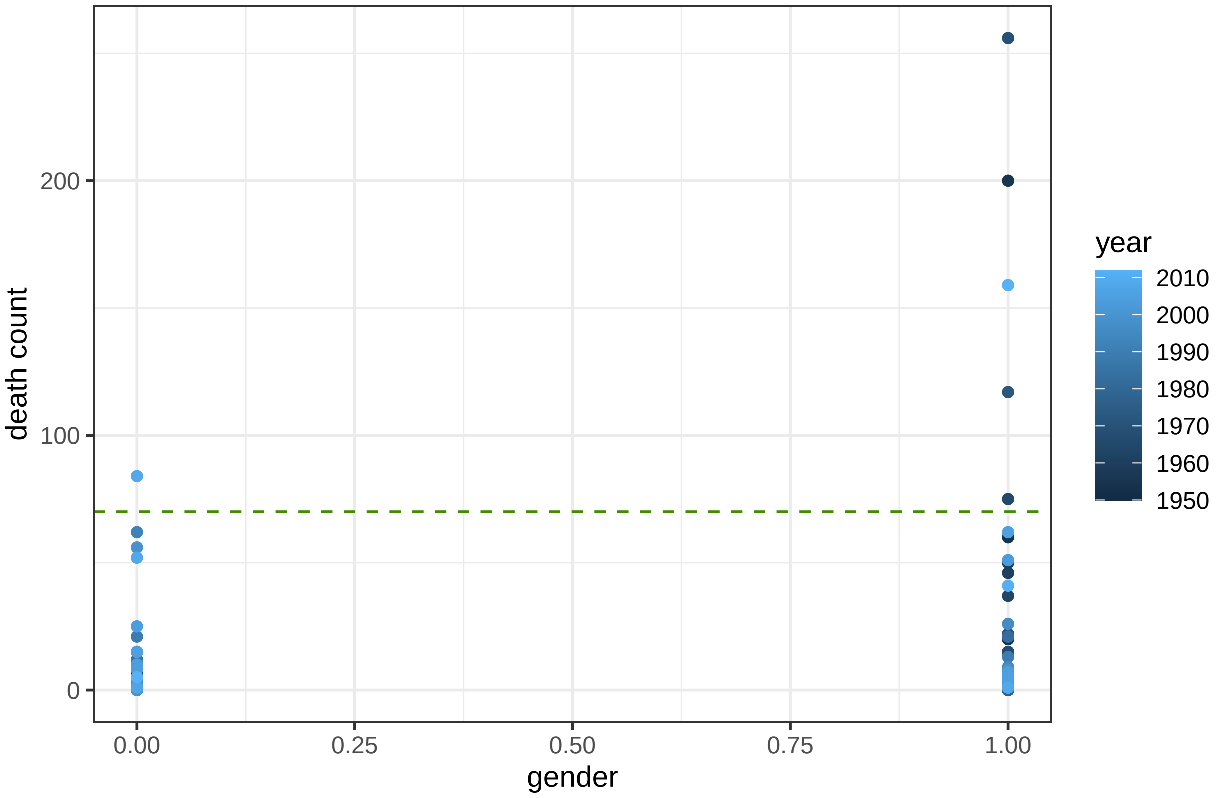
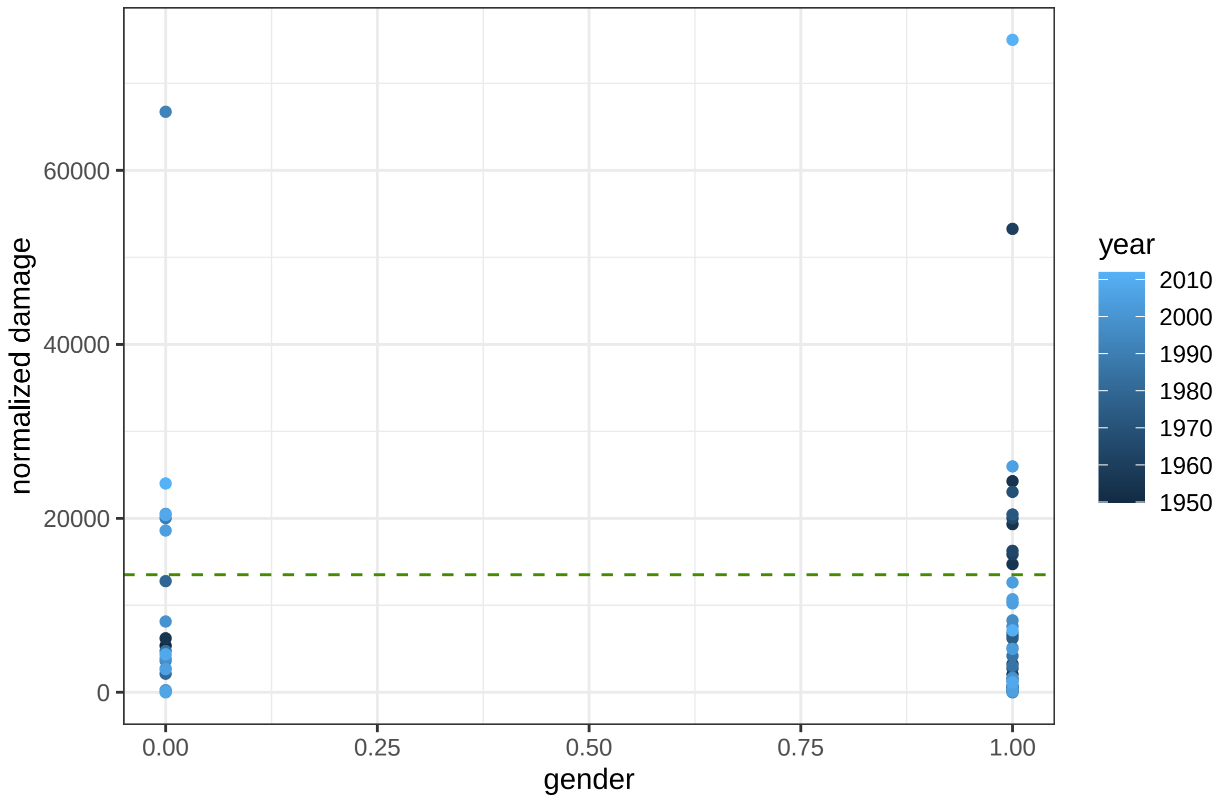


Figure 2.e.2: normalized damage vs gender



```
##
## Call:
## MASS::glm.nb(formula = alldeaths ~ z_min_pressure_a + zndam +
##     z_mas_fem + z_mas_fem:z_min_pressure_a + z_mas_fem:zndam,
##     data = hurr2, init.theta = 1.157979977, link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6529  -1.0374  -0.5222   0.5804   2.0240
##
## Coefficients:
##                              Estimate Std. Error z value Pr(>|z|)
## (Intercept)                 2.3015745  0.1444253  15.936  < 2e-16 ***
## z_min_pressure_a           -0.7660607  0.1932454  -3.964 7.36e-05 ***
## zndam                       0.2488859  0.1430168   1.740   0.0818 .
## z_mas_fem                  -0.0005155  0.1268688  -0.004   0.9968
## z_min_pressure_a:z_mas_fem  0.2177418  0.1656673   1.314   0.1887
## zndam:z_mas_fem             0.2707106  0.1387372   1.951   0.0510 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(1.158) family taken to be 1)
##
```

```
##     Null deviance: 113.953  on 53  degrees of freedom
## Residual deviance:  58.477  on 48  degrees of freedom
## AIC: 376.7
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  1.158
##           Std. Err.:  0.243
##
##  2 x log-likelihood:  -362.697
```

f.
- I am not convinced by the result. I think more data from the time range which both male and female hurricane names exist is needed. Due to the nature of the data, it might be necessary to look at world wide data instead of just US data, this should also helps remove the potential bias of English names only.
- As for the analyses, in model 2 (min pressure, ndam and MFI), MFI is not significant. It may be worth looking at a model without MFI or MFI interactions and compare to see how much MFI improves the model.

3 Vaccinations

This question 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.

- You can download the latest data on vaccination coverage here: https://data.cdc.gov/Vaccinations/COVID-19-Vaccinations-in-the-United-States-County/8xkx-amqh/data. Note that this is updated most days so depending on when you download it, it might be slightly different from others (that's okay). For the purposes of the assignment, please consider data from the 11th of January 2023. Also note that on the same webpage you should be able to find a data dictionary. We will be interested in people who have completed a primary vaccine series (have second dose of a two-dose vaccine or one dose of a single-dose vaccine), which refers to columns that have the `Series_Complete` prefix.
- The class repo has a dataset `acs` that contain 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`. For reference, the extraction code can be found in the repo (`acs.R`)

a) Perform some exploratory data analysis (EDA) using a dataset combining the vaccination and ACS data, and summarize your observations with the aid of 3-4 key tables or graphs.

b) 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 completed a primary vaccine series). There is no one right answer here, but you should justify the outcome measure you are using (e.g. counts, proportions, rates, etc) and your distributional assumptions about the outcome measure (e.g. binary, poisson, normal, etc). You should also discuss briefly your model building strategy; what covariates you considered and why (motivated by your EDA)[2], and how the candidate model was chosen. Interpret your findings, including visualizations where appropriate.

c) Use your model from b) to predict the proportion of the population aged 18+ in Ada County, Idaho. Briefly discuss how good you think this prediction is, and why.

d) Give a brief summary of your analysis. What other variables may be of interest to investigate in future?

e) Now consider the situation of analysing vaccination rates at the **state** level. Consider the three following options:

   1) Regression at the state level, outcome used is the total population 18+ fully vaccinated
   2) Regression at the state level, outcome used is the average of the county level full vaccination rates of 18+ population
   3) Regression at the county level, outcome used is the total population 18+ fully vaccinated, and include as a covariate a categorical variable (fixed effect) which indicates which state a county is in.

   Without performing these regressions, briefly discuss how you think these three approaches would differ in terms of the granularity of information used and the type of outcome measure. In your opinion which is the most appropriate analysis, or does it depend on the question being asked?

[2] Note that the vaccines dataset also has a `Metro` variable which you are welcome to use in your analyses.

a.

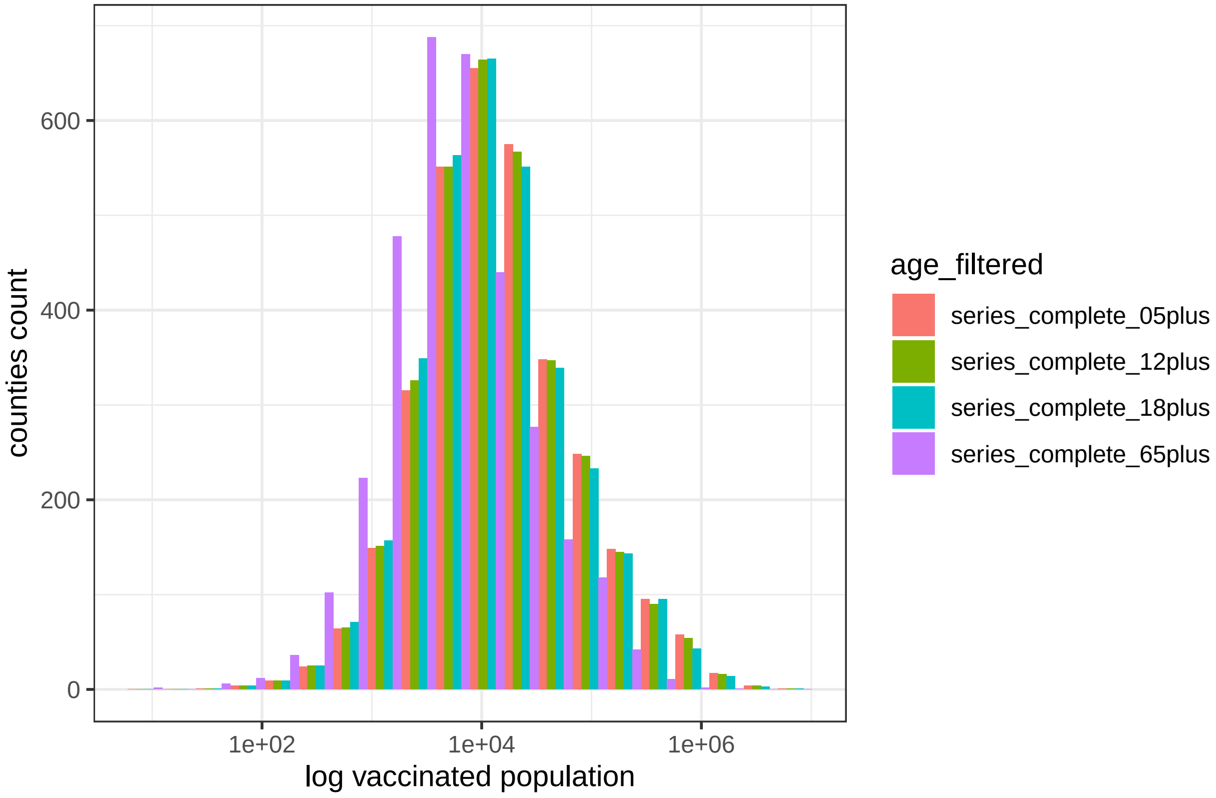
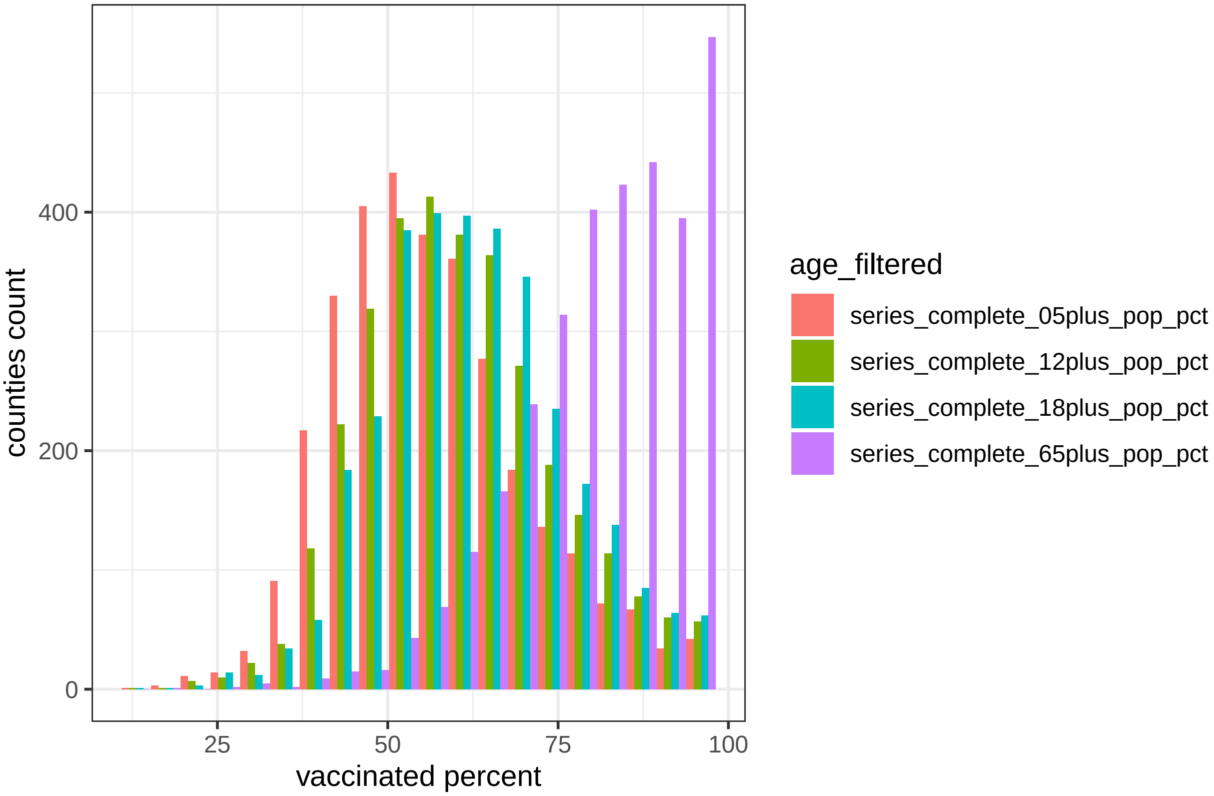


Figure 3.a.1: counties count by log vaccinated population

Figure 3.a.2: counties count by vaccinated percent



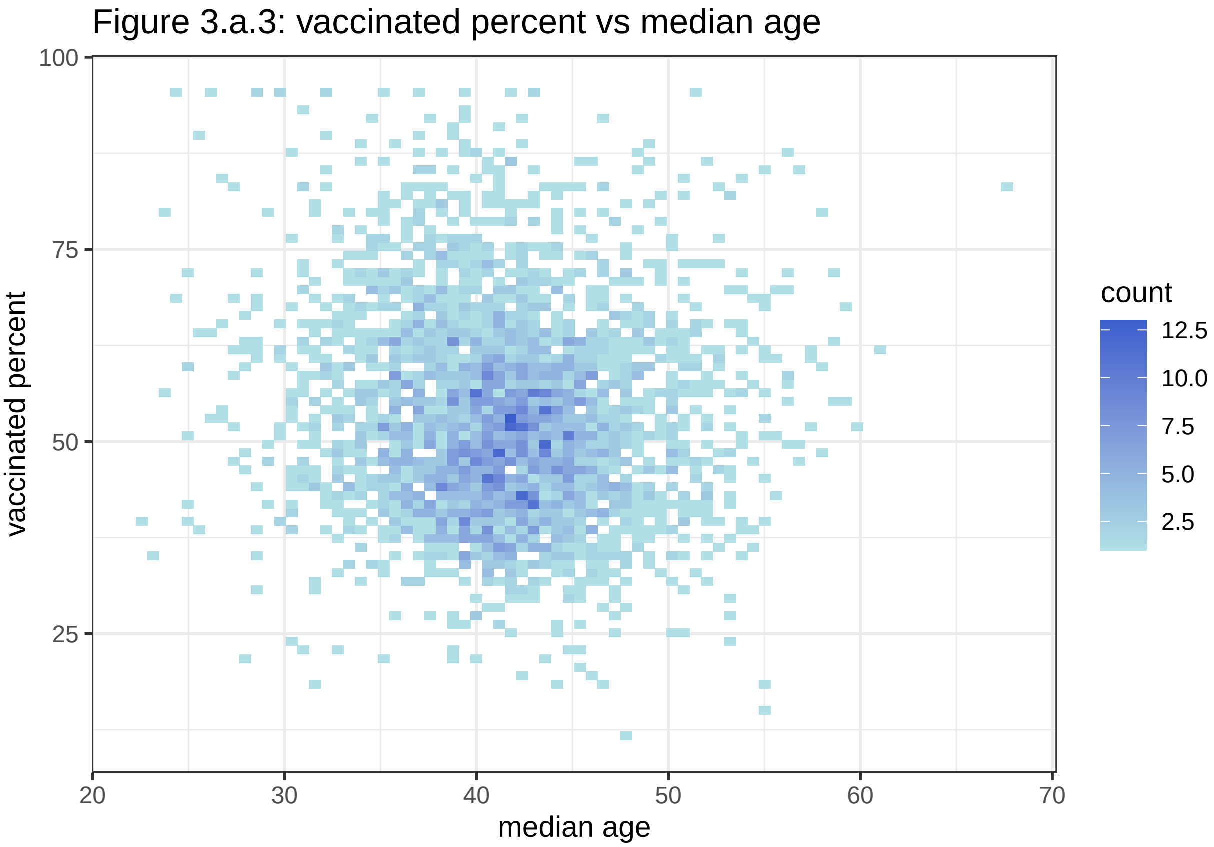
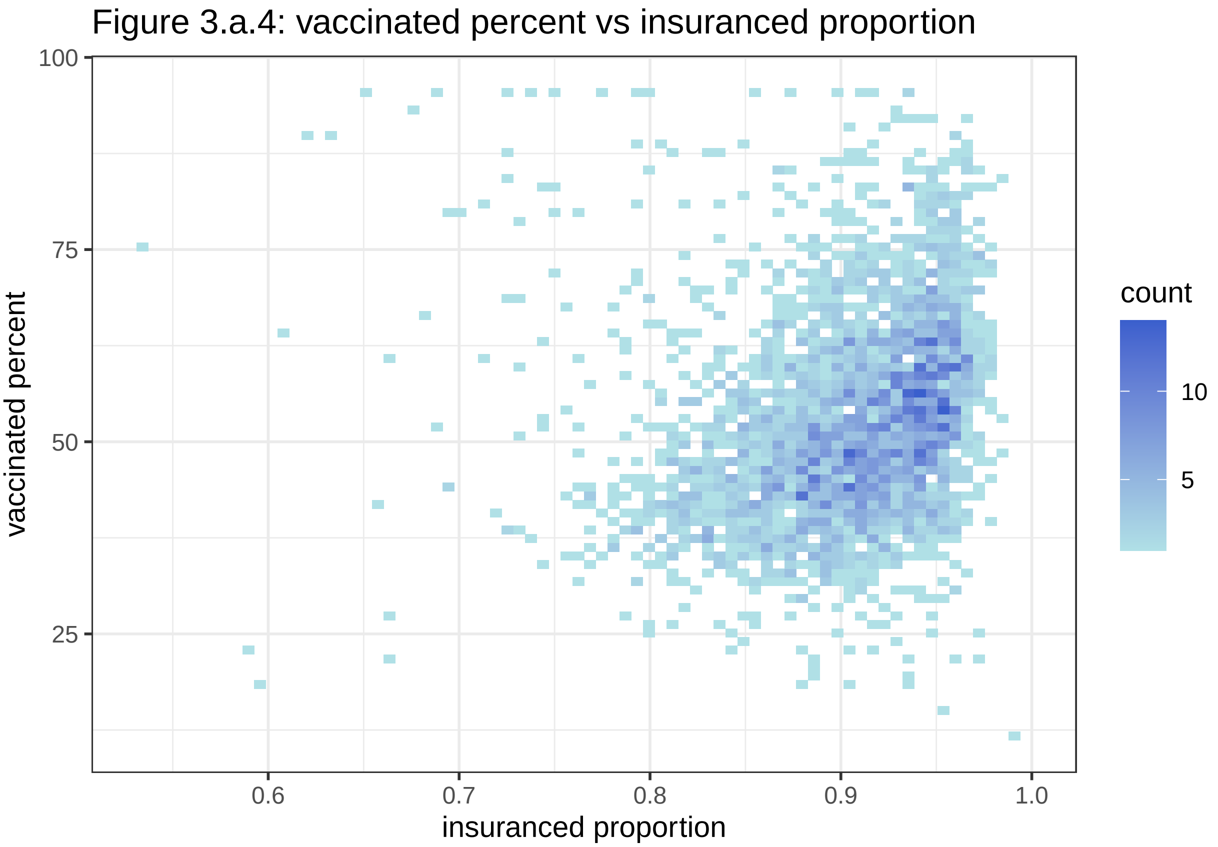


Figure 3.a.3: vaccinated percent vs median age
vaccinated percent
100
75
50
25
median age
20
30
40
50
60
70
count
12.5
10.0
7.5
5.0
2.5



- The log of series completed population for different counties is bell shaped for all age filters, ranging from less than 100 to more than 10^6.

- The percent of series completed across the counties is also bell shaped in general. However when we raise the filtering age (5+ to 12+ to 18+ to 65+), the distribution becomes more and more left skewed and denser at higher percentages. This becomes quite recognizable when filtered to 65+ only. (It may be more proper to calculate the percentage for age 5-11, 12-17, 18-64 and 65+ separately, but I think here the current graph draws a good enough picture to indicate that age may be worth looking at)

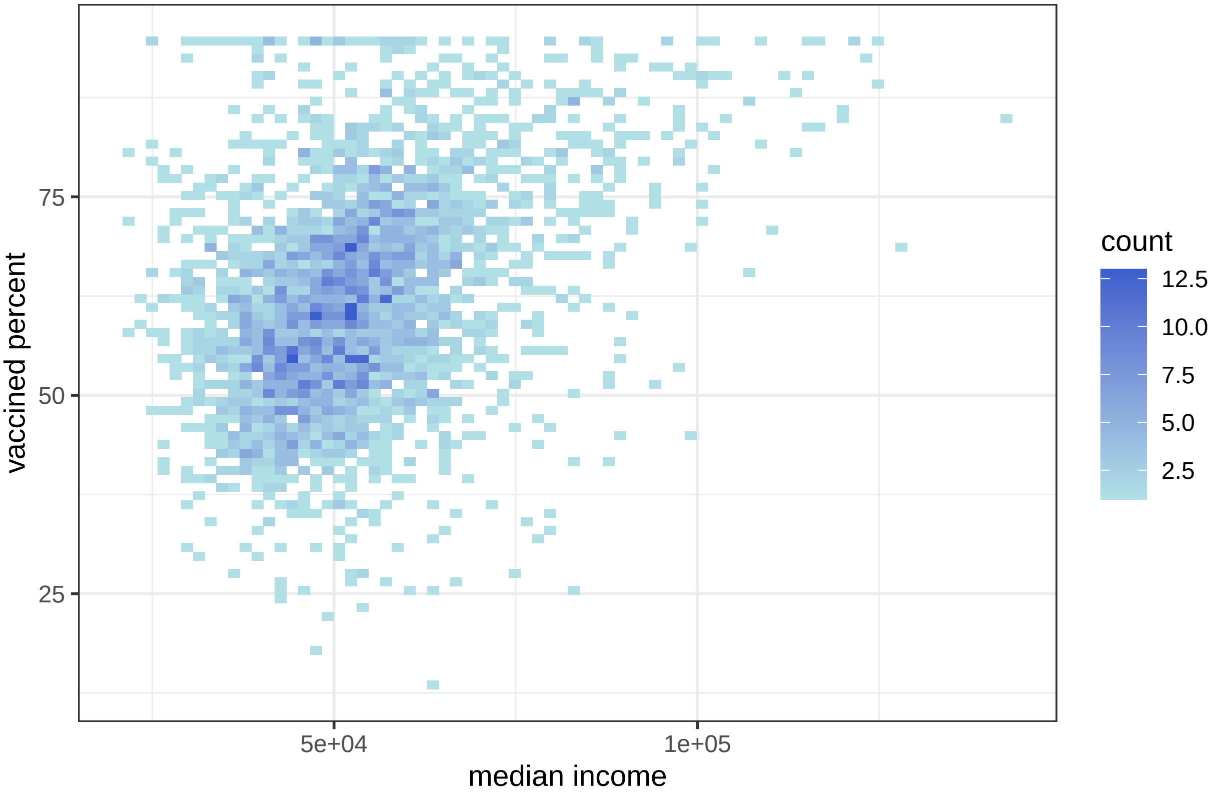
- So next I looked at the relationship between series completed percentage and median age of the counties, but there does not appear to be a relationship between the 2 variables. It appears that the age effect becomes unrecognizable when only looking at the median.

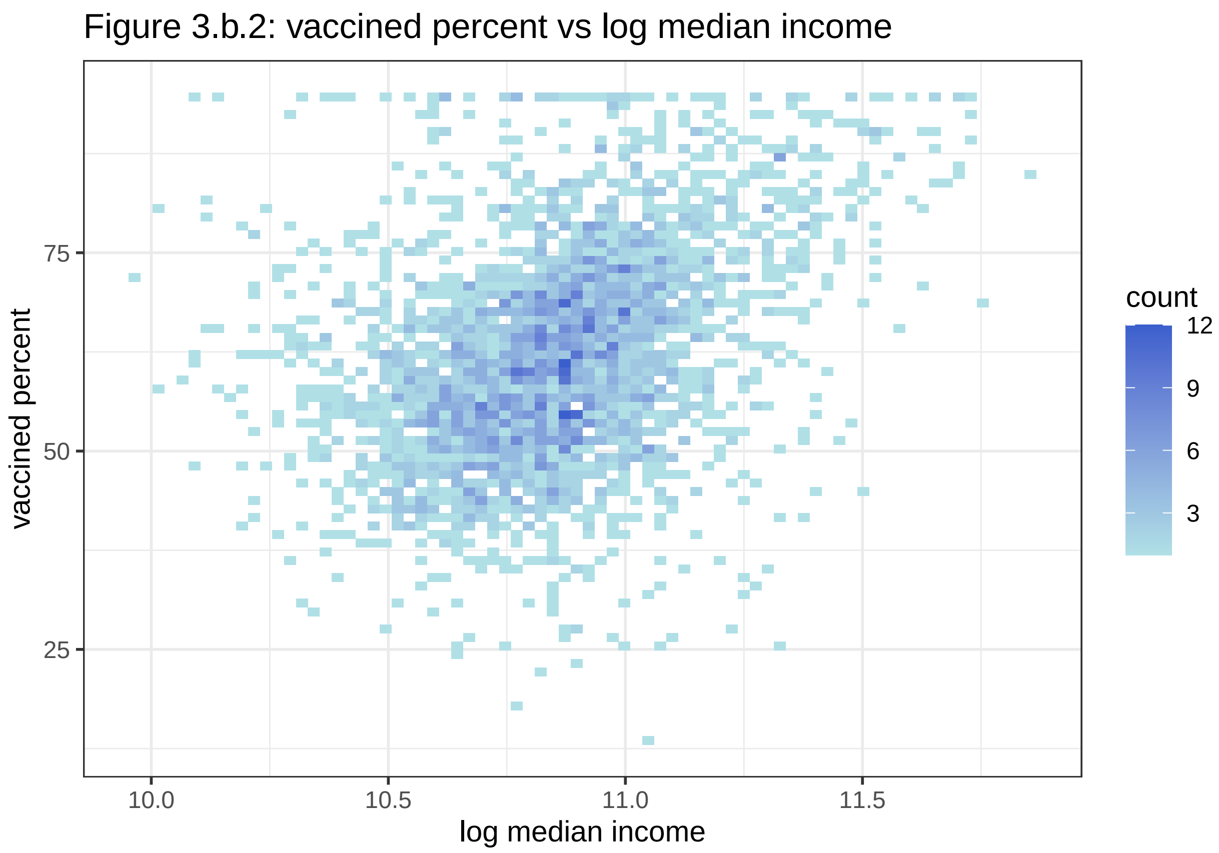
- Next I looked at another potential predictor, health insurance proportion, there appears to be a clear positive correlation between the two variables.

b. - Since we are investigating the patterns in 18+ full vaccination rate, I will use a binomial regression model here with 18+ full vaccination rate as outcome. As for predictors, since I have little technical knowledge about vaccination, ideally I should start with a full model and use model selection methods such as AIC or BIC to select several well fitted potential final models, then consult someone knowledgeable in vaccination to choose one as the final model. However, since this is an assignment and for demonstration purpose, I will pretend that I know something in vaccination, using information from the EDA and my common sense to select the "reasonable" predictors, then use the AIC method to select a final model.

- I decide to use prop_white, prop_foreign_born, median_age, median_income, prop_bachelor_above, prop_unemployed and prop_health_insurance as predictors before the AIC model selection.
- For predictors, it may be better to logistic transform the ratio predictors so that they match the transformation of the outcome variable and become easier to explain the relationships, however since some of them contain (and can reasonably contain) a ratio of 0 or 1, and I do not wish to lose these information after the logistic transform, they will be kept as their original form. For other variables, prop_less_than_hs is dropped because it is likely highly correlated to prop_bachelor_above, which we included, and prop_nilf is dropped for the same reason. For prop_low_ratio_ip, if my understanding of its description is correct, is potentially highly correlated to median_income, so it is also dropped. One more thing is to see if it is proper to log transform the financial term median_income:

Figure 3.b.1: vaccined percent vs median income





- Comparing the two plots, I think a log transformation on median_income is justifiable.
- It is worth noting here that after left joining the two table, for the same county, in rare cases, series_complete_18plus from vaccine table is bigger than total_pop_18plus from the acs table. This causes the glm(cbind(y, m - y) ~ .) binomial model to not function. Since the numbers from these two tables are off, any attempt that mixes the population or population proportion data from the two tables is also not justifiable, so I used series_complete_18plus_pop_pct and series_complete_18plus from only the vaccine table to calculate y and m for the glm.

```
##
## Call:
## glm(formula = cbind(series_complete_18plus, total_18plus - series_complete_18plus) ~
##     ., family = binomial, data = dat_q3_reg)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -377.41   -21.51    -3.73    14.76   441.88
##
## Coefficients:
##                         Estimate Std. Error z value Pr(>|z|)
## (Intercept)           -3.5085963  0.0104107 -337.02   <2e-16 ***
## prop_white             0.1985820  0.0013387  148.34   <2e-16 ***
## prop_foreign_born      2.8578692  0.0017689 1615.66   <2e-16 ***
## prop_bachelor_above    2.8058182  0.0022534 1245.14   <2e-16 ***
## prop_unemployed        6.5647581  0.0221256  296.70   <2e-16 ***
## prop_health_insurance  2.2566244  0.0042121  535.75   <2e-16 ***
## log_median_income      0.0228706  0.0010380   22.03   <2e-16 ***
## median_age             0.0163254  0.0000381  428.52   <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: 21215373  on 3125  degrees of freedom
## Residual deviance:  7702710  on 3118  degrees of freedom
##   (157 observations deleted due to missingness)
## AIC: 7734952
##
## Number of Fisher Scoring iterations: 4
```

- The model looks quite decent with all predictors being significant. So for the model selection process, I will not remove any terms and just select this one as the final model base on the p-values. I will double check using an automated step AIC method to see if any term could be deleted:

```
##
## Call:
## glm(formula = cbind(series_complete_18plus, total_18plus - series_complete_18plus) ~
##     prop_white + prop_foreign_born + prop_bachelor_above + prop_unemployed +
##         prop_health_insurance + log_median_income + median_age,
##     family = binomial, data = dat_q3_reg)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -377.41   -21.51    -3.73    14.76   441.88
##
## Coefficients:
##                         Estimate Std. Error z value Pr(>|z|)
## (Intercept)           -3.5085963  0.0104107 -337.02   <2e-16 ***
## prop_white             0.1985820  0.0013387  148.34   <2e-16 ***
## prop_foreign_born      2.8578692  0.0017689 1615.66   <2e-16 ***
## prop_bachelor_above    2.8058182  0.0022534 1245.14   <2e-16 ***
## prop_unemployed        6.5647581  0.0221256  296.70   <2e-16 ***
## prop_health_insurance  2.2566244  0.0042121  535.75   <2e-16 ***
## log_median_income      0.0228706  0.0010380   22.03   <2e-16 ***
## median_age             0.0163254  0.0000381  428.52   <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: 21215373  on 3125  degrees of freedom
## Residual deviance:  7702710  on 3118  degrees of freedom
##   (157 observations deleted due to missingness)
## AIC: 7734952
##
## Number of Fisher Scoring iterations: 4
```

- The results indeed match. This model is selected to be the final model.
- Now for the findings,
- Due to the properties of the predictors, the interpretations of findings will differ slightly:

$$\begin{aligned} rate &: e^{\beta(x+0.01)} = e^{\beta x} \cdot e^{0.01\beta} \\ log &: e^{\beta \cdot log(x+1)} = e^{\beta}(x+1) = e^{\beta}x + e^{\beta} \\ basic &: e^{\beta(x+1)} = e^{\beta x} \cdot e^{\beta} \end{aligned}$$

- Every 1 percent increase in proportion of white population will multiply the odds ratio for 18+ full vaccination rate by roughly 1.002. Similarly, every 1 percent increase in proportion of 18+ population who are foreign born, proportion of population aged 25+ who have at least a bachelor degree, proportion of population age 16+ who are unemployed and proportion of population with health insurance multiplies the odds ratio for 18+ full vaccination rate by roughly 1.029, 1.028, 1.068 and 1.023 respectively.
- Every 1 (2019 inflated adjusted) dollar increase the ratio by roughly 1.023, and every one year increase in median age multiply the ratio by roughly 1.016.

Predictor	Increment	Odds ratio for 18+ full vaccination rate
proportion white population	+ 1%	x 1.001988 (+ 0.20%)
proportion foreign born	+ 1%	x 1.028991 (+ 2.90%)
proportion at least bachelor	+ 1%	x 1.028456 (+ 2.85%)
proportion unemployed	+ 1%	x 1.06785 (+ 6.79%)
population health insurance	+ 1%	x 1.022823 (+ 2.28%)
(inflated adjusted) median income	+ 1$	+ 1.023134
median age	+ 1yr	x 1.016459 (+ 1.65%)

c.

```
## $fit
##         1
## 0.7403803
##
```

```
## $se.fit
##            1
## 5.777221e-05
##
## $residual.scale
## [1] 1
```

- The predicted value is 0.7403, which is 74.03%. As the model has good p-values, and se.fit is small, I think the prediction will be good. A comparison with the recorded percentage from the vaccine data, 76.9%, also seems very acceptable.

d.

- The fully vaccinated rate for 18+ population for a county is positively related to the county's 18+ population's proportion of white, proportion of foreign born, proportion of bachelor or above degree, proportion of unemployment, proportion of health insurance, log of median income and median age of the county, based on a binomial regression. All above predictors are statistically significant. Among the proportion predictors, unemployment proportion is the most impactful with an estimated beta of 6.56; the white proportion is the least significant with a estimated beta of 0.20; the rest of the proportion predictors all have betas between 2 and 3. For non-proportion predictors, log(median income) and median age have beta values 0.023 and 0.163 respectively.
- As for other variables of interest that may be investigated in the future, in the current data set there is no information about medical facilities, for example data about the clinics of a given county. I think data like this can potentially have a great impact on the vaccinated rate, and may be worth investigating in the future.

e.
- In terms of granularity of information used, option 2 needs the least granularity with only the vaccinated rate from the data. Option 1 needs more granularity than option 2 because calculating vaccinated rate requires more than vaccinated count, total population is also needed as the denominator. Option 3 needs the most granularity, as county is included as a new level of granularity.
- As for which model is appropriate, model 2 is probably not appropriate, assuming it is just an average and not a weighted average.
- For model 1 and 3, I think the answer somewhat depends.
  - If the data we have is at state level, then we can use model 1 to explain the relationship between a state's vaccinated rate and other variables of that state. Here model 3 is not viable due to the structure of the data.
  - If the data we have is at county level, model 3 is a viable option because we can perform a weighted average of outcome county vaccinated rate to get the state vaccinated rate. We cannot directly see the relationship between a state's vaccinated rate and other variables of that state this way due to the structure of the data, but we do get the state itself as a covariate included in our model, allowing us to see the its effect as a leveled factor. If we are able to use some averaging method to transform the data from county level into state level, model 1 can also be used here similar to other state level analyses.