**Stata Problem Set Suggestions**

Training introduction (assessment would be ideal)

Variable properties

* Generate bl, midline, endline prefixes for (variable TBD),
* Should all variables receive this prefix? If not, why?
* Remove bl prefix.
* Combining addressdur and addressdur\_unit, create a variable indicating the number of days a household has lived at its current address.
* For the variable educ, why are some values labeled while others remain unlabeled? *Hint: begin by using describe educ*
* Variables (A, B, and C) use the same value label, which has been removed for this problem set. In one line of code, label the values for all three variables (Y/N)
* Bonus: describe or provide syntax illustrating two ways to create a set of dummy variables
* Bonus: relabel using delimit
* Why/ When would you convert a variable from numeric to string? Vice versa?

Unique IDs and duplicates

* Determine whether surveyid is a unique identifier. If it is not unique, determine which values are duplicated.
* Drop observations that are duplicates on all variables. Does this change your answer to the previous question.

Explicit subscripting

* Why does sort order matter when using explicit subscripting?

For-loops and macros (NA)

**if** command vs. **if** qualifier

* Does the following expression include an if command or an if qualifier? Write the Stata commands reflected in this text.
* Why doesn’t display allow the if qualifier?
* How can you tell whether a command allows the if qualifier? Name a command that allows the if qualifier and another that does not.
* What is a realistic scenario in which the if command is helpful? The if qualifier?
* Bonus: Many Stata users indent once for every left brace. Why is this indentation useful?

Saved results (NA)

Recoding (explaining different interpretation, need for changing variable name, if 0/1 for gender rather than 1/2)

Checking skip patterns and logical consistency

* In your project, what critical skip patterns could you check using Stata.
  + What command would you use to check this skip pattern? Write the code below:
* How is the output of assert similar to isid?

**\_N**

**Other data checks**

* Have you encountered nonsensical data traceable to an unknown ***survey***(=questionnaire/field work/etc.) problem? Describe the process through which you identified the source of this error and how you resolved it.
* If you add the prefix “capture” to your command and the command produces an error, what do you expect to see? If the command does not produce an error?
* It can be difficult to tell whether your assertions are true if you use the prefix capture. What are two solutions to this problem?

**\_n**

* What is the difference between \_n and \_N
* **Refer to explicit subscripting.**
* …

**by**

* Keeping in mind the dataset for your project or a project you are familiar with, when would using “by” be helpful?
* What are “by-groups?”
* When thinking about by-able commands, it can help to break commands into four groups. What are these groups?
* Why, and when, is bysort
* We know that, without “by,” \_N is the total number of observations in the dataset and \_n is the specific number assigned to individual observations. The use of “by” can change the definition of \_N and \_n. Please explain.

**egen**

* What is the key difference between gen and egen? *(hint: different ways of calculating values for new variables)*

**String cleaning**

Discussion questions: embedded in the text, used during course

Problem set questions: everything else (may be discussed in groups during live trainings)