PriyankaVerma-assignment1-solution

Q1 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 $Var(Y) = \mu(1 + \mu\sigma^2)$.

$$E[Y] = E[E[Y|\theta]] \quad -Law \quad of \quad total \quad expectation$$

$$= E[E[\frac{e^{-(\mu\theta)}(\mu\theta)^y}{y!}]]$$

$$= E[\mu\theta]$$

$$= \mu E[\theta]$$

$$= \mu$$

$$Var[Y] = E[Var[Y|\theta]] + Var[E[Y|\theta]]-Law \quad of \quad total \quad variance$$

$$= E[\mu\theta] + Var[\mu\theta]$$

$$= \mu E[\theta] + \mu Var[\theta]$$

$$= \mu + \mu\sigma^2$$

$$= \mu(1 + \sigma^2)$$

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.

$$p(y) = p(y|\theta)p(\theta)d\theta$$

$$= \int \frac{e^{-(\mu\theta)}(\mu\theta)^y}{y!} * \frac{\theta^{\alpha-1}e^{-\frac{\theta}{\beta}}}{\beta^\alpha\Gamma(\alpha)}d\theta$$

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

$$= \frac{\mu^y}{y!\beta^\alpha\Gamma(\alpha)}\frac{\Gamma(\alpha+y)}{\left(\frac{\beta\mu+1}{\beta}\right)^{\alpha+y}}\int\frac{\left(\frac{\beta\mu+1}{\beta}\right)^{\alpha+y}}{\Gamma(\alpha+y)}e^{-(\frac{\beta\mu+1}{\beta})\theta}\theta^{y+\alpha-1}d\theta$$

$$= \frac{\mu^y}{y!\beta^\alpha\Gamma(\alpha)}\frac{\Gamma(\alpha+y)\beta^{\alpha+y}}{(\beta\mu+1)^{\alpha+y}}$$

$$= \frac{\Gamma(\alpha+y)}{\Gamma(\alpha)\Gamma(y+1)}\frac{\mu^y\beta^y}{(\beta\mu+1)^{\alpha+y}}$$

$$= \frac{\Gamma(\alpha+y)}{\Gamma(\alpha)\Gamma(y+1)}\left(\frac{\beta\mu}{\beta\mu+1}\right)^y\left(\frac{1}{\beta\mu+1}\right)^\alpha$$

$$= NB(\alpha, 1/(\beta\mu+1))$$

Hence, this is a negative binomial distribution.

c) In order for $E(Y) = \mu$ and $Var(Y) = \mu(1 + \mu\sigma^2)$, what must α and β equal?

For a Gamma distribution with parameters α and β, for a random variable X with this distribution, $E(X) = \alpha\beta$ and $Var[X] = \alpha\beta^2$. Proof:

$$E[X^a] = \int \frac{x^{a+\alpha-1}e^{-x/\beta}}{\beta^\alpha \Gamma(\alpha)} dx$$

$$= \frac{\beta^a}{\Gamma(\alpha)} \int (\frac{x}{\beta})^{a+\alpha-1} e^{(-x/\beta)} dx$$

$$= \frac{\beta^a}{\Gamma(\alpha)} \Gamma(a + \alpha)$$

$$E[X] = \frac{\beta}{\Gamma(\alpha)} \alpha\Gamma(\alpha) = \beta\alpha$$

Proof for variance-

$$E[X^2] = \beta^2 \frac{\Gamma(2 + \alpha)}{\Gamma(\alpha)}$$

$$= \beta^2 \alpha(\alpha + 1)$$

$$Var[X] = E[X^2] - E[X]^2 = \beta^2\alpha(\alpha + 1) - \alpha^2\beta^2 = \beta^2\alpha$$

Using part (a) we know that $E[Y] = \mu E[\theta]$. Therefore, combining it with the results of gamma distribution:

$$E[Y] = \mu E[\theta] = \mu\alpha\beta = \mu$$

$$=> \alpha\beta = 1$$

$$Var[Y] = \mu E[\theta] + Var[\mu\theta] = \mu E[\theta] + \mu^2 Var[\theta]$$

$$= \mu\alpha\beta + \alpha\mu^2\beta^2 = \alpha\mu\beta(1 + \mu\beta)$$

$$\alpha\beta(1 + \mu\beta) = (1 + \mu\sigma^2)$$

using $\alpha\beta = 1$, we get-

$$(1 + \mu\beta) = (1 + \mu\sigma^2)$$

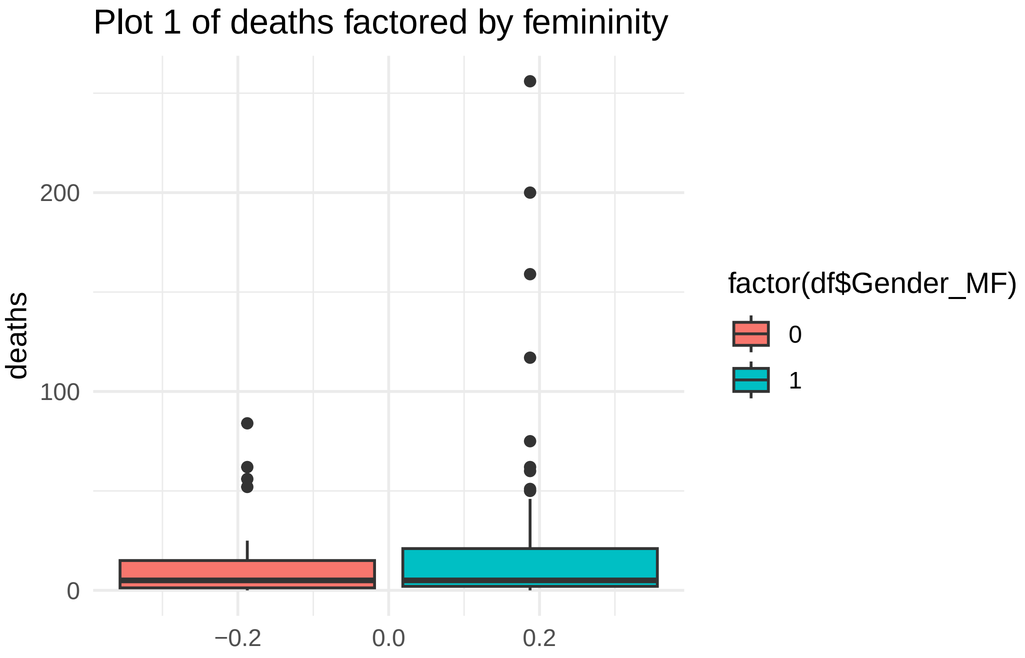
$$=> \beta = \sigma^2$$

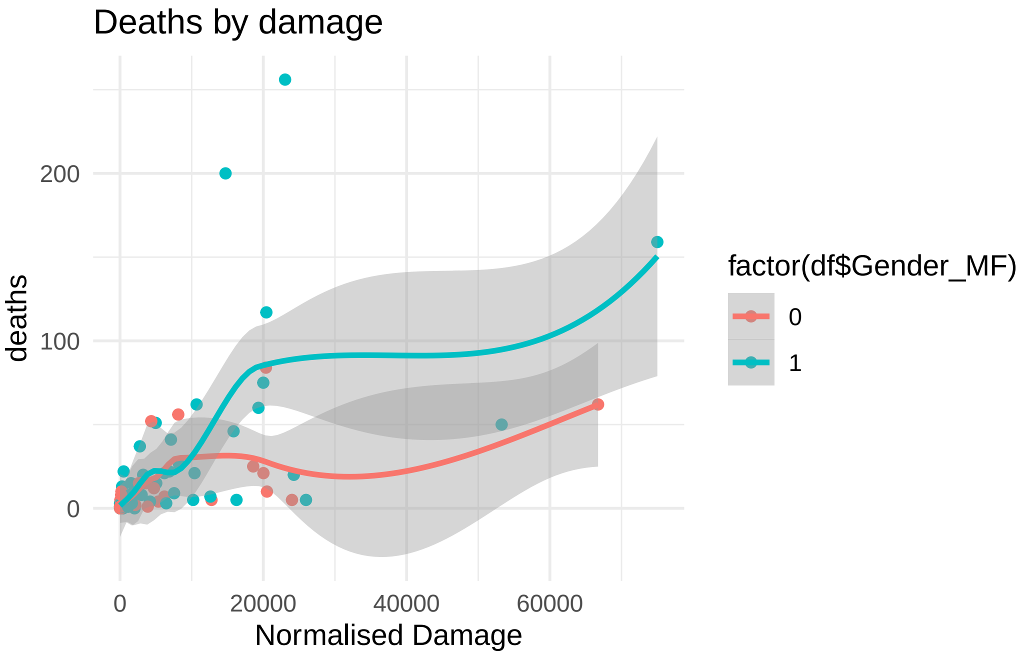
$$=> \alpha = 1/\sigma^2$$

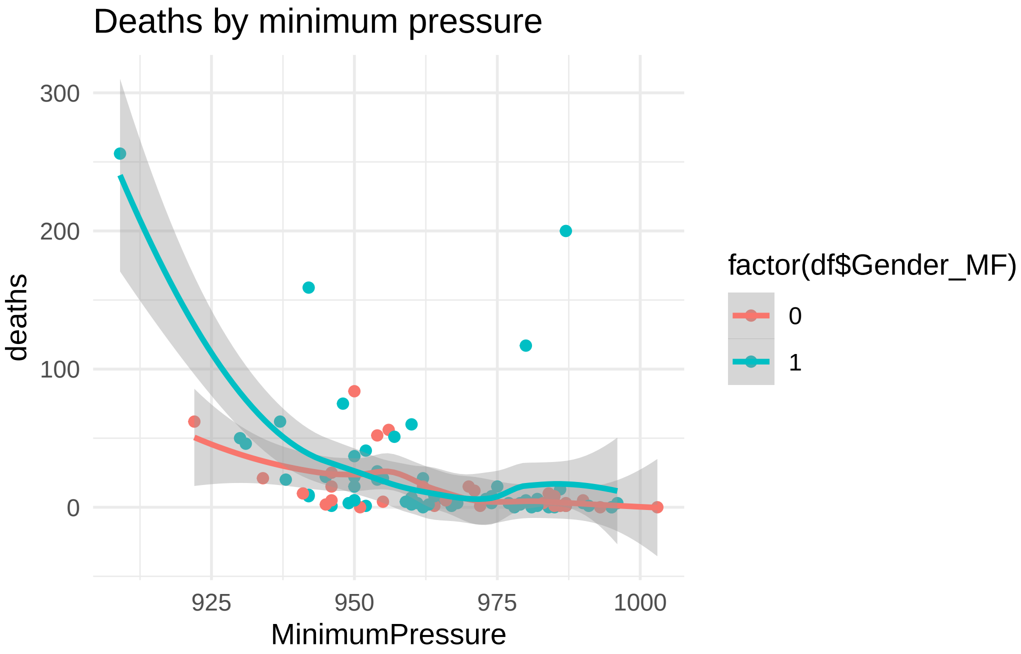
Q2 Hurricanes

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.

- The distribution of deaths is left skewed and not-normally distributed. Most of the deaths caused by hurricanes are under 80.
- The plot of death by femininity shows that the median deaths is same for both Masculine and Feminine hurricanes, however the feminine classified hurricanes have higher IQR and more outliers that have caused great number of deaths.
- The plot of death by normalised damage shows a positive correlation between deaths and normalised damage overall. The pattern exists individually too for both feminine- and masculine- classified hurricanes. We can also see that for hurricanes lower in normalized damage the death toll is similar for both masculine-named and feminine named hurricanes, whereas for hurricanes higher in normalized damage hurricanes with feminine names caused more deaths than those with masculine names.
- The plot of death by minimum pressure shows a negative correlation between deaths and normalised damage overall. The pattern exists individually too for both feminine- and masculine- classified hurricanes.







b) Run a Poisson regression with deaths as the outcome and femininity as the explanatory variable.

Interpretation of the resulting coefficient estimates For the poisson regression model, we note that femininity is statistically significant predictor with a p value less than 0.05. The intercept means that if the femininity of a hurricane is 0 then it would cause deaths of log (2.500370). Further, one unit change in feminine hurricanes has 1- exp(0.0738) = 7% more chance of causing deaths.

```
Call:
glm(formula = alldeaths ~ MasFem, family = poisson, data = df)

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
```

check for overdispersion - Null Hypthesis: the overdispersion is zero

```
n = dim(df)[1] #119
k = length(mod1$coefficients) # k=3
sum(rstandard(mod1)^2)/(n-k) # overdispersion factor
```

```
[1] 44.6563
```

```
1- pchisq(sum(rstandard(mod1)^2), n-k) #test value for overdispersion
```

```
[1] 0
```

- The chi-squared statistic value is 0 < 0.05 which means the null hypothesis can be rejected. thereby, implying that the overdispersion exists. As a result, I run a quasi Poisson regression model.
- However, the coefficient of femininity variable is not statistically significant as p-value > 0.05.

```
Call:
glm(formula = alldeaths ~ MasFem, family = quasipoisson, data = df)

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
```

c) Reproduce Model from the paper

```
df2 <- df %>% mutate_at(c('MasFem', 'NDAM', 'MinPressure_before'), ~(scale(.) %>% as.vecto
mod3 <- glm.nb(alldeaths ~ MinPressure_before + NDAM + MasFem + MasFem*MinPressure_before
summary(mod3)
```

```
Call:
glm.nb(formula = alldeaths ~ MinPressure_before + NDAM + MasFem + 
    MasFem * MinPressure_before + MasFem * NDAM, data = df2, 
    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 ***
MinPressure_before         -0.5521     0.1503  -3.673 0.000239 ***
NDAM                        0.8635     0.1445   5.976 2.28e-09 ***
MasFem                      0.1723     0.1238   1.392 0.163988    
MinPressure_before:MasFem   0.3948     0.1521   2.595 0.009453 ** 
NDAM:MasFem                 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 
```

```
#glm.nb(formula = alldeaths ~ min_pressure_before + ndam + mas_fem +   mas_fem * min_press
```

the estimated effect of femininity on deaths assuming a hurricane with median pressure and damage ratings

```
net_effect = (0.1723+0.3948 + 0.7051)
net_effect
```

```
[1] 1.2722
```

Therefore, if we keep the pressure and damage ratings constant, for one unit increase in femininity the expected death count would change by log of 1.2722.

d) Using Model 4, predict the number of deaths caused by Hurricane Sandy. Interpret your results.

```
death_sandy <- df2 |> filter(Name == 'Sandy')
predicted_death <- predict(mod3, death_sandy, type = "link")
print(predicted_death)
```

```
       1
9.943032
```

The model predicts that approximately 10 deaths by the hurricane Sandy, whereas the actual deaths in the data are 159. It is one of the outlier values of deaths, as can be seen from “Plot 1: of deaths factored by feminity”. Therefore, the model is not appropriate in predicting outlier values.

e) Describe at least two strengths and two weaknesses of this paper, focusing on the archival analysis. What was done well? What needed improvement?

Strengths One of the strengths of the paper is that the authors perform statistical analysis well, while there are limitations, with the available data and rigorously justified their approaches, for instance, they considered using different count models for model building before finally choosing the negative binomial model. Additionally, the paper is reproducible and the analysis is verifiable as the authors have publicly shared the data and model building approach. Lastly, by conducting this research, the authors have made an important point for practitioners to reconsider how hurricanes are named and communicated to the masses.

Weakness There are various limitations in the paper. First, their model does not take into account factors that influence the likelihood of people getting influenced by hurricanes, for instance population of an area, the route of hurricane, width of hurricane, etc. While some of the data might not have been available, data about the base population around the hurricane could have been included as an offset variable in the model. Additionally, there might have been temporal factors influencing the death rates, such as population changes over the years, development of better infrastructure to cope up with hurricanes, etc. Even with the data considered- the coefficient of the MFI index variable is not statistically significant (As p-value < 0.05), so the authors’ reasoning about the effect of gendered names on protective action does not seem appropriate. Further, the authors have not established the mechanism behind their claim- severe storms with more feminine names are deadlier. Moreover, there is no option of naming a hurricane as neutral, i.e. neither feminine nor masculine. I am curious why a 5-point likert scale was not considered for coding the MFI index, as opposed to using the scale adopted by authors (1 = very masculine, 11 = very feminine, and 1 = very man-like, 11 = very woman-like). Lastly, there is no mention of power analysis and effect size by the authors.

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.

I am not convinced by the results because of the aforementioned limitations. I would like to see data for safety infrastructure like storm shelters and population of an area, along with more meteorological information such as the route of hurricane, width of hurricane. Lastly, I would like to see a more rigorous analysis for the author's hypothesis "individuals systematically underestimate their vulnerability to hurricanes with more feminine names, avoiding or delaying protective measures", however due to limited statistical and meteorological knowledge I am not sure of the correct toolkit to test it- perhaps, causal inference models may help.

Q3 Vaccinations

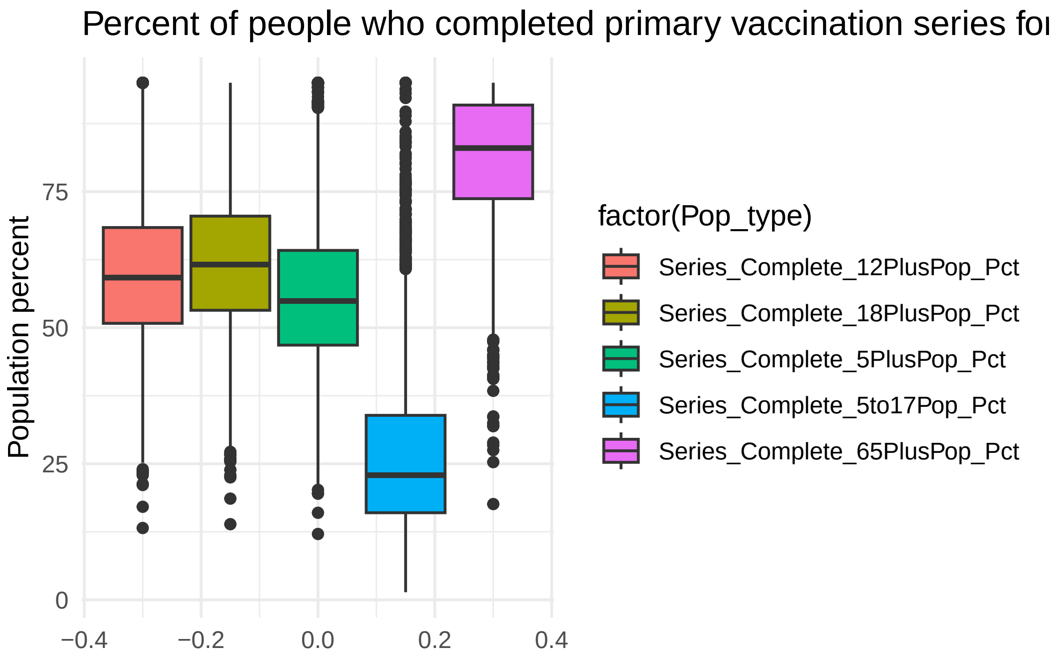
Data cleaning

- Clean data by removing NAs, converting character columns to integer, and by picking observation of most recent date (so we don't need to deal with temporal component in the model).

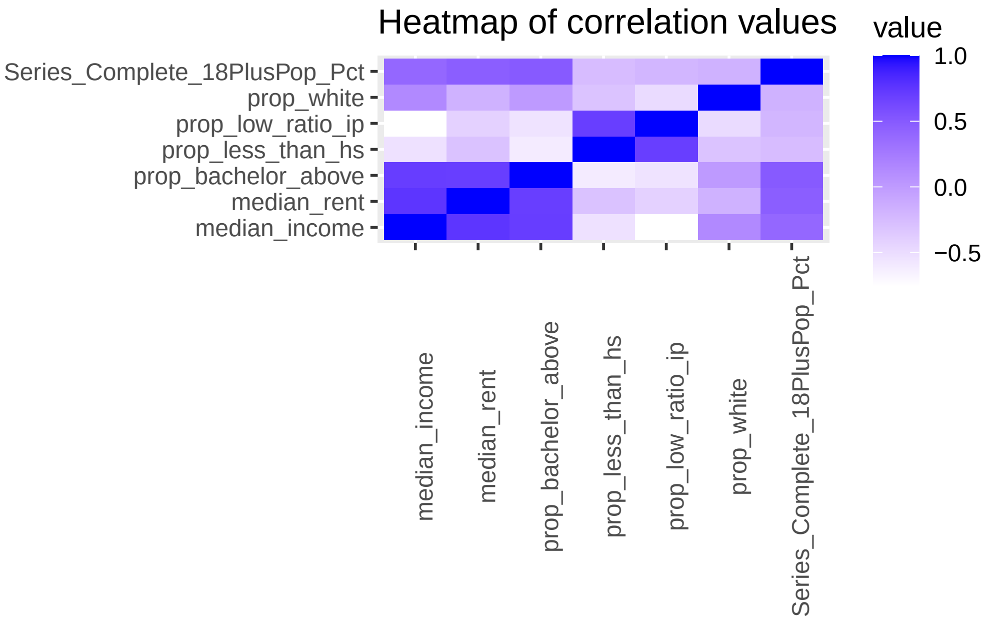
a) EDA

- The summary function helped me see the type of variables in the data, their central values. It is interesting to note that there are no significant outliers in the Series_Complete_18Plus population, as the max value is not too far from the 3rd quartile value of the variable.

- The box plot shows the percent of people of different age groups who have completed a primary vaccination series. It is interesting to note that the median is the highest for 65+ age category, whereas it is lowest for 5 to 17 age category. For 18+ population percentage there are various outlier values too, which implies that the association of age with vaccination rate is not the same across all the counties. Hence, median_age has been taken as a predictor variable in the model.



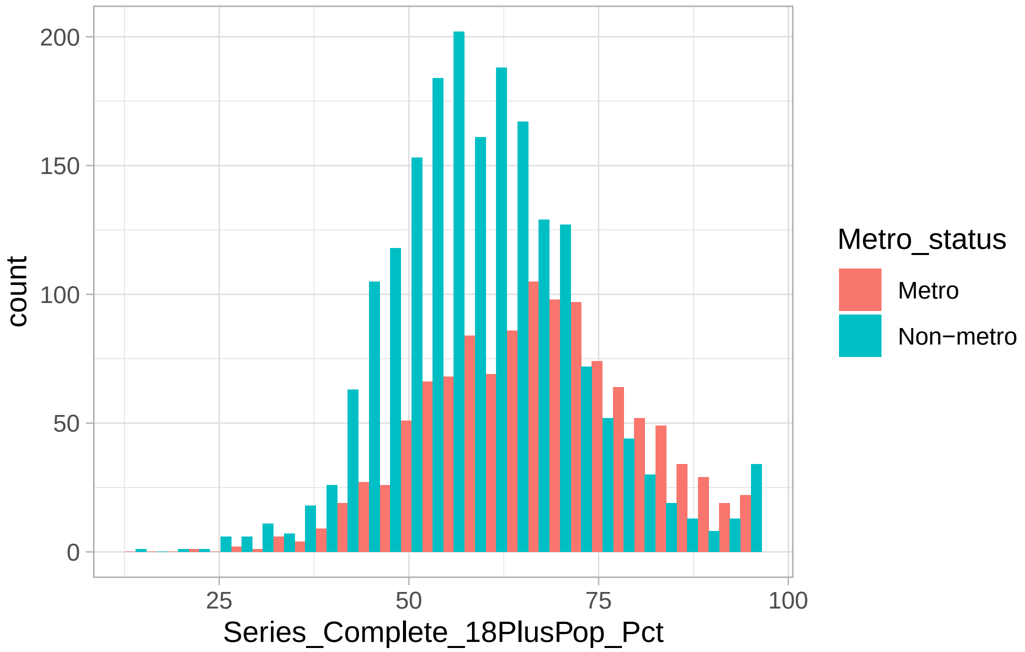
```r
ggplot(data = cordf, aes(x=var1, y=var2, fill=value)) +
  geom_tile() +
  #scale_fill_distiller(palette = "RdPu") +
  scale_fill_gradient(low="white", high="blue") +
  theme(axis.text.x = element_text(angle = 90))+
  labs(title = "Heatmap of correlation values") +
  xlab("")+ ylab("")
```



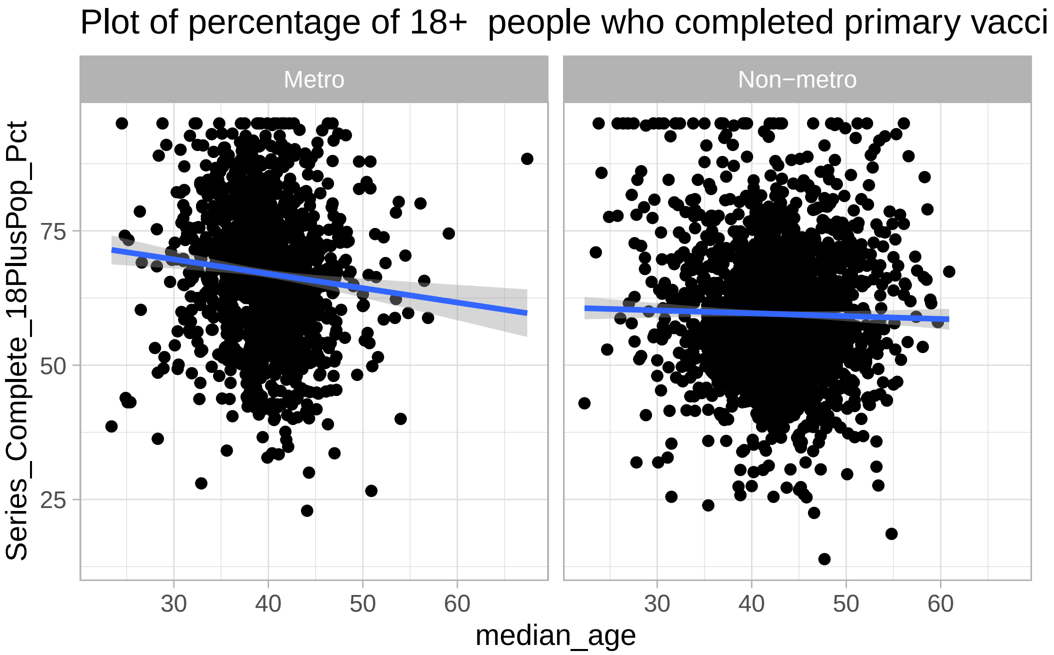
From the heat plot we can see some interesting and intuitive patterns like:

- prop_less_than_hs shows weakly negative correlation with the Series_Complete_18Plus variable. Intuitively, it means that lesser educated popluation are lesser likely to be vaccinated.
- Median income(median_income), median rent (median_rent) and proportion of people who completed bachelor or above education (prop_bachelor_above) are strongly positively correlated to each other and negatively correlated to prop_low_ratio_ip.
- Proportion of white (prop_white) people is weakly negatively correlated with the 18+ population who have completed vaccination series.

The histogram below shows that 'Series_Complete_18PlusPop_Pct' is normally distributed.



- The pattern for ‘median_age’ variable below looks different, especially the slopes differ substantially, for both Metro and non-metro counties, so interaction term would be added in the model. Intuitively, it means that people of the similar age would be influenced by whether they are located in non-metro or metro location, say due to the presence of more awareness or vaccination infrastructure in a metro county.



b) Model Building

distributional assumptions about the outcome measure - I have used a poisson model as each data point yi can equal 0, 1, 2,…, and the Poisson model is used for count data if no overdispersion would exist. I will use the (quasi)Poisson model, instead of the binomial model for count data because each data point yi does not have a natural limit and it is not based on a number of independent trials. - While the mean and variance of the 'Series_Complete_18Plus' variable are not equal, I have still used the poisson distribution instead of negative binomial distribution for simplicity and due to prior experience with poisson regression model.

consideration for covariates

- Based on the EDA, prop_low_ratio_ip and median_income are strongly correlated, therefore only 1 of them (median_income) has been taken into the model to avoid multicollinearity problems.
- Further, interaction terms were added in the model based on whether the slopes differed substantially when plotted with the variable of interest (as shown in the EDA section).
- prop_less_than_hs variable was included as people who are less educated are expected to have less awareness and belief in vaccinations, which also seems justified by its negative correlation with the Series_Complete_18Plus variable.

- Recip_State has been used as it would help control various fixed effects across counties within a particular state, such as vaccination centers, political inclinations of people of the state etc.
- Metro variable is taken into account as an interaction term as explained in EDA. It has not been taken into account separately as its effects would be incorporated through the state variable.

```
combined <- combined |> drop_na()
combined$Series_Complete_18Plus <- as.integer(combined$Series_Complete_18Plus)
```

```
mean(combined$Series_Complete_18Plus, na.rm = TRUE)
```

```
[1] 564.4685
```

```
var = (sd(combined$Series_Complete_18Plus, na.rm = TRUE))^2
print(var)
```

```
[1] 62781.05
```

- find out the overdispersion

```
n = dim(combined)[1] #119
k = length(modQ3_3$coefficients) # k=3
sum(rstandard(modQ3_3)^2)/(n-k) # overdispersion factor
```

```
[1] 50.70433
```

```
1- pchisq(sum(rstandard(modQ3_3)^2), n-k) #test value for overdispersion
```

```
[1] 0
```

- overdispersion exists so I switched to quasipoisson models. While Poisson models assumes equal variances and mean, running a quasi-poisson model is also better because it assumes variance is proportional to the mean.

Model training approach

- I have used a forward training approach with multiple iterations of adding/deleting variables. I have looked at whether the covariates are statistically significant or not, when considered into the model. I included (excluded) a variable in the model if it reduced (increased) the value of Residual deviance and AIC. To take into account the base population of each county and study the rate, instead of count, I have introduced an offset control "total_pop_18plus".

```
modQ3_13 <- glm(Series_Complete_18Plus ~  as.factor(Recip_State) + median_age*Metro_status
summary(modQ3_13)
```

```
Call:
glm(formula = Series_Complete_18Plus ~ as.factor(Recip_State) + 
    median_age * Metro_status + median_income + prop_white + 
    prop_less_than_hs, family = poisson, data = combined, offset = log(total_pop_18plus))

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-13.661   -3.034    0.000    2.857   14.974  

Coefficients:
                                   Estimate Std. Error z value Pr(>|z|)    
(Intercept)                      -1.007e+00  1.650e-01  -6.105 1.03e-09 ***
as.factor(Recip_State)CO         -1.653e-01  3.466e-02  -4.770 1.84e-06 ***
as.factor(Recip_State)GA         -5.300e-01  3.999e-02 -13.252  < 2e-16 ***
as.factor(Recip_State)ID         -5.146e-02  5.164e-02  -0.997 0.319006    
as.factor(Recip_State)KS         -3.360e-01  3.376e-02  -9.953  < 2e-16 ***
as.factor(Recip_State)KY         -2.387e-01  5.148e-02  -4.637 3.54e-06 ***
as.factor(Recip_State)MO         -3.603e-01  4.749e-02  -7.587 3.28e-14 ***
as.factor(Recip_State)MS         -3.642e-01  5.408e-02  -6.736 1.63e-11 ***
as.factor(Recip_State)MT         -5.018e-01  3.517e-02 -14.267  < 2e-16 ***
as.factor(Recip_State)ND         -6.247e-01  3.217e-02 -19.419  < 2e-16 ***
as.factor(Recip_State)NE         -5.774e-01  3.425e-02 -16.857  < 2e-16 ***
as.factor(Recip_State)NM          1.391e-01  4.148e-02   3.354 0.000795 ***
as.factor(Recip_State)NV         -5.250e-01  6.406e-02  -8.195 2.51e-16 ***
as.factor(Recip_State)OK         -3.628e-01  5.163e-02  -7.026 2.12e-12 ***
as.factor(Recip_State)OR         -3.089e-01  4.114e-02  -7.508 6.02e-14 ***
as.factor(Recip_State)SD         -3.362e-01  3.192e-02 -10.533  < 2e-16 ***
as.factor(Recip_State)TX         -3.918e-01  3.308e-02 -11.843  < 2e-16 ***
as.factor(Recip_State)UT         -5.562e-02  4.390e-02  -1.267 0.205182    
as.factor(Recip_State)WA         -4.251e-01  4.536e-02  -9.372  < 2e-16 ***
as.factor(Recip_State)WY         -5.182e-01  5.077e-02 -10.207  < 2e-16 ***
```

```
median_age                          1.421e-02  3.400e-03   4.181 2.90e-05 ***
Metro_statusNon-metro              -2.848e-02  1.605e-01  -0.177 0.859159
median_income                       8.197e-06  5.265e-07  15.569  < 2e-16 ***
prop_white                         -5.035e-01  5.590e-02  -9.009  < 2e-16 ***
prop_less_than_hs                  -6.175e-02  7.978e-02  -0.774 0.438909
median_age:Metro_statusNon-metro  2.251e-03  3.503e-03   0.643 0.520434
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 5492.4  on 110  degrees of freedom
Residual deviance: 2878.4  on  85  degrees of freedom
AIC: 3821.1

Number of Fisher Scoring iterations: 4
```

Interpretation of model

- The coefficients of most of the states are statistically significant, which implies that the state in which a county lies affects the population of 18+ people who complete vaccination. this is intuitive as the characteristics of state, like the infrastructure or state government's policies, would influence people's choice to get vaccinated.
- If the income in a county were to increase by one point, the difference in the log of expected 18+ people who completed vaccination series would change by 6.774e-06, while holding other variables in the model constant.
- If the proportion of white population in a county were to increase by one point, the difference in the log of expected 18+ people who completed vaccination series would change by -4.309e-01, while holding other variables in the model constant.
- The median age of the population also influence the people who complete vaccination series, and the effect interacts with whether the county is metro or non-metro. The coefficient of the interaction term (7.972e-03) represents the difference in slopes for the median_age, comparing with metro nature of the county.

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

```
numeric(0)
```

```
$fit
numeric(0)

$se.fit
numeric(0)

$residual.scale
[1] 1

numeric(0)
```

- the model predicts 211465.9 people of 18+ to be vaccinated, whereas the observed data shows 267230 people of 18+ to be vaccinated. The SE of the model is 10770.15. I believe that the model made a bad prediction as the actual value does not lie within 2 standard deviations, neither within 3 sd, of the estimate.

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

To sum up, I find various demographic variables influenced the vaccination rates. For instance, people with higher income were more likely to be vaccinated than the ones with lower income. Also, the higher the median age in the county more likely it would be vaccinated. People who were educated less than high school were had a negative association with the 18+ vaccination population. More the proportion of whites in a county, the lesser likely would be the rate of vaccination.

Further, the model can be trained to give better estimates. To strengthen the mnodel for future analysis, I would like to gather and understand the effect of non-demographic information such as–

- details about vaccination centers (the density or total number) in a county/ state
- political affiliation of people in a given state.
- the price of vaccines in different states (This may vary as the tax rate, government subsidies may vary for different states)

e) Now consider the situation of analysing vaccination rates at the state level.

- For regression 1), the granularity of the outcome variable and covariates would be lower than our model in part b. However, this may not necessarily be problematic, as such level of granularity might be appropriate based on the research questions being studied.

- For regression 2), the outcome variable won't be a count variable, as it won't vary in discrete fixed intervals. Therefore, to make it a count variable to use poisson distribution, we would need to transform it, say by rounding it-off, due to which we would lose some information. Thereby, the granularity would be lower of the outcome measure. Further, we would loose significant information in aggregating information from county to state level in other covariates.
- For regression 3- it would provide much better granularity of information, as compared to cases 1 and 2. Incorporating the fixed effect helps us account for likely time-invariant characteristics of the state that would vaccination rate; For instance, the categorical variable of state helps us to account when there are more vaccination centers in certain states as opposed to other states, which we doesn't get reflected in our census data. At the same time, when we study research questions that are about understanding different counties within a particular state we may choose to ignore the fixed-effect of state based on the assumption that the state in consideration has uniformly distributed characteristics (like vaccination centers) across the counties within it. So, ultimately the choice of variables depend on the question being asked.