BST210 Project Checkin2 Question 7 (Appendix)

Group Number: 7

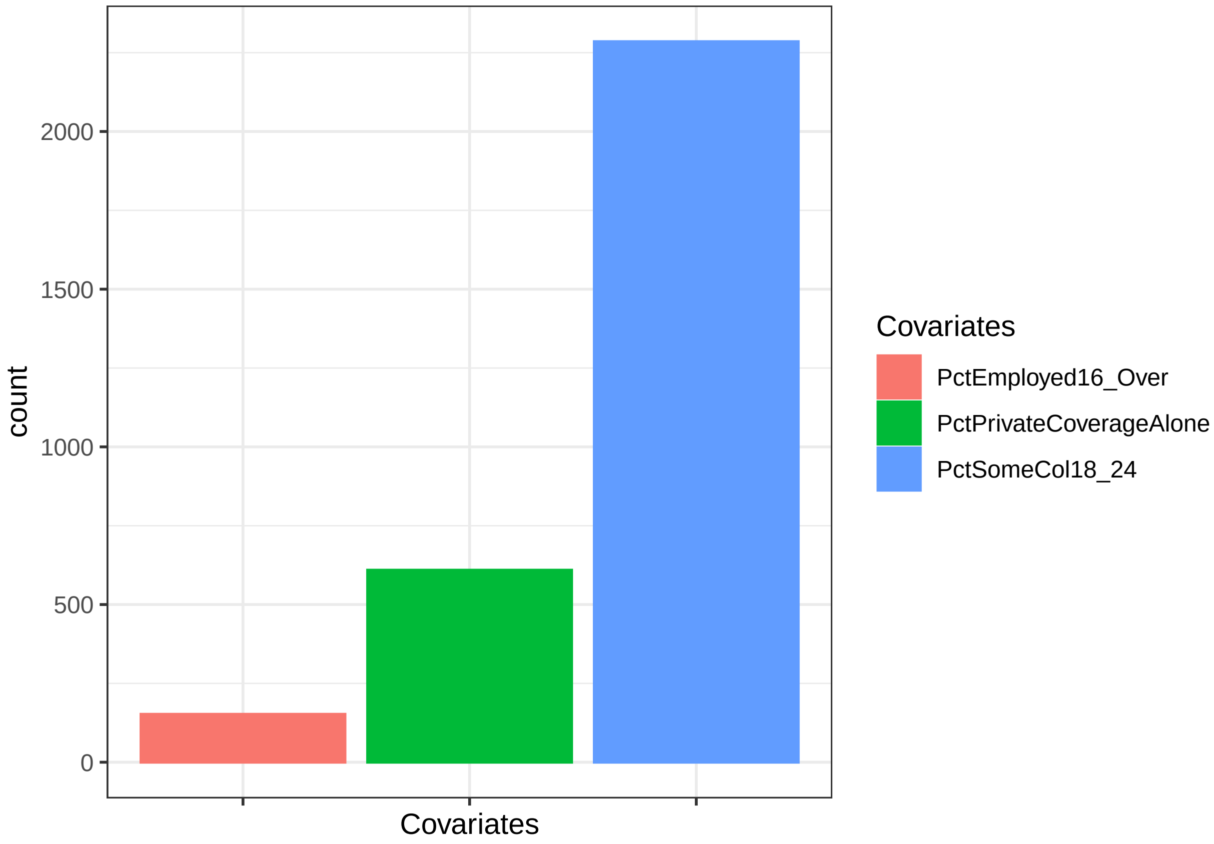
Group Name: Regression Heroes

Group Members: Ryan Wang, Stella Nam, Hongkai Wang

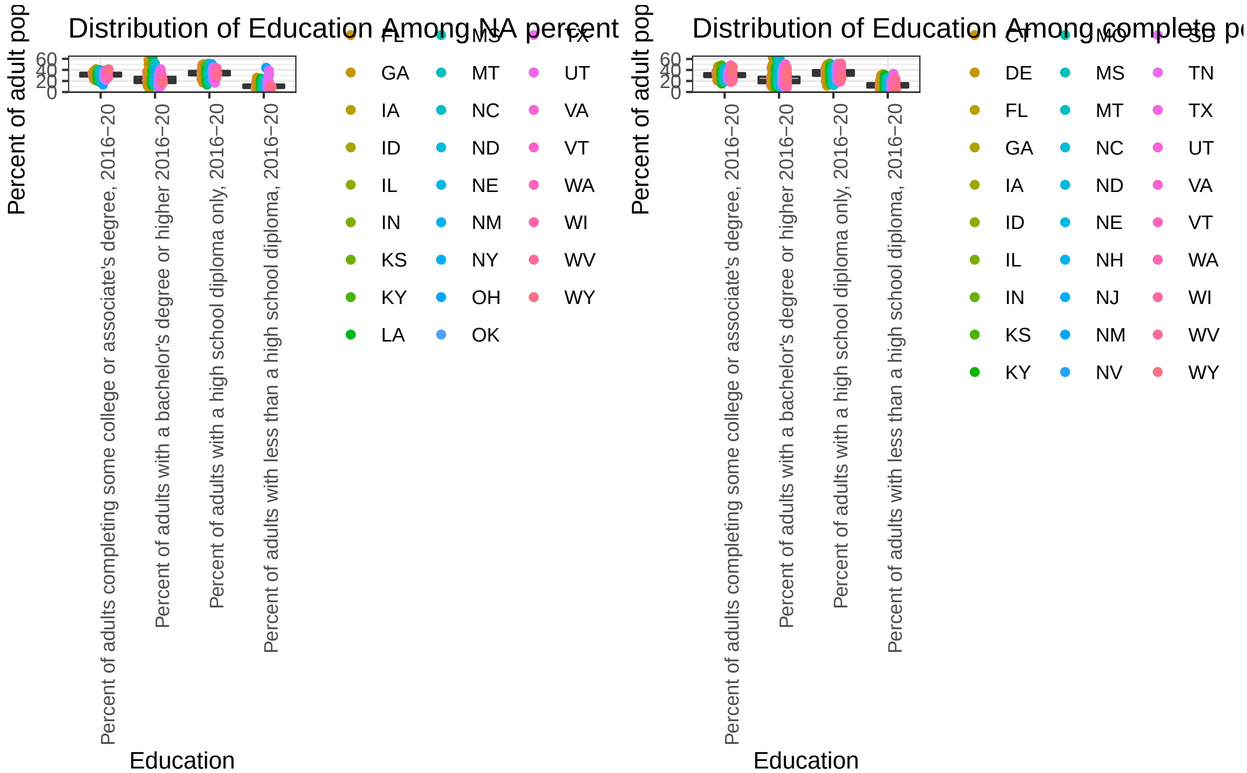
```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.3.6      v purrr   0.3.4
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.4.1
## v readr   2.1.2      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

4. Missing data

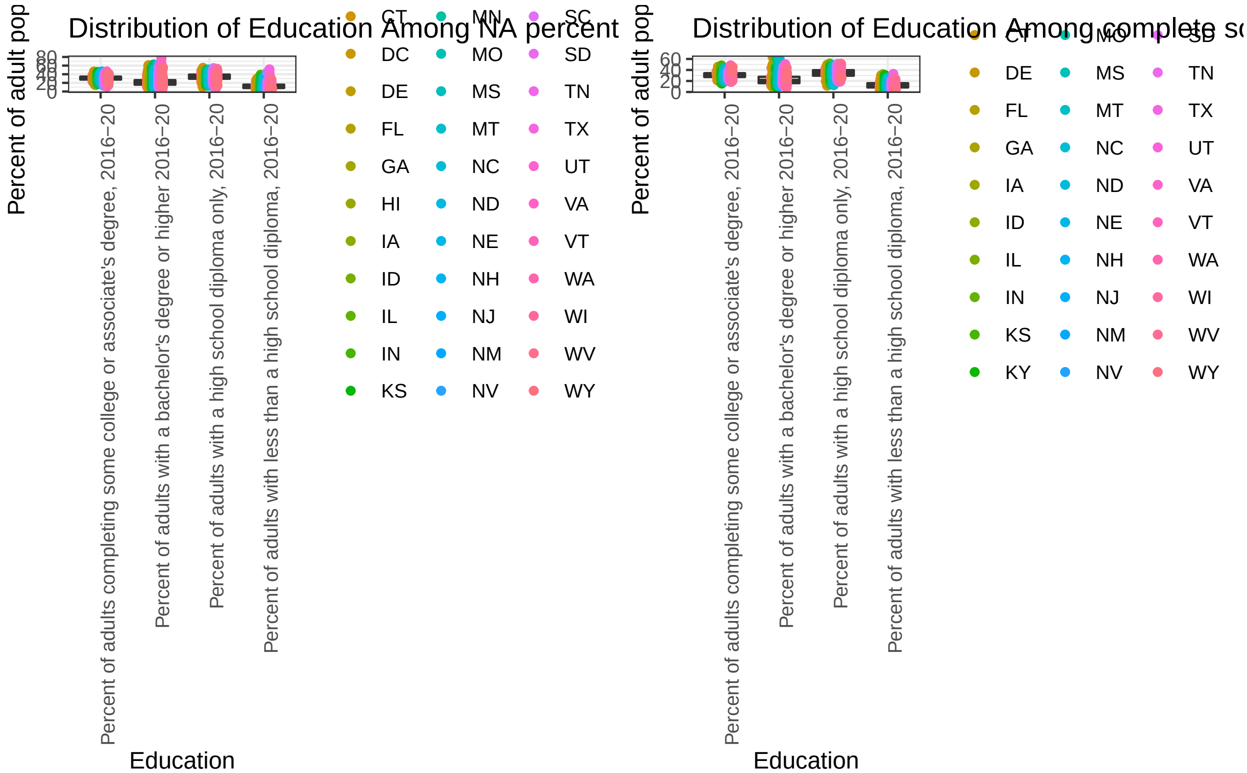
```
no_geodat <- dat %>% select(-c("Geography", "State Capital", "Region", "state", "State", "county_name",
no_geodat[!complete.cases(no_geodat),] %>%  # keep rows with NAs
    pivot_longer(colnames(no_geodat), names_to = "Covariates", values_to = "Values") %>% # pivot into l
    filter(is.na(Values)) %>% # filter out all the non-na's
    ggplot(aes(x = Covariates)) +
    geom_bar(position = "dodge", aes(col = Covariates, fill = Covariates)) +
    theme_bw() +
    # scale_fill_viridis_d() +
    # scale_color_viridis_d() +
    theme(axis.text.x = element_blank())
```



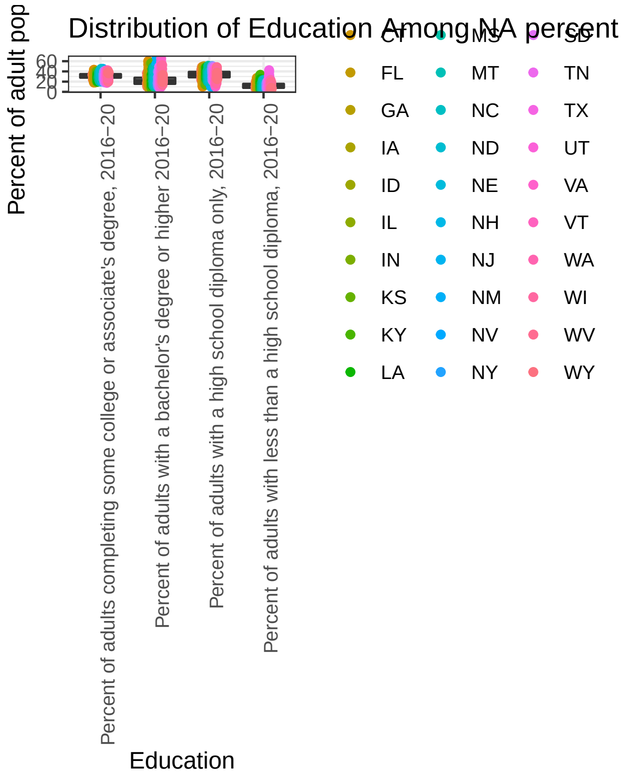

```
# NAs grouped by education
p1a <- dat[!complete.cases(dat),] %>%  # keep rows with NAs
    pivot_longer(colnames(ed)[8:11], names_to = "Education", values_to = "Values") %>% # pivot into lon
    filter(is.na(PctEmployed16_Over)) %>% # filter out all the non-na's
    ggplot(aes(x = Education, y = Values)) +
    geom_boxplot(show.legend = F, outlier.shape = NA) +
    geom_point(aes(col = State), position = position_jitterdodge(jitter.width=0, dodge.width = 0.3)) +
    theme_bw() +
    theme(axis.text.x = element_text(angle = 90, hjust=1)) +
    labs(y = "Percent of adult population", title = "Distribution of Education Among NA percent employe
    # scale_fill_viridis_d() +
    # scale_color_viridis_d() +

p1b <- dat[complete.cases(dat),] %>%  # keep rows with NAs
    pivot_longer(colnames(ed)[8:11], names_to = "Education", values_to = "Values") %>% # pivot into lon
    filter(!is.na(PctEmployed16_Over)) %>% # filter out all the non-na's
    ggplot(aes(x = Education, y = Values)) +
    geom_boxplot(show.legend = F, outlier.shape = NA) +
    geom_point(aes(col = State), position = position_jitterdodge(jitter.width=0, dodge.width = 0.3)) +
    theme_bw() +
    theme(axis.text.x = element_text(angle = 90, hjust=1)) +
    labs(y = "Percent of adult population", title = "Distribution of Education Among complete percent e

gridExtra::grid.arrange(p1a, p1b, ncol = 2)
```

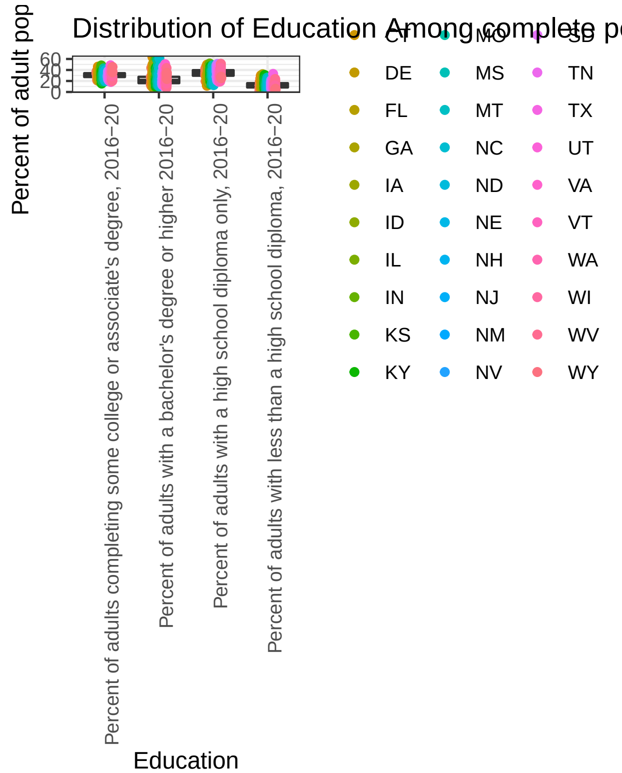



```r
# NAs grouped by education
p1a <- dat[!complete.cases(dat),] %>%  # keep rows with NAs
    pivot_longer(colnames(ed)[8:11], names_to = "Education", values_to = "Values") %>% # pivot into lon
    filter(is.na(PctSomeCol18_24)) %>% # filter out all the non-na's
    ggplot(aes(x = Education, y = Values)) +
    geom_boxplot(show.legend = F, outlier.shape = NA) +
    geom_point(aes(col = State), position = position_jitterdodge(jitter.width=0, dodge.width = 0.3)) +
    theme_bw() +
    theme(axis.text.x = element_text(angle = 90, hjust=1)) +
    labs(y = "Percent of adult population", title = "Distribution of Education Among NA percent some co
    # scale_fill_viridis_d() +
    # scale_color_viridis_d() +

p1b <- dat[complete.cases(dat),] %>%  # keep rows with NAs
    pivot_longer(colnames(ed)[8:11], names_to = "Education", values_to = "Values") %>% # pivot into lon
    filter(!is.na(PctSomeCol18_24)) %>% # filter out all the non-na's
    ggplot(aes(x = Education, y = Values)) +
    geom_boxplot(show.legend = F, outlier.shape = NA) +
    geom_point(aes(col = State), position = position_jitterdodge(jitter.width=0, dodge.width = 0.3)) +
    theme_bw() +
    theme(axis.text.x = element_text(angle = 90, hjust=1)) +
    labs(y = "Percent of adult population", title = "Distribution of Education Among complete some coll

gridExtra::grid.arrange(p1a, p1b, ncol = 2)
```




```
# NAs grouped by education
p1a <- dat[!complete.cases(dat),] %>%  # keep rows with NAs
    pivot_longer(colnames(ed)[8:11], names_to = "Education", values_to = "Values") %>% # pivot into lon
    filter(is.na(PctPrivateCoverageAlone)) %>% # filter out all the non-na's
    ggplot(aes(x = Education, y = Values)) +
    geom_boxplot(show.legend = F, outlier.shape = NA) +
    geom_point(aes(col = State), position = position_jitterdodge(jitter.width=0, dodge.width = 0.3)) +
    theme_bw() +
    theme(axis.text.x = element_text(angle = 90, hjust=1)) +
    labs(y = "Percent of adult population", title = "Distribution of Education Among NA percent alone p
    # scale_fill_viridis_d() +
    # scale_color_viridis_d() +

p1b <- dat[complete.cases(dat),] %>%  # keep rows with NAs
    pivot_longer(colnames(ed)[8:11], names_to = "Education", values_to = "Values") %>% # pivot into lon
    filter(!is.na(PctPrivateCoverageAlone)) %>% # filter out all the non-na's
    ggplot(aes(x = Education, y = Values)) +
    geom_boxplot(show.legend = F, outlier.shape = NA) +
    geom_point(aes(col = State), position = position_jitterdodge(jitter.width=0, dodge.width = 0.3)) +
    theme_bw() +
    theme(axis.text.x = element_text(angle = 90, hjust=1)) +
    labs(y = "Percent of adult population", title = "Distribution of Education Among complete percent a

gridExtra::grid.arrange(p1a, p1b, ncol = 2)
```






```
colMeans(is.na(no_geodat))*100
```

```
##                                                                           FIPS Code
##                                                                            0.000000
##                                         Less than a high school diploma, 2016-20
##                                                                            0.000000
##                                                High school diploma only, 2016-20
##                                                                            0.000000
##                                     Some college or associate's degree, 2016-20
##                                                                            0.000000
##                                            Bachelor's degree or higher, 2016-20
##                                                                            0.000000
##                         Percent of adults with less than a high school diploma, 2016-20
##                                                                            0.000000
##                                Percent of adults with a high school diploma only, 2016-20
##                                                                            0.000000
## Percent of adults completing some college or associate's degree, 2016-20
##                                                                            0.000000
##                              Percent of adults with a bachelor's degree or higher 2016-20
##                                                                            0.000000
##                                                                         avgAnnCount
##                                                                            0.000000
##                                                                     avgDeathsPerYear
##                                                                            0.000000
##                                                                     TARGET_deathRate
##                                                                            0.000000
##                                                                       incidenceRate
##                                                                            0.000000
```

```
##                                                    medIncome
##                                                     0.000000
##                                                   popEst2015
##                                                     0.000000
##                                               povertyPercent
##                                                     0.000000
##                                                  studyPerCap
##                                                     0.000000
##                                                    MedianAge
##                                                     0.000000
##                                                MedianAgeMale
##                                                     0.000000
##                                              MedianAgeFemale
##                                                     0.000000
##                                             AvgHouseholdSize
##                                                     0.000000
##                                               PercentMarried
##                                                     0.000000
##                                                 PctNoHS18_24
##                                                     0.000000
##                                                   PctHS18_24
##                                                     0.000000
##                                              PctSomeCol18_24
##                                                    74.991795
##                                              PctBachDeg18_24
##                                                     0.000000
##                                                 PctHS25_Over
##                                                     0.000000
##                                            PctBachDeg25_Over
##                                                     0.000000
##                                           PctEmployed16_Over
##                                                     4.988513
##                                         PctUnemployed16_Over
##                                                     0.000000
##                                           PctPrivateCoverage
##                                                     0.000000
##                                      PctPrivateCoverageAlone
##                                                    19.986872
##                                           PctEmpPrivCoverage
##                                                     0.000000
##                                            PctPublicCoverage
##                                                     0.000000
##                                       PctPublicCoverageAlone
##                                                     0.000000
##                                                     PctWhite
##                                                     0.000000
##                                                     PctBlack
##                                                     0.000000
##                                                     PctAsian
##                                                     0.000000
##                                                 PctOtherRace
##                                                     0.000000
##                                         PctMarriedHouseholds
##                                                     0.000000
```

```
##                                                                  BirthRate
##                                                                   0.000000
```

```r
ed_missing <- setdiff(unique(ed$county_name), unique(dat$county_name))
filter(ed, county_name %in% ed_missing)
```

```
## # A tibble: 191 x 11
##    FIPS ~1 State count~2 Less ~3 High ~4 Some ~5 Bache~6 Perce~7 Perce~8 Perce~9
##      <dbl> <chr> <chr>     <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>
##  1       0 US    United~  2.56e7  5.94e7  6.45e7  7.34e7   11.5     26.7    28.9
##  2    1000 AL    Alabama  4.39e5  1.01e6  1.01e6  8.77e5   13.1     30.3    30.3
##  3    1035 AL    Conecu~  1.33e3  4.12e3  2.15e3  1.22e3   15.0     46.7    24.4
##  4    2000 AK    Alaska   3.32e4  1.37e5  1.68e5  1.45e5    6.86    28.4    34.7
##  5    2013 AK    Aleuti~  3.54e2  1.08e3  7.77e2  4.38e2   13.4     40.7    29.4
##  6    2060 AK    Bristo~  3.5 e1  1.68e2  2.13e2  1.45e2    6.24    29.9    38.0
##  7    2063 AK    Chugac~  2.04e2  1.13e3  1.99e3  1.36e3    4.35    24.0    42.6
##  8    2066 AK    Copper~  8.6 e1  6.5 e2  5.51e2  5.93e2    4.57    34.6    29.3
##  9    2068 AK    Denali~  4.2 e1  5.63e2  5.31e2  7.3 e2    2.25    30.2    28.5
## 10    2105 AK    Hoonah~  1.05e2  6.35e2  6.62e2  3.42e2    6.02    36.4    38.0
## # ... with 181 more rows, 1 more variable:
## #   ‘Percent of adults with a bachelor’s degree or higher 2016-20‘ <dbl>, and
## #   abbreviated variable names 1: ‘FIPS Code‘, 2: county_name,
## #   3: ‘Less than a high school diploma, 2016-20‘,
## #   4: ‘High school diploma only, 2016-20‘,
## #   5: ‘Some college or associate’s degree, 2016-20‘,
## #   6: ‘Bachelor’s degree or higher, 2016-20‘, ...
```

```r
dat_missing <- setdiff(unique(dat$county_name), unique(ed$county_name))
filter(dat, county_name %in% dat_missing)
```

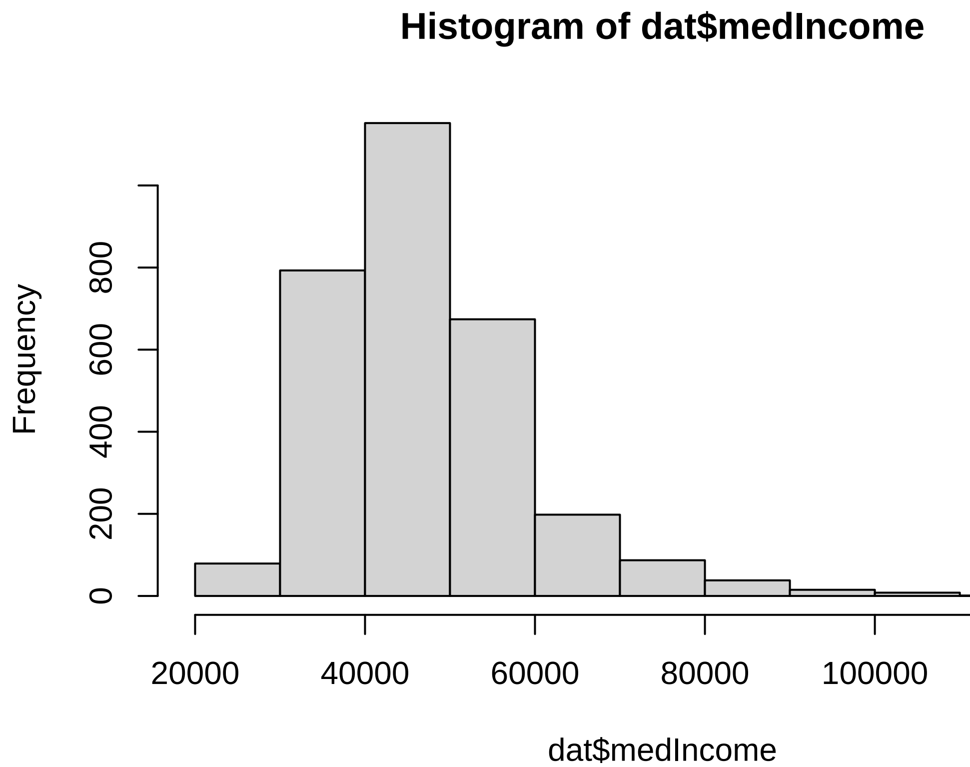
```
## # A tibble: 0 x 48
## # ... with 48 variables: FIPS Code <dbl>, State <chr>, county_name <chr>,
## #   Less than a high school diploma, 2016-20 <dbl>,
## #   High school diploma only, 2016-20 <dbl>,
## #   Some college or associate’s degree, 2016-20 <dbl>,
## #   Bachelor’s degree or higher, 2016-20 <dbl>,
## #   Percent of adults with less than a high school diploma, 2016-20 <dbl>,
## #   Percent of adults with a high school diploma only, 2016-20 <dbl>, ...
```

Here we show some difference in county representation within our two integrated cancer trial and socioeconomic dataset with a dataset of education attainment by county. Notably, a large difference in the counties from both datasets is the inclusino of Puerto Rico. While the education dataset includes Puerto Rico, the cancer trial data set does not. This means this missing data is **MAR** for our primary inference since it depends on a covariate *State* (or **MNAR** for our secondary as county is an outcome), however we will consider our analysis without Puerto Rico as it is a unique situation and not localized to the North American land mass. Other missing cancer data are at the county level, not found in the education dataset, similarly, as we are focused on the cancer data, we will disregard these education data (as we have education data for all cancer-statistic counties we have).

5. Modelling Approches

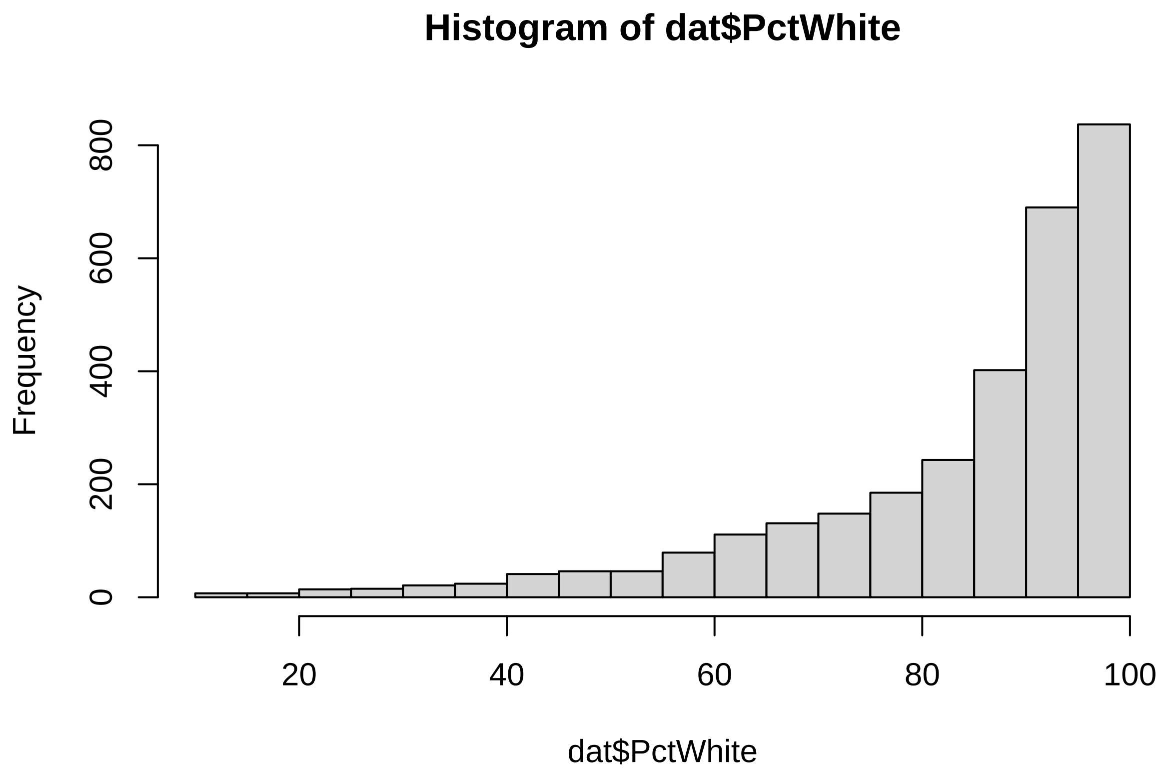
a. Fitting an linear model

```
hist(dat$medIncome)
```

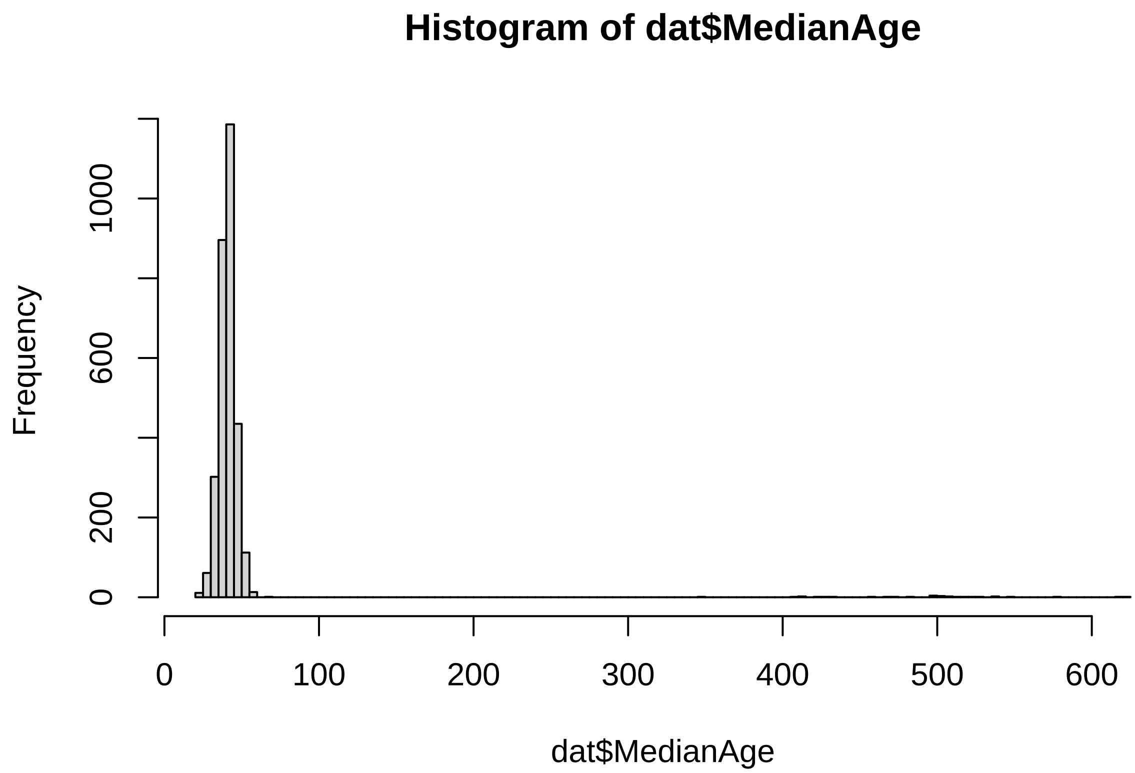


data transformation and cleaning:

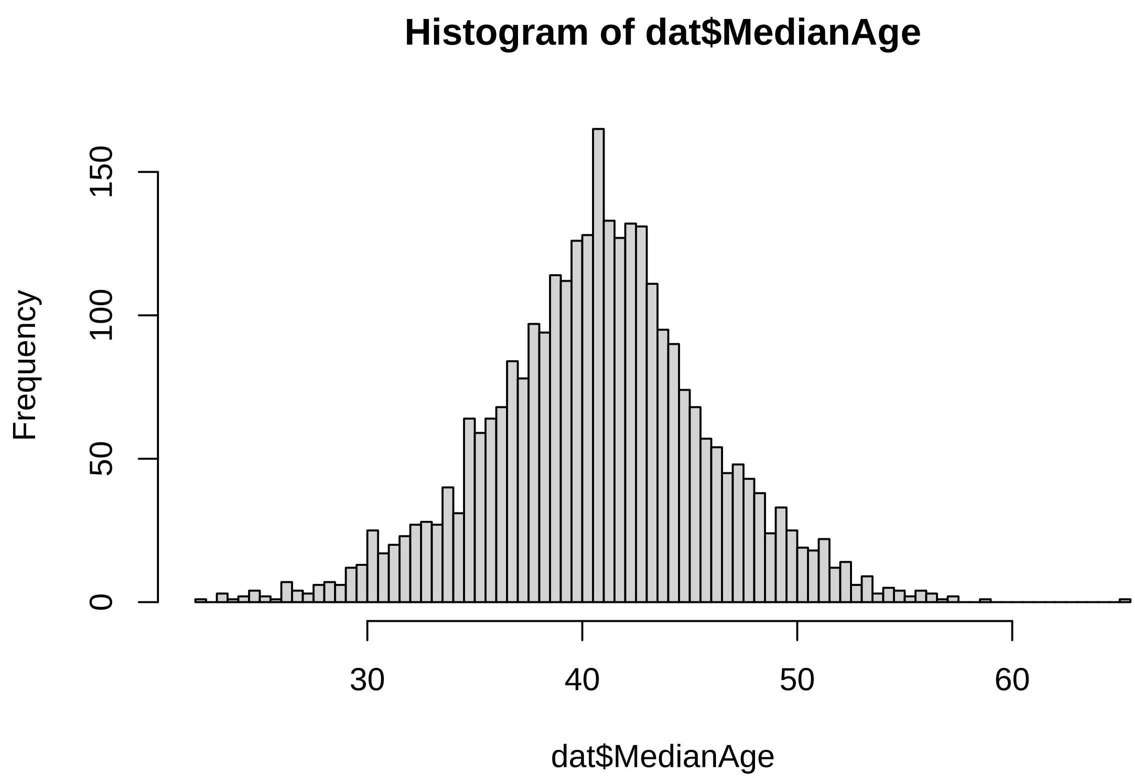
```
hist(dat$PctWhite)
```




```r
hist(dat$MedianAge, breaks = 100)
```




```
dat = dat %>% filter(MedianAge <= 100)

hist(dat$MedianAge, breaks = 100)
```




```r
mod1 = lm(data = dat, TARGET_deathRate ~ medIncome + MedianAge + PctWhite)
summary(mod1)
```

model fitting:

```
##
## Call:
## lm(formula = TARGET_deathRate ~ medIncome + MedianAge + PctWhite,
##     data = dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -123.117  -14.061    0.904   15.057  175.883
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.411e+02  4.250e+00  56.735  < 2e-16 ***
## medIncome   -9.492e-04  3.889e-05 -24.409  < 2e-16 ***
## MedianAge   -7.100e-02  9.591e-02  -0.740    0.459
## PctWhite    -1.785e-01  3.065e-02  -5.823  6.4e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 24.89 on 3013 degrees of freedom
## Multiple R-squared:  0.1954, Adjusted R-squared:  0.1946
## F-statistic:   244 on 3 and 3013 DF,  p-value: < 2.2e-16
```

```
mod1.1 = lm(data = dat, TARGET_deathRate ~ medIncome + PctWhite)
summary(mod1.1)
```

```
##
## Call:
## lm(formula = TARGET_deathRate ~ medIncome + PctWhite, data = dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -122.987  -14.167    0.874   15.145  175.870
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.387e+02  2.748e+00  86.894  < 2e-16 ***
## medIncome   -9.436e-04  3.813e-05 -24.746  < 2e-16 ***
## PctWhite    -1.876e-01  2.806e-02  -6.686 2.73e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 24.89 on 3014 degrees of freedom
## Multiple R-squared:  0.1953, Adjusted R-squared:  0.1947
## F-statistic: 365.7 on 2 and 3014 DF,  p-value: < 2.2e-16
```

```
anova(mod1.1, mod1)
```

```
## Analysis of Variance Table
##
## Model 1: TARGET_deathRate ~ medIncome + PctWhite
## Model 2: TARGET_deathRate ~ medIncome + MedianAge + PctWhite
##   Res.Df     RSS Df Sum of Sq      F Pr(>F)
## 1   3014 1867102
## 2   3013 1866763  1    339.57 0.5481 0.4592
```

```
mod1.2 = lm(data = dat, TARGET_deathRate ~ medIncome + PctWhite + MedianAge + medIncome*MedianAge)
summary(mod1.2)
```

```
##
## Call:
## lm(formula = TARGET_deathRate ~ medIncome + PctWhite + MedianAge +
##     medIncome * MedianAge, data = dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -123.331  -13.843    0.929   14.955  175.902
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          2.012e+02  1.625e+01  12.379  < 2e-16 ***
```

```
## medIncome           -7.301e-05  3.464e-04  -0.211   0.8331
## PctWhite            -1.806e-01  3.064e-02  -5.894  4.2e-09 ***
## MedianAge            9.390e-01  4.082e-01   2.301   0.0215 *
## medIncome:MedianAge -2.213e-05  8.693e-06  -2.546   0.0110 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 24.87 on 3012 degrees of freedom
## Multiple R-squared:  0.1972, Adjusted R-squared:  0.1961
## F-statistic: 184.9 on 4 and 3012 DF,  p-value: < 2.2e-16
```

```
anova(mod1.2, mod1)
```

```
## Analysis of Variance Table
##
## Model 1: TARGET_deathRate ~ medIncome + PctWhite + MedianAge + medIncome *
##     MedianAge
## Model 2: TARGET_deathRate ~ medIncome + MedianAge + PctWhite
##   Res.Df     RSS Df Sum of Sq      F  Pr(>F)
## 1   3012 1862754
## 2   3013 1866763 -1   -4008.3 6.4812 0.01095 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
mod1.3 = lm(data = dat, TARGET_deathRate ~ medIncome + PctWhite + MedianAge + PctWhite*MedianAge)
summary(mod1.3)
```

```
##
## Call:
## lm(formula = TARGET_deathRate ~ medIncome + PctWhite + MedianAge +
##     PctWhite * MedianAge, data = dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -123.122  -14.063    0.896   15.064  175.865
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          2.400e+02  1.824e+01  13.160   <2e-16 ***
## medIncome           -9.493e-04  3.890e-05 -24.401   <2e-16 ***
## PctWhite            -1.653e-01  2.123e-01  -0.779    0.436
## MedianAge           -4.166e-02  4.759e-01  -0.088    0.930
## PctWhite:MedianAge -3.438e-04  5.461e-03  -0.063    0.950
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 24.9 on 3012 degrees of freedom
## Multiple R-squared:  0.1954, Adjusted R-squared:  0.1944
## F-statistic: 182.9 on 4 and 3012 DF,  p-value: < 2.2e-16
```

```
anova(mod1.3, mod1)
```

```
## Analysis of Variance Table
##
## Model 1: TARGET_deathRate ~ medIncome + PctWhite + MedianAge + PctWhite *
##     MedianAge
## Model 2: TARGET_deathRate ~ medIncome + MedianAge + PctWhite
##   Res.Df     RSS Df Sum of Sq     F Pr(>F)
## 1   3012 1866760
## 2   3013 1866763 -1   -2.4565 0.004 0.9498
```

```
mod1.4 = lm(data = dat, TARGET_deathRate ~ medIncome + PctWhite + PctBlack + PctAsian + PctOtherRace)
summary(mod1.4)
```

```
##
## Call:
## lm(formula = TARGET_deathRate ~ medIncome + PctWhite + PctBlack +
##     PctAsian + PctOtherRace, data = dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -118.383  -13.918    0.866   14.279  174.216
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.364e+02  5.931e+00  39.869  < 2e-16 ***
## medIncome    -8.450e-04  4.302e-05 -19.640  < 2e-16 ***
## PctWhite     -1.917e-01  6.064e-02  -3.161  0.00159 **
## PctBlack      1.204e-01  6.406e-02   1.879  0.06033 .
## PctAsian     -2.622e-01  2.130e-01  -1.231  0.21854
## PctOtherRace -1.395e+00  1.414e-01  -9.865  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 24.33 on 3011 degrees of freedom
## Multiple R-squared:  0.2317, Adjusted R-squared:  0.2305
## F-statistic: 181.7 on 5 and 3011 DF,  p-value: < 2.2e-16
```

```
cor(dat$PctAsian, dat$PctWhite)
```

```
## [1] -0.2658648
```

```
cor(dat$PctBlack, dat$PctWhite)
```

```
## [1] -0.8312116
```

```
cor(dat$PctOtherRace, dat$PctWhite)
```

```
## [1] -0.2331931
```

```
mod1.5 = lm(data = dat, TARGET_deathRate ~ medIncome + PctWhite + PctAsian + PctOtherRace)
summary(mod1.5)
```

```
##
## Call:
## lm(formula = TARGET_deathRate ~ medIncome + PctWhite + PctAsian +
##     PctOtherRace, data = dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -118.382  -13.873    0.839   14.226  174.620
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.463e+02  2.763e+00  89.160   <2e-16 ***
## medIncome    -8.482e-04  4.301e-05 -19.723   <2e-16 ***
## PctWhite     -2.905e-01  3.025e-02  -9.604   <2e-16 ***
## PctAsian     -3.827e-01  2.032e-01  -1.883   0.0598 .
## PctOtherRace -1.495e+00  1.310e-01 -11.413   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 24.34 on 3012 degrees of freedom
## Multiple R-squared:  0.2308, Adjusted R-squared:  0.2298
## F-statistic:   226 on 4 and 3012 DF,  p-value: < 2.2e-16
```

```
anova(mod1.5, mod1.1)
```

```
## Analysis of Variance Table
##
## Model 1: TARGET_deathRate ~ medIncome + PctWhite + PctAsian + PctOtherRace
## Model 2: TARGET_deathRate ~ medIncome + PctWhite
##   Res.Df     RSS Df Sum of Sq     F    Pr(>F)
## 1   3012 1784592
## 2   3014 1867102 -2    -82510 69.63 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
mod1.6 = lm(data = dat, TARGET_deathRate ~ medIncome)
summary(mod1.6)
```

```
##
## Call:
## lm(formula = TARGET_deathRate ~ medIncome, data = dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -124.962  -14.433    0.937   15.098  177.402
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.250e+02  1.840e+00  122.29   <2e-16 ***
## medIncome   -9.856e-04  3.788e-05  -26.02   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 25.07 on 3015 degrees of freedom
## Multiple R-squared:  0.1833, Adjusted R-squared:  0.1831
## F-statistic: 676.9 on 1 and 3015 DF,  p-value: < 2.2e-16
```

```
mod1.6.1 = lm(data = dat, TARGET_deathRate ~ medIncome + I(medIncome ^2))
summary(mod1.6.1)
```

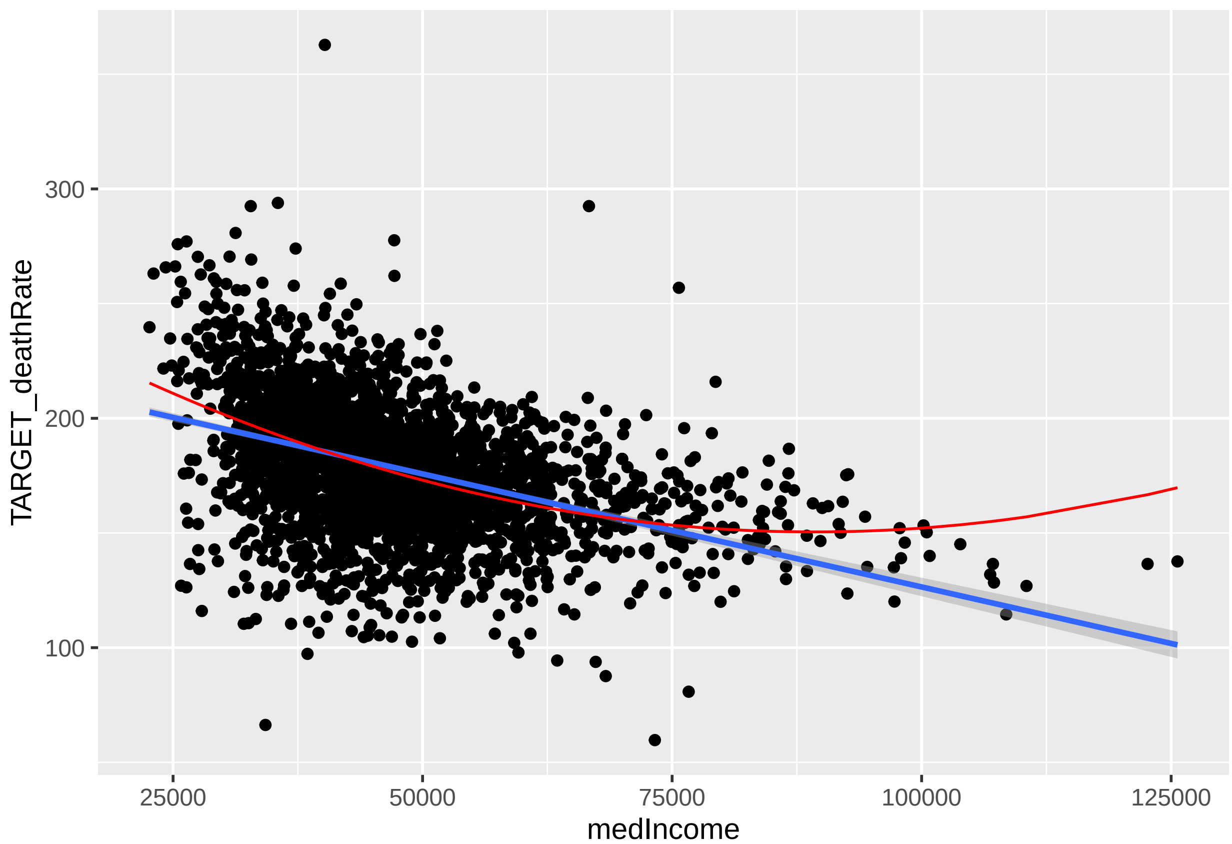
```
##
## Call:
## lm(formula = TARGET_deathRate ~ medIncome + I(medIncome^2), data = dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -128.419  -13.923    1.128   14.799  177.132
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     2.670e+02  4.952e+00  53.913   <2e-16 ***
## medIncome      -2.609e-03  1.822e-04 -14.318   <2e-16 ***
## I(medIncome^2)  1.461e-08  1.605e-09   9.104   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 24.74 on 3014 degrees of freedom
## Multiple R-squared:  0.2052, Adjusted R-squared:  0.2047
## F-statistic: 389.1 on 2 and 3014 DF,  p-value: < 2.2e-16
```

```
anova(mod1.6, mod1.6.1)
```

```
## Analysis of Variance Table
##
## Model 1: TARGET_deathRate ~ medIncome
## Model 2: TARGET_deathRate ~ medIncome + I(medIncome^2)
##   Res.Df     RSS Df Sum of Sq      F    Pr(>F)
## 1   3015 1894793
## 2   3014 1844082  1     50711 82.883 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
predict = data.frame(TARGET_deathRate = predict(mod1.6.1, dat), medIncome = dat$medIncome)

dat %>% ggplot(aes(medIncome, TARGET_deathRate)) + geom_point() + geom_smooth(method = "lm") + geom_lin
```

```
## `geom_smooth()` using formula 'y ~ x'
```



We now have this following core model:

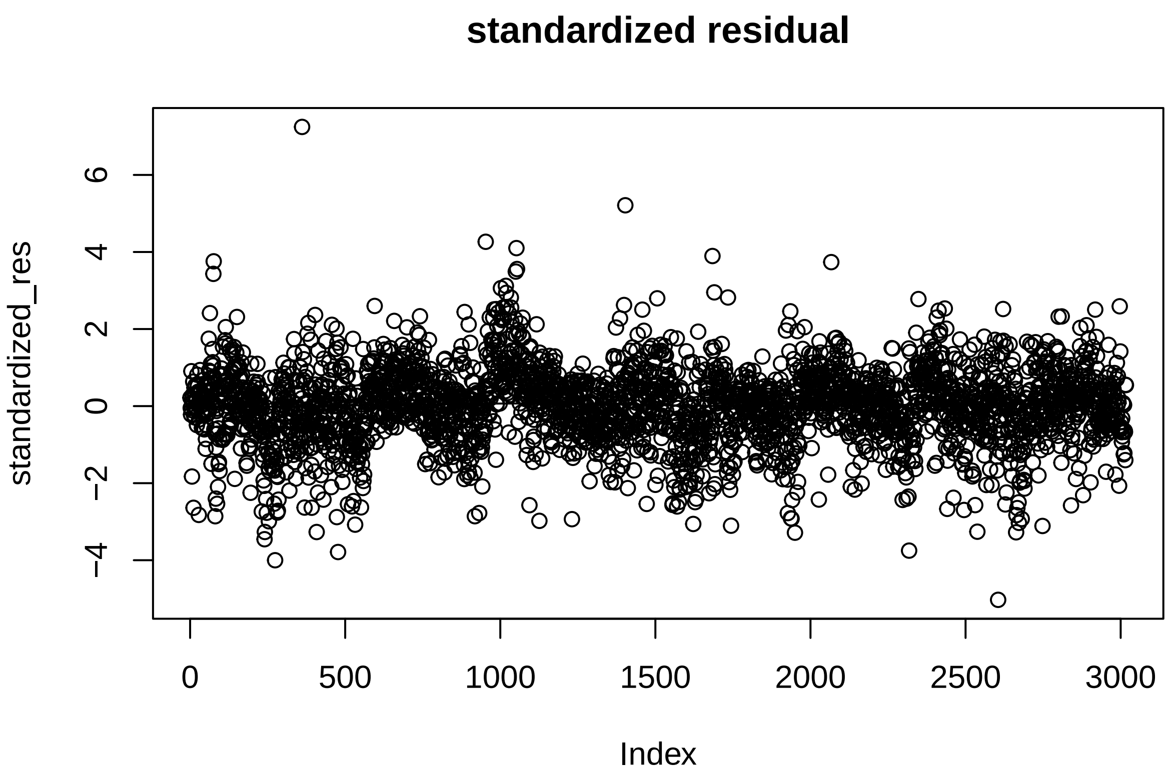
```
mod1_core = lm(data = dat, TARGET_deathRate ~ medIncome + I(medIncome ^2)+ PctWhite + PctAsian + PctOthe
summary(mod1_core)
```

```
##
## Call:
## lm(formula = TARGET_deathRate ~ medIncome + I(medIncome^2) +
##     PctWhite + PctAsian + PctOtherRace, data = dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -121.247  -13.750    1.302   14.195  174.871
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     2.756e+02  4.938e+00  55.821  < 2e-16 ***
## medIncome      -2.159e-03  1.885e-04 -11.456  < 2e-16 ***
## I(medIncome^2)  1.182e-08  1.656e-09   7.141 1.16e-12 ***
## PctWhite       -2.362e-01  3.095e-02  -7.632 3.08e-14 ***
## PctAsian       -5.664e-01  2.032e-01  -2.787  0.00535 **
## PctOtherRace   -1.419e+00  1.303e-01 -10.884  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 24.14 on 3011 degrees of freedom
```

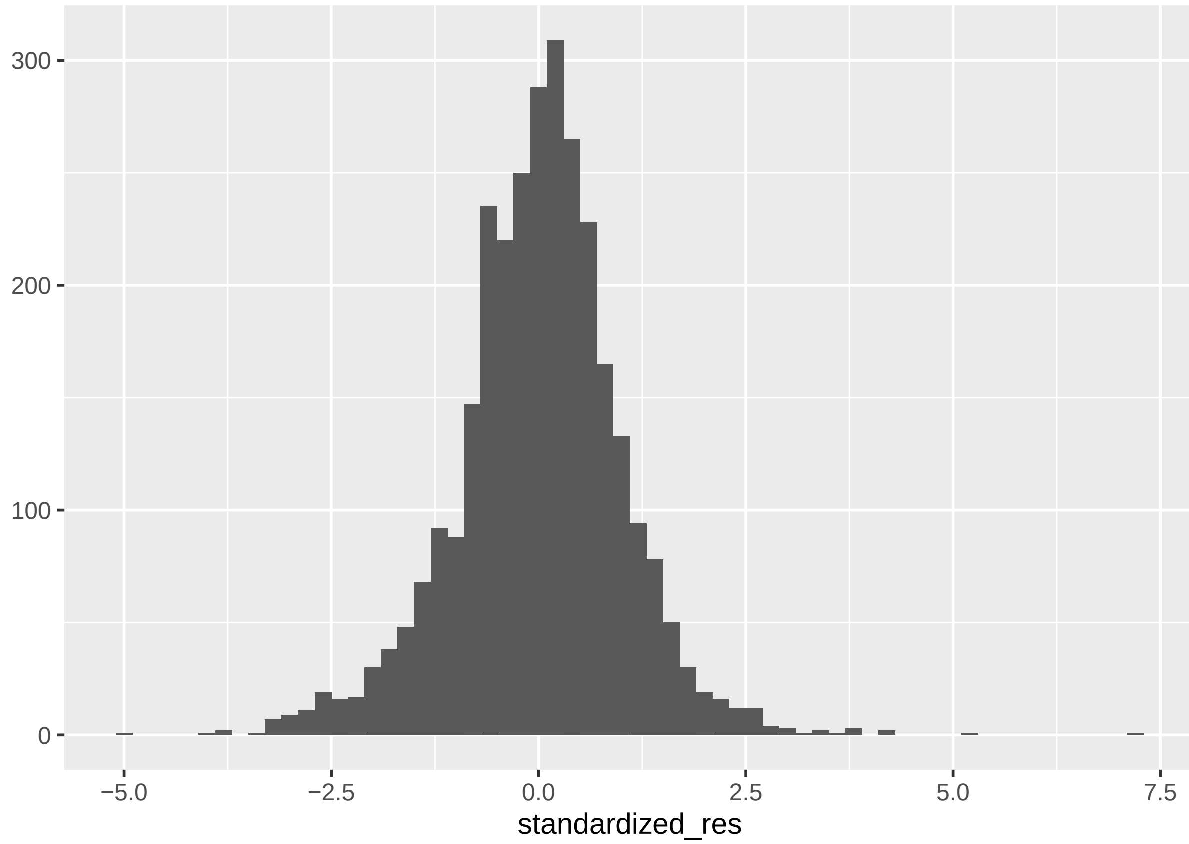
```
## Multiple R-squared:  0.2437, Adjusted R-squared:  0.2424
## F-statistic:   194 on 5 and 3011 DF,  p-value: < 2.2e-16
```

Let's evaluate the residual diagnostic to confirm the model's validity:

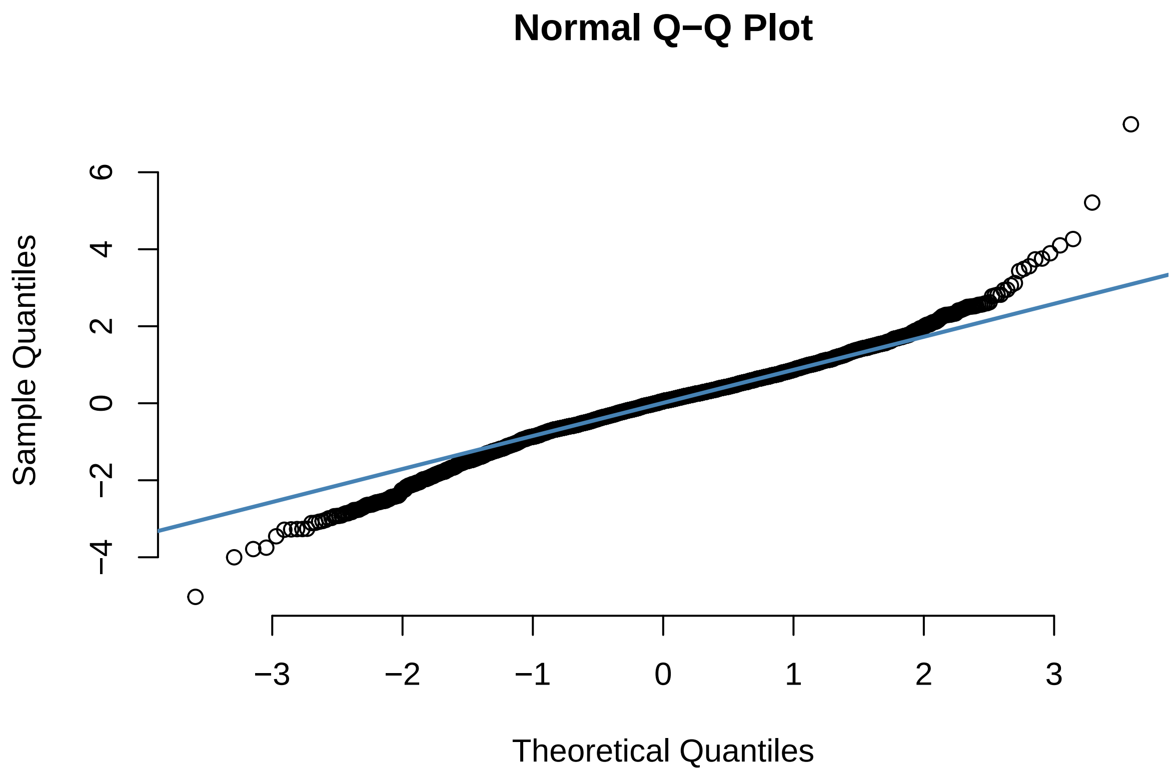
```
standardized_res = rstandard(mod1_core)
scatter.smooth(standardized_res, main = "standardized residual")
```




```
qplot(standardized_res, binwidth = 0.2)
```



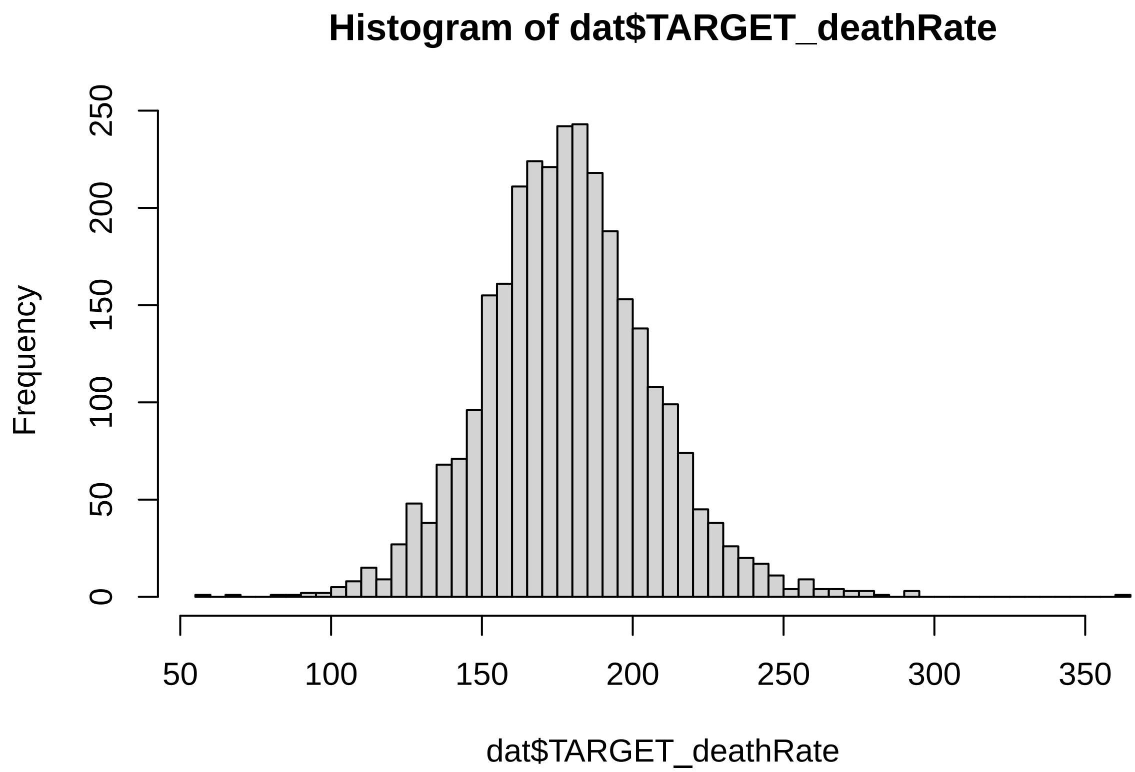

```
qqnorm(standardized_res, pch = 1, frame = FALSE)
qqline(standardized_res, col = "steelblue", lwd = 2)
```



b. Logistic/multinomial/ordinal regression

First, we can split the lung-cancer death rate into several categories to broadly access the healthcare system at each county. For example, we can artificially create three different categories in the death rate variable. Let's take a look at the death rate distribution.

```
hist(dat$TARGET_deathRate, breaks = 100)
```



We can make three bins, deathrate < 150, 150 <= deathrate < 200, 200 <= deathrate, and use them as a proxy to the quality of lung cancer prevention and quality of cancer care for each US county. The normal distribution of lung cancer death rate also suggests that dividing outcomes into bins doesn't really benefit us with our primary goal.

But let's do it anyway to test it out anyway:

```
#create the three bins
dat = dat %>% mutate(multi = case_when(TARGET_deathRate < 150 ~ 1, TARGET_deathRate < 200 ~ 2, T ~ 3))
# 3 is bad quality lung cancer prevention, 2 is medium, 1 is good quality.
```

```
library(nnet)
mod2.1 <- multinom(multi ~ medIncome + I(medIncome ^2), data = dat)
```

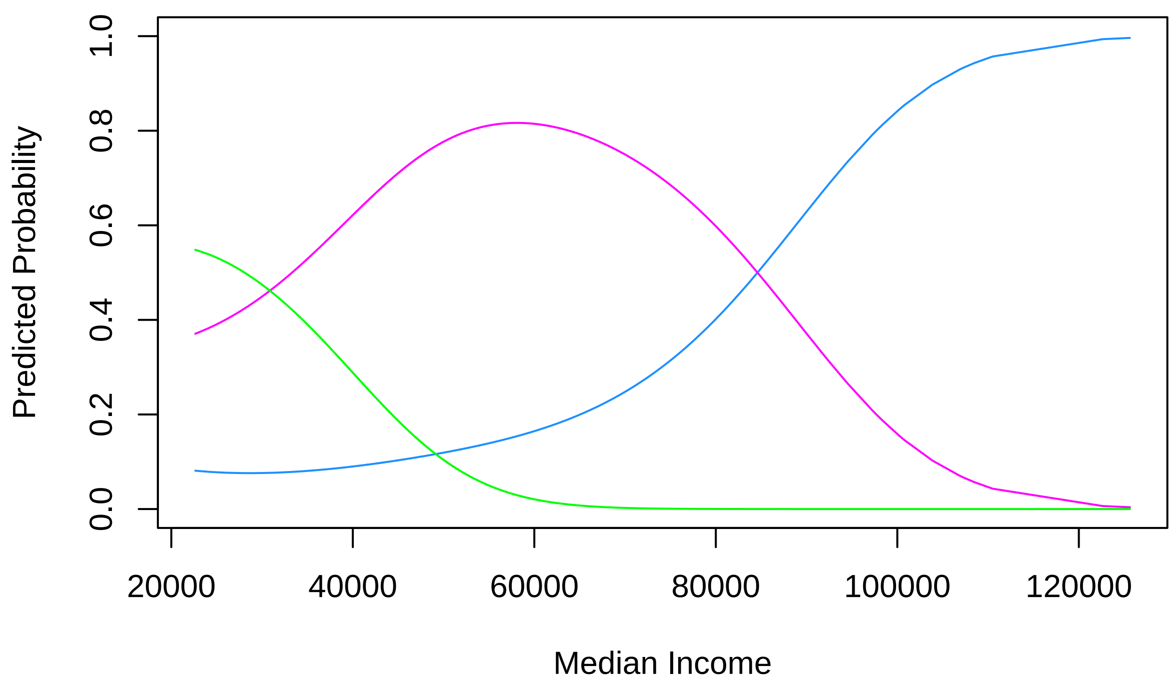
```
## # weights:  12 (6 variable)
## initial  value 3314.513275
## iter  10 value 2320.489209
## final  value 2316.095848
## converged
```

```
summary(mod2.1)
```

```
## Call:
## multinom(formula = multi ~ medIncome + I(medIncome^2), data = dat)
##
```

```
## Coefficients:
##    (Intercept)    medIncome I(medIncome^2)
## 2 2.417464e-09 9.165115e-05  -1.083448e-09
## 3 7.931352e-09 1.565121e-04  -3.185035e-09
##
## Std. Errors:
##    (Intercept)    medIncome I(medIncome^2)
## 2 6.249523e-21 2.523511e-16   1.799483e-11
## 3 1.659987e-20 7.128037e-16   3.143898e-11
##
## Residual Deviance: 4632.192
## AIC: 4644.192
```

```
plot(mod2.1$fitted.values[,1][order(dat$medIncome)] ~ sort(dat$medIncome), type="l", col="dodgerblue", :
points(mod2.1$fitted.values[,2][order(dat$medIncome)] ~ sort(dat$medIncome), type="l", col="magenta")
points(mod2.1$fitted.values[,3][order(dat$medIncome)]~sort(dat$medIncome), type="l", col="green")
```



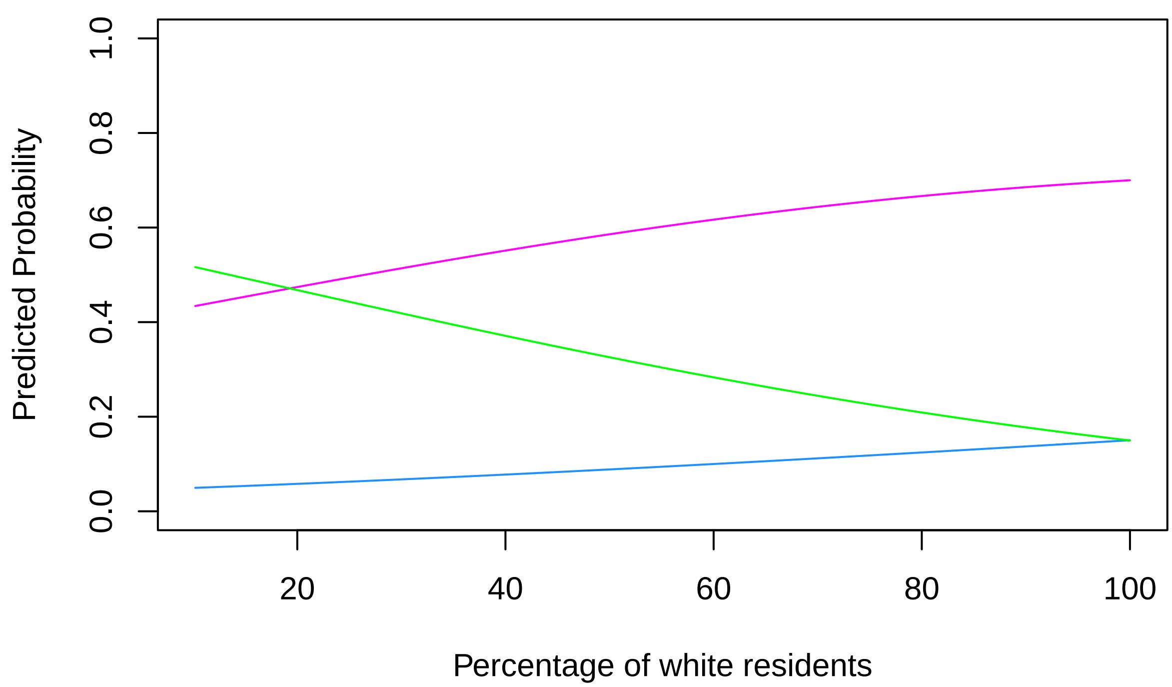

```
mod2.2 <- multinom(multi ~ PctWhite , data = dat)
```

```
## # weights:  9 (4 variable)
## initial  value 3314.513275
## iter  10 value 2552.935350
## iter  10 value 2552.935350
## final  value 2552.935350
## converged
```

```r
summary(mod2.2)
```

```
## Call:
## multinom(formula = multi ~ PctWhite, data = dat)
##
## Coefficients:
##   (Intercept)     PctWhite
## 2    2.241079 -0.007027921
## 3    2.609593 -0.026135639
##
## Std. Errors:
##   (Intercept)    PctWhite
## 2   0.3398884 0.003901638
## 3   0.3612030 0.004210873
##
## Residual Deviance: 5105.871
## AIC: 5113.871
```

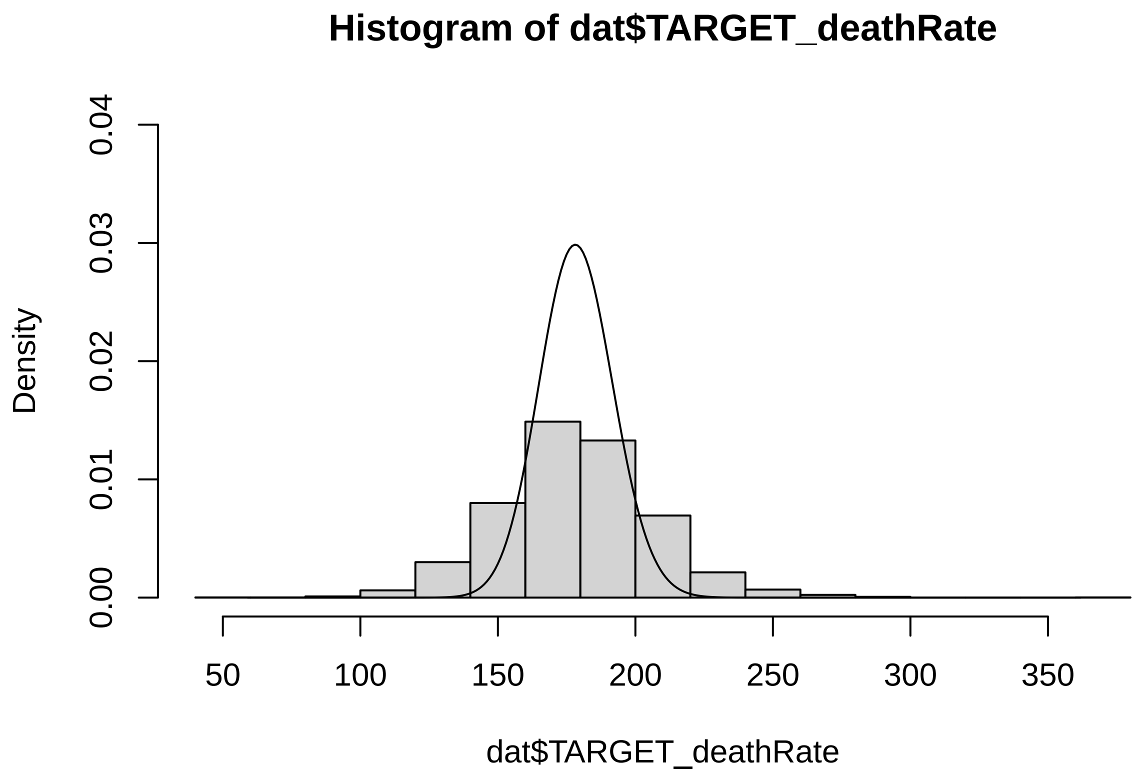
```r
plot(mod2.2$fitted.values[,1][order(dat$PctWhite)] ~ sort(dat$PctWhite), type="l", col="dodgerblue", xla
points(mod2.2$fitted.values[,2][order(dat$PctWhite)] ~ sort(dat$PctWhite), type="l", col="magenta")
points(mod2.2$fitted.values[,3][order(dat$PctWhite)]~sort(dat$PctWhite), type="l", col="green")
```



c. Poisson Regression

Over-dispersion

```
hist(dat$TARGET_deathRate, freq = F, ylim = c(0, 0.04))
lines(as.integer(min(dat$TARGET_deathRate)):as.integer(max(dat$TARGET_deathRate)), dpois(as.integer(min
```



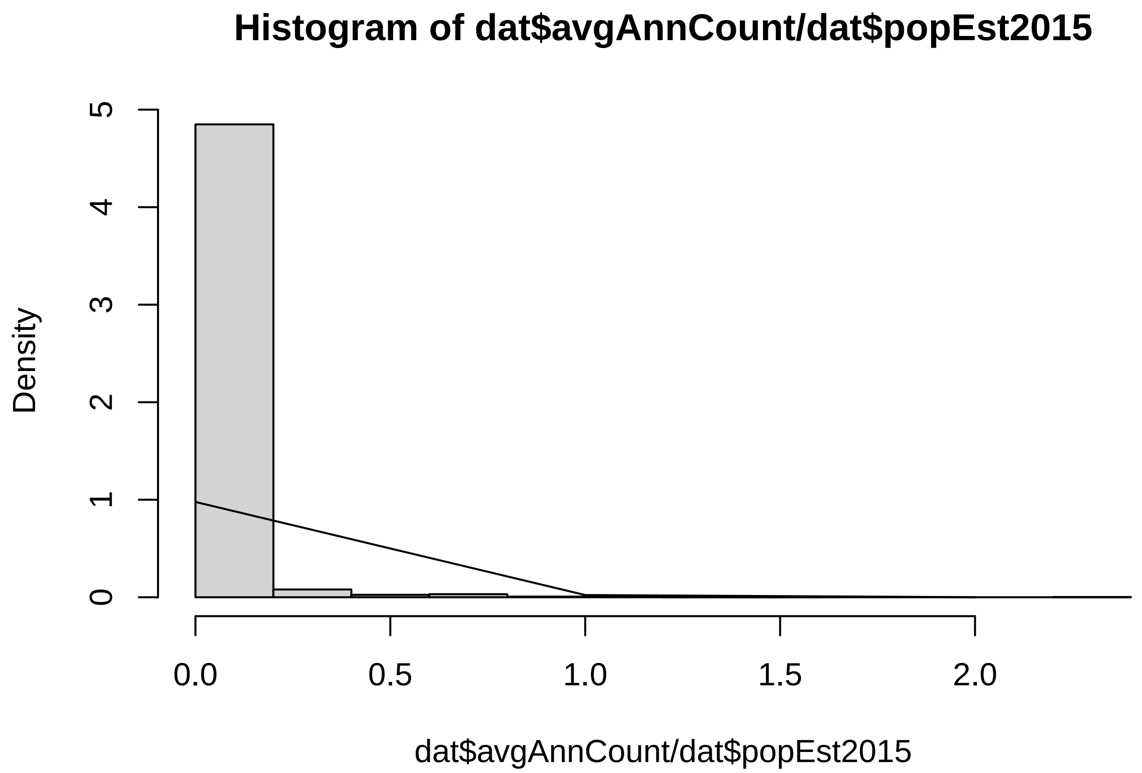

```
print(mean(dat$TARGET_deathRate))
```

```
## [1] 178.6452
```

```
print(var(dat$TARGET_deathRate))
```

```
## [1] 769.2961
```

```
hist(dat$avgAnnCount/dat$popEst2015, freq = F)#, ylim = c(0, 0.04))
lines(as.integer(min(dat$avgAnnCount/dat$popEst2015)):as.integer(max(dat$avgAnnCount/dat$popEst2015)), d
```



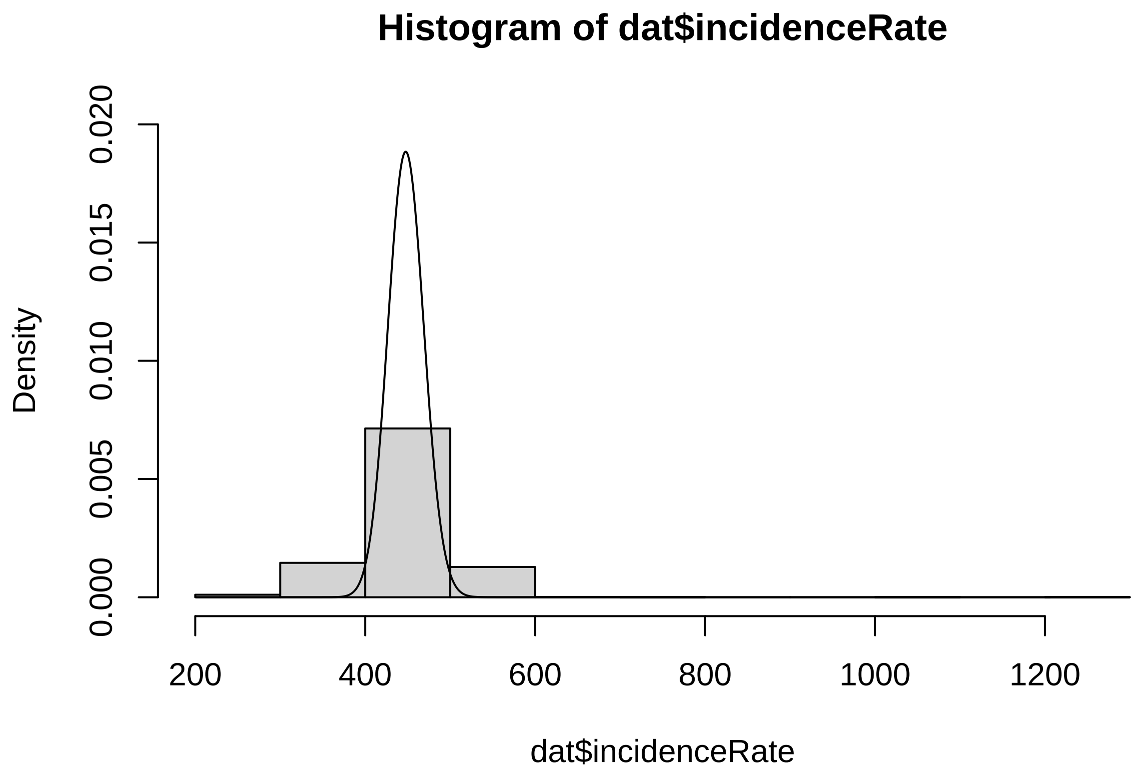

```r
mean(dat$avgDeathsPerYear/dat$popEst2015)
```

```
## [1] 0.002287129
```

```r
var(dat$avgDeathsPerYear/dat$popEst2015)
```

```
## [1] 3.729806e-07
```

```r
hist(dat$incidenceRate, freq = F, ylim = c(0, 0.02))
lines(as.integer(min(dat$incidenceRate)):as.integer(max(dat$incidenceRate)), dpois(as.integer(min(dat$i
```




```
print(mean(dat$incidenceRate))
```

```
## [1] 448.1764
```

```
print(var(dat$incidenceRate))
```

```
## [1] 2982.145
```

Model fits

```
# poisson fit
state_inc_pop_pois <- dat %>% glm(formula = TARGET_deathRate ~ medIncome + State + popEst2015, family=po
summary(state_inc_pop_pois)
```

```
##
## Call:
## glm(formula = TARGET_deathRate ~ medIncome + State + popEst2015,
##     family = poisson(), data = .)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -9.7899  -0.9173   0.0055   0.9062  11.6936
```

```
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)    
## (Intercept)  5.548e+00  1.957e-02 283.437  < 2e-16 ***
## medIncome   -4.502e-06  1.435e-07 -31.378  < 2e-16 ***
## StateAL     -1.111e-01  1.995e-02  -5.571 2.54e-08 ***
## StateAR     -7.953e-02  1.958e-02  -4.062 4.86e-05 ***
## StateAZ     -3.451e-01  2.765e-02 -12.482  < 2e-16 ***
## StateCA     -2.272e-01  2.063e-02 -11.015  < 2e-16 ***
## StateCO     -3.604e-01  2.065e-02 -17.454  < 2e-16 ***
## StateCT     -1.658e-01  3.318e-02  -4.998 5.78e-07 ***
## StateDC     -1.967e-02  5.528e-02  -0.356 0.722031    
## StateDE     -9.943e-02  4.658e-02  -2.134 0.032813 *  
## StateFL     -1.562e-01  1.995e-02  -7.829 4.90e-15 ***
## StateGA     -1.527e-01  1.871e-02  -8.160 3.35e-16 ***
## StateHI     -3.009e-01  4.533e-02  -6.637 3.19e-11 ***
## StateIA     -2.000e-01  1.917e-02 -10.433  < 2e-16 ***
## StateID     -3.036e-01  2.154e-02 -14.092  < 2e-16 ***
## StateIL     -1.067e-01  1.901e-02  -5.612 2.00e-08 ***
## StateIN     -8.772e-02  1.911e-02  -4.589 4.46e-06 ***
## StateKS     -2.028e-01  1.914e-02 -10.591  < 2e-16 ***
## StateKY      6.348e-04  1.881e-02   0.034 0.973082    
## StateLA     -7.386e-02  1.984e-02  -3.723 0.000197 ***
## StateMA     -1.453e-01  2.723e-02  -5.335 9.54e-08 ***
## StateMD     -7.295e-02  2.326e-02  -3.136 0.001710 ** 
## StateME     -1.317e-01  2.552e-02  -5.161 2.45e-07 ***
## StateMI     -1.645e-01  1.948e-02  -8.443  < 2e-16 ***
## StateMN     -2.181e-01  1.941e-02 -11.237  < 2e-16 ***
## StateMO     -1.165e-01  1.896e-02  -6.144 8.04e-10 ***
## StateMS     -7.530e-02  1.948e-02  -3.864 0.000111 ***
## StateMT     -2.559e-01  2.101e-02 -12.182  < 2e-16 ***
## StateNC     -1.783e-01  1.928e-02  -9.248  < 2e-16 ***
## StateND     -2.243e-01  2.068e-02 -10.847  < 2e-16 ***
## StateNE     -2.474e-01  1.968e-02 -12.571  < 2e-16 ***
## StateNH     -1.354e-01  2.982e-02  -4.540 5.63e-06 ***
## StateNJ     -9.329e-02  2.430e-02  -3.839 0.000123 ***
## StateNM     -3.131e-01  2.282e-02 -13.720  < 2e-16 ***
## StateNV     -1.216e-01  2.525e-02  -4.815 1.47e-06 ***
## StateNY     -1.560e-01  2.004e-02  -7.781 7.17e-15 ***
## StateOH     -9.885e-02  1.924e-02  -5.138 2.77e-07 ***
## StateOK     -8.282e-02  1.945e-02  -4.259 2.05e-05 ***
## StateOR     -1.967e-01  2.184e-02  -9.009  < 2e-16 ***
## StatePA     -1.547e-01  1.993e-02  -7.761 8.40e-15 ***
## StateRI     -1.563e-01  4.256e-02  -3.673 0.000239 ***
## StateSC     -1.265e-01  2.073e-02  -6.100 1.06e-09 ***
## StateSD     -2.384e-01  2.035e-02 -11.720  < 2e-16 ***
## StateTN     -6.076e-02  1.916e-02  -3.172 0.001515 ** 
## StateTX     -1.964e-01  1.830e-02 -10.731  < 2e-16 ***
## StateUT     -3.846e-01  2.404e-02 -15.999  < 2e-16 ***
## StateVA     -1.029e-01  1.873e-02  -5.491 3.99e-08 ***
## StateVT     -1.459e-01  2.668e-02  -5.468 4.54e-08 ***
## StateWA     -2.024e-01  2.149e-02  -9.420  < 2e-16 ***
## StateWI     -1.687e-01  1.968e-02  -8.569  < 2e-16 ***
## StateWV     -8.860e-02  2.018e-02  -4.390 1.13e-05 ***
```

```
## StateWY     -2.220e-01  2.404e-02  -9.235  < 2e-16 ***
## popEst2015  -1.060e-08  4.808e-09  -2.205 0.027479 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 13026.8  on 3016  degrees of freedom
## Residual deviance:  7555.3  on 2964  degrees of freedom
## AIC: Inf
##
## Number of Fisher Scoring iterations: 4
```

```
# neg bin fit
state_inc_pop_nb <- dat %>% MASS::glm.nb(formula = TARGET_deathRate ~ medIncome + State + popEst2015, da
summary(state_inc_pop_nb)
```

```
##
## Call:
## MASS::glm.nb(formula = TARGET_deathRate ~ medIncome + State +
##     popEst2015, data = ., init.theta = 118.3508188, link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -6.7153  -0.5830   0.0045   0.5662   6.8103
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  5.542e+00  3.153e-02 175.761  < 2e-16 ***
## medIncome   -4.428e-06  2.224e-07 -19.913  < 2e-16 ***
## StateAL     -1.088e-01  3.236e-02  -3.361 0.000776 ***
## StateAR     -7.718e-02  3.182e-02  -2.426 0.015280 *
## StateAZ     -3.421e-01  4.296e-02  -7.963 1.68e-15 ***
## StateCA     -2.255e-01  3.298e-02  -6.837 8.11e-12 ***
## StateCO     -3.588e-01  3.273e-02 -10.964  < 2e-16 ***
## StateCT     -1.662e-01  5.159e-02  -3.221 0.001277 **
## StateDC     -1.972e-02  8.826e-02  -0.223 0.823198
## StateDE     -9.806e-02  7.408e-02  -1.324 0.185601
## StateFL     -1.549e-01  3.222e-02  -4.808 1.52e-06 ***
## StateGA     -1.501e-01  3.030e-02  -4.954 7.28e-07 ***
## StateHI     -3.001e-01  6.833e-02  -4.392 1.12e-05 ***
## StateIA     -1.985e-01  3.091e-02  -6.422 1.34e-10 ***
## StateID     -3.020e-01  3.423e-02  -8.822  < 2e-16 ***
## StateIL     -1.046e-01  3.078e-02  -3.399 0.000676 ***
## StateIN     -8.608e-02  3.099e-02  -2.778 0.005476 **
## StateKS     -2.014e-01  3.088e-02  -6.521 6.97e-11 ***
## StateKY      2.029e-03  3.063e-02   0.066 0.947177
## StateLA     -7.152e-02  3.223e-02  -2.219 0.026472 *
## StateMA     -1.436e-01  4.300e-02  -3.339 0.000842 ***
## StateMD     -7.193e-02  3.733e-02  -1.927 0.053977 .
## StateME     -1.299e-01  4.107e-02  -3.163 0.001559 **
## StateMI     -1.627e-01  3.147e-02  -5.169 2.36e-07 ***
## StateMN     -2.161e-01  3.120e-02  -6.925 4.35e-12 ***
## StateMO     -1.144e-01  3.073e-02  -3.723 0.000197 ***
```

```
## StateMS     -7.231e-02  3.167e-02  -2.283 0.022432 *
## StateMT     -2.536e-01  3.359e-02  -7.548 4.43e-14 ***
## StateNC     -1.760e-01  3.116e-02  -5.650 1.61e-08 ***
## StateND     -2.228e-01  3.306e-02  -6.739 1.60e-11 ***
## StateNE     -2.452e-01  3.160e-02  -7.760 8.50e-15 ***
## StateNH     -1.336e-01  4.725e-02  -2.828 0.004683 **
## StateNJ     -9.247e-02  3.869e-02  -2.390 0.016846 *
## StateNM     -3.105e-01  3.615e-02  -8.589  < 2e-16 ***
## StateNV     -1.222e-01  4.045e-02  -3.022 0.002511 **
## StateNY     -1.541e-01  3.227e-02  -4.774 1.81e-06 ***
## StateOH     -9.731e-02  3.117e-02  -3.122 0.001799 **
## StateOK     -8.144e-02  3.157e-02  -2.580 0.009875 **
## StateOR     -1.942e-01  3.500e-02  -5.549 2.88e-08 ***
## StatePA     -1.528e-01  3.214e-02  -4.752 2.01e-06 ***
## StateRI     -1.558e-01  6.651e-02  -2.342 0.019186 *
## StateSC     -1.248e-01  3.357e-02  -3.717 0.000202 ***
## StateSD     -2.368e-01  3.261e-02  -7.263 3.78e-13 ***
## StateTN     -5.883e-02  3.114e-02  -1.890 0.058816 .
## StateTX     -1.936e-01  2.963e-02  -6.535 6.36e-11 ***
## StateUT     -3.825e-01  3.729e-02 -10.257  < 2e-16 ***
## StateVA     -1.011e-01  3.034e-02  -3.331 0.000865 ***
## StateVT     -1.434e-01  4.263e-02  -3.364 0.000768 ***
## StateWA     -2.004e-01  3.437e-02  -5.830 5.53e-09 ***
## StateWI     -1.674e-01  3.174e-02  -5.276 1.32e-07 ***
## StateWV     -8.637e-02  3.278e-02  -2.635 0.008423 **
## StateWY     -2.196e-01  3.800e-02  -5.778 7.58e-09 ***
## popEst2015  -1.010e-08  7.324e-09  -1.379 0.168041
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(118.3508) family taken to be 1)
##
##     Null deviance: 5232.5  on 3016  degrees of freedom
## Residual deviance: 3051.8  on 2964  degrees of freedom
## AIC: 27080
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  118.35
##           Std. Err.:  5.12
##
##  2 x log-likelihood:  -26971.87
```