Econ 452 Section 3 - STATA

Connor Cole

October 1, 2015

Variable Management, Continued

The 'Egen' Command, Continued

As we discussed in the previous section, the `egen` command allows creation of new variables on the basis of information over a whole column or row. For example, if we had data on income and sex, we could use the `egen` command to calculate standard deviations of income for men and women separately and save it in an additional variable that gives the relevant standard deviation for the sex of each individual. Or, we might want to add up all the values of some variable in a column and save it as an additional variable, or calculate the mean of all the variables in a column. In general, you will use the egen function by typing `egen newvariable =` and then putting in the relevant function, with the variables it takes as an argument in parentheses. The following section describes functions that will likely be of particular interest to you on your problem set.

If you want to calculate the mean of the values in a column and save it as an additional variable, you would type:

```
egen newvar = mean(variablename1)
```

If you want to calculate the standard deviation of the values in a column and save it as an additional variable, you would type:

```
egen newvar = sd(variablename1)twoway line seb1 nvals
```

If you want to calculate the max of the values in a column and save it as an additional variable, you would type:

```
egen newvar = max(variablename1)
```

If you want to calculate the value of some percentile of the distribution of values in a column, for example the median, and save it an additional variable you would type:

```
egen newvar = pctile(variablename1), p(50)
```

Note that if we added a conditioning argument after our `egen`, it would result in two restrictions on the command's function: first, it would result in restricting the calculations of the `egen` function to values from the relevant subgroup, second it would create a variable that only appears for the subgroup defined. Consider the following command:

```
egen newvar = mean(variablename1) if variablename2==1
```

This command would create a new variable that is only defined for observations in the `variablename2==1` category, and the new variable would be the mean of values of `variablename1`. Obviously, if you wanted to compute different means for different subgroups, then if you used this method, you would

Basic Statistical Tools

Calculating Correlations and Covariances among Variables

To calculate the sample covariance between two variables, we type:

```
correlate variablename1 variablename2 variablename3, covariance
```

Note that we can include many variables in this list.

To calculate the correlation coefficient between variables, which is the covariance between variables divided by the standard deviations of the two variables, we similarly type:

```
correlate variablename1 variablename2 variablename3
```

T-Tests for Equivalent Means

Another document available online should describe the basics of hypothesis testing. As this is not a probability theory course, if you need to review these topics, the appendicies to the Wooldridge textbook are very helpful.

To run a simple t-test for the mean of a particular variable being equal to a specific number, say 10, we type:

```
ttest variablename1==10
```

To run a t-test for the mean of a particular variable being equivalent across two subgroups, say sex, where we assume those two groups have equal variances of the variable in question, we type:

```
ttest variablename1, by(sex)
```

To run a t-test for the mean of a particular variable being equivalent across two subgroups, say sex, where we assume those two groups have unequal variances of the variable in question, we type:

```
ttest variablename1, by(sex) unequal
```

To run a simple t-test for the mean of a particular variable being equal to another variable, we type:

```
ttest variablename1==variablename2
```

STATA output for these t-tests provides a lot of different information about the test we have run and the outcomes from it. See the sample output for a single t-test in the picture below and the paragraphs underneath it describing every piece of the output. Here, I have tested the hypothesis that pre-tax labor income is equivalent by sex assuming equal variances.

Figure 1: T-Test Output

```
. ttest pretaxlaborincome, by(sex)

Two-sample t test with equal variances
```

Group	Obs	Mean	Std. Err.	Std. Dev.	[95% Conf.	Interval]
Male	3,438	30776.96	497.823	29189.58	29800.9	31753.02
Female	3,397	22170.89	399.5732	23288.64	21387.47	22954.32
combined	6,835	26499.74	323.781	26768.29	25865.03	27134.45
diff		8606.067	639.1978		7353.04	9859.094

1

```
    diff = mean(Male) - mean(Female)                              t =  13.4639
Ho: diff = 0                                     degrees of freedom =     6833

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 1.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 0.0000
```

2 3 4

1. Basic statistics about the relevant populations. Here, we see the number of observations, the mean, the standard deviation, the standard error associated with estimation of the mean, and a 95% condence interval. With this test, because there are two subgroups, there are two lines of data here on the variable being compared between the two groups. If we had tested equivalent means of two different variables, we would instead see these statistics for two variables, and if we had tested it for only one variable we would see just one row in this table. After these descriptive statistics, we see a row giving the same statistics about the variable if the two subgroups were combined. This row is unique to the output from this form of t-test, as it reflects the fact that we are testing the mean of a single variable for two subgroups. After this line, we see a row stating the mean difference between the two groups,

a standard error, and a 95% confidence interval for the difference. This same row is present also when running a t-test for equivalent means of two variables, but obviously would not be present when running a t-test looking at the value of the mean of a particular variable.

2. A statement of the null hypothesis and the alternative hypothesis.

3. The relevant test statistic.

4. P-values depending on the alternative hypothesis. Note that the center p-value is from a two-sided test, while the two p-values on either side are from alternate hypotheses in either direction around 0.