

Econ 400 Problem Set 6 Question 3 Guide

Travis Cao

March 8, 2021

This serves as a solution guide to PS 6 Q3; it doesn't provide the exact answer to each part of this question.

- (a) Before doing part (a), you need to install the `fre` command:

```
ssc install fre
```

and make sure you load the dataset first before running the do-file.

Look through all the tables generated by the `fre` command to see which variable has the most "Missing" type of data.

- (b) Notice that the recode of `income` is to transform it from integer to some numbers potentially with decimal points, so this might have something to do with transforming categorical data to something else.

"MD" is a label for values recorded in `rincome`. Look through all types of labels in `rincome` (browsing the dataset is fine): do you think something in `rincome` that's not coded as missing data should've been?

- (c) `mdpaeduc` is generated as a dummy variable to indicate whether `paeduc` had missing data for the specific entry. The

```
impute paeduc one, gen(paeduc2)
```

command is filling in the missing data with the mean level of `PAEDUC`, and save this version of filled-in data as a variable called `paeduc2`. Now think about why this way of filling in missing data is done, and why these variables are named the way they did.

- (d) Before running part 4's do-file, you need to run the following commands first:

```
gen male = (sex == 1) // generate male dummy variable  
gen white = (race == 1) // generate white dummy variable
```

Now think about what the regression implies if the coefficient on `MDPAEDUC` is zero, and what it implies if the coefficient on `MDPAEDUC` is nonzero.

Hint: If it's nonzero, then would it suggest some pattern on why certain data is missing? Perhaps the data isn't missing at random?