Open Log and Read in Data

- Set your working directory and create a log file to divert your codes and results

```
setwd("/Users/burrisfaculty/Desktop/DSCode/SOC686")
sink("Shepherd_asgn08.log", split=T)
```

- Read the external Stata data file gsscum7212Teach.dta into R.

```
> rm(list=ls(all=TRUE))

> mygss <- read.dta("gsscum7212teach.dta", convert.factor=F)
```

Manage Covariates Data

- Keep six variables, include mental health (`mntlhlth`), age (age), sex (`sex`), race (`race`), education (`educ`), and income (`inc1k`)
- Create dummy variables for sex and race. Note that the race variable has three categories, so please create three dummy variables for race (Alternatively, you can the factor function to turn the race variable into a factor variable and use it directly in the regression). Also, you need to be careful and clear about 1) how many of these three dummy variables, all measuring race, are usually used in a regression model and 2) how to interpret the results/corresponding coefficients (e.g., which group is the reference group?). Please also drop missing cases using listwise deletion (any case that has missing information for any of the six variables will be dropped from the sample data).
- Check the descriptive statistics of these variables using the table and the summary function when appropriate. Note that when there is too much output (e.g., tabulation of income), you can present representative information.

```
> #SELECT DATA
> useddta <- subset(mygss,
+                 select=c(mntlhlth, age, sex, race, educ, inc1k))

> #Create dummy variables female (male = 0)
> useddta$female <- as.numeric(useddta$sex==2)

> useddta$male <- as.numeric(useddta$sex == 1)

> #Create Binary Indicator Variables for Multi-Category Nomial Variables
>
> useddta$white <- ifelse(useddta$race == 1, 1, 0)

> useddta$black <- ifelse(useddta$race == 2, 1, 0)

> useddta$other <- ifelse(useddta$race == 3, 1, 0)

> nmdta <- useddta[complete.cases(useddta),] #no missing data

> #summarize data
> summary(nmdta$mntlhlth)
   Min.     1st Qu.  Median    Mean    3rd Qu.    Max.
  0.000     0.000    0.000     3.991   5.000      30.000

> summary(nmdta$inc1k)
   Min.     1st Qu.  Median    Mean       3rd Qu.    Max.
  0.245     15.921   28.157    39.404     48.475     155.140

> summary(nmdta$age)
```

```
   Min.     1st Qu.  Median    Mean       3rd Qu.    Max.
  18.00     31.00    42.00     41.81      51.00      84.00

> table(nmdta$female)

  0         1
  381       369

> table(nmdta$white)

  0         1
  160       590
```

Create Ordinal Response Variables

- Generate an ordered response variable based on the mental health variable such that the new ordered response variable is coded as one if the number of days for poor mental health is zero, two if the number of days for having poor mental health is one-seven, three if the number of days for having poor mental health is eight-fourteen, and four if the number of days for having poor mental health is fifteen and above. Label this new variable as `mntlhlthOrd`. Please crosstab `mntlhlthOrd` with `mntlhlth`.

```
> #Create Ordinal Response Variable
> nmdta$mntlhlthOrd <- cut(nmdta$mntlhlth, breaks = c(0,1,8,15,30), labels =
c(1,2,3,4), right = FALSE)

> #Used the table functions to verify the correctness of the ordinal variable
> table(nmdta$mntlhlthOrd)

  1    2    3   4
  399  226  44  58

> table(nmdta$mntlhlth)

0   1   2   3   4   5   6   7   8  10  12  14  15  16  18  20  21  25  27  30
399 34  62  37  29  39   6  19  2  35   3   4  22   1   2  21   2   9   1  23
```

Run Proportional Odds Model (Cumulative Logit)

- Run a proportional odds (cumulative ordered logit) regression of `mntlhlthOrd` on age (`age`), sex (sex; male is used as the reference category), race (race; white is used as the reference category), education (`educ`), and income (`inc1k`) (hint: when white is used as the reference category, the white dummy needs to be dropped from the equation and the other two dummy variables for race have to be retained).

```
> ordlogit.model1 <- polr(mntlhlthOrd ~ age + male + other + black + educ +
inc1k, data = nmdta, method = c("logistic"))

> summary(ordlogit.model1)
Call:
polr(formula = mntlhlthOrd ~ age + male + other + black + educ +
    inc1k, data = nmdta, method = c("logistic"))

Coefficients:
            Value             Std. Error        t value
age         -0.009002         0.005631          -1.5987
male        -0.499754         0.147603          -3.3858
other       -0.472610         0.284010          -1.6641
black       -0.575629         0.224130          -2.5683
educ        0.014756          0.028746          0.5133
inc1k       -0.002473         0.002246          -1.1013

Intercepts:
      Value       Std. Error  t value
1|2   -0.4317     0.4696      -0.9194
2|3   1.2236      0.4729      2.5872
3|4   1.8633      0.4793      3.8873

Residual Deviance: 1523.136
AIC: 1541.136
(23 observations deleted due to missingness)
```

Task 1 Calculate Predicted Probabilities for Estimation Sample

- Use the predict function to create four new variable that has the predicted probabilities for all cases in the estimation sample and then examine the descriptive statistics of these four new variable

```
> #TASK 1 Calculate Predicted Probabilities for Estimation Sample
> ordlogit.pred = predict(ordlogit.model1, type = "probs")

> summary(ordlogit.pred)
       1                2                 3                  4
 Min.   :0.3889   Min.   :0.1673   Min.   :0.02322   Min.   :0.02729
 1st Qu.:0.4787   1st Qu.:0.2837   1st Qu.:0.04904   1st Qu.:0.06153
 Median :0.5564   Median :0.3114   Median :0.05783   Median :0.07437
 Mean   :0.5488   Mean   :0.3110   Mean   :0.06047   Mean   :0.07969
 3rd Qu.:0.6058   3rd Qu.:0.3491   3rd Qu.:0.07334   3rd Qu.:0.09888
 Max.   :0.7822   Max.   :0.3802   Max.   :0.09420   Max.   :0.13670
```

Task 2 Calculate Predicted Probabilities for Hypothetical Cases

- Calculate the predicted probabilities of the four response levels of `mntlhlthOrd` for a 35-year-old white female with average (sample mean) education and sample median income. Please interpret the results.

```
> #TASK 2 Calculated Predicted Probabilities for Hypothetical Cases
> #35-year old white female with average education and median income and an
otherwise similar male
> hyp.person <- data.frame(age = 35, male = 0, black = 0, other = 0, educ =
mean(nmdta$educ), inc1k = median(nmdta$inc1k))

> hyp.pred = predict(ordlogit.model1, newdata = hyp.person, type = "probs")

> hyp.pred
         1              2           3            4
     0.43717960     0.36543077  0.08257299   0.11481665

> #Interpret results
```

Interpret Results: For a 35-year-old white woman with average education and median income, the predicted probabilities are as follows:

1. The predicted probability of this hypothetical person having zero mental health days during a month is 0.4372.
2. The predicted probability for having between one and seven, inclusive, mental health days is 0.3654.
3. The predicted probability for having between eight and fourteen, inclusive, mental health days is 0.0826.
4. The predicted probability for having more than fourteen mental health days is 0.1148.

Task 3 Calculate Differences in Predicted Probabilities

- Calculate the difference in the predicted probabilities of the four response levels of `mntlhlthOrd` for between a 35-year-old white female with average education and sample median income and an otherwise similar male. Please interpret the results.

```
> #TASK 3 Calculate Differences in Predicted Probabilities
> library(glm.predict)

> valuesw = c(35, 0, 0, 0, mean(nmdta$educ), median(nmdta$inc1k))

> valuesw
[1] 35.00000  0.00000  0.00000  0.00000 13.93200 28.15651

> valuesm = c(35,1,0,0,mean(nmdta$educ),median(nmdta$inc1k))

> valuesm
[1] 35.00000  1.00000  0.00000  0.00000 13.93200 28.15651

> dis.change = dc(ordlogit.model1,values1 = valuesw ,values2 =
valuesm,set.seed = 47)

> dis.change
Mean1        1:2.5%     1:97.5%   Mean2      2:2.5%     2:97.5%
1 0.43737154 0.38347272 0.4949230 0.56062814 0.49935992 0.62216254
2 0.36406560 0.32193012 0.4059347 0.30805107 0.26646937 0.35237992
3 0.08252519 0.05761795 0.1095573 0.05714878 0.03916635 0.07900081
4 0.11603768 0.08851998 0.1510008 0.07417201 0.05357920 0.10013140

Mean.Diff         diff:2.5%         diff:97.5%
-0.12325659       -0.19453381       -0.05457679
0.05601453        0.02321851        0.09138324
0.02537640        0.01121738        0.04329684
0.04186566        0.01860885        0.06964910
```

Interpret Results: We are 95% confident that that actual difference in predicted probabilities between a 35-year-old white woman with average education and median income and a similar man would be:

- Between -0.1945 and -0.05458 for no mental health days.
- Between 0.0232 and 0.0914 for between one and seven mental health days, inclusive.
- Between 0.01122 and 0.0433 for between eight and fourteen mental health days, inclusive
- Between 0.0186 and 0.0696 for more than fourteen mental health days.

Task 4 Compute AME

- Calculate the average marginal effects of education and interpret the results.

```
> #TASK 4: Calculate AME
> library(margins)

> summary(margins(ordlogit.model1))
 factor     AME         SE       z      p       lower    upper
  age       0.0022      0.0013   1.6651 0.0959  -0.0004  0.0047
  black     0.1384      0.0524   2.6407 0.0083   0.0357  0.2411
  educ      -0.0035     0.0070  -0.5044 0.6140  -0.0173  0.0102
  inc1k     0.0006      0.0005   1.0910 0.2753  -0.0005  0.0017
  male      0.1201      0.0341   3.5268 0.0004   0.0534  0.1869
  other     0.1136      0.0673   1.6884 0.0913  -0.0183  0.2455

> #Interpret results of average marginal effects of education
```

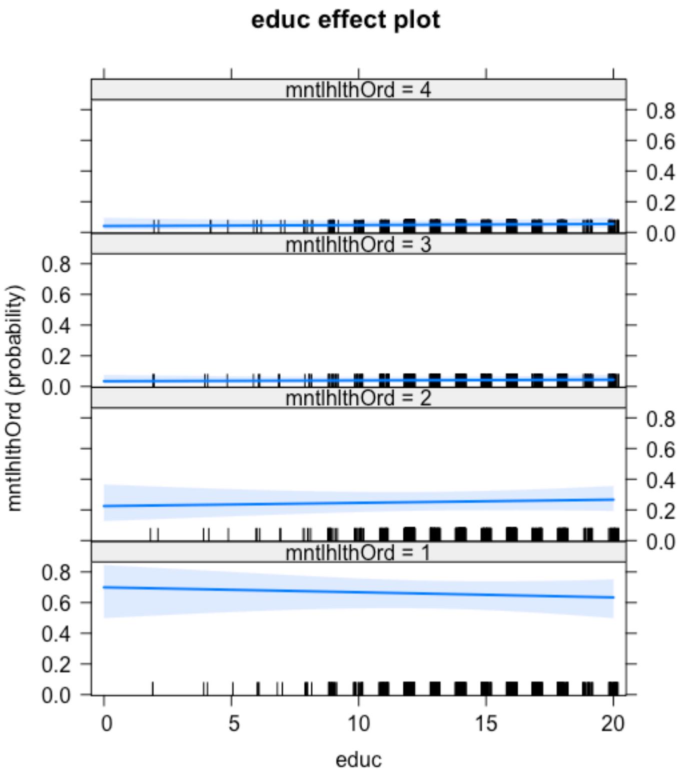
Interpretation: For each increase of one year of education, we would expect the ordinal variable for mental health to decrease by 0.0035 units.

Task 5 Plot Predicted Probabilities

- Plot the effects display of education for blacks with sample average age, sample average education, sample median income; that is to plot the effects of education on the predicted probabilities of the four response levels of `mntlhlthOrd` for blacks with sample average age, sample average education, sample median income. While displaying the effects of education, please restrict the values of education between 0 to 20. Please explain the probability plots and their confidence bands.

```
> #TASK 5 Plot the Predicted Probabilities
> require(effects)

> plot(effect("educ", ordlogit.model1, xlevels = list(educ =
0:20),given.values = c(black = 1, other = 0, age = mean(nmdta$age), educ =
mean(nmdta$educ), inc1k = median(nmdta$inc1k))))
```



Interpretation: For mntlhlthOrd values of 3 and 4, the plots are horizontal and the light blue envelopes indicating the confidence intervals are consistent in width and narrow. This would lead me to conclude education has very little effect for these values. For mntlhlthOrd value of 2, we see a slight increase in prevalence as education increases. Lastly, for mntlhlthOrd value of 1, we see a decrease as education increases. The light blue envelope is wider at the extremes for education, showing the effects are more variable near 0 and 20.

R-code

```
#
source("/Users/burrisfaculty/Desktop/DSCode/SOC686/Shepherd_Lab0
8_SOC686.r", echo=T, max.deparse.length=10000)
library(foreign)


#Open Log and read in data
setwd("/Users/burrisfaculty/Desktop/DSCode/SOC686")
sink("Shepherd_asgn08.log", split=T)
rm(list=ls(all=TRUE))
mygss <- read.dta("gsscum7212teach.dta", convert.factor=F)

#MANAGE DATA AND RUN LOGIT
#SELECT DATA
useddta <- subset(mygss,
                select=c(mntlhlth, age, sex, race, educ, inc1k))
```

```
#Create dummy variables female (male = 0)
useddta$female <- as.numeric(useddta$sex==2)
useddta$male <- as.numeric(useddta$sex == 1)

#Create Binary Indicator Variables for Multi-Category Nomial
Variables

useddta$white <- ifelse(useddta$race == 1, 1, 0)
useddta$black <- ifelse(useddta$race == 2, 1, 0)
useddta$other <- ifelse(useddta$race == 3, 1, 0)

nmdta <- useddta[complete.cases(useddta),] #no missing data

#summarize data
summary(nmdta$mntlhlth)
summary(nmdta$inc1k)
summary(nmdta$age)
table(nmdta$female)
table(nmdta$white)

#Create Ordinal Response Variable
nmdta$mntlhlthOrd <- cut(nmdta$mntlhlth, breaks =
c(0,1,8,15,30), labels = c(1,2,3,4), right = FALSE)

#Used the table functions to verify the correctness of the
ordinal variable
table(nmdta$mntlhlthOrd)
table(nmdta$mntlhlth)

#Run Proportional Odds Model
library (MASS)
ordlogit.model1 <- polr(mntlhlthOrd ~ age + male + other + black
+ educ + inc1k, data = nmdta, method = c("logistic"))
summary(ordlogit.model1)
ordlogit.model2 <- polr(mntlhlthOrd ~ male + other + black  +
inc1k, data = nmdta, method = c("logistic"))
summary(ordlogit.model2)

#TASK 1 Calculate Predicted Probabilities for Estimation Sample
ordlogit.pred = predict(ordlogit.model1, type = "probs")
summary(ordlogit.pred)

#TASK 2 Calculated Predicted Probabilities for Hypothetical
Cases
#35-year old white female with average education and median
income and an otherwise similar male
```

```
hyp.person <- data.frame(age = 35, male = 0, black = 0, other =
0, educ = mean(nmdta$educ), inc1k = median(nmdta$inc1k))
hyp.pred = predict(ordlogit.model1, newdata = hyp.person, type =
"probs")
hyp.pred
#Interpret results

#TASK 3 Calculate Differences in Predicted Probabilities
library(glm.predict)
valuesw = c(35, 0, 0, 0, mean(nmdta$educ), median(nmdta$inc1k))
valuesw
valuesm = c(35,1,0,0,mean(nmdta$educ),median(nmdta$inc1k))
valuesm
dis.change = dc(ordlogit.model1,values1 = valuesw ,values2 =
valuesm,set.seed = 47)
dis.change
#Interpret results
#TASK 4: Calculate AME
library(margins)
summary(margins(ordlogit.model1))

#Interpret results of average marginal effects of education

#TASK 5 Plot the Predicted Probabilities
require(effects)
plot(effect("educ", ordlogit.model1,
                    xlevels = list(educ = 0:20),
                    given.values = c(black = 1,
                         other = 0,
                         age = mean(nmdta$age),
                         educ = mean(nmdta$educ),
                         inc1k = median(nmdta$inc1k))))
#Close log
save(nmdta, file = "Assignment_08.rdata")
sink()
```

Log

```
> rm(list=ls(all=TRUE))

> mygss <- read.dta("gsscum7212teach.dta", convert.factor=F)

> #MANAGE DATA AND RUN LOGIT
> #SELECT DATA
> useddta <- subset(mygss,
+                  select=c(mntlhlth, age, sex, race, educ, inc1k))

> #Create dummy variables female (male = 0)
> useddta$female <- as.numeric(useddta$sex==2)
> useddta$male <- as.numeric(useddta$sex == 1)

> #Create Binary Indicator Variables for Multi-Category Nomial Variables
>
> useddta$white <- ifelse(useddta$race == 1, 1, 0)

> useddta$black <- ifelse(useddta$race == 2, 1, 0)

> useddta$other <- ifelse(useddta$race == 3, 1, 0)

> nmdta <- useddta[complete.cases(useddta),] #no missing data

> #summarize data
> summary(nmdta$mntlhlth)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.000   0.000   0.000   3.991   5.000  30.000

> summary(nmdta$inc1k)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.245  15.921  28.157  39.404  48.475 155.140

> summary(nmdta$age)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  18.00   31.00   42.00   41.81   51.00   84.00

> table(nmdta$female)

  0   1
381 369

> table(nmdta$white)

  0   1
160 590

> #Create Ordinal Response Variable
> nmdta$mntlhlthOrd <- cut(nmdta$mntlhlth, breaks = c(0,1,8,15,30), labels =
c(1,2,3,4), right = FALSE)

> #Used the table functions to verify the correctness of the ordinal variable
> table(nmdta$mntlhlthOrd)

  1   2   3   4
399 226  44  58

> table(nmdta$mntlhlth)

  0   1   2   3   4   5   6   7   8  10  12  14  15  16  18  20  21  25  27
30
399  34  62  37  29  39   6  19   2  35   3   4  22   1   2  21   2   9   1
23

> #Run Proportional Odds Model
> library (MASS)
> ordlogit.model1 <- polr(mntlhlthOrd ~ age + male + other + black + educ +
inc1k, data = nmdta, method = c("logistic"))

> summary(ordlogit.model1)
Call:
polr(formula = mntlhlthOrd ~ age + male + other + black + educ +
    inc1k, data = nmdta, method = c("logistic"))

Coefficients:
          Value Std. Error t value
age   -0.009002   0.005631 -1.5987
male  -0.499754   0.147603 -3.3858
other -0.472610   0.284010 -1.6641
black -0.575629   0.224130 -2.5683
educ   0.014756   0.028746  0.5133
inc1k -0.002473   0.002246 -1.1013

Intercepts:
    Value   Std. Error t value
1|2 -0.4317  0.4696    -0.9194
2|3  1.2236  0.4729     2.5872
3|4  1.8633  0.4793     3.8873

Residual Deviance: 1523.136
AIC: 1541.136
(23 observations deleted due to missingness)

> ordlogit.model2 <- polr(mntlhlthOrd ~ male + other + black  + inc1k, data =
nmdta, method = c("logistic"))

> summary(ordlogit.model2)
Call:
polr(formula = mntlhlthOrd ~ male + other + black + inc1k, data = nmdta,
    method = c("logistic"))

Coefficients:
          Value Std. Error  t value
male  -0.503079   0.147368  -3.414
other -0.435115   0.281733  -1.544
black -0.586781   0.223041  -2.631
inc1k -0.002645   0.002053  -1.288

Intercepts:
    Value   Std. Error t value
1|2 -0.2711  0.1360    -1.9939
2|3  1.3781  0.1498     9.1978
3|4  2.0173  0.1715    11.7595

Residual Deviance: 1526.105
AIC: 1540.105
(23 observations deleted due to missingness)

> #TASK 1 Calculate Predicted Probabilities for Estimation Sample
> ordlogit.pred = predict(ordlogit.model1, type = "probs")

> summary(ordlogit.pred)
       1                2                 3                  4
```

```
 Min.   :0.3889   Min.   :0.1673   Min.   :0.02322   Min.   :0.02729  
 1st Qu.:0.4787   1st Qu.:0.2837   1st Qu.:0.04904   1st Qu.:0.06153  
 Median :0.5564   Median :0.3114   Median :0.05783   Median :0.07437  
 Mean   :0.5488   Mean   :0.3110   Mean   :0.06047   Mean   :0.07969  
 3rd Qu.:0.6058   3rd Qu.:0.3491   3rd Qu.:0.07334   3rd Qu.:0.09888  
 Max.   :0.7822   Max.   :0.3802   Max.   :0.09420   Max.   :0.13670  

> #TASK 2 Calculated Predicted Probabilities for Hypothetical Cases
> #35-year old white female with average education and median income and an 
otherwise similar male
> hyp.person <- data.frame(age = 35, male = 0, black = 0, other = 0, educ = 
mean(nmdta$educ), inc1k = median(nmdta$inc1k))

> hyp.pred = predict(ordlogit.model1, newdata = hyp.person, type = "probs")

> hyp.pred
         1          2          3          4 
0.43717960 0.36543077 0.08257299 0.11481665 

> #Interpret results
> 
> #TASK 3 Calculate Differences in Predicted Probabilities
> library(glm.predict)

> valuesw = c(35, 0, 0, 0, mean(nmdta$educ), median(nmdta$inc1k))

> valuesw
[1] 35.00000  0.00000  0.00000  0.00000 13.93200 28.15651

> valuesm = c(35,1,0,0,mean(nmdta$educ),median(nmdta$inc1k))

> valuesm
[1] 35.00000  1.00000  0.00000  0.00000 13.93200 28.15651

> dis.change = dc(ordlogit.model1,values1 = valuesw ,values2 = 
valuesm,set.seed = 47)

> dis.change
       Mean1     1:2.5%   1:97.5%      Mean2     2:2.5%    2:97.5%   
Mean.Diff   diff:2.5%  diff:97.5%
1 0.43737154 0.38347272 0.4949230 0.56062814 0.49935992 0.62216254 -
0.12325659 -0.19453381 -0.05457679
2 0.36406560 0.32193012 0.4059347 0.30805107 0.26646937 0.35237992 
0.05601453  0.02321851  0.09138324
3 0.08252519 0.05761795 0.1095573 0.05714878 0.03916635 0.07900081 
0.02537640  0.01121738  0.04329684
4 0.11603768 0.08851998 0.1510008 0.07417201 0.05357920 0.10013140 
0.04186566  0.01860885  0.06964910

> #Interpret results
> #TASK 4: Calculate AME
> library(margins)

> summary(margins(ordlogit.model1))
 factor     AME     SE       z      p   lower  upper
    age  0.0022 0.0013  1.6651 0.0959 -0.0004 0.0047
  black  0.1384 0.0524  2.6407 0.0083  0.0357 0.2411
```

```
   educ -0.0035 0.0070 -0.5044 0.6140 -0.0173 0.0102
  inc1k  0.0006 0.0005  1.0910 0.2753 -0.0005 0.0017
   male  0.1201 0.0341  3.5268 0.0004  0.0534 0.1869
  other  0.1136 0.0673  1.6884 0.0913 -0.0183 0.2455

> #Interpret results of average marginal effects of education
>
> #TASK 5 Plot the Predicted Probabilities
> require(effects)

> plot(effect("educ", ordlogit.model1, xlevels = list(educ =
0:20),given.values = c(black = 1, other = 0, age = mean(nmdta$age), educ =
mean(nmdta$educ), inc1k = median(nmdta$inc1k))))

> #Close log
> save(nmdta, file = "Assignment_08.rdata")

> sink()
```