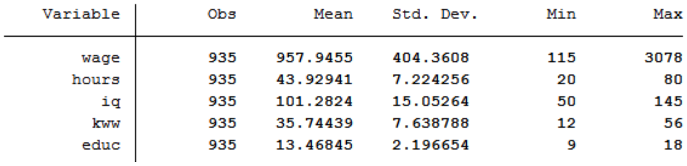
Basic Stata Commands

ECON113 Professor Spearot
TA Jae Hoon Choi

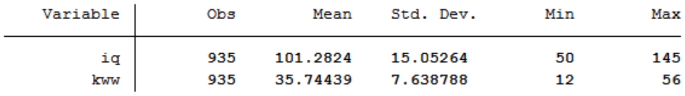
1 Basic Statistics

- `summarize`: gives us summary statistics
  - After opening the data file, running `summarize` will give us summary statistics, including number of observations, mean, standard deviation, minimum, and maximum, for all of the variables in the data file.
    `summarize`



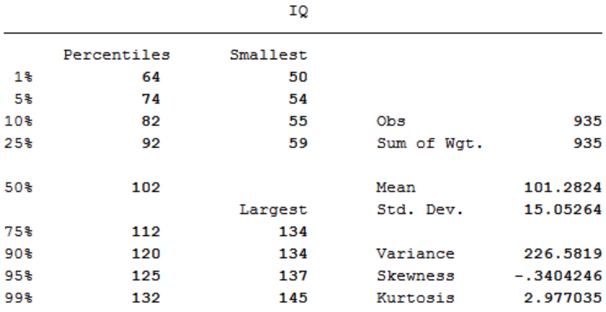
```
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
        wage |       935    957.9455    404.3608        115       3078
       hours |       935    43.92941    7.224256         20         80
          iq |       935    101.2824    15.05264         50        145
         kww |       935    35.74439    7.638788         12         56
        educ |       935    13.46845    2.196654          9         18
```

  - It is also possible to obtain summary statistics for specific variables.
    `summarize iq kww`



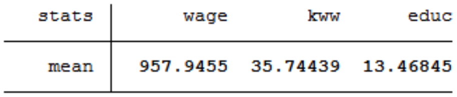
```
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
          iq |       935    101.2824    15.05264         50        145
         kww |       935    35.74439    7.638788         12         56
```

  - If we want to see more detailed summary statistics, we can use an option, `detail`.
    `summarize iq, detail`



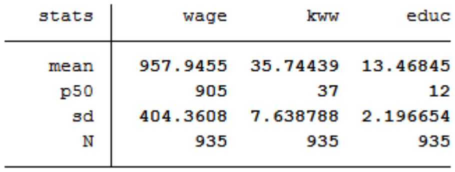
```
                             IQ
-------------------------------------------------------------
      Percentiles      Smallest
 1%           64             50
 5%           74             54
10%           82             55       Obs                 935
25%           92             59       Sum of Wgt.         935

50%          102                      Mean           101.2824
                        Largest       Std. Dev.      15.05264
75%          112            134
90%          120            134       Variance       226.5819
95%          125            137       Skewness      -.3404246
99%          132            145       Kurtosis       2.977035
```

- `tabstat`: displays table of summary statistics
  - Running `tabstat` without options simply provides us means of variables.
    `tabstat wage kww educ`



```
   stats |      wage       kww      educ
---------+------------------------------
    mean |  957.9455  35.74439  13.46845
----------------------------------------
```

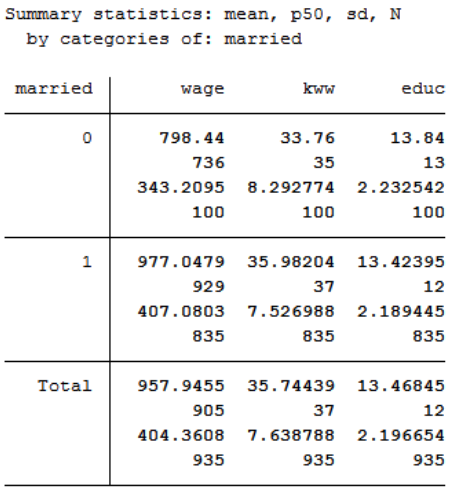
  - Adding an option `statistics( )` gives us more information on the variables
    `tabstat wage kww educ, statistics(mean median sd count)`



```
   stats |      wage       kww      educ
---------+------------------------------
    mean |  957.9455  35.74439  13.46845
     p50 |       905        37        12
      sd |  404.3608  7.638788  2.196654
       N |       935       935       935
----------------------------------------
```

    The statistics we can put in `statistics( )` are following: `mean` (mean), `count` (count of nonmissing observations), `n` (same as count), `sum` (sum), `max` (maximum), `min` (minimum), `range` (range = max - min), `sd` (standard deviation), and `variance` (variance).

- Adding an option `by( )` specifies that the statistics be displayed separately for each unique value of variable.
  `tabstat wage kww educ, by(married) statistics(mean median sd count)`



```
Summary statistics: mean, p50, sd, N
  by categories of: married

 married |      wage       kww      educ
---------+------------------------------
       0 |    798.44     33.76     13.84
         |       736        35        13
         |  343.2095  8.292774  2.232542
         |       100       100       100
---------+------------------------------
       1 |  977.0479  35.98204  13.42395
         |       929        37        12
         |  407.0803  7.526988  2.189445
         |       835       835       835
---------+------------------------------
   Total |  957.9455  35.74439  13.46845
         |       905        37        12
         |  404.3608  7.638788  2.196654
         |       935       935       935
----------------------------------------
```

  The top panel where `married = 0` shows the statistics of people who are not married.

2 Data Management

- `browse`: opens data editor to browse the data set
  - Through data editor you can see how the data set is built and also whether you have managed the data in a way that you want to work.
  - Using data editor, you can edit the values of observations, but I would not suggest doing so for this class or for your academic career. There are better ways to manage values of observations.
- `list`: lists values of variables
  - Adding variable names after command provides values of the specific variable
    ```
    list wage
    ```

    (This will list all observations – in our case, 935 observations. Unless you would like to stare at series of numbers, you can click "stop" button at the top of stata window to stop listing all numbers.)
- `generate`: creates or changes contents of variable
  - You can create a new variable using this command. The following example creates a new variable called `lnwage` with natural log values of `wage`.
    ```
    generate lnwage = ln(wage)
    ```
  - You can also create a new variable with an empty set.
    ```
    generate wage2 = .
    ```

    You can change values of this new variable (`wage2`) by using `replace` command.
    ```
    replace wage2 = wage^2
    ```

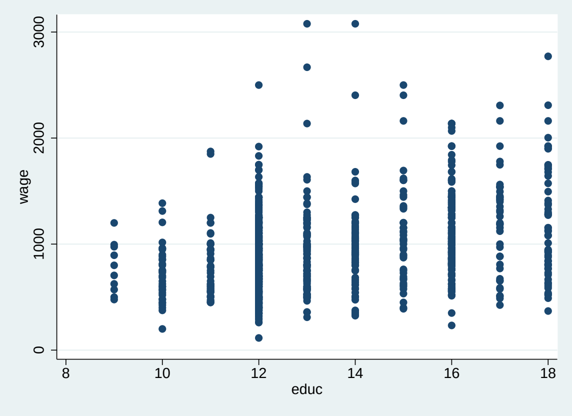
    Now `wage2` variable has values of (`wage`)^2.
- `drop`: eliminates variables or observations
  - You can eliminate the variable you just created.
    ```
    drop wage2
    ```

    (Be careful not to drop variables that you are using for your exercise. If you have accidentally dropped the variables you need, `clear` the memory and reopen the dataset.)
  - You can eliminate the observations by using `if`. The following command will eliminate the observations whose `wage` is greater than 3000. (Suppose you thought that people with `wage` more than 3000 are outliers)
    ```
    drop if wage > 3000
    ```

    (Again, be careful with this. Please `clear` the memory and reopen the original data set before you work on your homework.)

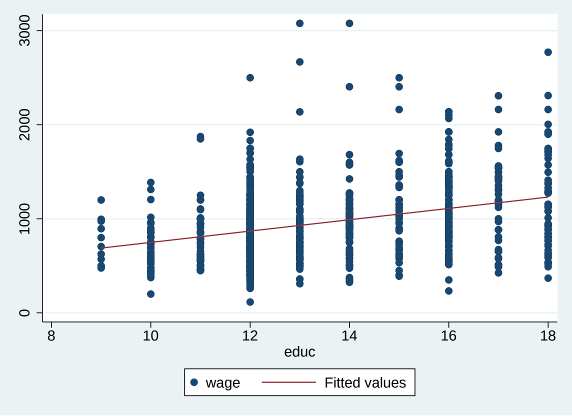
- `clear`: clears memory
- `graph twoway`: creates twoway graphs of scatter plots, line plots, etc.
  - You can investigate the scatter plots of two variables – since it's a twoway graph. The first variable you put after `scatter` will be on the y-axis and the second variable will be on the x-axis, as we will see in the next section, the dependent variable comes before the independent variables.

```
graph twoway scatter wage educ
```



  - You can also graph two different plots in one graph. While `scatter` graphs scatter plots, `lfit` graphs twoway linear prediction plots. We can merge these two plots using the following command:

```
graph twoway (scatter wage educ) (lfit wage educ)
```

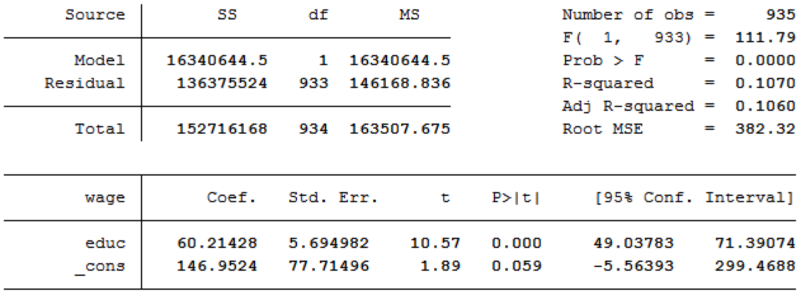


3 Regression

- `regress`: runs a linear regression
  - When using `regress`, after `regress` command put a dependent variable first and independent variable(s) after it. If you want to estimate the following regression specification:

    $$wage = \beta_0 + \beta_1 educ + u$$

    then you run the following command:
    `regress wage educ`



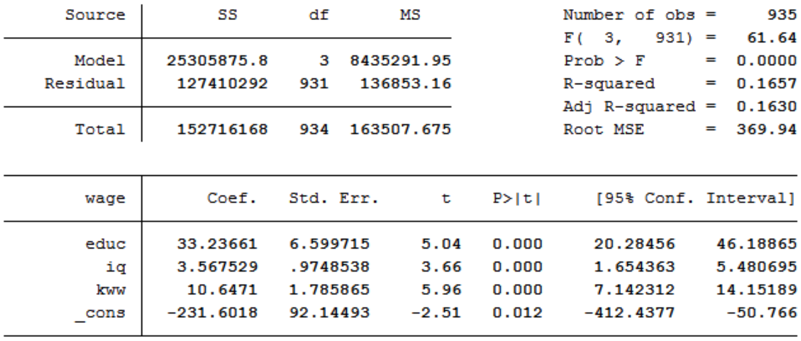
```
      Source |       SS       df       MS              Number of obs =     935
-------------+------------------------------           F(  1,   933) =  111.79
       Model |  16340644.5     1  16340644.5           Prob > F      =  0.0000
    Residual |   136375524   933  146168.836           R-squared     =  0.1070
-------------+------------------------------           Adj R-squared =  0.1060
       Total |   152716168   934  163507.675           Root MSE      =  382.32

------------------------------------------------------------------------------
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   60.21428   5.694982    10.57   0.000     49.03783    71.39074
       _cons |   146.9524   77.71496     1.89   0.059     -5.56393    299.4688
------------------------------------------------------------------------------
```

The result provides $\hat{\beta}_0$, $\hat{\beta}_1$, t-statistics, standard errors, and 95% confidence intervals of estimates, R^2, and many other statistical information of this regression.

  - For multivariate regression, you can just add more independent variables after dependent variable. For example, if you want to run a regression on the model

    $$wage = \beta_0 + \beta_1 educ + \beta_2 iq + \beta_3 kww + u$$

    you can use the following command:
    `regress wage educ iq kww`



```
      Source |       SS       df       MS              Number of obs =     935
-------------+------------------------------           F(  3,   931) =   61.64
       Model |  25305875.8     3  8435291.95           Prob > F      =  0.0000
    Residual |   127410292   931   136853.16           R-squared     =  0.1657
-------------+------------------------------           Adj R-squared =  0.1630
       Total |   152716168   934  163507.675           Root MSE      =  369.94

------------------------------------------------------------------------------
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   33.23661   6.599715     5.04   0.000     20.28456    46.18865
          iq |   3.567529   .9748538     3.66   0.000     1.654363    5.480695
         kww |    10.6471   1.785865     5.96   0.000     7.142312    14.15189
       _cons |  -231.6018   92.14493    -2.51   0.012    -412.4377     -50.766
------------------------------------------------------------------------------
```