Jianwen wu

Eco 20250

Prof. Foster

1)Group member: Crystal Hernandez, Keyi Long

2)Data:

> load("~/pums\_NY.RData")

> attach(dat\_pums\_NY)

> summary(income\_total[ (Hispanic == 1) & (Age>18) ])

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

  -6400    5000   15000   25300   33000  746000

> x <-c(income\_total[ (Hispanic == 1) & (Age>18) ])

> mean(x,na.rm=TRUE)

[1] 25298.76

> x1 <-mean(x,na.rm=TRUE)

> summary((Hispanic == 1) & (Age>18))

   Mode   FALSE    TRUE    NA's

logical  177653   18661       0

> n1 <- 18661

> sd(x)

[1] 38344.45

> sd1 <- sd(x)

> dat\_NYC <- subset(dat\_pums\_NY, (dat\_pums\_NY$in\_NYC == 1)&(dat\_pums\_NY$Age >= 18)&(dat\_pums\_NY$Asian == 1))

> attach(dat\_NYC)

> borough\_f <- factor((in\_Bronx + 2\*in\_Manhattan + 3\*in\_StatenI + 4\*in\_Brooklyn + 5\*in\_Queens), levels=c(1,2,3,4,5),labels = c("Bronx","Manhattan","Staten Island","Brooklyn","Queens"))

> summary(borough\_f)

        Bronx     Manhattan Staten Island      Brooklyn        Queens

          376           931           262          2268          4142

> xa <-c(income\_total[borough\_f])

> mean(xa,na.rm=TRUE)

[1] 30136.53

> x2 <-mean(xa,na.rm=TRUE)

> n2 <- 7979

> sd(xa)

[1] 37775.52

> sd2 <- sd(xa)

> (x1-x2)/sqrt((sd1^2/n1)+(sd2^2/n2))

[1] -9.53113

> z <- -9.53113

> pnorm(z)

[1] 7.779019e-22

> se <- sqrt(sd1\*sd1/n1+sd2\*sd2/n2)

> error <- qt(0.975,df=pmin(n1,n2)-1)\*se

> left <- (x1-x2)-error

> right <- (x1-x2)+error

> left

[1] -5832.751

> right

[1] -3842.789

**Explanation:**

we compare Hispanic (older than 18) to Asian (older than 18) in NYC. We are

assuming that they have Asian has higher total income than Hispanic. Based on the data above, the average total income for Hispanic (older than 18) is 28628.04, the average total income for Asian (older than 18) is 30136.53. Therefore, the difference in average is -4837.77. The standard error of this difference is 507.5757. The 95% confidence interval is (-5832.751, -5832.751). Since the P-value is less than significant level in this case, they are statistically significantly different.

3)

> cor(rent\_cost, income\_total)

[1] 0.07373553

> mean(rent\_cost)

[1] 624.7658

> mean(income\_total)

[1] 37069.07

**Explanation:**

There is weak correlation between rent cost and income total. In other word, the increased of total income is not necessary affect the rent cost.

4)

> rm(list = ls(all = TRUE)) # clear workspace

> setwd("~/Dropbox/CCNY/Statistics and Intro Econometrics/R Projects/PUMSdata-hw1")

> load("pums\_NY.RData")

> head(dat\_pums\_NY)

Age female PERNUM educ\_nohs educ\_hs educ\_smcoll educ\_as educ\_bach educ\_adv

1 43 1 1 0 0 0 0 0 1

2 45 0 2 0 0 1 0 0 0

3 33 0 1 0 1 0 0 0 0

4 57 0 1 0 1 0 0 0 0

5 52 1 2 1 0 0 0 0 0

6 26 0 3 0 0 0 1 0 0

ANCESTR1D ANCESTR2D immig Hispanic Hisp\_Mex Hisp\_PR Hisp\_Cuban Hisp\_DomR

1 2610 9990 0 1 0 1 0 0

2 511 9990 0 0 0 0 0 0

3 880 9990 0 0 0 0 0 0

4 7060 9990 1 0 0 0 0 0

5 7060 9990 1 0 0 0 0 0

6 7060 9990 1 0 0 0 0 0

white AfAm Amindian Asian race\_oth Married divwidsep unmarried veteran

1 1 0 0 0 0 1 0 0 0

2 1 0 0 0 0 1 0 0 0

3 1 0 0 0 0 0 0 1 0

4 0 0 0 1 0 1 0 0 0

5 0 0 0 1 0 1 0 0 0

6 0 0 0 1 0 0 0 1 0

has\_AnyHealthIns has\_PvtHealthIns Commute\_car Commute\_bus Commute\_subway

1 1 1 1 0 0

2 1 1 1 0 0

3 1 1 1 0 0

4 0 0 0 0 0

5 0 0 0 0 0

6 0 0 0 0 0

Commute\_rail Commute\_other below\_povertyline below\_150poverty

1 0 0 0 0

2 0 0 0 0

3 0 0 0 0

4 0 0 1 1

5 0 1 1 1

6 0 0 1 1

below\_200poverty foodstamps work\_fullyr income\_total income\_wagesal

1 0 1 1 110000 110000

2 0 1 1 39000 39000

3 0 1 1 72000 72000

4 1 1 0 0 0

5 1 1 0 7000 7000

6 1 1 0 0 0

HH\_income owner\_cost rent\_cost occ\_dum ind\_dum in\_NYC PUMA in\_Bronx

1 0 2850 0 1820 7860 0 3106 0

2 0 2850 0 1550 3390 0 3106 0

3 72000 0 430 4210 770 0 100 0

4 7000 0 900 0 0 1 4103 0

5 7000 0 900 4520 8980 1 4103 0

6 7000 0 900 0 0 1 4103 0

in\_Manhattan in\_StatenI in\_Brooklyn in\_Queens in\_Westchester in\_Nassau

1 0 0 0 0 1 0

2 0 0 0 0 1 0

3 0 0 0 0 0 0

4 0 0 0 1 0 0

5 0 0 0 1 0 0

6 0 0 0 1 0 0

ROOMS BUILTYR2 UNITSSTR

1 8 10 4

2 8 10 4

3 2 9 3

4 3 5 10

5 3 5 10

6 3 5 10

> norm\_varb <- function(X\_in) {

+ (X\_in - mean(X\_in, na.rm = TRUE))/sd(X\_in, na.rm = TRUE)

+ }

> dat\_NYC <- subset(dat\_pums\_NY, (dat\_pums\_NY$in\_NYC == 1)&(dat\_pums\_NY$Age > 20)&(dat\_pums\_NY$Age < 66))

> attach(dat\_NYC)

> borough\_f <- factor((in\_Bronx + 2\*in\_Manhattan + 3\*in\_StatenI + 4\*in\_Brooklyn + 5\*in\_Queens), levels=c(1,2,3,4,5),labels = c("Bronx","Manhattan","Staten Island","Brooklyn","Queens"))

> housing\_cost <- owner\_cost+rent\_cost

> norm\_inc\_tot <- norm\_varb(income\_total)

> norm\_housing\_cost <- norm\_varb(housing\_cost)

>

> data\_use <- data.frame(norm\_inc\_tot,norm\_housing\_cost)

> good\_obs\_data\_use <- complete.cases(data\_use,borough\_f)

> dat\_use <- subset(data\_use,good\_obs\_data\_use)

> y\_use <- subset(borough\_f,good\_obs\_data\_use)

> detach(dat\_NYC)

> set.seed(12345)

> NN\_obs <- sum(good\_obs\_data\_use == 1)

> select1 <- (runif(NN\_obs) < 0.9)

> train\_data <- subset(dat\_use,select1)

> test\_data <- subset(dat\_use,(!select1))

> cl\_data <- y\_use[select1]

> true\_data <- y\_use[!select1]

> summary(cl\_data)

Bronx Manhattan Staten Island Brooklyn Queens

5568 5546 2095 13443 11915

> prop.table(summary(cl\_data))

Bronx Manhattan Staten Island Brooklyn Queens

0.14437213 0.14380170 0.05432105 0.34856224 0.30894288

> summary(train\_data)

norm\_inc\_tot norm\_housing\_cost

Min. :-0.749438 Min. :-1.338548

1st Qu.:-0.535068 1st Qu.:-0.623880

Median :-0.270258 Median :-0.232514

Mean :-0.002506 Mean :-0.002493

3rd Qu.: 0.186855 3rd Qu.: 0.439614

Max. :13.900225 Max. : 8.552795

> require(class)

Loading required package: class

> for (indx in seq(44, 88, by= 4)) {

+ pred\_borough <- knn(train\_data, test\_data, cl\_data, k = indx, l = 0, prob = FALSE, use.all = TRUE)

+

+ num\_correct\_labels <- sum(pred\_borough == true\_data)

+ correct\_rate <- num\_correct\_labels/length(true\_data)

+ print(c(indx,correct\_rate))

+

+ }

[1] 44.0000000 0.3585084

[1] 48.0000000 0.3589804

[1] 52.0000000 0.3580363

[1] 56.0000000 0.3644088

[1] 60.0000000 0.3639367

[1] 64.0000000 0.3639367

[1] 68.0000000 0.3618126

[1] 72.0000000 0.3596885

[1] 76.0000000 0.3580363

[1] 80.0000000 0.3596885

[1] 84.0000000 0.3625207

[1] 88.0000000 0.3618126

> for (indx in seq(44, 55, by= 1)) {

+ pred\_borough <- knn(train\_data, test\_data, cl\_data, k = indx, l = 0, prob = FALSE, use.all = TRUE)

+

+ num\_correct\_labels <- sum(pred\_borough == true\_data)

+ correct\_rate <- num\_correct\_labels/length(true\_data)

+ print(c(indx,correct\_rate))

+

+ }

[1] 44.0000000 0.3596885

[1] 45.0000000 0.3603965

[1] 46.0000000 0.3587444

[1] 47.0000000 0.3582724

[1] 48.0000000 0.3582724

[1] 49.0000000 0.3568563

[1] 50.0000000 0.3589804

[1] 51.0000000 0.3596885

[1] 52.0000000 0.3570923

[1] 53.0000000 0.3578003

[1] 54.0000000 0.3627567

[1] 55.0000000 0.3618126

>

**Explanation:**

From the prob. Of every class, we assume most data is should be “Brooklyn”. And we want to predict the class after 44, and we get the output and match our prediction.