**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.

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

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

**Manage Data and Run Logit**

* 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).

> 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

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

+ data = nmdta)

> nmdta$logitpr <- predict(logit.model, type = "response")

> summary(nmdta$logitpr)

Min. 1st Qu. Median Mean 3rd Qu. Max.

0.2479 0.4071 0.4552 0.4680 0.5415 0.6316

**Task 1 Single-Coefficients/Variables Test 1**

* Test whether the effects of age on mental health is zero using the Wald test and interpret the results

> #Task 1

> #Wald Test: Effects of age on mental health

> HT01 = linearHypothesis(logit.model,c("age = 0"))

> HT01

Linear hypothesis test

Hypothesis:

age = 0

Model 1: restricted model

Model 2: mntlhc2 ~ age + male + other + black + educ + inc1k

Res.Df Df Chisq Pr(>Chisq)

1 744

2 743 1 1.7063 0.1915

> #Interpret Results in Assignment

**Conclusion**: With a and a p-value of 0.1915, we would fail to reject the null hypothesis at the usual 0.05 significance level that age has no effect on mental health. Because of this, we could remove age from the model.

**Task 2 Single-Coefficients/Variables Test 2**

* Test whether the effects of race (this is a little bit tricky) on mental health is zero using the score test and interpret the results

> #Task 2

> #Use Score test to see if the effects of race on mental health is 0

> noracelogit.model <- update(logit.model, .~. -other - black)

> anova(noracelogit.model, logit.model, test = "Rao")

Analysis of Deviance Table

Model 1: mntlhc2 ~ age + male + educ + inc1k

Model 2: mntlhc2 ~ age + male + other + black + educ + inc1k

Resid. Df Resid. Dev Df Deviance Rao Pr(>Chi)

1 745 1020.8

2 743 1016.0 2 4.7989 4.7593 0.09258 .

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> #Interpret Results in Assignment

**Conclusion:** With a p-value of 0.09258 we would fail to reject the null hypothesis at the 0.05 significance level. This would lead us to the conclusion that the variables involving race (other and black) do not have an effect on mental health. We could choose to work with the smaller logit model.

**Task 3 Multiple-Coefficients Test 1**

* Test whether the effects of education and income are simultaneous equal to zero using the Wald test and interpret the results.

> #Task 3

> #Use Wald test education and income equal to 0

> HT02 = linearHypothesis(logit.model, c("educ = 0", ("inc1k = 0")))

> HT02

Linear hypothesis test

Hypothesis:

educ = 0

inc1k = 0

Model 1: restricted model

Model 2: mntlhc2 ~ age + male + other + black + educ + inc1k

Res.Df Df Chisq Pr(>Chisq)

1 745

2 743 2 1.2919 0.5242

> #Interpret Results in Assignment

Conclusion: With a degree of freedom equal to 2 and chi-square of 1.2919, we would fail to reject the null hypothesis. So the variables education and income do not have a significant effect in the equation. We could choose to use the restricted model instead.

**Task 4 Multiple-Coefficients Test 2**

* Test whether the effects of education and income are simultaneous equal to zero using the LR test and interpret the results.

> #Task 4

> #Use LR Test for educ = 0 an inc1k = 0

>

> NoEducNoIncLogit.model <- update(logit.model, .~. - educ - inc1k)

> summary(NoEducNoIncLogit.model)

Call:

glm(formula = mntlhc2 ~ age + male + other + black, family = binomial(link = "logit"),

data = nmdta)

Deviance Residuals:

Min 1Q Median 3Q Max

-1.3648 -1.0899 -0.9163 1.1808 1.6304

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) 0.594429 0.266111 2.234 0.025498 \*

age -0.008629 0.005663 -1.524 0.127588

male -0.562113 0.149766 -3.753 0.000175 \*\*\*

other -0.226613 0.273693 -0.828 0.407681

black -0.467232 0.226784 -2.060 0.039375 \*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 1036.6 on 749 degrees of freedom

Residual deviance: 1017.3 on 745 degrees of freedom

AIC: 1027.3

Number of Fisher Scoring iterations: 4

> HT03 = anova(NoEducNoIncLogit.model, logit.model, test = "Chisq")

> HT03

Analysis of Deviance Table

Model 1: mntlhc2 ~ age + male + other + black

Model 2: mntlhc2 ~ age + male + other + black + educ + inc1k

Resid. Df Resid. Dev Df Deviance Pr(>Chi)

1 745 1017.3

2 743 1016.0 2 1.299 0.5223

> #Interpret Results in Assignment

Conclusion, based on the p-value of 0.5223, at the 0.05 significance level, we would fail to reject the null hypothesis that the effect of education and income on mental health could be simultaneously equal to 0.

**Task 5 Coefficients-Equality Test**

* Test whether the effects of education and income are equal using the Wald test and interpret the results.

> #Task 5

> #Test whether the effects of educ and inc are equal using Wald Test

>

> HT04 = linearHypothesis(logit.model, c("educ = inc1k"))

> HT04

Linear hypothesis test

Hypothesis:

educ - inc1k = 0

Model 1: restricted model

Model 2: mntlhc2 ~ age + male + other + black + educ + inc1k

Res.Df Df Chisq Pr(>Chisq)

1 744

2 743 1 0.6917 0.4056

> #Interpret Results in Assignment

Conclusion: At the 0.05 significance level with a p-value of 0.4056, we could not reject the null hypothesis that the effects of income and education on mental health are equal.

**R-Script**

# source("/Users/burrisfaculty/Desktop/DSCode/SOC686/Shepherd\_Lab05\_SOC686.r", echo=T, max.deparse.length=10000)

library(foreign)

library(carData)

library(car)

#Open Log and read in data

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

sink("Shepherd\_asgn05F.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)

logit.model <- glm(mntlhc2 ~ age + male + other + black + educ + inc1k, family = binomial(link = 'logit'),

data = nmdta)

nmdta$logitpr <- predict(logit.model, type = "response")

summary(nmdta$logitpr)

#Task 1

#Wald Test: Effects of age on mental health

HT01 = linearHypothesis(logit.model,c("age = 0"))

HT01

#Interpret Results in Assignment

#Task 2

#Use Score test to see if the effects of race on mental health is 0

noracelogit.model <- update(logit.model, .~. -other - black)

anova(noracelogit.model, logit.model, test = "Rao")

#Interpret Results in Assignment

#Task 3

#Use Wald test education and income equal to 0

HT02 = linearHypothesis(logit.model, c("educ = 0", ("inc1k = 0")))

HT02

#Interpret Results in Assignment

#Task 4

#Use LR Test for educ = 0 an inc1k = 0

NoEducNoIncLogit.model <- update(logit.model, .~. - educ - inc1k)

summary(NoEducNoIncLogit.model)

HT03 = anova(NoEducNoIncLogit.model, logit.model, test = "Chisq")

HT03

#Interpret Results in Assignment

#Task 5

#Test whether the effects of educ and inc are equal using Wald Test

HT04 = linearHypothesis(logit.model, c("educ = inc1k"))

HT04

#Interpret Results in Assignment

save(nmdta, file = "Assignment\_05.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

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

+ data = nmdta)

> nmdta$logitpr <- predict(logit.model, type = "response")

> summary(nmdta$logitpr)

Min. 1st Qu. Median Mean 3rd Qu. Max.

0.2479 0.4071 0.4552 0.4680 0.5415 0.6316

> #Task 1

> #Wald Test: Effects of age on mental health

> HT01 = linearHypothesis(logit.model,c("age = 0"))

> HT01

Linear hypothesis test

Hypothesis:

age = 0

Model 1: restricted model

Model 2: mntlhc2 ~ age + male + other + black + educ + inc1k

Res.Df Df Chisq Pr(>Chisq)

1 744

2 743 1 1.7063 0.1915

> #Interpret Results in Assignment

>

>

> #Task 2

> #Use Score test to see if the effects of race on mental health is 0

> noracelogit.model <- update(logit.model, .~. -other - black)

> anova(noracelogit.model, logit.model, test = "Rao")

Analysis of Deviance Table

Model 1: mntlhc2 ~ age + male + educ + inc1k

Model 2: mntlhc2 ~ age + male + other + black + educ + inc1k

Resid. Df Resid. Dev Df Deviance Rao Pr(>Chi)

1 745 1020.8

2 743 1016.0 2 4.7989 4.7593 0.09258 .

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> #Interpret Results in Assignment

>

> #Task 3

> #Use Wald test education and income equal to 0

> HT02 = linearHypothesis(logit.model, c("educ = 0", ("inc1k = 0")))

> HT02

Linear hypothesis test

Hypothesis:

educ = 0

inc1k = 0

Model 1: restricted model

Model 2: mntlhc2 ~ age + male + other + black + educ + inc1k

Res.Df Df Chisq Pr(>Chisq)

1 745

2 743 2 1.2919 0.5242

> #Interpret Results in Assignment

>

> #Task 4

> #Use LR Test for educ = 0 an inc1k = 0

>

> NoEducNoIncLogit.model <- update(logit.model, .~. - educ - inc1k)

> summary(NoEducNoIncLogit.model)

Call:

glm(formula = mntlhc2 ~ age + male + other + black, family = binomial(link = "logit"),

data = nmdta)

Deviance Residuals:

Min 1Q Median 3Q Max

-1.3648 -1.0899 -0.9163 1.1808 1.6304

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) 0.594429 0.266111 2.234 0.025498 \*

age -0.008629 0.005663 -1.524 0.127588

male -0.562113 0.149766 -3.753 0.000175 \*\*\*

other -0.226613 0.273693 -0.828 0.407681

black -0.467232 0.226784 -2.060 0.039375 \*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 1036.6 on 749 degrees of freedom

Residual deviance: 1017.3 on 745 degrees of freedom

AIC: 1027.3

Number of Fisher Scoring iterations: 4

> HT03 = anova(NoEducNoIncLogit.model, logit.model, test = "Chisq")

> HT03

Analysis of Deviance Table

Model 1: mntlhc2 ~ age + male + other + black

Model 2: mntlhc2 ~ age + male + other + black + educ + inc1k

Resid. Df Resid. Dev Df Deviance Pr(>Chi)

1 745 1017.3

2 743 1016.0 2 1.299 0.5223

> #Interpret Results in Assignment

>

>

> #Task 5

> #Test whether the effects of educ and inc are equal using Wald Test

>

> HT04 = linearHypothesis(logit.model, c("educ = inc1k"))

> HT04

Linear hypothesis test

Hypothesis:

educ - inc1k = 0

Model 1: restricted model

Model 2: mntlhc2 ~ age + male + other + black + educ + inc1k

Res.Df Df Chisq Pr(>Chisq)

1 744

2 743 1 0.6917 0.4056

> #Interpret Results in Assignment

>

> save(nmdta, file = "Assignment\_05.rdata")

> sink()