ECON 121 FA23 Problem Set 3

Solution

Question 2

16 percent of the sample reports being in fair or poor health, and 13 percent died before 2019. The sample has a median age of 49. (The mean age is less meaningful because age was top-coded at 85. You did not need to notice this.) 56 percent of the sample is female, perhaps surprisingly. This gender imbalance has two sources. First, men and women responded to the survey at different rates, so the gender imbalance would shrink when we use the sampling weights. Second, men die at higher rates than women, so the gender imbalance grows with age. For the outcome variables, 16 percent of the sample reports being in fair or poor health, while 13 percent dies by 2019.

```r
# generate fair/poor health dummy
table(nhis2010$health)
```

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

```r
nhis2010$fpoor <- if_else(nhis2010$health == "Fair" |
                          nhis2010$health == "Poor", 1, 0)

# summarize the dataset
summary(nhis2010)
```

```
##    sampweight         psu             hhnum           pernum      
##  Min.   :  853   Min.   :  1.0   Min.   :    1   Min.   : 1.000  
##  1st Qu.: 4339   1st Qu.:156.0   1st Qu.:10380   1st Qu.: 1.000  
##  Median : 6879   Median :306.5   Median :21096   Median : 1.000  
##  Mean   : 8214   Mean   :304.8   Mean   :21235   Mean   : 1.371  
##  3rd Qu.:10712   3rd Qu.:460.0   3rd Qu.:31968   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      :11724   Min.   :0.0000  
##  1st Qu.:37.00   1st Qu.:0.0000   Widowed      : 2549   1st Qu.:0.0000  
##  Median :49.00   Median :0.0000   Divorced     : 3988   Median :1.0000  
##  Mean   :50.79   Mean   :0.4381   Separated    : 1003   Mean   :0.5764  
##  3rd Qu.:63.00   3rd Qu.:1.0000   Never married: 5043   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   :0.1611   Mean   :0.1824   Mean   :0.06249   Mean   :0.01757  
##  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.   : 1.00   Working for pay at job/business    :13248
##  1st Qu.:13.00   Not in labor force                 : 8856
##  Median :14.00   Not employed                       : 1451
##  Mean   :13.81   With job, but not at work          :  563
##  3rd Qu.:16.00   Working, w/out pay, at job/business:  224
##  Max.   :19.00   (Other)                            :    0
##  NA's   :119     NA's                               :   14
##                incfam          health          mort              bmi
##  $0 - $34,999     :9737   Excellent:5953   Min.   :0.0000   Min.   : 9.89
##  $35,000 - $49,999:3469   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 - $99,999:2334   Fair     :2968   Mean   :0.1289   Mean   :27.91
##  $100,000 and over:3634   Poor     : 962   3rd Qu.:0.0000   3rd Qu.:30.86
##  NA's             :1333   NA's     :  14   Max.   :1.0000   Max.   :87.84
##                                            NA's   :362      NA's   :933
##    uninsured         cancerev         cheartdiev         heartattev
##  Min.   :0.0000   Min.   :0.00000   Min.   :0.00000   Min.   :0.00000
##  1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.00000
##  Median :0.0000   Median :0.00000   Median :0.00000   Median :0.00000
##  Mean   :0.1744   Mean   :0.09488   Mean   :0.05445   Mean   :0.03798
##  3rd Qu.:0.0000   3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:0.00000
##  Max.   :1.0000   Max.   :1.00000   Max.   :1.00000   Max.   :1.00000
##  NA's   :62       NA's   :20        NA's   :58        NA's   :25
##    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   :0.1272   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   :38       NA's   :16       NA's   :9745     NA's   :178
##     vig10fwk         hrsleep                           asad
##  Min.   : 0.000   Min.   : 3.000   None of the time    :17381
##  1st Qu.: 0.000   1st Qu.: 6.000   A little of the time: 3428
##  Median : 0.000   Median : 7.000   Some of the time    : 2428
##  Mean   : 1.494   Mean   : 7.158   Most of the time    :  650
##  3rd Qu.: 2.000   3rd Qu.: 8.000   All of the time     :  302
##  Max.   :28.000   Max.   :22.000   NA's                :  167
##  NA's   :309      NA's   :367
##      fpoor
##  Min.   :0.0000
##  1st Qu.:0.0000
##  Median :0.0000
##  Mean   :0.1614
##  3rd Qu.:0.0000
##  Max.   :1.0000
##  NA's   :14
```

Question 3

5-year mortality is higher for people with fair/poor health than for people with good/very good/excellent health. Thus, self-reported health status is predictive of mortality. In both groups, 5-year mortality rises non-linearly with age.

```
tbl <-
  nhis2010 |>
  drop_na(age, fpoor, mort) |>
  group_by(age, fpoor) |>
  summarise(mort = mean(mort))
```

```
## `summarise()` has grouped output by 'age'. You can override using the `.groups`
## argument.
```

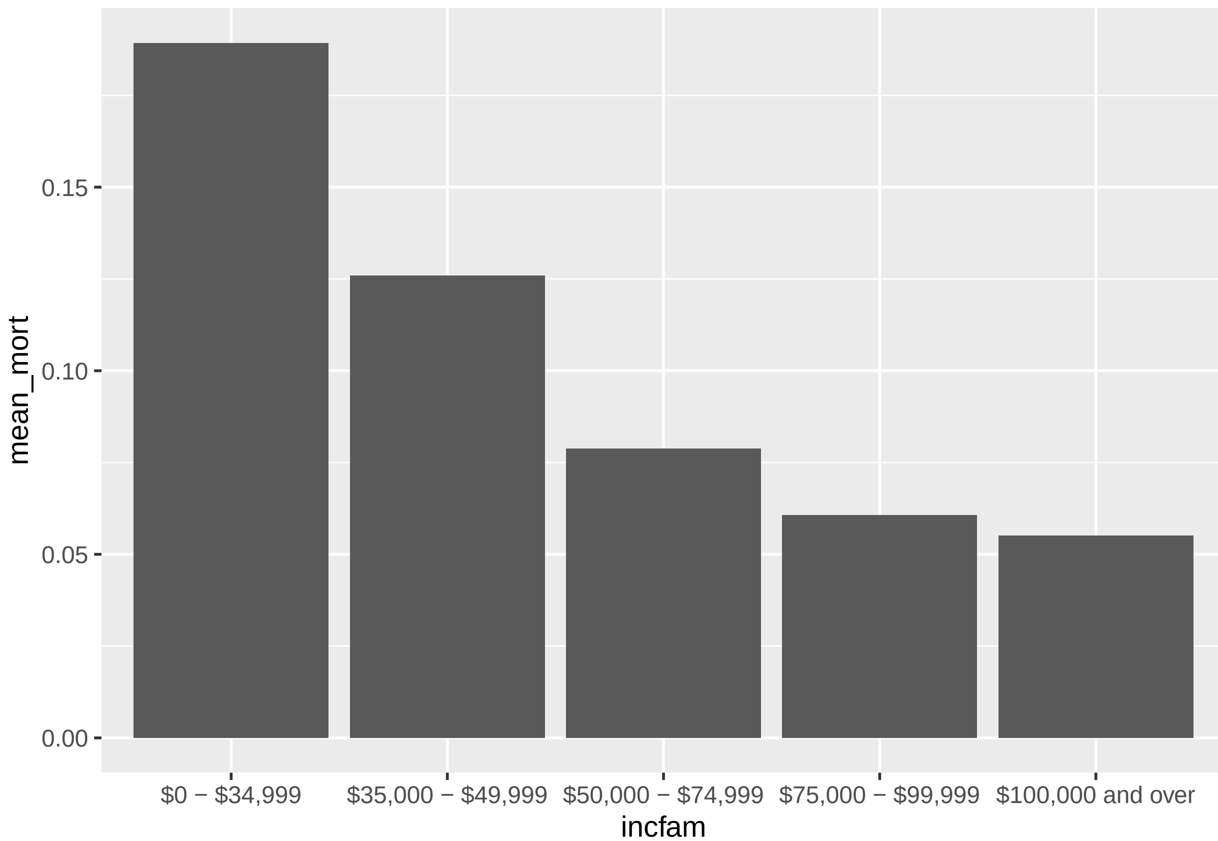
```
ggplot(tbl, aes(x = age, y = mort, color = factor(fpoor))) +
    geom_line() +
    labs(x="age", y="mortality rate")
```

Question 4

Rates of mortality and fair/poor health decline with family income. The same general pattern holds for education as well, although individuals with post- graduate education do not appear to be in worse health than college graduates.

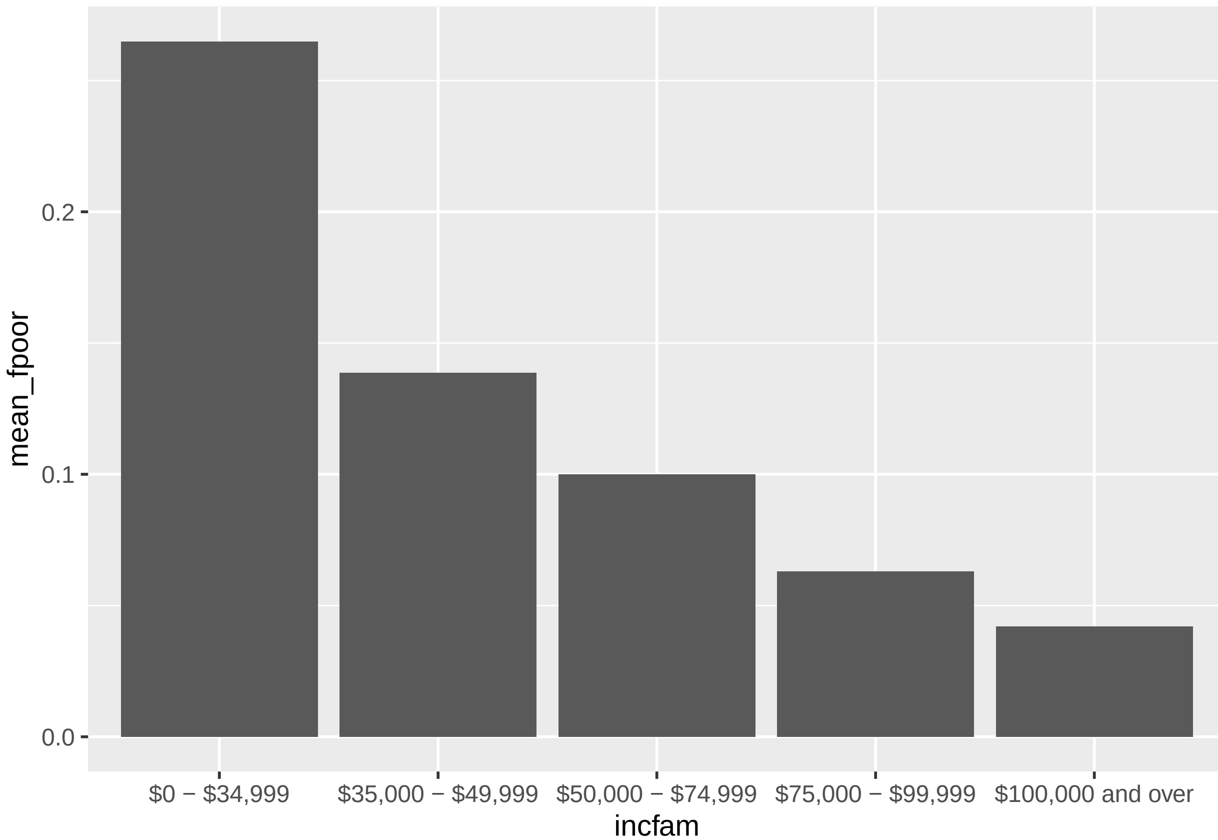
```r
# Mortality by income
tbl_mort_income <-
  nhis2010 |>
  drop_na(incfam, mort) |>
  group_by(incfam) |>
  summarize(mean_mort = mean(mort))

ggplot(tbl_mort_income, aes(x = incfam, y = mean_mort)) +
    geom_col()
```



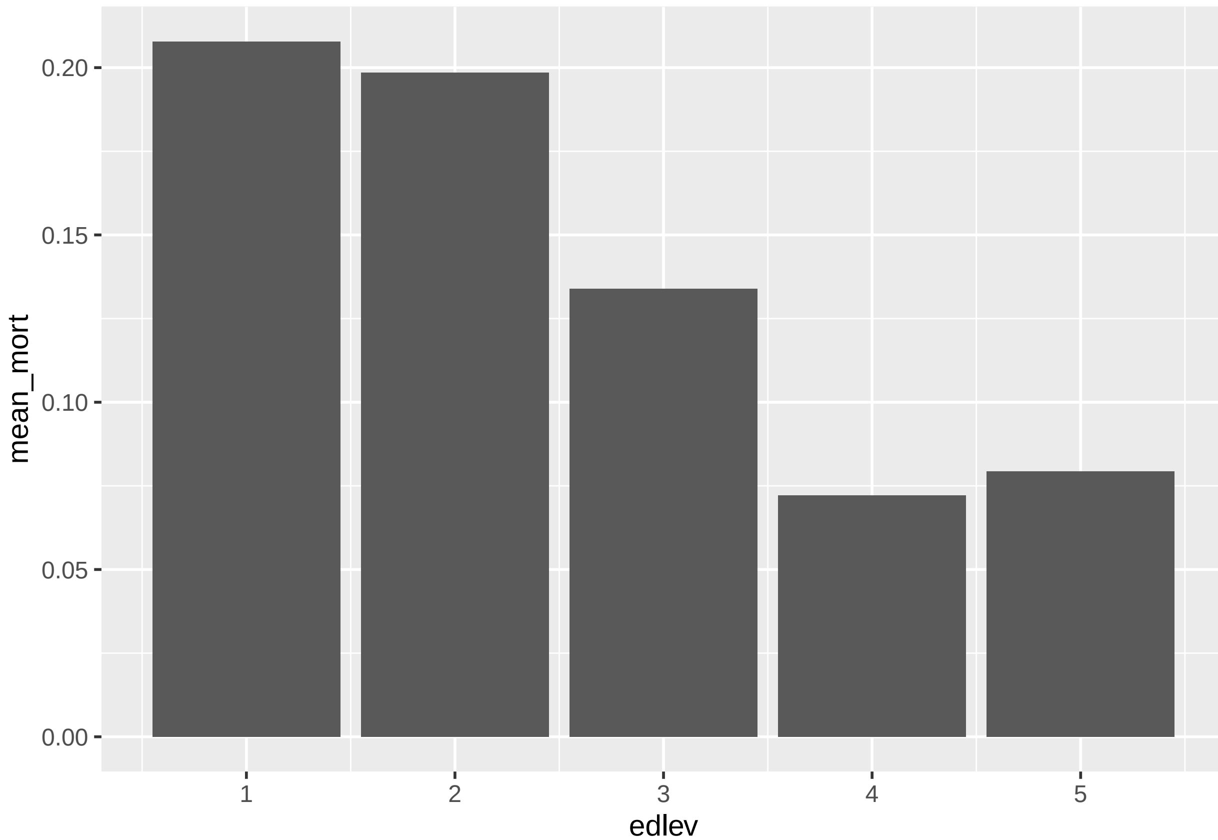

```r
# Health by income
tbl_health_income <-
  nhis2010 |>
  drop_na(incfam, fpoor) |>
  group_by(incfam) |>
  summarize(mean_fpoor = mean(fpoor))

ggplot(tbl_health_income, aes(x = incfam, y = mean_fpoor)) +
  geom_col()
```



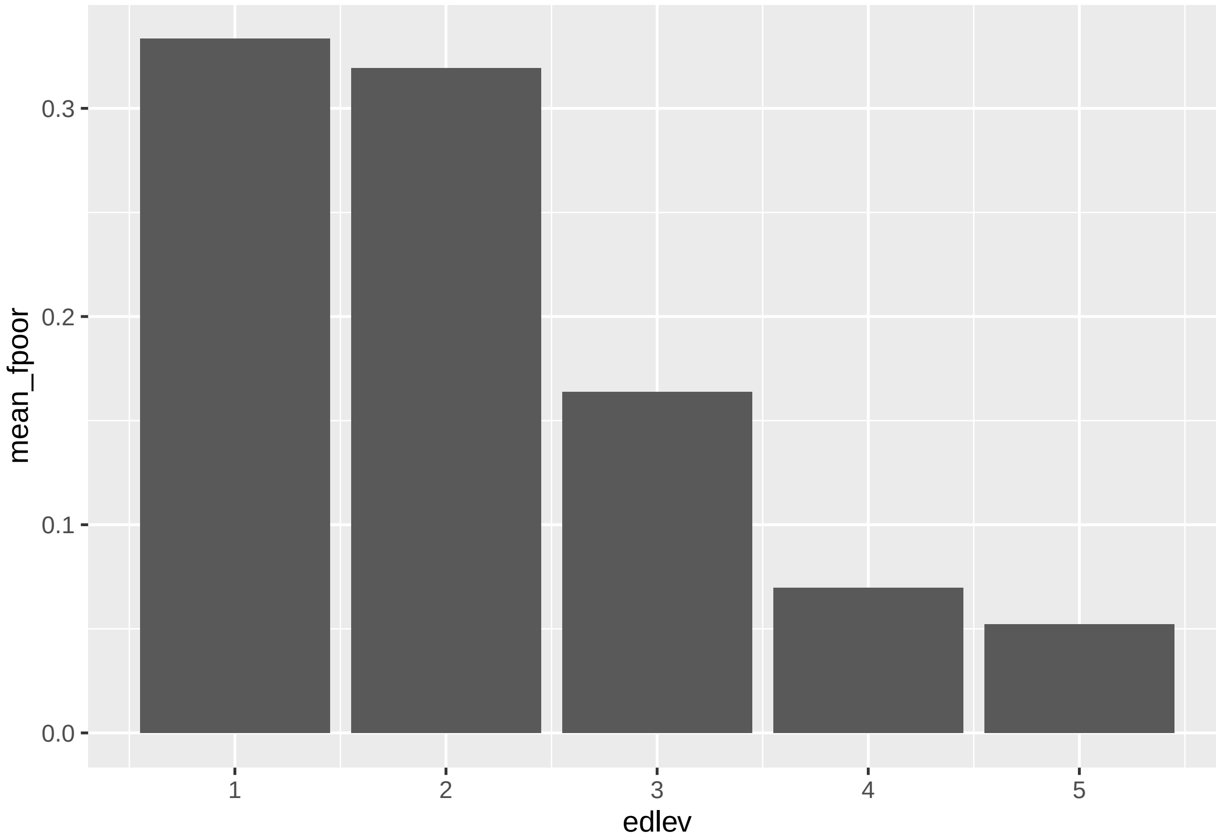

```
# Mortality by education
tbl_mort_education <-
  nhis2010 |>
  drop_na(edyrs, mort) |>
  mutate(edlev = case_when((edyrs<12)                ~ 1, # code edlev as
                           (edyrs==12)               ~ 2, # numeric so the
                           (edyrs>=13 & edyrs<15)    ~ 3, # graph is ordered
                           (edyrs==16)               ~ 4, # correctly
                           (edyrs>=17)               ~ 5)) |>
  group_by(edlev) |>
  summarize(mean_mort = mean(mort))

ggplot(tbl_mort_education, aes(x = edlev, y = mean_mort)) +
    geom_col()
```




```
# Health by education
tbl_health_education <-
  nhis2010 |>
  drop_na(edyrs, fpoor) |>
  mutate(edlev = case_when((edyrs<12)             ~ 1, # code edlev as
                          (edyrs==12)             ~ 2, # numeric so the
                          (edyrs>=13 & edyrs<15)  ~ 3, # graph is ordered
                          (edyrs==16)             ~ 4, # correctly
                          (edyrs>=17)             ~ 5)) |>
  group_by(edlev) |>
  summarize(mean_fpoor = mean(fpoor))

ggplot(tbl_health_education, aes(x = edlev, y = mean_fpoor)) +
    geom_col()
```


0.3
0.2
0.1
0.0
mean_fpoor
1
2
3
4
5
edlev

Question 5

Because age and education have non-linear relationships with health, I include a series of dummy variables for categories. I use the education categories from above, and 10-year age intervals.

For both outcomes and for all three models, the results show that mortality and fair/poor health decline with income, decline with education, and rise with age. One surprising result is that conditional on the socioeconomic variables, racial gaps in mortality are small and insignificant. There are larger racial gaps in fair/poor health. Another surprising result is that Hispanics have low mortality risk (conditional on the other covariates).

The linear probability results are similar to the probit and logit average marginal effects, although the similarity is much stronger for fair/poor health than for mortality. You did not need to comment on the reason in your response, but the larger difference in the case of mortality is probably due to the fact that mortality risk is exceptionally low across much of the age distribution, so that the marginal effect is calculated in the flatter part of the CDF.

```
# Generate age and education categories
nhis2010 <-
  nhis2010 %>%
  mutate(agecat = floor(age/10)*10,
         edlev = case_when((edyrs<12)                    ~ 1,
                           (edyrs==12)                   ~ 2,
                           (edyrs>=13 & edyrs<15)        ~ 3,
                           (edyrs==16)                   ~ 4,
                           (edyrs>=17)                   ~ 5))

# Mortality analyses
ols_mort <-
  feols(mort ~ incfam + factor(edlev) + factor(agecat) +
              black + hisp + asian + other,
        data = nhis2010, vcov = 'hetero')
```

```
## NOTE: 1,701 observations removed because of NA values (LHS: 362, RHS: 1,419).
```

```
summary(ols_mort)
```

```
## OLS estimation, Dep. Var.: mort
## Observations: 22,655
## Standard-errors: Heteroskedasticity-robust
##                          Estimate Std. Error    t value   Pr(>|t|)
## (Intercept)              0.073432   0.008105   9.059683  < 2.2e-16 ***
## incfam$35,000 - $49,999 -0.032004   0.005956  -5.372999 7.8202e-08 ***
## incfam$50,000 - $74,999 -0.054827   0.005510  -9.949666  < 2.2e-16 ***
## incfam$75,000 - $99,999 -0.057139   0.006096  -9.373067  < 2.2e-16 ***
## incfam$100,000 and over -0.057600   0.005674 -10.152461  < 2.2e-16 ***
## factor(edlev)2           0.010445   0.014579   0.716441 4.7373e-01
## factor(edlev)3          -0.018857   0.007588  -2.485204 1.2955e-02 *
## factor(edlev)4          -0.034375   0.008383  -4.100574 4.1356e-05 ***
## factor(edlev)5          -0.032430   0.009195  -3.527038 4.2107e-04 ***
## factor(agecat)30         0.013449   0.003125   4.304022 1.6843e-05 ***
## factor(agecat)40         0.028204   0.003677   7.670231 1.7860e-14 ***
## factor(agecat)50         0.077196   0.005002  15.433618  < 2.2e-16 ***
## factor(agecat)60         0.151830   0.006785  22.376300  < 2.2e-16 ***
## factor(agecat)70         0.325477   0.010798  30.143234  < 2.2e-16 ***
## factor(agecat)80         0.639263   0.013159  48.578255  < 2.2e-16 ***
## black                   -0.002238   0.005944  -0.376543 7.0652e-01
```

```
## hisp                   -0.046483   0.004988  -9.319681  < 2.2e-16 ***
## asian                  -0.037678   0.006817  -5.527279 3.2882e-08 ***
## other                  -0.000834   0.014384  -0.057952 9.5379e-01
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## RMSE: 0.284184   Adj. R2: 0.269659
```

```
probit_mort <-
  feglm(mort ~ incfam + factor(edlev) + factor(agecat) +
             black + hisp + asian + other,
        data = nhis2010, vcov = 'hetero', family = 'probit')
```

```
## NOTE: 1,701 observations removed because of NA values (LHS: 362, RHS: 1,419).
```

```
summary(probit_mort)
```

```
## GLM estimation, family = binomial(link = "probit"), Dep. Var.: mort
## Observations: 22,655
## Standard-errors: Heteroskedasticity-robust
##                         Estimate Std. Error    z value   Pr(>|z|)
## (Intercept)            -1.932348   0.085988 -22.472354  < 2.2e-16 ***
## incfam$35,000 - $49,999 -0.190400   0.036212  -5.257877 1.4573e-07 ***
## incfam$50,000 - $74,999 -0.357315   0.040155  -8.898419  < 2.2e-16 ***
## incfam$75,000 - $99,999 -0.391555   0.051932  -7.539749 4.7088e-14 ***
## incfam$100,000 and over -0.421553   0.046507  -9.064244  < 2.2e-16 ***
## factor(edlev)2           0.049972   0.073592   0.679050 4.9711e-01
## factor(edlev)3          -0.092318   0.039459  -2.339613 1.9304e-02 *
## factor(edlev)4          -0.233724   0.052412  -4.459321 8.2220e-06 ***
## factor(edlev)5          -0.198558   0.059864  -3.316843 9.1041e-04 ***
## factor(agecat)30         0.249896   0.086916   2.875136 4.0385e-03 **
## factor(agecat)40         0.495115   0.083473   5.931454 3.0026e-09 ***
## factor(agecat)50         0.941066   0.080090  11.750129  < 2.2e-16 ***
## factor(agecat)60         1.304036   0.079502  16.402509  < 2.2e-16 ***
## factor(agecat)70         1.821506   0.080705  22.570043  < 2.2e-16 ***
## factor(agecat)80         2.617116   0.084105  31.117342  < 2.2e-16 ***
## black                  -0.000115   0.035422  -0.003237 9.9742e-01
## hisp                   -0.334561   0.043756  -7.646122 2.0713e-14 ***
## asian                  -0.299337   0.068069  -4.397580 1.0946e-05 ***
## other                   0.018750   0.096302   0.194701 8.4563e-01
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## Log-Likelihood: -6,114.4   Adj. Pseudo R2: 0.287882
##            BIC: 12,419.4     Squared Cor.: 0.275609
```

```
avg_slopes(probit_mort) # probit marginal effects
```

```
## Warning: The `agecat` variable is treated as a categorical (factor) variable, but
##   the original data is of class numeric. It is safer and faster to convert
##   such variables to factor before fitting the model and calling a
##   `marginaleffects` function.
##
##   This warning appears once per session.
```

```
##
##    Term                          Contrast  Estimate Std. Error        z Pr(>|z|)
##   agecat 30 - 20                           1.01e-02    0.00323  3.12624  0.00177
##   agecat 40 - 20                           2.61e-02    0.00374  6.98816  < 0.001
```

```
##  agecat 50 - 20                         7.83e-02    0.00494 15.84384  < 0.001
##  agecat 60 - 20                         1.50e-01    0.00657 22.83655  < 0.001
##  agecat 70 - 20                         3.02e-01    0.01050 28.81040  < 0.001
##  agecat 80 - 20                         6.02e-01    0.01441 41.73642  < 0.001
##  asian  1 - 0                          -3.92e-02    0.00782 -5.01889  < 0.001
##  black  1 - 0                          -1.69e-05    0.00522 -0.00324  0.99742
##  edlev  2 - 1                           8.28e-03    0.01234  0.67111  0.50215
##  edlev  3 - 1                          -1.44e-02    0.00633 -2.27731  0.02277
##  edlev  4 - 1                          -3.43e-02    0.00778 -4.41678  < 0.001
##  edlev  5 - 1                          -2.96e-02    0.00882 -3.35708  < 0.001
##  hisp   1 - 0                          -4.49e-02    0.00527 -8.51433  < 0.001
##  incfam $100,000 and over - $0 - $34,999 -6.17e-02    0.00624 -9.89274  < 0.001
##  incfam $35,000 - $49,999 - $0 - $34,999 -3.08e-02    0.00568 -5.43109  < 0.001
##  incfam $50,000 - $74,999 - $0 - $34,999 -5.38e-02    0.00566 -9.51247  < 0.001
##  incfam $75,000 - $99,999 - $0 - $34,999 -5.81e-02    0.00690 -8.42709  < 0.001
##  other  1 - 0                           2.79e-03    0.01442  0.19321  0.84680
##      S    2.5 %   97.5 %
##    9.1  0.00376  0.01642
##   38.4  0.01879  0.03344
##  185.4  0.06863  0.08801
##  381.0  0.13720  0.16297
##  603.9  0.28184  0.32298
##    Inf  0.57328  0.62977
##   20.9 -0.05454 -0.02391
##    0.0 -0.01025  0.01022
##    1.0 -0.01590  0.03246
##    5.5 -0.02682 -0.00201
##   16.6 -0.04959 -0.01910
##   10.3 -0.04692 -0.01233
##   55.7 -0.05524 -0.03456
##   74.2 -0.07398 -0.04951
##   24.1 -0.04196 -0.01971
##   68.9 -0.06493 -0.04274
##   54.6 -0.07163 -0.04460
##    0.2 -0.02547  0.03105
##
## Columns: term, contrast, estimate, std.error, statistic, p.value, s.value, conf.low, conf.high
## Type:  response
```

```
logit_mort <-
  feglm(mort ~ incfam + factor(edlev) + factor(agecat) +
        black + hisp + asian + other,
        data = nhis2010, vcov = 'hetero', family = 'logit')
```

```
## NOTE: 1,701 observations removed because of NA values (LHS: 362, RHS: 1,419).
```

```
summary(logit_mort) # logit results
```

```
## GLM estimation, family = binomial(link = "logit"), Dep. Var.: mort
## Observations: 22,655
## Standard-errors: Heteroskedasticity-robust
##                            Estimate Std. Error    z value   Pr(>|z|)
## (Intercept)              -3.805985   0.209324 -18.182252  < 2.2e-16 ***
## incfam$35,000 - $49,999 -0.329330   0.066824  -4.928348 8.2928e-07 ***
## incfam$50,000 - $74,999 -0.666063   0.076015  -8.762250  < 2.2e-16 ***
```

```
## incfam$75,000 - $99,999 -0.752075   0.100595  -7.476249 7.6474e-14 ***
## incfam$100,000 and over -0.787600   0.090405  -8.711926  < 2.2e-16 ***
## factor(edlev)2           0.092624   0.135017   0.686018 4.9270e-01
## factor(edlev)3          -0.167995   0.071146  -2.361266 1.8213e-02 *
## factor(edlev)4          -0.434090   0.097586  -4.448303 8.6551e-06 ***
## factor(edlev)5          -0.359353   0.111136  -3.233459 1.2230e-03 **
## factor(agecat)30         0.605196   0.221428   2.733156 6.2731e-03 **
## factor(agecat)40         1.197880   0.210748   5.683950 1.3162e-08 ***
## factor(agecat)50         2.141488   0.201712  10.616549  < 2.2e-16 ***
## factor(agecat)60         2.821297   0.199807  14.120125  < 2.2e-16 ***
## factor(agecat)70         3.686436   0.200528  18.383665  < 2.2e-16 ***
## factor(agecat)80         4.979329   0.204565  24.341044  < 2.2e-16 ***
## black                   -0.014310   0.066288  -0.215869 8.2909e-01
## hisp                    -0.647340   0.083904  -7.715234 1.2076e-14 ***
## asian                   -0.623562   0.128269  -4.861373 1.1657e-06 ***
## other                    0.041697   0.184577   0.225908 8.2127e-01
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## Log-Likelihood: -6,116.8   Adj. Pseudo R2: 0.287604
##            BIC: 12,424.2     Squared Cor.: 0.275127
```

```
avg_slopes(logit_mort) # logit marginal effects
```

```
##
##    Term                         Contrast Estimate Std. Error      z Pr(>|z|)
##   agecat 30 - 20                          0.00955    0.00316  3.019  0.00253
##   agecat 40 - 20                          0.02602    0.00372  6.996  < 0.001
##   agecat 50 - 20                          0.07899    0.00494 15.984  < 0.001
##   agecat 60 - 20                          0.15079    0.00655 23.020  < 0.001
##   agecat 70 - 20                          0.29723    0.01036 28.681  < 0.001
##   agecat 80 - 20                          0.59276    0.01481 40.013  < 0.001
##   asian   1 - 0                          -0.04361    0.00773 -5.643  < 0.001
##   black   1 - 0                          -0.00114    0.00527 -0.216  0.82872
##   edlev   2 - 1                           0.00829    0.01222  0.678  0.49789
##   edlev   3 - 1                          -0.01417    0.00616 -2.300  0.02145
##   edlev   4 - 1                          -0.03445    0.00777 -4.432  < 0.001
##   edlev   5 - 1                          -0.02901    0.00884 -3.283  0.00103
##   hisp    1 - 0                          -0.04672    0.00540 -8.660  < 0.001
##   incfam $100,000 and over - $0 - $34,999 -0.06216    0.00648 -9.593  < 0.001
##   incfam $35,000 - $49,999 - $0 - $34,999 -0.02897    0.00570 -5.081  < 0.001
##   incfam $50,000 - $74,999 - $0 - $34,999 -0.05412    0.00575 -9.412  < 0.001
##   incfam $75,000 - $99,999 - $0 - $34,999 -0.05987    0.00709 -8.438  < 0.001
##   other   1 - 0                           0.00336    0.01501  0.224  0.82287
##      S    2.5 %   97.5 %
##    8.6  0.00335  0.01574
##   38.5  0.01873  0.03331
##  188.6  0.06931  0.08868
##  387.1  0.13795  0.16363
##  598.5  0.27692  0.31754
##    Inf  0.56372  0.62179
##   25.8 -0.05875 -0.02846
##    0.3 -0.01147  0.00919
##    1.0 -0.01567  0.03224
##    5.5 -0.02625 -0.00210
##   16.7 -0.04968 -0.01921
```

```
##    9.9 -0.04633 -0.01169
##   57.6 -0.05730 -0.03615
##   70.0 -0.07486 -0.04946
##   21.3 -0.04015 -0.01780
##   67.5 -0.06539 -0.04285
##   54.8 -0.07377 -0.04596
##    0.3 -0.02605  0.03277
##
## Columns: term, contrast, estimate, std.error, statistic, p.value, s.value, conf.low, conf.high
## Type:  response
```

```r
oddsratio <- exp(coef(logit_mort)) # logit odds ratios
ci <- exp(confint(logit_mort))
cbind(oddsratio, ci)
```

```
##                              oddsratio        2.5 %       97.5 %
## (Intercept)                 0.02223729   0.01475383   0.03351652
## incfam$35,000 - $49,999     0.71940576   0.63109333   0.82007624
## incfam$50,000 - $74,999     0.51372696   0.44261710   0.59626118
## incfam$75,000 - $99,999     0.47138737   0.38703579   0.57412276
## incfam$100,000 and over     0.45493515   0.38106297   0.54312807
## factor(edlev)2              1.09704919   0.84197568   1.42939632
## factor(edlev)3              0.84535799   0.73532767   0.97185263
## factor(edlev)4              0.64785375   0.53507169   0.78440794
## factor(edlev)5              0.69812808   0.56148249   0.86802851
## factor(agecat)30            1.83161153   1.18673525   2.82691594
## factor(agecat)40            3.31308713   2.19201709   5.00750946
## factor(agecat)50            8.51209139   5.73242693  12.63962030
## factor(agecat)60           16.79862430  11.35528025  24.85132661
## factor(agecat)70           39.90237508  26.93451630  59.11372304
## factor(agecat)80          145.37674878  97.35736292 217.08064447
## black                       0.98579235   0.86568717   1.12256089
## hisp                        0.52343649   0.44406313   0.61699731
## asian                       0.53603155   0.41687685   0.68924389
## other                       1.04257899   0.72609993   1.49699910
```

```r
# Fair/poor health analyses
ols_fpoor <-
  feols(fpoor ~ incfam + factor(edlev) + factor(agecat) +
              black + hisp + asian + other,
        data = nhis2010, vcov = 'hetero')
```

```
## NOTE: 1,426 observations removed because of NA values (LHS: 14, RHS: 1,419).
```

```r
summary(ols_fpoor)
```

```
## OLS estimation, Dep. Var.: fpoor
## Observations: 22,930
## Standard-errors: Heteroskedasticity-robust
##                          Estimate Std. Error    t value   Pr(>|t|)
## (Intercept)              0.241144   0.011218  21.496888  < 2.2e-16 ***
## incfam$35,000 - $49,999 -0.093818   0.007343 -12.776999  < 2.2e-16 ***
## incfam$50,000 - $74,999 -0.115876   0.006791 -17.062306  < 2.2e-16 ***
## incfam$75,000 - $99,999 -0.142442   0.007160 -19.894249  < 2.2e-16 ***
## incfam$100,000 and over -0.154658   0.006617 -23.372571  < 2.2e-16 ***
## factor(edlev)2          -0.014593   0.019831  -0.735868 4.6182e-01
```

```
## factor(edlev)3          -0.122761   0.009691 -12.668072  < 2.2e-16 ***
## factor(edlev)4          -0.160108   0.010542 -15.187872  < 2.2e-16 ***
## factor(edlev)5          -0.175008   0.010850 -16.129124  < 2.2e-16 ***
## factor(agecat)30         0.029379   0.006457   4.550053 5.3908e-06 ***
## factor(agecat)40         0.075986   0.007024  10.817946  < 2.2e-16 ***
## factor(agecat)50         0.158984   0.007890  20.150903  < 2.2e-16 ***
## factor(agecat)60         0.161410   0.008581  18.809394  < 2.2e-16 ***
## factor(agecat)70         0.165609   0.010816  15.311735  < 2.2e-16 ***
## factor(agecat)80         0.167053   0.013467  12.404813  < 2.2e-16 ***
## black                    0.059795   0.007383   8.099259 5.8030e-16 ***
## hisp                    -0.011195   0.006998  -1.599758 1.0967e-01
## asian                    0.010366   0.008622   1.202340 2.2924e-01
## other                    0.059887   0.018916   3.165876 1.5482e-03 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## RMSE: 0.346702   Adj. R2: 0.114901
```

```
probit_fpoor <-
  feglm(fpoor ~ incfam + factor(edlev) + factor(agecat) +
                black + hisp + asian + other,
        data = nhis2010, vcov = 'hetero', family = 'probit')
```

```
## NOTE: 1,426 observations removed because of NA values (LHS: 14, RHS: 1,419).
```

```
summary(probit_fpoor)
```

```
## GLM estimation, family = binomial(link = "probit"), Dep. Var.: fpoor
## Observations: 22,930
## Standard-errors: Heteroskedasticity-robust
##                          Estimate Std. Error    z value   Pr(>|z|)
## (Intercept)             -0.989583   0.054493 -18.159933  < 2.2e-16 ***
## incfam$35,000 - $49,999 -0.354259   0.031778 -11.147949  < 2.2e-16 ***
## incfam$50,000 - $74,999 -0.487437   0.033262 -14.654253  < 2.2e-16 ***
## incfam$75,000 - $99,999 -0.694225   0.046671 -14.874991  < 2.2e-16 ***
## incfam$100,000 and over -0.837998   0.045435 -18.444001  < 2.2e-16 ***
## factor(edlev)2          -0.022330   0.060059  -0.371802 0.71004013
## factor(edlev)3          -0.371939   0.031113 -11.954437  < 2.2e-16 ***
## factor(edlev)4          -0.627346   0.044523 -14.090392  < 2.2e-16 ***
## factor(edlev)5          -0.745788   0.055902 -13.340930  < 2.2e-16 ***
## factor(agecat)30         0.173580   0.051364   3.379405 0.00072643 ***
## factor(agecat)40         0.456868   0.049945   9.147343  < 2.2e-16 ***
## factor(agecat)50         0.825459   0.048763  16.928093  < 2.2e-16 ***
## factor(agecat)60         0.831443   0.050001  16.628616  < 2.2e-16 ***
## factor(agecat)70         0.815085   0.053535  15.225199  < 2.2e-16 ***
## factor(agecat)80         0.826804   0.059335  13.934393  < 2.2e-16 ***
## black                    0.236192   0.028609   8.255853  < 2.2e-16 ***
## hisp                    -0.002173   0.032986  -0.065863 0.94748712
## asian                    0.043994   0.050757   0.866760 0.38607379
## other                    0.271903   0.073617   3.693507 0.00022118 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## Log-Likelihood: -8,764.8   Adj. Pseudo R2: 0.136072
##            BIC: 17,720.3     Squared Cor.: 0.124998
```

```
avg_slopes(probit_fpoor) # probit marginal effects
```

```
##
##    Term                            Contrast  Estimate Std. Error        z Pr(>|z|)
##   agecat 30 - 20                             0.022485    0.00642   3.5034   <0.001
##   agecat 40 - 20                             0.070748    0.00705  10.0379   <0.001
##   agecat 50 - 20                             0.156646    0.00779  20.1151   <0.001
##   agecat 60 - 20                             0.158254    0.00840  18.8396   <0.001
##   agecat 70 - 20                             0.153874    0.00984  15.6360   <0.001
##   agecat 80 - 20                             0.157007    0.01207  13.0041   <0.001
##   asian  1 - 0                               0.009484    0.01112   0.8532    0.394
##   black  1 - 0                               0.053479    0.00687   7.7874   <0.001
##   edlev  2 - 1                              -0.006358    0.01703  -0.3734    0.709
##   edlev  3 - 1                              -0.093892    0.00857 -10.9580   <0.001
##   edlev  4 - 1                              -0.142910    0.01014 -14.0939   <0.001
##   edlev  5 - 1                              -0.161457    0.01093 -14.7755   <0.001
##   hisp   1 - 0                              -0.000461    0.00699  -0.0659    0.947
##   incfam $100,000 and over - $0 - $34,999 -0.163387    0.00686 -23.8344   <0.001
##   incfam $35,000 - $49,999 - $0 - $34,999 -0.086220    0.00721 -11.9625   <0.001
##   incfam $50,000 - $74,999 - $0 - $34,999 -0.111974    0.00693 -16.1649   <0.001
##   incfam $75,000 - $99,999 - $0 - $34,999 -0.145021    0.00767 -18.8980   <0.001
##   other  1 - 0                               0.063888    0.01892   3.3765   <0.001
##      S    2.5 %  97.5 %
##    11.1  0.00991  0.0351
##    76.3  0.05693  0.0846
##   296.5  0.14138  0.1719
##   260.6  0.14179  0.1747
##   180.7  0.13459  0.1732
##   126.0  0.13334  0.1807
##     1.3 -0.01230  0.0313
##    47.1  0.04002  0.0669
##     0.5 -0.03973  0.0270
##    90.4 -0.11069 -0.0771
##   147.4 -0.16278 -0.1230
##   161.7 -0.18287 -0.1400
##     0.1 -0.01416  0.0132
##   414.7 -0.17682 -0.1500
##   107.1 -0.10035 -0.0721
##   192.8 -0.12555 -0.0984
##   262.2 -0.16006 -0.1300
##    10.4  0.02680  0.1010
##
## Columns: term, contrast, estimate, std.error, statistic, p.value, s.value, conf.low, conf.high
## Type:  response
```

```r
logit_fpoor <-
  feglm(fpoor ~ incfam + factor(edlev) + factor(agecat) +
                black + hisp + asian + other,
        data = nhis2010, vcov = 'hetero', family = 'logit')
```

```
## NOTE: 1,426 observations removed because of NA values (LHS: 14, RHS: 1,419).
```

```r
summary(logit_fpoor)
```

```
## GLM estimation, family = binomial(link = "logit"), Dep. Var.: fpoor
## Observations: 22,930
## Standard-errors: Heteroskedasticity-robust
```

```
##                          Estimate Std. Error    z value   Pr(>|z|)
## (Intercept)             -1.751317   0.103826 -16.867805  < 2.2e-16 ***
## incfam$35,000 - $49,999 -0.629661   0.057292 -10.990381  < 2.2e-16 ***
## incfam$50,000 - $74,999 -0.881977   0.062021 -14.220517  < 2.2e-16 ***
## incfam$75,000 - $99,999 -1.310141   0.092879 -14.105905  < 2.2e-16 ***
## incfam$100,000 and over -1.609333   0.093835 -17.150666  < 2.2e-16 ***
## factor(edlev)2          -0.056218   0.101020  -0.556497 5.7787e-01
## factor(edlev)3          -0.640555   0.053029 -12.079315  < 2.2e-16 ***
## factor(edlev)4          -1.118474   0.082504 -13.556632  < 2.2e-16 ***
## factor(edlev)5          -1.360534   0.109570 -12.417022  < 2.2e-16 ***
## factor(agecat)30         0.343186   0.102040   3.363262 7.7027e-04 ***
## factor(agecat)40         0.884451   0.098060   9.019491  < 2.2e-16 ***
## factor(agecat)50         1.557965   0.095008  16.398234  < 2.2e-16 ***
## factor(agecat)60         1.548108   0.096927  15.971821  < 2.2e-16 ***
## factor(agecat)70         1.502036   0.102043  14.719707  < 2.2e-16 ***
## factor(agecat)80         1.496835   0.110544  13.540656  < 2.2e-16 ***
## black                    0.412738   0.050307   8.204395 2.3176e-16 ***
## hisp                    -0.013578   0.058954  -0.230319 8.1784e-01
## asian                    0.068276   0.093771   0.728118 4.6654e-01
## other                    0.495755   0.130652   3.794460 1.4797e-04 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## Log-Likelihood: -8,763.6   Adj. Pseudo R2: 0.136188
##            BIC: 17,718.0     Squared Cor.: 0.125558
```

```
avg_slopes(logit_fpoor) # logit marginal effects
```

```
##
##    Term                         Contrast Estimate Std. Error       z Pr(>|z|)
##  agecat 30 - 20                           0.02259    0.00643   3.515   <0.001
##  agecat 40 - 20                           0.07194    0.00710  10.137   <0.001
##  agecat 50 - 20                           0.16085    0.00785  20.487   <0.001
##  agecat 60 - 20                           0.15932    0.00839  18.980   <0.001
##  agecat 70 - 20                           0.15227    0.00963  15.814   <0.001
##  agecat 80 - 20                           0.15148    0.01156  13.099   <0.001
##  asian   1 - 0                            0.00824    0.01149   0.717    0.473
##  black   1 - 0                            0.05257    0.00682   7.712   <0.001
##  edlev   2 - 1                           -0.00913    0.01629  -0.560    0.575
##  edlev   3 - 1                           -0.09130    0.00829 -11.009   <0.001
##  edlev   4 - 1                           -0.14111    0.01004 -14.051   <0.001
##  edlev   5 - 1                           -0.16085    0.01093 -14.711   <0.001
##  hisp    1 - 0                           -0.00161    0.00697  -0.231    0.817
##  incfam $100,000 and over - $0 - $34,999 -0.16624    0.00682 -24.393   <0.001
##  incfam $35,000 - $49,999 - $0 - $34,999 -0.08621    0.00719 -11.993   <0.001
##  incfam $50,000 - $74,999 - $0 - $34,999 -0.11255    0.00696 -16.173   <0.001
##  incfam $75,000 - $99,999 - $0 - $34,999 -0.14773    0.00767 -19.266   <0.001
##  other   1 - 0                            0.06613    0.01928   3.430   <0.001
##      S    2.5 %  97.5 %
##   11.2  0.00999  0.0352
##   77.8  0.05803  0.0858
##  307.4  0.14546  0.1762
##  264.4  0.14287  0.1758
##  184.7  0.13339  0.1711
##  127.8  0.12881  0.1741
##    1.1 -0.01429  0.0308
```

```
##   46.2  0.03921  0.0659
##    0.8 -0.04106  0.0228
##   91.2 -0.10755 -0.0750
##  146.6 -0.16079 -0.1214
##  160.3 -0.18228 -0.1394
##    0.3 -0.01528  0.0121
##  434.2 -0.17960 -0.1529
##  107.7 -0.10030 -0.0721
##  193.0 -0.12619 -0.0989
##  272.3 -0.16276 -0.1327
##   10.7  0.02835  0.1039
##
## Columns: term, contrast, estimate, std.error, statistic, p.value, s.value, conf.low, conf.high
## Type:  response
```

```r
oddsratio <- exp(coef(logit_fpoor)) # logit odds ratios
ci <- exp(confint(logit_fpoor))
cbind(oddsratio, ci)
```

```
##                          oddsratio     2.5 %    97.5 %
## (Intercept)              0.1735452 0.1415910 0.2127108
## incfam$35,000 - $49,999 0.5327724 0.4761838 0.5960857
## incfam$50,000 - $74,999 0.4139636 0.3665805 0.4674713
## incfam$75,000 - $99,999 0.2697820 0.2248818 0.3236470
## incfam$100,000 and over 0.2000210 0.1664191 0.2404076
## factor(edlev)2           0.9453334 0.7755258 1.1523218
## factor(edlev)3           0.5269996 0.4749762 0.5847211
## factor(edlev)4           0.3267780 0.2779876 0.3841318
## factor(edlev)5           0.2565237 0.2069480 0.3179755
## factor(agecat)30         1.4094312 1.1539515 1.7214729
## factor(agecat)40         2.4216534 1.9982197 2.9348150
## factor(agecat)50         4.7491453 3.9422525 5.7211914
## factor(agecat)60         4.7025646 3.8889286 5.6864283
## factor(agecat)70         4.4908240 3.6767768 5.4851032
## factor(agecat)80         4.4675262 3.5972622 5.5483278
## black                    1.5109489 1.3690788 1.6675202
## hisp                     0.9865136 0.8788639 1.1073491
## asian                    1.0706612 0.8909104 1.2866786
## other                    1.6417367 1.2708436 2.1208741
```

Question 6

It is possible to use the hypotheses() function with coefficients on
the categories of factor variables, but it's easier to work with dummy variables. For pedagogical purposes, I will generate income and education category dummies, and then I will re-run the model.

```
nhis2010 <-
  nhis2010 %>%
  mutate(inc_35_50 = ifelse(incfam=="$35,000 - $49,999",1,0),
         inc_50_75 = ifelse(incfam=="$50,000 - $74,999",1,0),
         inc_75_100 = ifelse(incfam=="$75,000 - $99,999",1,0),
         inc_gt_100 = ifelse(incfam=="$100,000 and over",1,0),
         ed_12 = ifelse(edlev==2,1,0),
         ed_13_15 = ifelse(edyrs>12 & edyrs<16,1,0),
         ed_16 = ifelse(edyrs==16,1,0),
         ed_gt16 = ifelse(edyrs>16,1,0))

logit_mort2 <-
  feglm(mort ~ inc_35_50 + inc_50_75 + inc_75_100 + inc_gt_100 +
             ed_12 + ed_13_15 + ed_16 + ed_gt16 + factor(agecat) +
             black + hisp + asian + other,
        data = nhis2010, vcov = 'hetero', family = 'logit')
```

```
## NOTE: 1,701 observations removed because of NA values (LHS: 362, RHS: 1,419).
```

```
summary(logit_mort2)
```

```
## GLM estimation, family = binomial(link = "logit"), Dep. Var.: mort
## Observations: 22,655
## Standard-errors: Heteroskedasticity-robust
##                    Estimate Std. Error    z value   Pr(>|z|)
## (Intercept)       -3.805985   0.209324 -18.182252  < 2.2e-16 ***
## inc_35_50         -0.329330   0.066824  -4.928348 8.2928e-07 ***
## inc_50_75         -0.666063   0.076015  -8.762250  < 2.2e-16 ***
## inc_75_100        -0.752075   0.100595  -7.476249 7.6474e-14 ***
## inc_gt_100        -0.787600   0.090405  -8.711926  < 2.2e-16 ***
## ed_12              0.092624   0.135017   0.686018 4.9270e-01
## ed_13_15          -0.167995   0.071146  -2.361266 1.8213e-02 *
## ed_16             -0.434090   0.097586  -4.448303 8.6551e-06 ***
## ed_gt16           -0.359353   0.111136  -3.233459 1.2230e-03 **
## factor(agecat)30   0.605196   0.221428   2.733156 6.2731e-03 **
## factor(agecat)40   1.197880   0.210748   5.683950 1.3162e-08 ***
## factor(agecat)50   2.141488   0.201712  10.616549  < 2.2e-16 ***
## factor(agecat)60   2.821297   0.199807  14.120125  < 2.2e-16 ***
## factor(agecat)70   3.686436   0.200528  18.383665  < 2.2e-16 ***
## factor(agecat)80   4.979329   0.204565  24.341044  < 2.2e-16 ***
## black             -0.014310   0.066288  -0.215869 8.2909e-01
## hisp              -0.647340   0.083904  -7.715234 1.2076e-14 ***
## asian             -0.623562   0.128269  -4.861373 1.1657e-06 ***
## other              0.041697   0.184577   0.225908 8.2127e-01
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## Log-Likelihood: -6,116.8   Adj. Pseudo R2: 0.287604
##            BIC: 12,424.2     Squared Cor.: 0.275127
```

This model is the same as the one above. We just coded the categorical variables as dummies. The difference in log odds between Groups A and B is given by:

```
hypotheses(logit_mort2, "asian - black - ed_16 - inc_gt_100 = 0")
```

```
##
##                                     Term Estimate Std. Error    z Pr(>|z|)   S
##  asian - black - ed_16 - inc_gt_100 = 0    0.612      0.187 3.27  0.00108 9.9
##  2.5 % 97.5 %
##  0.245   0.98
##
## Columns: term, estimate, std.error, statistic, p.value, s.value, conf.low, conf.high
```

Since this is positive, we conclude that the poorer, less- educated Asian group have higher mortality risk than richer, more-educated Black group. The difference is statistically significant at the 5% level. If we exponentiate this difference, we get:

```
exp(.612)
```

```
## [1] 1.844116
```

which implies that the odds of dying are 84 percent higher for the poorer, less-educated, Asian group. You did not need to state this quantity in your answer.

It's likely that this model is not the best for testing differences between these groups. It would be better to include interactions of race and income.

Question 7

We probably should not interpret these results as causal. One problem is that there are many confounding variables that we do not observe but may jointly determine health and income, for instance place of birth. Another problem is that there may be reverse causality, i.e. health may affect income.

Question 8

I use the logit model again, and I exponentiated the coefficients for interpretability. I control for insurance status, smoking status, exercise, bacon consumption, and obesity. To keep the samples the same in the regressions with and without the additional control variables, I run the long regression and the short regression on the same sample, which required subsetting the data first. This was not required.

Smoking, exercise, and obesity predicted mortality: ever smoking doubled the odds of death, ever exercising reduced the odds by 38%, ever binge drinking raised the odds by 20%, and obesity raised the odds by 17%. In contrast, uninsurance did not significanty associate with mortality. The patterns explain part of the socioeconomic gradient in health. After controlling for these variables, the odds ratio on the highest income category rose from 0.44 to 0.51 and that on the >16 years of education dummy rose from 0.65 to 0.87. Because the odds ratios are
moving closer to 1, mortality gaps are smaller after controlling for health behaviors. Health behavior explains a larger share of the education-mortality relationship than the income-mortality relationship.

```
# Recode behavior variables as 0/1 dummies
nhis2010 <-
  nhis2010 %>%
  mutate(exev = ifelse(vig10fwk > 0, 1, 0), # ever exercise
         bingev = ifelse(alc5upyr > 0, 1, 0), # ever binge drink
         obese = ifelse(bmi >= 30, 1, 0) ) # obese if bmi>=30

# Subset data to non-missing obs
nhis_subset <-
  nhis2010 %>%
  drop_na(mort, incfam, edlev, agecat, white, black, hisp,
          uninsured, smokev, exev, bingev, obese)

# Run regression without health behaviors, report odds ratios and CIs
logit_model1 <-
  feglm(mort ~ incfam + factor(edlev) + factor(agecat) +
                black + hisp + asian + other,
        data = nhis_subset, vcov = 'hetero', family = 'logit')
cbind(exp(coef(logit_model1)), exp(confint(logit_model1)))
```

```
##                          exp(coef(logit_model1))        2.5 %       97.5 %
## (Intercept)                           0.02748355   0.01661693   0.04545637
## incfam$35,000 - $49,999               0.66322885   0.54218269   0.81129942
## incfam$50,000 - $74,999               0.46099871   0.37185553   0.57151177
## incfam$75,000 - $99,999               0.48557368   0.37244428   0.63306598
## incfam$100,000 and over               0.44455744   0.35282746   0.56013589
## factor(edlev)2                        0.96438136   0.60932015   1.52634275
## factor(edlev)3                        0.77210731   0.59801353   0.99688331
## factor(edlev)4                        0.59568643   0.43914635   0.80802747
## factor(edlev)5                        0.65164663   0.46880033   0.90580851
## factor(agecat)30                      1.57269409   0.95956029   2.57760429
## factor(agecat)40                      2.63158819   1.63964737   4.22362549
## factor(agecat)50                      6.78294289   4.33401583  10.61563132
## factor(agecat)60                     12.36096673   7.92124302  19.28908102
## factor(agecat)70                     27.68028609  17.61740927  43.49097111
## factor(agecat)80                    168.38644050 104.27751900 271.90897534
## black                                 1.26354391   1.03731708   1.53910818
## hisp                                  0.52159771   0.39581901   0.68734489
## asian                                 0.57683989   0.37377527   0.89022544
## other                                 0.89862068   0.51457521   1.56929271
```

```r
# Run regression with health behaviors
logit_model2 <-
  feglm(mort ~ incfam + factor(edlev) + factor(agecat) +
                black + hisp + asian + other +
                uninsured + smokev + exev + bingev + obese,
        data = nhis_subset, vcov = 'hetero', family = 'logit')
cbind(exp(coef(logit_model2)), exp(confint(logit_model2)))
```

```
##                          exp(coef(logit_model2))       2.5 %       97.5 %
## (Intercept)                           0.01833452  0.01075589   0.03125309
## incfam$35,000 - $49,999               0.69135905  0.56413112   0.84728058
## incfam$50,000 - $74,999               0.48912590  0.39364126   0.60777201
## incfam$75,000 - $99,999               0.52185988  0.39624955   0.68728844
## incfam$100,000 and over               0.50827402  0.40075975   0.64463179
## factor(edlev)2                        1.00298351  0.64435663   1.56120985
## factor(edlev)3                        0.84234194  0.65295651   1.08665728
## factor(edlev)4                        0.75235950  0.55353964   1.02259131
## factor(edlev)5                        0.87234154  0.62405405   1.21941323
## factor(agecat)30                      1.53659614  0.93473555   2.52598470
## factor(agecat)40                      2.45293857  1.52163988   3.95422577
## factor(agecat)50                      5.80128711  3.68565060   9.13134093
## factor(agecat)60                     10.47090258  6.64464479  16.50047582
## factor(agecat)70                     23.79172112 14.92407194  37.92838820
## factor(agecat)80                    163.19634302 99.23440763 268.38520035
## black                                 1.26439402  1.03558439   1.54375854
## hisp                                  0.56489548  0.42772439   0.74605731
## asian                                 0.62696144  0.40244087   0.97674136
## other                                 0.80265160  0.45323793   1.42143793
## uninsured                             0.90131106  0.71650301   1.13378675
## smokev                                2.00578259  1.72220459   2.33605451
## exev                                  0.61552368  0.52115416   0.72698143
## bingev                                1.19625658  1.00757484   1.42027147
## obese                                 1.17465753  1.00388539   1.37447994
```