Stata How-to: OLS Regressions

Patrick Blanchenay

2021-12-27

Contents

1. Basic specifications (univariate and multivariate)

To estimate the following simple regression:

$$y_i = \alpha + \beta x_i + e_i \tag{1}$$

using Ordinary Least Squares, the command in Stata is:

```
reg y x, robust
```

The official command is **regress** but everyone uses its abbreviated form **reg**.

Note the use of the **robust** option, which allows for the possibility of heteroskedasticity (Lecture 5). In practice, since there are very few situtations where the error term should be homoskedastic, we always use the **robust** option.[1]

To estimate the same model but with additional controls w_1 and w_2, that is:

$$y_i = \alpha + \beta x_i + \gamma_1 w_{1i} + \gamma_2 w_{2i} + e_i \tag{2}$$

we would use:

```
reg y x w1 w2, robust
```

It is important to remember that Stata treats variables x, w_1 and w_2 in the same way. Indeed, from a statistical point of view, the treatment variable and control variables are just the same: a regressor. Only the context and the specific research question will determine which one we consider the treatment variable.

[1] It is also possible to account for even more complex standard errors, but this is beyond the scope of this document.

2. Discrete (categorical) variables

Discrete (categorical) variables are variables that take a limited (countable) number of values. For instance, a variable coding "male" or "female" can take two values.

If the variable takes more than one value, the variable can be ordinal or not. A variable is ordinal whenever it makes sense to talk about the order of values. For instance, a variable coding individuals' age by decade (0-9,10-19,...) has an ordinal meaning: belonging to the 30-39 category makes you older than a person belonging to the 20-29 category. A categorical variable is NOT ordinal if the categories do not have a natural order. For instance, data on Canadian companies might record the province they belong to; there is no order betwen ON, AL, QC....

3. Using discrete (categorical) variables as dummy variables

If the variable is ordinal then it can be included like any continuous regressor. If our variable is already coded as 1 or 0, we can use it as a normal regressor. For instance, a variable `female` equal to 1 for women and 0 for men, could just be included directly.

```
reg lincome female, robust
```

If the variable is not ordinal, using it like a normal regressor would lead to non-sensical results (the coefficient would not have a logical interpretation). In such case, we should convert such variable to a series of dummy variables.

Imagine we would like to run the regression in lecture 4, explaining future (log)-earnings $\ln Y_i$ with attendance of a private university P_i:

$$\ln Y_i = \alpha + \beta P_i + e_i \tag{3}$$

As in MM ch.2, we would like to control for the potential of students, by grouping them according to the universities they applied to and were accepted to. There are 151 such "potential" groups, and each student belongs to one of these. For each student, the variable `group` takes a value between 1 and 151 based on the particular group this student belongs to.

Since the `group` variable is not necessarily ordinal, it doesn't make sense to include it directly in our regression; however we would like to include one dummy for each group in our regression, that is we would like to estimate:

$$\ln Y_i = \alpha + \beta P_i + \sum_{j=2}^{151} \left(\gamma_j \cdot \text{Group-}j_i\right) + e_i \tag{4}$$

where Group-$j_i = 1$ if student i belongs to group j, and 0 otherwise.[2] One tedious way to do this in Stata would be to manually create one dummy for each unique value for the `group` variable, that is create:

- a variable `group2` equal to 1 whenever variable`group` is equal to 2, and 0 otherwise
- a variable `group3` equal to 1 whenever variable `group` is equal to 3, and 0 otherwise
- ...
- a variable `group151` equal to 1 whenever variable `group` is equal to 151, and 0 otherwise

We would then run the regression including those dummies as controls:

[2] Remember that with mutually exclusive and exhaustive dummies, we should always leave one out, here Group-1.

```
// Generating the dummies individually (tedious)
gen group2 = (group == 2) // the part in brackets equals  1 when group is equal to 2, and 0 otherwise
gen group3 = (group == 3)
...
gen group151 = (group == 151)

// Running the OLS regression
reg logearnings private group2 group3 ... group151 , robust
```

There is however a much shorter way to do this in Stata, which is to use the **i.** indicator operator. Stata then creates dummy variables "on the fly", one per value of the original variable.

Stata transforms any discrete variable into a series of dummies, one for each unique value of the original variable. The dummy variables are generated automatically, used for the regression, then erased automatically once the regression is run.

Our previous tedious example could be replaced by a single line, using **i.** in front of the `group` variable:

```
// Running the OLS regression with automatically generated dummies
reg logearnings private i.group , robust
```

As usual with dummy variables, one should be left out to avoid "dummy trap". For instance, if we have dummies for seasons, we should leave one out (for instance Winter).

By default, the dummy corresponding to `group==1` is left out, and group 1 is used as the reference group. It is also possible to use a different value of `group` as a reference group (or baseline level), which would then be left out. To do that, we use **ib[#group].** in front of `group`, where **[#group]** is the value to be used as a reference. For instance, if we want to use group 151 as a reference group (instead of group 1), we type:

```
// Changing the baseline group
reg logearnings private ib151.group , robust
```

which would include dummies for groups 1, 2, ..., 150, and leave out the dummy for group 151.

4. Interaction terms (heterogeneous effects) – Lecture 6

In the same way, Stata can create interaction terms "on the fly". To estimate

$$\ln Y_i = \alpha + \beta P_i + \gamma \text{Female}_i + \delta \left(P_i \times \text{Female}_i\right) + e_i \tag{5}$$

we could create the interaction term manually and use it as a regression (the tedious way):

```
// Creating interaction terms manually (tedious)
gen privatefemale = private * female
label variable privatefemale "private x female"
reg logearnings private female privatefemale, robust
```

or we can let Stata create them on the fly using the **#** operator placed between the two regressors:

```
// Using interaction terms on the fly (recommended)
reg logearnings private female private#female, robust
```

If you want to include each variable separately as well as their product, you can use the **##** operator for a more compact command. The command below performs exactly the same thing as the command above.

```
// Using interaction terms on the fly (recommended)
reg logearnings private##female, robust
```

Optional but encouraged: specify if variables are continuous or categorical/dummy, by using the **`c.`** for continuous variables, and **`i.`** prefix for categorical variables, including dummies. For instance:

```
// Using interaction terms on the fly
reg logearnings i.private##c.parentalincome, robust
```

See section 5.2 for an example.

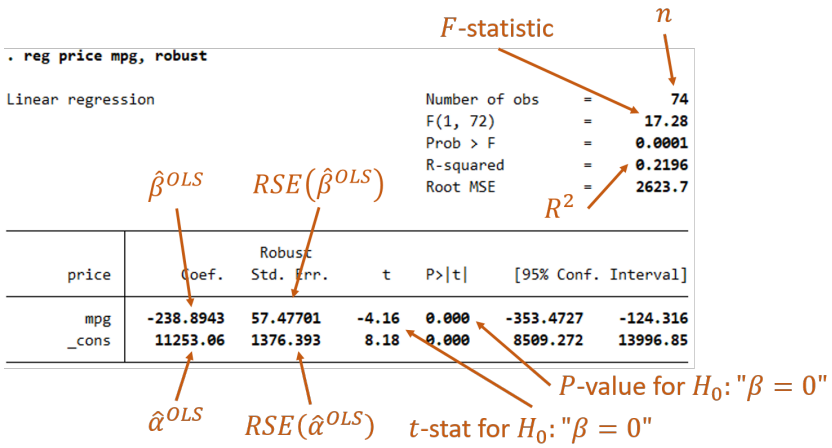
5. Example

Let's use the system dataset `auto` supplied with Stata about prices and characteristics of 72 cars (**`sysuse auto, clear`**). We use it to see how car prices (`price`) depends on their range, measured in miles-per-gallon (`mpg`). (Cars with higher `mpg` can drive longer distances before having to refuel.)

We would like to estimate the following model:

$$\text{Price}_i = \alpha + \beta \text{MPG}_i + u_i \tag{6}$$

```
sysuse auto, clear // Clears memory and loads a default dataset included in Stata
reg price mpg, robust
```



At the top of the output, a number of information is displayed about the model you estimated, in particular the sample size n, the F-statistic (test of all coefficients jointly null), and the R^2 (fraction of the variation in the dependent variable explained by variation in the regressors).

Next, a panel reports information on each regressor in the model. The column "coef." reports the estimated coefficients for each regressor. Here since we estimate a simple regression, we have two coefficients, $\hat{\alpha}$ (intercept) and $\hat{\beta}$ (slope). Stata refers to the intercept as **`_cons`**. The column next to it reports, for each coefficient, the associated (robust) standard error. The column "t" reports the t-statistic associated with the null hypothesis H_0 that this specific coefficient is zero, and the two-sided alternative hypothesis that is different from 0. Stata also displays the associated P-value for each coefficient. (As a reminder, in this case the P-value is the probability, if

the null hypothesis (no effect) is true, of observing a coefficient as far from zero as the one obtained here. The lowest the P-value, the stronger is the evidence in favour of the alternative hypothesis.)

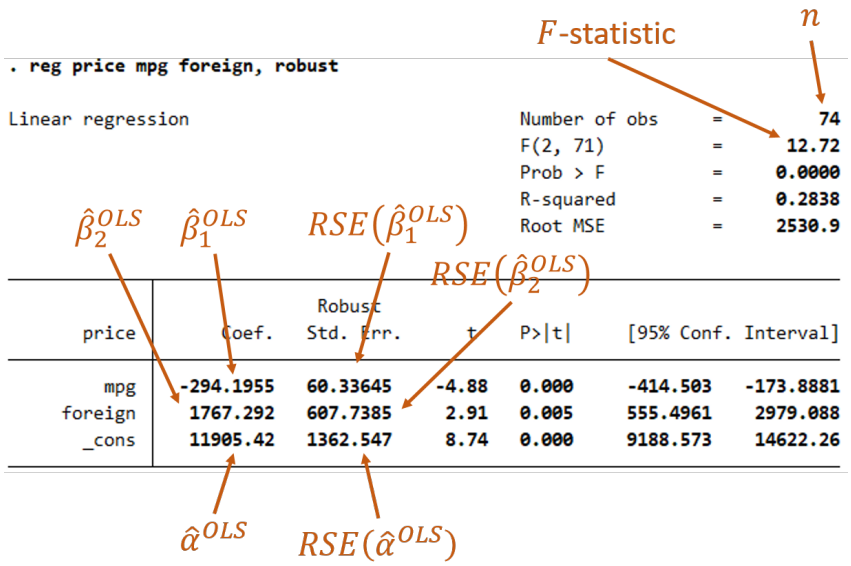
If we want to control for whether a car is domestic or foreign (as imported cars might be more expensive), we might want to estimate the following model:

$$\text{Price}_i = \alpha + \beta_1 \text{MPG}_i + \beta_2 \text{Foreign}_i + e_i \tag{7}$$

which is done with the following command:

```
reg price mpg i.foreign, robust
```

Note that since the variable `foreign` is already a dummy variable, it is not necessary to use the **`i.`** operator. The output now contains an additional line for the extra regressor that we added:




```
. reg price mpg foreign, robust

Linear regression                               Number of obs     =         74
                                                F(2, 71)          =      12.72
                                                Prob > F          =     0.0000
                                                R-squared         =     0.2838
                                                Root MSE          =     2530.9

------------------------------------------------------------------------------
             |               Robust
       price |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         mpg |  -294.1955   60.33645    -4.88   0.000    -414.503   -173.8881
     foreign |    1767.29   607.7385     2.91   0.005    555.4961    2979.088
       _cons |   11905.42   1362.547     8.74   0.000    9188.573    14622.26
------------------------------------------------------------------------------
```

5.1. Example with categorical dummies

Let's use the system dataset `nlsw88`, which contains earning of women, as well information about their occupation and their level of education. We could try to find out whether more educated earn more by estimating:

$$\text{wage}_i = \alpha + \beta \text{grade}_i + \varepsilon_i \tag{8}$$

with the following command:

```
reg wage grade, robust // regressing wage on grade
```

which gives the following output:

```
Linear regression                               Number of obs     =      2,244
                                                F(1, 2242)        =     286.47
                                                Prob > F          =     0.0000
                                                R-squared         =     0.1059
                                                Root MSE          =     5.4454

------------------------------------------------------------------------------
             |               Robust
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       grade |   .7431729   .0439086    16.93   0.000     .6570671    .8292786
       _cons |  -1.965886   .5558785    -3.54   0.000    -3.055976   -.8757954
------------------------------------------------------------------------------
```

The estimate for β is $\hat{\beta} = 0.743$ and is statistically signifcant at the 95% confidence level: more educated women do earn more. Is this because more educated women tend to have different occupations than less educated ones? Let's control for occupations, that is, estimate the following:

$$\text{wage}_i = \alpha + \beta \text{grade}_i + \sum_{j=2}^{13} \left(\gamma_j \cdot \text{Occup-}j_i\right) + e_i \tag{9}$$

We generate the series of occupation dummies "on the fly" by using the **i.** operator:

```
reg wage grade i.occupation, robust
```

```
Linear regression                               Number of obs     =      2,235
                                                F(12, 2221)       =          .
                                                Prob > F          =          .
                                                R-squared         =     0.1767
                                                Root MSE          =     5.2448
```

wage	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
grade	.5762547	.0589149	9.78	0.000	.4607207	.6917887
occupation						
Managers/admin	.9404403	.5697132	1.65	0.099	-.1767859	2.057667
Sales	-2.420281	.4070469	-5.95	0.000	-3.218513	-1.622049
Clerical/unskilled	-1.172196	.9131987	-1.28	0.199	-2.963009	.6186161
Craftsmen	-2.286248	.6041603	-3.78	0.000	-3.471026	-1.10147
Operatives	-2.991275	.4633233	-6.46	0.000	-3.899867	-2.082683
Transport	-5.01789	.5061537	-9.91	0.000	-6.010474	-4.025306
Laborers	-3.956103	.4351453	-9.09	0.000	-4.809437	-3.102769
Farmers	-3.377415	.3585346	-9.42	0.000	-4.080513	-2.674317
Farm laborers	-4.439484	.8993999	-4.94	0.000	-6.203236	-2.675731
Service	-2.955104	.7514278	-3.93	0.000	-4.428678	-1.481529
Household workers	-2.735038	.4027501	-6.79	0.000	-3.524844	-1.945232
Other	-3.03948	.4661365	-6.52	0.000	-3.953589	-2.125371
_cons	2.20887	.9115622	2.42	0.015	.4212671	3.996473

The estimate now is $\hat{\beta} = 0.576$, showing indeed that women of different educational levels work in different occupations.[3]

5.2. Example with an interaction term

Suppose we want to know whether the relationship between education and wages is different for married versus non-married women. We want to estimate:

$$\text{wage}_i = \alpha + \beta \text{grade}_i + \gamma \text{married}_i + \delta \left(\text{grade}_i \times \text{married}_i\right) + \varepsilon_i \tag{10}$$

where married_i is a dummy equal to 1 for married women. Indeed, in this specification, one additional grade is associated with hourly wages higher by $\$\beta$ for unmarried women; and with wages higher by $\$(\beta+\delta)$ for married women.[4]

We estimate equation (10), which contains an interaction term, by using the **##** operator:

```
reg wage c.grade##married, robust
```

We use the **c.** prefix in **c.grade** to tell Stata that grade is a continuous variable (not a categorical variable).

[3] We will explain this reasoning in much more details in class.

[4] Set married equal to 0 in equation (10); the slope is β. Now set married equal to 1; the slope is $(\beta + \delta)$.

```
Linear regression                               Number of obs     =      2,244
                                                F(3, 2240)        =     100.05
                                                Prob > F          =     0.0000
                                                R-squared         =     0.1082
                                                Root MSE          =     5.4407

------------------------------------------------------------------------------
             |               Robust
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       grade |   .7923671   .0783891    10.11   0.000     .6386442    .9460899
             |
     married |
     married |   .4982313   1.187612     0.42   0.675    -1.830703    2.827166
             |
married#c.gr |
     married |  -.0793654   .0937022    -0.85   0.397    -.2631177    .1043869
             |
       _cons |  -2.261485   .9959107    -2.27   0.023    -4.214489   -.3084807
------------------------------------------------------------------------------
```

We would obtain exactly the same results by manually creating the interaction term:

```
generate gradeXmarried = grade * married
reg wage grade married gradeXmarried, robust
```

6. Practice

Using a well-structured do-file, use the system dataset `nlsw88` to answer the following:

1. ☐ Do wages depend on tenure (how long individuals have been in their current position)?
2. ☐ What elements of this estimation change when you use the **`robust`** option?
3. ☐ Does the relationship change once we control for the industry in which individuals work?
4. ☐ Does the relationship between tenure and wages depend on whether the individuals are unionized or not?