Assignment1

Haoluan Chen

Q1

a)

Assume $E[\theta] = 1$ and $Var(\theta) = \sigma^2$

By the Law of total expectation and $Y|\theta \sim$ Poisson($\mu\theta$):

$E(Y) = E[E(Y|\theta)] = E[\mu\theta] = \mu E[\theta] = \mu$

By the Law of total variance and $Y|\theta \sim$ Poisson($\mu\theta$):

$Var(Y|\theta) = E[Var(Y|\theta)] + Var[E(Y|\theta)] = E[\mu\theta] + Var(\mu\theta) = \mu + \mu^2\sigma^2 = \mu(1 + \mu\sigma^2)$

b)

Assume $\theta \sim Gamma(\alpha, \beta)0$

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

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

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

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

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

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

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

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

c)

Since

$E(Y) = \mu = \frac{\alpha(1-\frac{\mu\beta}{\mu\beta+1})}{\frac{\mu\beta}{\mu\beta+1}} = \alpha\mu\beta => \alpha\beta = 1$

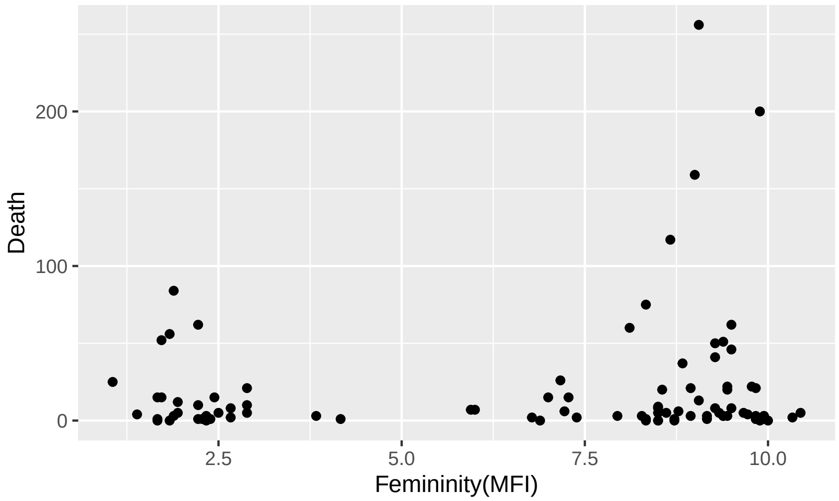
$Var(Y) = \mu(1+\mu\sigma^2) = \frac{\alpha(1-\frac{\mu\beta}{\mu\beta+1})}{(\frac{\mu\beta}{\mu\beta+1})^2} = \alpha\mu\beta + \alpha\mu^2\beta^2 => \alpha\beta^2 = \sigma^2$

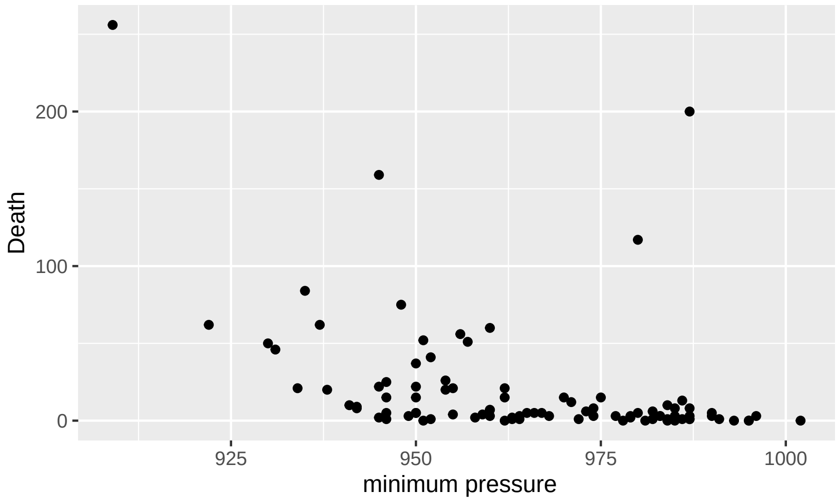
Then

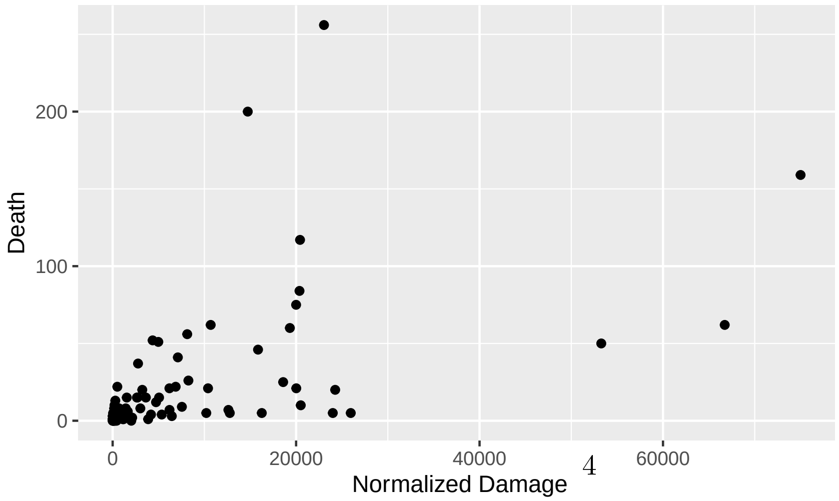
$\alpha = 1/\sigma^2, \beta = \sigma^2$

Q2

a)


Death
200
100
0
2.5
5.0
7.5
10.0
Femininity(MFI)


Death
200
100
0
925
950
975
1000
minimum pressure


Death
200
100
0
0
20000
40000
60000
Normalized Damage

From the death by femininity scatter plot, it looks like there is two cluster. A group centered around femininity value of 2 and a group centered around femininity value of 8.5. Higher femininity value has higher variability on the number of deaths. One extreme value of over 200 deaths. For minimum pressure, there is a slightly increasing trend as the minimum pressure goes below 950. Lastly, we see an increasing in deaths as normalized damage increase, the variation also increase.

b)

Fitting Poisson model(Estimates are exponentiated) :

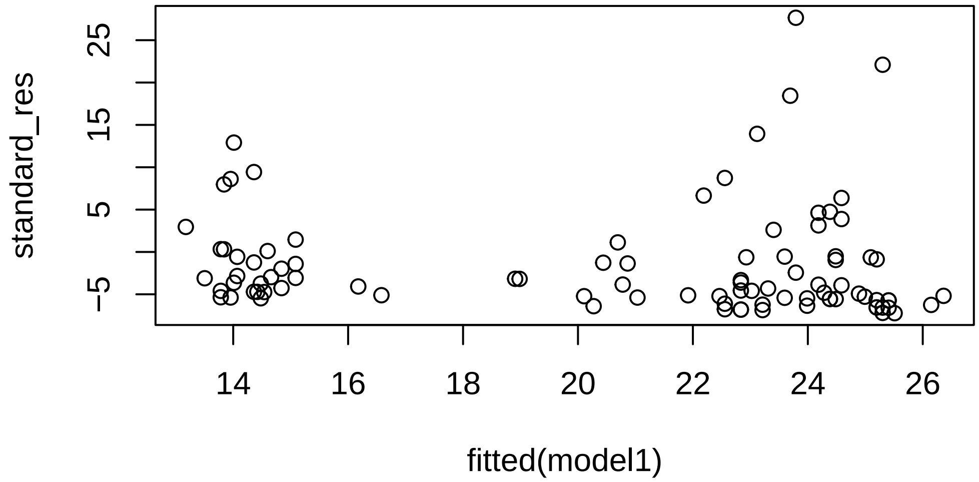
```
model1 <- glm(alldeaths~MasFem, family = poisson, data = q2)
est <- data.frame(summary(model1)$coefficients) %>%
  mutate(Estimate = exp(Estimate))
kable(round(est, 4))
```

	Estimate	Std..Error	z.value	Pr...z..
(Intercept)	12.1870	0.0633	39.5021	0
MasFem	1.0767	0.0079	9.3620	0

The poisson model suggested that as the MFI increase by one unit, the death count increase by a factor of 1.0767

Checking for overdispersion:

```
standard_res <- rstandard(model1)
plot(fitted(model1), standard_res)
```



```
n = 92
k = 2
sum(standard_res^2)/(n-k)
```

```
[1] 44.6563
```

```
1-pchisq(sum(standard_res^2), n-k)
```

```
[1] 0
```

There is an overdispersion!

Fitting quasi-poisson model(Estimates are exponentiated):

	Estimate	Std..Error	t.value	Pr...t..
(Intercept)	12.1870	0.5437	4.5987	0.0000
MasFem	1.0767	0.0678	1.0899	0.2787

Assuming the significant level to be 0.05. The quasi-poisson suggest that the MFI does not affect on the death count.

c)

Model 4(Estimates are exponentiated):

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	2.4756	0.1222	20.2605	0.0000
ZMasFem	0.1723	0.1238	1.3918	0.1640
ZMinPressure_A	-0.5521	0.1503	-3.6734	0.0002
ZNDAM	0.8635	0.1445	5.9764	0.0000
ZMasFem:ZMinPressure_A	0.3948	0.1521	2.5952	0.0095
ZMasFem:ZNDAM	0.7051	0.1501	4.6988	0.0000

```
exp(0.1723)
```

```
[1] 1.188034
```

Assuming a hurricane with median pressure and damage ratings, the estimated effect of one unit increase in MFI on death count is 18.8%

d)

```
d <- q2%>% filter(Name == "Sandy")
sandy <- d[12:14]
predict(cmodel, sandy, type="response")
```

```
       1
20806.74
```

The predicted death count for Sandy is 20807. However, the actual death count is only 159. The predicted death count is so high because Sandy has highest damage.

e)

weakness:

1. Only 9 independent coder were included in determine the MFI, which may be biased. More coder can be included.

2. P-value of the models were not include in the table.

strength:

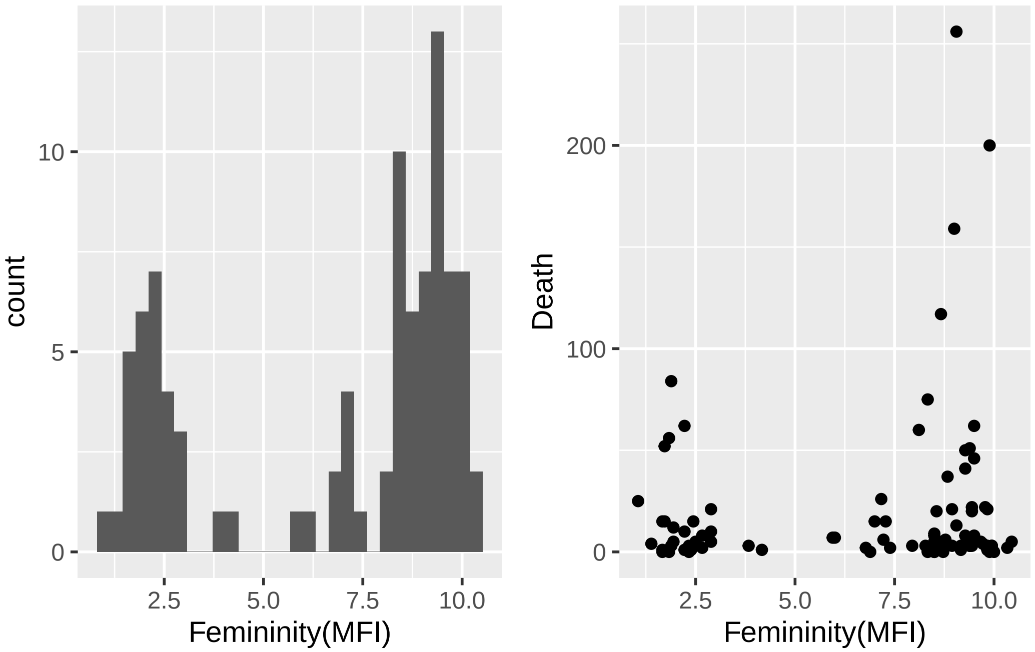
1. Recognizing the confounding variable: effect of gendered names on protective action, not simply conclude that Feminine-named hurricanes cause significantly more deaths.
2. Many experiment were carried out to test difference aspect about the perceived risk of the hurricanes, predicted intensity and evacuation intention. This wide range of experiment helps convince reader that gendered hurricanes names will affect how people feel and act.
3. Data set are available for reproducible

f)

I think I'm convinced by the experiments result that people perceive male named hurricanes are slightly more risk/intensity and will more likely to follow evacuation plan. However, I'm not convinced that the name of the hurricanes have impact on the total death. Firstly, both the quasi-poisson and negative binomial model suggested that there is not effect of the MFI on the total death. Furthermore, there are about 2/3 of the hurricanes with feminine names and 1/3 of the hurricanes with masculine names, and it looks like there are four extreme values for hurricanes with more feminine names. It may be due to chance that these sever hurricane got a feminine name. Lastly, I noticed that there are duplicates in the names, and most of the time, the later hurricane caused more damage and deaths(Bob, Bonnie, Charley, Danny, Floyd, Irene), maybe calling two hurricanes same is not a good idea? People may let their guard down.

```r
f1 <- q2 %>% ggplot(aes(MasFem))+ geom_histogram() + xlab("Femininity(MFI)")
f2 <- q2 %>% ggplot(aes(x=MasFem, y=alldeaths)) + geom_point() +
  xlab("Femininity(MFI)") + ylab("Death")
grid.arrange(f1, f2, ncol=2)
```

```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Q3

Loading and combining two datasets

```
Rows: 3283 Columns: 80
-- Column specification --------------------------------------------------------
Delimiter: ","
chr  (6): Date, FIPS, Recip_County, Recip_State, SVI_CTGY, Metro_status
dbl (47): MMWR_week, Completeness_pct, Administered_Dose1_Pop_Pct, Administe...
num (27): Administered_Dose1_Recip, Administered_Dose1_Recip_5Plus, Administ...

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
Rows: 37704 Columns: 4
-- Column specification --------------------------------------------------------
Delimiter: ","
chr (3): fips, county_name, variable
dbl (1): value

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

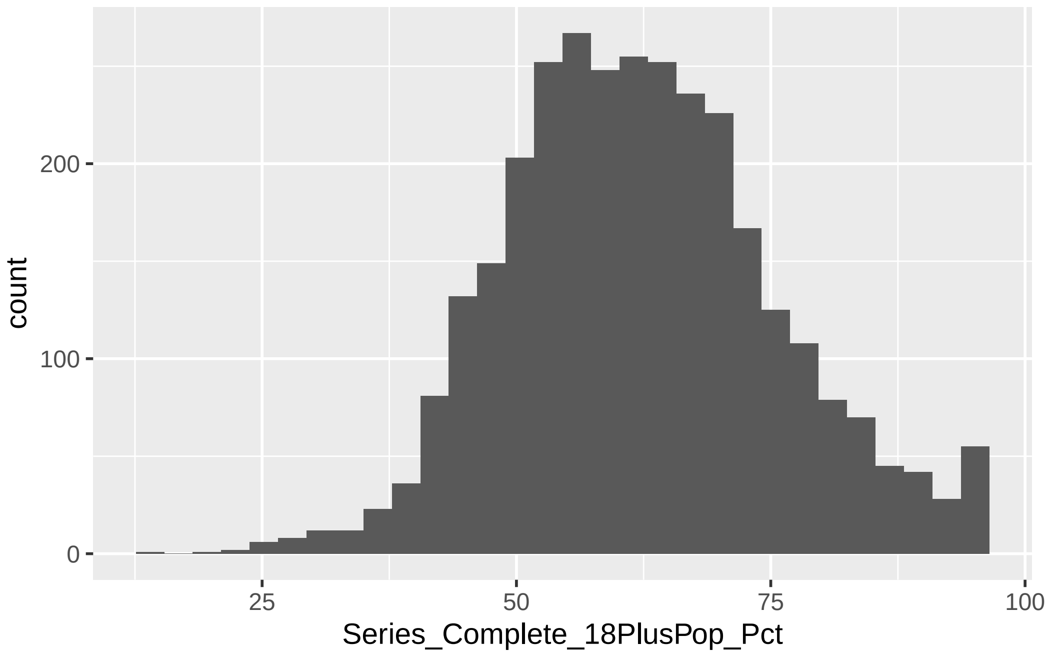
a)

There are very few data points that are missing. Most of the variable has complete rate of 99%. I removed the missing assuming they are missing completely at random.

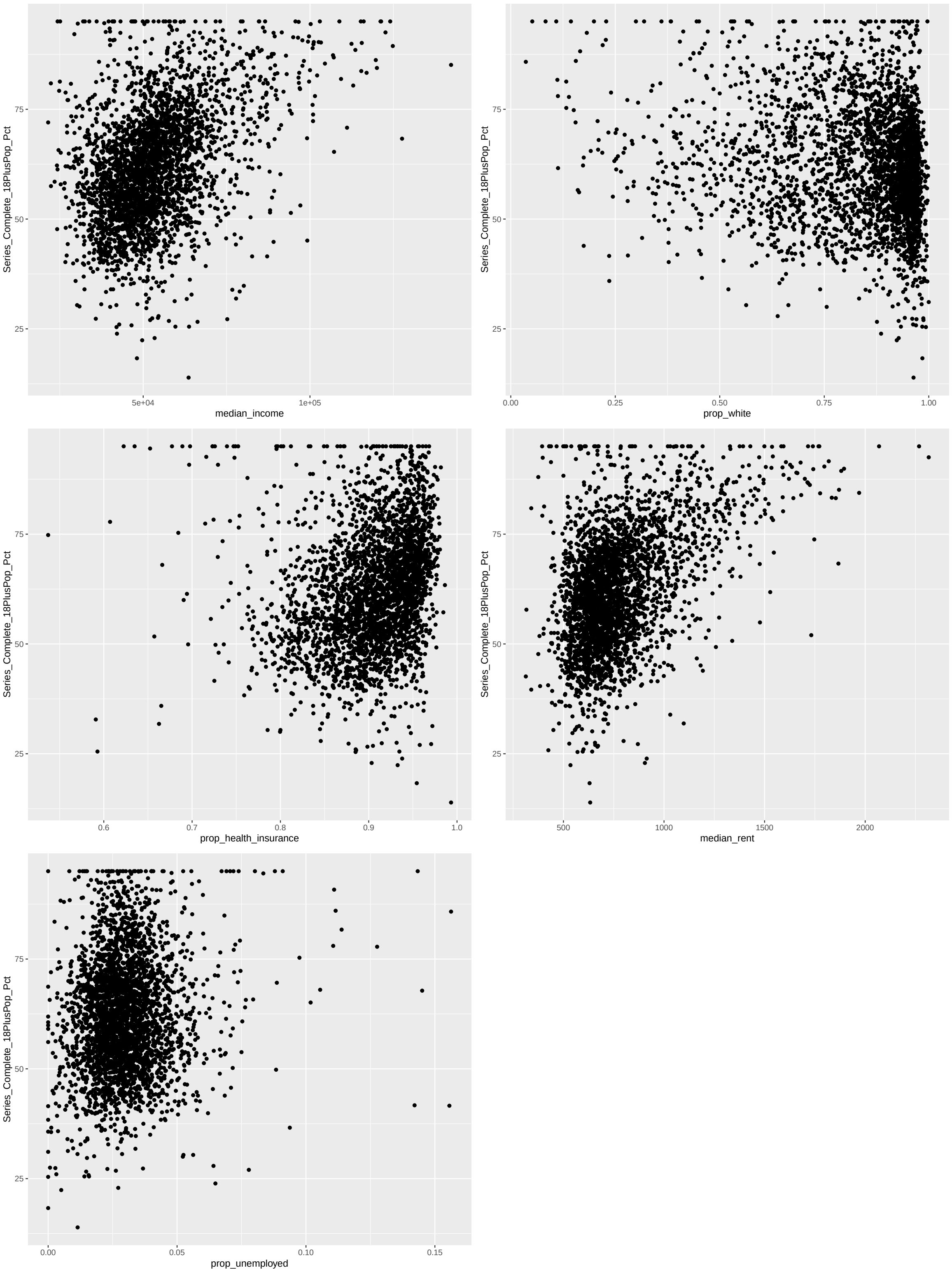
```
combined <- na.omit(combined)
```

Check distribution of the count

```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

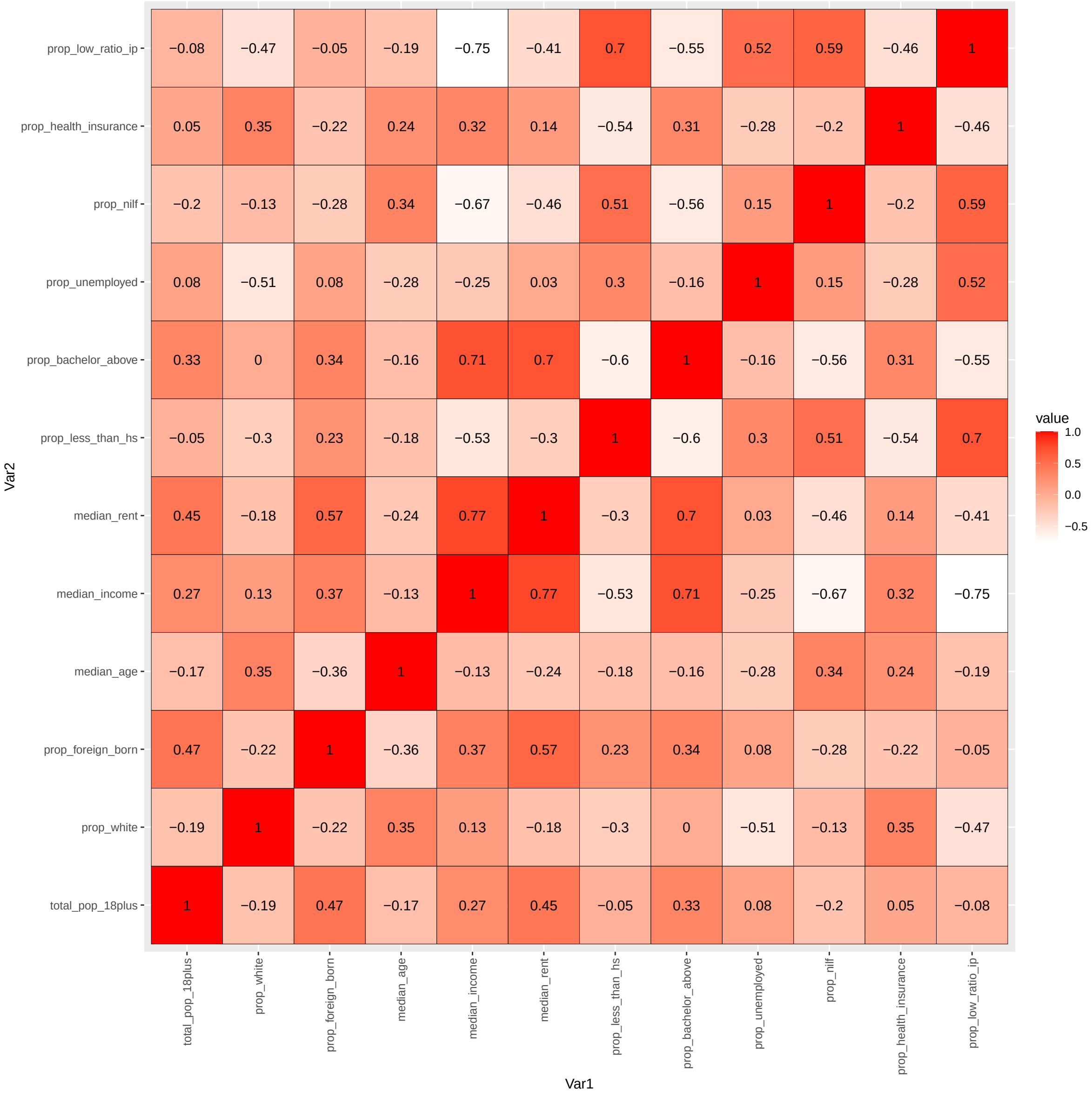


The distribution looks pretty normal, but with a small mode at 95%.


Series_Complete_18PlusPop_Pct
75
50
25
5e+04
1e+05
median_income
Series_Complete_18PlusPop_Pct
75
50
25
0.00
0.25
0.50
0.75
1.00
prop_white
Series_Complete_18PlusPop_Pct
75
50
25
0.6
0.7
0.8
0.9
1.0
prop_health_insurance
Series_Complete_18PlusPop_Pct
75
50
25
500
1000
1500
2000
median_rent
Series_Complete_18PlusPop_Pct
75
50
25
0.00
0.05
0.10
0.15
prop_unemployed

There is some increasing trend for some of the variables, such as median income and median rent. And maybe a slight decreasing in vaccination rate as the population white increase. There is no obvious trend for health insurance or unemployed.

Correlation heat map for acs data.



b)

I chose to use binomial model, since the proportion is a probability between 0 and 1. And I assume the outcome follows a binomial distribution, one person is completed the vaccine is consider as a success, each county is a binomial sample.

I chose covariates that I am interested in(proportion of white people, proportion of health insurance, proportion of foreign born) and the covariates that has a decreasing/increasing impact on vaccine completion rate based on my EDA(median income). Also, I avoided including multiple covariates that are highly correlated(median income, median rent, bachelor above).

In term of choosing the candidate model, I fitted a large model and a small model. The larger model includes the covariates that I'm interested in. And the covariates in small model are selected by backward elimination, which includes covariates that are significant in the large model.

```
modeldata3b <- combined %>%
  mutate(Series_Complete_18PlusPop_Pct_model = Series_Complete_18PlusPop_Pct/100)
model3b1 <- glm(Series_Complete_18PlusPop_Pct_model ~
                  prop_white + prop_foreign_born+
                 median_income +prop_unemployed + prop_nilf +
                 prop_health_insurance + prop_low_ratio_ip,
               family = binomial, data = modeldata3b)
```

```
Warning in eval(family$initialize): non-integer #successes in a binomial glm!
```

```
kable(round(summary(model3b1)$coefficients, 4))
```

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-2.6367	0.9802	-2.6901	0.0071
prop_white	-0.6614	0.2908	-2.2740	0.0230
prop_foreign_born	2.3742	0.7013	3.3856	0.0007
median_income	0.0000	0.0000	2.0797	0.0376
prop_unemployed	5.3454	3.4930	1.5303	0.1259
prop_nilf	-0.4858	0.6571	-0.7393	0.4597
prop_health_insurance	3.2692	0.8815	3.7086	0.0002
prop_low_ratio_ip	0.2877	1.1202	0.2569	0.7973

```
model3b2 <- glm(Series_Complete_18PlusPop_Pct_model ~
                  prop_white + prop_foreign_born+
                 median_income +prop_unemployed +
                 prop_health_insurance,
               family = binomial, data = modeldata3b)
```

```
Warning in eval(family$initialize): non-integer #successes in a binomial glm!
```

```
summary(model3b2)
```

```

Call:
glm(formula = Series_Complete_18PlusPop_Pct_model ~ prop_white + 
    prop_foreign_born + median_income + prop_unemployed + prop_health_insurance, 
    family = binomial, data = modeldata3b)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-1.0532  -0.1355  -0.0020   0.1342   1.1398  

Coefficients:
                        Estimate Std. Error z value Pr(>|z|)    
(Intercept)           -2.818e+00  7.610e-01  -3.704 0.000212 ***
prop_white            -6.705e-01  2.745e-01  -2.442 0.014590 *  
prop_foreign_born      2.439e+00  6.868e-01   3.551 0.000384 ***
median_income          1.239e-05  3.447e-06   3.595 0.000324 ***
prop_unemployed        5.794e+00  3.269e+00   1.772 0.076340 .  
prop_health_insurance  3.231e+00  8.718e-01   3.706 0.000211 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 247.53  on 3120  degrees of freedom
Residual deviance: 172.59  on 3115  degrees of freedom
AIC: 3329.5

Number of Fisher Scoring iterations: 4
```

```
kable(round(summary(model3b2)$coefficients, 4))
```

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-2.8184	0.7610	-3.7037	0.0002
prop_white	-0.6705	0.2745	-2.4424	0.0146
prop_foreign_born	2.4389	0.6868	3.5510	0.0004
median_income	0.0000	0.0000	3.5950	0.0003
prop_unemployed	5.7943	3.2693	1.7723	0.0763
prop_health_insurance	3.2309	0.8718	3.7059	0.0002

```
est <- data.frame(summary(model3b2)$coefficients) %>%
  mutate(Estimate = exp(Estimate))
kable(round(est, 4))
```

	Estimate	Std..Error	z.value	Pr...z..
(Intercept)	0.0597	0.7610	-3.7037	0.0002
prop_white	0.5114	0.2745	-2.4424	0.0146
prop_foreign_born	11.4609	0.6868	3.5510	0.0004
median_income	1.0000	0.0000	3.5950	0.0003
prop_unemployed	328.4166	3.2693	1.7723	0.0763
prop_health_insurance	25.3019	0.8718	3.7059	0.0002

Assuming 0.05 significance level. We see that everything in the small model are significant expect for prop_unemployed. Among the significant covariates, the median income has the odds of one, this means that the median income does not have an impact on the vaccination rate. The proportion of foreign born and health insurance have positive impact on the vaccine rate, with odd ratio of 11.46, 328.31 and 25.30 compared to our baseline respectively. Additionally, the proportion of white people does have an negative effect on the vaccination rate, about 50% reduction in odds ratio compared to the basedline.

c)

```
Ada <- modeldata3b %>% filter(county_name == "Ada County, Idaho")
Ada <- dplyr::select(Ada, prop_white, prop_foreign_born,
                 median_income,prop_unemployed, prop_nilf,
                 prop_health_insurance, prop_low_ratio_ip, total_pop_18plus)
```

```
dplyr::select(modeldata3b, county_name, Series_Complete_18PlusPop_Pct) %>%
  filter(county_name == "Ada County, Idaho")
```

```
# A tibble: 1 x 2
  county_name       Series_Complete_18PlusPop_Pct
  <chr>                                     <dbl>
1 Ada County, Idaho                          76.9
```

```
predict(model3b2, Ada, type = "response")
```

```
        1 
0.6635257 
```

The prediction is about 10% off. I guess it is pretty good considering the variability in the data

d)

In summary, our model suggest that the proportion of foreign born and health insurance have positive impact on the vaccine rate. Income related estimates does not seen to have an impact. Interestingly, the proportion of white people have an negative effect on the vaccination rate. However, our model is only based on current available data, there may be some confounding variable that are not included in our analysis. For example, maybe the county with higher proportion of white people are mostly elders, which may be more reluctant or worried about the side affect of the vaccine. Therefore, age may be of interest to investigate in future.

e)

For first option, I think it has the second highest granularity of information, since it is combining the count of 18+ fully vaccinated in county level into state level. The Second option has least granularity of information since it is averaging the county results. And the last option is not only using the information in county level(mode detailed level), but also includes county as a covariate.

In my opinion, the the third option is not appropriate in most of the case because it contains way too many covariates, we will not be able to extract useful information from it. First and second option depends on weather or not you are interested in the total count or average of the county level. Note that the average will suffer from problem of outlier