> # Author: William Eddy, M.A. Candidate East Carolina University, william.sebastian.eddy@gmail.com

> # Alternative name provided by Professor Poopypants algorithm: Booger Burger? Go Poopypants.

>

>

> ##########

> ##

> ## Use these packages

> ##

> ##########

> library(ggplot2); library(cowplot); library(rstudioapi); library(compareGroups)

> library(tidyverse); library(psych); library(mdscore); library(car)

>

> ## Set working directiory where this R file is

> setwd(dirname(rstudioapi::getActiveDocumentContext()$path))

> #Import raw data

> data <- read.table("../data/Log-06.dat")

> # Name columns

> colnames(data) <- c("SOCPROB","REPEAT","ADDSC","DROPOUT")

>

> ###########

> ##

> ## Test linearity of the logit using the Box-Tidwell test

> ## Done before labelling dichotomous variable values

> ##

> ###########

>

> library(car)

> boxTidwell(DROPOUT ~ ADDSC, ~ SOCPROB + REPEAT, data=data)

MLE of lambda Score Statistic (z) Pr(>|z|)

1.6168 0.3226 0.747

iterations = 5

>

> ##########

> ##

> ## Implement data labels and categorize dichotomous variables as factors

> ##

> ##########

>

> # Label variable values - this also sets them as dichotomous variables (factors in r)

> data$SOCPROB <- factor(data$SOCPROB,

+ levels = c(0,1),

+ labels = c("0 - Yes 9th grade social problems", "1 - No 9th grade social problems"))

>

> data$REPEAT <- factor(data$REPEAT,

+ levels = c(0,1),

+ labels = c("0 - Didn't repeat a grade", "1 - Repeated a grade"))

>

> data$DROPOUT <- factor(data$DROPOUT,

+ levels = c(0,1),

+ labels = c("0 - Completed high school", "1 - Dropped out of high school"))

>

>

>

> ###########

> ##

> ## Create and view the intercept-only model

> ##

> ###########

>

> LM1 <- glm(DROPOUT ~ 1, data=data, family="binomial")

> summary(LM1)

Call:

glm(formula = DROPOUT ~ 1, family = "binomial", data = data)

Deviance Residuals:

Min 1Q Median 3Q Max

-0.4493 -0.4493 -0.4493 -0.4493 2.1649

Coefficients:

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

(Intercept) -2.2425 0.2147 -10.45 <2e-16 \*\*\*

---

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

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 158.1 on 249 degrees of freedom

Residual deviance: 158.1 on 249 degrees of freedom

AIC: 160.1

Number of Fisher Scoring iterations: 5

> # Return -2 Log Liklihood

> LM1\_minus2log <- -2\*(logLik(LM1))

>

>

>

> #####################################

> ##

> ## Create and view logistic regression using all variables to predict DROPOUT

> ##

> #####################################

>

> LM2 <- glm(DROPOUT ~ ., data=data, family="binomial")

> summary(LM2)

Call:

glm(formula = DROPOUT ~ ., family = "binomial", data = data)

Deviance Residuals:

Min 1Q Median 3Q Max

-1.2701 -0.4338 -0.3277 -0.2417 2.5511

Coefficients:

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

(Intercept) -5.64662 1.24657 -4.530 5.91e-06 \*\*\*

SOCPROB1 - No 9th grade social problems 1.06114 0.61209 1.734 0.08298 .

REPEAT1 - Repeated a grade 1.41725 0.48413 2.927 0.00342 \*\*

ADDSC 0.05287 0.02140 2.470 0.01350 \*

---

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

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 158.10 on 249 degrees of freedom

Residual deviance: 136.19 on 246 degrees of freedom

AIC: 144.19

Number of Fisher Scoring iterations: 6

> anova(LM2, test="Chisq")

Analysis of Deviance Table

Model: binomial, link: logit

Response: DROPOUT

Terms added sequentially (first to last)

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

NULL 249 158.10

SOCPROB 1 4.0785 248 154.02 0.0434323 \*

REPEAT 1 11.3320 247 142.69 0.0007618 \*\*\*

ADDSC 1 6.4972 246 136.19 0.0108046 \*

---

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

> ?anova

> # Return -2 Log Liklihood

> LM2\_minus2log <--2\*(logLik(LM2))

>

> # Return odds ratios

> exp(LM2$coefficients)

(Intercept) SOCPROB1 - No 9th grade social problems REPEAT1 - Repeated a grade

0.003529429 2.889671894 4.125746804

ADDSC

1.054290272

>

> ## Now calculate the overall "Pseudo R-squared" and its p-value

> ll.null.2 <- LM2$null.deviance/-2

> ll.proposed.2 <- LM2$deviance/-2

>

> ## McFadden's Pseudo R^2 = [ LL(Null) - LL(Proposed) ] / LL(Null)

> McFaddenR2 <- (ll.null.2 - ll.proposed.2) / ll.null.2

> McFaddenR2

[1] 0.1385666

> sqrt(McFaddenR2)

[1] 0.3722453

>

> ## The p-value for the R^2 above

> 1 - pchisq(2\*(ll.proposed.2 - ll.null.2), df=(length(LM2$coefficients)-1))

[1] 6.818136e-05

>

> # Create confusion matrix; Calculate sensitivity and specificity

>

> threshold=0.1

> predicted\_values<-ifelse(predict(LM2,type="response")>threshold,1,0)

> actual\_values<-data$DROPOUT

> conf\_matrix<-table(predicted\_values,actual\_values)

> conf\_matrix

actual\_values

predicted\_values 0 - Completed high school 1 - Dropped out of high school

0 172 11

1 54 13

>

> 13/(24) # % dropouts correctly predicted - sensitivity

[1] 0.5416667

> 172/(226) # % non-dropouts correctly predicted - specificity

[1] 0.7610619

>

>

>

> #####################################

> ##

> ## Create models without each variable and calculate delta(-2 log likelihood)

> ##

> #####################################

>

> LM3 <- glm(DROPOUT ~ SOCPROB + REPEAT, data=data, family="binomial")

> # Return -2 Log Liklihood

> LM3\_minus2log <- -2\*(logLik(LM3))

> # Find delta(-2 log likelihood) for ADDSC

> LM3\_minus2log - LM2\_minus2log

'log Lik.' 6.497175 (df=3)

>

> LM4 <- glm(DROPOUT ~ SOCPROB + ADDSC, data=data, family="binomial")

> # Return -2 Log Liklihood

> LM4\_minus2log <- -2\*(logLik(LM4))

> # Find delta(-2 log likelihood) for REPEAT

> LM4\_minus2log - LM2\_minus2log

'log Lik.' 7.963908 (df=3)

>

> LM5 <- glm(DROPOUT ~ ADDSC + REPEAT, data=data, family="binomial")

> # Return -2 Log Liklihood

> LM5\_minus2log <- -2\*(logLik(LM5))

> # Find delta(-2 log likelihood) for SOCPROB

> LM5\_minus2log - LM2\_minus2log

'log Lik.' 2.702681 (df=3)

>

>

>

> ##################

> ##

> ## Plot the logistic regression

> ##

> ##################

>

> predicted.data <- data.frame(

+ probability.of.DROPOUT=LM2$fitted.values,

+ DROPOUT=data$DROPOUT)

> predicted.data <- predicted.data[

+ order(predicted.data$probability.of.DROPOUT, decreasing=FALSE),]

> predicted.data$rank <- 1:nrow(predicted.data)

>

> ggplot(data=predicted.data, aes(x=rank, y=probability.of.DROPOUT)) +

+ geom\_point(aes(color=DROPOUT), alpha=1, shape=4, stroke=2) +

+ xlab("Probability rank") +

+ ylab("Predicted probability of dropping out of high school")

>

> Save plot as png / pdf

Error: unexpected symbol in " Save plot"

> ggsave("Dropout\_Probabilities.png")

Saving 7.81 x 10.8 in image

>

> ##################

> ##

> ## Use CompareGroups package to show percentages across DROPOUT variable

> ##

> ## CompareGroups package Output includes some tests where results

> ## differ from better understood individual tests in next section

> ## and so are ignored

> ##

> ## Conduct t test and chi square comparisons

> ##

> ##################

>

> allgroups\_compare <- compareGroups(DROPOUT ~ . , data = data)

Warning messages:

1: In chisq.test(xx, correct = correction) :

Chi-squared approximation may be incorrect

2: In chisq.test(xx, correct = correction) :

Chi-squared approximation may be incorrect

> allgroups\_compare

-------- Summary of results by groups of 'DROPOUT'---------

var N p.value method selection

1 SOCPROB 250 0.038\*\* categorical ALL

2 REPEAT 250 0.001\*\* categorical ALL

3 ADDSC 250 <0.001\*\* continuous normal ALL

-----

Signif. codes: 0 '\*\*' 0.05 '\*' 0.1 ' ' 1

> summary(allgroups\_compare)

--- Descriptives of each row-variable by groups of 'DROPOUT' ---

-------------------

row-variable: SOCPROB

0 - Yes 9th grade social problems 1 - No 9th grade social problems

[ALL] 229 21

0 - Completed high school 210 16

1 - Dropped out of high school 19 5

0 - Yes 9th grade social problems (row%) 1 - No 9th grade social problems (row%) p.overall

[ALL] 91.6 8.4

0 - Completed high school 92.92035 7.079646 0.037593

1 - Dropped out of high school 79.16667 20.83333

OR OR.lower OR.upper

0 - Yes 9th grade social problems 1

1 - No 9th grade social problems 3.484965 1.024386 10.19516

-------------------

row-variable: REPEAT

0 - Didn't repeat a grade 1 - Repeated a grade 0 - Didn't repeat a grade (row%)

[ALL] 213 37 85.2

0 - Completed high school 199 27 88.0531

1 - Dropped out of high school 14 10 58.33333

1 - Repeated a grade (row%) p.overall

[ALL] 14.8

0 - Completed high school 11.9469 0.000669

1 - Dropped out of high school 41.66667

OR OR.lower OR.upper

0 - Didn't repeat a grade 1

1 - Repeated a grade 5.225922 2.048259 13.01565

-------------------

row-variable: ADDSC

N mean sd lower upper p.overall

[ALL] 250 51.78 11.71771 50.32039 53.23961

0 - Completed high school 226 50.99115 11.66304 49.46236 52.51994 0.000487

1 - Dropped out of high school 24 59.20833 9.596101 55.15626 63.26041

OR OR.lower OR.upper

[1,] 1.066966 1.025123 1.110517

>

> allgroups\_table <- createTable(allgroups\_compare, show.ratio=TRUE)

> allgroups\_table

--------Summary descriptives table by 'DROPOUT'---------

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

0 - Completed high school 1 - Dropped out of high school OR p.ratio p.overall

N=226 N=24

¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯

SOCPROB: 0.038

0 - Yes 9th grade social problems 210 (92.9%) 19 (79.2%) Ref. Ref.

1 - No 9th grade social problems 16 (7.08%) 5 (20.8%) 3.48 [1.02;10.2] 0.046

REPEAT: 0.001

0 - Didn't repeat a grade 199 (88.1%) 14 (58.3%) Ref. Ref.

1 - Repeated a grade 27 (11.9%) 10 (41.7%) 5.23 [2.05;13.0] 0.001

ADDSC 51.0 (11.7) 59.2 (9.60) 1.07 [1.03;1.11] 0.001 <0.001

¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯¯

>

> # Some descriptive statistics on DROPOUT by group

> describe(Dropouts$ADDSC)

Error in describe(Dropouts$ADDSC) : object 'Dropouts' not found

> describe(Graduates$ADDSC)

Error in describe(Graduates$ADDSC) : object 'Graduates' not found

>

> # export2csv(allgroups\_table, "../tables/allgroups\_table.csv")

>

> # Independent samples t-test - ADDSC by DROPOUT groups

>

> # Standard deviations are similar, but N is different - reported separate variance t-test p value but

> # confidence interval from shared variance test as suggested by professor

> Dropouts <- subset(data, DROPOUT == "1 - Dropped out of high school")

> Graduates <- subset(data, DROPOUT == "0 - Completed high school")

> # Equal variances

> t.test(Dropouts$ADDSC, Graduates$ADDSC, var.equal = TRUE)

Two Sample t-test

data: Dropouts$ADDSC and Graduates$ADDSC

t = 3.332, df = 248, p-value = 0.0009938

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

3.35994 13.07443

sample estimates:

mean of x mean of y

59.20833 50.99115

> # Separate variances

> t.test(Dropouts$ADDSC, Graduates$ADDSC, var.equal = FALSE)

Welch Two Sample t-test

data: Dropouts$ADDSC and Graduates$ADDSC

t = 3.9002, df = 30.705, p-value = 0.0004874

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

3.918582 12.515784

sample estimates:

mean of x mean of y

59.20833 50.99115

>

> # Chi squared tests

> chisq.test(data$REPEAT, data$DROPOUT, correct = FALSE)

Pearson's Chi-squared test

data: data$REPEAT and data$DROPOUT

X-squared = 15.197, df = 1, p-value = 9.684e-05

Warning message:

In chisq.test(data$REPEAT, data$DROPOUT, correct = FALSE) :

Chi-squared approximation may be incorrect

> chisq.test(data$SOCPROB, data$DROPOUT, correct = FALSE)

Pearson's Chi-squared test

data: data$SOCPROB and data$DROPOUT

X-squared = 5.3339, df = 1, p-value = 0.02091

Warning message:

In chisq.test(data$SOCPROB, data$DROPOUT, correct = FALSE) :

Chi-squared approximation may be incorrect

> ?chisq.test

>

> ##################

> ##

> ## Other tests - Not all reported

> ##

> ##################

>

> # Likelihood ratio test

> library(lmtest)

> lrtest(LM1, LM2)

Likelihood ratio test

Model 1: DROPOUT ~ 1

Model 2: DROPOUT ~ SOCPROB + REPEAT + ADDSC

#Df LogLik Df Chisq Pr(>Chisq)

1 1 -79.051

2 4 -68.097 3 21.908 6.818e-05 \*\*\*

---

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

>

> # Wald test (FAILED TO IMPLEMENT)

> library(survey)

> regTermTest(LM2, ~DROPOUT +., data=data)

Error in regTermTest(LM2, ~DROPOUT + ., data = data) :

unused argument (data = data)

>

> # Hosmer-Lemeshow (likely implemented incorrectly rendering result unusable)

> library(ResourceSelection)

> Hosmer\_Lem <- hoslem.test(data$DROPOUT,LM2$fitted.values)

Warning message:

In Ops.factor(1, y) : ‘-’ not meaningful for factors

> Hosmer\_Lem

Hosmer and Lemeshow goodness of fit (GOF) test

data: data$DROPOUT, LM2$fitted.values

X-squared = 250, df = 8, p-value < 2.2e-16

>

> # Use anova function to compare the two models

> anova(LM1, LM2, test="Chisq")

Analysis of Deviance Table

Model 1: DROPOUT ~ 1

Model 2: DROPOUT ~ SOCPROB + REPEAT + ADDSC

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

1 249 158.10

2 246 136.19 3 21.908 6.818e-05 \*\*\*

---

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

>

> # Power analysis

> library(pwr)

> pwr.f2.test(1, 249, .08, .05)

Multiple regression power calculation

u = 1

v = 249

f2 = 0.08

sig.level = 0.05

power = 0.993856

> pwr.t.test(n=250, sig.level=.05, power=.08)

Two-sample t test power calculation

n = 250

d = 0.04548517

sig.level = 0.05

power = 0.08

alternative = two.sided

NOTE: n is number in \*each\* group

> pwr.chisq.test(N = 250, df = 1, sig.level = 0.05, power = .08)

Chi squared power calculation

w = 0.03209176

N = 250

df = 1

sig.level = 0.05

power = 0.08

NOTE: N is the number of observations

>

> exp(1.06114)

[1] 2.889663

>

> ##################

> ##

> ## Evaluating the model

> # Wald, Hosmer, McFadden model tests

> # https://www.r-bloggers.com/evaluating-logistic-regression-models/

> ##

> ## Power analysis in R https://www.statmethods.net/stats/power.html

> ##

> ##################

> install.packages("sink")

WARNING: Rtools is required to build R packages but is not currently installed. Please download and install the appropriate version of Rtools before proceeding:

https://cran.rstudio.com/bin/windows/Rtools/

Installing package into ‘C:/Users/Sunil/Documents/R/win-library/3.6’

(as ‘lib’ is unspecified)

Warning in install.packages :

package ‘sink’ is not available (for R version 3.6.0)

> ?sink

> sink(output.txt)

Error in sink(output.txt) : object 'output.txt' not found

> sink("output.txt")