Problem Set 1
General Feedback

Nick Miller

17.871

3/4/2013

Comment your code!

- This is more than just /*Part 1*/, /*Part 2*/, etc. and more than /*Collapsing dataset*/
- Explain precisely what you are doing
- In future problem sets, I will take off points for lack of descriptive comments

Example

Good:

/*Generate a variable that equals 1 if the respondent reported waiting more than 30 minutes (when the variable= 4 or 5)*/

Bad:

/*Generating variable*/

Dealing with Missing Values

- Always use tabulate command (with the missing option) to get a sense of variables before analysis, and *especially* before transforming them or using them to create new variables
- Missing values, or 'don't know' answers on surveys should be treated as containing no information (i.e. should not be recoded as zero or as anything other than missing when transformed for use in other variables).

```
. tab q5

                        mode of voting |      Freq.     Percent        Cum.
---------------------------------------+-----------------------------------
in person on election day (at polling p |      6,103       64.93       64.93
  in person before election day (early) |      1,679       17.86       82.80
            voted by mail (or absentee) |      1,613       17.16       99.96
                             don't know |          4        0.04      100.00
---------------------------------------+-----------------------------------
                                  Total |      9,399      100.00

. tab q5, missing

                        mode of voting |      Freq.     Percent        Cum.
---------------------------------------+-----------------------------------
in person on election day (at polling p |      6,103       61.03       61.03
  in person before election day (early) |      1,679       16.79       77.82
            voted by mail (or absentee) |      1,613       16.13       93.95
                             don't know |          4        0.04       93.99
                                      . |        601        6.01      100.00
---------------------------------------+-----------------------------------
                                  Total |     10,000      100.00
```

Use caution when generating variables

- Stata treats the '.' as higher than any integer value.
- gen newvar=1 if oldvar>5
  - This will turn missing values on oldvar into 1s on new var. This is bad.
- gen newar=1 if oldvar>5 &oldvar!=.
  - This will avoid that problem if missing values are dots (but not if they are 999, for example)
- Takeaway point: know how missing values are coded and tailor accordingly

More missing value pitfalls

- gen newvar1=0
  - This will code every observation as zero, regardless of missing data on existing variables. After all, Stata does not know in advance which old variable you will use to code the new variable
- gen newvar1=.
  - This is generally safer because you start with a blank slate and recode based on specific values of an existing variable

Collapse command

collapse (mean) thirtyplus_2008 (count) n= thirtyplus_2008 [aweight=weight], by(inputstate)

- The mean of a binary variable=the percentage of 1s. No additional variables/calculations needed.
- Counting the same variable used to calculate the mean ensures we are getting the correct sample size (need to generate a new variable in order to avoid duplicate names)