

```

#Name: Lauren Done
#Homework -- Midterm Exam Questions

#Question 1

#H(A): People with at least an associates degree and that live in the
Northeast + Midwest have higher vaccination rates than people with at least
an associates degree that live in the South + West.
#H(o): People with at least an associates degree and that live in the
Northeast + Midwest do not have high vaccination rate than people with with
at least an associates degree that live in the South + West.

attach(Household_Pulse_data)
Df <- Household_Pulse_data
detach
attach(Df)
#Step 1... Look at data

typeof(Df$RRACE)
#integer
typeof(Df$REGION)
#integer
typeof(Df$RECDVACC)
#integer

#Create alt dataset where "N/A" answers in RECDVACC are considered "NO".

#Count # of NAs

summary(Df$RECDVACC)
#851 NAs

levels(Df$RECDVACC)
#3 levels
class(Df$RECDVACC)
#is factor

Df$RECDVACC[Df$RECDVACC=="NA"] <- "no did not get vaxx"

#check for NAs
summary(Df$RECDVACC)

#Now all "NAs" are "no did not get vaxx"
#Next, we create subsets of data for the variables we want to observe.

levels(Df$REGION)

#Change Vax status to TRUE and FALSE -- Logical values for subset.
levels(Df$EEDUC)
levels(Df$RECDVACC) <- c(FALSE,TRUE,FALSE)

class(RRACE)
summary(EEDUC)

```

```

class(Df$EEDUC)

Df$EEDUC <- as.numeric(factor(Df$EEDUC, levels = c("less than hs", "some
hs", "HS diploma", "some coll", "assoc deg", "bach deg", "adv deg")))
#Therefore, (Less than hs = 1, Some HS = 2, HS diploma = 3, Some college =
4, Assoc Deg = 5, Bach deg = 6, Adv deg = 7)

class(Df$EEDUC)
Atleast_Asoc <- Df[Df$EEDUC >= 5,]
-----

testAA <- Atleast_Asoc

#group vectors of REGION together using levels ... There South and West US
inhabitants are not classified as SouthWest and Northeast and Midwest
inhabitants are now classified as Northeast_Midwest.
levels(testAA$REGION) <- c("Northeast_Midwest", "South_West",
"Northeast_Midwest", "South_West")
testAA$EEDUC <- as.factor(testAA[testAA$EEDUC])

class(testAA$EEDUC)

model <- lm(testAA$RECVDVACC == "yes got vaxx" ~ testAA$REGION )
summary(model)
stargazer(model, title = "Likelihood of Vaccination Among College
Educated", type = "text", dep.var.labels = "Vaccination Rate")

#Provide: Estimate, standard error, t-stat and a p-value for the hypothesis
test and a confidence interval.
summary(model)

#Estimate:
#intercept -- 0.9155 , South_West -- -0.0075

#Standard error: intercept -- .0022 , South_West -- .0028

#t-stat: intercept -- 406.95 , South_West -- -2.698

#p-value: .00698

confint(model)

#Using the vaccination rates of people residing in the North and Midwestern
regions of the US as the standard of comparison, people residing in the
Southern and Western regions are about -.75% less likely to be vaccinated.
#However, the the findings are significant enough to allow us to conclude
that we can reject the null hypothesis and expect a slight difference in
vaccination rates between people with college degrees living in
Southern+Western US and Northern_Midwestern US.

#Midterm Question #2

```

```

#For these questions I tried a different approach and format.

#Ha: Females that have bachelors degrees and advanced degrees have higher
vaccination rates among themselves when compared to men with the same
educational level.
#Ho: There is no difference in vaccination rates between genders.

#create a new data to work with-- to avoid screwing with original data set

attach(Household_Pulse_data)
levels(EEDUC)
use_varb <- (!GENID_DESCRIBE == "NA") & (!EEDUC == "less than hs") & (!
EEDUC == "some hs") & (!EEDUC == "HS diploma") & (!EEDUC == "some coll")

dat_use <- subset(Household_Pulse_data, use_varb)

detach()
attach(dat_use)
#Testing for errors in the subgroups
summary(GENID_DESCRIBE)
summary(EEDUC)
summary(RECVDVACC)
#Lets change all NAs to "No" in the data set for whether someone received
vaccinations or not.
dat_use$RECVDVACC[dat_use$RECVDVACC == "NA"] <- "no did not get vaxx"
summary(dat_use$RECVDVACC)

#Great, now my data only contains individuals I we want to observe to test
our hypothesis.
#
#
#Let's create a linear regression model ---

model_ver1 <- lm(RECVDVACC == "yes got vaxx" ~ -1 + GENID_DESCRIBE, data =
dat_use)
#see dummy variable trap notes to understand why I put -1 in lm formula.
Essentially it was to remove the intercept of the mean of all the variables
so that I can see "males". Before the lm calculation was removing males.
summary(model_ver1)

require(stargazer)
#stargazer package is loaded
stargazer(model_ver1, type = "text")

#Confidence intervals calculations
confint(model_ver1)
sd(model_ver1)
summary(model_ver1)

```

#According to this data, people that are male at the most likely to have the highest portion of thier population vaccinated. The vaccination rate amoungst men with each test falls between 92.8% and 93.2%...
#The findings were seen to be significant, to where we can reject the null hypothesis. There is likely to be an onservable difference at least 97.3% of the time.

#Question Number 3

#For question number 3, I'd like to assess the gravity of child hunger and which state has the highest rates of child hunger. I'd also like to examine the racial make up and marital statuses of the households experiencing the highest rates of child hunger.
#This is to see if there is a correlation between location, race or marital status and child hunger.

```
#CHILDFOOD
#EST_ST
#RRACE
#RHISPANIC
```

```
detach()
attach(
  Household_Pulse_data
)
```

```
use_varb3 <- (!CHILDFOOD== "kids got enough food") & (!CHILDFOOD== "NA")
```

```
dat_use3 <- subset(Household_Pulse_data, use_varb3)
```

```
use_varb3_logical <- as.logical(use_varb3)
dat_use_3_logical <- subset(Household_Pulse_data,use_varb3_logical)
```

```
detach()
attach(dat_use3)
```

#check for obvious errors and NAs in data

```
summary(CHILDFOOD)
summary(EST_ST)
summary(RRACE)
summary(RHISPANIC)
levels(CHILDFOOD)
```

#Linear Regression Model

```
regress1 <- lm(CHILDFOOD == "often kids not eating enough because couldnt afford" ~ 0+ (RRACE) + (INCOME) + (EST_ST))
regress2 <- lm(CHILDFOOD == "sometimes kids not eating enough" ~ 0+ (RRACE) + (INCOME) + (EST_ST))
library(stargazer)
```

```
stargazer(regress1, type = "text", title = "Independant Variables Affecting Childhood Hunger")
```

```
stargazer(regress2, type = "text", title = "Independant Variables Affecting  
Childhood Hunger Ver. 2")
```

```
install.packages("AER")  
require(AER)
```

```
detach()
```

```
attach(dat_use3)  
#Plotting the subset  
NNobs <- length(CHILDFOOD)  
set.seed(12345)  
graph_obs <- (runif(NNobs) > 0.1)  
dat_graph <- subset(dat_use3,graph_obs)
```

```
plot(CHILDFOOD ~ RRACE + INCOME, pch = 16, col = rgb(0.5,0.5,0.5, alpha =  
0.2), data = dat_graph)  
plot(CHILDFOOD ~ RRACE + INCOME, pch = 16, col = rgb(0.5,0.5,0.5, alpha =  
0.2), ylim = c(0,1),data = dat_graph)
```

```
#Change line to fit regression  
levels(RRACE)  
to_be_predicted2 <- data.frame(RRACE == "White", INCOME=="HH income $25k -  
$34.9k" )
```

```
to_be_predicted2$yhat <- predict(regress1, newdata = to_be_predicted2)
```

```
lines(yhat ~ RRACE, data = to_be_predicted2)
```

```
#My model seemed to find that although a majority of the families  
experiencing child hunger were White, that of these families did not make  
25k or less. The model predicted them making more than that on avg.
```