

Phương pháp Nghiên cứu Kinh tế

cuu duong than cong. com

TS. Kiều Thanh Nga
Viện Hàn lâm Khoa học Xã hội Việt Nam
Email: kieuthanhnga@iames.gov.vn
Tel: 0986654176

Chương 7: Nhập và xử lý số liệu trên một số phần mềm cơ bản

7.1. Nhập và xử lý số liệu trên phần mềm Stata

- Những vấn đề cơ bản về phần mềm Stata
- Phân tích dữ liệu bằng Stata

cuduongthancong.com

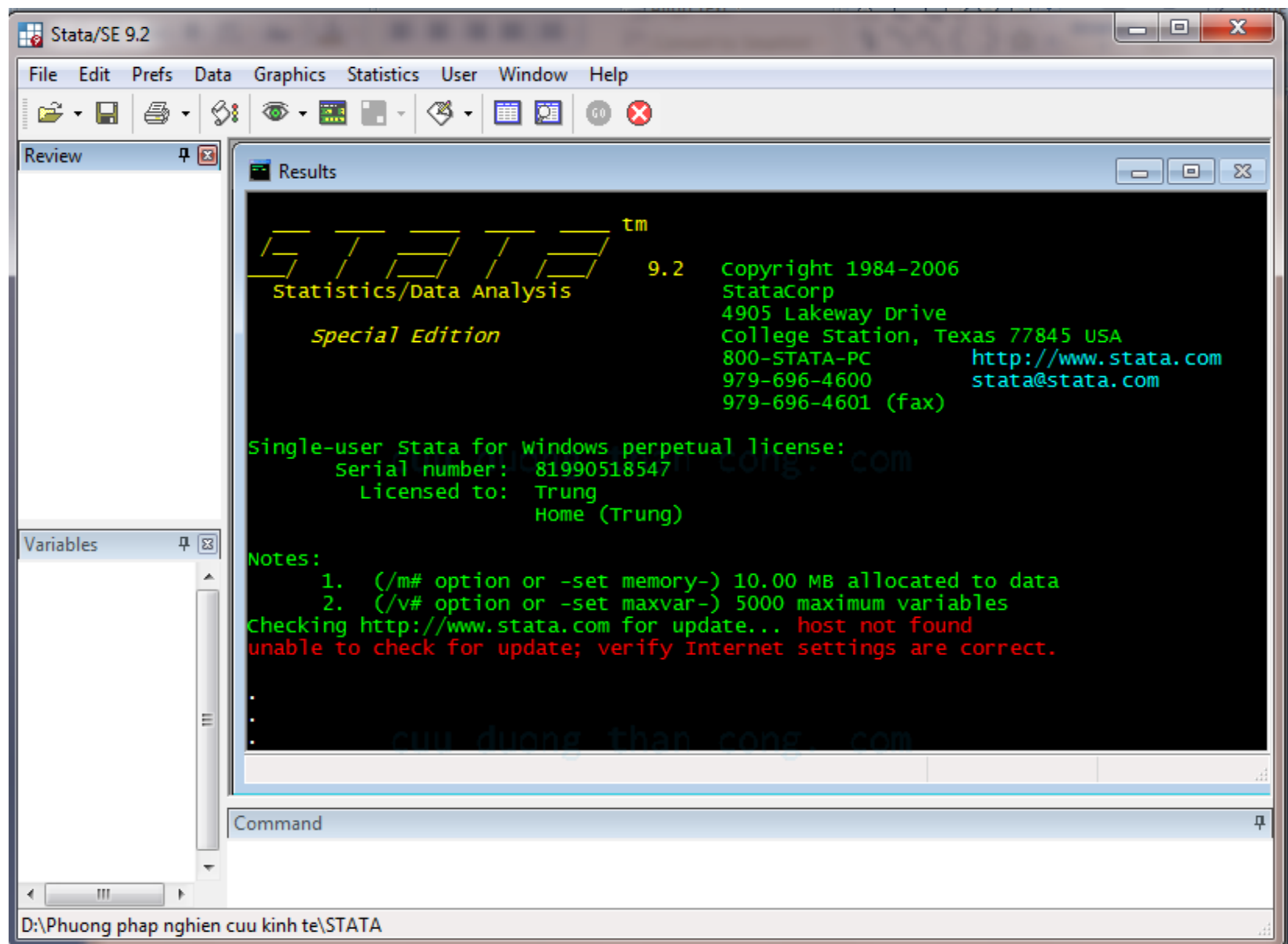
7.2. So sánh tính năng của phần mềm Stata với một số loại phần mềm khác

- Phần mềm SPSS
- Ưu/nhược điểm của các phần mềm
- Cách khắc phục

cuduongthancong.com

Giới thiệu về Stata

- ❑ Stata là phần mềm thống kê để quản lý, phân tích và vẽ đồ thị của số liệu. Sức mạnh lớn nhất của Stata là hồi quy. Ưu điểm: dùng để phân tích dữ liệu theo mẫu, có khả năng áp dụng chúng trong phân tích số liệu điều tra bởi các công cụ hồi quy. Nhược điểm: Khả năng phân tích phương sai và phân tích nhiều chiều kém.
- ❑ Có 4 loại cửa sổ trên Stata: Command, Review, Variables và Results
 - Cửa sổ Command cho phép đánh các lệnh
 - Cửa sổ Review liệt kê các lệnh sử dụng gần đây
 - Cửa sổ Variables liệt kê các biến (variables) trong file dữ liệu
 - Cửa sổ Results là màn hình chính hiển thị các kết quả thực hiện lệnh



Giới thiệu về Stata

- ❑ Ngoài ra, Stata còn có một số cửa sổ khác sẽ hiện lên khi ta chọn chúng trong Menu Windows, thanh công cụ hoặc thực hiện các lệnh liên quan đến các cửa sổ này.
- Cửa sổ Graph: hiển thị các đồ thị
- Cửa sổ Viewer: hiển thị trợ giúp hoặc xem nội dung các file văn bản
- Cửa sổ Data Editor: cho phép hiệu đính file dữ liệu dưới dạng bảng như Excel.
- Cửa sổ Do-file Editor: soạn thảo các file chương trình
- Cửa sổ Log: Để ghi nhật ký 1 buổi làm việc
- Cửa sổ Data Browse: Để xem tập dữ liệu đang hoạt động

Các Menu trên Stata

❖ File:

Open: Mở file số liệu Stata

View: Xem các file của Stata trong cửa sổ Viewer

Save: Lưu file số liệu với tên đang có

Save as: Lưu file số liệu với tên mới

File Name: Chọn tên file để đưa vào cửa sổ lệnh

Log: đóng, mở hoặc xem file Log

Save Graph: Lưu đồ thị

Print Graph: in đồ thị

Print Results: in kết quả

Exit: Ra khỏi Stata

Các Menu trên Stata

❖ Edit:

Copy text: copy văn bản đã đánh dấu

Copy Table: copy bảng biểu đã đánh dấu

Paste: Dán thông tin đã copy vào chỗ yêu cầu

Table Copy options: tùy chọn copy bảng số liệu

Graph copy options: tùy chọn copy trong đồ thị

❖ Prefs:

Tùy chọn về màu sắc, font chữ, kích cỡ chữ

Các Menu trên Stata

❖ Data:

Describe data: Cho biết thông tin về biến, 1 số thống kê trên biến

Data editor: mở cửa sổ hiệu đính dữ liệu

Data browser: mở cửa sổ xem dữ liệu

Creat or change: tạo biến mới hoặc thay đổi nội dung biến

Sort: sắp xếp, phân tổ dữ liệu

Combine Datasets: Kết nối các file dữ liệu

Label & Notes: Dán nhãn cho biến, cho trị số hoặc ghi lời chú cho tập dữ liệu

Variable Utilities: Đổi tên biến, so sánh hai biến

Matrices: Một số lệnh trên về ma trận

Other Utilities: Một số lệnh khác về biến và ma trận

Các Menu trên Stata

❖ Graphs

Easy graph: Vẽ các đồ thị đơn giản: Scatter Plot, Line Graph, Bar Chat, Pie Chat...

Twoway Graphs: Vẽ các đồ thị hai chiều

Overlay Graphs: Vẽ nhiều đồ thị trên một khung

Bar chat: Đồ thị cột

Pie chat: đồ thị bánh xe

Histogram: đồ thị tần số

Box plots: đồ thị hộp

Scatter matrix: ma trận các đồ thị phân tán

Các Menu trên Stata

❖ Statistics:

Summaries, tables & tests: lập bảng và kiểm định

Linear regression and related: hồi quy tuyến tính và các lệnh liên quan

Binary Outcomes: Hồi quy logistic

Ordinal Outcomes: Hồi quy logistic thứ tự

Categorical outcomes: Hồi quy logistic bội

Selection models: Mô hình Heckman

Generalized linear models: Mô hình tuyến tính tổng quát

Nonparametric Analysis: phân tích phi tham số

Time series: Phân tích chuỗi thời gian

Multivariate time series: Phân tích chuỗi thời gian chéo

Survival analysis: phân tích nguy cơ

Other multivariate analysis: phân tích nhiều chiều khác

.....

Cấu trúc lệnh, các phép toán và hàm số

✓ Cấu trúc lệnh:

[by varlist:] command [varlist] [if exp] [in range] [weight] [,options]

Trong đó

By varlist: thực hiện lặp lại câu lệnh đối với từng giá trị của danh sách biến. Các biến phải được sắp xếp trước đó

Command: tên câu lệnh

Varlist: danh sách biến mà câu lệnh command sẽ thực hiện trên đó

If exp: exp là biểu thức logic, những quan sát trong file số liệu thỏa mãn biểu thức sẽ được đưa vào xử lý

In range: range chỉ ra giới hạn một tập liên tiếp các quan sát sẽ được đưa vào xử lý

Weight: quyền số trong điều tra mẫu.

Options: các tùy chọn khác

Ví dụ: .list in 20/1: đọc dữ liệu các biến từ quan sát thứ 20 đến cuối tập dữ liệu

Regress Yi Xi: Hồi quy tuyến tính biến Yi Xi

Cấu trúc lệnh, các phép toán và hàm số

✓ Các phép toán:

+ Cộng - trừ * nhân / chia ^ lũy thừa

> Lớn hơn < nhỏ hơn >= lớn hơn hoặc bằng <= nhỏ hơn hoặc bằng

== bằng != không bằng

✓ Hàm số

Hàm toán học

Hàm thống kê

Hàm ngẫu nhiên

Hàm ký tự

Hàm đặc biệt

Hàm ngày tháng

Hàm chuỗi thời gian

Hàm ma trận

Phân tích dữ liệu trên Stata

Nhập liệu từ Stata: Có ba cách chính

- Vào Menu Data sau đó chọn Data Editor (hoặc dùng lệnh Edit trên cửa sổ Command) rồi nhập liệu trực tiếp
- Nhập liệu trên Excel sau đó lưu file dưới dạng csv (comma delimited). Sau đó từ Stata vào File => Import => ASCII data created by a spreadsheet rồi chọn file. Chú ý là phải chọn file type là All để hiển thị file cần chọn.
- Nhập liệu trên Excel. Mở đồng thời Excel và Stata. Sau khi nhập liệu xong chọn bảng cần sử dụng. Vào Stata, chọn Menu Data sau đó chọn Data Editor (hoặc dùng lệnh Edit trên cửa sổ Command) rồi nhấn chuột phải để Paste (hay Ctrl + V).

Phân tích dữ liệu trên Stata

- Sau khi nhập liệu, có thể save file với lệnh save hoặc vào File rồi chọn Save as. File sẽ được xếp với đuôi là .dta.
- Mở file .dta bằng cách chọn File rồi Open.
- Mục Help của Stata rất tiện dụng để tra cứu các câu lệnh cần thiết.

cuu duong than cong. com

Bảng phân tích

- Giả sử chúng ta muốn biết sở hữu xe máy theo hộ theo tổng số hộ. (file Eg1)
- Lập bảng phân tích

cuu duong than cong. com

cuu duong than cong. com

Kiểm định giá trị trung bình:

Cú pháp: Test varname ==[in range]

Ví dụ: Kiểm định giá trị trung bình số hộ có trung bình 1,6 xe máy

Ta lập bảng như sau:

Bảng phân tích

X= Số xe máy sở hữu	h= Tần số tuyệt đối (số hộ sở hữu xe máy)	f=h/n (quan hệ tần suất)	Tỷ lệ (%)
0	3	0,03	3
1	45	0,45	45
2	37	0,37	37
3	11	0,11	11
4	4	0,04	4
Tổng (n)	100	1,00	100

Kiểm định giả thuyết thống kê

- Bài tập:

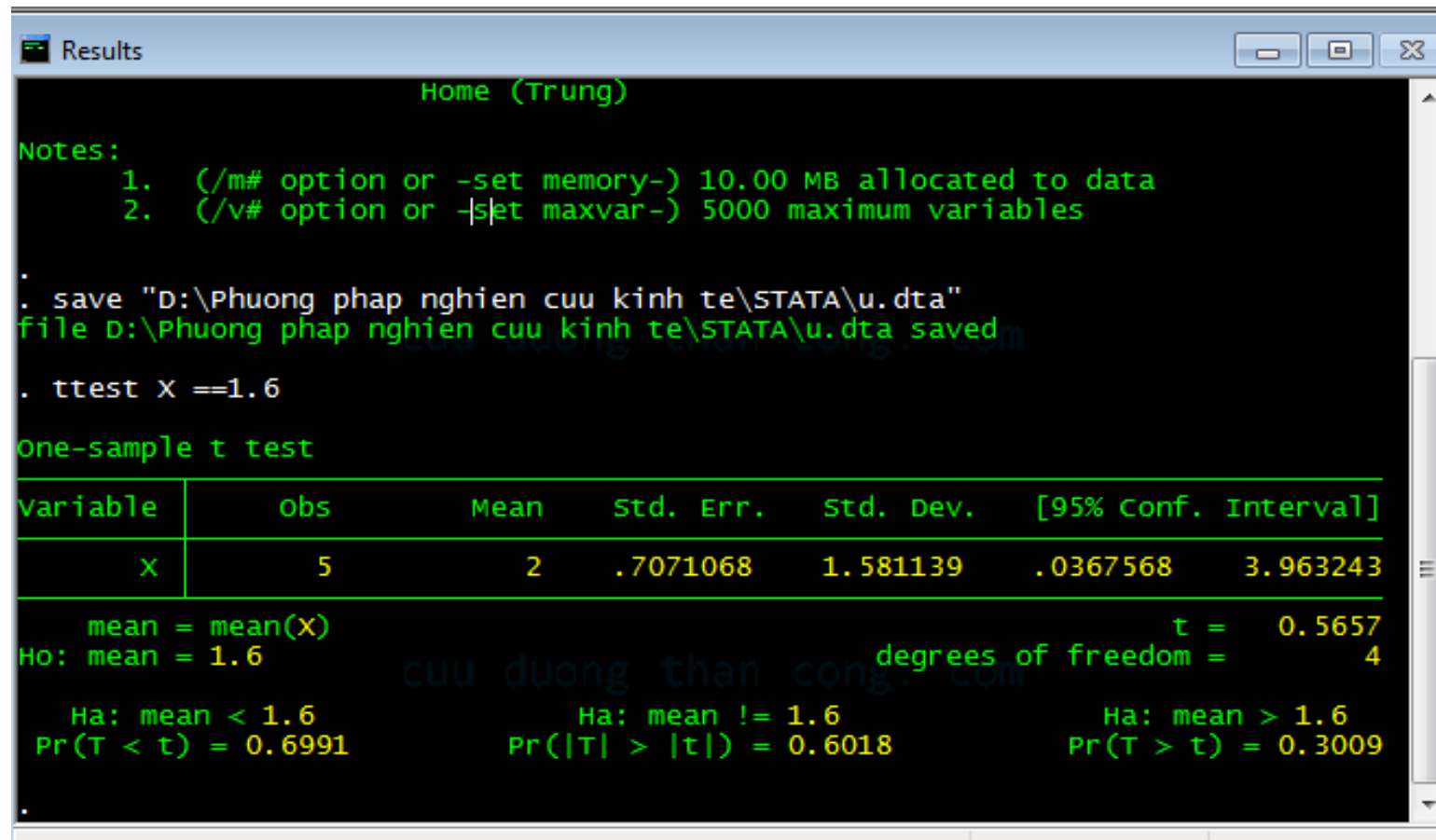
Kiểm định giả thuyết là một hộ gia đình có trung bình 1,6 xe máy, 1,5 xe máy, 1,7 xe máy.

Câu lệnh Stata: `ttest X==1.6`

cuu duong than cong. com

cuu duong than cong. com

Kết quả như sau:



```
Results
Home (Trung)

Notes:
1. (/m# option or -set memory-) 10.00 MB allocated to data
2. (/v# option or -set maxvar-) 5000 maximum variables

. save "D:\Phuong phap nghien cuu kinh te\STATA\u.dta"
file D:\Phuong phap nghien cuu kinh te\STATA\u.dta saved

. ttest x ==1.6

One-sample t test

Variable | Obs   Mean   Std. Err.   Std. Dev.   [95% Conf. Interval]
-----+-----
x        |    5     2    .7071068    1.581139    .0367568    3.963243

      mean = mean(x)                                t = 0.5657
Ho: mean = 1.6                                     degrees of freedom = 4

      Ha: mean < 1.6                                Ha: mean != 1.6                                Ha: mean > 1.6
Pr(T < t) = 0.6991                                Pr(|T| > |t|) = 0.6018                                Pr(T > t) = 0.3009

.
```

Phân tích hồi quy tuyến tính đơn giản

- Phương trình biểu diễn tương quan giữa hai biến (độc lập và phụ thuộc) là phương trình hồi quy đơn giản.
- Giả sử X là biến độc lập, Y là biến phụ thuộc
- $Y = \alpha X + \beta$ là phương trình hồi quy tuyến tính
- Câu lệnh Stata: *regress Y X*

Phân tích hồi quy tuyến tính đơn giản

Year	Thu nhập quốc dân (Yi)	Vốn đầu tư (Xi)
2000	20	10
2001	22	11
2002	25	12
2003	27	13
2004	30	14
2005	32	15
2006	33	16
2007	35	17
2008	36	18
2009	37	19

Phân tích hồi quy tuyến tính đơn giản

Phân tích: Thu nhập quốc dân (Y_i): biến phụ thuộc

Vốn đầu tư (X_i): biến độc lập

Câu lệnh Stata:

`regress Yi Xi`

`scatter Yi Xi`

Muốn kiểm tra xem 1 biến độc lập có ý nghĩa thống kê hay không thì ta nhìn vào chỉ số t. Nếu t-value của biến độc lập > 2 (Hoặc $> 1,96$) thì có thể kết luận là có mối quan hệ về mặt thống kê.

Ý nghĩa thống kê: thay đổi của biến độc lập có thể ảnh hưởng đến biến phụ thuộc hay không.

Phân tích hồi quy tuyến tính đơn giản

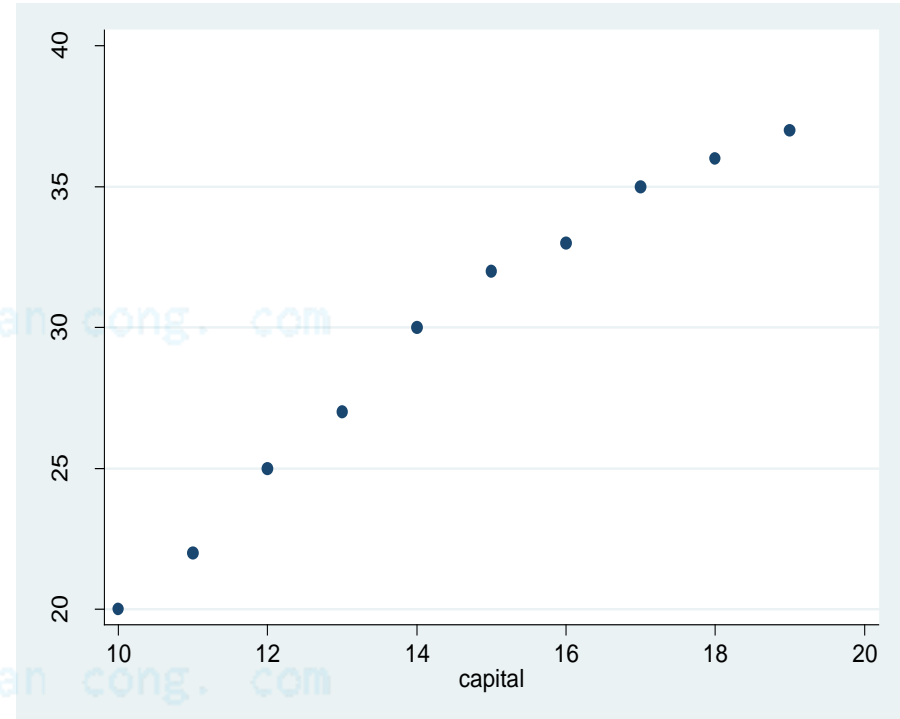
- Stata cho kết quả: $t\text{-stat}=17.8 \Rightarrow$ biến số capital có ý nghĩa thống kê
- $R^2=0,976 \Rightarrow 97,6\%$ độ biến thiên của thu nhập quốc dân có thể được giải thích bằng độ biến thiên của vốn

regress income capital

Source	SS	df	MS	Number of obs = 10		
Model	312.245455	1	312.245455	F(1, 8) = 318.03		
Residual	7.85454545	8	.981818182	Prob > F = 0.0000		
Total	320.1	9	35.5666667	R-squared = 0.9755		
				Adj R-squared = 0.9724		
				Root MSE = .99087		
income	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
capital	1.945455	.1090909	17.83	0.000	1.69389	2.197019
_cons	1.490909	1.612554	0.92	0.382	-2.227647	5.209465

Phân tích hồi quy tuyến tính đơn giản

- *regress income capital*
- Vẽ đồ thị:
- *scatter income capital =>*



Phân tích hồi quy đa biến

- Mô hình hồi quy đa biến có dạng $Y=f(X)$
- Với các mô hình phi tuyến tính có thể chuyển thành dạng tuyến tính. Ví dụ như với dạng hàm số mũ có thể chuyển thành tuyến tính bằng cách lấy logarithm hai vế
- Hàm sản xuất: $Y = AX^\alpha L^\beta$ trong đó X , L là vốn và lao động. Hàm này có thể được chuyển thành dạng tuyến tính như sau:

$$\ln(Y) = \ln(A) + \alpha \ln(X) + \beta \ln(L)$$

$$\text{hay } y = A_0 + \alpha x_1 + \beta x_2$$

Phân tích hồi quy đa biến

Year	Thu nhập quốc dân (Yi)- tỷ USD	Vốn (Xi)- tỷ USD	Lao động (Li) - triệu người
2000	20	10	10
2001	22	11	10.5
2002	25	12	11
2003	27	13	11.7
2004	30	14	12
2005	32	15	12.1
2006	33	16	12.2
2007	35	17	12.5
2008	36	18	12.6
2009	37	19	12.8

Phân tích hồi quy đa biến

- Cú pháp câu lệnh trong STATA. Lệnh *gen* (viết tắt của generate) nhằm tạo ra biến mới.
- `gen y=ln(Yi)`
- `gen x1=ln(Xi)`
- `gen x2=ln(x2)`
- `regress y x1 x2`

[cuu duong than cong. com](http://cuuduongthancong.com)

[cuu duong than cong. com](http://cuuduongthancong.com)

Phân tích hồi quy đa biến

- Kết quả mô hình: lưu ý ý nghĩa thống kê các biến, R^2

```
. regress y x1 x2
```

Source	SS	df	MS	Number of obs = 10		
Model	.40779056	2	.20389528	F(2, 7) = 560.03		
Residual	.002548547	7	.000364078	Prob > F = 0.0000		
Total	.410339107	9	.045593234	R-squared = 0.9938		
				Adj R-squared = 0.9920		
				Root MSE = .01908		

y	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
x1	.5086284	.1222477	4.16	0.004	.2195585	.7976984
x2	1.271582	.3172915	4.01	0.005	.5213064	2.021857
_cons	-1.106403	.4723957	-2.34	0.052	-2.223441	.0106354

Phân tích hồi quy đa biến

- Câu hỏi

- Trong mô hình, vốn và lao động có đóng góp tới thu nhập không? Các hệ số của vốn và lao động có ý nghĩa thống kê không?
- Hệ số R^2 có ý nghĩa gì?
- Nếu vốn tăng 1% thì tăng trưởng kinh tế tăng bao nhiêu %?
Nếu lao động tăng 1% thì tăng trưởng kinh tế tăng bao nhiêu %.

Phân tích hồi quy đa biến

- Hồi quy với biến giả (dummy variable)
- Cũng mô hình và số liệu như trên, giả sử chúng ta dự đoán là việc VN tham gia WTO năm 2007 dẫn tới thay đổi mô hình tăng trưởng.
- Áp dụng mô hình với biến giả $wto = 1$ với các năm từ 2007-2009 và bằng 0 với các năm từ 2000-2006.
- Cú pháp trong STATA:

gen wto=0

replace wto=1 if year>=2007

regress y x1 x2 wto

Phân tích hồi quy đa biến

- Câu hỏi: Biến WTO có ý nghĩa thống kê không?
- Viết phương trình hồi quy
- Nếu mức ý nghĩa thống kê là 1% thì biến nào có ý nghĩa thống kê

```
gen wto=0
replace wto=1 if year>=2007
(3 real changes made)
edit y x1 x2 wto
preserve
regress y x1 x2 wto
```

Source	SS	df	MS	Number of obs =	10
Model	.408202471	3	.13606749	F(3, 6) =	382.10
Residual	.002136636	6	.000356106	Prob > F =	0.0000
Total	.410339107	9	.045593234	R-squared =	0.9948
				Adj R-squared =	0.9922
				Root MSE =	.01887

y	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
x1	.6542899	.1815491	3.60	0.011	.2100553 1.098524
x2	1.006043	.399284	2.52	0.045	.0290301 1.983055
wto	-.0272618	.025348	-1.08	0.323	-.0892861 .0347624
_cons	-.8315596	.5325188	-1.56	0.169	-2.134586 .471467

Bài tập 1

- Có số liệu như hình bên:
- Sử dụng Stata, nhập dữ liệu bằng hai cách
 - Nhập trực tiếp bằng Stata (lệnh Edit)
 - Nhập vào Excel rồi Import từ Stata
- Chạy hồi quy điểm theo thu nhập gia đình
- Rút ra các nhận xét về ý nghĩa thống kê, R^2
- Vẽ biểu đồ điểm (trục Y) theo thu nhập (trục X) trên Stata trong đó có đường thẳng thể hiện phương trình hồi quy

Điểm thi	Thu nhập gia đình
10	40
9	10
9	30
8	40
8	30
7	20
7	30
6	25
6	15
5	20
5	15
4	12
3	10
2	15

Bài tập 2

Một sinh viên đã tiến hành nghiên cứu mối quan hệ giữa Giá thuê nhà (triệu/tháng) và Số phòng của ngôi nhà. Dữ liệu thu thập từ mẫu gồm 10 ngôi nhà cho thuê và được kết quả như sau:

- Có thể dựa vào số phòng ngôi nhà để dự đoán giá thuê của ngôi nhà không?
- Rút ra ý nghĩa thống kê của biến X (Số phòng).
- Hệ số R^2 trong mô hình có ý nghĩa gì?
- Giả sử bạn có thể thu thập thêm dữ liệu để xác định các biến số có thể ảnh hưởng tới Giá thuê nhà. Bạn hãy thử liệt kê ba biến số có thể ảnh hưởng tới Giá thuê nhà, lý giải tại sao và dự đoán về mối quan hệ giữa các biến số này với Giá thuê nhà (thuận chiều hay ngược chiều).

STT	X = Số phòng	Y = Giá thuê nhà (triệu/tháng)
1	1	2
2	3	4
3	2	6
4	2	5
5	3	5
6	4	6
7	2	3
8	4	15
9	2	7
10	5	12

So sánh Stata với SPSS, Sas

Trên thế giới hiện đang có 3 chương trình phân tích thống kê thông dụng, đó là Stata, Spss và Sas.

Sas là chương trình mạnh nhất nhưng bản quyền đắt nhất, những người có trình độ cao ưa thích, rất khó học

Stata thông dụng trong các trường học, có đến phiên bản Stata12, vừa dễ học lại rất mạnh, các lệnh thực hiện trực tiếp và dễ dàng

Spss dễ sử dụng.

+ ***Quản lý dữ liệu:***

Sas quản lý dữ liệu tốt, cho phép thao tác dữ liệu hầu như với cách có thể.

Spss có một soạn thảo dữ liệu tương tự excel, tuy nhiên quản lý dữ liệu không mạnh.

Stata quản lý dữ liệu kém hơn Sas nhưng tốt hơn Spss.