//use a public health individual panel data

use http://www.stata-press.com/data/r11/nhanes2f.dta

\*\*this dataset contains health condition, heart attack, diabetes, and 20 other health related medical information.

//delete unnecessary variables of this project

drop highlead age2 sizplace

drop smsa2 smsa3 smsa1

drop region1 region2 region3 region4

drop hsiz1 hsiz2 hsiz3 hsiz4 hsiz5

drop region smsa

drop stratid psuid

\*\*I kept age, race, sex, and health records.

//do some descriptive statistics

tabulate health

mean bpsystol bpdiast hgb hct tibc iron albumin vitaminc zinc copper porphyrn height2weight

by race, sort: correlate heartatk health diabetes age

\*\* there are relationships between heart attack, health, diabetes, and age.

//visualize the comparison

graph bar (mean) health, over(sex) by(race)

graph bar (mean) heartatk diabetes, over(sex) by(race)

//perform a hypothesis test of the difference of heart attck rate between male and femal

ztest heartatk, by(sex)

\*\*heart attack rate of male is significantly different from female at alpha of 0.05.

//run a regression to predict health

generate height2weight = height / weight

mlogit health sex race age bpsystol bpdiast hgb hct tibc iron corpuscl trnsfern albumin vitaminc zinc copper porphyrn height2weight

predict healthhat

/\*some observed facts from the regression summary:

1.lack of vitamin c is very likely related to the poor health condition

2.height to weight ratio plays an important role, which means obesity is an influencial factor

3.health variable is pretty balanced according to the despriptive statistics, however, the health data

of different race groups is not balanced.

\*/