# CRJ604 Advanced Statistical Analysis
## Stata - Basics by an example

Goals of this exercise:

- Run Stata
- See 4 windows
- Download data from www.public.asu.edu/~gasweete/crj604/data: smoke.dta
- Describe the content of "smoke" data in Stata
- Get basic statistics for all variables
- Show the relationship between two variables: cigarette smoking and birth weight (cigs & bwght)
- Run a regression model using birth weight as the dependent variable and cigarettes smoked as the independent variable

Access Stata.

You should see 4 windows when Stata is opened.

- You will type in commands in the Stata **Command** window.
- These commands will be echoed in the **Review** window. They may be used again simply by clicking on the desired command in the **Review** window, which brings it back to the Stata **Command** window.
- Once a dataset is opened, variable names will appear in the **Variables** window. This is especially helpful when one has complicated variable names as it is possible to click on the variable name to enter it into a Stata Command rather than typing it out.
- Finally, results will appear in the **Results** window. Whenever output from a command exceeds one page, you will see the word **more** at the bottom of the **Results** window. Hit the spacebar to see the next page of output. You can change this setting by typing "`set more off`" in the command window. This will result in no interruption in output.

In what follows, Stata commands and output are reported in **`Courier`** font to distinguish it from my explanations. Basic commands are described using an example. You can also use the drop down menu to do a lot of these tasks. [Important note: Stata is case sensitive, so **`help`** is not the same as **`Help.]`**

Download Stata file "smoke.dta" from http://www.public.asu.edu/~gasweete/crj604/data/

Use *file*, then *open* in Stata's windowing environment to locate the data file on your computer. When you have done this successfully, you will see the **use** command echoed in the results screen as below:

```
. use "…\smoke.dta", clear
```

**`use`** reads in Stata format datasets. Stata format datasets already have variable labels and are ready for use. Stata format datasets have a **`.dta`** extension. Obviously, most data does not automatically arrive in Stata format. Very often, data is in a raw ascii format. Other times, it may be in the format of some other statistical software or as a spreadsheet file. There are a variety of ways to convert data into Stata format. When the data are in ascii (raw) format, a form of **`infile`** is probably the best way to read the data. If the data are already in some other proprietary format (**SAS, Excel, Dbase**, etc.), Stata provides data translation software called **StatTransfer** that is most useful. For now, all the data we'll use will be in Stata format. The option **`clear`**, which follows the comma (,) instructs Stata to clear whatever might be in memory prior to the use command. If there is another dataset already in memory, Stata will clear it.

### . des

**`des`**(cribe) is a command that briefly describes the contents of the dataset. Output from this command is as follows.

```
Contains data from C:\misc\CRJ 604\Wooldridge Datasets\BWGHT.dta
  obs:         1,388
 vars:            14                          3 Jun 1997 13:47
 size:        49,968
-------------------------------------------------------------------------------------
              storage   display    value
variable name   type    format     label      variable label
-------------------------------------------------------------------------------------
--
faminc          float   %9.0g                 1988 family income, $1000s
cigtax          float   %9.0g                 cig. tax in home state, 1988
cigprice        float   %9.0g                 cig. price in home state, 1988
bwght           int     %8.0g                 birth weight, ounces
fatheduc        byte    %8.0g                 father's yrs of educ
motheduc        byte    %8.0g                 mother's yrs of educ
parity          byte    %8.0g                 birth order of child
male            byte    %8.0g                 =1 if male child
white           byte    %8.0g                 =1 if white
cigs            byte    %8.0g                 cigs smked per day while preg
lbwght          float   %9.0g                 log of bwght
bwghtlbs        float   %9.0g                 birth weight, pounds
packs           float   %9.0g                 packs smked per day while preg
lfaminc         float   %9.0g                 log(faminc)
-------------------------------------------------------------------------------------
Sorted by:
```

### . sum

**`sum`**(marize) produces basic statistics. It is always a good idea to summarize your data before proceeding to fancier things. It will give you a basic idea of what your variables look like. If the min, max or mean does not make sense, you've probably read your data in incorrectly, or there is some other error in definition or data. For example, if the min of sales had been negative, it would indicate a problem! Summarize can be used to produce additional statistics like the median, quantiles, etc. To see how to get these options, look at the help on summarize. This is what you should get:

```
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
      faminc |      1388    29.02666    18.73928         .5         65
      cigtax |      1388    19.55295    7.795598          2         38
    cigprice |      1388     130.559    10.24448      103.8      152.5
       bwght |      1388    118.6996    20.35396         23        271
    fatheduc |      1192    13.18624    2.745985          1         18
-------------+--------------------------------------------------------
    motheduc |      1387    12.93583    2.376728          2         18
      parity |      1388    1.632565    .8940273          1          6
        male |      1388    .5208934    .4997433          0          1
       white |      1388    .7845821    .4112601          0          1
        cigs |      1388    2.087176    5.972688          0         50
-------------+--------------------------------------------------------
      lbwght |      1388    4.760031    .1906622   3.135494   5.602119
    bwghtlbs |      1388    7.418723    1.272123     1.4375    16.9375
       packs |      1388    .1043588    .2986344          0        2.5
     lfaminc |      1388    3.071271    .9180645  -.6931472   4.174387
```

## . sum , detail

**`sum`**`(marize), detail` produces more detailed statistics including not only the mean and standard deviation but several measures of central tendency, and percentiles

```
. summ bwght, detail

                    birth weight, ounces
-------------------------------------------------------------
      Percentiles      Smallest
 1%           61             23
 5%           86             30
10%           93             35       Obs                1388
25%          107             38       Sum of Wgt.        1388

50%          120                      Mean           118.6996
                        Largest       Std. Dev.      20.35396
75%          132            172
90%          143            176       Variance       414.2839
95%          149            192       Skewness      -.1458657
99%          161            271       Kurtosis       6.147639
```

## . tab(ulate)

**`Tab(ulate)`** provides a frequency table enumerating all the values for a certain variable.

```
. tab packs

packs smked |
    per day |
 while preg |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |      1,176       84.73       84.73
```

```
        .05 |          3        0.22       84.94
         .1 |          4        0.29       85.23
        .15 |          7        0.50       85.73
         .2 |          9        0.65       86.38
        .25 |         19        1.37       87.75
         .3 |          6        0.43       88.18
        .35 |          4        0.29       88.47
         .4 |          5        0.36       88.83
        .45 |          1        0.07       88.90
         .5 |         55        3.96       92.87
         .6 |          5        0.36       93.23
        .75 |         19        1.37       94.60
          1 |         62        4.47       99.06
        1.5 |          5        0.36       99.42
          2 |          6        0.43       99.86
        2.3 |          1        0.07       99.93
        2.5 |          1        0.07      100.00
------------+-----------------------------------
      Total |      1,388      100.00
```

It can also provide cross-tabs along with a chi-square test for independence of the two variables:

```
. tab packs white, chi

     packs |
 smked per |
 day while |      =1 if white
      preg |         0          1 |     Total
-----------+----------------------+----------
         0 |       252        924 |     1,176
       .05 |         0          3 |         3
        .1 |         1          3 |         4
       .15 |         2          5 |         7
        .2 |         4          5 |         9
       .25 |         4         15 |        19
        .3 |         2          4 |         6
       .35 |         1          3 |         4
        .4 |         1          4 |         5
       .45 |         0          1 |         1
        .5 |        12         43 |        55
        .6 |         0          5 |         5
       .75 |         2         17 |        19
         1 |        13         49 |        62
       1.5 |         1          4 |         5
         2 |         2          4 |         6
       2.3 |         1          0 |         1
       2.5 |         1          0 |         1
```

```
-----------+----------------------+----------
     Total |       299      1,089 |     1,388

         Pearson chi2(17) =  15.2013   Pr = 0.581
```

Any command can be performed on a subset of the dataset by conditioning with an “if” statement. Note that the if statement goes before the comma. Command options go after the comma:

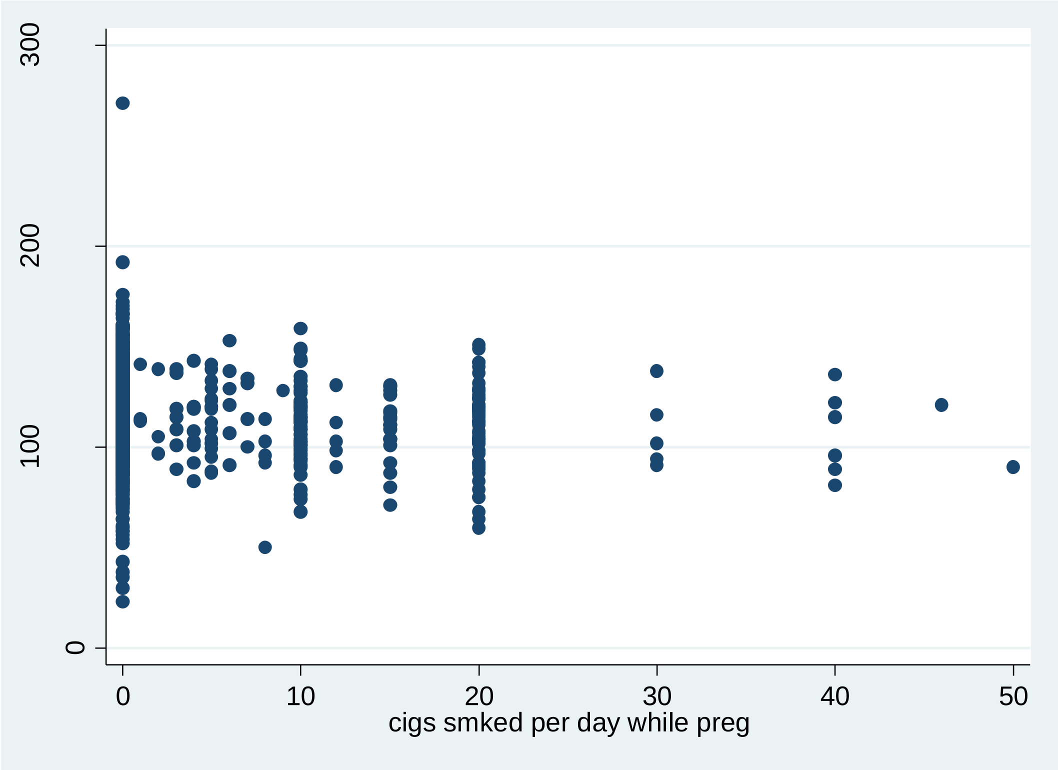
```
. summ bwght if cigs>0, detail

                    birth weight, ounces
-------------------------------------------------------------
      Percentiles      Smallest
 1%           64             50
 5%           79             60
10%           89             64       Obs                 212
25%         98.5             68       Sum of Wgt.         212

50%          112                      Mean           111.1462
                        Largest       Std. Dev.      19.18141
75%          123            149
90%          137            151       Variance       367.9264
95%          143            153       Skewness       -.192946
99%          151            159       Kurtosis       3.048944
```

```
. twoway scatter bwght cigs
```

**twoway** is the command to produce graphs of all types - scatter plots, histograms, etc, for two variables.



```
. regress bwght cigs
```

**regress** is the command to estimate linear regression models. The syntax is set up as **regress y** (bwght) **x** (cigs). The output is as follows.

```
      Source |       SS       df       MS              Number of obs =    1388
-------------+------------------------------           F(  1,  1386) =   32.24
       Model |  13060.4194     1  13060.4194           Prob > F      =  0.0000
    Residual |    561551.3  1386  405.159668           R-squared     =  0.0227
-------------+------------------------------           Adj R-squared =  0.0220
       Total |   574611.72  1387  414.283864           Root MSE      =  20.129

------------------------------------------------------------------------------
       bwght |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        cigs |  -.5137721   .0904909    -5.68   0.000    -.6912861   -.3362581
       _cons |   119.7719   .5723407   209.27   0.000     118.6492    120.8946
```