

Assignment 2

For this assignment, you will be using the Consumer Expenditures Survey (CE). Specifically, you will use two separate datasets from the survey: a dataset containing information on household annual expenditures and a dataset containing information on estimated federal and state taxes. (Note that the survey actually includes many more files; I am giving you only two.) Please follow the instructions below carefully. Make sure to upload **three** files when you turn in your assignment: a do file, a log file, and a pdf file with your answers. To make things clear, the sentences in **bold** below require explicit responses from you. Finally, make sure to provide comments in your do file to allow me to follow everything you do.

You can find the unique identifiers (“primary keys”) in Table 3 of the “ce Information” document. You may also find the other two documents useful; one is the questionnaire and the other is the data dictionary.

1. **Give a brief description of each file.** What is included in each? You may have to look through Table 3 of the ce Information document or the dictionary/questionnaire.
2. First, open the tax file. What is the unique identifier?
3. Summarize all of the variables. Some variables have fewer valid observations than there are total observations in the dataset. **What explains this?**
4. **Find the average and median of each of the following variables:** taxpayer wage and salary income; state adjusted gross income (AGI) in the current year AND previous year; property taxes paid; and rent paid. (Hint: you may find it helpful to sort by File when using the dictionary to find the variables you need.)
5. Take the (natural) log of wage and salary income and of the current year’s alternative minimum tax liability. Summarize both. **What do you notice? Why did this happen?**
6. Regress AMT liability on wage and **interpret the coefficient.**
7. One way to take logs while not ignoring zero values is to take the log of the variable plus one. Generate two new variables this way. Regress the new AMT liability on the new wage variable and **interpret the coefficient. Did the results change appreciably?**
8. Now, open the expenditures file.
9. Summarize the following variables: total amount of family pre-tax income in the last 12 months (collected data); family size; educational expenses in the current quarter; new car expenditures in the current quarter; used car expenditures in the current quarter; and total amount of student loans owed one year ago. **Are there any missing values for any of the variables? If so, which ones? Why are they missing?** (Hint: you may want to look at the “flag” variable.)
10. If you were going to use the student loan variable in a regression, what would you do with the missing values? Would you replace them all with zeros? Would you impute them all? **Discuss how you would deal with the missing values, with a particular focus on whether you would deal with all missing values the same. If you decide you would impute, discuss how you would impute, including what kinds of variables you would use (if any).**
11. **Discuss some of the advantages and disadvantages of different options for dealing with missing values. How do these different options affect standard errors and inference?**

12. Merge the two files together using the appropriate command. You may have to use Table 3 of the ce Information document. **Were there any matching failures?**
13. Finally, **why might we want to merge these two files together?** In other words, what would merging allow us to do that having them separately would not?