**R Script**

**Cleaning / Deleting / Renaming the Columns**

Install “readr”

1-# import the dataset to RStudio by running this script:

obesity <- read\_csv("~/Zeynep/~UMD\_Grad/INFM600/obesity.csv")

2- View(obesity)

3- # check if the columns YearStart and YearEnd are the same (find the difference)

obesity$year\_start\_end\_diff <- obesity$YearEnd - obesity$YearStart

unique(obesity$year\_start\_end\_diff)

4- # delete column YearEnd

obesity$YearEnd <- NULL

5- # rename column YearStart as Year

names(obesity)[names(obesity) == 'YearStart'] <- 'Year'

6- # delete column LocationAbbr is the abbreviations of the state names which is unnecessary information. There is another column called “LocationDesc” which is the state names that we want to use and keep.

obesity$LocationAbbr <- NULL

7- # delete column Datasource. We know that all the entries are the same, and we know that our data source is survey.

obesity$Datasource <- NULL

8- # delete column Class. Class had three sub-categories (Fruits and vegetables, Obesity/Weight Status, Physical Activity) since we are not interested in Class, we are interested in specific questions.

obesity$Class <- NULL

9- # delete column Topic. Topic had three sub-categories (Fruits and vegetables, Obesity/Weight Status, Physical Activity) since we are not interested in Topic, we are interested in specific questions.

obesity$Topic <- NULL

10- # check if the values are same for Data\_Value and Data\_Value\_Alt

obesity$data\_value\_diff <- obesity$Data\_Value - obesity$Data\_Value\_Alt

unique(obesity$data\_value\_diff)

11- # the values are same for Data\_Value and Data\_Value\_Alt; so, delete Data\_Value\_Alt

obesity$Data\_Value\_Alt <- NULL

12- View(obesity)

13- # check if there is any data other than "Value" in Data\_Value\_Type

unique(obesity$Data\_Value\_Type)

14- # there is no data other than "Value" in Data\_Value\_Type; so, delete column Data\_Value\_Type

obesity$Data\_Value\_Type <- NULL

15- # check if there is any useful data in Data\_Value\_Unit, Data\_Value\_Footnote, and Data\_Value\_Footnote\_Symbol by running the following scripts:

unique(obesity$Data\_Value\_Unit)

unique(obesity$Data\_Value\_Footnote)

unique(obesity$Data\_Value\_Footnote\_Symbol)

Or this:

unique(c(obesity$Data\_Value\_Unit, obesity$Data\_Value\_Footnote, obesity$Data\_Value\_Footnote\_Symbol))

16- # there is no useful data in Data\_Value\_Unit, Data\_Value\_Footnote, and Data\_Value\_Footnote\_Symbol; so, delete them all by using the following scripts:

obesity$Data\_Value\_Unit <- NULL

obesity$Data\_Value\_Footnote <- NULL

obesity$Data\_Value\_Footnote\_Symbol <- NULL

17- # delete all of the following columns since we are not going to use them for our analyses, and they are repetition of other entries: GeoLocation, ClassID, TopicID, QuestionID, DataValueTypeID, LocationID, StratificationCategory1, Stratification1, StratificationCategoryId1, and StratificationID1.

obesity$GeoLocation <- NULL

obesity$ClassID <- NULL

obesity$TopicID <- NULL

obesity$QuestionID <- NULL

obesity$DataValueTypeID <- NULL

obesity$LocationID <- NULL

obesity$StratificationCategory1 <- NULL

obesity$Stratification1 <- NULL

obesity$StratificationCategoryId1 <- NULL

obesity$StratificationID1 <- NULL

18- Also # delete the following columns since we will not need them for our analyses: Low\_confidence\_Limit, High\_confidence\_Limit, Sample\_Size, Total, `Age(years)`, Education, Gender, `Race/Ethnicity`, and year\_start\_end\_diff

obesity$Low\_Confidence\_Limit <- NULL

obesity$High\_Confidence\_Limit <- NULL

obesity$Sample\_Size <- NULL

obesity$Total <- NULL

obesity$`Age(years)` <- NULL

obesity$Education <- NULL

obesity$Gender <- NULL

obesity$`Race/Ethnicity` <- NULL

obesity$year\_start\_end\_diff <- NULL

**Cleaning and Deleting the Rows**

19- Find all the rows # not including “Data not reported”, and delete those rows.

obesity <- obesity[!(obesity$Income %in% c("Data not reported")),]

20- View(obesity)

21- Find all the rows # not including NA, and delete those rows.

obesity <- obesity[!(obesity$Income %in% c(NA)),]

22- Find rows # including only “Percent of adults aged 18 years and older who have obesity”, and keep those rows.

obesity <- obesity[(obesity$Question %in% c("Percent of adults aged 18 years and older who have obesity")),]

Updated script for 11/14/17 due date

# Create data frame "obesity" from CSV file w/headers

obesity <- read.csv("C:/Users/mgiezeman/Desktop/obesity.csv", header = T)

R Script for Data Preparation (rows 4-69)

#check if the columns YearStart and YearEnd are the same (find the difference)

obesity$year\_start\_end\_diff<-obesity$YearEnd - obesity$YearStart

#display the unique values of the difference; see there is only 0 difference

unique(obesity$year\_start\_end\_diff)

#delete column YearEnd

obesity$YearEnd<-NULL

#rename column YearStart as Year

names(obesity)[names(obesity)=='YearStart']<-'Year'

#delete column LocationAbbr, as we only need LocationDesc that has the full state name

obesity$LocationAbbr<-NULL

#delete column Datasource. We know all entries are the same, and that our data source is the survey

obesity$Datasource<-NULL

#delete column Class. Class had three sub-categories (Fruits and vegetables, Obesity/Weight Status, Physical Activity) since we are not interested in Class, we are interested in specific questions.

obesity$Class<-NULL

#delete column Topic. Topic had three sub-categories (Fruits and vegetables, Obesity/Weight Status, Physical Activity) since we are not interested in Topic, we are interested in specific questions.

obesity$Topic<-NULL

#check if the values are same for Data\_Value and Data\_Value\_Alt. Get difference, then view unique values.

obesity$data\_value\_diff<-obesity$Data\_Value - obesity$Data\_Value\_Alt

unique(obesity$data\_value\_diff)

#values are all the same, so delete Data\_Value\_Alt

obesity$Data\_Value\_Alt<-NULL

#check if there is any data other than "Value" in Data\_Value\_Type

unique(obesity$Data\_Value\_Type)

#there is no data other than "Value" in Data\_Value\_Type, so delete column

obesity$Data\_Value\_Type<-NULL

#check if there is any useful data in columns Data\_Value\_Unit, Data\_Value\_Footnote, and Data\_Value\_Footnote\_Symbol

unique(obesity$Data\_Value\_Unit)

unique(obesity$Data\_Value\_Footnote)

unique(obesity$Data\_Value\_Footnote\_Symbol)

#there is nothing we will use, so we will delete these 3 columns

obesity$Data\_Value\_Unit<-NULL

obesity$Data\_Value\_Footnote<-NULL

obesity$Data\_Value\_Footnote\_Symbol<-NULL

#delete all of the following columns since we are not going to use them for our analyses, and they are repetition of other entries: GeoLocation, ClassID, TopicID, QuestionID, DataValueTypeID, LocationID

obesity$GeoLocation <- NULL

obesity$ClassID <- NULL

obesity$TopicID <- NULL

obesity$QuestionID <- NULL

obesity$DataValueTypeID <- NULL

obesity$LocationID <- NULL

#delete columns created for this cleanup process

obesity$year\_start\_end\_diff <- NULL

obesity$data\_value\_diff <- NULL

**R Script for Descriptive Statistics**

#Get a subset of rows where the Question is % obesity, as this is the main indicator for our analyses

percentobesity <- subset(obesity,Question=="Percent of adults aged 18 years and older who have obesity")

#find the highest, lowest, median, and mean in % obesity, to get a general sense of how wide-ranging the values are across different categories, and understand which values represent categories with above-average and below-average rates of obesity

summary(percentobesity$Data\_Value)

#Make a histogram of this subset to visualize where most categories fall, and the highs and lows

hist(percentobesity$Data\_Value, main="Histogram of % Obesity in Stratifications")

#get a subset of obesity rates for the most recent year, 2015, so we can analyze the most up-to-date data. Only select columns showing state and subset of people.

obesity2015<-subset(percentobesity,Year==2015,select=c(LocationDesc,Data\_Value,Stratification1,StratificationCategory1))

#find which 20 categories have the highest rates of obesity in 2015

head(obesity2015[order(obesity2015$Data\_Value,decreasing=T),],n=20)

#find which 20 categories have the lowest rates of obesity in 2015

head(obesity2015[order(obesity2015$Data\_Value,decreasing=F),],n=20)

#we want to look at how income in particular correlates with obesity rates, so we create a subset that only includes income categories, not including rows where Income was not reported.

obesity2015income <- subset(obesity2015,StratificationCategory1=='Income' & !(Stratification1=="Data not reported"))

#in the income subset, get the highest rate of obesity per state

obesity2015income.agg<-aggregate(Data\_Value ~ LocationDesc,obesity2015income,max)

#merge with original to see the income category of that highest rate

obesity2015income.max<-merge(obesity2015income.agg,obesity2015income)

#in the income subset, get the lowest rate of obesity per state

obesity2015income.agg2<-aggregate(Data\_Value ~ LocationDesc,obesity2015income,min)

#merge with original to see the income category of that lowest rate

obesity2015income.min<-merge(obesity2015income.agg2,obesity2015income)

#we realized "National" was still a value for state, and we need to remove it, since the data is represented elsewhere.

obesity2015income<-obesity2015income[!(obesity2015income$LocationDesc=="National"),]

**R Script for Analyses**

#look at relationship between obesity value and income

regObesityIncome<-lm(Data\_Value ~ Stratification1,obesity2015income)

#View regression

regObesityIncome

#plot regression to see relationship between income and obesity

plot(regObesityIncome)

#summary to see relationship between income and obesity

summary(regObesityIncome)

#add a new column with a numerical value (1-6) to represent income level, so we can perform analysis more clearly. We have assigned numbers to income categories

$75,000 or greater 6  
$50,000 - $74,999 5  
$35,000 - $49,999 4  
$25,000 - $34,999 3  
$15,000 - $24,999 2  
Less than $15,000 1

This way, it is easier to analyse the relationship between these income intervals and obesity rates. The following graph from Excel shows that Data\_Value for obesity rates decreases as Income levels increase.

obesity2015income$incomelevel1<-ifelse(obesity2015income$Stratification1=="Less than $15,000",1,

ifelse(obesity2015income$Stratification1=="$15,000 - $24,999",2,

ifelse(obesity2015income$Stratification1=="$25,000 - $34,999",3,

ifelse(obesity2015income$Stratification1=="$35,000 - $49,999",4,

ifelse(obesity2015income$Stratification1=="$50,000 - $74,999",5,

6)))))

#replace regression with one done with numerical value

regObesityIncome<-lm(Data\_Value ~ incomelevel1,obesity2015income)

#view new regression

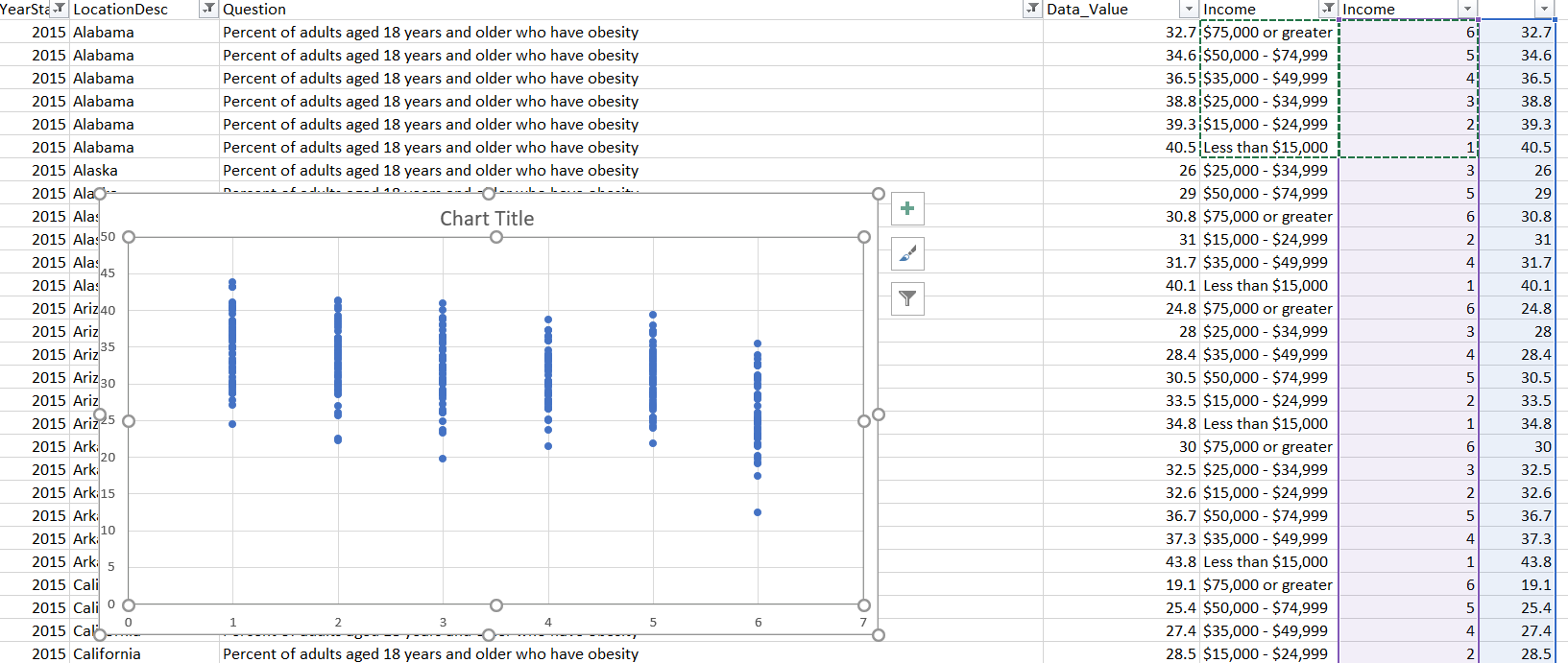
regObesityIncome

#plot regression to see relationship between income and obesity

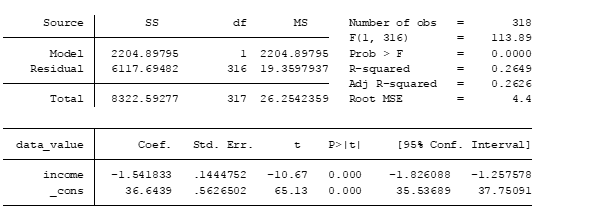
plot(regObesityIncome)

#summary

summary(regObesityIncome)



When we do “reg data\_value income” in Stata, the numbers are:



So, the -1.54 coefficient is significant because p>ItI is low, which shows that the relationship that we found between income and obesity is significant. We are looking for R commands to find these values for now but as a cross check, we have looked at other statistical packages such as Excel and Stata.