Quick Stata Guide

by Liz Foster

Table of Contents

Part 1: Top Ten Stata Commands

describe

This command tells you information about the variables in your dataset – how big they are, what they represent, units, what different codes stand for – if this information is available.

Example

```
. describe

Contains data from example.dta
  obs:           281                          Child Support Awards Santa
                                                Clara County California
 vars:             5                          18 Nov 2004 15:52
 size:         4,496 (99.6% of memory free)
-------------------------------------------------------------------------------
              storage  display     value
variable name   type   format      label      variable label
-------------------------------------------------------------------------------
award           int    %8.0g                  Child support award
earndad         float  %9.0g                  Father's monthly earnings
earnmom         float  %9.0g                  Mother's monthly earnings
nkids           byte   %8.0g                  Number of kids
petmom          byte   %8.0g       yesno      Was it the mother who
                                                petitioned for divorce?
-------------------------------------------------------------------------------
Sorted by:
```

Options

You can select only certain variables by listing them, for example:
`describe earnmom earndad`

---

generate

This command generates new variables. In particular, it can generate dummy variables and interaction terms. It can be abbreviated **`gen`**.

Example

```
. gen richmom = (earnmom >= 2500)

. table richmom, c(freq min earnmom max earnmom mean earnmom)

----------------------------------------------------------------------
  richmom |      Freq.   min(earnmom)   max(earnmom)  mean(earnmom)
----------+-----------------------------------------------------------
        0 |         239              0        2491.67       1205.992
        1 |          42           2500           5250       2950.984
----------------------------------------------------------------------
```

This creates a new binary variable equal to 1 if the mother earns more than $2500 a month. Or, we could generate a variable that indicated whether the mother earned more than the father:
`gen richermom = (earnmom > earndad)`

If you want to see whether child support is a quadratic function of the number of children,

rather than linear, you need to add an *nkids*^{2} term.

```
. gen nkidssq = nkids * nkids

. reg award nkids nkidssq, r

Regression with robust standard errors                Number of obs =     281
                                                      F(  2,   278) =   31.92
                                                      Prob > F      =  0.0000
                                                      R-squared     =  0.1921
                                                      Root MSE      =   218.1

------------------------------------------------------------------------------
             |               Robust
       award |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       nkids |    446.319   111.7826     3.99   0.000     226.2711    666.3669
     nkidssq |  -80.59451   32.09875    -2.51   0.013     -143.782   -17.40702
       _cons |  -106.9748   85.40495    -1.25   0.211    -275.0973    61.14779
------------------------------------------------------------------------------
```

The coefficient on the squared term is significant, so the quadratic form fits the data better.

To see whether the effect of the mother being the petitioner is different for mothers who earn more than their husbands, we need an interaction term *richermom * petmom*.

```
. gen richermom_X_petmom = richermom * petmom

. reg award richermom petmom richermom_X_petmom, r

Regression with robust standard errors                Number of obs =     281
                                                      F(  3,   277) =   17.06
                                                      Prob > F      =  0.0000
                                                      R-squared     =  0.1970
                                                      Root MSE      =  217.83

------------------------------------------------------------------------------
             |               Robust
       award |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   richermom |  -257.8725   66.34101    -3.89   0.000    -388.4691   -127.2759
      petmom |  -150.4746   64.45085    -2.33   0.020    -277.3503    -23.5989
richermom_~m |   87.16457   74.13835     1.18   0.241    -58.78159    233.1107
       _cons |   596.4483   57.13669    10.44   0.000      483.971    708.9256
------------------------------------------------------------------------------
```

The interaction term is not significant.

Options

The command **gen** can be combined with the command **tab** to generate a set of indicator variables for the categories of a category variable. For example:

```
. tab nkids, gen(nkids_)
  Number of |
       kids |      Freq.     Percent        Cum.
------------+-----------------------------------
          1 |        143       50.89       50.89
          2 |        117       41.64       92.53
          3 |         20        7.12       99.64
          4 |          1        0.36      100.00
------------+-----------------------------------
```

```
      Total |        281      100.00

. sum nkids_*

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
     nkids_1 |       281    .5088968    .5008128          0          1
     nkids_2 |       281    .4163701    .4938359          0          1
     nkids_3 |       281    .0711744    .2575746          0          1
     nkids_4 |       281    .0035587     .059655          0          1
```

This creates four new indicator variables. For example, *nkids_2* is equal to 2 if the family has two children, and 0 otherwise. We can now regress the child support award on the number of children in the most flexible way possible without assuming the relationship to be linear or quadratic.

```
. reg award nkids_*, r

Regression with robust standard errors                Number of obs =     281
                                                      F(  3,   277) =   54.72
                                                      Prob > F      =  0.0000
                                                      R-squared     =  0.1954
                                                      Root MSE      =  218.04
------------------------------------------------------------------------------
             |               Robust
       award |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     nkids_1 |   60.06993   13.61372     4.41   0.000     33.27045    86.86941
     nkids_2 |   258.4444   23.15878    11.16   0.000     212.8549     304.034
     nkids_3 |     334.95   74.62313     4.49   0.000     188.0495    481.8505
     nkids_4 |  (dropped)
       _cons |        200          .        .       .            .           .
------------------------------------------------------------------------------
```

For help in interpreting these results, see the page for **`test`**.

`regress`

This command runs an OLS regression. The first variable is the dependant one (Y) the following are the independent ones (Xs). Can be abbreviated **`reg`**.

Example

```
. reg award nkids

      Source |       SS       df       MS              Number of obs =     281
-------------+------------------------------           F(  1,   279) =   55.87
       Model |  2730761.85     1  2730761.85           Prob > F      =  0.0000
    Residual |  13636695.1   279  48877.0432           R-squared     =  0.1668
-------------+------------------------------           Adj R-squared =  0.1639
       Total |  16367456.9   280  58455.2033           Root MSE      =  221.08

------------------------------------------------------------------------------
       award |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       nkids |   154.1658   20.62522     7.47   0.000      113.565    194.7666
       _cons |   120.0708   34.95281     3.44   0.001     51.26609    188.8755
------------------------------------------------------------------------------
```

Options
The option **, r** is added so that Stata allows for heteroskedasticity and calculates the correct standard errors. According to Watson, you should always use it. It changes the format of the output a little:

```
. reg award nkids, r

Regression with robust standard errors                Number of obs =     281
                                                      F(  1,   279) =   35.12
                                                      Prob > F      =  0.0000
                                                      R-squared     =  0.1668
                                                      Root MSE      =  221.08

------------------------------------------------------------------------------
             |               Robust
       award |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       nkids |   154.1658   26.01487      5.93   0.000     102.9554    205.3761
       _cons |   120.0708   36.85967      3.26   0.001     47.51243    192.6292
------------------------------------------------------------------------------
```

So we have the result that award = 120.1 + 154.2 * nkids with standard errors of 26.0 and 36.9 on the two coefficients.

Stata no longer automatically displays the adjusted R-squared. To make Stata display it, use:
display _result(8)

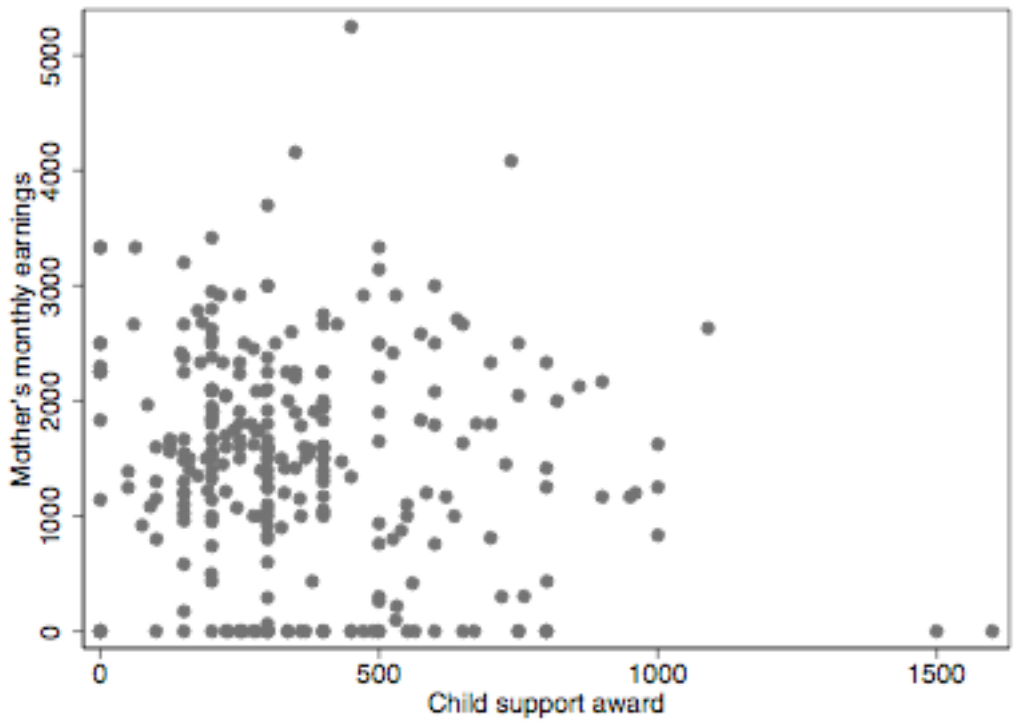
---

scatter

Produces basic scatter plots of data.

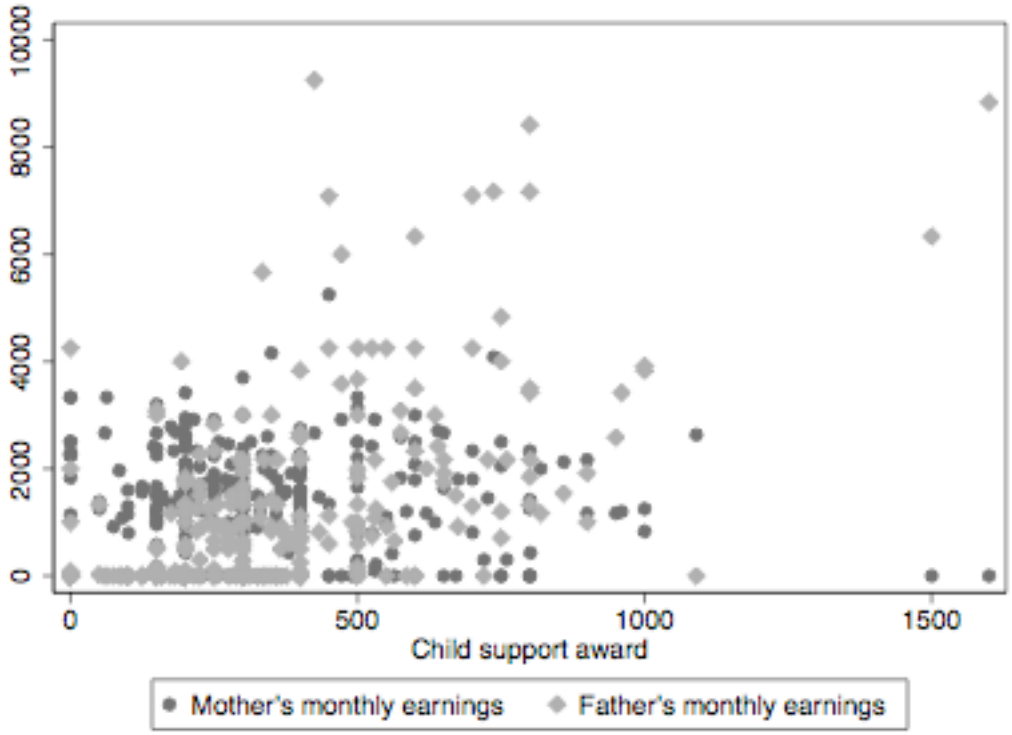
Example

```
. scatter earnmom award
```



Options
Add more variables. The last variable will always be on the x-axis, the other variables on the y-axis, represented by different colored dots.

```
. scatter earnmom earndad award
```



There are dozens of other options – read Stata help. If you want to change something about the scatter plot, you can.

`sort`

This commands sorts your data by the values of a specific variable. It must be run before you can use the prefix **`by :`**

Example
The command
`sort nkids`
produces no output, but if you now run **`describe`** it will tell you that your dataset is sorted by *nkids*.

`summarize`

If run with no arguments, this command produces a basic summary of every variable in your data set. It may be abbreviated **`sum`**.

Example

```
. summarize

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
       award |       281    362.0178    241.7751          0       1600
     earndad |       281    1363.912    2514.409          0   28333.33
     earnmom |       281    1466.809    962.5245          0       5250
       nkids |       281    1.569395    .6405823          1          4
      petmom |       281    .7793594    .4154184          0          1
```

This data set has 5 variables, called *award*, *earndad*, *earnmom*, *nkids*, and *petmom*. The first column *Obs* tells you the number of observations you have for each variable – here we have 281 for each. The second column *Mean* tells you the average value of each variable in the dataset. The third column *Std. Dev.* tells you the standard deviation of the variable. The fourth and fifth columns *Min* and *Max* tell you the smallest and largest value of the variable in the dataset.

Options
You can add a list of variables to produce summary stats for those variables only. For example:

```
. summarize earnmom earndad

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
     earnmom |       281    1466.809    962.5245          0       5250
     earndad |       281    1363.912    2514.409          0   28333.33
```

You can add the option **`, detail`** to produce more detailed statistics for one or more variables.

```
. summarize earnmom, detail

                          earnmom
-------------------------------------------------------------
      Percentiles      Smallest
 1%            0              0
 5%            0              0
10%            0              0       Obs                 281
25%          900              0       Sum of Wgt.         281

50%         1500                     Mean           1466.809
                        Largest      Std. Dev.      962.5245
75%         2100           3700
90%      2666.67        4083.33      Variance       926453.4
95%         2950        4158.33      Skewness       .2368765
99%      4083.33           5250      Kurtosis       3.094632
```

`table`

When given a list of variables, produces tables showing the frequency of combinations of values of those variables.

Examples
One variable:

```
. table nkids

----------------------
    nkids |      Freq.
----------+-----------
        1 |         143
        2 |         117
        3 |          20
        4 |           1
----------------------
```

Two variables:

```
. table nkids petmom

----------------------
          |   petmom
    nkids |    0     1
----------+-----------
        1 |   26   117
        2 |   32    85
        3 |    4    16
        4 |           1
----------------------
```

Three variables:

```
. table award petmom nkids

-------------------------------------------------------------------
          |                    nkids and petmom
          | ---- 1 ---    ---- 2 ---    ---- 3 ---    ---- 4 ---
    award |    0     1       0     1       0     1       0     1
----------+--------------------------------------------------------
        0 |    1    10             3
       50 |           1                           1
       60 |                        1
       63 |           1
       75 |           1
       85 |           1
       90 |           1
      100 |           2             2
      101 |                        1
...
```

Options
The power of table lies in its ability to present a wide range of other statistics in these tables instead of simple frequencies. This is done by an option of the form
`, c(stat1 var1 stat2 var2 ... )`

```
. table nkids petmom, c(mean earnmom)

------------------------------
          |       petmom
    nkids |        0         1
----------+-------------------
        1 | 1811.013  1610.766
        2 | 1099.323   1324.06
```

```
        3 | 605.8325  1523.801
        4 |               2100
------------------------------
```

This table presents the average mother's earnings broken down by number of kids and whether the mother was the petitioner. The next table gives the average value and standard deviation of the award.

```
. table nkids  petmom, c(mean award sd award)

------------------------------
          |       petmom
    nkids |        0         1
----------+-------------------
        1 |  308.846   249.231
          | 182.9812  156.0083
          |
        2 |  578.906   413.094
          | 302.8179  211.4426
          |
        3 |   478.75       549
          |  274.099  360.9759
          |
        4 |                200
          |
------------------------------
```

Some other statistics you may find useful include freq (frequency), sum, median, max and min. Note that freq is not followed by a variable name.

Note
`table` and **`tabulate`** overlap greatly in what they do. In particular, the following two commands
`tabulate var1, sum(var2)`
`table var1, c(mean var2 sd var2 freq)`
produce exactly the same information. **`table`** allows you much more flexibility in exactly what information you present and the number of variables you can work with at once. **`tabulate`** is quicker – significantly quicker for largish datasets.

`tabulate`

When this command is give one variable, it creates a table showing the values that variable takes on. It can be abbreviated **`tab`**.

Example

```
. tabulate nkids

      nkids |      Freq.     Percent        Cum.
------------+-----------------------------------
          1 |        143       50.89       50.89
          2 |        117       41.64       92.53
          3 |         20        7.12       99.64
          4 |          1        0.36      100.00
------------+-----------------------------------
      Total |        281      100.00
```

This shows that *nkids* takes on values from 1 to 4. *Freq* (*Percent*) tells you the number (percentage) of observations with each number of kids. *Cum.* tells you the total percentage of observations with less than or equal to that number of kids.

Options
The option **`, sum(var)`** can be added to the end, so that instead of just giving the percentage / cumulative percentages for each value, Stata gives you the average value of another variable.

```
. tab nkids, sum(earnmom)

            |         Summary of earnmom
      nkids |        Mean   Std. Dev.       Freq.
------------+------------------------------------
          1 |   1647.1741   950.28916         143
          2 |   1262.5931   940.80927         117
          3 |   1340.2075   979.98086          20
          4 |        2100           0           1
------------+------------------------------------
      Total |   1466.8094   962.52452         281
```

Thus we see that for families with 2 kids, the average mother's earnings are $1262.59.

Two variables can be given to create a table that shows how different variables are distributed together.

```
. tab nkids petmom

           |        petmom
     nkids |         0          1 |     Total
-----------+----------------------+----------
         1 |        26        117 |       143
         2 |        32         85 |       117
         3 |         4         16 |        20
         4 |         0          1 |         1
-----------+----------------------+----------
     Total |        62        219 |       281
```

The numbers in the table are frequencies. There were 85 families with two kids where the mother was the petitioner.

These two options can be used together.

```
. tab nkids petmom, sum(award)

            Means, Standard Deviations and Frequencies of award

           |       petmom
     nkids |         0          1 |     Total
-----------+----------------------+----------
         1 | 308.84615  249.23077 | 260.06993
           | 182.98124  156.00825 | 162.20165
           |        26        117 |       143
-----------+----------------------+----------
         2 | 578.90625  413.09412 | 458.44444
           | 302.81788  211.44264 | 249.78092
           |        32         85 |       117
-----------+----------------------+----------
         3 |    478.75        549 |    534.95
```

```
           | 274.09898   360.9759 | 339.94868
           |         4         16 |        20
-----------+----------------------+----------
         4 |         .        200 |       200
           |         .          0 |         0
           |         0          1 |         1
-----------+----------------------+----------
     Total | 459.19355  334.50685 | 362.01779
           | 284.94871   221.1655 | 241.77511
           |        62        219 |       281
```

---

test

After a regression, this command can be used to test various hypotheses about the coefficients of the regression, including an F-test for joint significance.

Example

```
. capture tab nkids, gen(nkids_)

. reg award nkids_*, r

Regression with robust standard errors                Number of obs =     281
                                                      F(  3,   277) =   54.72
                                                      Prob > F      =  0.0000
                                                      R-squared     =  0.1954
                                                      Root MSE      =  218.04

------------------------------------------------------------------------------
             |               Robust
       award |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     nkids_1 |   60.06993   13.61372     4.41   0.000     33.27045    86.86941
     nkids_2 |   258.4444   23.15878    11.16   0.000     212.8549     304.034
     nkids_3 |     334.95   74.62313     4.49   0.000     188.0495    481.8505
     nkids_4 |  (dropped)
       _cons |        200          .        .       .            .           .
------------------------------------------------------------------------------

. test nkids_1 nkids_2 nkids_3

 ( 1)  nkids_1 = 0
 ( 2)  nkids_2 = 0
 ( 3)  nkids_3 = 0

       F(  3,   277) =   54.72
            Prob > F =    0.0000
```

This tells you that the indicator variables of the number of kids are jointly significant – not surprising, since they're all individually highly significant.

Options

In order to use `*` to avoid having to type out all the names of the indicator random variables, you can use the command **`testparam`**.

```
. testparam nkids_*

 ( 1)  nkids_1 = 0
 ( 2)  nkids_2 = 0
```

```
 ( 3)  nkids_3 = 0
 ( 4)  nkids_4 = 0
       Constraint 4 dropped

       F(  3,   277) =   54.72
            Prob > F =    0.0000
```

You can test basically any statement about the coefficients. For example – the child support award is increasing the father's earnings and decreasing in mother's earnings, but are the two effects of the same size?

```
. reg award earnmom earndad, r

Regression with robust standard errors                Number of obs =     281
                                                      F(  2,   278) =    5.06
                                                      Prob > F      =  0.0069
                                                      R-squared     =  0.2256
                                                      Root MSE      =  213.53

------------------------------------------------------------------------------
             |               Robust
       award |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     earnmom |  -.0392945   .0144494    -2.72   0.007    -.0677385   -.0108504
     earndad |   .0437977    .016691     2.62   0.009     .0109409    .0766545
       _cons |    359.919   25.24921    14.25   0.000     310.2151     409.623
------------------------------------------------------------------------------

. test earnmom = -earndad

 ( 1)  earnmom + earndad = 0

       F(  1,   278) =    0.07
            Prob > F =    0.7914
```

Answer: could well be.

---

ttest

Performs a statistical test as to whether a variable has a specified mean, whether two variables are equal, or whether the mean of one variable is equal across values of another variable.

Examples

First case: test whether in half the cases the mother is the petitioner – i.e., whether the mean of *petmom* is 0.5.

```
. ttest petmom==0.5

One-sample t test

------------------------------------------------------------------------------
Variable |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
  petmom |     281    .7793594    .0247818    .4154184    .7305772    .8281417
------------------------------------------------------------------------------
Degrees of freedom: 280
```

```
                          Ho: mean(petmom) = 0.5

    Ha: mean < 0.5               Ha: mean != 0.5                 Ha: mean > 0.5
      t =  11.2728                 t =  11.2728                   t =  11.2728
  P < t =   1.0000           P > |t| =   0.0000           P > t =   0.0000
```

This presents data on the values of *petmom* – the mean, standard deviation and a 95% confidence interval for the mean. It also specifically tests whether the mean of *petmom* is equal to 0.5 and finds that against the alternative that it's not equal to 0.5 (middle column on the bottom) the t-stat is 11.27 which corresponds to a p-value of 0 – so we reject the hypothesis that in half the cases the petitioner is the mother.

Now we test whether for each family, mother's earnings are equal to father's earnings on average. This basically creates a variable that for each family is the difference in earnings, and tests whether it has mean 0.

```
. ttest earnmom == earndad

Paired t test

------------------------------------------------------------------------------
Variable |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
 earnmom |     281    1466.809     57.4194    962.5245    1353.781    1579.838
 earndad |     281    1363.912     149.997    2514.409    1068.647    1659.177
---------+--------------------------------------------------------------------
    diff |     281    102.8973     158.193    2651.798   -208.5013    414.2959
------------------------------------------------------------------------------

               Ho: mean(earnmom - earndad) = mean(diff) = 0

  Ha: mean(diff) < 0         Ha: mean(diff) != 0         Ha: mean(diff) > 0
       t =   0.6505               t =   0.6505               t =   0.6505
   P < t =   0.7420           P > |t| =   0.5159           P > t =   0.2580
```

Here, we can't reject the hypothesis that mothers and father have the same earnings on average.

Now we test whether or not awards are equal in cases where the mother is the petitioner versus cases where she isn't. We tell Stata to break the data up into groups based on the value of *petmom* and test if the mean of *award* is the same in all these groups.

```
. ttest award, by(petmom)

Two-sample t test with equal  variances

------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       0 |      62    459.1935    36.18852    284.9487    386.8301     531.557
       1 |     219    334.5068    14.94498    221.1655    305.0517     363.962
---------+--------------------------------------------------------------------
combined |     281    362.0178    14.42309    241.7751    333.6263    390.4093
---------+--------------------------------------------------------------------
    diff |            124.6867    34.03465                57.68939     191.684
------------------------------------------------------------------------------
Degrees of freedom: 279

                     Ho: mean(0) - mean(1) = diff = 0
```

```
    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
      t =   3.6635                 t =   3.6635                 t =   3.6635
  P < t =   0.9999             P > |t| =   0.0003             P > t =   0.0001
```

Options

Note that in the last example, Stata assumes equal variances (homoskedasticity). If we want to allow that the variance of the award might differ depending on whether or not the mother is the petitioner we want to add the option unequal as in:

`ttest award, by(petmom) unequal`

This accomplishes the same thing as adding the option **`, r`** to a regression command.

The prefix **`by var:`** can also be used to break up the data further – see the explanation for this prefix.

Part 2: Prefixes and Notes

by :

If you have data sorted by the values of a variable, you can run a command separately for each value of the variable. Simply add **by var:** before the command.

Example

```
. sort petmom

. by petmom: reg award nkids, r

_____________________________________________________________________________
-> petmom = no

Regression with robust standard errors                Number of obs =      62
                                                      F(  1,    60) =   10.17
                                                      Prob > F      =  0.0023
                                                      R-squared     =  0.1446
                                                      Root MSE      =  265.73

------------------------------------------------------------------------------
             |               Robust
       award |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       nkids |   179.6584   56.34751     3.19   0.002     66.94663    292.3702
       _cons |   163.6265   84.32099     1.94   0.057    -5.040636    332.2935
------------------------------------------------------------------------------


_____________________________________________________________________________
-> petmom = yes

Regression with robust standard errors                Number of obs =     219
                                                      F(  1,   217) =   23.93
                                                      Prob > F      =  0.0000
                                                      R-squared     =  0.1755
                                                      Root MSE      =  201.28

------------------------------------------------------------------------------
             |               Robust
       award |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       nkids |   142.4462   29.12188     4.89   0.000      85.0483    199.8442
       _cons |   114.0079   40.80712     2.79   0.006     33.57882    194.4369
------------------------------------------------------------------------------
```

This shows us that when the father is the petitioner for divorce, each child increases the award by about $179.66 while if the mother is the petitioner for divorce, each child increases the award by about $142.44. (However, if you look at the confidence intervals, they overlap significantly.)

Note

You must have sorted the data by the variable you wish to use before using **by :**, if not, you'll get the error "not sorted".

---

capture

Makes Stata suppress the normal output from a command.

Example
The command
`capture tab earnmom, gen(em_cat)`
will create a indicator variable for each category of mother's earnings we have in the data set, but without printing out each level of earnings.

Note
The command **`quietly`** seems to do exactly the same thing.

---

*

The asterisk can be used to include a set of variables without having to type them all out. In many commands (**`sum`**, **`describe`** and **`reg`**) instead of a variable use **`var*`** to include all variables whose names start with *var*.

Example

```
. tab nkids, gen(nkids_)

  Number of |
      kids |      Freq.     Percent        Cum.
-----------+-----------------------------------
         1 |        143       50.89       50.89
         2 |        117       41.64       92.53
         3 |         20        7.12       99.64
         4 |          1        0.36      100.00
-----------+-----------------------------------
     Total |        281      100.00

. reg award nkids_*, r

Regression with robust standard errors                Number of obs =     281
                                                      F(  3,   277) =   54.72
                                                      Prob > F      =  0.0000
                                                      R-squared     =  0.1954
                                                      Root MSE      =  218.04

------------------------------------------------------------------------------
             |               Robust
       award |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     nkids_1 |   60.06993   13.61372     4.41   0.000     33.27045    86.86941
     nkids_2 |   258.4444   23.15878    11.16   0.000     212.8549     304.034
     nkids_3 |     334.95   74.62313     4.49   0.000     188.0495    481.8505
     nkids_4 |  (dropped)
       _cons |        200          .        .       .            .           .
------------------------------------------------------------------------------
```

This is particularly useful when you have a lot of indicators variables.

Notes
Be sure to type **`nkids_*`** and not **`nkids*`** so that Stata does not include the original variable *nkids*

This does not work with the command **`test`**, but you can use the command **`testparam`**

that does an F-test for joint significance and takes a variable expression with `*`.

Explanation of the Dataset

The dataset used in the examples is based on a data set we used in 508c last year. It presents information on 281 child support cases in Santa Clara county, California. In all these cases, the mother has physical custody of the children, and the child support payment is to be paid by the father. It has 5 variables:
award – the amount of the child support that was awarded
earnmom – the mother's monthly earnings
earndad – the father's monthly earnings
nkids – the number of children in the family
petmom – a binary variable equal to 1 if the mother petitioned for divorce and 0 if the father petitioned for divorce.

Part 3: More Examples

Interaction Term

First we'll generate a dummy variable to indicate if the mother earns more than the father. Then we'll look at the average award for each group of people.

```
. gen richermom = (earnmom > earndad)

. table richermom petmom, c(mean award)

----------------------------
          |Was it the mother
          |  who petitioned
          |   for divorce?
richermom |      no      yes
----------+-----------------
        0 | 596.448  445.974
        1 | 338.576  275.266
----------------------------
```

What happens if we regress the child support award on whether or not the mother was the petitioner and whether or not she earns more than the father?

```
. reg award richermom petmom, r

Regression with robust standard errors                Number of obs =     281
                                                      F(  2,   278) =   25.18
                                                      Prob > F      =  0.0000
                                                      R-squared     =  0.1915
                                                      Root MSE      =  218.18

------------------------------------------------------------------------------
             |               Robust
       award |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   richermom |  -191.3874   29.79194    -6.42   0.000    -250.0339    -132.741
      petmom |  -101.5843    35.1742    -2.89   0.004    -170.8259   -32.34273
       _cons |    561.061   39.18008    14.32   0.000     483.9337    638.1884
------------------------------------------------------------------------------
```

Based on these regression coefficients, we can predict how much should be awarded to each type of mother.

		Mother petitioned?	
		yes	no
Mother earns more?	yes	561.061	561.061 - 101.5843 = 459.4767
	no	561.061 - 191.3874 = 369.6736	561.061 - 101.5843 - 191.3874 = 268.0893

Note that these numbers are close the averages we calculated, but not exact. In particular, from the first table, it looks like it makes more of a difference who petitions for divorce when the father is richer. To allow for this type of phenomenon, we need an interaction term.

```
. gen richer_pet = richermom * petmom

. reg award richermom petmom richer_pet, r
```

```
Regression with robust standard errors                Number of obs =     281
                                                      F(  3,   277) =   17.06
                                                      Prob > F      =  0.0000
                                                      R-squared     =  0.1970
                                                      Root MSE      =  217.83

------------------------------------------------------------------------------
             |               Robust
       award |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   richermom |  -257.8725   66.34101    -3.89   0.000    -388.4691   -127.2759
      petmom |  -150.4746   64.45085    -2.33   0.020    -277.3503    -23.5989
  richer_pet |   87.16457   74.13835     1.18   0.241    -58.78159    233.1107
       _cons |   596.4483   57.13669    10.44   0.000      483.971    708.9256
------------------------------------------------------------------------------
```

Now if we recalculate our table of predicted values from the regression result, we exactly replicate the sample averages.

		Mother petitioned?	
		yes	no
Mother earns more?	yes	596.4483	596.4483 - 150.4746 = 445.9737
	no	596.4483 - 257.8725 = 338.5758	596.4483 - 257.8725 - 150.4746 + 87.16457 = 275.26577

Notice however that the interaction term is not statistically significant, so that difference we saw may just be an artifact of the data.

Graphing the Regression Line

If we look at the data, we see that there are two outliers of fathers who make more than $15,000 a month – that's more than $180,000 a year.

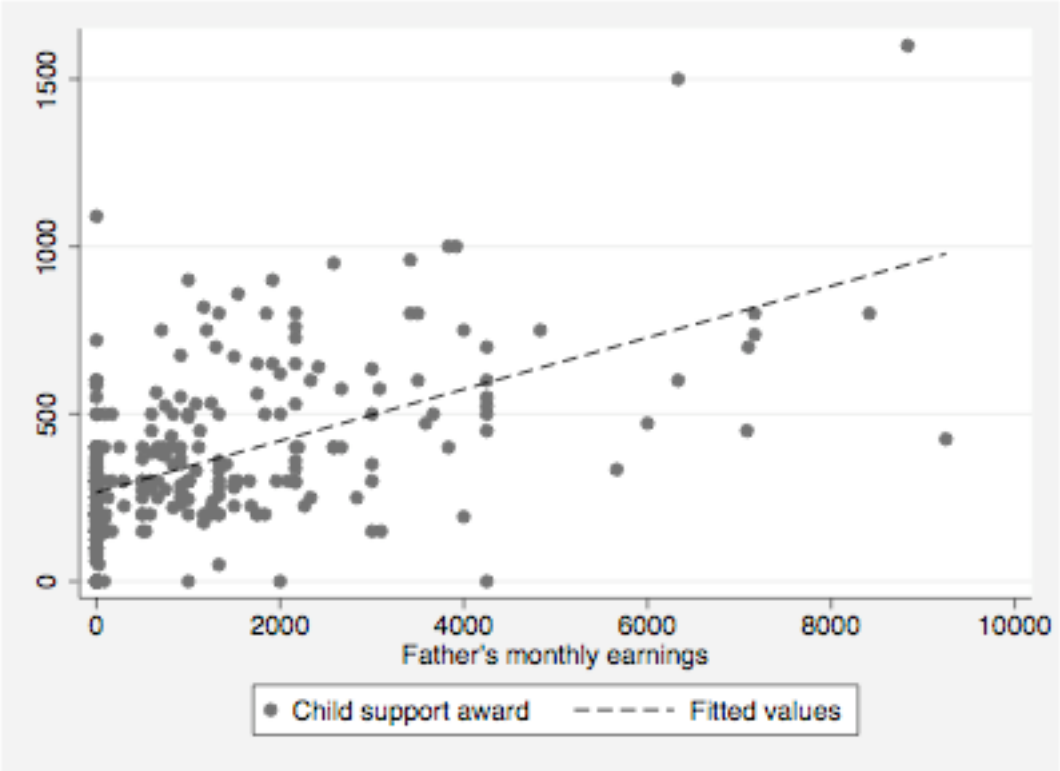
We're going to drop these two observations to make our graphs easier to see. Note that this does change the regression results, so if you were really doing this you would want to think before just dropping them.

```
. drop if earndad > 10000

. reg award earndad, r

Regression with robust standard errors                Number of obs =     279
                                                      F(  1,   277) =   43.07
                                                      Prob > F      =  0.0000
                                                      R-squared     =  0.2979
                                                      Root MSE      =  200.91

------------------------------------------------------------------------------
             |               Robust
       award |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     earndad |   .0770272   .0117365     6.56   0.000     .0539231    .1001313
       _cons |   265.7285   14.62653    18.17   0.000     236.9352    294.5218
------------------------------------------------------------------------------

. predict award_hat
```

```
(option xb assumed; fitted values)

. scatter award earndad || scatter award_hat earndad, c(l) sort m(i)
```

(note that's the letter 'l' not the number '1' after the c.)

And we get this result:



Key Terms and Concepts

Note: Page numbers refer to Stock and Watson.
A regression coefficient is **significant** if

- its value is more than twice its standard error (quick rule for significance at 5% level)
- its **t-stat is greater than** 2.58 / **1.96** / 1.64 (significance at 1% / 5% / 10% level)
- its **p-value is less than** 0.01 / **0.05** / 0.10 (significance at 1% / 5% / 10% level)

(p 112-114)

A **95% confidence interval** for a coefficient is the **estimated value ± 1.96 × standard error** (p 117-118)

Interpretation of a regression coefficient:

- **linear-linear** ($Y = \beta_0 + \beta_1 X$): an increase of one unit of X is associated with an increase of β_1 units of Y
- **log-linear** ($\ln Y = \beta_0 + \beta_1 X$): an increase of one unit of X is associated with a $100 \times \beta_1$ % increase in Y
- **linear-log** ($Y = \beta_0 + \beta_1 \ln X$): a 1% increase in X is associated with a $0.01 \times \beta_1$ unit increase in Y
- **log-log** ($\ln Y = \beta_0 + \beta_1 \ln X$): a 1 % increase in X is associated with a β_1 % increase in Y (p 210-214)

In a regression $Y = \beta_0 + \beta_1 X$: **Y is the dependent variable** and **X is the independent variable.** (p 94)

The **R^2** of a regression is the **fraction of variation in the dependent variable (Y) that is explained by the regression.** (The **adjusted- R^2** or **R-bar-squared** is the same thing with a small technical adjustment which means you can compared it across regressions with different numbers of variables.) (p 122, 176)

The **standard error of the regression (SER)** is an estimator of the **standard error of the regression error u**. It has the same units as Y. ESS, TSS and SSR are less important. (p 122-123)

An **F-test** tests whether multiple coefficients could all be 0.

- If the **p-value is less than 0.05** (or the F-stat is greater than the critical value) then we can **reject the possibility that all the coefficients are 0.**
- If the **p-value is greater than 0.05** then we **cannot reject that all the coefficients might be 0.** (p 165-169, inside back cover)

There is evidence of a **non-linear effect of X on Y** if when the variable X^2 is added to the regression, its coefficient is significant. (p 206)

For a regression $Y = \beta_0 + \beta_1 X + \beta_2 W + \beta_3 X \times W$: there is evidence for an **interaction of between X and W** (or that **the effect of X on Y depends on W**) if the coefficient of the **interaction term X × W is significant**. (If X is included as X, X^2, X^3 … interaction terms between W and all the powers of X must be included and an F-test done on all the interaction terms) (p 218-229)

Difference in Differences
The following table gives average or predicted values of Y for 4 groups:

	X=1	X=0
W=1	μ_1	μ_2
W=0	μ_3	μ_4

The difference in differences is $(\mu_1 - \mu_3) - (\mu_2 - \mu_4)$. This should be the same as the value of β_3 in the regression $Y = \beta_0 + \beta_1 X + \beta_2 W + \beta_3 X \times W$

The error term is
- **homoskedastic** if the variance of Y is the same for different values of X (ie, the variance of test scores is the same for kids in small classes as large classes).
- **heteroskedastic** is the variance of Y is different for different values of X.

If you assume homoskedasticity wrongly, your standard errors will be too small (but your coefficient unbiased). If you allow for heteroskedasticity, your standard error will be right even if the error term is really homoskedastic. (p 124-126, 129)

You have **omitted variable bias** is there is some factor that is correlated with your independent variables (X etc) and influences the dependent variable Y but is not included in the regression.
- This means that **your coefficients are, on average, wrong.**
- If X is truly randomly assigned then you don't have OVB. (p 144-147)

To **derive an OLS estimator** for a regression $Y = f(X,\beta) + u$ where you get to pick β:
1. Set up what you want to minimize : $\Sigma (Y - f(X, \beta))^2$
2. Take the derivative with respect to β and set = 0.
3. Solve for β.

To **prove an estimator β-hat is unbiased**
1. Use $Y = f(X,\beta) + u$ to write β-hat in terms of β, X and u – preferably as β + some expression in X and u.
2. Take expectations.
3. Use LIE to replace u with E[u | X].
4. Use the first assumption to say that E[u | X] = 0 and simplify. (p 135-137)

Remember:
$E[a+bX] = a + bE[X]$ $\quad var(a+bX) = b^2 var(X))$ $\quad var(X) = E[(X-\mu)^2] = E[X^2] - (E[X])^2$

If A and B are normally distributed with means ma and mb and variance sa and sb (standard errors ea and eb) then A-B is normally distributed with mean ma-mb and variance sa+sb (standard error $\sqrt{(ea + eb)}$)