ECON 121 HW2

Wanjun Gu

2/11/2020

```
# Import the required packages
knitr::opts_chunk$set(echo = TRUE, fig.width=12, fig.height=8)
library(here)
```

```
## here() starts at C:/Users/wanju/Desktop/ECON 121 HW2
```

```
library(readstata13)
library(ggplot2)
library(gridExtra)
library(margins)
library(knitr)
library(MASS)

# Clear workspace
graphics.off()
rm(list = ls())

# Import data
suppressWarnings({
  nhis = read.dta13(file = "nhis2000.dta", nonint.factors = TRUE)
})
```

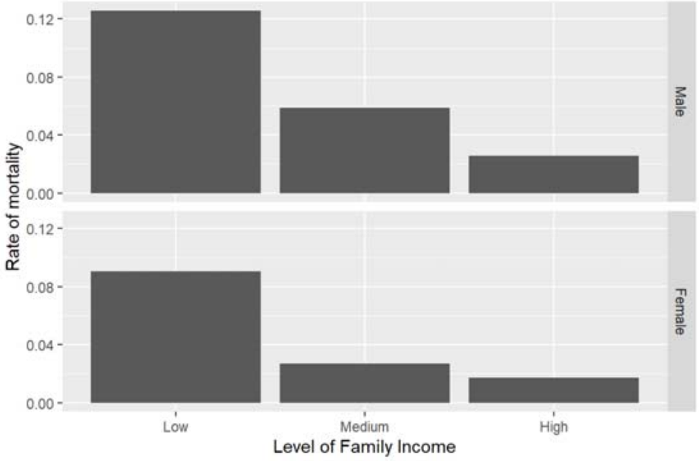
Problem 1

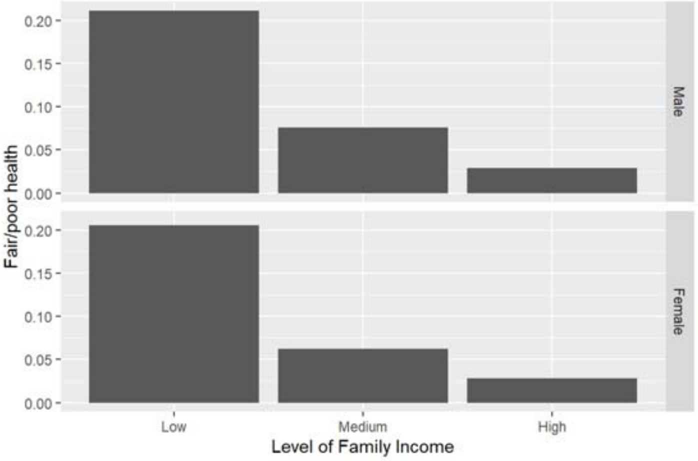
```
nhis$unhealthy = ifelse((nhis$health == "Poor" |
                           nhis$health == "Fair"), 1, 0)
```

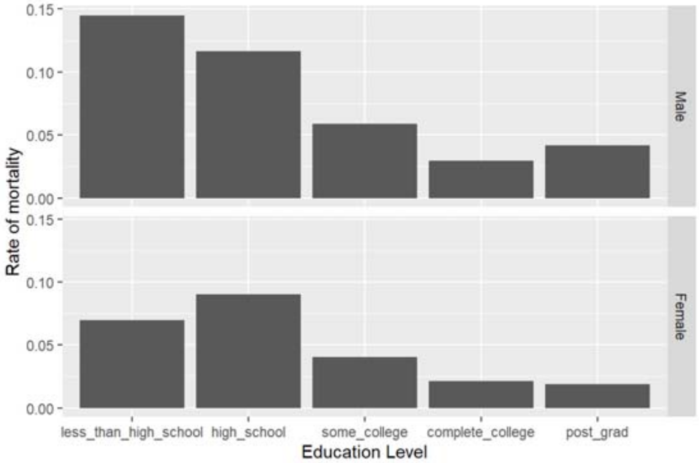
Problem 2

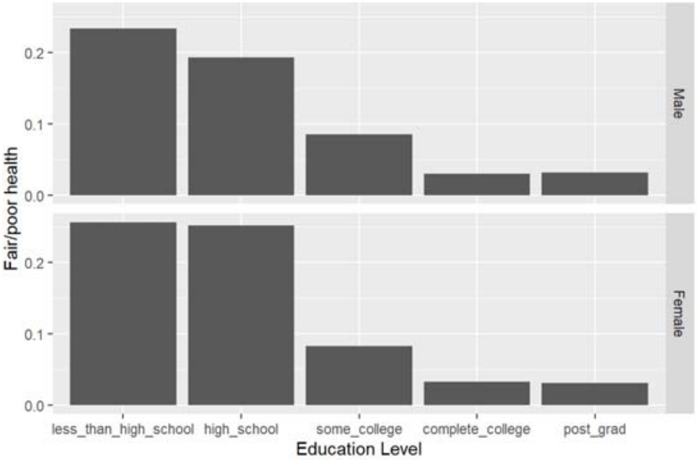
```
nhis$faminc_level = ifelse(nhis$faminc_gt75 == 1, "High",
                           ifelse(nhis$faminc_20t75 == 1, "Medium", "Low"))
nhis$faminc_level = factor(nhis$faminc_level,
                           levels = c("Low", "Medium", "High"))
nhis$educ_level = ifelse(nhis$edyrs < 12, "less_than_high_school",
                         ifelse(nhis$edyrs == 12, "high_school",
                                ifelse(nhis$edyrs >= 13 & nhis$edyrs <= 15, "some_college",
                                       ifelse(nhis$edyrs == 16, "complete_college",
                                              ifelse(nhis$edyrs > 16, "post_grad", NA)))))
nhis$educ_level = factor(nhis$educ_level,
                         levels = c("less_than_high_school",
                                    "high_school",
                                    "some_college",
                                    "complete_college",
                                    "post_grad"))

nhis = na.omit(nhis)
p1 = ggplot(data = nhis, aes(x = faminc_level, y = mort5)) +
  geom_bar(stat = "summary", fun.y = "mean") +
  facet_grid(rows = nhis$sex) +
  xlab("Level of Family Income") +
  ylab("Rate of mortality")
p2 = ggplot(data = nhis, aes(x = faminc_level, y = unhealthy)) +
  geom_bar(stat = "summary", fun.y = "mean") +
  facet_grid(rows = nhis$sex) +
  xlab("Level of Family Income") +
  ylab("Fair/poor health")
p3 = ggplot(data = nhis, aes(x = educ_level, y = mort5)) +
  geom_bar(stat = "summary", fun.y = "mean") +
  facet_grid(rows = nhis$sex) +
  xlab("Education Level") +
  ylab("Rate of mortality")
p4 = ggplot(data = nhis, aes(x = educ_level, y = unhealthy)) +
  geom_bar(stat = "summary", fun.y = "mean") +
  facet_grid(rows = nhis$sex) +
  xlab("Education Level") +
  ylab("Fair/poor health")
p5 = grid.arrange(p1,p2,p3,p4, nrow = 2)
```









Problem 3

```
nhis$age = as.numeric(nhis$age)
nhis$age[is.na(nhis$age)] = 85
nhis$faminc_20less = ifelse(nhis$faminc_gt75 == 1, 0,
                            ifelse(nhis$faminc_20t75 == 1, 0, 1))

fm1 = (mort5 ~ age + white + black + hisp + edyrs + faminc_level)
lp1 = summary(lm(data = nhis, fm1))$coefficient
pb1 = summary(glm(data = nhis, fm1, family = binomial(link = "probit")))$coefficient
pbm1 = margins(glm(data = nhis, fm1, family = binomial(link = "probit")))
lg1 = summary(glm(data = nhis, fm1, family = binomial(link = "logit")))$coefficient
lgm1 = margins(glm(data = nhis, fm1, family = binomial(link = "logit")))

fm2 = (unhealthy ~ age + white + black + hisp + edyrs + faminc_level)
lp2 = summary(lm(data = nhis, fm2))$coefficient
pb2 = summary(glm(data = nhis, fm2, family = binomial(link = "probit")))$coefficient
pbm2 = margins(glm(data = nhis, fm2, family = binomial(link = "probit")))
lg2 = summary(glm(data = nhis, fm2, family = binomial(link = "logit")))$coefficient
lgm2 = margins(glm(data = nhis, fm2, family = binomial(link = "logit")))
```

For both models to predict rate of mortality and fair/poor health, I used the formula: `mort5 ~ age + white + black + hisp + edyrs + faminc_level` and `unhealthy ~ age + white + black + hisp + edyrs + faminc_level` respectively. I believe age, race and years of education should be included in the regression as control variables because they are all potentially correlated with mortality and unhealthy rates. As the results have proven, most of the coefficients of these controlled variables are significant. Furthermore, as race/ethnicity is included, “other” is omitted to avoid perfect collinearity. Family income level is a factor with three different levels (High, Medium, Low). In this syntax, each single factor is paneled (Low is omitted again to avoid perfect collinearity). The marginal effects of the probit and logit models are calculated. As expected, the results of LP, probit and logit models are similar since the outcome is binary. With that said, the intercepts are not numerically the same since they represent different meanings in different regressions. An unexpecting complication is revealed that when predicting rates of mortality using all three models, neither the coefficient of white nor black is significantly different from zero. Interestingly, the coefficient of hisp is significant. This indicates that after controlling for education and age, race and ethnicity does not explain much variation of rates of mortality. However, Hispanics tend to have lower mortality. On the same note, for the prediction of unhealthy rate, race tends to have a significant impact, especially if the subject is white or black.

Problem 4

In this question, we are assuming that the population demographics are either white or black (mutually exclusive). Using the logit model from problem 3, we can acquire the difference in coefficients in the regression. However, the t score is smaller than the significant threshold thus the difference cannot be proven significant from the regression model. we know that it is hard to compare high-income African Americans with low-income whites in terms of mortality rates because the results on the race/ethnicity are not significant. However, with that said, I also believe that using the regression model to compare the effect size is not appropriate since the regression model used in problem 3 controlled for years of education and age, which are likely demographic characteristics of the white and African Americans. Thus, not controlling for these variables is a better way to make passive prediction.

```
kable(lg1)
```

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-6.0067462	0.3521191	-17.0588496	0.0000000
age	0.0784773	0.0029241	26.8377863	0.0000000
white	-0.1574562	0.2046597	-0.7693560	0.4416820
black	0.1396901	0.2366743	0.5902207	0.5550427
hisp	-0.3971912	0.1644742	-2.4149139	0.0157389
edyrs	-0.0552623	0.0158838	-3.4791554	0.0005030
faminc_levelMedium	-0.4647844	0.0960671	-4.8381239	0.0000013
faminc_levelHigh	-0.8566055	0.1601550	-5.3486038	0.0000001

```
black_white_diff = lg1[3,1] - (lg1[4,1] + lg1[8,1])
black_white_std = (lg1[3,1]^2 + lg1[4,1]^2 + lg1[8,1]^2)^0.5
t_score = black_white_diff/black_white_std
print(t_score)
```

```
## [1] 0.6342444
```

```
black_rich = nhis[nhis$black == 1 & nhis$faminc_level == "High",]$mort5
white_poor = nhis[nhis$white == 1 & nhis$faminc_level == "Low",]$mort5
t.test(black_rich, white_poor)
```

```
##
##  Welch Two Sample t-test
##
## data:  black_rich and white_poor
## t = -8.4992, df = 399.04, p-value = 3.851e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.12378003 -0.07727443
## sample estimates:
## mean of x mean of y
## 0.0130719 0.1135991
```

As the result shows, whites with low income has a significantly higher mortality rate.

Problem 5

The coefficients should not be considered as causal. This is because as much as we would like to control for all extraneous variables to eliminate omitted variable bias, we are never certain that there are no other variables that effects mortality rates and are correlated with family income.

Problem 6

```
fm3 = (mort5 ~ age + white + black + hisp + edyrs + faminc_level + hypertenev + smokev + diabeti
cev)
lg3 = glm(data = nhis, fm3, family = binomial(link = "logit"))
summary(lg3)$coefficients
```

```
##                               Estimate  Std. Error     z value      Pr(>|z|)
## (Intercept)                -6.79048638 0.371108050 -18.2978687  8.604983e-75
## age                         0.07554267 0.003107461  24.3100943 1.533193e-130
## white                      -0.10338592 0.206688701  -0.5002011  6.169335e-01
## black                       0.13678412 0.238978100   0.5723709  5.670707e-01
## hisp                       -0.31805975 0.166660614  -1.9084278  5.633595e-02
## edyrs                      -0.04553643 0.016278428  -2.7973481  5.152399e-03
## faminc_levelMedium         -0.43426368 0.096627818  -4.4941891  6.983556e-06
## faminc_levelHigh           -0.77030210 0.160850616  -4.7889285  1.676742e-06
## hypertenevYes               0.30226085 0.088784530   3.4044315  6.630193e-04
## smokevYes                   0.84390061 0.095118271   8.8721189  7.176707e-19
## diabeticevYes or mentioned  0.62312867 0.123682669   5.0381244  4.701160e-07
```

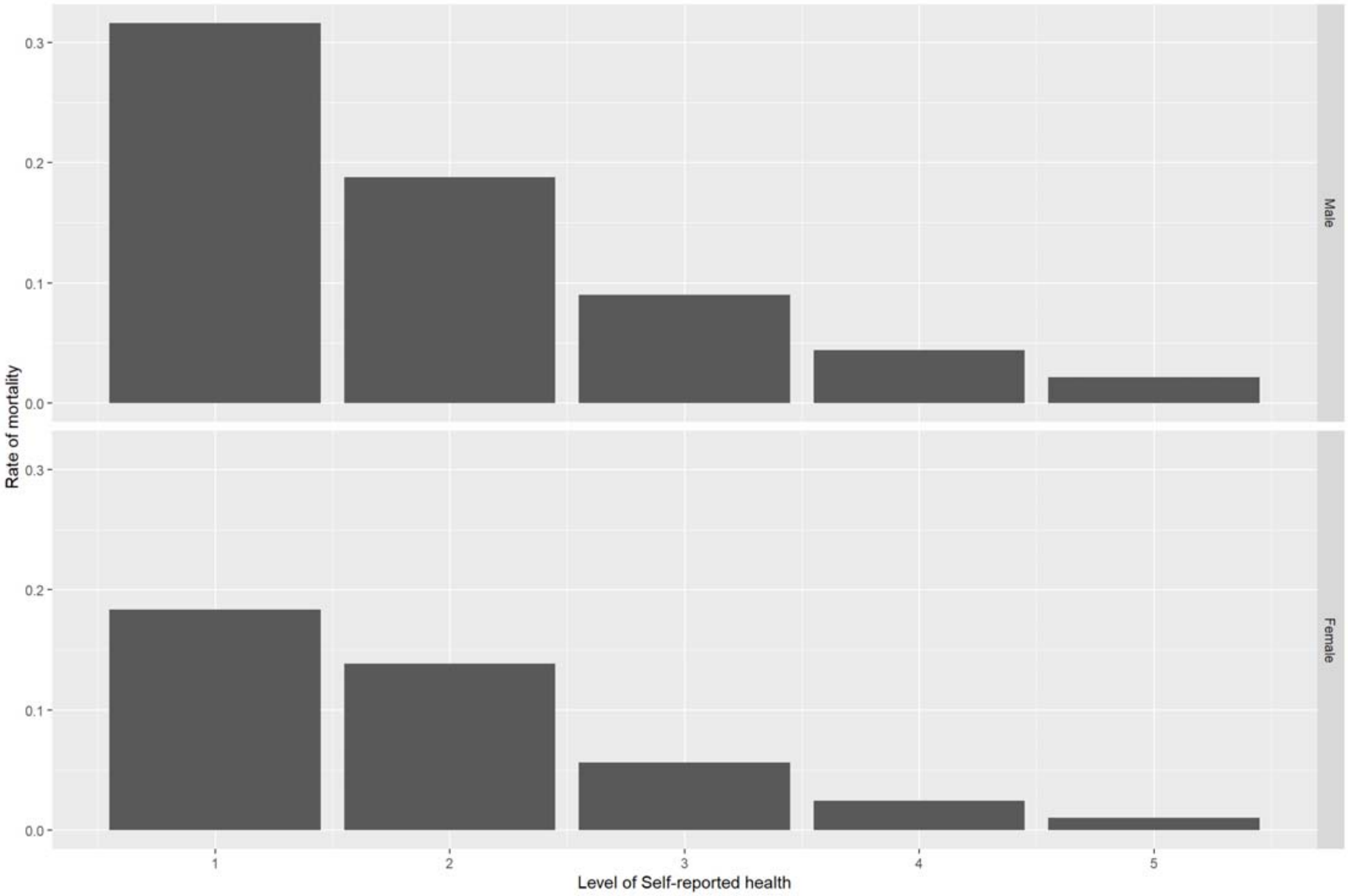
On top the variables that I used from problem 3, I added three more variables into the logit regression: hypertension, smoke and diabetic, each significantly correlated to the mortality rate. Moreover, all three variables have positive coefficients. This means that smoking, hypertension and diabetics are all correlated to high mortality rates. In particular, the presence of smoking, hypertension and diabetics increase the odd ratio of dying within five years after the study by 30, 84 and 62 percent.

Problem 7

```
nhis$health_level = ifelse(nhis$health == "Poor", 1,
                           ifelse(nhis$health == "Fair", 2,
                                  ifelse(nhis$health == "Good", 3,
                                         ifelse(nhis$health == "Very Good", 4,
                                                ifelse(nhis$health == "Excellent", 5, NA)))))
fm4 = mort5 ~ health_level
summary(lm(data = nhis, fm4))$coefficient
```

```
##                Estimate  Std. Error   t value      Pr(>|t|)
## (Intercept)   0.2191483 0.007277461  30.11328 3.770744e-193
## health_level -0.0433210 0.001805059 -23.99977 8.579927e-125
```

```
p6 = ggplot(data = nhis, aes(x = health_level, y = mort5)) +
  geom_bar(stat = "summary", fun.y = "mean") +
  facet_grid(rows = nhis$sex) +
  xlab("Level of Self-reported health") +
  ylab("Rate of mortality")
p6
```



As the regression shows, self-reported health condition is strongly correlated with mortality rate: self-reported healthier people tend to have lower mortality rates. This is also demonstrated in the graph. The graph shows that not only the correlated is substantiated, the relationship is as well monotonic. The healthier people report themselves to be, the lower their mortality rates tend to be.

Problem 8

```
fm5 = (health_level ~ age + white + black + hisp + edyrs + faminc_level)
nhis$health_level = as.factor(nhis$health_level)
pb3 = polr(data = nhis, fm5)
kable(summary(pb3)$coefficient)
```

```
##
## Re-fitting to get Hessian
```

	Value	Std. Error	t value
age	-0.0277485	0.0010963	-25.3111500
white	0.3289719	0.0642349	5.1213928
black	-0.0448920	0.0791118	-0.5674501
hisp	-0.0779682	0.0510379	-1.5276537
edyrs	0.1245529	0.0067048	18.5765521
faminc_levelMedium	0.6008476	0.0436976	13.7501138
faminc_levelHigh	1.1046861	0.0554953	19.9059272
1|2	-3.0422897	0.1418458	-21.4478598
2|3	-1.2793076	0.1304310	-9.8083062
3|4	0.4948096	0.1293980	3.8239357
4|5	2.1118848	0.1303544	16.2011011

As the new regression shows, the results are similar to the probit regression result from question 3, where all the health status levels are concatenated into binary variables. In fact, the results from this question should only be comparable to the results from the probit regression from question 3.