ECON 121 FA23 Problem Set 3

Robert Tso

Question 1

Verbal: list group members.

Robert Tso - A13829791

Tomas Lopez - A16798775

Stephanie Nguyen - A16215540

Akash Juwadi - A16372772

Question 2

Code: Load packages and dataset, generate variables, summarize data.

Verbal: Interpret the summary statistics.

The summary statistics show a seemingly near population pool of survey answers, except there is a slight weight towards female answers, as the population split of male/female in the US is closer to 49/51% as opposed to the 43/57% split in the dataset. The dummy variables we generated for fair and poor health show that 16% of the survey judge themselves in this classification. It is worth nothing that this percent is near and above the range of percentage answers for diabetic(12%) and alcohol use(11%).

```r
# The PDF will show the code you write here but not the output.
# Load packages and dataset, generate variables here.
#install.packages("mfx")
library(mfx)

#install.packages("betareg")
library(betareg)

library(tidyverse)
library(fixest)
library(car)

#load(url("https://github.com/tvogl/econ121/raw/main/data/nhis2010.Rdata"))
load("D:/Documents/Class/Econ 121/econ121/data/nhis2010.Rdata")
#view(nhis2010)

# drop observations with health missing/NA.
nhis2010 <- nhis2010 %>% drop_na(health)

# generate a variable that equals one if fair or poor health, zero otherwise.
table(nhis2010$health)
```

```
##
## Excellent Very Good      Good      Fair      Poor
##      5953      7447      7012      2968       962
```

```r
nhis2010$health_dummy <- ifelse(nhis2010$health %in% c("Fair", "Poor"), 1, 0)

#the sum of Fair and Poor should be the same as 1
table(nhis2010$health_dummy)
```

```
##
##     0     1
## 20412  3930
```

```r
# The PDF will show the code AND output here.
# Summarize the data here.
summary(nhis2010)
```

```
##    sampweight          psu             hhnum            pernum
##   Min.   :  853   Min.   :  1.0   Min.   :    1   Min.   : 1.000
```

```
##  1st Qu.: 4338   1st Qu.:156.0   1st Qu.:10383   1st Qu.: 1.000
##  Median : 6878   Median :306.5   Median :21098   Median : 1.000
##  Mean   : 8213   Mean   :304.8   Mean   :21238   Mean   : 1.371
##  3rd Qu.:10710   3rd Qu.:460.0   3rd Qu.:31969   3rd Qu.: 2.000
##  Max.   :65899   Max.   :600.0   Max.   :43208   Max.   :12.000
##
##       age             male                  marstat          white
##  Min.   :25.00   Min.   :0.0000   Married      :11719   Min.   :0.0000
##  1st Qu.:37.00   1st Qu.:0.0000   Widowed      : 2545   1st Qu.:0.0000
##  Median :49.00   Median :0.0000   Divorced     : 3985   Median :1.0000
##  Mean   :50.78   Mean   :0.4382   Separated    : 1003   Mean   :0.5763
##  3rd Qu.:63.00   3rd Qu.:1.0000   Never married: 5041   3rd Qu.:1.0000
##  Max.   :85.00   Max.   :1.0000   NA's         :   49   Max.   :1.0000
##
##      black             hisp             asian              other
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.00000   Min.   :0.00000
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:0.00000
##  Median :0.0000   Median :0.0000   Median :0.00000   Median :0.00000
##  Mean   :0.1612   Mean   :0.1824   Mean   :0.06253   Mean   :0.01754
##  3rd Qu.:0.0000   3rd Qu.:0.0000   3rd Qu.:0.00000   3rd Qu.:0.00000
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.00000   Max.   :1.00000
##
##      edyrs                                       empstat
##  Min.   : 1.0   Working for pay at job/business    :13244
##  1st Qu.:13.0   Not in labor force                 : 8848
##  Median :14.0   Not employed                       : 1451
##  Mean   :13.8   With job, but not at work          :  563
##  3rd Qu.:16.0   Working, w/out pay, at job/business:  224
##  Max.   :19.0   (Other)                            :    0
##  NA's   :116    NA's                               :   12
##                incfam           health          mort             bmi
##  $0 - $34,999     :9730   Excellent:5953   Min.   :0.0000   Min.   : 9.89
##  $35,000 - $49,999:3468   Very Good:7447   1st Qu.:0.0000   1st Qu.:23.72
##  $50,000 - $74,999:3849   Good     :7012   Median :0.0000   Median :26.69
##  $75,000 - $99,999:2333   Fair     :2968   Mean   :0.1288   Mean   :27.91
##  $100,000 and over:3634   Poor     : 962   3rd Qu.:0.0000   3rd Qu.:30.86
##  NA's             :1328                    Max.   :1.0000   Max.   :87.84
##                                            NA's   :362      NA's   :930
##    uninsured         cancerev         cheartdiev        heartattev
##  Min.   :0.0000   Min.   :0.00000   Min.   :0.00000   Min.   :0.000
##  1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.000
##  Median :0.0000   Median :0.00000   Median :0.00000   Median :0.000
##  Mean   :0.1743   Mean   :0.09473   Mean   :0.05448   Mean   :0.038
##  3rd Qu.:0.0000   3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:0.000
##  Max.   :1.0000   Max.   :1.00000   Max.   :1.00000   Max.   :1.000
##  NA's   :61       NA's   :19        NA's   :57        NA's   :24
##    hypertenev       diabeticev        alc5upyr          smokev
##  Min.   :0.0000   Min.   :0.0000   Min.   :  0.00   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:  0.00   1st Qu.:0.0000
##  Median :0.0000   Median :0.0000   Median :  0.00   Median :0.0000
##  Mean   :0.3571   Mean   :0.1271   Mean   : 10.95   Mean   :0.4202
##  3rd Qu.:1.0000   3rd Qu.:0.0000   3rd Qu.:  2.00   3rd Qu.:1.0000
##  Max.   :1.0000   Max.   :1.0000   Max.   :365.00   Max.   :1.0000
##  NA's   :37       NA's   :15       NA's   :9733     NA's   :176
```

```
##      vig10fwk         hrsleep                           asad
##  Min.   : 0.000   Min.   : 3.000   None of the time    :17373
##  1st Qu.: 0.000   1st Qu.: 6.000   A little of the time: 3426
##  Median : 0.000   Median : 7.000   Some of the time    : 2427
##  Mean   : 1.494   Mean   : 7.158   Most of the time    :  649
##  3rd Qu.: 2.000   3rd Qu.: 8.000   All of the time     :  301
##  Max.   :28.000   Max.   :22.000   NA's                :  166
##  NA's   :307      NA's   :365
##   health_dummy
##  Min.   :0.0000
##  1st Qu.:0.0000
##  Median :0.0000
##  Mean   :0.1614
##  3rd Qu.:0.0000
##  Max.   :1.0000
##
```

Question 3

Code: Draw graph with two line plots.

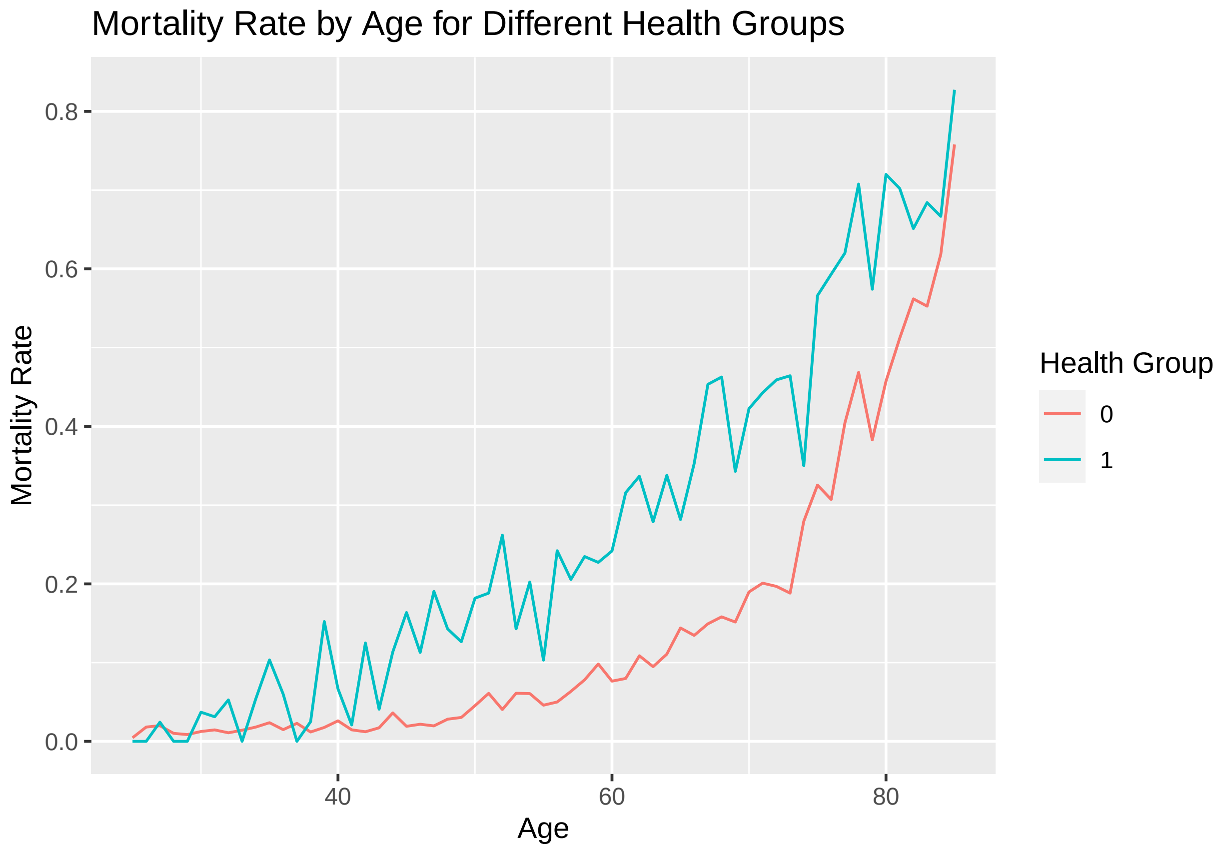
Verbal: Interpret.

Risk of death for both categories go up with age, however the greatest difference of mortality between the self-reported health groups is more pronounced between ages 40 and 80, with a clear observation of lower risk of death among those with self-reported good-to-excellent health. In the beginning and end of the data, both groups have very similar mortality rates.

```
# All question 3 code here.

# Compute mortality rates by age for both groups
mortality_data <- nhis2010 %>%
  drop_na(age, mort)%>%
  group_by(health_dummy, age) %>%
  summarise(mortality_rate = mean(mort))
```

```
## ‘summarise()‘ has grouped output by ’health_dummy’. You can override using the
## ‘.groups‘ argument.
```

```
# Create separate line plots for the two groups
ggplot(mortality_data, aes(x = age, y = mortality_rate, color = factor(health_dummy))) +
  geom_line() +
  labs(
    x = "Age",
    y = "Mortality Rate",
    title = "Mortality Rate by Age for Different Health Groups",
    color = "Health Group"
  )
```


Mortality Rate by Age for Different Health Groups
Mortality Rate
0.8
0.6
0.4
0.2
0.0
40
60
80
Age
Health Group
0
1

Question 4

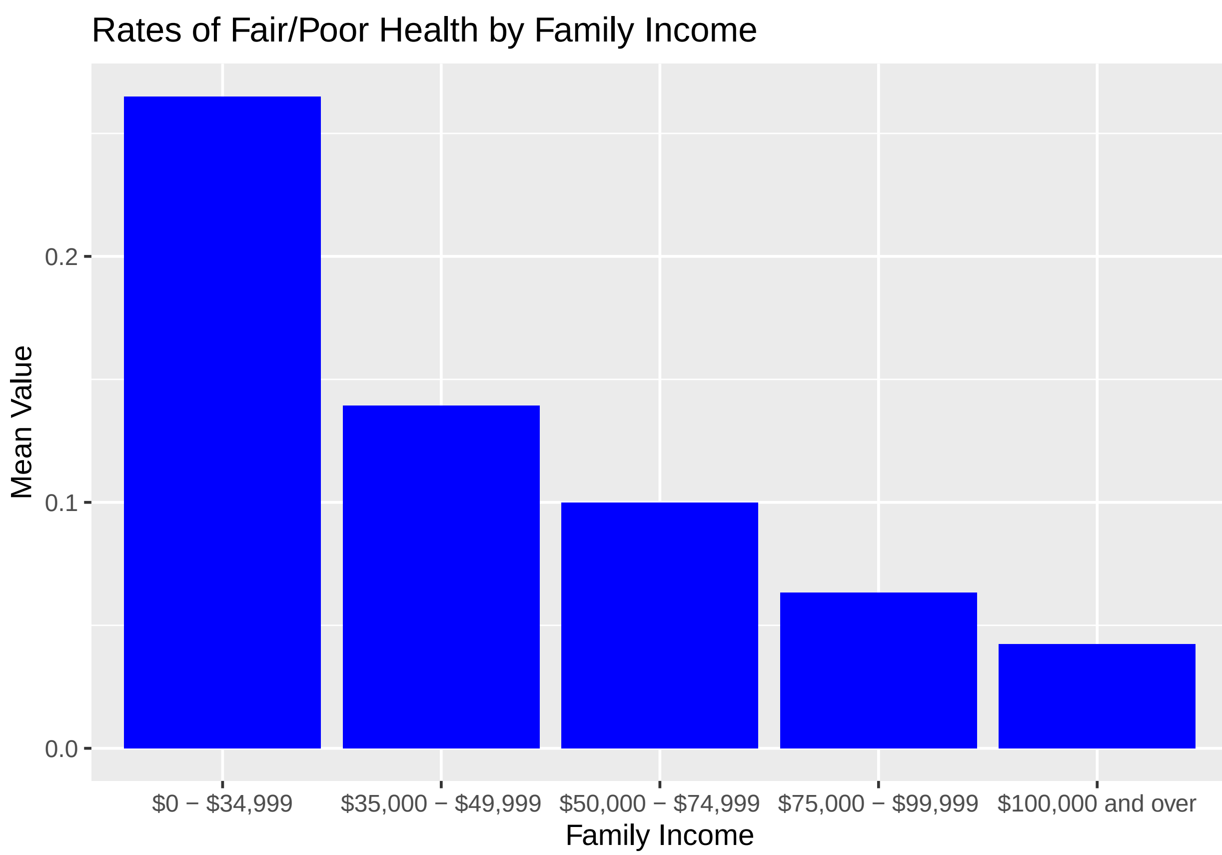
Code: Draw bar graphs.

Verbal: Interpret your results.

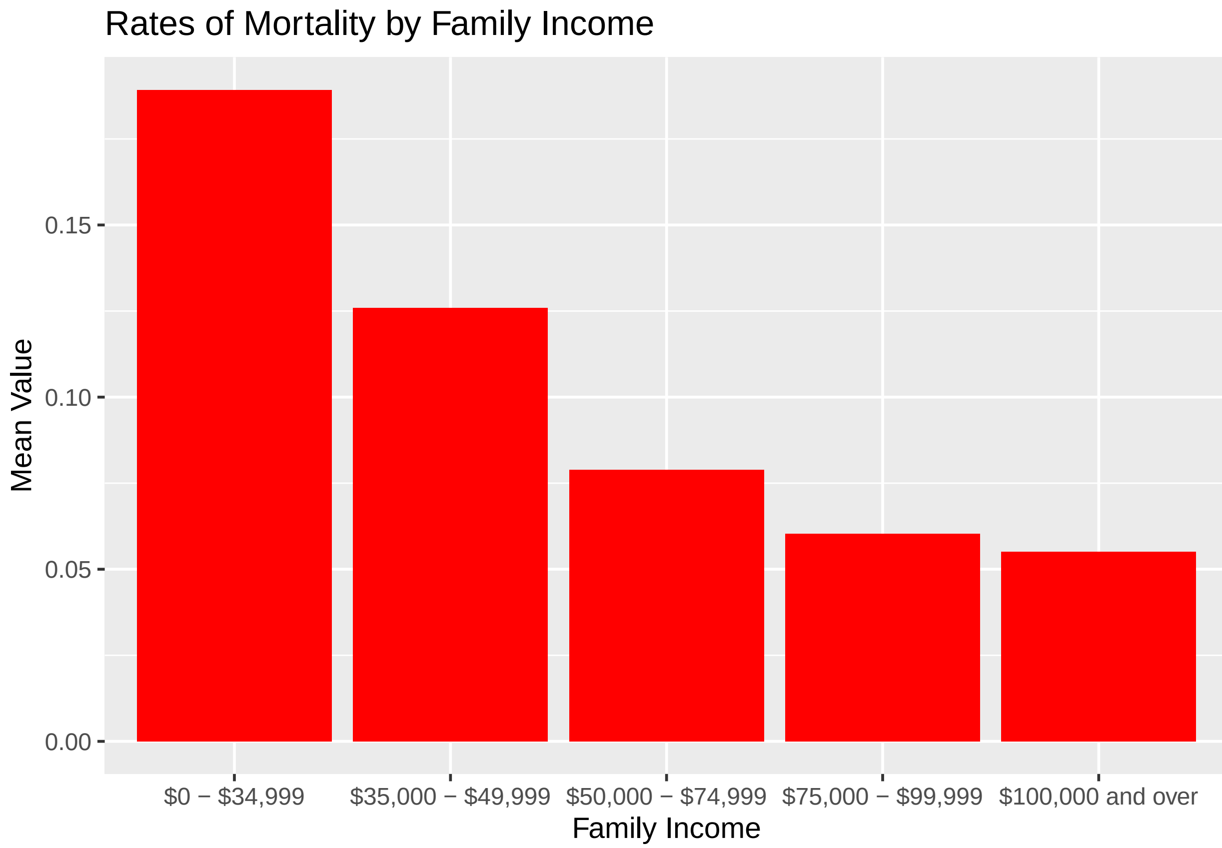
In the case of family income, as income increases, rates of mortality and poor health decreases. Similarly as education level increases, rates of mortality and poor health decreases. We are not certain of the cross interaction of education and income on either mortality or health, but it would not be surprising if the interactive case was true. Looking at the odds ratios, all the incomes higher than $35,000 have lower odds of mortality and better reported health, meanwhile all education lower than college have higher odds. In regards to race differences, Asians and Hispanics have lower mortality rates compared to Whites, and Blacks and Other have a higher rate to claim poor/fair health relative to Whites.

```r
# All question 4 code here

# Create table for fair/poor health and mortality by family income
graph_a <- nhis2010 %>%
  drop_na(incfam,mort) %>%
  group_by(incfam) %>%
  summarise(mean_fair_poor_health = mean(health_dummy),
            mean_mortality = mean(mort))

# Create bar plots for family income
fam_health <- ggplot(graph_a, aes(x = incfam)) +
  geom_bar(aes(y = mean_fair_poor_health), stat = "identity", fill = "blue", position = "dodge") +
  labs(x = "Family Income", y = "Mean Value", title = "Rates of Fair/Poor Health by Family Income")

fam_mort <- ggplot(graph_a, aes(x = incfam)) +
  geom_bar(aes(y = mean_mortality), stat = "identity", fill = "red", position = "dodge") +
  labs(x = "Family Income", y = "Mean Value", title = "Rates of Mortality by Family Income")

# Categorize years of education into five categories
nhis2010 <- nhis2010 %>%
  drop_na(edyrs)%>%
  mutate(education_category = case_when(
    edyrs < 12 ~ "Less than High School",
    edyrs == 12 ~ "High School Completion",
    edyrs >= 13 & edyrs <= 15 ~ "Some College",
    edyrs == 16 ~ "College Completion",
    edyrs > 16 ~ "Post-graduate Study"
  ))

# Create table for fair/poor health and mortality by education category
graph_b <- nhis2010 %>%
  drop_na(mort)%>%
  group_by(education_category) %>%
  summarise(mean_fair_poor_health = mean(health_dummy),
            mean_mortality = mean(mort))

# Create bar plots for education
edu_health <- ggplot(graph_b, aes(x = education_category)) +
  geom_bar(aes(y = mean_fair_poor_health), stat = "identity", fill = "blue", position = "dodge") +
  labs(x = "Education Level", y = "Mean Value", title = "Rates of Fair/Poor Health by Education Level")
  theme(axis.text.x = element_text(angle = 35, hjust = 1))  # Rotate x-axis labels
```

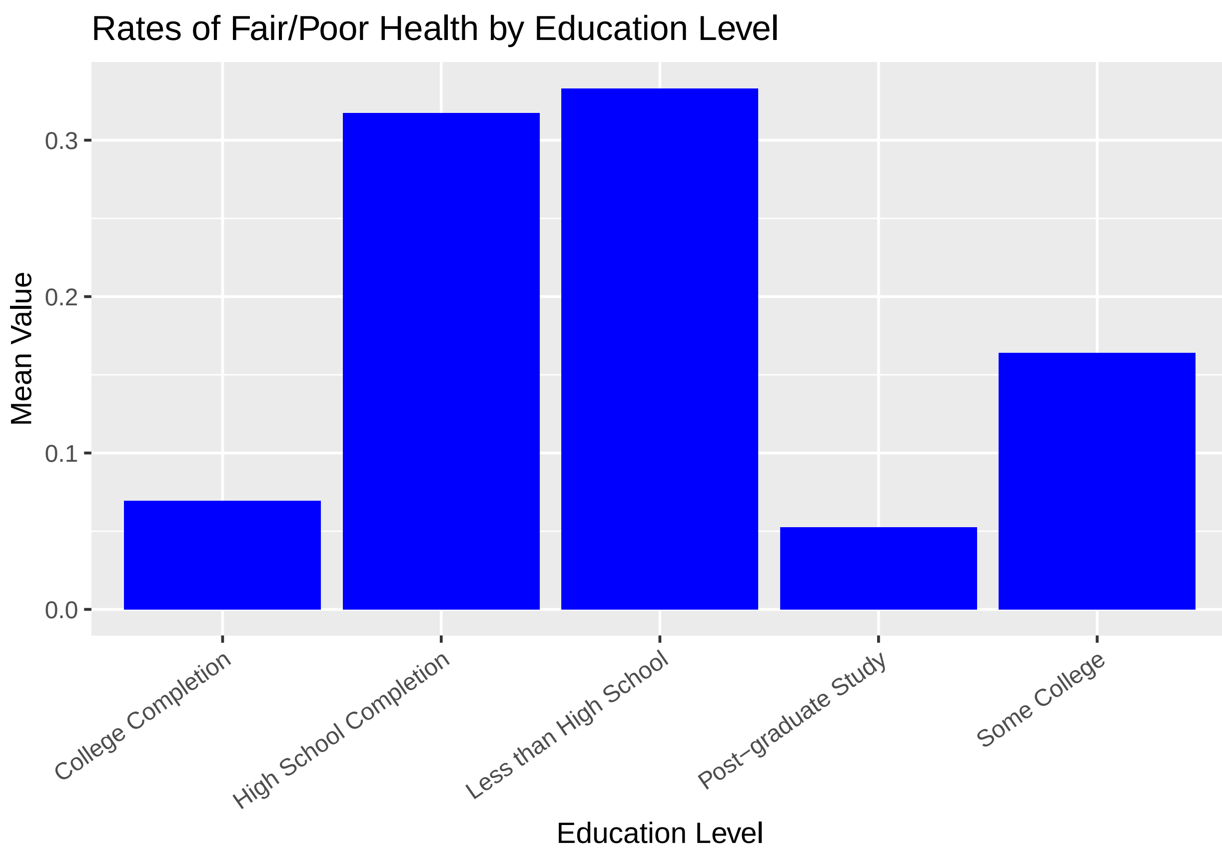
```
edu_mort <- ggplot(graph_b, aes(x = education_category)) +
  geom_bar(aes(y = mean_mortality), stat = "identity", fill = "red", position = "dodge") +
  labs(x = "Education Level", y = "Mean Value", title = "Rates of Mortality by Education Level") +
  theme(axis.text.x = element_text(angle = 35, hjust = 1))  # Rotate x-axis labels


fam_health
```



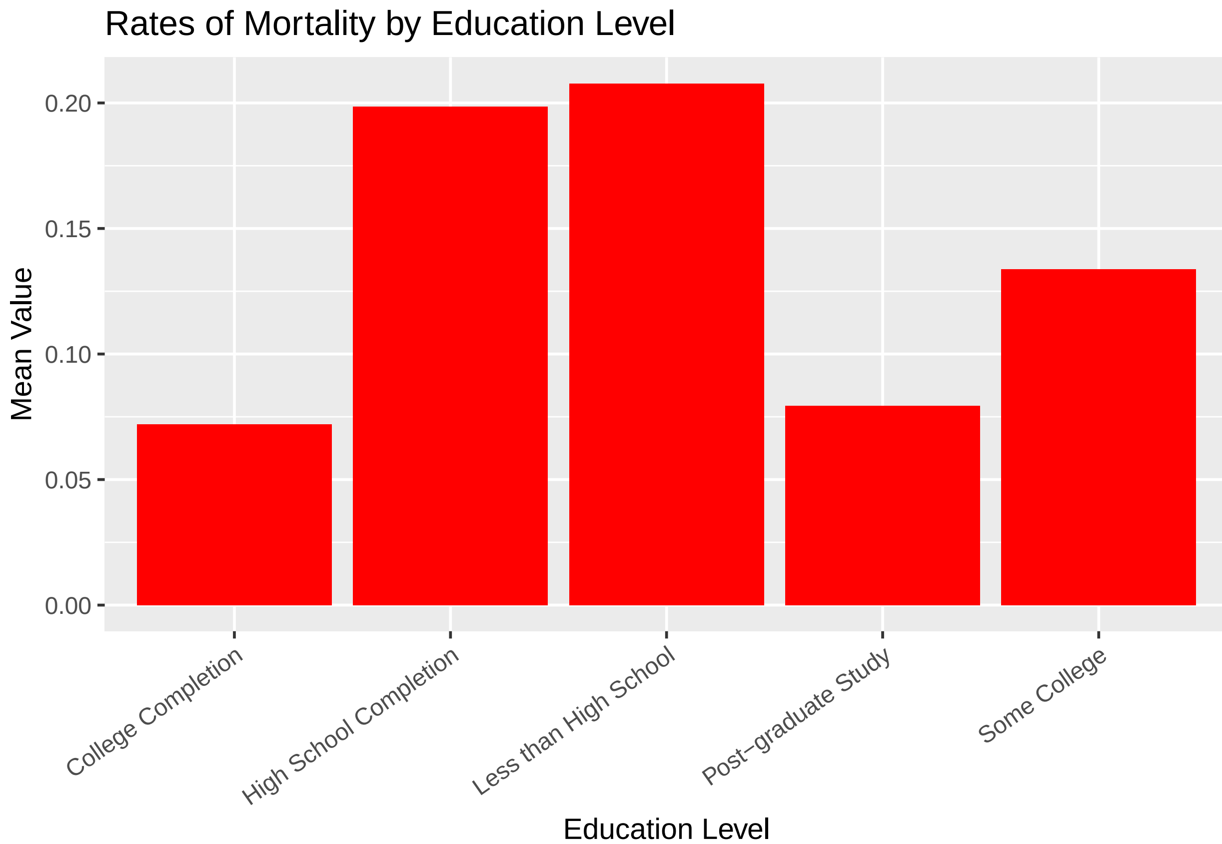

```
fam_mort
```


Rates of Mortality by Family Income
Mean Value
0.15
0.10
0.05
0.00
$0 − $34,999
$35,000 − $49,999
$50,000 − $74,999
$75,000 − $99,999
$100,000 and over
Family Income

```
edu_health
```


Rates of Fair/Poor Health by Education Level
Mean Value
0.3
0.2
0.1
0.0
College Completion
High School Completion
Less than High School
Post−graduate Study
Some College
Education Level

```
edu_mort
```


Rates of Mortality by Education Level
0.20
0.15
0.10
0.05
0.00
Mean Value
College Completion
High School Completion
Less than High School
Post−graduate Study
Some College
Education Level

Question 5

Code: Estimate regressions.

Verbal: Interpret and compare.

Based on the summary statistics OLS does not look like a good predictive model because it has negative values on the lower end of the model. It also has a dampened maximum value at close to 50%, compared to the 66-70%+ of logit and probit.

```r
# All question 5 code here

#dropping empty rows
nhis2010 <- nhis2010 %>%
  drop_na(mort,health_dummy,incfam,age,education_category,black,hisp,asian,other)

#generating dummy variables for education categories
LessHS <- ifelse(nhis2010$education_category == "Less than High School", 1, 0)
HSGrad <- ifelse(nhis2010$education_category == "High School Completion", 1, 0)
SomeCol <- ifelse(nhis2010$education_category == "Some College", 1, 0)
ColGrad <- ifelse(nhis2010$education_category == "College Completion", 1, 0)
PostGrad <- ifelse(nhis2010$education_category == "Post-graduate Study", 1, 0)

#generating dummy variables for income categories
Low <- ifelse(nhis2010$incfam == "$0 - $34,999", 1, 0)
LowMed <- ifelse(nhis2010$incfam == "$35,000 - $49,999", 1, 0)
Med <- ifelse(nhis2010$incfam == "$50,000 - $74,999", 1, 0)
MedHigh <- ifelse(nhis2010$incfam == "$75,000 - $99,999", 1, 0)
High <- ifelse(nhis2010$incfam == "$100,000 and over", 1, 0)


#add all the dummy variables to nhis2010
nhis2010 <- nhis2010 %>%
  mutate(
    LessHS=LessHS,
    HsGrad=HSGrad,
    SomeCol=SomeCol,
    ColGrad=ColGrad,
    PostGrad=PostGrad,
    Low=Low,
    LowMed=LowMed,
    Med=Med,
    MedHigh=MedHigh,
    High=High
  )

#view(nhis2010)

ols_model_pf <- feols(health_dummy ~  age +
                      Low + LowMed + Med + MedHigh + High +
                      LessHS + HsGrad + SomeCol + ColGrad + PostGrad +
                      black + hisp + asian + other,
                      data = nhis2010,
                      vcov = 'hetero')
```

```
## The variables 'High' and 'PostGrad' have been removed because of collinearity (see $collin.var).
```

```
nhis2010$ols_predict_pf <- predict(ols_model_pf, nhis2010, type="response")


ols_model_mort <- feols(mort ~  age +
                        Low + LowMed + Med + MedHigh + High +
                        LessHS + HsGrad + SomeCol + ColGrad + PostGrad +
                        black + hisp + asian + other,
                        data = nhis2010,
                        vcov = 'hetero')
```

```
## The variables 'High' and 'PostGrad' have been removed because of collinearity (see $collin.var).
```

```
nhis2010$ols_predict_mort <- predict(ols_model_mort, nhis2010, type="response")

probit_model_pf <- feglm(health_dummy ~  age +
                        Low + LowMed + Med + MedHigh + High +
                        LessHS + HsGrad + SomeCol + ColGrad + PostGrad +
                        black + hisp + asian + other,
                        data = nhis2010,
                        vcov = 'hetero',
                        family = 'probit')
```

```
## The variables 'High' and 'PostGrad' have been removed because of collinearity (see $collin.var).
```

```
nhis2010$probit_predict_pf <- predict(probit_model_pf, nhis2010, type="response")

probit_model_mort <- feglm(mort ~  age +
                          Low + LowMed + Med + MedHigh + High +
                          LessHS + HsGrad + SomeCol + ColGrad + PostGrad +
                          black + hisp + asian + other,
                          data = nhis2010,
                          vcov = 'hetero',
                          family = 'probit')
```

```
## The variables 'High' and 'PostGrad' have been removed because of collinearity (see $collin.var).
```

```
nhis2010$probit_predict_mort <- predict(probit_model_mort, nhis2010, type="response")


logit_model_pf <- feglm(health_dummy ~  age +
                        Low + LowMed + Med + MedHigh + High +
                        LessHS + HsGrad + SomeCol + ColGrad + PostGrad +
                        black + hisp + asian + other,
                        data = nhis2010,
                        vcov = 'hetero',
                        family = 'logit')
```

```
## The variables 'High' and 'PostGrad' have been removed because of collinearity (see $collin.var).
```

```r
nhis2010$logit_predict_pf <- predict(logit_model_pf, nhis2010, type="response")


logit_model_mort <- feglm(mort ~  age +
                          Low + LowMed + Med + MedHigh + High +
                          LessHS + HsGrad + SomeCol + ColGrad + PostGrad +
                          black + hisp + asian + other,
                          data = nhis2010,
                          vcov = 'hetero',
                          family = 'logit')
```

```
## The variables 'High' and 'PostGrad' have been removed because of collinearity (see $collin.var).
```

```r
nhis2010$logit_predict_mort <- predict(logit_model_mort, nhis2010, type="response")

# ols_model_pf
# probit_model_pf
# logit_model_pf

#summary statistics of health
nhis2010 %>%
  select(ols_predict_pf,logit_predict_pf,probit_predict_pf) %>%
  summary()
```

```
##  ols_predict_pf     logit_predict_pf  probit_predict_pf
##  Min.   :-0.07673   Min.   :0.01348   Min.   :0.008328
##  1st Qu.: 0.06437   1st Qu.:0.06107   1st Qu.:0.058816
##  Median : 0.14761   Median :0.12551   Median :0.127067
##  Mean   : 0.16249   Mean   :0.16249   Mean   :0.162221
##  3rd Qu.: 0.25054   3rd Qu.:0.22965   3rd Qu.:0.234210
##  Max.   : 0.52742   Max.   :0.66867   Max.   :0.643540
```

```r
#summary statistics of mortality
nhis2010 %>%
  select(ols_predict_mort,logit_predict_mort,probit_predict_mort) %>%
  summary()
```

```
##  ols_predict_mort     logit_predict_mort probit_predict_mort
##  Min.   :-0.1869300   Min.   :0.001979   Min.   :0.0004103
##  1st Qu.: 0.0001547   1st Qu.:0.015015   1st Qu.:0.0122731
##  Median : 0.1076756   Median :0.046755   Median :0.0507455
##  Mean   : 0.1266779   Mean   :0.126678   Mean   :0.1273415
##  3rd Qu.: 0.2352395   3rd Qu.:0.156094   3rd Qu.:0.1756179
##  Max.   : 0.5048590   Max.   :0.744464   Max.   :0.6998198
```

```r
#marginal effect of IVs for poor and fair health using logit
logitmfx(health_dummy ~  incfam + age + education_category + black + hisp + asian + other,
                   data = nhis2010,
                  atmean = TRUE,
                  robust = TRUE)
```

```
## Call:
## logitmfx(formula = health_dummy ~ incfam + age + education_category +
##     black + hisp + asian + other, data = nhis2010, atmean = TRUE,
##     robust = TRUE)
##
## Marginal Effects:
##                                              dF/dx   Std. Err.        z
## incfam$35,000 - $49,999                 -0.05550324  0.00452754 -12.2590
## incfam$50,000 - $74,999                 -0.07262887  0.00445077 -16.3183
## incfam$75,000 - $99,999                 -0.09120304  0.00472132 -19.3173
## incfam$100,000 and over                 -0.11135986  0.00461556 -24.1270
## age                                      0.00297078  0.00012975  22.8956
## education_categoryHigh School Completion  0.15855836  0.02256822   7.0257
## education_categoryLess than High School   0.15895215  0.01505960  10.5549
## education_categoryPost-graduate Study    -0.02375731  0.01059333  -2.2427
## education_categorySome College            0.05301995  0.00722165   7.3418
## black                                    0.05668476  0.00687244   8.2481
## hisp                                     0.00179775  0.00640647   0.2806
## asian                                    0.00650231  0.01050546   0.6189
## other                                    0.07356644  0.02043691   3.5997
##                                            P>|z|
## incfam$35,000 - $49,999                 < 2.2e-16 ***
## incfam$50,000 - $74,999                 < 2.2e-16 ***
## incfam$75,000 - $99,999                 < 2.2e-16 ***
## incfam$100,000 and over                 < 2.2e-16 ***
## age                                     < 2.2e-16 ***
## education_categoryHigh School Completion 2.129e-12 ***
## education_categoryLess than High School  < 2.2e-16 ***
## education_categoryPost-graduate Study    0.0249183 *
## education_categorySome College           2.107e-13 ***
## black                                   < 2.2e-16 ***
## hisp                                    0.7790061
## asian                                   0.5359523
## other                                   0.0003186 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## dF/dx is for discrete change for the following variables:
##
##  [1] "incfam$35,000 - $49,999"
##  [2] "incfam$50,000 - $74,999"
##  [3] "incfam$75,000 - $99,999"
##  [4] "incfam$100,000 and over"
##  [5] "education_categoryHigh School Completion"
##  [6] "education_categoryLess than High School"
##  [7] "education_categoryPost-graduate Study"
##  [8] "education_categorySome College"
##  [9] "black"
## [10] "hisp"
## [11] "asian"
## [12] "other"
```

```
#marginal effect of IVs for poor and fair health using probit
probitmfx(health_dummy ~  incfam + age + education_category + black + hisp + asian + other,
```

```r
                   data = nhis2010,
                  atmean = TRUE,
                  robust = TRUE)
```

```
## Call:
## probitmfx(formula = health_dummy ~ incfam + age + education_category +
##     black + hisp + asian + other, data = nhis2010, atmean = TRUE,
##     robust = TRUE)
##
## Marginal Effects:
##                                              dF/dx   Std. Err.        z
## incfam$35,000 - $49,999                -0.06240593  0.00496601 -12.5666
## incfam$50,000 - $74,999                -0.08061979  0.00474515 -16.9899
## incfam$75,000 - $99,999                -0.09842525  0.00489991 -20.0872
## incfam$100,000 and over                -0.11828044  0.00463461 -25.5211
## age                                     0.00336596  0.00013924  24.1734
## education_categoryHigh School Completion  0.16257580  0.02176088   7.4710
## education_categoryLess than High School  0.16156441  0.01395017  11.5815
## education_categoryPost-graduate Study  -0.02276621  0.01036076  -2.1973
## education_categorySome College          0.05466151  0.00715288   7.6419
## black                                   0.06002296  0.00719825   8.3386
## hisp                                    0.00251821  0.00691967   0.3639
## asian                                   0.00699541  0.01087395   0.6433
## other                                   0.07474414  0.02067813   3.6146
##                                            P>|z|
## incfam$35,000 - $49,999                < 2.2e-16 ***
## incfam$50,000 - $74,999                < 2.2e-16 ***
## incfam$75,000 - $99,999                < 2.2e-16 ***
## incfam$100,000 and over                < 2.2e-16 ***
## age                                    < 2.2e-16 ***
## education_categoryHigh School Completion 7.958e-14 ***
## education_categoryLess than High School  < 2.2e-16 ***
## education_categoryPost-graduate Study    0.0279955 *
## education_categorySome College         2.141e-14 ***
## black                                  < 2.2e-16 ***
## hisp                                   0.7159177
## asian                                  0.5200178
## other                                  0.0003008 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## dF/dx is for discrete change for the following variables:
##
##  [1] "incfam$35,000 - $49,999"
##  [2] "incfam$50,000 - $74,999"
##  [3] "incfam$75,000 - $99,999"
##  [4] "incfam$100,000 and over"
##  [5] "education_categoryHigh School Completion"
##  [6] "education_categoryLess than High School"
##  [7] "education_categoryPost-graduate Study"
##  [8] "education_categorySome College"
##  [9] "black"
## [10] "hisp"
```

```
## [11] "asian"
## [12] "other"
```

```r
#marginal effect of IVs for mortality using logit
logitmfx(mort ~  incfam + age + education_category + black + hisp + asian + other,
                  data = nhis2010,
                 atmean = TRUE,
                 robust = TRUE)
```

```
## Call:
## logitmfx(formula = mort ~ incfam + age + education_category +
##     black + hisp + asian + other, data = nhis2010, atmean = TRUE,
##     robust = TRUE)
##
## Marginal Effects:
##                                               dF/dx   Std. Err.        z
## incfam$35,000 - $49,999                 -0.01562624  0.00289180  -5.4036
## incfam$50,000 - $74,999                 -0.02856166  0.00283943 -10.0589
## incfam$75,000 - $99,999                 -0.03003861  0.00327570  -9.1701
## incfam$100,000 and over                 -0.03209107  0.00321894  -9.9695
## age                                      0.00475195  0.00011079  42.8924
## education_categoryHigh School Completion  0.03228714  0.01116808   2.8910
## education_categoryLess than High School   0.02532388  0.00666249   3.8010
## education_categoryPost-graduate Study     0.00292105  0.00581803   0.5021
## education_categorySome College            0.01379211  0.00391163   3.5259
## black                                   -0.00017489  0.00345400  -0.0506
## hisp                                    -0.02801763  0.00318520  -8.7962
## asian                                   -0.02458421  0.00411724  -5.9710
## other                                    0.00336710  0.01038915   0.3241
##                                             P>|z|
## incfam$35,000 - $49,999                 6.530e-08 ***
## incfam$50,000 - $74,999                 < 2.2e-16 ***
## incfam$75,000 - $99,999                 < 2.2e-16 ***
## incfam$100,000 and over                 < 2.2e-16 ***
## age                                     < 2.2e-16 ***
## education_categoryHigh School Completion 0.0038399 **
## education_categoryLess than High School  0.0001441 ***
## education_categoryPost-graduate Study    0.6156197
## education_categorySome College           0.0004220 ***
## black                                   0.9596163
## hisp                                    < 2.2e-16 ***
## asian                                   2.358e-09 ***
## other                                   0.7458643
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## dF/dx is for discrete change for the following variables:
##
##  [1] "incfam$35,000 - $49,999"
##  [2] "incfam$50,000 - $74,999"
##  [3] "incfam$75,000 - $99,999"
##  [4] "incfam$100,000 and over"
##  [5] "education_categoryHigh School Completion"
##  [6] "education_categoryLess than High School"
```

```
##  [7] "education_categoryPost-graduate Study"
##  [8] "education_categorySome College"
##  [9] "black"
## [10] "hisp"
## [11] "asian"
## [12] "other"
```

```r
#marginal effect of IVs for poor and fair healt usin probit
 probitmfx(mort ~  incfam + age + education_category + black + hisp + asian + other,
                  data = nhis2010,
                 atmean = TRUE,
                 robust = TRUE)
```

```
## Call:
## probitmfx(formula = mort ~ incfam + age + education_category +
##     black + hisp + asian + other, data = nhis2010, atmean = TRUE,
##     robust = TRUE)
##
## Marginal Effects:
##                                                 dF/dx   Std. Err.        z
## incfam$35,000 - $49,999                   -0.02205377  0.00359167  -6.1403
## incfam$50,000 - $74,999                   -0.03760271  0.00342030 -10.9940
## incfam$75,000 - $99,999                   -0.03864816  0.00384539 -10.0505
## incfam$100,000 and over                   -0.04287636  0.00373740 -11.4722
## age                                        0.00572909  0.00012184  47.0215
## education_categoryHigh School Completion   0.03983485  0.01342397   2.9674
## education_categoryLess than High School    0.03213077  0.00815816   3.9385
## education_categoryPost-graduate Study      0.00271792  0.00713541   0.3809
## education_categorySome College             0.01647089  0.00481464   3.4210
## black                                     -0.00074839  0.00430289  -0.1739
## hisp                                      -0.03473070  0.00393068  -8.8358
## asian                                     -0.02758041  0.00551634  -4.9998
## other                                      0.00322048  0.01270953   0.2534
##                                               P>|z|
## incfam$35,000 - $49,999                   8.239e-10 ***
## incfam$50,000 - $74,999                   < 2.2e-16 ***
## incfam$75,000 - $99,999                   < 2.2e-16 ***
## incfam$100,000 and over                   < 2.2e-16 ***
## age                                       < 2.2e-16 ***
## education_categoryHigh School Completion  0.0030029 **
## education_categoryLess than High School   8.200e-05 ***
## education_categoryPost-graduate Study     0.7032736
## education_categorySome College            0.0006239 ***
## black                                     0.8619225
## hisp                                      < 2.2e-16 ***
## asian                                     5.740e-07 ***
## other                                     0.7999659
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## dF/dx is for discrete change for the following variables:
##
##  [1] "incfam$35,000 - $49,999"
##  [2] "incfam$50,000 - $74,999"
```

```
##  [3] "incfam$75,000 - $99,999"
##  [4] "incfam$100,000 and over"
##  [5] "education_categoryHigh School Completion"
##  [6] "education_categoryLess than High School"
##  [7] "education_categoryPost-graduate Study"
##  [8] "education_categorySome College"
##  [9] "black"
## [10] "hisp"
## [11] "asian"
## [12] "other"
```

```
#odds ratio of logit mortality
logitor(mort ~incfam+ age + education_category + black + hisp + asian + other,
        data = nhis2010,
        robust = TRUE)
```

```
## Call:
## logitor(formula = mort ~ incfam + age + education_category +
##     black + hisp + asian + other, data = nhis2010, robust = TRUE)
##
## Odds Ratio:
##                                          OddsRatio Std. Err.       z     P>|z|
## incfam$35,000 - $49,999                  0.7206713 0.0474596 -4.9742 6.553e-07
## incfam$50,000 - $74,999                  0.5220870 0.0388154 -8.7418 < 2.2e-16
## incfam$75,000 - $99,999                  0.4824160 0.0481695 -7.3004 2.869e-13
## incfam$100,000 and over                  0.4697881 0.0419379 -8.4628 < 2.2e-16
## age                                      1.0944066 0.0020794 47.4784 < 2.2e-16
## education_categoryHigh School Completion 1.6427473 0.2323163  3.5099 0.0004483
## education_categoryLess than High School  1.5177021 0.1451930  4.3610 1.295e-05
## education_categoryPost-graduate Study    1.0559042 0.1122582  0.5117 0.6088857
## education_categorySome College           1.3039927 0.0993173  3.4850 0.0004922
## black                                    0.9966820 0.0654827 -0.0506 0.9596555
## hisp                                     0.5337707 0.0443977 -7.5476 4.434e-14
## asian                                    0.5594329 0.0692336 -4.6933 2.688e-06
## other                                    1.0642138 0.1990525  0.3327 0.7393304
##
## incfam$35,000 - $49,999                  ***
## incfam$50,000 - $74,999                  ***
## incfam$75,000 - $99,999                  ***
## incfam$100,000 and over                  ***
## age                                      ***
## education_categoryHigh School Completion ***
## education_categoryLess than High School  ***
## education_categoryPost-graduate Study
## education_categorySome College           ***
## black
## hisp                                     ***
## asian                                    ***
## other
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
#odds ratio of logit poor or fair health
logitor(health_dummy ~incfam+ age + education_category + black + hisp + asian + other,
        data = nhis2010,
        robust = TRUE)
```

```
## Call:
## logitor(formula = health_dummy ~ incfam + age + education_category +
##     black + hisp + asian + other, data = nhis2010, robust = TRUE)
##
## Odds Ratio:
##                                           OddsRatio Std. Err.        z     P>|z|
## incfam$35,000 - $49,999                   0.5496799 0.0313801 -10.4824 < 2.2e-16
## incfam$50,000 - $74,999                   0.4412566 0.0271175 -13.3126 < 2.2e-16
## incfam$75,000 - $99,999                   0.3037446 0.0279438 -12.9521 < 2.2e-16
## incfam$100,000 and over                   0.2316135 0.0212599 -15.9350 < 2.2e-16
## age                                       1.0278032 0.0011901  23.6844 < 2.2e-16
## education_categoryHigh School Completion 2.8263803 0.3223966   9.1087 < 2.2e-16
## education_categoryLess than High School  2.9857002 0.2467247  13.2369 < 2.2e-16
## education_categoryPost-graduate Study    0.7903508 0.0893542  -2.0811   0.03743
## education_categorySome College           1.6486892 0.1149386   7.1718 7.404e-13
## black                                     1.5922317 0.0806761   9.1800 < 2.2e-16
## hisp                                      1.0166677 0.0596534   0.2817   0.77815
## asian                                     1.0606386 0.0989633   0.6310   0.52807
## other                                     1.7468168 0.2280598   4.2724 1.934e-05
##
## incfam$35,000 - $49,999                   ***
## incfam$50,000 - $74,999                   ***
## incfam$75,000 - $99,999                   ***
## incfam$100,000 and over                   ***
## age                                       ***
## education_categoryHigh School Completion ***
## education_categoryLess than High School  ***
## education_categoryPost-graduate Study    *
## education_categorySome College           ***
## black                                     ***
## hisp
## asian
## other                                     ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Question 6

Code: Use the results from the mortality logit model to compare the two groups.

Verbal: Interpret your results.

Given these scenarios, Group A has a greater mortality rate, which makes sense since education and income are stronger predictors of mortality than race. We should include interaction terms because Asian adults with a low education and low income and Black adults with college graduate education and over $100k incomes since both collectively represent less than 1% of the population of the data and the basic logit models would not be good predictors of such small data subsets.

```
# All question 6 code here

#looking at the percent of people in the survey that are of Group A or B
nhis2010 %>%
  filter(asian*LessHS*Low==1) %>%
  count()/count(nhis2010)
```

```
##             n
## 1 0.003753091
```

```
#0.375% are Group A

nhis2010 %>%
  filter(black*ColGrad*High==1) %>%
  count()/count(nhis2010)
```

```
##             n
## 1 0.002472625
```

```
#0.247% are Group B


logit_model_mort
```

```
## GLM estimation, family = binomial(link = "logit"), Dep. Var.: mort
## Observations: 22,648
## Standard-errors: Heteroskedasticity-robust
##              Estimate Std. Error    t value   Pr(>|t|)
## (Intercept) -7.796191   0.152869 -50.999148  < 2.2e-16 ***
## age          0.090212   0.001901  47.464295  < 2.2e-16 ***
## Low          0.755474   0.089297   8.460258  < 2.2e-16 ***
## LowMed       0.427901   0.098123   4.360880 1.2954e-05 ***
## Med          0.105552   0.101514   1.039781 2.9844e-01
## MedHigh      0.026525   0.118888   0.223110 8.2345e-01
## LessHS       0.362800   0.108953   3.329872 8.6886e-04 ***
## HsGrad       0.441973   0.150705   2.932703 3.3603e-03 **
## SomeCol      0.211033   0.091466   2.307226 2.1042e-02 *
## ColGrad     -0.054397   0.106347  -0.511510 6.0899e-01
## black       -0.003324   0.065721  -0.050571 9.5967e-01
## hisp        -0.627789   0.083202  -7.545322 4.5117e-14 ***
## asian       -0.580832   0.123794  -4.691934 2.7063e-06 ***
```

```
## other        0.062236   0.187099   0.332639 7.3941e-01
## ... 2 variables were removed because of collinearity (High and PostGrad)
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## Log-Likelihood: -6,079.4   Adj. Pseudo R2: 0.29214
##            BIC: 12,299.1     Squared Cor.: 0.280405
```

```r
deltaMethod(logit_model_mort,"(asian + LessHS + Low) - (black + ColGrad)",rhs=0)
```

```
##                                             Estimate      SE   2.5 %  97.5 %
## (asian + LessHS + Low) - (black + ColGrad)  0.59516 0.18363 0.23524 0.95508
##                                             Hypothesis z value Pr(>|z|)
## (asian + LessHS + Low) - (black + ColGrad)    0.00000   3.241 0.001191 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
#manually
#Group A is an asian adult with less than 12 years or education and family income less the 35k
GroupA <- (-0.580831612  #asian coef
          +0.755474   #less than $35k
          +0.362800)     #less than 12 years edu

# Group B: Black adults with 16 years of education and family incomes over $100k
GroupB <- (-0.003324     #black coef
          +0.211033 # 16 years of edu
          +0) #family income over 100k, 0 due to collinearity

GroupA-GroupB
```

```
## [1] 0.3297334
```

Question 7

Verbal: Assess causality.

No, there may be an omitted variable that predicts both income and health, such as motivation to work harder in health and in career. This motivation variable may be impacted by socioeconomic or purely random and normally distributed through the population.

Question 8

Code: Assess how much health behavior can explain the mortality logit results.

Verbal: Interpret your results.

When measured separately, being Uninsured has a coefficient of -0.074457 while smoking has a coefficient of 0.544864. Measured together uninsured is -0.101424 and smoking is 0.544725, showing that smoking has a relatively large increase in mortality, while being uninsured has a lower, but negative effect. Using the odds ratio we see that smoking has 1.7241335, which means smoking has a 72% higher likelihood of mortality versus those that do not smoke. Similarly being uninsured has a 10% higher likelihood of mortality.

```
# All question 8 code here

#model for uninsured on mortality
logit_model_uninsur <- feglm(mort ~  age + uninsured +
                            Low + LowMed + Med + MedHigh + High +
                            LessHS + HsGrad + SomeCol + ColGrad + PostGrad +
                            black + hisp + asian + other,
                            data = nhis2010,
                            vcov = 'hetero',
                            family = 'logit')
```

```
## NOTE: 35 observations removed because of NA values (RHS: 35).
```

```
## The variables 'High' and 'PostGrad' have been removed because of collinearity (see $collin.var).
```

```
#odds ratio of logit uninsured
logitor(logit_model_uninsur,
        data = nhis2010,
        robust = TRUE)
```

```
## Call:
## logitor(formula = logit_model_uninsur, data = nhis2010, robust = TRUE)
##
## Odds Ratio:
##           OddsRatio Std. Err.       z     P>|z|
## age       1.0936670 0.0022023 44.4648 < 2.2e-16 ***
## uninsured 0.9282473 0.0813687 -0.8494 0.3956585
## Low       2.1679875 0.1961122  8.5542 < 2.2e-16 ***
## LowMed    1.5617150 0.1540008  4.5207 6.164e-06 ***
## Med       1.1290013 0.1147953  1.1933 0.2327508
## MedHigh   1.0386032 0.1235373  0.3184 0.7501531
## LessHS    1.4430962 0.1572128  3.3669 0.0007603 ***
## HsGrad    1.5608641 0.2354746  2.9513 0.0031643 **
## SomeCol   1.2298618 0.1124157  2.2636 0.0236007 *
## ColGrad   0.9407892 0.0999868 -0.5743 0.5657665
## black     0.9927381 0.0652777 -0.1108 0.9117425
## hisp      0.5279722 0.0442075 -7.6282 2.381e-14 ***
## asian     0.5553476 0.0690857 -4.7279 2.268e-06 ***
## other     1.0635173 0.1988562  0.3293 0.7418918
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
logit_model_uninsur
```

```
## GLM estimation, family = binomial(link = "logit"), Dep. Var.: mort
## Observations: 22,613
## Standard-errors: Heteroskedasticity-robust
##              Estimate Std. Error    t value   Pr(>|t|)
## (Intercept) -7.758016   0.158782 -48.859506  < 2.2e-16 ***
## age          0.089536   0.002014  44.450634  < 2.2e-16 ***
## uninsured   -0.074457   0.087687  -0.849126 3.9581e-01
## Low          0.773799   0.090488   8.551449  < 2.2e-16 ***
## LowMed       0.445785   0.098642   4.519214 6.2070e-06 ***
## Med          0.121333   0.101712   1.192916 2.3290e-01
## MedHigh      0.037877   0.118984   0.318335 7.5023e-01
## LessHS       0.366791   0.108977   3.365772 7.6330e-04 ***
## HsGrad       0.445240   0.150911   2.950348 3.1742e-03 **
## SomeCol      0.206902   0.091435   2.262832 2.3646e-02 *
## ColGrad     -0.061036   0.106314  -0.574112 5.6589e-01
## black       -0.007288   0.065777  -0.110805 9.1177e-01
## hisp        -0.638712   0.083758  -7.625693 2.4273e-14 ***
## asian       -0.588161   0.124441  -4.726438 2.2849e-06 ***
## other        0.061582   0.187041   0.329242 7.4197e-01
## ... 2 variables were removed because of collinearity (High and PostGrad)
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## Log-Likelihood: -6,066.1   Adj. Pseudo R2: 0.292063
##            BIC: 12,282.6     Squared Cor.: 0.280249
```

```r
#model for smoking on mortality
logit_model_smoke <- feglm(mort ~  age + smokev +
                            Low + LowMed + Med + MedHigh + High +
                            LessHS + HsGrad + SomeCol + ColGrad + PostGrad +
                            black + hisp + asian + other,
                            data = nhis2010,
                            vcov = 'hetero',
                            family = 'logit')
```

```
## NOTE: 139 observations removed because of NA values (RHS: 139).
## The variables 'High' and 'PostGrad' have been removed because of collinearity (see $collin.var).
```

```r
#odds ratio of logit drinking
logitor(logit_model_smoke,
        data = nhis2010,
        robust = TRUE)
```

```
## Call:
## logitor(formula = logit_model_smoke, data = nhis2010, robust = TRUE)
##
## Odds Ratio:
##         OddsRatio Std. Err.       z     P>|z|
## age     1.0963597 0.0021588 46.7213 < 2.2e-16 ***
## smokev  1.7243731 0.0823702 11.4064 < 2.2e-16 ***
## Low     2.0545309 0.1842930  8.0272 9.971e-16 ***
```

```
## LowMed  1.4999680 0.1484319  4.0972 4.182e-05 ***
## Med     1.0705600 0.1094655  0.6668 0.5048930
## MedHigh 0.9981641 0.1196897 -0.0153 0.9877728
## LessHS  1.3677785 0.1501472  2.8530 0.0043307 **
## HsGrad  1.4444386 0.2173248  2.4440 0.0145239 *
## SomeCol 1.1863284 0.1102410  1.8387 0.0659597 .
## ColGrad 0.9464554 0.1023455 -0.5089 0.6108145
## black   1.0563424 0.0700536  0.8265 0.4085093
## hisp    0.5969013 0.0500301 -6.1564 7.444e-10 ***
## asian   0.6344022 0.0797354 -3.6207 0.0002938 ***
## other   1.0088626 0.1933132  0.0460 0.9632716
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
logit_model_smoke
```

```
## GLM estimation, family = binomial(link = "logit"), Dep. Var.: mort
## Observations: 22,509
## Standard-errors: Heteroskedasticity-robust
##              Estimate Std. Error    t value   Pr(>|t|)
## (Intercept) -8.150218   0.163249 -49.925218  < 2.2e-16 ***
## age          0.091995   0.001970  46.706965  < 2.2e-16 ***
## smokev       0.544864   0.047784  11.402751  < 2.2e-16 ***
## Low          0.720048   0.089729   8.024654 1.0181e-15 ***
## LowMed       0.405444   0.098988   4.095873 4.2058e-05 ***
## Med          0.068182   0.102283   0.666598 5.0503e-01
## MedHigh     -0.001838   0.119948  -0.015320 9.8778e-01
## LessHS       0.313188   0.109810   2.852095 4.3432e-03 **
## HsGrad       0.367721   0.150505   2.443249 1.4556e-02 *
## SomeCol      0.170863   0.092956   1.838109 6.6046e-02 .
## ColGrad     -0.055031   0.108170  -0.508750 6.1093e-01
## black        0.054812   0.066339   0.826253 4.0866e-01
## hisp        -0.516003   0.083843  -6.154406 7.5359e-10 ***
## asian       -0.455072   0.125725  -3.619581 2.9508e-04 ***
## other        0.008824   0.191677   0.046034 9.6328e-01
## ... 2 variables were removed because of collinearity (High and PostGrad)
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## Log-Likelihood: -5,977.8   Adj. Pseudo R2: 0.299146
##            BIC: 12,105.9     Squared Cor.: 0.287117
```

```
#model for uninsured and smoking on mortality
logit_model_unsur_smoke <- feglm(mort ~  age + uninsured + smokev +
                            Low + LowMed + Med + MedHigh + High +
                            LessHS + HsGrad + SomeCol + ColGrad + PostGrad +
                            black + hisp + asian + other,
                            data = nhis2010,
                            vcov = 'hetero',
                            family = 'logit')
```

```
## NOTE: 174 observations removed because of NA values (RHS: 174).
## The variables 'High' and 'PostGrad' have been removed because of collinearity (see $collin.var).
```

```r
logitor(logit_model_unsur_smoke,
        data = nhis2010,
        robust = TRUE)
```

```
## Call:
## logitor(formula = logit_model_unsur_smoke, data = nhis2010, robust = TRUE)
##
## Odds Ratio:
##           OddsRatio Std. Err.       z     P>|z|
## age       1.0953696 0.0022779 43.8028 < 2.2e-16 ***
## uninsured 0.9035494 0.0795042 -1.1527 0.2490465
## smokev    1.7241335 0.0824595 11.3896 < 2.2e-16 ***
## Low       2.1000245 0.1906494  8.1727 3.017e-16 ***
## LowMed    1.5307410 0.1521341  4.2838 1.837e-05 ***
## Med       1.0887273 0.1114625  0.8303 0.4063450
## MedHigh   1.0102700 0.1211775  0.0852 0.9321141
## LessHS    1.3738312 0.1508272  2.8929 0.0038166 **
## HsGrad    1.4500078 0.2184666  2.4662 0.0136563 *
## SomeCol   1.1812746 0.1096947  1.7940 0.0728116 .
## ColGrad   0.9398872 0.1015706 -0.5737 0.5661864
## black     1.0525452 0.0698768  0.7714 0.4404773
## hisp      0.5913583 0.0498515 -6.2317 4.614e-10 ***
## asian     0.6295491 0.0795079 -3.6641 0.0002482 ***
## other     1.0084635 0.1930705  0.0440 0.9648873
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
logit_model_unsur_smoke
```

```
## GLM estimation, family = binomial(link = "logit"), Dep. Var.: mort
## Observations: 22,474
## Standard-errors: Heteroskedasticity-robust
##              Estimate Std. Error    t value   Pr(>|t|)
## (Intercept) -8.097249   0.168742 -47.986061  < 2.2e-16 ***
## age          0.091092   0.002080  43.788385  < 2.2e-16 ***
## uninsured   -0.101424   0.088021  -1.152278 2.4921e-01
## smokev       0.544725   0.047843  11.385654  < 2.2e-16 ***
## Low          0.741949   0.090815   8.169866 3.0873e-16 ***
## LowMed       0.425752   0.099420   4.282364 1.8492e-05 ***
## Med          0.085009   0.102414   0.830060 4.0650e-01
## MedHigh      0.010218   0.119986   0.085156 9.3214e-01
## ... 8 coefficients remaining (display them with summary() or use argument n)
## ... 2 variables were removed because of collinearity (High and PostGrad)
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## Log-Likelihood: -5,964.8   Adj. Pseudo R2: 0.299059
##            BIC: 12,089.8     Squared Cor.: 0.28697
```