Quiz #1

To help illustrate the tools presented in this and related chapters, we apply many of the entries to the HELP RCT data.

Data input and output

1. Read in the HELP csv file (available on Bb)
2. Create a dataset that only has the following variables:
   `"cesd","female","i1","i2","id","treat","f1a","f1b","f1c","f1d", "f1e","f1f","f1g","f1h","f1i","f1j","f1k","f1l","f1m","f1n", "f1o","f1p","f1q","f1r","f1s","f1t"`
3. Provide a summary of the dataset including a count of the number of observations and variables along with their types
4. Displaying the first few rows of data, say 5 observations, so we can get a more concrete sense of what is in the dataset.
5. Save your new data set in a foreign format, say either Microsoft Excel or comma separated value format. This will allow access to other tools for analysis and display.

Data display

1. Consider the CESD (Center for Epidemiologic Statistics) measure of depressive symptoms for this sample at baseline. Print the first ten observations of this variable (only this variable).
2. It may be useful to know how many high values there are and to which observations they belong. Print the observation numbers and their corresponding cesd values for anyone with a cesd value greater than 56.

Derived variables and data manipulation

1. Suppose the dataset arrived with only the individual CESD questions and not the sum. We would need to create the CESD score. We'll need to recode the four questions which are asked "backwards," meaning that high values of the response are counted for fewer points. To create the score, we'll need to generate the sum of the non-missing items, which effectively imputes 0 for the missing values, as well as a version that imputes the mean of the observed values instead.
   a. Flip the order of the backwards questions: `f1d, f1h, f1l, f1p` (0s should become 3s, 1s become 2s, 2s become 1s, and 3s become 0s).
   b. Create a variable called **`newcesd`** that is the sum of the non-missing questions `"f1a","f1b","f1c","f1d","f1e","f1f","f1g","f1h","f1i","f1j" ,"f1k","f1l","f1m","f1n","f1o","f1p","f1q","f1r","f1s", "f1t"` (Remember `f1d, f1h, f1l, f1p` should be backwards)
   c. Create a variable called **`nmisscesd`** by counting the number of missing values, per person
   d. Impute the cesd value, using a variable called imputmeancesd, that calculates the average of `"f1a","f1b","f1c","f1d","f1e","f1f","f1g","f1h","f1i","f1j" ,"f1k","f1l","f1m","f1n","f1o","f1p","f1q","f1r","f1s", "f1t"` and multiplying by 20.

2. It is always prudent to review the results when deriving variables. Print the first 20 observations of your newly derived variables and check your recreated CESD score against the one which came with the dataset. To make sure that the missing data was correctly coded, only print the subjects with any missing questions. If you did it right, your output should match the output below.

Obs	cesd	newcesd	nmisscesd	imputemeancesd
4	15	15	1	15.7895
17	19	19	1	20.0000
87	44	44	1	46.3158
101	17	17	1	17.8947
154	29	29	1	30.5263
177	44	44	1	46.3158
229	39	39	1	41.0526

3. The output shows that the original dataset was created with unanswered questions counted as if they had been answered with a zero. This conforms to the instructions provided with the CESD, but might be questioned on theoretical grounds.

   It is often necessary to create a new variable using logic. In the HELP study, many subjects reported extreme amounts of drinking (as the baseline measure was taken while they were in detox). We should create an ordinal measure of alcohol consumption (abstinent, moderate, high-risk) using information about average consumption per day in the 30 days prior to detox (i1, measured in standard drink units) and maximum number of drinks per day in the 30 days prior to detox (i2). The number of drinks required for each category differs for men and women according to NIAAA guidelines for physicians.

   a. Create a new variable called drinkstat as follows:
   "abstinent" if `i1` is equal to 0,
   "moderate" for females: if `i1` is greater than 0 and less than or equal to 1
   and `i2` is less than or equal to 3
   for males: if `i1` is greater than 0 and `i1` is less than or equal to 2
   and `i2` is less than or equal to 4,
   "highrisk" for females if `i1` is greater than 1 or `i2` is greater than 3
   For males: if `i1` is greater than 2 or `i2` is greater than 4

4. It is always prudent to check the results of derived variables. Print observations 365-370 and compare your output to the output below.

Obs	i1	i2	female	drinkstat
365	6	15	1	highrisk
366	6	19	1	highrisk
367	0	44	1	abstinent
368	0	17	1	abstinent
369	8	29	1	highrisk
370	32	44	1	highrisk

Sorting and subsetting datasets

1. It is often useful to sort datasets by the order of a particular variable (or variables). Sort your by CESD and drinking and print the id, cesd, and i1 variables for the first 5 observations of your sorted list.