Homework 2

PHPM 672 - Spring 2012

Eric A. Booth, Texas A&M University, ebooth@tamu.edu

January 25, 2012

Overview

In this homework, you will work with the 2010 National Survey on Drug Use and Health (NSDUH) dataset which is conducted by SAMHSA and housed at the Inter-university Consortium for Political and Social Research (ICPSR) at the University of Michigan.

This dataset contains survey measures for the prevalence and correlates of drug use in the United States, such as information on illicit drugs, alcohol, and tobacco use, substance abuse treatment history, DSM diagnostic criteria, personal and family income sources and amounts, health care access and coverage, illegal activities and arrest record, neighborhood safety/context, and respondent demographics. The goals of this homework are for you to download and import this data into Stata; clean and recode this data using the survey documentation as a guide; and find some basic, descriptive patterns in the data as instructed below.

Directions

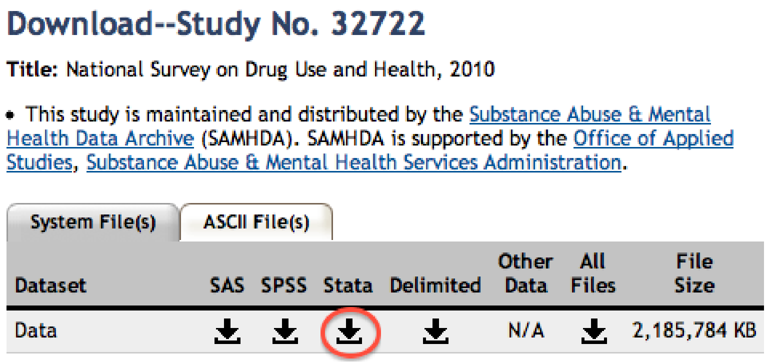
- This homework is due in two weeks (Feb 7th at 5pm) by email to ebooth@tamu.edu. No late work will be accepted.
- For this assignment you must, at minimum, turn in the following:
  1. A do-file (.do) with ONLY the commands you used to create this homework. I should be able to run your do-file from start to stop on the original assigned dataset. (So, do not put comments, output, etc. into this file UNLESS it is properly commented out)
  2. A annotated file (*e.g.*, a word document, text file, etc.) (Assuming you do not include all of your commentary/annotations and output into the file above). In order to get full credit on each question, it must contain (where relevant) the following:
     (a) the code you ran;
     (b) the output it produced, and
     (c) your description/comments about why you ran what your ran and/or what the output tells you. Remember that, when answering a question on the HW, it is better to be wrong than vague.
  3. Do not send me the ICPSR dataset.
- Please be sure to read the assigned chapters in the textbook, the assigned outside reading (e.g., any articles or webpages that are included on the syllabus), the lecture slides/notes, and the help files (including the PDF documentation) for each of the commands included in the homework for the week.
- This homework is graded on a 0 to 100 pt scale. The value of each question/step is included in [brackets] next to the question text.

1 Download and Import Data

1.1 Accessing the ICPSR Data

Data housed at ICPSR are generally well maintained and documented, and they usually have Stata versions of the datasets available for download (which saves you from having to convert and label your data).

Go to this url: `http://dx.doi.org/10.3886/ICPSR32722.v1` and click on the "Download select files" link on this page. Next you will click the icon under Stata that looks like:



Download--Study No. 32722

Title: National Survey on Drug Use and Health, 2010

- This study is maintained and distributed by the Substance Abuse & Mental Health Data Archive (SAMHDA). SAMHDA is supported by the Office of Applied Studies, Substance Abuse & Mental Health Services Administration.

Dataset	SAS	SPSS	Stata	Delimited	Other Data	All Files	File Size
Data	⬇	⬇	⬇	⬇	N/A	⬇	2,185,784 KB

At this point, the website will ask you to create a free log in (probably only if you are off campus – if you are on campuses it *should* recognize your IP address on campus) and agree to terms of service.

Your browser should start downloading a .zip file named something like"ICPSR_32722.zip" You will need to unzip (on Windows right click the file and choose "Extract all ..." or "unzip") these files to somewhere on your hard drive. Make note of the location of this folder.

In the folder, there is a subfolder (called "DS0001") that contains the dataset documentation and Stata dataset (.dta). This dataset, called "32722-0001-Data.dta", is the dataset you will use in this assignment.

1.2 Use Stata dataset

[5 points]

This is a large dataset - so `use` ing the dataset should be done with care. If you load this dataset fully into Stata you will have:

```
obs:   57,873
vars:   3,112
size:   364,831,392 or 364MB
```

If you own Stata `IC` or lower, this dataset will be too big to load all at once since this version of Stata only supports 2,047 variables. That is okay since we certainly do not need to work with 3112 variables all at once. You can modify your `use` command to only use the variables you need for this assignment.

Consult the `help use` documentation on how to open this dataset with only the following variables that are indicators for substance use, treatment, insurance coverage, race, gender, health status, age, work status, and county of residence:

caseid, cigever, alcever, cocever, mjever, txever, medicare, caidchip, champus, prvhltin, income, health, irsex, catage, newrace2, coutyp2, wrkhrsw2

These are the variables we will use for this assignment.

1.3 Describing the Data

1.3.1 Describe the dataset, including the number of observations and variables that are in your subset of data.

[5 points]

1.3.2 Check your data.

[5 points]

(a) Examine each variable for its range (min to max) or number of categories. (b) Discuss any anomalies you see in this data (such as an abnormally high or low values) Be sure to read the codebook documentation provided by ICPSR. (Tell me why there are odd values/categories in some variables and not others)

2 Cleaning and Manipulating the Data

Preparing your data for analysis usually involves:

- Checking and correcting the data formats;
- Adding or changing variable labels, variable names, and value labels;
- Checking and coding missing values;
- Calculating new variables or manipulating existing variables so that you can use them for analysis (using `generate, replace, or egen` or other functions (see `help functions`).

2.1 Checking and correcting the data formats

[5 points]

In this dataset, the data are already checked for formatting and there are no time-date variables to format. For the purposes of this assignment, convert the id variable to string format.

2.2 Adding or changing variable labels, variable names, and value labels

[10 points]

Rename all your variables to lowercase versions of the same variable name (i.e., rename "CASEID" to "caseid")

Label[1] the variable caseid as “ID variable for the sample.” Display the variable labels for all variables in this dataset.

2.3 Checking and coding missing values

[10 points]

Use the codebook to determine which variables have missing value codes. For those variables, replace the value of those codes with “extended” missing value codes so that they will affect analysis (like calculation of a mean).[2]

2.4 Calculating new variables or manipulating existing variables

[20 points]

In this section you will calculate some variables that we can use in the next section for descriptive analysis:

Generate a new variable, called `use_substance`, that is equal to 1 if the respondent has ever had any substance in our sub-dataset (ever responded YES to using cigarette, alcohol, cocaine, marijuana) and equal to zero otherwise.

Generate a new variable, called `type_insurance`, that is 0 if the respondent has no insurance of any kind, 1 if the respondent has government-based/public insurance (medicaid, medicare, or champus/tricare/va insurance), and 2 if the respondent has private insurance. If the person has both public and private insurance, code them as 1.

3 Descriptive Analysis of the Data

In this section you will perform some descriptive analysis of continuous and categorial data in the subset of data.[3]

3.1 Create a summary table for any continuous variables using `tabstat`

[10 points]

Include the mean, standard error of the mean (semean), p99, and N (total Number) values in this table for each continuous variable.

3.2 Create these cross-tabulations for categorical variables

Several commands can be used to examine categorical data. These commands can be used to examine descriptive/bivariate relationships between variables in the form of cross-tabulations. Use some of these commands to explore the following:

[1] See the label help file (`help label`) and Chapter 7 in the Juul book.

[2] So, if the value ’999’ were ”not sure”, try converting it to ’.a’ and then see how it changes the calculation of the mean for that variable.

[3] Lecture #2 Slides contained some commands for examining categorical and continuous data, including for running cross-tabulations

a. Ever Used any Substance - `use_substance`

[10 points]

Explore the relationship between `use_substance` and any 2 of the respondent characteristics (age, gender, income, county type, race/ethnicity).

For instance, what percent and/or number of males have ever used any of the substances versus females? What are the differences between substance use across race/ethnicity categories?

b. Alcohol or Drug Treatment - `txever`

[10 points]

Answer the following: what percent of respondents who have ever used a substance (viz. `use_substance`==1) and make $75,000 or more dollars have been to Alcohol/Drug Treatment?

c. Overall Health Status - `health`

[10 points]

What is breakdown of Health Status (e.g. % Excellent, %V.Good, %Good, %Fair, %Poor) for respondents with the following characteristics/attributes: female, over the age of 25, they have used a substance, they are Hispanic or NonHispanic White, and they worked more than 25 hours a week?

3.3 Save your Dataset in Stata format