AGRODEP Stata Training

April 2013

# Module 3

# Linear Regressions

Manuel Barron[1] and Pia Basurto[2]

[1] University of California, Berkeley, Department of Agricultural and Resource Economics
[2] University of California, Santa Cruz, Department of Economics

AGRODEP Stata Training documents are designed to give AGRODEP members a brief overview of basic Stata commands needed in AGRODEP training courses. These documents have been reviewed but have not been subject to a formal external peer review via IFPRI's Publications Review Committee; any opinions expressed are those of the author(s) and do not necessarily reflect the opinions of AGRODEP or of IFPRI.

# Module 3 – Linear Regressions

In this module we will cover basic commands that you will need in order to run linear regressions and two-stage least squares (2SLS). We will show you how to generate predicted values of the dependent variable and residuals. We will also give examples on how to use the *outreg* command.

For this module we will use hhmembers_2.dta, available in the AGRODEP website. This dataset has some variables from the Ethiopian Demographic and Health Survey.

## 1. Linear Regressions

### 1.1 regress

Stata has a very large set of commands that implement different types of estimations. This module will introduce you to basic linear regressions. The command in Stata to run a linear regression is *regress.*

*** Do-file or Command Window**

```
help regress
```

The help window will appear. Let's see how to read a Stata help file.

***Help File**

```
regress depvar [indepvars] [if] [in] [weight] [, options]
```

In the syntax for *regress*, *depvar* is the dependent variable or left-hand-side variable (usually denoted as Y in econometrics textbooks) and *indepvars* includes all the independent, or right-hand-side variables you want to include in the regression (usually denoted as X in econometrics textbooks).

Note that only the first three letters are underlined. As noted in Module 1, this means that you can type *reg* or *regress* interchangeably and Stata will know that you are referring to the *regress* command. We will use *regress* in the first example and *reg* in the rest.

As usual, you can run this estimation for a subset of observations using the "if" and "in" options, and you can also include weights for the observations.

The most common options for *regress* are *noconstant*, which drops the constant from the estimation, and [vce(*vcetype*)], which specifies the type of covariance matrix matrix to be used for calculating the standard errors of the coefficients. vcetype can be robust (the Huber-White "sandwich" estimator), cluster, bootstrap, jackknife, hc2, or hc3.

Let's use "hhmembers_2.dta" to run a regression of hours worked in the past week explained by sex, age and years of education.

In your do-file or in the command window, type:

*** Do-file or Command Window**

```
use hhmembers_2.dta, clear

regress hours_worked sex age years_education
```

Stata will produce the following output:

***Stata Output**

```
      Source |       SS       df       MS              Number of obs =     954
-------------+------------------------------           F(  3,   950) =   11.35
       Model |  13848.0172     3  4616.00572           Prob > F      =  0.0000
    Residual |  386476.856   950  406.817743           R-squared     =  0.0346
-------------+------------------------------           Adj R-squared =  0.0315
       Total |  400324.873   953  420.068073           Root MSE      =   20.17
------------------------------------------------------------------------------
hours_worked |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sex |   3.275817   1.307172     2.51   0.012     .7105386    5.841095
         age |   .7718197   .3105787     2.49   0.013     .1623201    1.381319
years_educ~n |  -1.987406   .3746954    -5.30   0.000    -2.722733    -1.25208
       _cons |   10.08002   3.106916     3.24   0.001     3.982807    16.17723
------------------------------------------------------------------------------
.
```

| | | |
|---|---|---|
| SS | = | Sum of squares |
| df | = | Degrees of freedom |
| MS | = | Mean squares |
| Number of obs | = | Number of observations used in the regression |
| F() | = | F value from the joint test of significance of the model |
| Prob > F | = | p-value of the F test |
| R-squared | = | Model’s R-Squared |
| Adj R-squared | = | Model’s Adjusted R-squared |
| Root MSE | = | Root Mean Squared Error |

As you can see, the name of the “years_education” variable is too long to fit in the output. Stata will shorten it to the first 10 characters, followed by ~ and the last character.

When you run a regression, Stata stores the estimation results in its memory until you run a new regression. This is quite useful if, for example, you want to generate the predicted values of the dependent variable, or the residual of the model for each observation. To do this, use the *predict* command followed by the name of the new variable you want to generate (containing the predicted values). We usually want to do in-sample predictions only, so the command will usually be issued with the “if e(sample)” option.

***DO-file or command window**

```
predict yhat if e(sample)
```

If you use the *res* option, you will generate the residuals. In this case we named the residuals “ehat”.

***DO-file or command window**

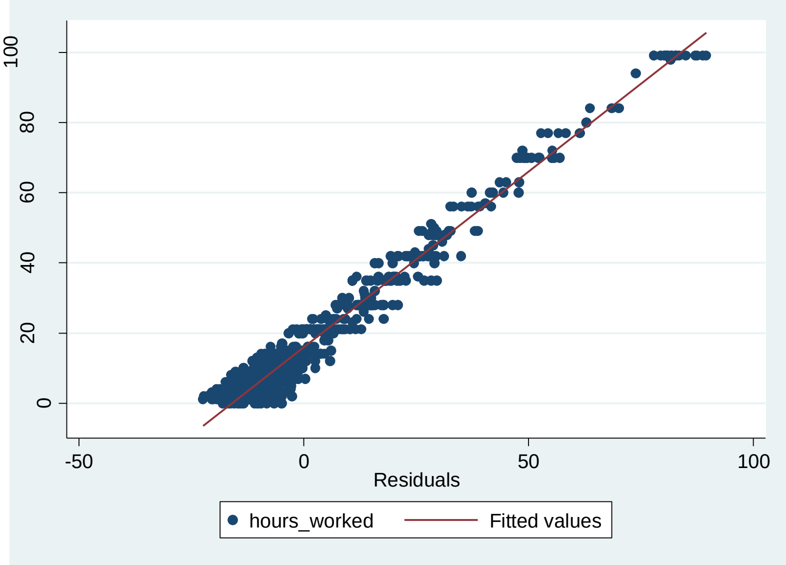
```
predict ehat if e(sample), res
```

### 1.2 Graphical Analysis

You may want to graph the dependent variable with the error term:

***DO-file or command window**

```
graph twoway (scatter hours_worked ehat ) (lfit hours_worked ehat )
```



Where the *lfit* command fits a line through the scatterplot, as shown in the graph.

### 1.3 vcetype

If you don’t specify *vcetype*, Stata will assume homoscedasticity. You may want to use the *robust* option to calculate standard errors that are robust to heteroscedasticity (Huber-White sandwich estimator of the residual covariance matrix). Note our use of *reg* as short form for *regress*:

***DO-file or command window**

```
reg hours_worked sex age years_education [pweight = hh_weight], vce(robust)
```

*Stata output

```
Linear regression                                      Number of obs =     954
                                                       F(  3,   950) =    4.98
                                                       Prob > F      =  0.0020
                                                       R-squared     =  0.0271
                                                       Root MSE      =  19.444

------------------------------------------------------------------------------
             |               Robust
hours_worked |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sex |   4.655033   1.654987     2.81   0.005      1.40718    7.902886
         age |   .4384048   .3818904     1.15   0.251    -.3110415    1.187851
years_educ~n |   -1.33224   .4474231    -2.98   0.003    -2.210292   -.4541881
       _cons |   12.15444   3.942635     3.08   0.002     4.417159    19.89172
------------------------------------------------------------------------------
.
```

### 1.4 areg

There may be instances when you want to run a regression with several region dummies and you are not interested in their coefficients. You may use *regress* as in the previous examples and include dummies in the regression, but the output window will be cluttered (try it!). A useful alternative is the *areg* command.

*** Do-file or Command Window**

```
help areg
```

The help window will appear. Let’s review the Stata help file.

*Help File

```
areg depvar [indepvars] [if] [in] [weight], absorb(varname) [options]
```

Now type the following regression in your do-file or command window:

*** Do-file or Command Window**

```
areg hours_worked sex age years_education , absorb(region)
```

*** Stata output**

```
Linear regression, absorbing indicators                Number of obs =     954
                                                       F(  3,   940) =    8.41
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.1018
                                                       Adj R-squared =  0.0894
                                                       Root MSE      =  19.558

------------------------------------------------------------------------------
hours_worked |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sex |   3.861399   1.283259     3.01   0.003     1.343015    6.379783
         age |   .6815821   .3050005     2.23   0.026     .0830214    1.280143
years_educ~n |   -1.53328   .3792266    -4.04   0.000    -2.277509   -.7890515
       _cons |   9.825308   3.034284     3.24   0.001     3.870554    15.78006
-------------+----------------------------------------------------------------
      region |        F(10, 940) =      7.035   0.000          (11 categories)
------------------------------------------------------------------------------
.
```

### 1.5 xi: reg

If you are interested in the coefficients, you may find the *xi: reg* command useful. It allows us to use dummies in our regression without creating them by hand. This command is also quite advanced, so we present an example of its basic use, but will not discuss its more advanced options. For more details, see *help xi*.

*** Do-file or Command Window**

```
xi: reg hours_worked sex age years_education i.region
```

***Stata output**

```
      Source |       SS       df       MS              Number of obs =     954
-------------+------------------------------           F( 13,   940) =    8.20
       Model |  40758.8983    13  3135.29987           Prob > F      =  0.0000
    Residual |  359565.975   940  382.516995           R-squared     =  0.1018
-------------+------------------------------           Adj R-squared =  0.0894
       Total |  400324.873   953  420.068073           Root MSE      =  19.558

------------------------------------------------------------------------------
hours_worked |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sex |   3.861399   1.283259     3.01   0.003     1.343015    6.379783
         age |   .6815821   .3050005     2.23   0.026     .0830214    1.280143
years_educ~n |   -1.53328   .3792266    -4.04   0.000    -2.277509   -.7890515
   _Iregion_2 |  -1.401783   3.491143    -0.40   0.688     -8.25312    5.449554
   _Iregion_3 |   2.294028   2.172206     1.06   0.291    -1.968907    6.556964
   _Iregion_4 |  -1.150637   2.422096    -0.48   0.635    -5.903978    3.602704
   _Iregion_5 |   16.71634   3.224495     5.18   0.000      10.3883    23.04438
   _Iregion_6 |   -2.85313   2.797126    -1.02   0.308    -8.342464    2.636203
   _Iregion_7 |  -3.034862   2.482459    -1.22   0.222    -7.906665     1.83694
  _Iregion_12 |  -8.287519   2.971516    -2.79   0.005    -14.11909   -2.455945
  _Iregion_13 |   -1.53776   4.137092    -0.37   0.710    -9.656766    6.581246
  _Iregion_14 |  -4.961679   3.744623    -1.33   0.185    -12.31047    2.387109
  _Iregion_15 |  -9.805673   3.370196    -2.91   0.004    -16.41965   -3.191693
       _cons |    10.4676   3.394033     3.08   0.002     3.806837    17.12835
------------------------------------------------------------------------------
.
```

## 2. Instrumental variables

### 2.1 ivregress

Stata has several commands to implement instrumental variables. The two most common commands are *ivregress* and *ivreg2*. We will focus on *ivregress*.

***Do-file or command window**

```
help ivregress
```

***Help file**

```
ivregress estimator depvar [varlist1] (varlist2 = varlist_iv) [if] [in] [weight] [,
options]
```

*ivregress* allows for three types of estimators:

| | | |
|---|---|---|
| 2sls | : | two-stage least squares (2SLS) |
| liml | : | limited-information maximum likelihood (LIML) |
| gmm | : | generalized method of moments (GMM) |

To indicate the endogenous variable to be instrumented, you need to put it between parentheses, followed by an equal sign and the exogenous instrument(s).

*** Do-file or Command Window**

```
ivregress 2sls hours_worked sex age (years_education = head_sex)
```

***Stata output**

```
Instrumental variables (2SLS) regression               Number of obs =     954
                                                       Wald chi2(3)  =    0.80
                                                       Prob > chi2   =  0.8492
                                                       R-squared     =       .
                                                       Root MSE      =  55.529

------------------------------------------------------------------------------
hours_worked |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
years_educ~n |  -31.68254   215.2345    -0.15   0.883    -453.5345    390.1694
         sex |   5.715456   18.04513     0.32   0.751    -29.65236    41.08327
         age |   14.40654   98.82894     0.15   0.884    -179.2946    208.1077
       _cons |  -76.92911   630.7051    -0.12   0.903    -1313.088     1159.23
------------------------------------------------------------------------------
Instrumented:  years_education
Instruments:   sex age head_sex

.
```

To report the first stage statistics you may issue the *estat* command:

*** Do-file or Command Window**

```
estat firststage
```

*** Stata output**

```
  First-stage regression summary statistics
  --------------------------------------------------------------------------
               |            Adjusted      Partial
      Variable |   R-sq.       R-sq.        R-sq.      F(1,950)   Prob > F
  -------------+------------------------------------------------------------
  years_educ~n |  0.3077      0.3055       0.0000       .021822    0.8826
  --------------------------------------------------------------------------


  Minimum eigenvalue statistic = .0218225

  Critical Values                      # of endogenous regressors:    1
  Ho: Instruments are weak             # of excluded instruments:     1
  ---------------------------------------------------------------------
                                     |    5%     10%     20%     30%
  2SLS relative bias                 |         (not available)
  -----------------------------------+---------------------------------
                                     |   10%     15%     20%     25%
  2SLS Size of nominal 5% Wald test  |  16.38    8.96    6.66    5.53
  LIML Size of nominal 5% Wald test  |  16.38    8.96    6.66    5.53
  ---------------------------------------------------------------------
```

## 2.2 ivreg2

*ivreg2* is an extended instrumental variables command. You first need to install the program. You may do so by typing *"findit ivreg2"* and following the on-screen instructions or by typing "*ssc install ivreg2"*. Once the program is installed, you can access the help file by typing "*help ivreg2"*.

*ivreg2* will produce the exact same results as *ivregress*, but it has some advanced options. We will not cover them in these introductory notes.

*** Do-file or Command Window**

```
ivreg2 hours_worked sex age (years_education = head_sex)
```

*** Stata output**

```
IV (2SLS) estimation
--------------------

Estimates efficient for homoskedasticity only
Statistics consistent for homoskedasticity only

                                                      Number of obs =      954
                                                      F(  3,   950) =     0.27
                                                      Prob > F      =   0.8500
Total (centered) SS     =  400324.8732                Centered R2   =  -6.3481
Total (uncentered) SS   =       644197                Uncentered R2 =  -3.5663
Residual SS             =   2941611.56                Root MSE      =    55.53

------------------------------------------------------------------------------
hours_worked |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
years_educ~n |  -31.68254   215.2345    -0.15   0.883    -453.5345    390.1694
         sex |   5.715456   18.04513     0.32   0.751    -29.65236    41.08327
         age |   14.40654   98.82894     0.15   0.884    -179.2946    208.1077
       _cons |  -76.92911   630.7051    -0.12   0.903    -1313.088     1159.23
------------------------------------------------------------------------------
Underidentification test (Anderson canon. corr. LM statistic):           0.022
                                                   Chi-sq(1) P-val =    0.8823
------------------------------------------------------------------------------
Weak identification test (Cragg-Donald Wald F statistic):                0.022
Stock-Yogo weak ID test critical values: 10% maximal IV size             16.38
                                         15% maximal IV size              8.96
                                         20% maximal IV size              6.66
                                         25% maximal IV size              5.53
Source: Stock-Yogo (2005).  Reproduced by permission.
------------------------------------------------------------------------------
Sargan statistic (overidentification test of all instruments):           0.000
                                                 (equation exactly identified)
------------------------------------------------------------------------------
Instrumented:         years_education
Included instruments: sex age
Excluded instruments: head_sex
------------------------------------------------------------------------------
```

### 3. outreg

When you are running multiple regressions in Stata and you want to present the results in a nice looking table you may find it useful to use the *outreg* command. This is a user-written package and you need to install it before you using it for the first time.

Type in your do-file or command window:

***Do-file or command window**

```
ssc install outreg
```

Now that *outreg* is installed you can look at the help file to learn about all the options that *outreg* allows you to do. The most common options are: include a title, include additional statistics like the mean or a p-value of a T-test using *addstat*, and report standard errors instead of t-statistics. In addition, if you want to have several regressions in different columns of the same table, you may use the option "append" instead of "replace". This is a very advanced command and you will only see a very simple example here.

***Do-file or command window**

```
reg hours_worked sex age years_education

outreg using reg_module3, replace  se ctitle("Example 1: Hours worked")
```

Now you will have a file with the extension .out in the directory you have been working on. You can open this file from excel or with a text editor like Notepad or Word. Note that the default is a .doc file. The stored results will look like this:

| Linear Regression | |
|---|---|
| | hours_worked |
| Sex | 3.276 |
| | (1.307)* |
| Age | 0.772 |
| | (0.311)* |
| years_education | -1.987 |
| | (0.375)** |
| Constant | 10.08 |
| | (3.107)** |
| Observations | 954 |
| R-squared | 0.03 |
| Standard errors in parentheses | |
| * significant at 5%; ** significant at 1% | |

By default the coefficients that are statistically significant will have a star (or two) next to them indicating the significance level of 5% (or 1%).

## 4. Wrapping up

In this module we have covered linear regressions and instrumental variables methods through two stage least squares. We showed you how to generate predicted values of the dependent variable and residuals, and we illustrated the use of the *outreg* command.