Development Data Boot Camp
Linear Regression: Understand OLS

Ge Sun

University of Notre Dame

May 19, 2023

Outline

Explore the relationship between Y and X

Linear regression models

Digressions: the GSS dataset

For five decades, the General Social Survey (GSS) has studied the growing complexity of American society. It is the only full-probability, personal-interview survey designed to monitor changes in both social characteristics and attitudes currently being conducted in the United States.

- GSS official website

Outline

Explore the relationship between Y and X

Linear regression models

How to explore the relationship between two variables

- Suppose there are two variables: Y representing wages, and X representing years of education.
- We are interested in "explaining Y in terms of X," or in "studying how Y varies with changes in X."
- Suppose we can observe a group of data, indexed by $1, 2, \ldots, n$:

$$(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$$

- How would we get the sample above in reality?
  - * RCT
  - * Sample survey: the Quarterly Labour Force Survey (QLFS) database (randomly drawing $6,550$ people from the working population)
  - * Administrative data

How to explore the relationship between two variables

- Y is often called dependent variable, outcome variable, explained variable, and predicted variable.
- X is often called independent variable, regressor, explanatory variable, and covariate.

Terminology for Simple Regression

y	x
Dependent variable	Independent variable
Explained variable	Explanatory variable
Response variable	Control variable
Predicted variable	Predictor variable
Regressand	Regressor

Figure 1: Terminology for simple regression

How to explore the relationship between two variables

- Draw a scatter figure of Y and X for 12 artificial observations:

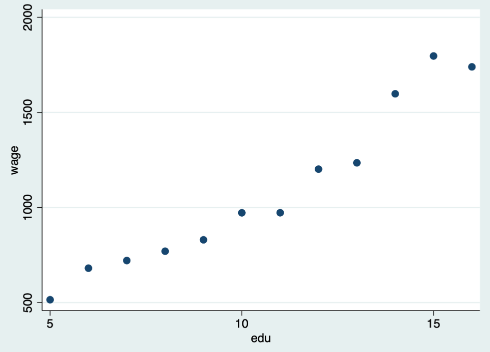
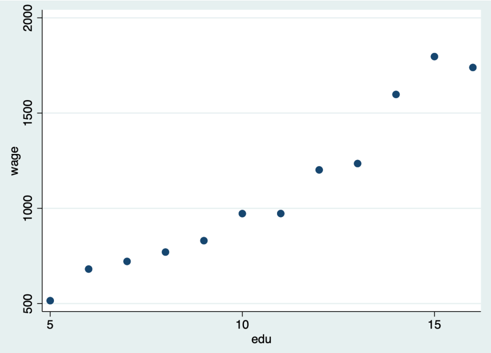


Figure 2: Relationship between wages and years of education

- Question of interest: Can I generate some general rules based on my observations?

How to explore the relationship between two variables



- A natural way to explore the relationship between wages and years of education, is to add a best fitted **line** on the figure such that all points are "closely" enough to the line.

Outline

Explore the relationship between Y and X

Linear regression models

OLS

- The mathematical way to write down a line:

$$y = \beta_0 + \beta_1 x$$

- In our wage and education example, we are trying to write a line as:

$$\hat{\text{wage}} = \beta_0 + \beta_1 edu$$

- Then, the original wage level is:

$$\begin{aligned} \text{wage} &= \hat{\text{wage}} + u \\ &= \beta_0 + \beta_1 edu + u \end{aligned}$$

- The definition of “closely” / “bested fitted”:

$$\min \sum_i (\text{wage}_i - \hat{\text{wage}}_i)^2 \equiv \min \sum_i (u_i)^2$$

- the method to get β_0 and β_1: **Ordinary Least Square (OLS)**

Implement OLS in Stata

- The command regression can be used to obtain the coefficients of the best fitted line. Type

reg wage edu

```
. reg wage edu

      Source |       SS           df       MS      Number of obs   =        12
-------------+----------------------------------   F(1, 10)        =    147.13
       Model |  1912605.78         1  1912605.78   Prob > F        =    0.0000
    Residual |  129997.996        10  12999.7996   R-squared       =    0.9364
-------------+----------------------------------   Adj R-squared   =    0.9300
       Total |  2042603.78        11  185691.252   Root MSE        =    114.02

------------------------------------------------------------------------------
        wage | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
         edu |   115.6498   9.534552    12.13   0.000     94.40545    136.8941
       _cons |  -128.0852   105.3845    -1.22   0.252    -362.8964    106.7261
------------------------------------------------------------------------------
```

Figure 3: Regression results

Implement OLS in Stata

- The best fitted line is:

$$wage = -128.0852 + 115.65 * edu$$

- and the corresponding plot in the figure is

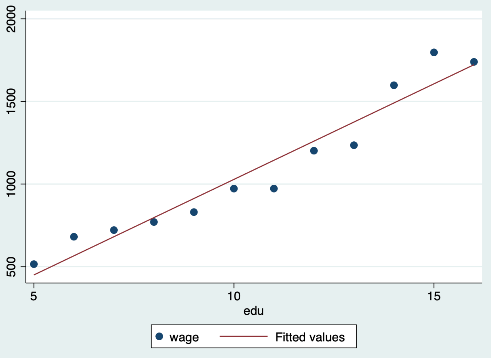


Figure 4: Fitted line

Understanding the regression results

```
. reg wage edu

      Source |       SS           df       MS      Number of obs   =        12
-------------+----------------------------------   F(1, 10)        =    147.13
       Model |  1912605.78         1  1912605.78   Prob > F        =    0.0000
    Residual |  129997.996        10  12999.7996   R-squared       =    0.9364
-------------+----------------------------------   Adj R-squared   =    0.9300
       Total |  2042603.78        11  185691.252   Root MSE        =    114.02

------------------------------------------------------------------------------
        wage | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
         edu |   115.6498   9.534552    12.13   0.000     94.40545    136.8941
       _cons |  -128.0852   105.3845    -1.22   0.252    -362.8964    106.7261
------------------------------------------------------------------------------
```

Std.err.

- It is short for “Standard error”
- Basically it tries to evaluate how accurate our estimation is.
- The smaller, the better.

The coefficients and standard errors are the most common things that authors will report in their paper.

Understanding Std.err.

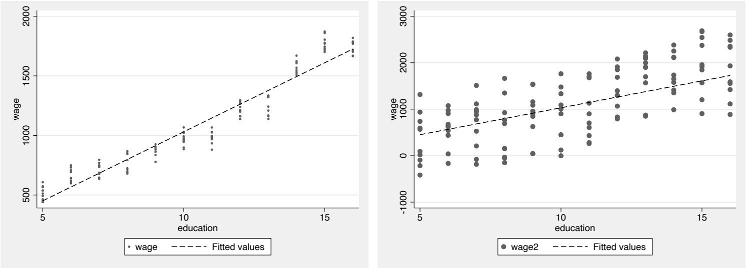


Figure 5: Sample with small and large variance

Understanding the regression results

. reg wage edu

Source	SS	df	MS			
Model	1912605.78	1	1912605.78	Number of obs	=	12
Residual	129997.996	10	12999.7996	F(1, 10)	=	147.13
				Prob > F	=	0.0000
				R-squared	=	0.9364
				Adj R-squared	=	0.9300
Total	2042603.78	11	185691.252	Root MSE	=	114.02

wage	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
edu	115.6498	9.534552	12.13	0.000	94.40545	136.8941
_cons	-128.0852	105.3845	-1.22	0.252	-362.8964	106.7261

t, P>|t|, 95% conf.interval

- t: t-statistics

$$t = \frac{\beta}{Std.err.}$$

- P>|t|: p-value, calculated by t-statistics
  - helps us to identify the “significance level”
  - the smaller, the better

Understanding the regression results

```
. reg wage edu

      Source |       SS           df       MS      Number of obs   =        12
-------------+----------------------------------   F(1, 10)        =    147.13
       Model |  1912605.78         1  1912605.78   Prob > F        =    0.0000
    Residual |  129997.996        10  12999.7996   R-squared       =    0.9364
-------------+----------------------------------   Adj R-squared   =    0.9300
       Total |  2042603.78        11  185691.252   Root MSE        =    114.02

------------------------------------------------------------------------------
        wage | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
         edu |   115.6498   9.534552    12.13   0.000     94.40545    136.8941
       _cons |  -128.0852   105.3845    -1.22   0.252    -362.8964    106.7261
------------------------------------------------------------------------------
```

t, P>|t|, 95% conf.interval

- 95% conf.interval:
  - We have 95% confidence that the true coefficient lies within this interval.

Understanding the regression results

```
. reg wage edu

      Source |       SS           df       MS      Number of obs   =        12
-------------+----------------------------------   F(1, 10)        =    147.13
       Model |  1912605.78         1  1912605.78   Prob > F        =    0.0000
    Residual |  129997.996        10  12999.7996   R-squared       =    0.9364
-------------+----------------------------------   Adj R-squared   =    0.9300
       Total |  2042603.78        11  185691.252   Root MSE        =    114.02

------------------------------------------------------------------------------
        wage | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
         edu |   115.6498   9.534552    12.13   0.000     94.40545    136.8941
       _cons |  -128.0852   105.3845    -1.22   0.252    -362.8964    106.7261
------------------------------------------------------------------------------
```

- t-statistics, $P > |t|$, and 95% confidence intervals: measure how confident we can say that explanatory variable does impact on explained variables.

$$|t| \uparrow \Rightarrow \text{more confident}$$

$$P > |t| \downarrow \Rightarrow \text{more confident}$$

$$0 \text{ is far from 95\% confident intervals} \Rightarrow \text{more confident}$$

Understanding the regression results

```
. reg wage edu

      Source |       SS           df       MS      Number of obs   =        12
-------------+----------------------------------   F(1, 10)        =    147.13
       Model |  1912605.78         1  1912605.78   Prob > F        =    0.0000
    Residual |  129997.996        10  12999.7996   R-squared       =    0.9364
-------------+----------------------------------   Adj R-squared   =    0.9300
       Total |  2042603.78        11  185691.252   Root MSE        =    114.02

------------------------------------------------------------------------------
        wage | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
         edu |   115.6498   9.534552    12.13   0.000     94.40545    136.8941
       _cons |  -128.0852   105.3845    -1.22   0.252    -362.8964    106.7261
------------------------------------------------------------------------------
```

The upper-right corner

- R-squared: the proportion of wage variation that can be explained by edu variation.
- Prob $>$ F: the overall significance level of this model. The smaller, the better
- Root MSE: root of MSE

$$MSE = \frac{1}{n}\sum_{i=1}^{N}(wage_i - \hat{wage}_i)^2$$

Understanding the regression results

```
. reg wage edu

      Source |       SS           df       MS      Number of obs   =        12
-------------+----------------------------------   F(1, 10)        =    147.13
       Model |  1912605.78         1  1912605.78   Prob > F        =    0.0000
    Residual |  129997.996        10  12999.7996   R-squared       =    0.9364
-------------+----------------------------------   Adj R-squared   =    0.9300
       Total |  2042603.78        11  185691.252   Root MSE        =    114.02

------------------------------------------------------------------------------
        wage | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
         edu |   115.6498   9.534552    12.13   0.000     94.40545    136.8941
       _cons |  -128.0852   105.3845    -1.22   0.252    -362.8964    106.7261
------------------------------------------------------------------------------
```

The upper-left table

- df: degree of freedom
  - Regression df: he number of independent variables in our regression model (just edu)
  - Residual df: total number of observations of the dataset subtracted by the number of variables being estimated (12-2)
- SS: sum of squares
- MS: mean squared errors
- R-squared $= Model_{SS}/Total_{SS}$

External resources: How to read a regression table

Multi-variate linear regression

- Education level (schooling years) is not the only determinant for people's future earnings.
- Mincer earnings function:

$$\log(wage) = \beta_0 + \beta_1 schooling + \beta_2 exp + \beta_3 exp_{sq}$$

- This seems more complex, but the simple *reg* command can help us get β_0, β_1 and β_2 all at once.

```
reg log_wage edu exp exp_sq
```

How to interpret β_1?

- Recall the linear regression models

$$Y = \beta_0 + \beta_1 X + \beta_2 Z + \beta_3 W + u$$

- Fix u, W, Z,

$$\begin{aligned} Y &= \beta_0 + \beta_1 X + \beta_2 Z + \beta_3 W + u \\ Y + \Delta Y &= \beta_0 + \beta_1 (X + \Delta X) + \beta_2 Z + \beta_3 W + u \\ \Rightarrow \Delta Y &= \beta_1 \Delta X \\ \Rightarrow \beta_1 &= \frac{\Delta Y}{\Delta X} \end{aligned}$$

- Thus, if we fix u and Z and W (that is, holding experience and other factors unchanged), β_1 measures the impact of one unit increment of X on Y.
- If Y is the (log)wage measured in dollars per month and X is years of education, then β_1 measures the change in monthly wage given another year of education, holding all other factors, including experience, fixed.

Implement multi-variate OLS in Stata

- Now, it is time to try by your self! Use our South African Labour Force data,

reg logwage edyears exp exp_sq