Applied Statistics II HW1 STA2201S Winter 2023

1 Overdispersion

a) The conditional distribution $Y|\theta \sim \text{Pois}(\mu\theta)$ results in $E(Y|\theta) = \mu\theta$, $\text{Var}(Y|\theta) = \mu\theta$. Using the laws of total expectation and variance,

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

b) When $\theta \sim \text{Gamma}(\alpha, \beta)$, the marginal distribution of Y is, according to the definition,

$$\begin{aligned}
p(Y) &= \int p(Y|\theta)p(\theta)d\theta \\
&= \int \text{Pois}(\mu\theta)\text{Gamma}(\theta; \alpha, \beta)d\theta \\
&= \int \frac{e^{-\mu\theta}(\mu\theta)^k}{k!} \cdot \frac{\theta^{\alpha-1}e^{-\beta\theta}\beta^\alpha}{\Gamma(\alpha)}d\theta \\
&= \frac{\mu^k\beta^\alpha}{\Gamma(\alpha)k!} \int \theta^{\alpha+k-1}e^{-(\mu+\beta)\theta}d\theta \qquad (3)
\end{aligned}$$

Let $t = (\mu+\beta)\theta$, $z = \alpha+k$, so $d\theta = dt/(\mu+\beta)$. Use the definition of the Gamma function, $\Gamma(z) = \int t^{z-1}e^{-t}dt$, we can simplify $p(Y)$.

$$\begin{aligned}
p(Y) &= \frac{\mu^k\beta^\alpha}{\Gamma(\alpha)k!} \int \left(\frac{t}{\mu+\beta}\right)^{z-1} e^{-t}\frac{dt}{\mu+\beta} \\
&= \frac{1}{\Gamma(\alpha)k!} \cdot \frac{\mu^k\beta^\alpha}{(\mu+\beta)^{\alpha+k}} \int t^{z-1}e^{-t}dt \\
&= \frac{\Gamma(\alpha+k)}{\Gamma(\alpha)k!} \cdot \left(\frac{\beta}{\mu+\beta}\right)^\alpha \cdot \left(\frac{\mu}{\mu+\beta}\right)^k \qquad (4)
\end{aligned}$$

Let $p = \frac{\beta}{\mu+\beta}$, therefore $\frac{\mu}{\mu+\beta} = 1 - p$, then $p(Y)$ can be further simplified to,

$$\begin{aligned}
p(Y) &= \frac{\Gamma(\alpha+k)}{\Gamma(\alpha)k!}p^\alpha(1-p)^k \\
&\sim \text{NB}(\alpha, p) \qquad (5)
\end{aligned}$$

It is a negative binomial (NB) distribution with parameters α and p.

c) Let $E(Y) = \mu$ and $\text{Var}(Y) = \mu(1 + \mu\sigma^2)$, then using the properties of the NB distribution,

$$E(Y) = \frac{\alpha(1-p)}{p} = \mu \tag{6}$$

$$\text{Var}(Y) = \frac{\alpha(1-p)}{p^2} = \mu(1 + \mu\sigma^2) \tag{7}$$

$$\frac{E(Y)}{\text{Var}(Y)} = p = \frac{1}{1 + \mu\sigma^2} \tag{8}$$

From **b)** we know that,

$$p = \frac{\beta}{\mu + \beta} = \frac{1}{1 + \mu/\beta} \tag{9}$$

Therefore, comparing Eqs. (8) and (9), we find that $\beta = 1/\sigma^2$. Then from Eq. (6),

$$\alpha \cdot \frac{\mu/(\mu+\beta)}{\beta/(\mu+\beta)} = \frac{\mu}{\beta} = \mu$$

$$\alpha = \beta = \frac{1}{\sigma^2} \tag{10}$$

Therefore, the Gamma distribution that satisfies the required form of the expected value and variance is $\text{Gamma}(1/\sigma^2, 1/\sigma^2)$.

2 Hurricanes

Loading the data from the hurricane paper,

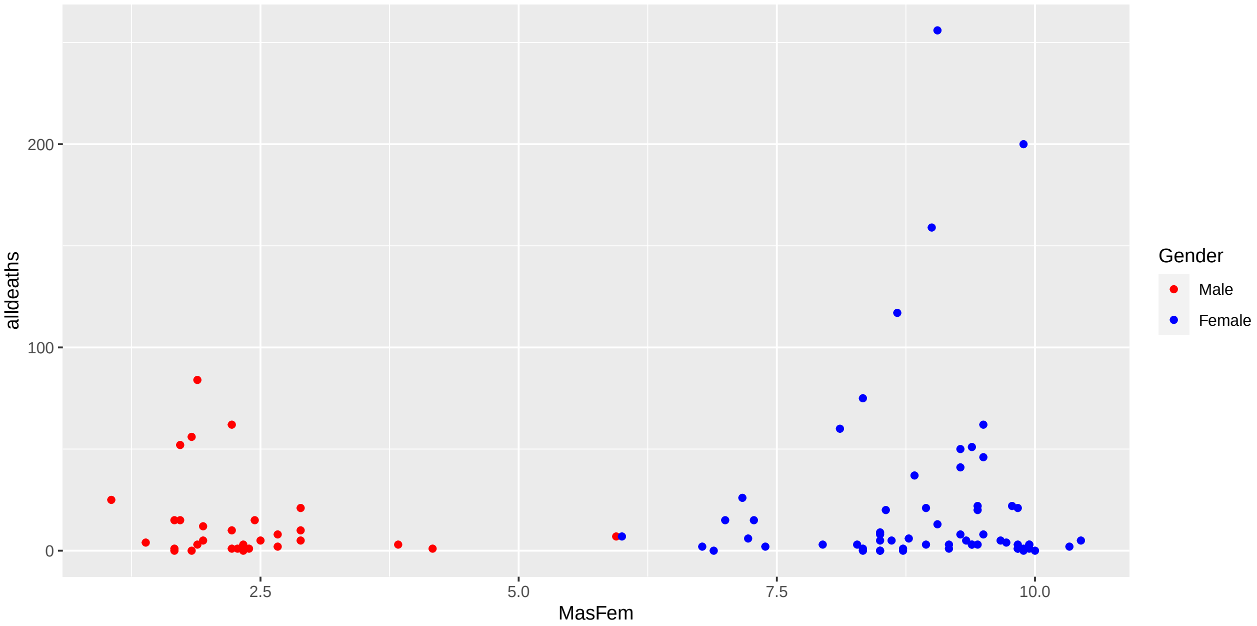
```
head(hurricanes, 3)
```

```
## # A tibble: 3 x 14
##   Year  Name  MasFem MinPressure~1 Minpr~2 Gende~3 Categ~4 allde~5  NDAM Elaps~6
##   <chr> <chr>  <dbl>         <dbl>   <dbl>   <dbl>   <dbl>   <dbl> <dbl>   <dbl>
## 1 1950  Easy    6.78           958     960       1       3       2  1590      63
## 2 1950  King    1.39           955     955       0       3       4  5350      63
## 3 1952  Able    3.83           985     985       0       1       3   150      61
## # ... with 4 more variables: Source <chr>, ZMasFem <dbl>, ZMinPressure_A <dbl>,
## #   ZNDAM <dbl>, and abbreviated variable names 1: MinPressure_before,
## #   2: `Minpressure_Updated 2014`, 3: Gender_MF, 4: Category, 5: alldeaths,
## #   6: `Elapsed Yrs`
## # i Use `colnames()` to see all variable names
```

a) Visualization

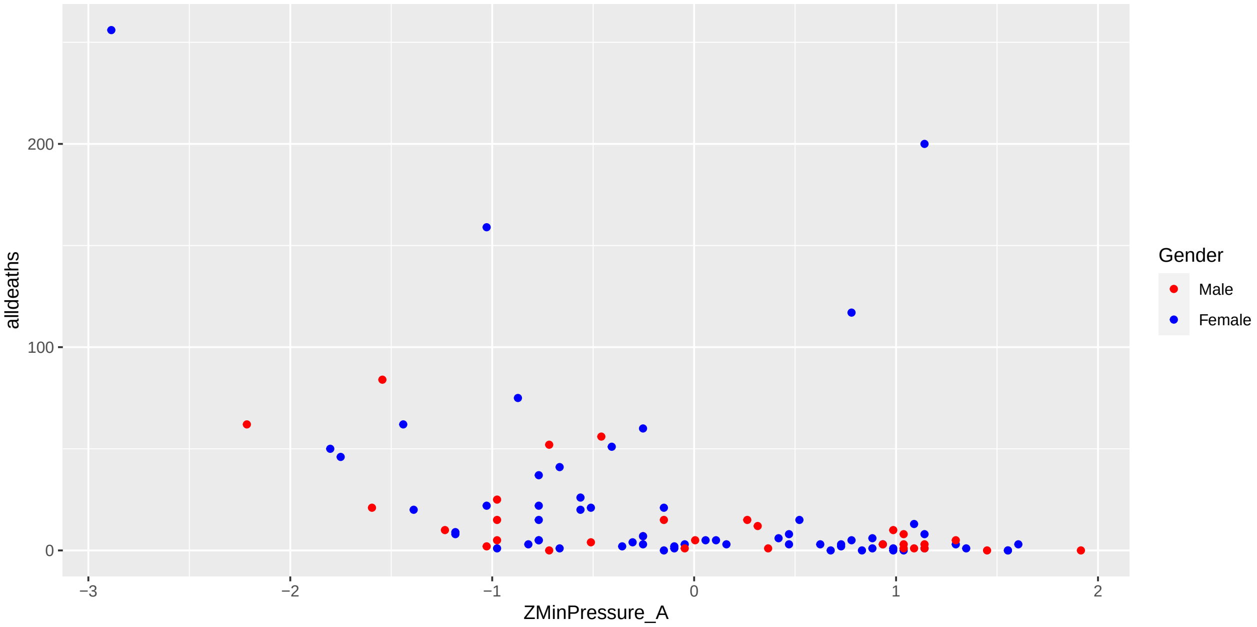
Masculinity-femininity index (MFI) vs deaths

```
# MFI vs hurricane-caused deaths
hurricanes %>%
    ggplot(aes(x = MasFem, y = alldeaths, color = as.factor(Gender_MF))) + geom_point() +
    labs(color = "Gender") + scale_color_manual(labels = c("Male", "Female"), values = c("red",
    "blue"))
```



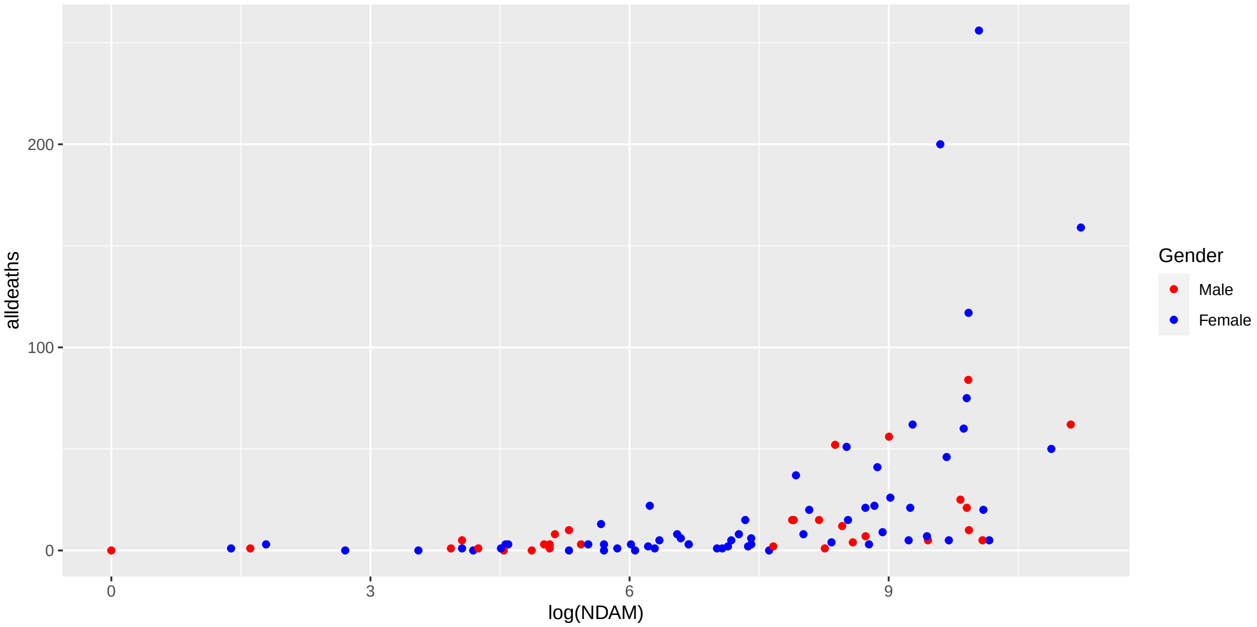
Minimum pressure of the hurricane vs deaths

```
# Minimum pressure of the hurricane vs hurricane-caused deaths
hurricanes %>%
    ggplot(aes(x = ZMinPressure_A, y = alldeaths, color = as.factor(Gender_MF))) +
    geom_point() + labs(color = "Gender") + scale_color_manual(labels = c("Male",
    "Female"), values = c("red", "blue"))
```



Normalized damage (on log scale) vs deaths

```
# Normalized damage vs hurricane-caused deaths
hurricanes %>%
    ggplot(aes(x = log(NDAM), y = alldeaths, color = as.factor(Gender_MF))) + geom_point() +
    labs(color = "Gender") + scale_color_manual(labels = c("Male", "Female"), values = c("red",
    "blue"))
```



The log scale is used on the third graph to reduce the effects of the extreme outliers. All three graphs show similar distributions between hurricanes with particularly masculine and feminine names (as used in the color). The hurricanes with more feminine names have a few more extreme outliers (about 4) than those with more masculine names.

b) Poisson regression and overdispersion check

```
poisson_fit <- glm(alldeaths ~ MasFem, data = hurricanes, family = "poisson")
summary(poisson_fit)
```

```
##
## Call:
## glm(formula = alldeaths ~ MasFem, family = "poisson", data = hurricanes)
##
## 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
```

Assessment of overdispersion by checking for the mean and the variance.

```
h_mean <- mean(hurricanes$alldeaths, na.rm = TRUE)
h_var <- var(hurricanes$alldeaths, na.rm = TRUE)
sprintf("Mean is %f, variance is %f", h_mean, h_var)
```

```
## [1] "Mean is 20.652174, variance is 1673.152413"
```

Variance > mean, there is overdispersion. Fit the GLM using quasi-poisson distribution.

```
quasipoisson_fit <- glm(alldeaths ~ MasFem, data = hurricanes, family = "quasipoisson")
summary(quasipoisson_fit)
```

```
##
## Call:
## glm(formula = alldeaths ~ MasFem, family = "quasipoisson", data = hurricanes)
##
## 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 4 from the paper

Model 4 fits the data to a GLM with negative binomial distribution using the following explanatory variables: minimum pressure (ZMinPressure_A), normalized damage (ZNDAM), MFI (ZMasFem), MFI × minimum pressure (ZMasFem × ZMinPressure_A), MFI × normalized damage (ZMasFem × ZNDAM).

```
negbin_fit <- MASS::glm.nb(alldeaths ~ ZMinPressure_A + ZNDAM + ZMasFem + ZMasFem:ZMinPressure_A +
    ZMasFem:ZNDAM, data = hurricanes)
summary(negbin_fit)
```

```
##
## Call:
## MASS::glm.nb(formula = alldeaths ~ ZMinPressure_A + ZNDAM + ZMasFem +
##     ZMasFem:ZMinPressure_A + ZMasFem:ZNDAM, data = hurricanes,
##     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
```

```
coeffs <- negbin_fit$coefficients
median_pressure <- median(hurricanes$ZMinPressure_A, na.rm = TRUE)
median_ndam <- median(hurricanes$ZNDAM, na.rm = TRUE)
fem_affect = coeffs["ZMasFem"] + coeffs["ZMinPressure_A:ZMasFem"] * median_pressure +
    coeffs["ZNDAM:ZMasFem"] * median_ndam
fem_affect
```

```
##    ZMasFem
## -0.1626157
```

Assuming a hurricane with median pressure and damage ratings, a decrease by one of the MFI leads to a decrease of deaths on log scale of about 0.163.

d) Death prediction for Hurricane Sandy

Use Model 4 (`negbin_fit`) to predict the number of deaths from Hurricane Sandy.

```
sandy = hurricanes %>%
    filter(Name == "Sandy")
predict.glm(negbin_fit, newdata = sandy %>%
    dplyr::select(ZMinPressure_A, ZNDAM, ZMasFem), type = "response")
```

```
##        1
## 20806.74
```

The predicted deaths from Hurricane Sandy is huge, but the actual reported death from Hurricane Sandy is

```
sandy$alldeaths
```

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

The fitted model (without Sandy) significantly overestimate the deaths, since Hurricane Sandy is an outlier (although not as extreme as Hurricane Katrina or Aubrey) and therefore isn't fit well by the model.

e) Appraisal of the paper

Strengths:

(1) One highlight of the paper is the empirically constructed MFI for (hurricane) names by using survey of volunteer opinions. Since some names may have an unclear gender assignment (e.g. Ashley, Leslie, Sam, etc), pooling opinions from a number of people increases the limitation of binary gender assignment for names.

(2) The authors conducted a number of variable selection methods, including choosing which higher-order features (i.e. interaction terms) to keep, before finalizing the structure of the model.

Weaknesses:

(1) The types of explanatory variables for the model is limited. Regarding the information about the hurricane damage on land, the main missing factors are the geographical region that the hurricane hits, including the size, population density, maximum wind speed, rainfall dat, etc. These information is absent from the features and they can naturally be included in the GLM.

(2) The authors implicitly chose the hurricanes that landed on the continent. Some hurricanes not landed are also given anthropomorphic names, since the naming is decided much earlier than landfall. Moreover, hurricane names are often reused and only those that caused significant damages are not reused. These aspect likely complicates the data and modelling process, yet neither of which was discussed by the authors.

(3) In terms of the distributional assumption, the authors may consider inclusion of extreme statistical models that better account for the tail of the distribution.

f) Opinions about the results

Although this paper does raise an interesting point in considering the implicit bias in gender-based naming and in media reporting, I'm not convinced that the results fully justify the main argument from the paper as is also the title. First of all, the argument seems to have inherent flaws since the death counting method (indirect and direct) suggests a more nuanced link between the casualties and the impact of the hurricane, which is not yet fully explored in the paper. What the authors can also do is to use a hypothesis testing framework to examine the difference in distributions between deaths and damages from hurricanes of differently gendered names. As has been discussed at the EDA stage earlier, this difference seems minute, if at all present beyond the extreme outliers, according to visual inspection.

The authors merged the data before and after 1979 to improve the statistical power, but the hurricane naming system was altered in these two time periods, which may not justify this treatment. For small datasets such as analyzed in this paper, improving the statistical power can also be achieved by the collection of more data, which in this context refers to the varieties of quantifiers of hurricane-caused destruction. Deaths and dollar-valued damages are the only two quantifiers used here. These variables are in their aggregated form, although their breakdown seems to be available, and that they are inherently heterogeneous (e.g. indirect and direct deaths or damage). The simplest way to improve the argument of the paper is therefore to use the constituent values of these variables to fit a more elaborate model. In addition, taking into consideration of other (currently ignored) covariates, as mentioned earlier in e), can also increase the statistical power and model robustness.

## 3 Vaccinations

Loading the covid-19 vaccination data

```
head(vax, 3)
```

```
## # A tibble: 3 x 80
##   Date     FIPS  MMWR_~1 Recip~2 Recip~3 Compl~4 Admin~5 Admin~6 Admin~7 Admin~8
##   <chr>    <chr>   <dbl> <chr>   <chr>     <dbl>   <dbl>   <dbl>   <dbl>   <dbl>
## 1 01/11/2~ 38079       2 Rolett~ ND         92.7   14539    95     14204    95
## 2 01/11/2~ 48391       2 Refugi~ TX         98.9    4253    61.2    4248    64.8
## 3 01/11/2~ 53025       2 Grant ~ WA         96     62848    64.3   62667    69.5
## # ... with 70 more variables: Administered_Dose1_Recip_12Plus <dbl>,
## #   Administered_Dose1_Recip_12PlusPop_Pct <dbl>,
## #   Administered_Dose1_Recip_18Plus <dbl>,
## #   Administered_Dose1_Recip_18PlusPop_Pct <dbl>,
## #   Administered_Dose1_Recip_65Plus <dbl>,
## #   Administered_Dose1_Recip_65PlusPop_Pct <dbl>, Series_Complete_Yes <dbl>,
## #   Series_Complete_Pop_Pct <dbl>, Series_Complete_5Plus <dbl>, ...
## # i Use `colnames()` to see all variable names
```

Load the acs data

```
head(acs, 3)
```

```
## # A tibble: 3 x 14
##   fips  county~1 total~2 prop_~3 prop_~4 media~5 media~6 media~7 prop_~8 prop_~9
##   <chr> <chr>      <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>
## 1 01001 Autauga~   42175   0.768  0.0286    38.2   58731     986  0.115    0.266
## 2 01003 Baldwin~  166595   0.862  0.0426    43     58320    1020  0.0919   0.319
## 3 01005 Barbour~   20054   0.468  0.0325    40.4   32525     576  0.268    0.116
## # ... with 4 more variables: prop_unemployed <dbl>, prop_nilf <dbl>,
## #   prop_health_insurance <dbl>, prop_low_ratio_ip <dbl>, and abbreviated
## #   variable names 1: county_name, 2: total_pop_18plus, 3: prop_white,
## #   4: prop_foreign_born, 5: median_age, 6: median_income, 7: median_rent,
## #   8: prop_less_than_hs, 9: prop_bachelor_above
## # i Use `colnames()` to see all variable names
```

Join the two datasets by the county names, followed by some data cleaning

```
county_data <- vax %>%
    merge(acs, by.x = "FIPS", by.y = "fips")
# Clean the 18+ vaccinated population data
county_data$Series_Complete_18Plus = county_data$Series_Complete_18Plus %>%
    str_remove_all(",") %>%
    as.integer()
# Calculate vaccination rate
county_data$vax_rate = with(county_data, Series_Complete_18Plus/total_pop_18plus)
county_data = county_data %>%
    filter(vax_rate <= 1)

## Not printing because output is quite long head(county_data, 3)
```

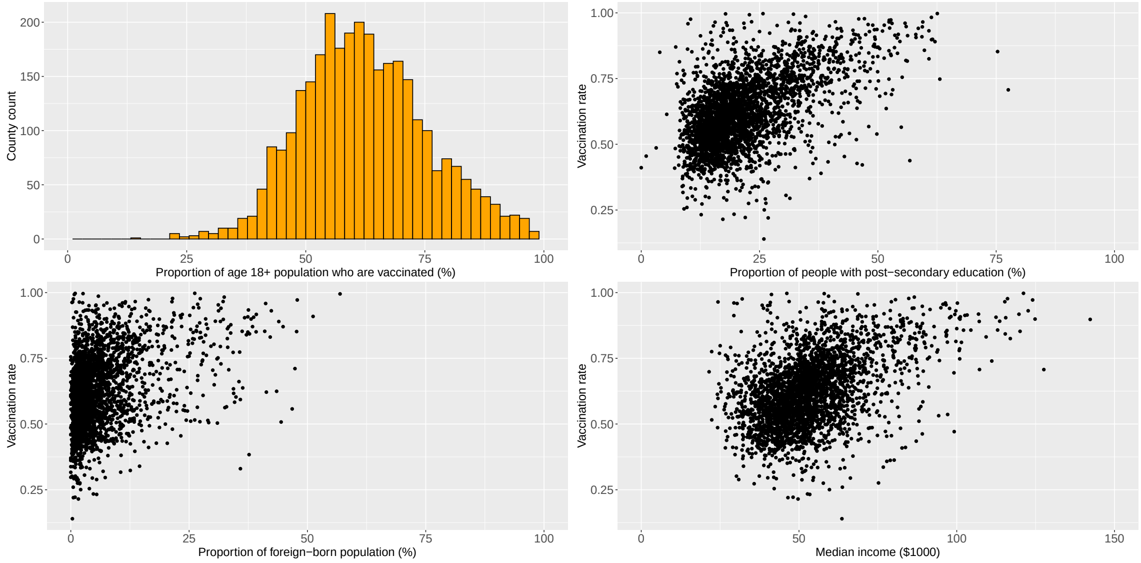
a) Exploratory data analysis

The following plots show the distribution of the vaccination rates in the age 18+ population over counties and its relationship to several variables present in the census data, including education level, foreign born proportion, median income, health insurance coverage, employment status, ethnicity, and median age.

```
par(mfrow = c(2, 2), mai = c(0.9, 0.9, 0.2, 0.2))
theme_large_text <- theme(axis.text = element_text(size = 15), axis.title = element_text(size = 15))

county_data %>%
    ggplot(aes(x = vax_rate * 100)) + geom_histogram(bins = 50, color = "black",
    fill = "orange") + labs(x = "Proportion of age 18+ population who are vaccinated (%)",
    y = "County count") + xlim(0, 100) + theme_large_text
```

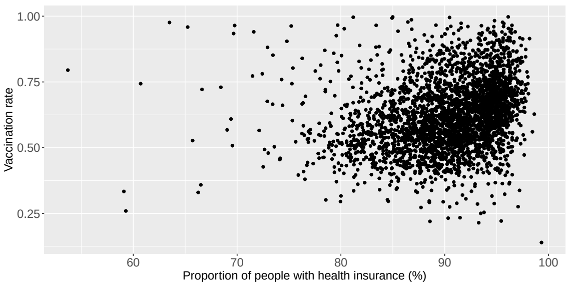
```
## Warning: Removed 2 rows containing missing values (`geom_bar()`).
```

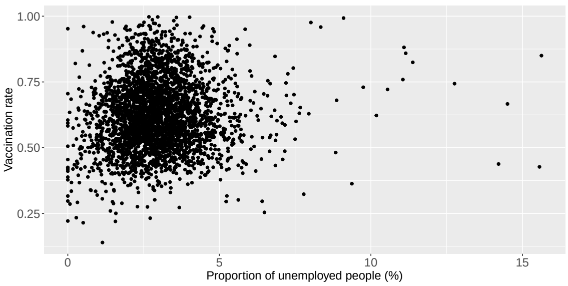
```
county_data %>%
    ggplot(aes(x = prop_bachelor_above * 100, y = vax_rate)) + geom_point(stroke = 0.5) +
    labs(x = "Proportion of people with post-secondary education (%)", y = "Vaccination rate") +
    xlim(0, 100) + theme_large_text

county_data %>%
    ggplot(aes(x = prop_foreign_born * 100, y = vax_rate)) + geom_point(stroke = 0.5) +
    labs(x = "Proportion of foreign-born population (%)", y = "Vaccination rate") +
    xlim(0, 100) + theme_large_text

county_data %>%
    ggplot(aes(x = median_income/1000, y = vax_rate)) + geom_point(stroke = 0.5) +
    labs(x = "Median income ($1000)", y = "Vaccination rate") + xlim(0, 150) + theme_large_text
```



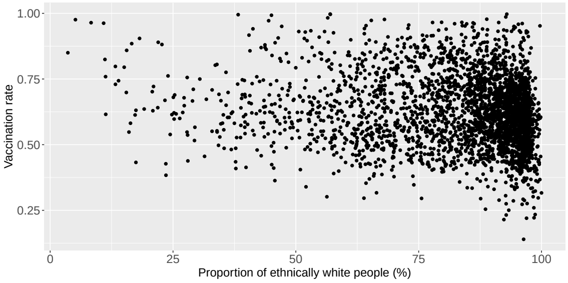
```
par(mfrow = c(1, 2), mai = c(0.9, 0.9, 0.2, 0.2))
county_data %>%
    ggplot(aes(x = prop_health_insurance * 100, y = vax_rate)) + geom_point(stroke = 0.5) +
    labs(x = "Proportion of people with health insurance (%)", y = "Vaccination rate") +
    theme_large_text
```

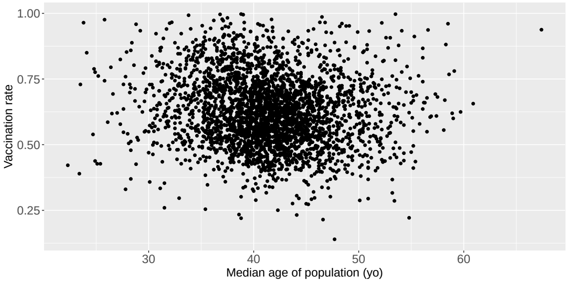
```
county_data %>%
    ggplot(aes(x = prop_unemployed * 100, y = vax_rate)) + geom_point(stroke = 0.5) +
    labs(x = "Proportion of unemployed people (%)", y = "Vaccination rate") + theme_large_text
```





```
par(mfrow = c(1, 2), mai = c(0.9, 0.9, 0.2, 0.2))
county_data %>%
    ggplot(aes(x = prop_white * 100, y = vax_rate)) + geom_point(stroke = 0.5) +
    labs(x = "Proportion of ethnically white people (%)", y = "Vaccination rate") +
    theme_large_text

county_data %>%
    ggplot(aes(x = median_age, y = vax_rate)) + geom_point(stroke = 0.5) + labs(x = "Median age of popul
    y = "Vaccination rate") + theme_large_text
```





```
vr_mean <- mean(county_data$vax_rate, na.rm = TRUE)
vr_var <- var(county_data$vax_rate, na.rm = TRUE)
sprintf("Mean is %f, variance is %f", vr_mean, vr_var)
```

```
## [1] "Mean is 0.624039, variance is 0.017476"
```

The vaccination rate data doesn't exhibit overdispersion

b) Regression model at the county level for vaccination rates of people aged 18+

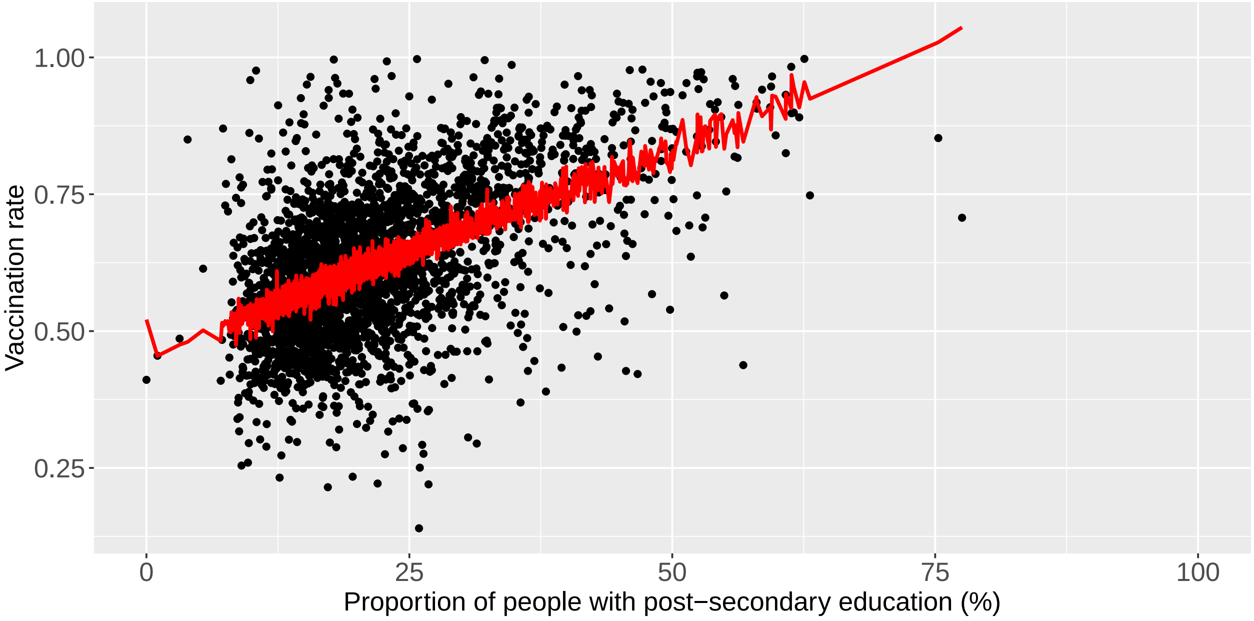
Build a GLM using the county-level explanatory variables such as the proportion of people who received post-secondary level of education (`prop_bachelor_above`), of people who has health insurance (`prop_health_insurance`), and people's median income (`median_income`). These variables positively correlate with the vaccination rates as found in the EDA.

```
vax_gau_fit <- glm(vax_rate ~ prop_bachelor_above + prop_health_insurance + median_income,
    data = county_data, family = "gaussian")
summary(vax_gau_fit)
```

```
##
```

```
## Call:
## glm(formula = vax_rate ~ prop_bachelor_above + prop_health_insurance +
##     median_income, family = "gaussian", data = county_data)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -0.53460  -0.06982   0.00268   0.06746   0.48835
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)           2.865e-01  3.672e-02   7.801 8.34e-15 ***
## prop_bachelor_above   6.056e-01  2.995e-02  20.225  < 2e-16 ***
## prop_health_insurance 1.614e-01  4.232e-02   3.814 0.000139 ***
## median_income         1.109e-06  2.017e-07   5.501 4.09e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.01228647)
##
##     Null deviance: 54.124  on 3097  degrees of freedom
## Residual deviance: 38.014  on 3094  degrees of freedom
## AIC: -4831.2
##
## Number of Fisher Scoring iterations: 2
```

```
pred_gau_vax_rate <- predict(vax_gau_fit)
ggplot(county_data, aes(x = prop_bachelor_above * 100, y = vax_rate)) + geom_point(stroke = 0.5) +
    labs(x = "Proportion of people with post-secondary education (%)", y = "Vaccination rate") +
    xlim(0, 100) + theme_large_text + geom_line(aes(y = pred_gau_vax_rate), linewidth = 1,
    color = "red")
```



c) Model prediction for Ada County, Idaho

```
glm_pred <- predict.glm(vax_gau_fit, county_data %>%
    filter(county_name == "Ada County, Idaho"), type = "response")

ada_value <- county_data %>%
    filter(county_name == "Ada County, Idaho") %>%
    dplyr::select(vax_rate)

sprintf("Vaccination rate for Ada county predicted by the GLM is %f, and the true value is %f",
    glm_pred, ada_value)
```

```
## [1] "Vaccination rate for Ada county predicted by the GLM is 0.741398, and the true value is 0.819704
```

d) Summary of analysis

A GLM was built to predict county-level vaccination rate (continuous random variable valued between 0 and 1) using census data. The initial EDA show a few variables with county-level information that positively correlates with the vaccination rates in the age 18+ population. These variables concern the education history and socioeconomic status of the county-level population. The GLM was used to predict the vaccination rate of residents of Ada County, Idaho, and the result shows good agreement with the actual value.

e) Potential outcomes from alternative models

1) Regression at the state level, outcome used is the total population 18+ fully vaccinated

   The model needs the age 18+ population of each state as a covariate. Otherwise, states similar in total population distribution but different in the proportion of the age 18+ population would be indistinguishable.

2) Regression at the state level, outcome used is the average of the county level full vaccination rates of 18+ population

   The model can make use of average vaccination rate at the state level over its constituent counties. It needs less granular information than 1) and 3).

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.

   The model would require the total population of each county as covariates. The same reasoning in 1) applies here.