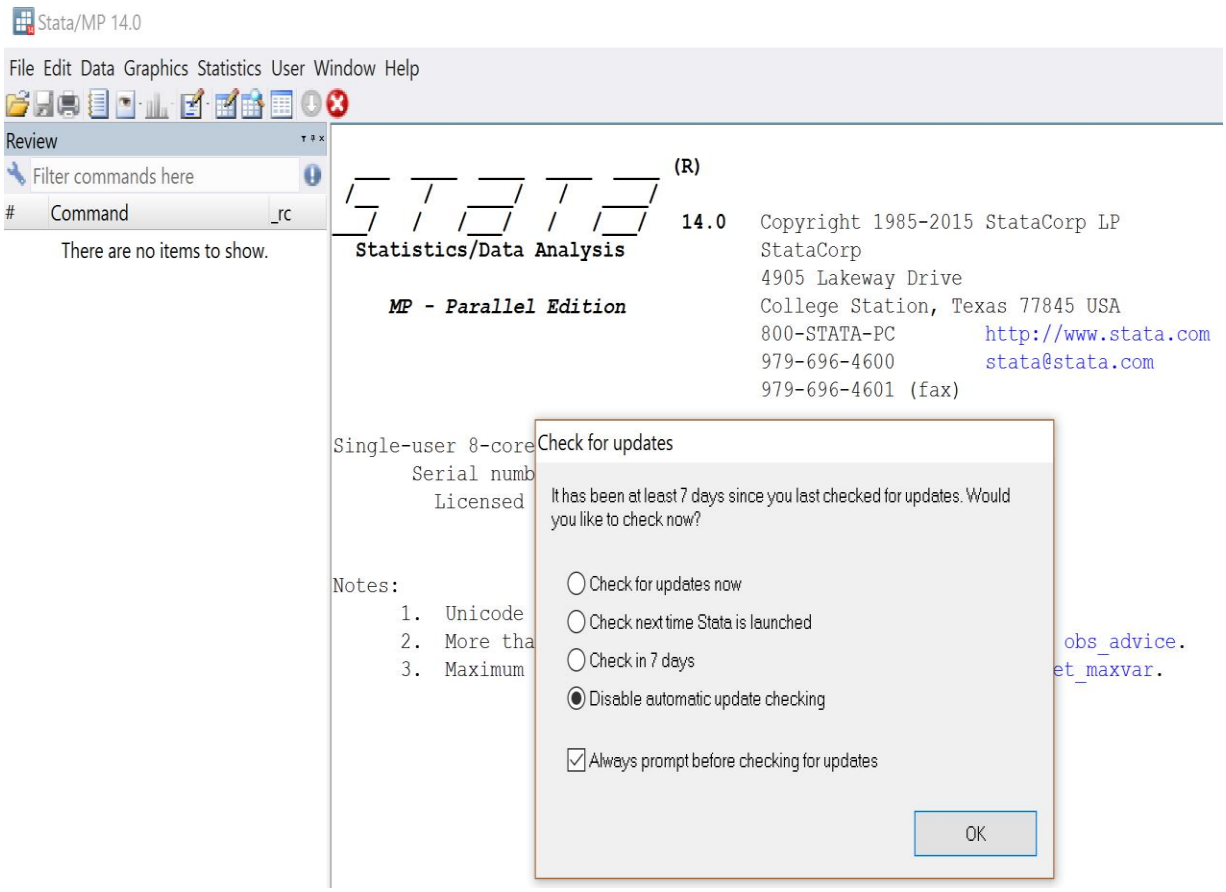
HƯỚNG DẪN SỬ DỤNG STATA 14

M&B – 5/2017

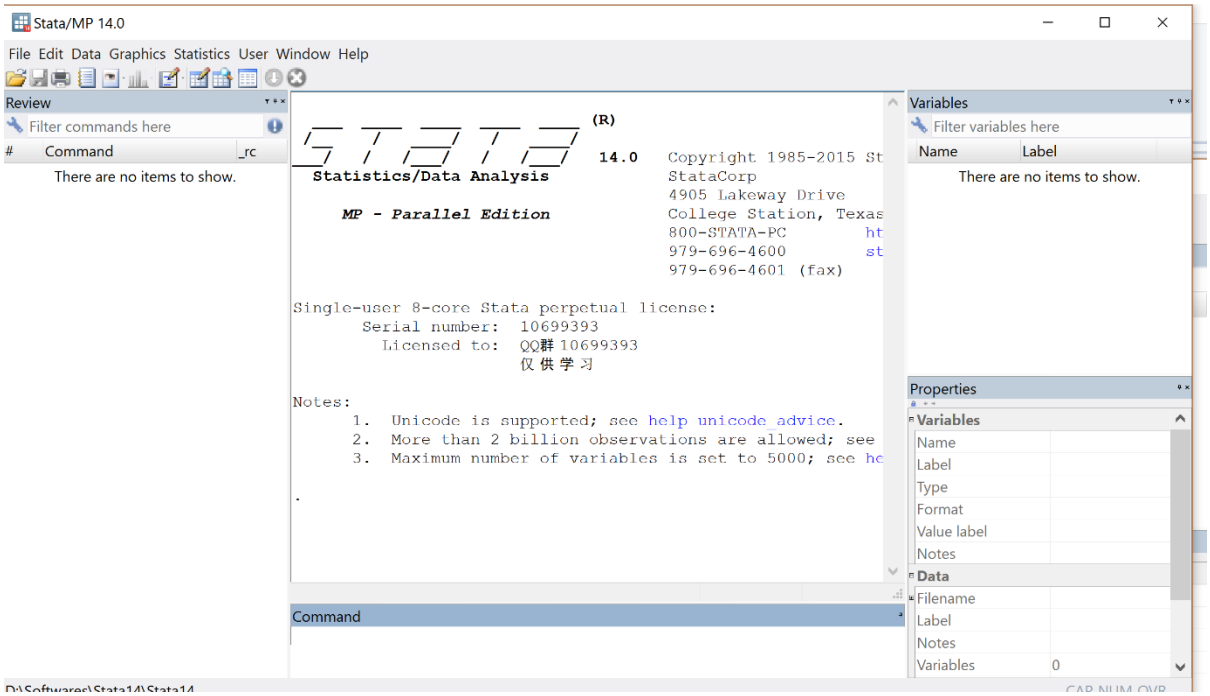
1. GIỚI THIỆU PHẦN MỀM STATA 14

- Mở Folder Stata14, copy và dán STATA.lic đè lên cái STATA.lic hiện có, và double click vào StataMP-64.exe, lần đầu tiên máy sẽ hỏi:



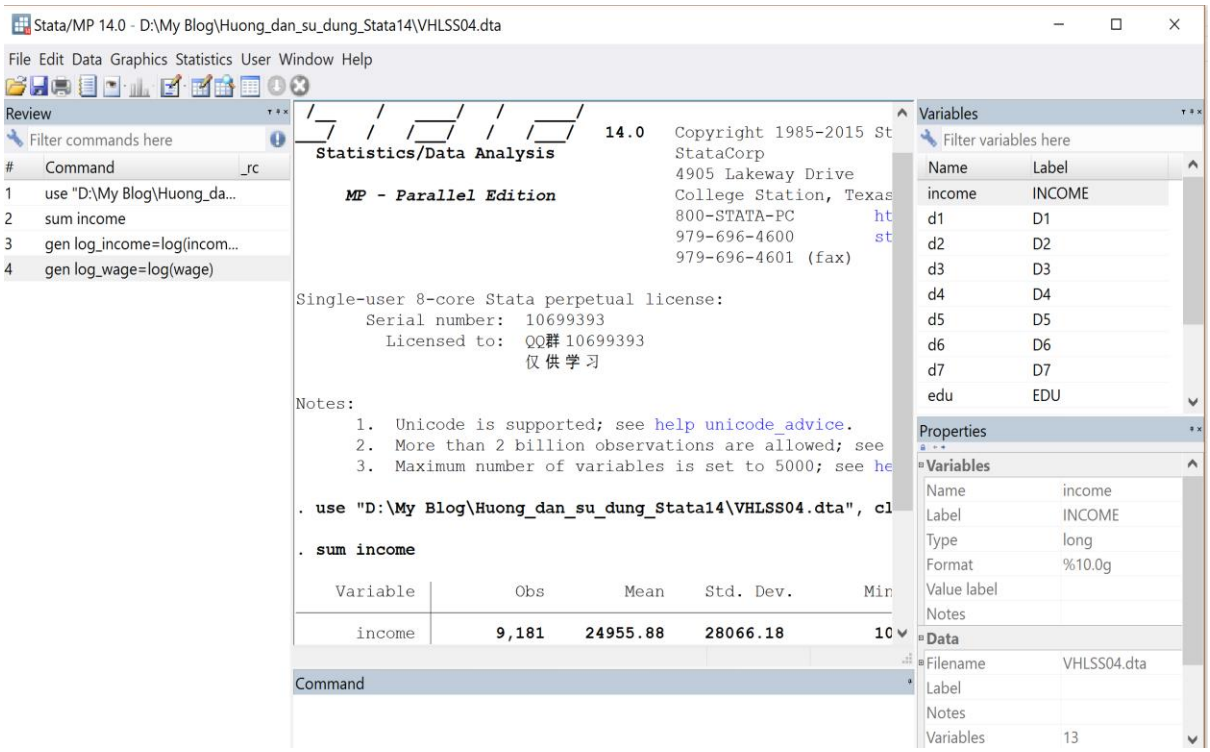
Đừng updates nhé, vì đây là phần mềm do mấy anh tàu của mình bẻ khóa, share với mấy thằng em của nó ở VN. Nên em chọn như trong hình trên nhé.

- Sau đó, màn hình chính của Stata sẽ như sau:



Command ở dưới cùng là màn hình lệnh, nghĩa là em gỏ từng lệnh riêng lẻ khi phân tích dữ liệu (ví dụ hồi quy OLS giữa hai biến y và x thì gỏ reg y x rồi enter). Sau đó, dòng lệnh vừa được gỏ vào sẽ được lưu lại ở chổ **Review Command** ở bên trái màn hình chính, và kết quả hồi quy sẽ xuất hiện ở màn hình giữa (từ từ mình nói cái này nha). Khi mở Stata file (tên file.dta) thì em sẽ thấy các biến chứa trong file xuất hiện ở **Variables** bên phải.

- Ví dụ cụ tỷ (nghĩa là cụ thể một cách tỷ mỉ): mở tập tin VHLSS04.dta, sẽ thấy các biến như income, d1, d2, …, d7, edu … ở **Variables**; và phần bên dưới mô tả thuộc tính của biến được chọn như Tên (Name), Nhãn (Label), … Phần mô tả này do người nhập và quản lý dữ liệu thực hiện.



Em thấy ở **Review** hiện các dòng lệnh đã được thực hiện: Dòng 1 là lệnh mở tập tin, Dòng 2 là lệnh về thống kê mô tả, Dòng 3 là tạo biến mới từ biến có sẵn. Các dòng này có thể được copy và lưu lại dưới dại do.file để sau này dùng lại. Sẽ nói cách tạo do.file sau em nhé. Và màn hình chính hiện ra các kết quả đã được thực hiện.

- Thanh công cụ hay dùng Dtrên màn hình Stata:

  - File: Để mở tập tin mới, lưu tập tin đang làm việc, đóng tập tin đang làm việc, in ấn, … Tuy nhiên, ta thường dùng các biểu tượng phía dưới nhanh hơn.

- Data: Để quản lý dữ liệu như nhập từ Excel qua, biên tập dữ liệu, tạo hoặc thay đổi biến, quản lý các biến, … Tuy nhiên, ta thường dùng ba hai biểu tượng phía dưới hơn: Biểu tượng cây viết và tờ giấy là tạo một DO File mới, Biểu tượng có cây viết và ô kẻ là để mở dạng bảng tính trên Excel để copy hoặc cắt dán, Biểu tượng có kính lúp và ô kẻ là để xem dữ liệu trong bảng tính.

- Graphics: Là vẻ đồ thị hoặc dùng biểu tượng đồ thị phía dưới cũng được. Mới làm quen Stata thì cần vào Graphics để vẻ các loại đồi thị, nhưng khi quen rồi thì gỏ trực tiếp ở màn hình lệnh (Command). Cuốn A Visual Guide to Stata Graphics hướng dẫn rất cụ tỉ về từng loại đồ thị, em không cần nhớ lệnh, khi nào cần vẻ loại đồ thị nào thì cứ tham khảo, làm theo.

- Statistics: Tất cả công việc liên quan đến thống kê mô tả và hồi quy đều nằm ở đây. Tuy nhiên, khi quen rồi thì cứ gỏ lệnh ở màn hình lệnh thôi.

- User và Help: Rất hữu ích, vì cung cấp cho người sử dụng tất cả các lệnh mình muốn biết. Nguyên bộ hướng dẫn đều được lập trình vào đấy. Dù bằng tiếng Anh, nhưng các câu lệnh thì mình không cần quá giỏi tiếng Anh cũng hiểu được. Đừng lo.

2. CÁC LỆNH CƠ BẢN CỦA STATA 14

- **describe** hoặc **des** để xem đặc điểm các biến trong tập tin. Ví dụ, với tập tin VHLSS04.dta, ở màn hình lệnh, em gỏ des rồi enter, thấy như sau:

```
. des

Contains data from D:\My Blog\Huong_dan_su_dung_Stata14\VHLSS04.dta
  obs:         9,189
 vars:            13                          17 May 2017 22:21
 size:       229,725
-------------------------------------------------------------------------------
              storage   display    value
variable name   type    format     label      variable label
-------------------------------------------------------------------------------
income          long    %10.0g                INCOME
d1              byte    %8.0g                 D1
d2              byte    %8.0g                 D2
d3              byte    %8.0g                 D3
d4              byte    %8.0g                 D4
d5              byte    %8.0g                 D5
d6              byte    %8.0g                 D6
d7              byte    %8.0g                 D7
edu             byte    %10.0g                EDU
wage            long    %10.0g                WAGE
region          byte    %8.0g                 REGION
log_income      float   %9.0g
log_wage        float   %9.0g
-------------------------------------------------------------------------------
Sorted by:
     Note: Dataset has changed since last saved.
```

- **rename** hoặc **ren** để đổi tên biến, ví dụ em muốn đổi tên biến income thành INCOME (lưu ý: Stata xem chữ thường và chữ HOA là khác nhau, không giống như Eviews đâu), thì em chỉ cần gỏ trên màn hình lệnh là:

```
ren income INCOME
```

  Thì em thấy trong VARIABLES không còn biến income nữa, mà giờ sẽ thành INCOME.

- **label** hoặc **lab** là để đặt tên nhãn của biến. Ví dụ, em muốn cho người khác hiểu dữ liệu của mình rõ hơn (ví dụ **income** là gì, đơn vị tính ra sao, …, thì ở màn hình lệnh em gỏ như sau:

```
label var income "Tổng thu nhập của hộ, triệu đồng"
```

Lưu ý: Nhớ là dấu " chứ không phải dấu ' nha. Đặc biệt, chỉ từ Stata 14 mới cho phép gỏ tiếng Việt UniCode, chứ các phiên bản trước chỉ gỏ không dấu hoặc tiếng Anh thôi.

- **generate** hoặc **gen** dùng để tạo biến mới từ biến có sẵn. Ví dụ, em muốn tạo biến log(income) từ biến income, thì trên màn hình lệnh em gỏ:

```
gen log_income=log(income)
```

Lưu ý: Chuyển hóa dạng biến sang logarithm là việc rất thường xuyên trong phân tích dữ liệu và hồi quy. Với Stata thì buộc phải tạo thêm biến mới, chứ không có lệnh trực tiếp như với Eviews.

Trong nhiều trường hợp phải kết hợp với lệnh `if` để gán thêm điều kiện ràng buộc. Ví dụ, `gen log_income=log(income) if income>0`.

- **drop** dùng để bỏ một hoặc một số biến, hay một hoặc một số quan sát không cần thiết để dễ dàng cho việc quản lý dữ liệu, nhất là bộ dữ liệu quá lớn. Lưu ý: Cách này nên dùng cẩn thận (Cho nên người phân tích dữ liệu với Stata thường dùng DO file hơn, vì không đụng chạm đến dữ liệu gốc, nếu như lỡ tay lưu chồng lên tập tin hiện hành sau khi sử dụng lệnh drop hoặc lệnh **keep**). Ví dụ, sau khi tạo ra biết log_income, như em thấy biến log_income không tốt bằng biến income ban đầu, để cho bộ dữ liệu gọn nhẹ, em quyết định bỏ biến log_income. Ở màn hình lệnh, em gỏ:

```
drop log_income
```

- **keep** là lệnh rất hay dùng trong việc quản lý dữ liệu, nhất là bộ dữ liệu quá lớn, mỗi lần xem khó khăn. Với lệnh này, em chỉ cần giữ lại những biến em cần. Ví dụ,

em chỉ muốn giữ các biến income, edu, urban, thì ở màn hình lệnh em gỏ:

```
keep income edu urban
```

Rủi ro cũng giống như lệnh **drop**, nên phải hết sức cẩn thận khi dùng lệnh này nhé.

- **summarize** hay **sum** dừng để lập bảng thống kê mô tả (bao gồm tên biến, số quan sát, giá trị trung bình mẫu, độ lệch chuẩn của mẫu, giá trị nhỏ nhất, giá trị lớn nhất). Ví dụ, em muốn xem thống kê mô tả của 3 biến income, edu, và wage, thì trên màn hình lệnh em gỏ:

```
sum income edu wage
```

Kết quả như sau:

```
sum income wage edu
```

Variable	Obs	Mean	Std. Dev.	Min	Max
income	9,181	24955.88	28066.18	10	1380200
wage	9,188	7649.812	12813.04	0	169300
edu	9,188	2.183827	.8290547	1	3

Ngoài ra, nếu muốn biết chi tiết hơn về phân phối của mỗi biến, em gỏ thêm detail vào sau dấu phẩy:

```
sum income edu wage, detail
```

Kết quả như sau:

```
. sum income wage edu, detail
```

tổng thu nhập của hộ, triệu đồng

Percentiles		Smallest		
1%	1971	10		
5%	4926	20		
10%	7057	20	Obs	9,181
25%	11525	20	Sum of Wgt.	9,181
50%	18584		Mean	24955.88
		Largest	Std. Dev.	28066.18
75%	30000	412790		
90%	47200	469400	Variance	7.88e+08
95%	64600	471274	Skewness	15.60589
99%	114914	1380200	Kurtosis	624.666

Nếu trường hợp có nhiều số lẻ quá làm cho bảng mô tả không được đẹp, thì em phải dùng lệnh **format** trước, sau đó mới dùng lệnh sum:

```
Format income %9.2f
Format wage %9.2f
```

Lưu ý: 9 là hàng trăm triệu, và 2 là số con số thập phân. Cho nên nếu biết giá trị lớn nhất của income trong dữ liệu chỉ có hàng triệu thôi, thì em dùng %7.2f.

```
format income %9.2f
format wage %9.2f
sum income wage
```

Variable	Obs	Mean	Std. Dev.	Min	Max
income	9,181	24955.88	28066.18	10	1380200
wage	9,188	7649.812	12813.04	0	169300

Bây giờ giá trị mean trong bảng có vẻ 'vuông vức' hơn so với bảng ở trên.

- **tabulate** hay **tab** thường dùng để lập bảng tần suất cho loại biến rời rạc. Nếu chỉ một biết gọi là tab oneway (lập bảng 1 chiều), và hai biến là tab twoway (lập bảng 2 chiều). Ví dụ chỉ một biến edu.

```
tab edu
```

```
. tab edu
```

giáo dục chủ hộ	Freq.	Percent	Cum.
1	2,468	26.86	26.86
2	2,563	27.90	54.76
3	4,157	45.24	100.00
Total	9,188	100.00	

```
tab edu, plot
```

```
. tab edu, plot
```

giáo dục chủ hộ	Freq.	
1	2,468	**********
2	2,563	**********
3	4,157	**********
Total	9,188	

```
tab edu urban
```

```
. tab edu urban
```

giáo dục chủ hộ	URBAN 0	1	Total
1	1,982	486	2,468
2	1,982	581	2,563
3	2,974	1,183	4,157
Total	6,938	2,250	9,188

tab edu urban, row

```
. tab edu urban, row
```

Key
frequency *row percentage*

giáo dục chủ hộ	URBAN 0	1	Total
1	**1,982** **80.31**	**486** **19.69**	**2,468** **100.00**
2	**1,982** **77.33**	**581** **22.67**	**2,563** **100.00**
3	**2,974** **71.54**	**1,183** **28.46**	**4,157** **100.00**
Total	**6,938** **75.51**	**2,250** **24.49**	**9,188** **100.00**

Tab edu urban, col

```
. tab edu urban, col
```

Key
frequency *column percentage*

giáo dục chủ hộ	URBAN 0	1	Total
1	**1,982** **28.57**	**486** **21.60**	**2,468** **26.86**
2	**1,982** **28.57**	**581** **25.82**	**2,563** **27.90**
3	**2,974** **42.87**	**1,183** **52.58**	**4,157** **45.24**
Total	**6,938** **100.00**	**2,250** **100.00**	**9,188** **100.00**

```
tab edu urban, row col
```

giáo dục chủ hộ	URBAN 0	1	Total
1	1,982 80.31 28.57	486 19.69 21.60	2,468 100.00 26.86
2	1,982 77.33 28.57	581 22.67 25.82	2,563 100.00 27.90
3	2,974 71.54 42.87	1,183 28.46 52.58	4,157 100.00 45.24
Total	6,938 75.51 100.00	2,250 24.49 100.00	9,188 100.00 100.00

- **correlate** hay **cor** dùng để lập bảng ma trận hệ số tương quan (khái niệm hệ số tương quan em có thể đọc ở phần thống kê nhé).

Cor income wage

```
. cor income familysize edu urban
(obs=9,181)
```

	income	family~e	edu	urban
income	1.0000			
familysize	0.1752	1.0000		
edu	0.0415	-0.0407	1.0000	
urban	0.2773	-0.0595	0.0864	1.0000

Lưu ý: Trong bảng này em thấy cái dở của việc đặt tên biến familysize quá dài chưa. Khoảng 8 ký tự là tối đa.

- **regress** hay **reg** dùng để chạy mô hình hồi quy theo phương pháp OLS (các phương pháp khác sẽ trình bày sau, vì chương trình căn bản em chỉ giới hạn với OLS thôi. Ngoài ra, các loại kiểm định hậu hồi quy đã hướng dẫn lồng trong các tài liệu đã upload trước đây). Khi nào học tới đó, mình thảo luận tiếp nhé. Ví dụ, chạy mô hình hồi quy (dĩ nhiên đây không phải là một mô hình có ý nghĩa về mặt lý

thuyết, nhưng chỉ mang tính minh họa cho lệnh trên Stata thôi. Ngoài ra, mình cũng chưa cần bàn đến các con số trong bảng kết quả hồi quy OLS):

$$income_i = b_0 + b_1edu_i + b_2urban_i + b_3familysize_i + e_i$$

Ở màn hình lệnh, em gỏ như sau:

```
reg income edu urban familysize
```

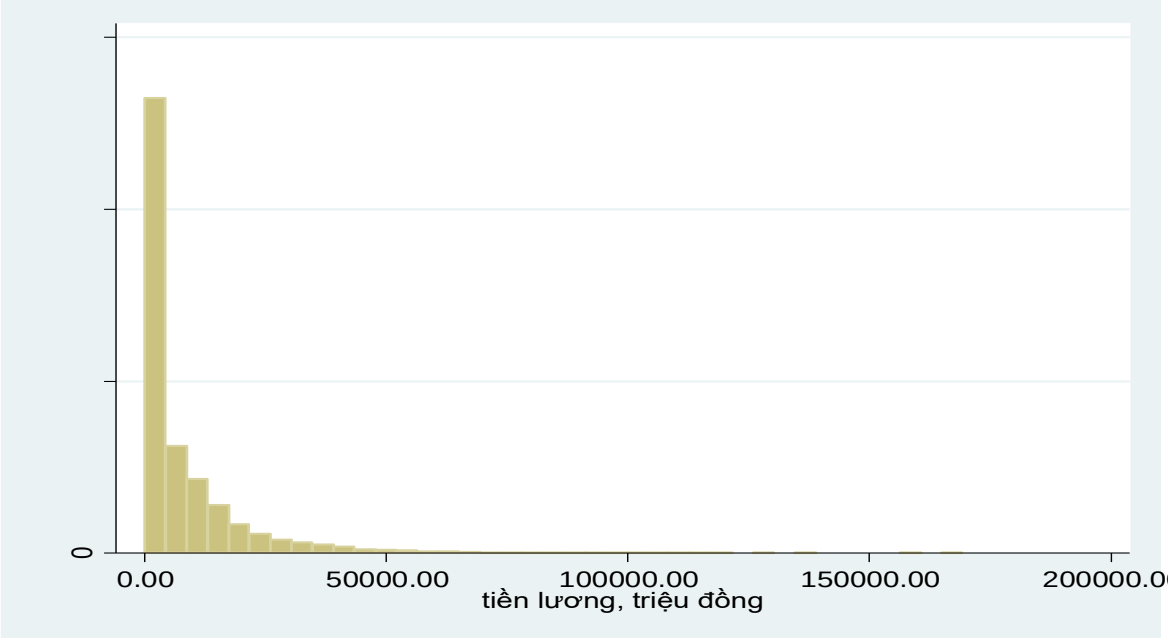
```
. reg income edu urban familysize
```

Source	SS	df	MS	Number of obs	=	9,181
				F(3, 9177)	=	395.11
Model	8.2717e+11	3	2.7572e+11	Prob > F	=	0.0000
Residual	6.4040e+12	9,177	697832924	R-squared	=	0.1144
				Adj R-squared	=	0.1141
Total	7.2312e+12	9,180	787710271	Root MSE	=	26417

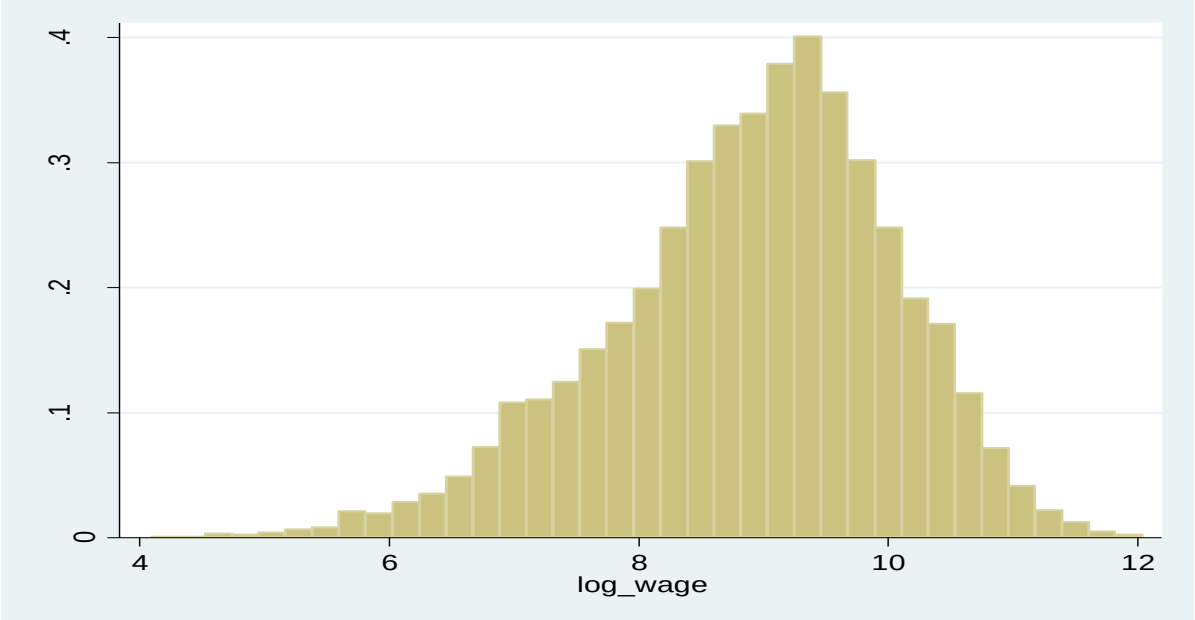
income	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
edu	831.4589	334.0071	2.49	0.013	176.7306	1486.187
urban	18715.86	644.6752	29.03	0.000	17452.15	19979.57
familysize	3132.289	159.6162	19.62	0.000	2819.405	3445.172
_cons	4774.736	1075.281	4.44	0.000	2666.945	6882.527

- **histogram** hay **hist** dùng để vẻ đồ thị phân phối tần suất của một biến, ví dụ wage và log_wage:

```
hist wage
```




```
hist log_wage
```



- **scatter** dùng để vẻ đồ thị phân tán giữa hai biến.

```
scatter income wage hoặc twoway (scatter income wage)
```