**Task 1 Run Logit and Calculate Predicted Probabilities for Within-Sample Cases**

* 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). This is a review what we did in our last assignment.
* Use the predict function to create a new variable that has the predicted probabilities for all cases in the estimation sample and then examine the descriptive statistics of this new variable

> logit.model <- glm(mntlhc2 ~ age + male + white + 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 2 Calculate the Predicted Probability for A Hypothetical Individual**

* Calculate the predicted probability of having poor mental health (P(y = 1|x) and its associated precision estimates (hint: use expected predicted probability) for a 35-year-old white female with average (sample mean) education and income. Please interpret the results.

> z.out <-zelig(mntlhc2 ~ age + female + white + black + educ + inc1k, + data = nmdta, model = "logit")

How to cite this model in Zelig:

R Core Team. 2007.

logit: Logistic Regression for Dichotomous Dependent Variables

in Christine Choirat, Christopher Gandrud, James Honaker, Kosuke Imai, Gary King, and Olivia Lau,

"Zelig: Everyone's Statistical Software," http://zeligproject.org/

> x.out <-setx(z.out, age = 35, female = 1, black = 0, white =1, educ = mean(nmdta$educ), inc1k = mean(nmdta$inc1k))

> set.seed(47306)

> s.out <- sim(z.out, x = x.out)

> summary(s.out)

sim x :

-----

ev

mean sd 50% 2.5%

[1,] 0.5683742 0.02923825 0.5696898 0.5136599

97.5%

[1,] 0.6242061

pv

0 1

[1,] 0.462 0.538

> #Interpret results in document

**Interpretation:**

We are 95% confident that the expected predicted probability lies somewhere between 0.514 and 0.624.

**Task 3 Calculate the Difference in Predicted Probabilities**

* Calculate the difference in the predicted probabilities of having poor mental health (P(y = 1|x) and its associated precision estimates (hint: use expected predicted probability) between a 35-year-old white female with average education and income and an otherwise similar male. Please interpret the results.

> x.start <- setx(z.out, age = 35, female = 1, white = 1, black = 0, educ = mean(nmdta$educ), inc1k = mean(nmdta$inc1k) )

> x.end <- setx(z.out, age = 35, female = 0, white = 1, black = 0,

educ = mean(nmdta$educ), inc1k = mean(nmdta$inc1k))

> s.out <- sim(z.out, x = x.start, x1 = x.end)

> summary(s.out)

sim x :

-----

ev

mean sd 50% 2.5%

[1,] 0.5682234 0.02983913 0.5676251 0.5101918

97.5%

[1,] 0.6242775

pv

0 1

[1,] 0.423 0.577

sim x1 :

-----

ev

mean sd 50% 2.5%

[1,] 0.4334699 0.02864437 0.4335726 0.3778286

97.5%

[1,] 0.4881381

pv

0 1

[1,] 0.574 0.426

fd

mean sd 50% 2.5%

[1,] -0.1347535 0.03716771 -0.1341693 -0.2065442

97.5%

[1,] -0.0612533

**Interpretation:**

We are 95% confident that the expected difference between a 35-year-old white female with average education and income and a similar male would be between -0.2065 and -0.0613.

**Task 4 Compute Partial Change/Marginal Effect**

* Calculate the average marginal effect of education and interpret the results.

#Calculate average marginal effects of education

>

> require(margins)

> summary(margins(logit.model))

factor AME SE z p lower

age -0.0018 0.0014 -1.3121 0.1895 -0.0045

black -0.0660 0.0815 -0.8094 0.4183 -0.2257

educ 0.0055 0.0071 0.7773 0.4370 -0.0084

inc1k -0.0006 0.0006 -1.0709 0.2842 -0.0017

male -0.1312 0.0353 -3.7126 0.0002 -0.2005

white 0.0511 0.0664 0.7689 0.4420 -0.0791

` upper

0.0009

0.0938

0.0195

0.0005

-0.0619

0.1812

**Interpretation:**

The marginal effect of education on mental health with covariates held constant for age, black, income, male, and white held at the mean is 0.0055.

**Task 5 Plot Predicted Probabilities**

* Plot the effects display of age for blacks with the sample median income and other covariates set to their corresponding sample averages; that is to plot the effects of age on the predicted probabilities of having poor mental health for blacks with the sample median income and other covariates set to their corresponding sample averages. While displaying the effects of age, please restrict the values of age between 18 to 65. Please explain the probability plot and its confidence bands.

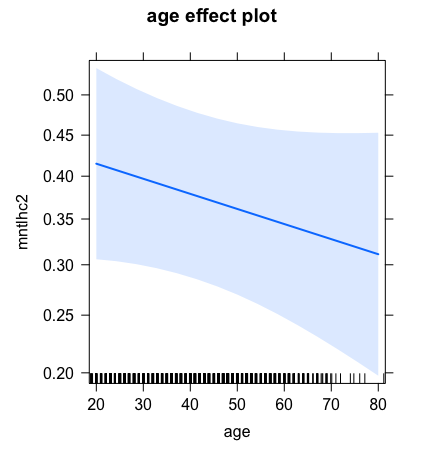
> #Task 5 Plot Predicted Probabilities

>

>

> plot(effect("age", logit.model,xlevels = list(18:65),

+ given.values = c(black = 1, white = 0, inc1k = median(nmdta$inc1k), educ = mean(nmdta$educ)) ))



**Interpretation:**

In the plot above, the 95% confidence intervals of effects for mental health are shown for the range of ages from 18 to 65 for blacks with a median income and average education.

**R Script**

#source("/Users/burrisfaculty/Desktop/DSCode/SOC686/Shepherd\_Lab04\_SOC686.r", echo=T, max.deparse.length=10000)

library(foreign)

library(carData)

library(Zelig)

#Open Long and read in data

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

sink("Shepherd\_asgn04.log", split=T)

rm(list=ls(all=TRUE))

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

library(foreign)

library(carData)

library(Zelig)

#SELECT DATA

useddta <- subset(mygss,

select=c(mntlhlth, age, sex, race, educ, inc1k))

#SUMMARIZE USING lapply instead of tables this time

lapply(useddta[,-6], table, useNA="ifany")

lapply(useddta, summary, na.rm=T)

lapply(useddta, mean, na.rm=T)

#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)

#Graph Bivariate Scatter Plot

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

#Create Scatter plot matrix

#scatterplotMatrix(~ mntlhlth + age + sex + race +

# educ + inc1k + mntlhc2,

# smooth = list(span = 0.7), data = nmdta)

#scatterplot(nmdta$educ,nmdta$mntlhc2)

#Task 1 Run Logit and Calculate Predicted Probabilities for Within Sample Cases

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

data = nmdta)

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

summary(nmdta$logitpr)

#Task 2 Calculate the Predicted Probability for a Hypothetical Individual

# 35 year-old white woman with average education and income

z.out <-zelig(mntlhc2 ~ age + female + white + black + educ + inc1k,

data = nmdta, model = "logit")

x.out <-setx(z.out, age = 35, female = 1, black = 0, white = 1, educ = mean(nmdta$educ), inc1k = mean(nmdta$inc1k))

set.seed(47306)

s.out <- sim(z.out, x = x.out)

summary(s.out)

#Interpret results in document

#Task 3 Calculate the Difference in Predicted Probabilities

x.start <- setx(z.out, age = 35, female = 1, white = 1, black = 0, educ = mean(nmdta$educ), inc1k = mean(nmdta$inc1k) )

x.end <- setx(z.out, age = 35, female = 0, white = 1, black = 0, educ = mean(nmdta$educ), inc1k = mean(nmdta$inc1k))

s.out <- sim(z.out, x = x.start, x1 = x.end)

summary(s.out)

#Interpret results in document

#Task 4 Computer Partial Change/Marginal Effect

#Calculate average marginal effects of education

require(margins)

summary(margins(logit.model))

#Interpret results in document

#Task 5 Plot Predicted Probabilities

plot(effect("age", logit.model,xlevels = list(18:65),

given.values = c(black = 1, white = 0, inc1k = median(nmdta$inc1k), educ = mean(nmdta$educ)) ))

save(nmdta, file = "Assignment\_03.rdata")

sink()

**Log**

> library(carData)

> library(Zelig)

> #SELECT DATA

> useddta <- subset(mygss,

+ select=c(mntlhlth, age, sex, race, educ, inc1k))

> #SUMMARIZE USING lapply instead of tables this time

> lapply(useddta[,-6], table, useNA="ifany")

$mntlhlth

0 1 2 3 4 5 6 7 8 10

401 34 62 37 29 39 6 19 2 35

12 14 15 16 18 20 21 25 27 30

3 4 22 1 2 21 2 9 1 23

<NA>

4954

$age

18 19 20 21 22 23 24 25 26 27

27 92 91 93 80 119 111 123 114 125

28 29 30 31 32 33 34 35 36 37

151 124 118 117 143 126 136 114 131 107

38 39 40 41 42 43 44 45 46 47

121 91 102 111 95 123 115 83 112 89

48 49 50 51 52 53 54 55 56 57

99 117 99 91 98 76 80 80 73 72

58 59 60 61 62 63 64 65 66 67

74 69 81 70 77 78 71 74 58 72

68 69 70 71 72 73 74 75 76 77

67 71 62 49 58 49 54 37 37 43

78 79 80 81 82 83 84 85 86 87

46 25 21 26 9 23 22 21 16 14

88 89 <NA>

10 35 18

$sex

1 2

2480 3226

$race

1 2 3

4644 770 292

$educ

0 1 2 3 4 5 6 7 8 9

20 7 15 25 33 30 85 90 251 213

10 11 12 13 14 15 16 17 18 19

216 350 1817 479 580 249 679 167 189 91

20 <NA>

102 18

> lapply(useddta, summary, na.rm=T)

$mntlhlth

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

0.00 0.00 0.00 3.98 5.00 30.00

NA's

4954

$age

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

18.00 31.00 43.00 45.57 59.00 89.00

NA's

18

$sex

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

1.000 1.000 2.000 1.565 2.000 2.000

$race

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

1.000 1.000 1.000 1.237 1.000 3.000

$educ

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

0.0 12.0 12.0 12.7 15.0 20.0

NA's

18

$inc1k

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

0.245 12.481 22.605 30.279 37.226 162.607

> lapply(useddta, mean, na.rm=T)

$mntlhlth

[1] 3.980053

$age

[1] 45.57331

$sex

[1] 1.56537

$race

[1] 1.237294

$educ

[1] 12.69849

$inc1k

[1] 30.27912

> #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)

> #Graph Bivariate Scatter Plot

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

> #Create Scatter plot matrix

>

> #scatterplotMatrix(~ mntlhlth + age + sex + race +

> # educ + inc1k + mntlhc2,

> # smooth = list(span = 0.7), data = nmdta)

> #scatterplot(nmdta$educ,nmdta$mntlhc2)

>

> #Task 1 Run Logit and Calculate Predicted Probabilities for Within Sample Cases

>

> logit.model <- glm(mntlhc2 ~ age + male + white + 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 2 Calculate the Predicted Probability for a Hypothetical Individual

> # 35 year-old white woman with average education and income

>

>

> z.out <-zelig(mntlhc2 ~ age + female + white + black + educ + inc1k,

+ data = nmdta, model = "logit")

How to cite this model in Zelig:

R Core Team. 2007.

logit: Logistic Regression for Dichotomous Dependent Variables

in Christine Choirat, Christopher Gandrud, James Honaker, Kosuke Imai, Gary King, and Olivia Lau,

"Zelig: Everyone's Statistical Software," http://zeligproject.org/

> x.out <-setx(z.out, age = 35, female = 1, black = 0, white = 1, educ = mean(nmdta$educ), inc1k = mean(nmdta$inc1k))

> set.seed(47306)

> s.out <- sim(z.out, x = x.out)

> summary(s.out)

sim x :

-----

ev

mean sd 50% 2.5%

[1,] 0.5683742 0.02923825 0.5696898 0.5136599

97.5%

[1,] 0.6242061

pv

0 1

[1,] 0.462 0.538

> #Interpret results in document

>

> #Task 3 Calculate the Difference in Predicted Probabilities

> x.start <- setx(z.out, age = 35, female = 1, white = 1, black = 0, educ = mean(nmdta$educ), inc1k = mean(nmdta$inc1k) )

> x.end <- setx(z.out, age = 35, female = 0, white = 1, black = 0, educ = mean(nmdta$educ), inc1k = mean(nmdta$inc1k))

> s.out <- sim(z.out, x = x.start, x1 = x.end)

> summary(s.out)

sim x :

-----

ev

mean sd 50% 2.5%

[1,] 0.5682234 0.02983913 0.5676251 0.5101918

97.5%

[1,] 0.6242775

pv

0 1

[1,] 0.423 0.577

sim x1 :

-----

ev

mean sd 50% 2.5%

[1,] 0.4334699 0.02864437 0.4335726 0.3778286

97.5%

[1,] 0.4881381

pv

0 1

[1,] 0.574 0.426

fd

mean sd 50% 2.5%

[1,] -0.1347535 0.03716771 -0.1341693 -0.2065442

97.5%

[1,] -0.0612533

> #Interpret results in document

>

> #Task 4 Computer Partial Change/Marginal Effect

> #Calculate average marginal effects of education

>

> require(margins)

> summary(margins(logit.model))

factor AME SE z p lower

age -0.0018 0.0014 -1.3121 0.1895 -0.0045

black -0.0660 0.0815 -0.8094 0.4183 -0.2257

educ 0.0055 0.0071 0.7773 0.4370 -0.0084

inc1k -0.0006 0.0006 -1.0709 0.2842 -0.0017

male -0.1312 0.0353 -3.7126 0.0002 -0.2005

white 0.0511 0.0664 0.7689 0.4420 -0.0791

upper

0.0009

0.0938

0.0195

0.0005

-0.0619

0.1812

> #Interpret results in document

>

> #Task 5 Plot Predicted Probabilities

>

>

> plot(effect("age", logit.model,xlevels = list(18:65),

+ given.values = c(black = 1, white = 0, inc1k = median(nmdta$inc1k), educ = mean(nmdta$educ)) ))

> save(nmdta, file = "Assignment\_04.rdata")

> sink()