**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.

>setwd("/Users/burrisfaculty/Desktop/DSCode/SOC686")

>sink("Shepherd\_asgn07.log", split=T)

>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.

> 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

**Task 1 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.

> #Task 1 Create Ordinal Response Variable

> # 1 number of days is 0

> # 2 days 1 to 7

> # 3 days 8-14

> # 4 days 15+

>

> 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

**Task 2 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). Let us call this model Model 1 in the following tasks (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).
* Then run a second proportional odds (cumulative ordered logit) regression of mental health ( mntlhlthOrd) 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 2.

> #Task 2 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 3 Test Hypothesis Using the Likelihood Ratio Test**

* Test whether the effects of education and income are simultaneous equal to zero in Model 1 using the Wald test. Please also provide a brief interpretation of the results.
* Test whether the effects of race are zero in Model 2 using the LR test. Please also provide a brief interpretation of the results.

> #Task 3 Test Hypothesis Using the Likelihood Ratio Test

> #Wald Test

> library(car)

> waldtest = linearHypothesis(ordlogit.model1, c("age = 0","educ = 0" ))

> waldtest

Linear hypothesis test

Hypothesis:

age = 0

educ = 0

Model 1: restricted model

Model 2: mntlhlthOrd ~ age + male + other + black + educ + inc1k

Res.Df Df Chisq Pr(>Chisq)

1 720

2 718 2 2.9446 0.2294

**Wald Test Interpretation:** After completing the test, with a degree of freedom of 2 and a test statistic, we have a p-value of 0.2294, so we would fail to reject the null hypothesis at the 0.05 significance level. This would lead us to conclude there is no significant advantage of using the restricted model over the original model.

> #Test the same hypothesis using LR test

>

> lrTest = anova(ordlogit.model2, ordlogit.model1, test = "Chisq")

> print(lrTest)

Likelihood ratio tests of ordinal regression models

Response: mntlhlthOrd

Model Resid.df Resid.Dev Test Df LR stat Pr(Chi)

1 male + other + black + inc1k 720 1526.105

2 age + male + other + black + educ + inc1k 718 1523.136 1 vs 2 2 2.968299 0.226695

**LR Test Interpretation:** After completing the test, with a degree of freedom of 2 and a test statistic, we have a p-value of 0.2267, so we would fail to reject the null hypothesis at the 0.05 significance level. This would lead us to conclude there is no significant advantage of using the second model (ordlogit.model2) over the original model (ordlogit.model1)

**Task 4 Test Hypothesis Using AIC and BIC**

* Compare models 1-2 using AIC, and interpret the results.
* Compare models 1-2 using BIC, and interpret the results.

> #TASK 4 Test using AIC and BIC

> library(stats4)

> AIC(ordlogit.model1,ordlogit.model2)

df AIC

ordlogit.model1 9 1541.136

ordlogit.model2 7 1540.105

**AIC Interpretation:** From the table below, the second model will fit better because it has a smaller AIC. However, the difference in AICs between model1(1541.136) and model2(1540.105) is less than 2, so it is weak evidence that model2 has a significant advantage over the original model.

> BIC(ordlogit.model1,ordlogit.model2)

df BIC

ordlogit.model1 9 1582.437

ordlogit.model2 7 1572.227

**BIC Interpretation:** By looking strictly at the BIC numbers, the second model would fit the data better because it has the smaller BIC. When we check the difference between the two BIC values, 1582.437 – 1572.227 = 10.21. Since this difference greater than 10, we would conclude there is strong evidence that the second model is the best-fitting model between the two.

**Task 5 Compare Models Using R-squared**

* Compare models 1-2 using the McFadden's pseudo R-squared, and interpret the results.
* Compare models 1-2 using the Nagel's (Nagelkerke / Cragg & Uhler's) pseudo R-squared, and interpret the results.

> #TASK 5 Compare Using Pseudo R-Squared

> library(DescTools)

> PseudoR2(ordlogit.model1, c("McFadden", "Nagel"))

McFadden Nagelkerke

0.01542965 0.03666645

> PseudoR2(ordlogit.model2, c("McFadden", "Nagel"))

McFadden Nagelkerke

0.01351092 0.03217211

**Interpretation:** With both McFadden’s pseudo-R2 and Nagel’s pseudo-R2, we would arrive at the same conclusion that the original model (ordlogit.model1) is the better fitting model because the pseudo-R2 values are greater for the original model than the second model (ordlogit.model2).

**R-File**

#source("/Users/burrisfaculty/Desktop/DSCode/SOC686/Shepherd\_Lab07\_SOC686.r", echo=T, max.deparse.length=10000)

library(foreign)

#Open Log and read in data

setwd("/Users/burrisfaculty/Desktop/DSCode/SOC686")

sink("Shepherd\_asgn07.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)

#Task 1 Create Ordinal Response Variable

# 1 number of days is 0

# 2 days 1 to 7

# 3 days 8-14

# 4 days 15+

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)

#Task 2 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 3 Test Hypothesis Using the Likelihood Ratio Test

#Wald Test

library(car)

waldtest = linearHypothesis(ordlogit.model1, c("age = 0","educ = 0" ))

waldtest

#Test the same hypothesis using LR test

lrTest = anova(ordlogit.model2, ordlogit.model1, test = "Chisq")

print(lrTest)

#TASK 4 Test using AIC and BIC

library(stats4)

AIC(ordlogit.model1,ordlogit.model2)

BIC(ordlogit.model1,ordlogit.model2)

#TASK 5 Compare Using Pseudo R-Squared

library(DescTools)

PseudoR2(ordlogit.model1, c("McFadden", "Nagel"))

PseudoR2(ordlogit.model2, c("McFadden", "Nagel"))

#Close log

save(nmdta, file = "Assignment\_07.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

> #Task 1 Create Ordinal Response Variable

> # 1 number of days is 0

> # 2 days 1 to 7

> # 3 days 8-14

> # 4 days 15+

>

> 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

> #Task 2 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 3 Test Hypothesis Using the Likelihood Ratio Test

> #Wald Test

> library(car)

> waldtest = linearHypothesis(ordlogit.model1, c("age = 0","educ = 0" ))

> waldtest

Linear hypothesis test

Hypothesis:

age = 0

educ = 0

Model 1: restricted model

Model 2: mntlhlthOrd ~ age + male + other + black + educ + inc1k

Res.Df Df Chisq Pr(>Chisq)

1 720

2 718 2 2.9446 0.2294

> #Test the same hypothesis using LR test

>

> lrTest = anova(ordlogit.model2, ordlogit.model1, test = "Chisq")

> print(lrTest)

Likelihood ratio tests of ordinal regression models

Response: mntlhlthOrd

Model Resid. df Resid. Dev Test Df LR stat. Pr(Chi)

1 male + other + black + inc1k 720 1526.105

2 age + male + other + black + educ + inc1k 718 1523.136 1 vs 2 2 2.968299 0.226695

> #TASK 4 Test using AIC and BIC

> library(stats4)

> AIC(ordlogit.model1,ordlogit.model2)

df AIC

ordlogit.model1 9 1541.136

ordlogit.model2 7 1540.105

> BIC(ordlogit.model1,ordlogit.model2)

df BIC

ordlogit.model1 9 1582.437

ordlogit.model2 7 1572.227

> #TASK 5 Compare Using Pseudo R-Squared

> library(DescTools)

> PseudoR2(ordlogit.model1, c("McFadden", "Nagel"))

McFadden Nagelkerke

0.01542965 0.03666645

> PseudoR2(ordlogit.model2, c("McFadden", "Nagel"))

McFadden Nagelkerke

0.01351092 0.03217211

> #Close log

> save(nmdta, file = "Assignment\_07.rdata")

> sink()