Project 1: Ambulatory Care Medical Data

McCourt School of Public Policy, Georgetown University

Week 2: Testing and Complex Variable Creation

Key Ideas:

- test command after regression
- egen command
- bysort prefix

Overview

- We will continue with the project we started last week
- You may have to re-download the data set and documentation from Blackboard
- Like last week, save all project files to a dedicated folder
- Download the provided solutions from last week to your project folder
- Review the solution and ask questions if there is anything you don't understand
- For the questions this week, do not start a new do-file
- Continue adding commands, either to your previous do-file from last week or to the solution do-file from Blackboard

Questions

1.11. Confirm Sample

- Last week we restricted our data set to patients age 18 and older.
- Re-run the do-file and verify that you are working with the restricted data set.
- After running the do-file, your data set should have 3,885 observations.

1.12. Recode missing

- We will be working with the variables bpsys, bpdias, htin, wtlb this week.
- For each variable, check for negative values
- Recode -7, -8, and -9 to missing for each of these variables

1.13. Regressions on dummies

- The question uses the three indicator variables that you created in Question 1.8
- Verfy your variable creation:

  . tab overwt current_tobac

```
           |     current_tobac
    overwt |         0          1 |     Total
-----------+----------------------+----------
         0 |       454        113 |       567
         1 |       594        135 |       729
-----------+----------------------+----------
```

```
     Total |     1,048        248 |     1,296
```

. tab overwt_current_tobac

```
overwt_curr |
  ent_tobac |      Freq.     Percent        Cum.
------------|-----------------------------------
          0 |      1,161       89.58       89.58
          1 |        135       10.42      100.00
------------|-----------------------------------
      Total |      1,296      100.00
```

- Run a regression of systolic blood pressure on the indicators for current tobacco use and overweight

- The `test` command allows you to perform hypothesis tests using results from the most recent regression
- `test` performs an F-test, which you will learn about in Quant class. For now, just focus on p-values
- Test the hypothesis that the coefficient on current_tobac is equal to zero.
- test current_tobac==0
- Compare to the p-value reported in the regression results

- Unless you run another regression, any test commands will continue to apply to the last one run.

- Test the null-hypothesis that the coefficient on current_tobac is equal to 7. Can you reject this hypothesis?
- Test the null-hypothesis that the coefficient on current_tobac is equal to 2. Can you reject this hypothesis?
- What null-hypothesis is tested if you don't specify a number? Run the command: `test current_tobac`

1.14 Testing with multiple restrictions

- The test command can be used to test hypotheses with multiple variables
- Run a regression of diastolic blood pressure on the indicators for current tobacco use and overweight.
- What null-hypothesis is tested by the following command? `test current_tobac overwt`
- Test the null-hypothesis that the coefficient on `current_tobac` is equal to the coefficient on `overwt`
- Run the test command: `test current_tobac = overwt = 0`. How is test compare to the previous two tests?

1.15 Regressions with dummies and interaction with testing

- Run a regression of diastolic blood pressure on current_tobac, overwt, and the indicator for overwt and current_tobac.
- Is the current_tobac indicator significantly different from zero? How about the indicator for overwt and current_tobac?
- Test the null-hypothesis that both of the two current_tobac indicators are jointly equal to zero.

1.16 Quadratic terms

- Create a new variable equal to age-squared.
- Run a regression of systolic blood pressure on current_tobac, overwt, age, and age-squared.
- Test the null-hypothesis that the coefficients on age and age-squared are jointly equal to zero.
- Test the null-hypothesis that the coefficient on current_tobac coefficient on overwt are equal to each other.

1.17 egen Variable Creation with multiple variables

- Create a new variable called `bpave` equal to the average of systolic and diastolic blood pressure.
- Another way to create this variable is with egen: `egen bpave2 = rowmean(bpsys bpdias)`
- Egen gives you access to many different functions that make complex variable creation easier.
- Try out the following egen commands.
- Look up the description of each function on the help page: `help egen`

- How many imaging tests were performed on each patient?

- The perfomance of imaging tests is described in the variables xray-othimage.
- Each variable is equal to 1 or 0 , so create a sum of the number of imaging tests.

```
browse xray-othimage
describe xray-othimage
tab1 xray-othimage , nolabel missing
egen numimage = rowtotal(xray-othimage)
browse xray-othimage numimage
```

- The variables med1-med8 describe the medications received by each patient.

- Construct a variable giving the count medications received by each patient.
- First recode "No Entry Made" to Stata missing.
- Then count the number of non-missing values for each patient.
- Fill in the appropriate egen function below to count the number of non-missing values.

```
browse med1-med8
describe med1-med8
tab med1 if med1 < 0
tab med1 if med1 < 0 , nol
mvdecode med1-med8 , mv(-9)
egen nummeds = ???
browse med1-med8 nummeds
```

1.18 egen Variable Creation with multiple observations

- egen also gives you ways to create variables that use data from all observations
- Create a standardized version of wtlb (mean 0, st. dev. 1)

```
sum wtlb
egen meanwtlb = mean(wtlb)
egen sdstlb = sd(wtlb)
browse wtlb meanwtlb sdwtlb
gen stdwtlb = (wtlb-meanwtlb) / sdwtlb
sum wtlb stdwtlb
```

- Generate another a standardized version of height.
- Use either the method above, or another egen function.

1.19 bysort

- The bysort prefix allows you to repeat a single command over different groups within your data set.

- Imagine you wanted summary statistics for males and females separately.
- Here are two methods to accomplish this: ``` summarize htin wtlb bpsys bpdias if sex==1 summarize htin wtlb bpsys bpdias if sex==2
- Alternative Method bysort sex: summarize htin wtlb bpsys bpdias ```

- You can define bysort categories with more than one variable bysort sex raceun : tab current_tobac

- Summarize bpdias and bpsys for each combination of current_tobac and overwt

1.20 egen and bysort

- Combining egen and bysort lets you do some very complicated variable creation
- You can find observations that are above average within their category
- For example, find males and females of above average height for their gender

```
bysort sex: egen mfaveht = mean(htin)
gen mftall = .
replace mftall = 1 if htin > mfaveht
replace mftall = 0 if htin <= mfaveht
browse sex htin mfaveht mftall
```

- Create a new indicator variable marking individuals that have above average weight for their age.