Instructions

Data quality assurance is key to our success at IPA. When you submit the test, we want to see a complete record of your work. Please attach a **do-file** that completes all problems, as well as a **log file** that shows the output of your code. The do-file should run completely without error from beginning to end. Moreover, attach any other relevant files that you think would be useful for us to evaluate your work. With each module, in addition to the questions*,* please check for data quality and any potential problems with the data that should be flagged and include this in the do-file you submit.

Remember that you will be graded not only on the answers you give, but also on the process that you take to arrive at your answers (i.e. your do-file and comments). This test is meant to judge how well you can solve common problems in data collection and cleaning, your ability to communicate potential problems and the decisions you make, as well as how well you can use Stata to do so. If there are any questions that you cannot answer using Stata, please explain what steps you would have taken had you known the appropriate commands, and you will be given the appropriate partial credit. Keep in mind that some questions are somewhat open-ended/ambiguous and that there is no "correct" answer (although that does not mean that there are no incorrect answers!). Please make sure to explain your reasoning about any decision you make and provide as much detail as possible in your code’s documentation.

Using outside resources is encouraged (and often necessary!), but please mention what resources you have used if you do so, and do **not** consult with other people about the test.

The Data

The following datasets include household survey data on consumption in Sierra Leone. The surveys were conducted over four rounds starting in 2016, with follow ups in 2017, 2018, 2019. These data contain information from the first and second rounds on topics ranging from political preferences in an upcoming election to self-reported measures of expenditure. Please contact us if you have any questions.

**Replacement Workflow/Basic Data Cleaning**

Files: hh\_round1.dta

**Instructions:** hh\_round1.dta is the new wave of data collection that you have just received. There are a few issues with the data that you need to resolve before running checks.

1. Variables *read* and *write* need to be in numeric format. Replace these variables with a numeric variable.
2. What percent of respondents have a female head of household?
3. Recode head\_gender so that female = 0. Adjust the value label as well.
4. After talking to your team leaders, you learn that an enumerator made a few mistakes when entering in the data. They send you an email with these changes:
   1. For hhid 15064, the household head gender should be female.
   2. For hhid 29038, the household head education should be none.
   3. For hhid 53024, the number of members who can read should be 3.
5. Change the label of the variable “hhid” to “Household ID”.
6. Rename the variable “size” to “hh\_size”.
7. The Principal Investigator wants to see if there is a relationship between household head gender and household size. Find the average household size by household head gender.
8. Save this new dataset as “hh\_round1\_clean.dta”.

**Import and Export Excel**

Files: round2.xlsx

**Instructions:** A new round of data collection has just been sent to you. Import the data into Stata and export an excel document with two sheets using Stata: one that contains a list of responses with duplicate household IDs (hhid) and another that contains a list of enumerators that recorded duplicate IDs and how many duplicates each enumerator had.

**Survey Design Reconciliation**

Files: hh\_nr\_round1.dta, hh\_round1.dta

The survey data for the Northern region of the survey area used a different variable format in the agriculture module. All respondents in both areas were asked if they grew the following crops, or if they grew any other crop:

1. Rice
2. Cassava
3. Millet
4. Groundnut
5. Sweet Potato
6. Wheat
7. Sorghum

Please combine these datasets in a consistent manner. The affected variables are crop\_l1 through crop\_l9 in the hh\_round1.dta file and crop\_l in hh\_nr\_round1.dta file.

**Note:** The hh\_nr\_round1.dta has been cleaned by another RA and may not be fully clean. In addition, please provide justifications for any decisions you need to make about the data to perform the task.