Chak Wong

Computation and Statistics I  
Homework 2

1) What is a census tract? How many census tracts are in New York County?

**Answer:** A census tract is a semi-permanant subdivision of a county or equivalent entity that is updated by local participants prior to each decennial census, and the census tract's primary purpose is to provide a stable set of geographic units for the presentation of statistical data. According to the link provided by the

homework (American FactFinder), there are 319 census tracts in New York County.

2) Describe one advantage and one disadvantage of computing estimates after combining 5-years data.

**Answer:** One advantage of computing estimates after combining 5-years data allows possible longitudinal trends that come up that otherwise don't show up in 1-year data. One disadvantage is that with a combined 5-year data, it's difficult to view what happened in one year, since the data had already been merged together.

3) Download the 2008-2012 5-year estimates for all New York county census tracts for the following variables: 1) unemployment; 2) housing tenure; 3) no vehicles; 4) low occupancy. These variables are found in Table DP03 (economic characteristics) and Table DP04 (housing characteristics). First, merge the tables into one data frame, each row representing a census tract, each column representing one of the Townsend variables. Second, for each variable construct a histogram and compute the following summary statistics: mean, median, standard deviation, maximum and minimum. Describe the shape of each histogram.

#load each dataset to a respective table

x <- read.csv("ACS\_12\_5YR\_DP03\_with\_ann.csv", sep=",", header=TRUE,

stringsAsFactors=TRUE)

y <- read.csv("ACS\_12\_5YR\_DP04\_with\_ann.csv", sep=",", header=TRUE,

stringsAsFactors=TRUE)

#Extract select variables from each table into one data frame

xy <- data.frame(x$GEO.display.label, x$HC03\_VC08, y$HC03\_VC64, y$HC03\_VC82,

y$HC03\_VC110)

#Redefine each column name.

#Low Occupany to Over Crowded transformation performed later.

colnames(xy) <- c("Census.Tract", "Unemployment", "House.Rented",

"No.Vehicles", "Over.Crowded")

#Remove the first row, which has unncessary data, from the data frame

xy <- xy[-c(1),]

#Convert columns into 'character' type

xy$Census.Tract <- as.character(xy$Census.Tract)

xy$Unemployment <- as.character(xy$Unemployment)

xy$House.Rented <- as.character(xy$House.Rented)

xy$No.Vehicles <- as.character(xy$No.Vehicles)

xy$Over.Crowded <- as.character(xy$Over.Crowded)

#Reconvert variables with economic and housing characteristics into 'numeric' type.

xy$Unemployment <- as.numeric(xy$Unemployment)

xy$House.Rented <- as.numeric(xy$House.Rented)

xy$No.Vehicles <- as.numeric(xy$No.Vehicles)

xy$Over.Crowded <- as.numeric(xy$Over.Crowded)

# Transform Low Occupany to represent Over Crowdedness

xy$Over.Crowded <- 100 - xy$Over.Crowded

#Summary characteristics

summary(xy)

apply(xy[,2:5], 2, sd, na.rm=T)

#Figure with 4 plots (2 rows, 2 columns)

par(mfrow=c(2,2))

# Plot 4 histograms and save onto .pdf

hist(xy$Unemployment, las=TRUE, col="cadetblue", density=30, angle=50,

border="black", main="Percentage of Unemployment\nin New York County",

xlab="Unemployment", ylab="Frequency")

hist(xy$House.Rented, las=TRUE, col="cadetblue", density=30, angle=50,

border="black", main="Percentage of Houses Rented\nin New York County",

xlab="Houses Rented", ylab="Frequency")

hist(xy$No.Vehicles, las=TRUE, col="cadetblue", density=30, angle=50,

border="black", main="Percentage of Household without

vehicle in New York County", xlab="No Vehicle", ylab="Frequency")

hist(xy$Over.Crowded, las=TRUE, col="cadetblue", density=30, angle=50,

border="black", main="Over Crowdedness\nin New York County",

xlab="Over Crowded", ylab="Frequency")

pdf("HW2\_Question3\_Histograms.pdf", width=8, height=8)

dev.off() #R knows plot is complete

**Answer:**

**Unemployment**: mean=5.845; median=5.150; sd=4.066738; max=50.00; min=0.00

histogram=positively-skewed.

**House Rented**: mean= 78.17; median=81.30; sd=18.585493; max=100.00; min=10.40

histogram=negatively-skewed.

**No Vehicles**: mean=77.54; median=79.10; sd=9.736394; max=94.70; min=21.10

histogram="generally" negatively-skewed.

**Over Crowded**: mean=5.889; median=4.60; sd=4.994355; max=23.80; min=0.00

histogram=positively-skewed.

4) How many observations are missing for each variable? What percentage of census tracts do not have complete data? Is this a problem for our analysis? Justify your answer. (Note: Do not delete tracks with missing data.)

#Searching for missing observations for each variable

xy[!complete.cases(xy),2:5]

#Obtained missing rates by dividing rows with incomplete cases by total rows

xy.incomplete <-xy[!complete.cases(xy),]

nrow(xy.incomplete)/nrow(xy)

**Answer:** 1) Unemployment is missing 6 observations; Housing Rented, No Vehicles, and Over Crowded each are # missing 9 observations. 2) 3.125% of census tracts do not have complete data. 3) Whether each variable's missing data is problematic for our analysisdepends on the extent of what is considered problematic. On the one hand,3% missing data might be considered insignificant when assessing the entireNew York County. On the other hand, only four variables from the 5-year census tracts are used to compute the Townsend Material Deprivation Index - specifically,the Townsend Material Deprivation Index is the product of one economiccharacteristic, Unemployment, and three housing characteristics, Housing Tenure,No Vehicles, and Low Occupancy. This could be a problem when discovering trends over a 5-year period or when analyzing a specific region withinNew York County.

5) Construct a scatterplot matrix of the 4 variables. Are they linearly related? Next, transform the variables from step a), include them in the data frame, and make a scatterplot matrix. Are they linearly related? Then, construct a correlation matrix of the transformed variables and describe your results. Are high or low correlations preferable when constructing an index? Justify your answer.

#Scatterplot Matrix of the four variables

pairs(xy[,c("Unemployment", "House.Rented", "No.Vehicles", "Over.Crowded")])

pdf("HW2\_Question5\_ScatterPlot\_Untransformed.pdf", width=8, height=8)

**Answer:** Scatterplot Matrix results suggests most variables are nonlinearly related.

#Transform each variable into a new variable

T.Unemployment <- log(xy$Unemployment + 1)

T.HouseRented <- log(xy$House.Rented + 1)

T.Car <- sqrt(xy$No.Vehicles)

T.OverCrowded <- log(xy$Over.Crowded + 1)

#Create new Data frame with existing and new variables

xy.plusfour <- xy

xy.plusfour$T.Unemployment <- T.Unemployment

xy.plusfour$T.HouseRented <- T.HouseRented

xy.plusfour$T.Car <- T.Car

xy.plusfour$T.OverCrowded <- T.OverCrowded

#Examine complete data frame

dim(xy.plusfour)

colnames(xy.plusfour)

#Scatterplot Matrix of the four variables

pairs(xy.plusfour[,c("T.Unemployment", "T.HouseRented", "T.Car", "T.OverCrowded")])

pdf("HW2\_Question5\_ScatterPlot\_Transformed.pdf", width=8, height=8)

**Answer:** Scatterplot Matrix of transformed variables shows more linear relationships, though the entire scatterlot matrix lacks a single, general direction.

#Correlation matrix of transformed variables

cor(xy.plusfour[,6:9], use= "complete.obs")

dev.off() # reset plotting parameters

**Answer:** Results show transformed variable's correlation values ranging from (+.22 to +.52). When constructing an index, it's generally preferred that all variables have high correlations. In this case, however, since our correlations ranges greatly from weak to moderate throughout the matrix (i.e., different relationships between different pairs of variables), it's unclear if the Townsend Index will have strong validity.

6) Compute the Townsend index for each census tract, and add it to your data frame. Which census tract is the most deprived and which census tract is the least deprived (give the census tract number and deprivation index level?

#Standardized variables for each region

z <- scale(xy[,c("Unemployment", "House.Rented", "No.Vehicles", "Over.Crowded")],

center=TRUE, scale = apply(xy[,c("Unemployment", "House.Rented",

"No.Vehicles", "Over.Crowded")], 2, sd, na.rm=T))

#Set z as data.frame

z <- data.frame(z)

#Obtain Townsend Index by adding standarized variable for each census tract

Index <- apply(z,1,sum,na.rm=T)

#Add to main data frame (xy.final)

xy.final <- xy.plusfour

xy.final$Index <-Index

#Obtain most and least deprived Index and respective census tract

xy.final[order(xy.final$Index),]

**Answer:** Census Tract 112.02 is the least deprived, with a Townsend Index = -9.74875, and Census Tract 285 is the most deprived, with a Townsend Index = 6.303566. Another Census Tract had a higher Townsend Index, but it was ignored because it was missing three of the four variables required to construct the Index.

7) The ACS data also includes the estimates' margin of error, which was ignored in the calculations. What are the implications?

**Answer:** Margin of Error indicates an estimate's accuracy - relative to the estimate value, the larger the Margin of Error lessens the estimate's accuracy, and thus lowers our confidence in estimate's value, while the smaller the Margin of Error increases the estimate's accuracy, thereby boosting our confidence in the

value. Since the Townsend Index is the summation of four standardized variable for its region and economic and housing characteristics, having large margins of error provides inaccurate and incorrect estimations of the Census Tract's actual estimates, standardized variables, and Townsend Index.

8) Construct a map color-coded by the deprivation index value. In your map, go from red for the most deprived areas to blue for the least deprived. Include a legend and plot title. Describe the patterns you see, especially in relation to what you know about neighborhoods in New York City What does the large rectangle in the middle of the map represent?

#Load packages to plot maps

library(maps)

library(maptools)

library(RColorBrewer)

library(rgdal)

map <- readOGR(dsn="Map", layer="tl\_2012\_36\_tract")

#rownames for data frame

row.names(xy.final) <- x[-1,2]

#Extract map area representing New York County census tracts

row.names(map) <- as.character(map$GEOID)

map <- map[is.element(row.names(map), row.names(xy.final)),]

xy.final <- xy.final[row.names(map),]

#Attach data to map

map <- spCbind(map, xy.final)

plot(map)

#Save New York County Map as pdf file

pdf("NYC\_heatmap.pdf", height=10, width=12)

library(RColorBrewer)

#Variable to color map by

plot.variable <- map@data$Index

#Construct range of values for each color

breaks.factor <-cut(plot.variable, breaks=seq(6.303566, -9.748751, length=6))

length(levels(breaks.factor))

#Assign Color

color.palette <- brewer.pal(length(levels(breaks.factor)), "Spectral")

color.coding <- color.palette[as.numeric(breaks.factor)]

#Plot map

plot(map, col=color.coding)

#Improve Legend

legend("topleft", legend=attributes(breaks.factor)$levels,

fill=color.palette, cex=.8, bty="n", y.intersp=1.2, ncol=3)

#Plot Title

map@data$GEOID

text(-95.85102, 51, cex=1.5,

labels="Townsend Index in New York County")

#label highest and lowest Index Census Tract Regions

##subset(xy.final, Index==max(xy.final$Index))

##text(coordinates(map[81,]), labels="285\n(highest)", cex=1.2)

##subset(xy.final, Index==min(xy.final$Index))

##text(coordinates(map[257,]), labels="112.02\n(lowest)", cex=1.2)

**Answer:** The patterns shown in the map meets some of my expectations and knowledge of New York County (i.e. Manhattan). For example, the blue regions are located around Chinatown and the Lower East Side, which is known for cheap housing and is a residential for incoming immigrants, while the Financial District and the area above Washington Heights are some examples of least deprived areas. However, it should also be noted that areas marked as some of the most deprived areas are located around Midtown West. The large rectangle in the middle of the map represents Central Park.

9) In which census tract is Lowenstein? What is the deprivation level rank (where a rank of 1 is the most deprived)? Mark it on the map and add it to your legend.

#Define Census Tract 145 as variable 'Find'

Find <- "Census Tract 145, New York County, New York"

match(Find, xy.final$Census.Tract)

274 #Row 274 contains county that contains Lowenstein

#Lowenstein Index

Low <- xy.final[274,10]

#Sort Index values and find Index value rank for Lowenstein

Sorted <- sort(xy.final$Index, decreasing=T)

match(Low,Sorted)

257 #The deprivation level rank is 257, with 1 being the highest

**Answer:** Lowenstein is located in Census Tract 145, its Deprivation Index level is -3.049149, and its Deprivation Index Level Rank, in comparison to the entire New York County, where rank 1 is the most deprived, is 257 out of 288. In other words, Lowenstein Building is located in one of the least deprived areas of New York County.

#Mark Index Census Tract Region for Lowenstein Building and place in Legend

subset(xy.final, Index==match(Low,Sorted))

text(coordinates(map[17,]), labels="145\n(Lowenstein Building)", cex=1.2, pch=19)

legend("bottomright", legend="Lowenstein Building", bty="n")

pdf("HW2\_NYCounty\_HeatMap.pdf", width=8, height=8)

10) New York County is an urban county, however New York state has roughly 22 counties classified as rural. Would it make sense to compute the Townsend index values for all census tracts within New York state combined? Why or why not?

**Answer:** It's illogical to combine all census tracts within New York state because the economic and housing characteristics that makes up the Townsend Index differs as the geographical area changes. For example, overcrowding is considered as high deprivation, but both Midtown East and Washington Heights in New York country contains the some of the highest deprivation indices, two areas that are on opposite spectrums of the Townsend Deprivation Index continuum. In a rural area, however, issues regarding overcrowding should be considered differently because there's less people living in houses and the houses are farther apart.