Lab 6 - Simple Regression*

Spring 2018

Contents

*Updated from Stata tutorials provided by Prof. Cichello

1 Thinking About Regression

- Today we will look at the following simple linear regression:

$$wage = \beta_0 + \beta_1 educ + u \tag{1}$$

  which models the relationship between education and earnings in a linear fashion.

- It is important to keep in mind that
  - the fact that more educated people tend to earn more than less educated people does not mean necessarily that schooling causes earnings to increase
  - even without resolving the causality problem, it is clear that education predicts earnings, at least in a narrow statistical sense

- This *predictive power* is summarized by the so-called **conditional expectations function (CEF)**, which is the expectation (or population average) of *wage*, with *educ* held fixed:

$$CEF = E(wage|educ) \tag{2}$$

  Note that in a dataset you will have multiple observations for your dependent and independent variables and usually, when looking at the relationship between education and earnings, observations are at the individual level. Indexing each individual with the letter i, we can rewrite equation (1) as:

$$wage_i = \beta_0 + \beta_1 educ_i + u_i \tag{3}$$

  For simplicity, we omit the individual index, but it is always important to know what we are dealing with!

- One can prove that:

$$wage = E(wage|educ) + u \tag{4}$$

  where $E(u|educ)=0$, which implies that the error term is uncorrelated with any function of the variable *educ*.

  !! In words, equation (4) simply states that our random variable *wage* can be decomposed into a piece that is explained by *educ*, i.e. the CEF, and a piece left over u. Of course, this decomposition applies for general random y, x and u.

- Going back to the linear regression in equation (1), we now understand that it is a particular case of equation (4), where the CEF is assumed to be a linear function.

- **When you run a regression in Stata to estimate $\hat{\beta}_0$ and $\hat{\beta}_1$, it is always important to understand what you are doing and what you are trying to estimate**!

- If you are interested in understanding more about econometric methods, I suggest you to have a look at "Mostly Harmless Econometrics: An Empiricist's Companion" by Angrist-Pischke, Princeton University Press, 2008.

2 Regression Output

- We will use `wage1`, load the dataset using `bcuse` command.
- Since we are estimating the equation using **Ordinary Least Squares (OLS)** and we are in a simple bivariate case where there are a single regressor and a constant term, the coefficients in equation (1) will be equal to:

$$\beta_1 = \frac{Cov(wage, educ)}{Var(educ)} \tag{5}$$

and

$$\beta_0 = E(wage) - \beta_1 E(educ) \tag{6}$$

- What do you expect the sign of the slope coefficient (β_1) to be?
- Regress `wage` on `educ`

```
. reg wage educ
```

```
      Source |       SS       df       MS              Number of obs =     526
-------------+------------------------------           F(  1,   524) =  103.36
       Model |  1179.73204     1  1179.73204           Prob > F      =  0.0000
    Residual |  5980.68225   524  11.4135158           R-squared     =  0.1648
-------------+------------------------------           Adj R-squared =  0.1632
       Total |  7160.41429   525  13.6388844           Root MSE      =  3.3784

------------------------------------------------------------------------------
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .5413593    .053248    10.17   0.000     .4367534    .6459651
       _cons |  -.9048516   .6849678    -1.32   0.187    -2.250472    .4407687
------------------------------------------------------------------------------
```

- First let's look at the coefficient output (today we will focus only on the coefficients output):

```
------------------------------------------------------------------------------
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .5413593    .053248    10.17   0.000     .4367534    .6459651
       _cons |  -.9048516   .6849678    -1.32   0.187    -2.250472    .4407687
------------------------------------------------------------------------------
```

Here `wage` is the dependent variable, while `educ` is the independent variable. A constant term (`_cons`) is added automatically without the need to specify.

- The estimated beta coefficients are:

$$\hat{\beta}_0 = -.90$$

$$\hat{\beta}_1 = .54$$

- How would you interpret the estimate of the slope coefficient ($\hat{\beta}_1$)?

  β_1 measures the change in the hourly wage for an additional year of education, holding all other factors fixed. Note that the linearity of (1) implies that a one-unit change in education has the same effect on hourly wage, regardless of the initial value of the variable education.

  !! If you are asked to interpret one of the regression coefficients in a problem set or quiz always make sure to mention the following:

  - magnitude of the coefficient – put a number!
  - units both of the dependent variable and independent variable of interest
  - change in dependent variable is ceteris paribus – or holding all other factors fixed
  - generally no causal interpretation

- Make a scatter plot of wage and education, including a fitted line. To do so use the following command (Learnt in Lab 4):

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

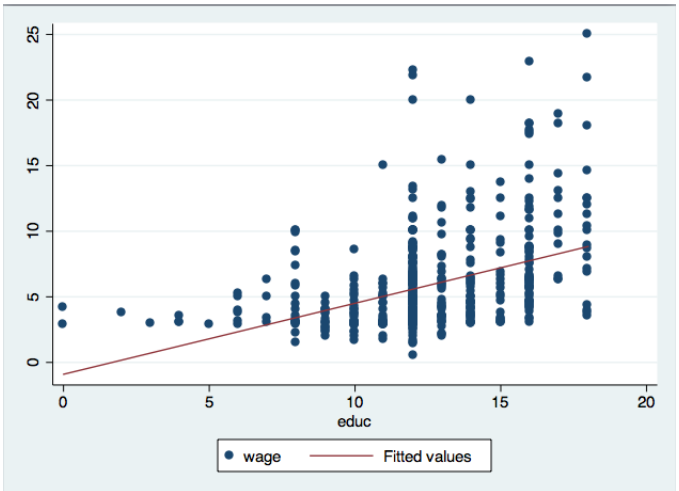

Figure 1: Wage and education relationship

The fitted line is the same as the regression line we estimated in the regression above: it has an intercept equal to $\hat{\beta}_0 = -.90$ and slope $\hat{\beta}_1 = .54$.

3 Fitted Values

- After regression estimation we can construct fitted values ($\widehat{wage}$). For each observation i in the dataset, the fitted values are constructed as:

$$\widehat{wage}_i = \hat{\beta}_0 + \hat{\beta}_1 * educ_i$$

- To obtain the "predicted values" in Stata we use the command `predict` *newvar* after a regression estimation. Type in:

```
predict wagehat
```

where *wagehat* is the name of the new variable that will be created by the `predict` command (you choose what you want to call the newly created variable). Open the Data Browser to see the new variable.

- Now make a scatter plot of *wagehat* and *educ*, including a fitted line for *wage* and *educ*. The command for this is:

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

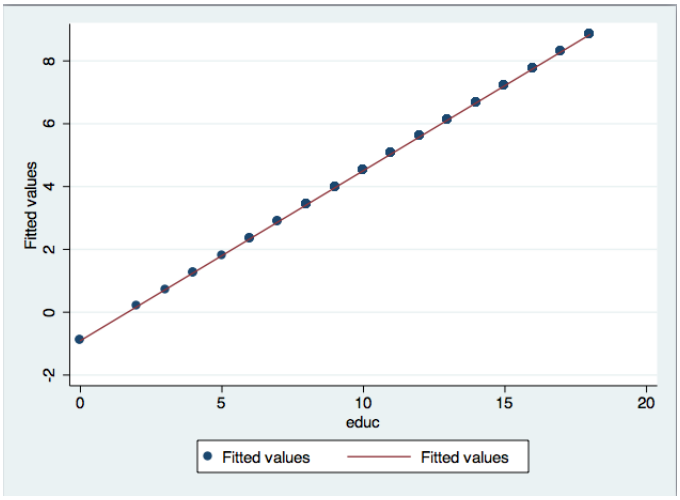

Figure 2: Relationship between predicted wage and education

By construction, each fitted value of $\widehat{wage}_i$ is on the OLS regression line!

4 Residuals

- The residual for observation i is the difference between the actual value of the dependent variable ($wage_i$) and its fitted value ($\widehat{wage}_i$):

$$\hat{u}_i = wage_i - \widehat{wage}_i$$

- In Stata to obtain residuals we can again use the `predict` command specifying the option `residual`. For example to create variable *resid* to contain the residuals from the above regression type:

```
predict resid, residuals
```

Note `resid` is only a variable name, and you can choose any other name for residuals.

- Look at the first 10 observations for *wage*, *wagehat*, and *resid*:

```
. list wage wagehat resid in 1/10
```

	wage	wagehat	resid
1.	3.1	5.0501	-1.9501
2.	3.24	5.591459	-2.35146
3.	3	5.0501	-2.0501
4.	6	3.426023	2.573977
5.	5.3	5.591459	-.2914593
6.	8.75	7.756896	.9931036
7.	11.25	8.839615	2.410385
8.	5	5.591459	-.5914595
9.	3.6	5.591459	-1.991459
10.	18.18	8.298256	9.881744

Above we can see that the residual takes on both positive and negative values. When the residual (*resid*) is positive, the fitted value is less than the actual value of *wage*, i.e. the fitted value underpredicts $wage_i$. If the residual is negative, then $\widehat{wage}_i > wage_i$. The ideal case for observation i is when $\hat{u}_i = 0$ but in most cases each residual is NOT equal to zero.

- Can you tell the sign of the residuals by looking at Figure 1?

- Let's look at a scatter plot of the residuals

  `scatter resid educ`

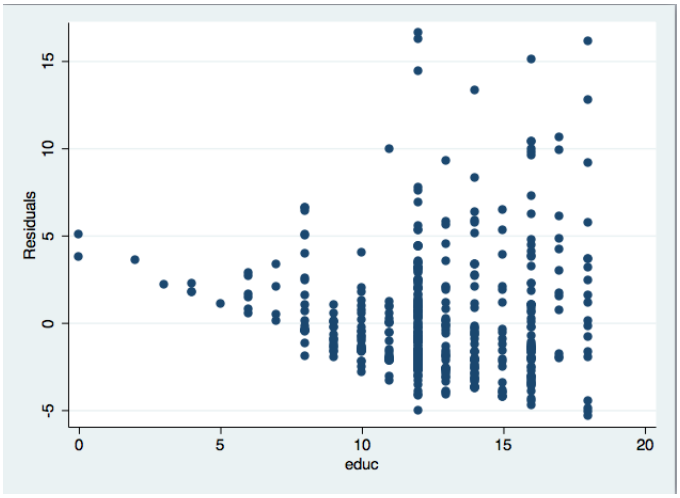

Figure 3: Scatter plot of residuals

You can combine your observations, the fitted values and the regression line in a single plot with the following command:
`twoway (scatter wage wagehat educ) (lfit wage educ)`

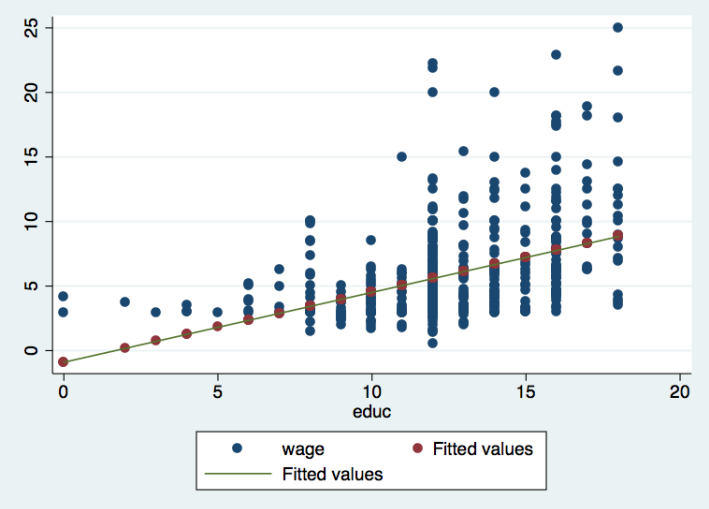

Figure 4: Scatter plot of residuals

Another way to obtain residuals is to use the predicted values for wage we obtained above:
`gen resid2 = wage - wagehat`

Summarize `resid` and `resid2`:

```
. sum resid resid2

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
       resid |       526    4.43e-09     3.37517  -5.339615   16.60854
      resid2 |       526    2.79e-08     3.37517  -5.339615   16.60854
```

- What do you think the mean of the residuals would be? What is it?
- Note the means of both variables are in fact the same.

5 Functional Forms

- So far we have focused on *linear* relationships between the dependent and independent variables. For many economics applications linear relationships are not enough and we may want to introduce non linearities in our econometric model.
- One simple way to do this is by appropriately redefining the dependent and independent variables.
- When we estimated the equation

$$wage = \beta_0 + \beta_1 educ + U \tag{7}$$

we implicitly assumed that an additional year of education has the same effect not the hourly wage for either the first year of education or the twentieth (recall the interpretation of the β_1). A more reasonable characterization may be that each year of education increases wage by a constant *percentage*. A model that gives (approximately) a **constant percentage effect** is

$$\ln(wage) = \beta_0 + \beta_1 educ + u \tag{8}$$

- How would you interpret β_1 in this case? $100\beta_1$ measures the percentage change in wage given one additional year of education.

```
. regress lwage educ

      Source |       SS       df       MS              Number of obs =     526
-------------+------------------------------           F(  1,   524) =  119.58
       Model |  27.5606296     1  27.5606296           Prob > F      =  0.0000
    Residual |  120.769132   524  .230475443           R-squared     =  0.1858
-------------+------------------------------           Adj R-squared =  0.1843
       Total |  148.329762   525   .28253288           Root MSE      =  .48008

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95\% Conf. Interval]
```

```
-------------+----------------------------------------------------------------
        educ |   .0827444   .0075667    10.94   0.000     .0678796    .0976092
       _cons |   .5837726   .0973358     6.00   0.000     .3925562     .774989
-------------------------------------------------------------------------------
```

In this case, one additional year of education increases wages by approximately 8.27%.

- Another important use in economics of the natural log is in obtaining the **constant elasticity model**. To obtain this, we take the logarithm of both the dependent and independent variable. For this example, let's use the dataset on CEO salaries and firm sales (which you can access with `bcuse ceosal1`). The constant elasticity model for CEO salaries and sales is:

$$\ln(salary) = \beta_0 + \ln(sales) + u \tag{9}$$

```
. reg lsalary lsales

      Source |       SS       df       MS              Number of obs =     209
-------------+------------------------------           F(  1,   207) =   55.30
       Model |  14.0661688     1  14.0661688           Prob > F      =  0.0000
    Residual |  52.6559944   207  .254376785           R-squared     =  0.2108
-------------+------------------------------           Adj R-squared =  0.2070
       Total |  66.7221632   208  .320779631           Root MSE      =  .50436

------------------------------------------------------------------------------
     lsalary |      Coef.   Std. Err.      t    P>|t|     [95\% Conf. Interval]
-------------+----------------------------------------------------------------
      lsales |   .2566717   .0345167     7.44   0.000     .1886224    .3247209
       _cons |   4.821997   .2883396    16.72   0.000     4.253538    5.390455
------------------------------------------------------------------------------
```

Now β_1 is the elasticity of salary with respect to sales. A 1% increase in sales increases CEO salary by 0.257%.