Regression Analysis:

Regression analysis is a simple method for investigating the functional relationship among the variables. Some of the real world examples for regression analysis are: we may wish to examine whether cigarette consumption is related to various socioeconomic and demographic variables such as age, education, income, price of cigarette etc. The relationship is expressed in the equation form also known as model connecting the response of dependent variable with a set of predictors or explanatory variables.

Regression Analysis in Stata:

Let's dive in and perform regression analysis in Stata. In this chapter, and in subsequent chapters, we will be using a data file that was created by randomly sampling of 400 elementary schools from the California Department of Education's API 2000 dataset. This data file contains a measure of school academic performance as well as other attributes of the elementary schools, such as, class size, enrolment, poverty, etc.

You can access this data file over the web from within Stata with the Stata `use` command as shown below:

```
.use http://www.ats.ucla.edu/stat/stata/webbooks/reg/elemapi
```

As an example we first perform the regression analysis by using academic performance of the school (`api00`) as our response variable and using the average class size in kindergarten through 3rd grade (`acs_k3`), the percentage of students receiving free meals (meals) - which is an indicator of poverty, and the percentage of teachers who have full teaching credentials (full) as a set of explanatory variables.

We expect that better academic performance would be associated with lower class size, fewer students receiving free meals, and a higher percentage of teachers having full teaching credentials.

Stata command for testing this regression followed by the Stata output is given below:

REGRESSION ANALYSIS

```
. regress api00 acs_k3 meals full

      Source |       SS       df       MS              Number of obs =     313
-------------+------------------------------           F(  3,   309) =  213.41
       Model |  2634884.26     3  878294.754            Prob > F      =  0.0000
    Residual |  1271713.21   309  4115.57673           R-squared     =  0.6745
-------------+------------------------------           Adj R-squared =  0.6713
       Total |  3906597.47   312  12521.1457           Root MSE      =  64.153

------------------------------------------------------------------------------
       api00 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      acs_k3 |  -2.681508   1.393991    -1.92   0.055    -5.424424    .0614074
       meals |  -3.702419   .1540256   -24.04   0.000    -4.005491   -3.399348
        full |   .1086104    .090719     1.20   0.232    -.0698947    .2871154
       _cons |   906.7392   28.26505    32.08   0.000     851.1228    962.3555
------------------------------------------------------------------------------
```

The R-squared of 0.6745 means that approximately 67% of the variance of `api00` is accounted for by the variables in the model. Next we check whether our explanatory variables are individually statistically significant and, if so, the direction of the relationship. The average class size (`acs_k3`, b=-2.68), is not significant (p=0.055>0.05), but only just so. Next, the effect of `meals` (b=-3.70) and is significant (p=.000<0.05). Finally, the percentage of teachers with full credentials (`full`, b=0.11) is not related to the academic performance of the school on average because the p-value is (p=0.

The only significant variable is `meals` and its coefficient is negative indicating that the greater the proportion students receiving free meals, the lower the academic performance. It is important to note that we are not saying that free meals are causing lower academic performance. The meals variable is highly related to income level and functions more as a proxy for poverty. Thus, higher levels of poverty are associated with lower academic performance.

You may be wondering what a -2.68 change in `acs_k3` or the average class size really means, and how you might compare the strength of that coefficient to the coefficient of another variable, say `meals`. To address this problem, we can add an option to the `regress` command called `beta`, which will give us the standardized regression coefficients. The beta coefficients are used by some researchers to compare the relative strength of the various predictors within the model. Because the beta coefficients are all measured in standard deviations, instead of the units of the variables, they can be compared to one another. In other words, the beta coefficients are the coefficients that you would obtain if the outcome and predictor variables were all transformed standard scores, also called z-scores, before running the regression. The Stata command and output is given below:

REGRESSION ANALYSIS

```
. regress api00 acs_k3 meals full, beta

      Source |       SS       df       MS              Number of obs =     313
-------------+------------------------------           F(  3,   309) =  213.41
       Model |  2634884.26     3  878294.754            Prob > F      =  0.0000
    Residual |  1271713.21   309  4115.57673           R-squared     =  0.6745
-------------+------------------------------           Adj R-squared =  0.6713
       Total |  3906597.47   312  12521.1457           Root MSE      =  64.153

------------------------------------------------------------------------------
       api00 |      Coef.   Std. Err.      t    P>|t|                     Beta
-------------+----------------------------------------------------------------
      acs_k3 |  -2.681508   1.393991    -1.92   0.055                -.0635654
       meals |  -3.702419   .1540256   -24.04   0.000                -.8075094
        full |   .1086104    .090719     1.20   0.232                 .0408765
       _cons |   906.7392   28.26505    32.08   0.000                        .
------------------------------------------------------------------------------
```

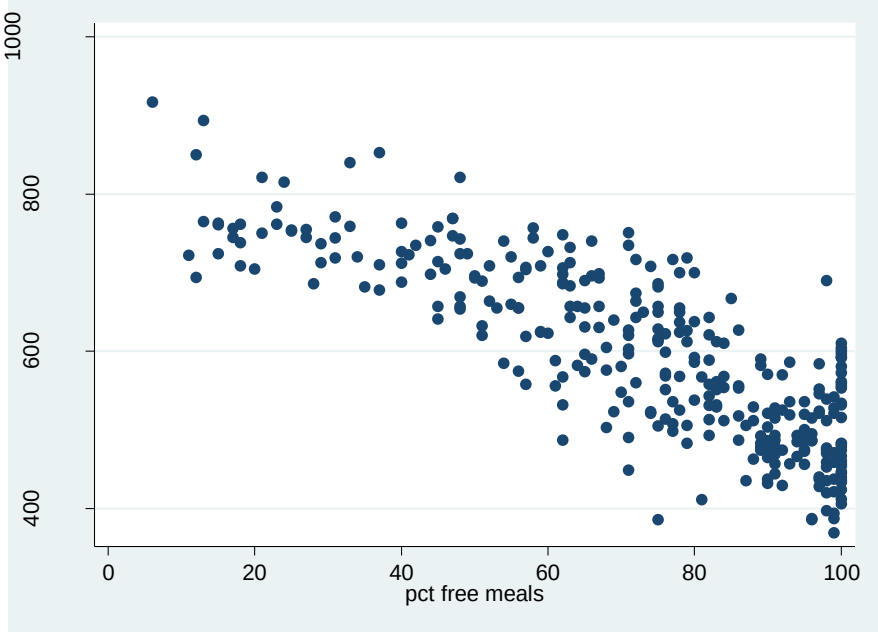
Since, the coefficients in the Beta column are all in the same standardized units you can compare these coefficients to assess the relative strength of each of the predictors. In this example, `meals` has the largest `Beta` coefficient, 0.807 (in absolute value), and full has the smallest `Beta`, 0.041. Thus, a one standard deviation increase in `meals` leads to a 0.81 standard deviation decrease in predicted `api00`, with the other variables held constant. And, a one standard deviation increase in `acs_k3`, in turn, leads to a 0.041 standard deviation increase in predicted `api00` with the other variables in the model held constant.

In interpreting this output, remember that the difference between the numbers listed in the `Coef.` column and the `Beta` column is in the units of measurement. For example, to describe the raw coefficient for `meals` you would say "A one-unit decrease in `meals` would yield a 3.70-unit increase in the predicted `api00`." However, for the standardized coefficient (`Beta`) you would say, "A one standard deviation decrease in `meals` would yield a 0.81 standard deviation increase in the predicted `api00`."

Finally, as part of doing a regression analysis you might be interested in seeing the correlations among the variables in the regression model. You can do this with the `correlate` command as shown below.

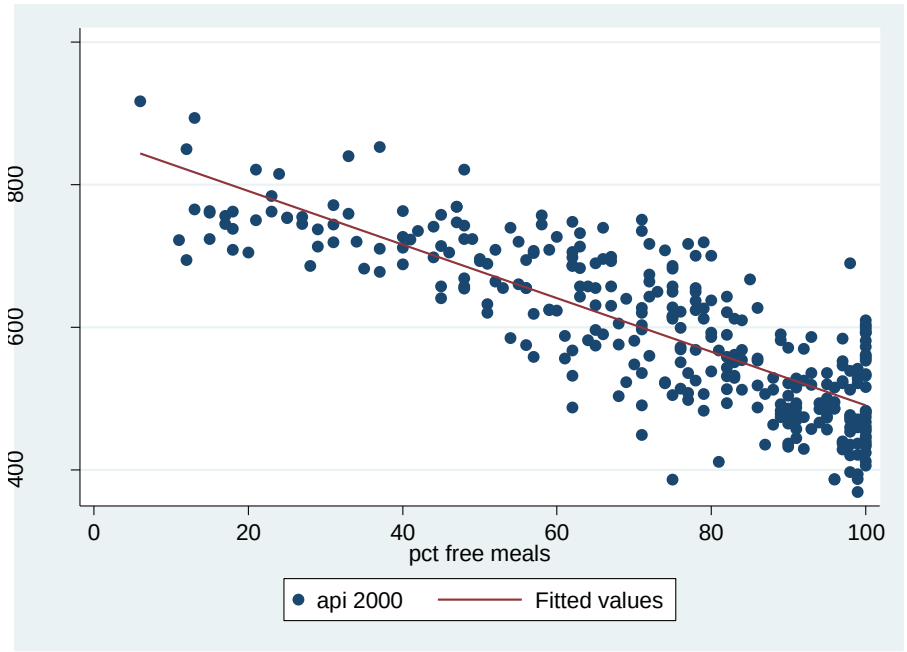
```
. correlate api00 acs_k3 meals full
(obs=313)

             |   api00   acs_k3   meals    full
-------------+------------------------------------
       api00 |  1.0000
      acs_k3 | -0.0641  1.0000
       meals | -0.8184  0.0097  1.0000
        full |  0.2328  0.1789 -0.2518  1.0000
```

In addition to getting the regression table, it can be useful to see a scatter plot of the predicted and outcome variables with the regression line plotted. Below we can show a scatter plot of the outcome variable, `api00` and the predictor, `meals` by typing `scatter api00 meals` in the command window:



We can combine **scatter** with **lfit** to show a scatter plot with fitted values by typing the following command in the command window:

```
twoway (scatter api00 meals) (lfit api00 meals)
```



The .do file content:

```
. // # 6 Regression Analysis in Stata


. capture log close
. log using RegressionAnalysis_output.txt, text replace

. use http://www.ats.ucla.edu/stat/stata/webbooks/reg/elemapi , clear


. regress api00 acs_k3 meals full

. regress api00 acs_k3 meals full, beta


. correlate api00 acs_k3 meals full

. scatter api00 meals

. twoway (scatter api00 meals) (lfit api00 meals)

. log close
```