Open Log and Read in Data

- Set your working directory and create a log file to divert your codes and results
- Read the external Stata data file gsscum7212Teach.dta into R.

```
#Open Log and read in data
setwd("/Users/burrisfaculty/Desktop/DSCode/SOC686")
sink("Shepherd_asgn06F.log", split=T)
rm(list=ls(all=TRUE))
mygss <- read.dta("gsscum7212teach.dta", convert.factor=F)
```

Manage Data and Run Logit Models

- Keep six variables, include mental health (mntlhlth), age (age), sex (sex), race (race), education (educ), and income (inc1k)
- Dichotomize the mental health variable such that the new binary response variable is coded as one (1 = having poor mental health) if the number of days for poor mental health is greater than zero, otherwise it's coded as zero (0 = having excellent mental health). Label this new variable as mntlhlthc2.
- 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.
- Run a logit model of mntlhlthc2 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). Let's call this Model 1.
- Run a second binary logit regression of mntlhlthc2 on sex (sex; male is used as the reference category), race (race; white is used as the reference category), education (educ), and income (inc1k), with age excluded from the explanatory variables. Let's call this model Model 2. Then run a third binary logit regression of mental health (mntlhlthc2) on sex (sex; male is used as the reference category), race (race; white is used as the reference category), and income (inc1k), with age and educ excluded from the explanatory variables. Let's call this model Model 3.

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

> #Create Binary Response Variables
> # 1 = poor mental health mntlhl > 0
> useddta$mntlhc2 <- ifelse(useddta$mntlhlth > 0, 1, 0)
```

```
> #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(useddta$mntlhlth)
   Min. 1st Qu.  Median    Mean     3rd Qu.    Max.    NA's
   0.00    0.00    0.00    3.98      5.00      30.00   4954

> summary(useddta$inc1k)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.245  12.481  22.605  30.279  37.226 162.607

> summary(useddta$age)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
  18.00   31.00   43.00   45.57   59.00   89.00      18

> table(useddta$female)

   0    1
2480 3226

> table(useddta$white)

   0    1
1062 4644

> # Create 2 logit models
> logit.model <- glm(mntlhc2 ~ age + male + other + black + educ + inc1k,
family = binomial(link = 'logit'),data = nmdta)

> logit.model2 <- glm(mntlhc2 ~ male + other + black + educ + inc1k, family =
binomial(link = 'logit'),data = nmdta)

> logit.model3 <- glm(mntlhc2 ~ male + other + black + inc1k, family =
binomial(link = 'logit'),data = nmdta)
```

Task 1 Compare Models Using AIC

- Compare models 1-3 using AIC, and interpret the results.

```
> library(stats4)

> AIC(logit.model, logit.model2, logit.model3)
             df      AIC
logit.model   7   1029.969
logit.model2  6   1029.680
logit.model3  5   1028.425

> #Interpret results in document
```

So, looking strictly at the numbers, the model that excludes age and education would fit the data best, because the AIC is the smallest for this model. However, the difference between the AICs of the other models is less than 2, so it would be considered weak evidence that the third model has a significant advantage over the other two.

Task 2 Compare Models Using BIC

- Compare models 1-3 using BIC, and interpret the results.

```
> BIC(logit.model,logit.model2, logit.model3)
             df      BIC
logit.model   7   1062.310
logit.model2  6   1057.401
logit.model3  5   1051.525

> #Interpret results in document
```

According to the BICs of the three models, the third model which excludes age and education is the best-fitting model. The difference in BICS between model 1, which includes all the variables, and the third model, which excludes education and age, is greater than 10, which indicates a strong evidence based on Raftery's guidelines. This would be a clear indication to use the third model.

Task 3 Compare Models Using McFadden's Pseudo R-Squared

- Compare models 1-3 using the McFadden's pseudo R-squared, and interpret the results.

```
> #Compare using McFadden's and Tjur's
>
> library(DescTools)

> PseudoR2(logit.model, c("McFadden", "Tjur"))
  McFadden       Tjur
0.01994656  0.02747576

> PseudoR2(logit.model2, c("McFadden", "Tjur"))
  McFadden       Tjur
0.01829571  0.02514880

> PseudoR2(logit.model3, c("McFadden", "Tjur"))
  McFadden       Tjur
0.01757798  0.02418605
```

According to McFadden's Pseudo R-Squared, the best choice would be the original model, logit.model, because it has the pseudo R^2 closest to 1. However, since the differences in the Pseudo R^2 among all three models is small, it is fairly weak evidence.

Task 4 Compare Models Using Tjur's Pseudo R-Squared

- Compare models 1-3 using the Tjur's pseudo R-squared, and interpret the results.

```
> #Compare using McFadden's and Tjur's
>
> library(DescTools)

> PseudoR2(logit.model, c("McFadden", "Tjur"))
  McFadden       Tjur
0.01994656  0.02747576

> PseudoR2(logit.model2, c("McFadden", "Tjur"))
  McFadden       Tjur
0.01829571  0.02514880

> PseudoR2(logit.model3, c("McFadden", "Tjur"))
  McFadden       Tjur
0.01757798  0.02418605
```

Using Tjur's Pseudo R^2, we arrive at the same conclusion as the McFadden's R^2, that the original model (logit.model) is the best choice.

Task 5 Compare Models Using the Hosmer-Lemeshow Test of Goodness of Fit

- Test the goodness of fit of the full model using the Hosmer-Lemeshow Test and interpret the results.

```
> library(generalhoslem)

> HL1 = logitgof(nmdta$mntlhc2, fitted(logit.model), g = 10)

> HL1

      Hosmer and Lemeshow test (binary model)

data:  nmdta$mntlhc2, fitted(logit.model)
X-squared = 6.993, df = 8, p-value = 0.5374
```

With a degree of freedom of 8 and a test statistic $X^2 = 6.993$, we would have a p-value of 0.5374. Using the usual significance level of 0.05, we would not find that the difference between the expected and observed are statistically significant. Therefore, we would conclude that our model fits the data fairly well.

R-Script

```
#
source("/Users/burrisfaculty/Desktop/DSCode/SOC686/Shepherd_Lab06_SOC686.r",
echo=T, max.deparse.length=10000)
library(foreign)


#Open Log and read in data
setwd("/Users/burrisfaculty/Desktop/DSCode/SOC686")
sink("Shepherd_asgn06.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 Binary Response Variables
# 1 = poor mental health mntlhl > 0
useddta$mntlhc2 <- ifelse(useddta$mntlhlth > 0, 1, 0)

#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(useddta$mntlhlth)
summary(useddta$inc1k)
summary(useddta$age)
table(useddta$female)
table(useddta$white)
# Create 2 logit models
logit.model <- glm(mntlhc2 ~ age + male + other + black + educ + inc1k,
family = binomial(link = 'logit'),
                   data = nmdta)
logit.model2 <- glm(mntlhc2 ~ male + other + black + educ + inc1k, family =
binomial(link = 'logit'),
                    data = nmdta)
logit.model3 <- glm(mntlhc2 ~ male + other + black + inc1k, family =
binomial(link = 'logit'),
                    data = nmdta)
#Task 1
#Compare using AIC
library(stats4)
AIC(logit.model, logit.model2, logit.model3)

#Interpret results in document

#Task 2
#Compare Using BIC
BIC(logit.model,logit.model2, logit.model3)

#Interpret results in document

#Task 3 and 4
#Compare using McFadden's and Tjur's

library(DescTools)
```

```
PseudoR2(logit.model, c("McFadden", "Tjur"))
PseudoR2(logit.model2, c("McFadden", "Tjur"))
PseudoR2(logit.model3, c("McFadden", "Tjur"))


#Task 5
#Compare using H-S GOF
library(generalhoslem)
HL1 = logitgof(nmdta$mntlhc2, fitted(logit.model), g = 10)
HL1

HL2 = (logitgof(nmdta$mntlhc2, fitted(logit.model2), g = 10))
HL2

HL3 = (logitgof(nmdta$mntlhc2, fitted(logit.model3), g = 10))
HL3
save(nmdta, file = "Assignment_06.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 Binary Response Variables
> # 1 = poor mental health mntlhl > 0
> useddta$mntlhc2 <- ifelse(useddta$mntlhlth > 0, 1, 0)

> #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(useddta$mntlhlth)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
   0.00    0.00    0.00    3.98    5.00   30.00    4954

> summary(useddta$inc1k)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
```

```
  0.245  12.481  22.605  30.279  37.226 162.607

> summary(useddta$age)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
  18.00   31.00   43.00   45.57   59.00   89.00      18

> table(useddta$female)

   0    1
2480 3226

> table(useddta$white)

   0    1
1062 4644

> # Create 2 logit models
> logit.model <- glm(mntlhc2 ~ age + male + other + black + educ + inc1k,
family = binomial(link = 'logit'),
+                    data = nmdta)

> logit.model2 <- glm(mntlhc2 ~ male + other + black + educ + inc1k, family =
binomial(link = 'logit'),
+                     data = nmdta)

> logit.model3 <- glm(mntlhc2 ~ male + other + black + inc1k, family =
binomial(link = 'logit'),
+                     data = nmdta)

> #Task 1
> #Compare using AIC
> library(stats4)

> AIC(logit.model, logit.model2, logit.model3)
             df      AIC
logit.model   7 1029.969
logit.model2  6 1029.680
logit.model3  5 1028.425

> #Interpret results in document
>
> #Task 2
> #Compare Using BIC
> BIC(logit.model,logit.model2, logit.model3)
             df      BIC
logit.model   7 1062.310
logit.model2  6 1057.401
logit.model3  5 1051.525

> #Interpret results in document
>
> #Task 3 and 4
> #Compare using McFadden's and Tjur's
>
> library(DescTools)

> PseudoR2(logit.model, c("McFadden", "Tjur"))
```

```
  McFadden       Tjur
0.01994656 0.02747576

> PseudoR2(logit.model2, c("McFadden", "Tjur"))
  McFadden       Tjur
0.01829571 0.02514880

> PseudoR2(logit.model3, c("McFadden", "Tjur"))
  McFadden       Tjur
0.01757798 0.02418605

> #Task 5
> #Compare using H-S GOF
> library(generalhoslem)

> HL1 = logitgof(nmdta$mntlhc2, fitted(logit.model), g = 10)

> HL1

      Hosmer and Lemeshow test (binary model)

data:  nmdta$mntlhc2, fitted(logit.model)
X-squared = 6.993, df = 8, p-value = 0.5374


> HL2 = (logitgof(nmdta$mntlhc2, fitted(logit.model2), g = 10))

> HL2

      Hosmer and Lemeshow test (binary model)

data:  nmdta$mntlhc2, fitted(logit.model2)
X-squared = 7.9949, df = 8, p-value = 0.434


> HL3 = (logitgof(nmdta$mntlhc2, fitted(logit.model3), g = 10))

> HL3

      Hosmer and Lemeshow test (binary model)

data:  nmdta$mntlhc2, fitted(logit.model3)
X-squared = 2.561, df = 8, p-value = 0.9588


> save(nmdta, file = "Assignment_06.rdata")

> sink()
```