

Khai thác dữ liệu khảo sát mức sống hộ gia đình Việt Nam (VHLSS) để làm đề tài nghiên cứu – sử dụng phần mềm Stata

- 1. Giới thiệu về Bộ dữ liệu khảo sát mức sống hộ gia đình Việt Nam (VHLSS 2008)**
 - 2. Khởi động Stata 11**
 - 3. Một vài lệnh quản lý dữ liệu đơn giản**
 - 4. Tạo bảng tần số**
 - 5. Tính các thống kê mô tả**
 - 6. Sơ lược về tương quan & hồi quy**
 - 7. Nối hai file dữ liệu bằng lệnh Merge**
 - 8. Trợ giúp**
- Phụ lục 1. Mở rộng về hồi quy bội*
- Phụ lục 2. Một số lệnh quản lý dữ liệu nâng cao*
- Phụ lục 3. Mô hình Logit*
- Phụ lục 4. Cấu trúc lệnh cơ bản trong Stata, vấn đề trọng số trong VHLSS*
- Phụ lục 5. Kiểu dữ liệu; một số lệnh, hàm, toán tử thường dùng*

1. Giới thiệu về Dữ liệu khảo sát mức sống hộ gia đình

Để cung cấp thông tin về mức sống dân cư phục vụ việc xây dựng, đánh giá chính sách ... Đến nay, Tổng cục thống kê đã tiến hành 6 cuộc điều tra mức sống lớn với 2 tên gọi khác nhau: khảo sát mức sống dân cư (1993-1994, 1997-1998); khảo sát mức sống hộ gia đình (năm 2002, năm 2004, năm 2006, năm 2008). Có lẽ, khoảng gần 2 năm nữa anh chị mới có được dữ liệu VHLSS của năm 2010!

Gần đây nhất là cuộc khảo sát/điều tra mức sống (thường được viết tắt là KSMS) hộ gia đình năm 2008. Dữ liệu điều tra từ cuộc điều tra này được lưu trữ trong bộ dữ liệu khảo sát mức sống hộ gia đình năm 2008 (thường gọi là VHLSS 2008). Chúng ta có thể khai thác bộ dữ liệu này để làm đề tài nghiên cứu/ bài viết chính sách. *(Bạn có thể liên hệ với Vụ Xã hội & Môi trường – Tổng cục thống kê về vấn đề bản quyền trong việc sử dụng bộ dữ liệu này, hỏi các thông tin cần thiết ...)*

Để tìm hiểu chi tiết về cuộc điều tra này, về cách chọn mẫu, tổ chức điều tra, phiếu điều tra, các khái niệm ..., chúng ta cần đọc thêm tài liệu **“Sổ tay khảo sát mức sống hộ gia đình 2008”** do Tổng cục Thống kê biên soạn. Dường như, người phân tích VHLSS nào cũng cần có quyển sổ tay này bên cạnh.

Chúng ta tìm hiểu sơ lược một số thông tin chung về KSMS 2008

1.1 Mục đích của khảo sát mức sống 2008

Thu thập các thông tin làm căn cứ đánh giá mức sống, đánh giá tình trạng nghèo đói và phân hoá giàu nghèo để phục vụ công tác hoạch định các chính sách, kế hoạch và các chương trình mục tiêu quốc gia của Đảng và Nhà nước nhằm không ngừng nâng cao mức sống dân cư trong cả nước, các vùng và các địa phương.

Cung cấp số liệu để tính quyền số chỉ số giá tiêu dùng.

Ngoài ra, thu thập thông tin phục vụ nghiên cứu, phân tích một số chuyên đề về quản lý điều hành và quản lý rủi ro và phục vụ tính toán tài khoản quốc gia.

1.2 Nội dung của khảo sát mức sống 2008

KSMS 2008 gồm những nội dung chủ yếu phản ánh mức sống của các hộ gia đình trên cả nước và những điều kiện kinh tế xã hội cơ bản (đặc điểm của xã/phường...) có tác động đến mức sống của người dân nơi họ sinh sống. Các nội dung cụ thể bao gồm:

a. Đối với hộ gia đình

- Một số đặc điểm về nhân khẩu học của các thành viên trong hộ, gồm: Tuổi, giới tính, dân tộc, tình trạng hôn nhân.

- Thu nhập của hộ gia đình, gồm: Mức thu nhập; thu nhập phân theo nguồn thu (tiền công, tiền lương; hoạt động sản xuất tự làm nông nghiệp, lâm nghiệp, thủy sản; hoạt động ngành nghề sản xuất kinh doanh dịch vụ tự làm của hộ gia đình; thu khác); thu nhập phân theo khu vực kinh tế và ngành kinh tế.

- Chi tiêu hộ gia đình: mức chi tiêu, chi tiêu phân theo mục đích chi và khoản chi (chi cho ăn, mặc, ở, đi lại, giáo dục, y tế, văn hoá, v.v... và chi khác theo danh mục các nhóm/khoản chi tiêu để tính quyền số chỉ số giá tiêu dùng).

- Trình độ học vấn, trình độ chuyên môn kỹ thuật của từng thành viên hộ gia đình.

- Tình trạng ốm đau, bệnh tật và sử dụng các dịch vụ y tế.
- Tình trạng việc làm, thời gian làm việc.
- Tài sản, nhà ở và các tiện nghi như đồ dùng, điện, nước, điều kiện vệ sinh.
- Tham gia chương trình xóa đói giảm nghèo, tình hình tín dụng.
- Quản lý điều hành và quản lý rủi ro

b. Đối với xã

- Một số tình hình chung về nhân khẩu, dân tộc.
- Kết cấu hạ tầng kinh tế - xã hội chủ yếu, gồm: hiện trạng điện, đường, trường học, trạm y tế, chợ, buro điện, nguồn nước.
- Tình trạng kinh tế, gồm: Tình hình sản xuất nông nghiệp (đất đai, xu hướng và nguyên nhân tăng giảm sản lượng các cây trồng chính, các điều kiện hỗ trợ phát triển sản xuất như tưới tiêu, khuyến nông); cơ hội việc làm phi nông nghiệp.
- Một số thông tin cơ bản về trật tự an toàn xã hội và bảo vệ môi trường.

1.3 Mẫu khảo sát**a. Đối tượng, phạm vi, thời điểm khảo sát**

Đối tượng khảo sát gồm các hộ gia đình, các thành viên hộ gia đình và các xã có các hộ gia đình được khảo sát. Đơn vị khảo sát gồm hộ gia đình và xã được chọn khảo sát.

Phạm vi khảo sát bao gồm tất cả các địa bàn, các xã được chọn thuộc 64 tỉnh, thành phố trực thuộc trung ương (sau đây gọi tắt là tỉnh/thành phố).

Thời điểm khảo sát gồm hai kỳ vào tháng 5 và tháng 9 năm 2009. Thời gian thu thập thông tin tại địa bàn mỗi kỳ kéo dài 2 tháng.

b. Mẫu khảo sát

Mẫu 1: Mẫu để khảo sát mức sống 2008 và để tính quyền số chỉ số giá tiêu dùng (CSGTD). Mẫu này chọn từ dàn mẫu chủ thiết kế cho các cuộc KSMS giai đoạn 2000-2010 gồm 3.063 xã/phường, mỗi xã/phường chọn 3 địa bàn từ các địa bàn của Tổng điều tra Dân số và Nhà ở năm 1999.

Cỡ của Mẫu 1 gồm 45.945 hộ được chọn từ 3.063 địa bàn của dàn mẫu chủ, chia làm 2 loại:

- *Mẫu thu nhập và quyền số CSGTD* gồm 36.756 hộ để thu thập các nội dung thông tin đã nêu trên và quyền số CSGTD, trừ chi tiêu của hộ gia đình để đánh giá mức sống ở cấp quốc gia, vùng và tỉnh/thành phố, đồng thời để tính quyền số CSGTD. Mẫu này phỏng vấn Phiếu số 1A-PVH/KSMS08;

- *Mẫu thu nhập chi tiêu* gồm 9.189 hộ để thu thập đầy đủ các nội dung thông tin đánh giá, phân tích mức sống một cách sâu hơn ở cấp quốc gia và vùng (không có thông tin để tính quyền số CSGTD). Mẫu này phỏng vấn Phiếu số 1B-PVH/KSMS08.

Mẫu 2: Mẫu chỉ để tính quyền số CSGTD, gồm 2 phần, **Phần 1** gồm 9.189 hộ gia đình được chọn thêm từ 3.063 địa bàn của Mẫu 1, mỗi địa bàn chọn 3 hộ gia đình; và **Phần 2** gồm 15.000 hộ

được chọn từ 1000 địa bàn của Tổng điều tra Dân số và Nhà ở năm 1999 ngoài mẫu chủ. Mẫu này phỏng vấn Phiếu số 1C-PVH/QS08.

c. Các bước chọn mẫu

Đối với Mẫu 1:

Bước 1: Chọn địa bàn.

Các địa bàn của Mẫu 1 sẽ được chọn theo cách luân phiên, cụ thể: chọn lại 50% số địa bàn của KSMS 2006 (trong đó có một nửa số địa bàn đã được khảo sát cả trong KSMS 2004 và 2006 và nửa số địa bàn còn lại chỉ được khảo sát trong KSMS 2006) và 50% số địa bàn còn lại được chọn mới hoàn toàn từ dàn mẫu chủ, phần chưa được chọn vào mẫu của KSMS 2004 và 2006.

Vụ Thống kê Xã hội và Môi trường chịu trách nhiệm chọn và gửi danh sách địa bàn đã chọn cho các Cục Thống kê để rà soát và cập nhật, trong đó có gửi kèm cả sơ đồ và bảng kê của Tổng điều tra Dân số và Nhà ở năm 1999 của các địa bàn mới. Các Cục Thống kê tỉnh/thành phố có thể xem xét, đề nghị điều chỉnh một số địa bàn cho phù hợp hơn với các đặc điểm địa lý, kinh tế, xã hội thực tế của địa phương, nhưng số địa bàn đề nghị điều chỉnh không vượt quá 5% tổng số địa bàn của tỉnh/thành phố và phải được sự đồng ý của TCTK (Vụ XHMT) trước khi tiến hành khảo sát.

Bước 2: Chọn hộ.

Cục Thống kê chọn hộ khảo sát, cụ thể:

- Đối với những địa bàn chọn lại từ KSMS 2006, chọn tất cả 15 hộ, trong đó 12 hộ đã khảo sát thu nhập (hộ thu nhập) năm 2006 để khảo sát thu nhập cho KSMS 2008 và 3 hộ đã khảo sát thu nhập chi tiêu (hộ thu nhập chi tiêu) năm 2006 để khảo sát thu nhập chi tiêu cho KSMS 2008. Trong trường hợp có những hộ đã được khảo sát năm 2004 hoặc 2006 nhưng nay đã di dời địa bàn thì phải chọn hộ dự bị thay thế để có đủ số lượng 12 hộ thu nhập và 3 hộ thu nhập chi tiêu ở mỗi địa bàn khảo sát.

- Đối với những địa bàn mới, chọn 20 hộ từ danh sách hộ đã cập nhật của địa bàn. Từ 20 hộ được chọn, chọn 15 hộ (12 hộ chính thức, 3 hộ dự phòng) để khảo sát thu nhập; 5 hộ còn lại (3 chính thức và 2 dự phòng) để khảo sát thu nhập chi tiêu.

Việc chọn hộ khảo sát được thực hiện theo phương pháp nêu trong Sổ tay hướng dẫn nghiệp vụ KSMS 2008.

Đối với Mẫu 2:

- Đối với Phần 1 của Mẫu 2: Chọn 5 hộ (3 hộ chính thức và 2 hộ dự bị) từ danh sách hộ đã cập nhật của mỗi địa bàn trong 3.063 địa bàn của Mẫu 1 (trừ các hộ đã được chọn vào Mẫu 1) để thu thập thông tin tính quyền số CSGTD..

- Đối với Phần 2 của Mẫu 2: chọn 20 hộ từ danh sách hộ đã cập nhật của mỗi địa bàn trong 1.000 địa bàn của Phần 2 Mẫu 2. Từ 20 hộ được chọn, chọn 15 hộ chính thức và 5 hộ dự phòng để thu thập thông tin tính quyền số CSGTD.

Cục Thống kê tỉnh/thành phố sẽ chia số địa bàn được phân bổ của từng khu vực thành thị/nông thôn và vùng địa lý cho 2 kỳ khảo sát vào tháng 5 và tháng 9 như sau: 2/3 địa bàn của Mẫu 1, kể cả 3 hộ của Phần 1 Mẫu 2 để khảo sát vào kỳ tháng 5; số địa bàn còn lại khảo sát vào kỳ tháng 9. Các xã có địa bàn được chọn phỏng vấn hộ sẽ đồng thời tiến hành phỏng vấn Phiếu phỏng vấn xã.

Danh sách địa bàn và hộ được chọn sẽ được lưu giữ tại 2 địa chỉ: Cục Thống kê tỉnh/thành phố và Vụ Thống kê Xã hội và Môi trường để phục vụ việc tổ chức thực hiện và theo dõi, kiểm tra, giám sát.

Mẫu thu nhập và mẫu thu nhập chi tiêu được phân bổ cho 2 thời điểm khảo sát như sau:

Thời gian thu thập số liệu	Mẫu 1 Thu nhập và quyền số CSGTD	Mẫu 1 Thu nhập chi tiêu	Mẫu 2 Phần 1	Mẫu 2 Phần 2	Cộng
TỔNG SỐ					
<i>Chia ra:</i>	36.756	9.189	9.189	15.000	70.134
Tháng 5-6/2008	24.504	6.126	6.126		36.756
Tháng 9-10/2008	12.252	3.063	3.063	15.000	33.378

1.4 Phương pháp thu thập dữ liệu

Cuộc khảo sát này sử dụng hai loại phiếu phỏng vấn: loại phiếu phỏng vấn hộ gia đình và loại phiếu phỏng vấn xã. Loại phiếu phỏng vấn hộ gia đình gồm: Phiếu phỏng vấn thu nhập chi tiêu (áp dụng cho mẫu thu nhập chi tiêu) bao gồm tất cả các thông tin của nội dung khảo sát; Phiếu phỏng vấn thu nhập và quyền số CSGTD (áp dụng cho mẫu thu nhập và quyền số CSGTD) gồm các thông tin của nội dung khảo sát trừ các thông tin về chi tiêu của hộ và thêm thông tin để tính quyền số CSGTD; và Phiếu quyền số CSGTD (áp dụng cho mẫu chỉ thu thập thông tin để tính quyền số CSGTD). Phiếu phỏng vấn được thiết kế tương đối chi tiết giúp điều tra viên ghi chép thuận lợi, đồng thời tránh bỏ sót các khoản mục và tăng tính thống nhất giữa các điều tra viên, từ đó nâng cao chất lượng số liệu khảo sát.

Cuộc khảo sát áp dụng phương pháp phỏng vấn trực tiếp. Điều tra viên đến hộ, gặp chủ hộ và những thành viên trong hộ có liên quan để phỏng vấn và ghi thông tin vào phiếu phỏng vấn hộ gia đình. Đội trưởng đội khảo sát phỏng vấn lãnh đạo xã và các cán bộ địa phương có liên quan và ghi thông tin vào phiếu phỏng vấn xã. Để bảo đảm chất lượng thông tin thu thập, cuộc khảo sát không chấp nhận phương pháp khảo sát gián tiếp hoặc sao chép các thông tin từ các nguồn có sẵn khác vào phiếu phỏng vấn.

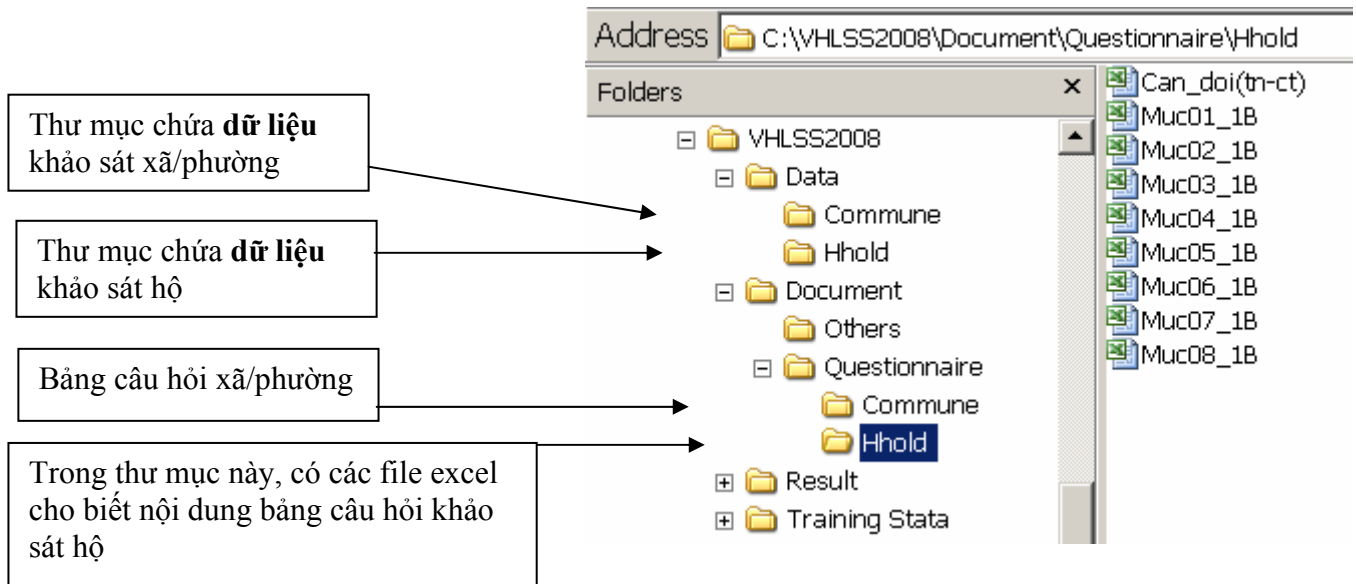
1.5 Dữ liệu

Có 2 loại dữ liệu chính: dữ liệu khảo sát xã, và dữ liệu khảo sát hộ. Chúng ta sẽ tìm hiểu về dữ liệu khảo sát hộ, vì nó được sử dụng khá phổ biến. Dữ liệu khảo sát xã cũng được khai thác tương tự.

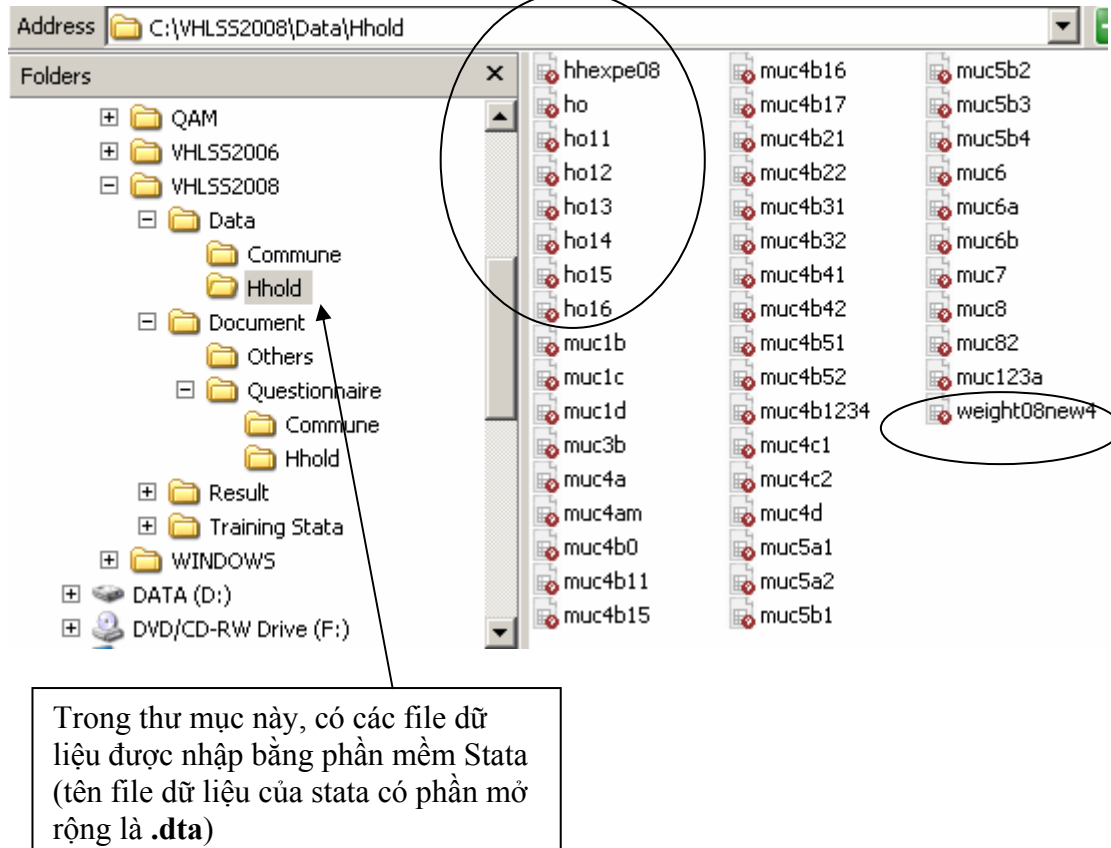
Trong dữ liệu khảo sát hộ, những người làm nghiên cứu thường hay sử dụng mẫu thu nhập và chi tiêu (9189 hộ) để thực hiện phân tích vì có đầy đủ dữ liệu về tất cả các biến.

Dữ liệu VHLSS2008 do tổng cục thống kê cung cấp thường được lưu trong đĩa CD. Sau khi chép sang ổ đĩa C của máy tính, có dạng như Hình 1.

Hình 1

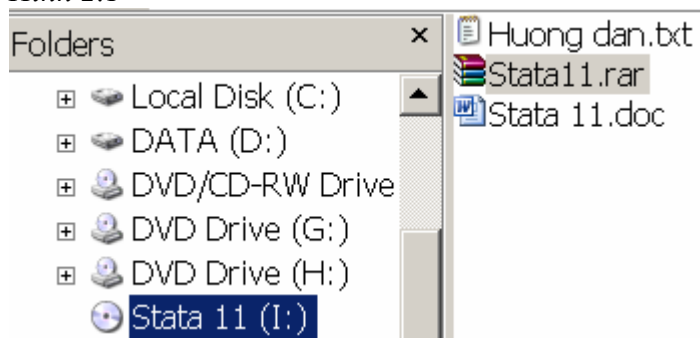


Hình 2

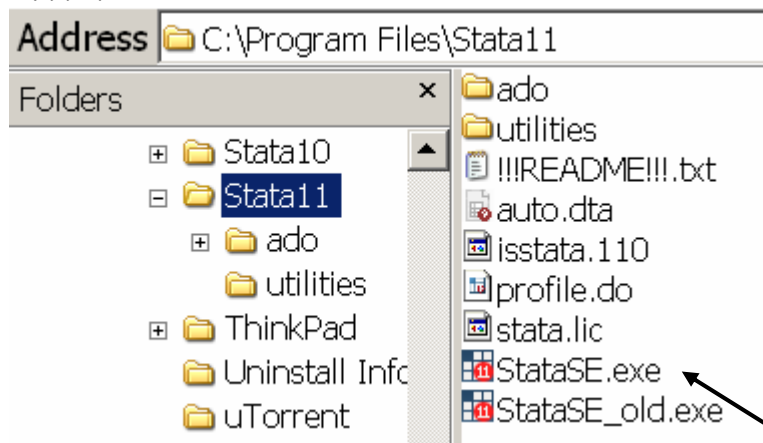


2. Khởi động

Hình 2.1



Hình 2.2



- **Khởi động Stata?**

Để khởi động Stata11, đơn giản, bạn hãy double-click vào biểu tượng StataSE.exe, hoặc double-click vào biểu tượng Shortcut của Stata trên desktop

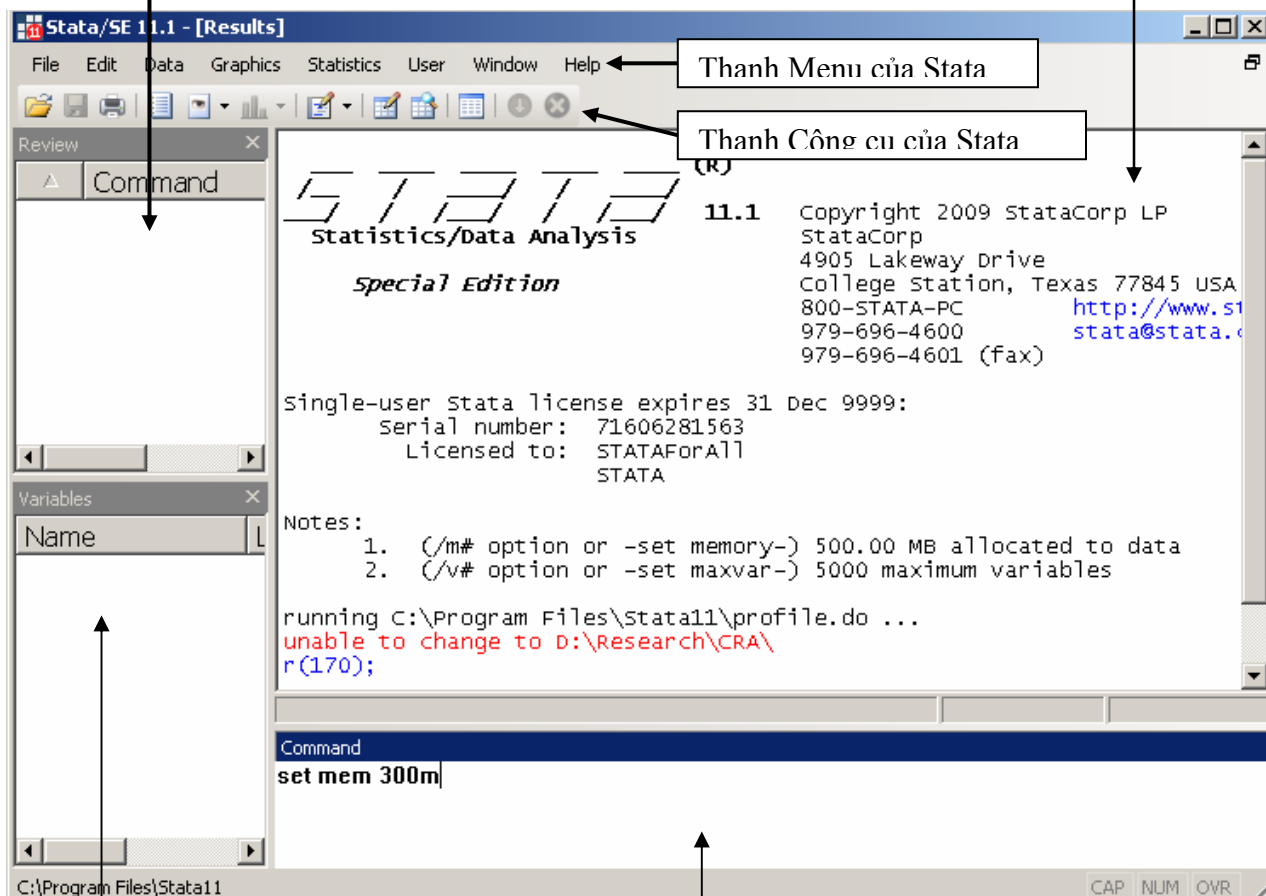


- **Màn hình STATA?**

Cửa sổ Review: cửa sổ này sẽ liệt kê các lệnh trong quá khứ bạn đã sử dụng

Cửa sổ Results: cửa sổ này hiện các kết quả tính toán, các thông báo của Stata...

Hình 2.3



Cửa sổ Variables: Cửa sổ này sẽ liệt kê danh sách các biến của file dữ liệu mà bạn đang mở

Cửa sổ Command: dùng để gõ các lệnh của Stata

- **Thoát khỏi Stata?**

\- Hãy thử gõ lệnh **exit** vào cửa sổ lệnh! Hoặc Bấm nút **✕** trong Hình 2.3

3. Một vài lệnh Quản lý dữ liệu đơn giản

- **Khai báo dung lượng bộ nhớ dành cho Stata?**

- Trong cửa sổ lệnh ở Hình 2.3, bạn đang gõ câu lệnh **set mem 300m**
Khi gõ lệnh này, bạn muốn máy tính dành cho Stata 300 megabytes bộ nhớ

Cấu trúc lệnh cơ bản: **set mem #[b|k|m|g]**

Với # là số bytes, kilobytes, megabytes, hay gigabytes (tương ứng với b, k, m, hay g được gõ phía sau), mặc định là k

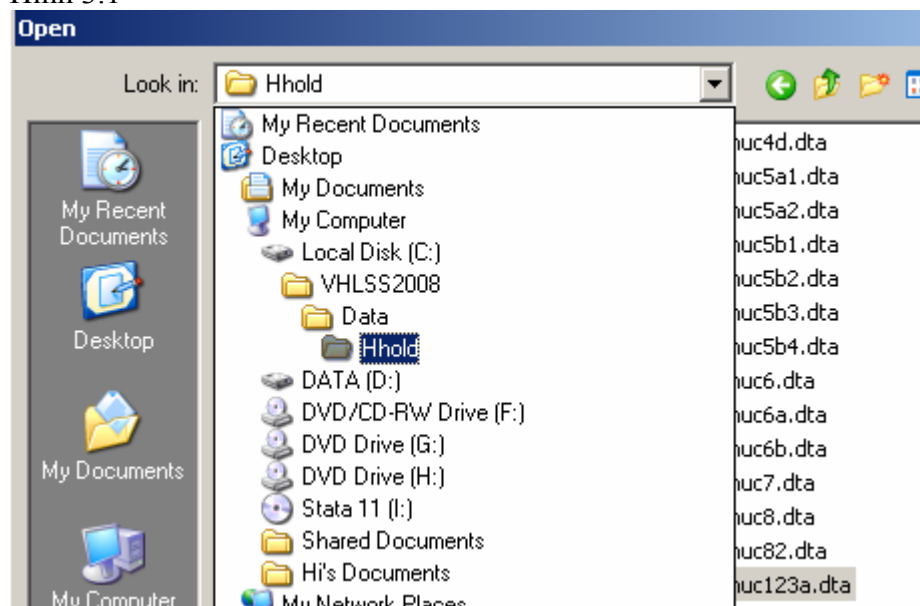
- **Mở 1 file dữ liệu?**

Cách 1

Từ thanh Menu của Stata, chọn **File\Open** → Chỉ đường dẫn đến file cần mở → **Open**

Ví dụ. Hình 3.1 chỉ ra đường dẫn của file dữ liệu **muc123a.dta** trong thư mục
C:\VHLSS2008\Data\Hhold

Hình 3.1



Cách 2

Bạn hãy gõ lệnh sau vào cửa sổ lệnh của Stata:

use "C:\VHLSS2008\Data\Hhold\muc123a.dta", clear

- **Xem thông tin sơ bộ về các biến (tên biến, nhãn biến, kiểu dữ liệu...)?**

- Bạn hãy kéo thanh trượt ở **Cửa sổ Variables** → Có những biến gì trong file **muc123a.dta** nhỉ?

- Gõ lệnh **des** vào cửa sổ lệnh bạn sẽ thấy những thông tin sau ở Cửa sổ kết quả:

Bảng 3.1

Contains data from C:\VHLSS2008\Data\Hhold\muc123a.dta

obs: 38,253
 vars: 56 11 Mar 2010 15:26
 size: 4,934,637 (98.4% of memory free)

variable name	storage type	display format	value label	variable label
tinhh	int	%8.0g		
huyen	byte	%8.0g		
xa	double	%8.0g		
diaban	int	%8.0g		
hoso	int	%8.0g		
matv	byte	%8.0g		Mã hiệu
mlac2	byte	%8.0g	M1AC2	2. Giới tính
mlac3	byte	%14.0g	M1AC3	3. Quan hệ
mlac4a	byte	%8.0g		4. Tháng sinh
mlac4b	int	%8.0g		Năm sinh
mlac5	int	%8.0g		5. Tuổi
mlac6	byte	%10.0g	M1AC6	6. Hôn nhân
mlac7	byte	%8.0g		7. Tháng ở hộ
mlac7a	byte	%26.0g	M1AC7A	7a. Lý do
mlac8	byte	%16.0g	M1AC8	8. Hộ khẩu
mlac9	int	%18.0g	M1AC9	9. Nơi đang ky HK
mlac10a	int	%8.0g		10. Năm ở tỉnh
mlac10b	byte	%8.0g		10. Tháng ở tỉnh
m2ac1	byte	%26.0g	M2AC1	1.Học hết lớp
m2ac2	byte	%8.0g	M2AC2	2.Biết đọc, biết viết
m2ac3a	byte	%11.0g	M2AC3A	3.Bằng cấp cao nhất - GDPT
m2ac3b	byte	%14.0g	M2AC3B	Bằng cấp cao nhất - GDNN
m2ac4	byte	%8.0g	M2AC4	4.Loại trường đã TN
m2ac5	byte	%8.0g	M2AC5	5.Hiện có đi học
m2ac6	byte	%8.0g	M2AC6	6.12 tháng qua có đi học
m2ac7	byte	%17.0g	M2AC7	7.Lý do bỏ đi học
m2ac8	byte	%14.0g	M2AC8	8.Hệ/cấp/bậc đang học
m2ac9	byte	%8.0g	M2AC9	9.Loại trường
m2ac10	byte	%8.0g	M2AC10	10.Có miễn giảm
m2ac11a	byte	%18.0g	M2AC11A	11.Lý do miễn giảm học phí
m2ac11b	byte	%18.0g	M2AC11B	Lý do miễn giảm đóng góp
m2ac12a	int	%8.0g		12.% miễn giảm học phí
m2ac12b	int	%8.0g		% miễn giảm đóng góp
m2ac13a	long	%12.0g		13a.Chi học phí
m2ac13b	long	%12.0g		13b.Chi trái tuyến
m2ac13c	long	%12.0g		13c.Chi đóng góp
m2ac13d	long	%12.0g		13d.Chi quỹ
m2ac13e	long	%12.0g		13e.Chi đồng phục
m2ac13f	long	%12.0g		13f.Chi sách giáo khoa
m2ac13g	long	%12.0g		13g.Chi dụng cụ học tập
m2ac13g1	long	%12.0g		13g1. Giấy vở, sổ
m2ac13g2	long	%12.0g		13g2. Cặp, bút
m2ac13g3	long	%12.0g		13g3. Máy tính, sách đ.tử
m2ac13h	long	%12.0g		13h.Chi học thêm
m2ac13i	long	%12.0g		13i.Chi giáo dục khác
m2ac13i1	long	%12.0g		13i1.Chi nhà trọ SV
m2ac13k	long	%12.0g		13k.Tổng số (a+b+...+i)
m2ac14	long	%12.0g		14.Các khoản nhận
m2ac15	long	%12.0g		15.Giá trị học bổng
m2ac16	long	%12.0g		16.Chi giáo dục-đào tạo khác
m3c1	byte	%8.0g	M3C1	1. 4 tuần, có bị ốm/bệnh
m3c2	byte	%8.0g	M3C2	2. 12 tháng, có bị ốm/bệnh
m3c3a	int	%8.0g		3. Số ngày nằm 1 chỗ
m3c3b	int	%8.0g		Số ngày nghỉ việc
m3c4	byte	%8.0g	M3C4	4. Có BHYT miễn phí
m3c5	byte	%29.0g	M3C5	5. Loại BHYT

Sorted by: tinhh huyen xa diaban hoso

- Bạn hãy mở các file (và tìm đến sheet trong file tương ứng) bảng câu hỏi điều tra (*các file Excel, ví dụ Muc01_1B.xls, Muc02_1B.xls, Muc03_1B.xls*) liên quan đến các biến ở file dữ liệu mà bạn đang mở (*Ví dụ, file muc123a.dta*) để biết thêm chi tiết về các biến.

Hình 3.2

Muc01_1B.xls													
	A	B	C	D	E	F	G	H	I	J	K	L	
1	MỤC 1A. DANH SÁCH THÀNH VIÊN HỘ GIA ĐÌNH												
2													
3		1. Xin [ÔNG/BÀ] vui lòng cho biết họ và tên của từng thành viên trong hộ, bắt đầu từ chủ hộ	2	Giới tính của [TÊN].	3	Quan hệ của ...[TÊN]... với chủ hộ?	4	Tháng, năm sinh của ...[TÊN]... THEO DƯƠNG LỊCH KHÔNG BIẾT THÁNG SINH	5	Tuổi của [TÊN] là bao nhiêu?	6	Tình trạng hôn nhân của [TÊN]?	7
4	M	Thành viên trong hộ là những người ăn, ở chung từ 6 tháng trở lên trong 12 tháng qua và chung quỹ thu, chi											
5	A												
6	T												
7	H												
8	A												
9	N												
10	H												
11	V												
12	I												
13	E												
14	N	GHI HỌ TÊN BẰNG CHỮ IN HOA VÀ THEO THỨ TỰ GIA ĐÌNH HẠT NHÂN (ĐIV LUY Y CAC TRUONG HUP ĐANG BIET)	NAM.....1 NỮ.....2	ÔNG/BÀ NỘI/NGOẠI.....5 CHÁU NỘI/NGOẠI.....6 QUAN HỆ KHÁC.....7									
15													
16	1												
17	2												
18	3												
19	4												
20	5												
21	6												

Hình 3.3

Muc02_1B.xls														
	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	MỤC 2. GIÁO DỤC													
2	Xin [ông/bà] vui lòng cho biết một số thông tin về giáo dục của các thành viên trong hộ													
3														
4		1	2	3	4	5	6							
5	M	...[TÊN]...	...[TÊN]...	Bằng cấp cao nhất mà [TÊN]	Trường mà [TÊN] học	Hiện nay ...[TÊN]...	Trong 12 tháng qua							
6	A	đã học hết lớp mấy?	có biết đọc, biết viết không?	đã đạt được?	thuộc loại nào?	có đi học không?	...[TÊN]...							
7				KHÔNG CÓ BẰNG CẤP.....0			...[TÊN]...							
8	T			TIỂU HỌC.....1			có đi học không?							
9	H	QUI ĐỔI LỚP THEO HỆ		TRUNG HỌC CƠ SỞ.....2										
10	A	12 NĂM		TRUNG HỌC PHỔ THÔNG.....3										
11	H			SƠ CẤP NGHỀ.....4										
12	N			TRUNG CẤP NGHỀ.....5										
13	H	CHƯA HẾT LỚP 1 HOẶC CHƯA BAO GIỜ ĐI HỌC		TRUNG HỌC CHUYÊN NGHIỆP.....6										
14	V	GHI 00		CAO ĐẲNG NGHỀ.....7										
15				CAO ĐẲNG.....8										
16	I			ĐẠI HỌC.....9										
17	E	TỪ LỚP 5 TRỞ LÊN >>3	CÓ.....1 KHÔNG.....2	TIẾN SĨ.....10										
18	N			KHÁC (GHI RÕ.....).....11										
19				GIÁO DỤC PHỔ THÔNG VÀ CAO ĐẲNG TRỞ LÊN	GIÁO DỤC NGHỀ NGHIỆP									
20														
21														
22	1													
23	2													
24	3													

- Trong Hình 3.2 (file dữ liệu thể hiện **Mục 1A** của bảng câu hỏi), bạn hãy thử xem **câu 2** là câu hỏi về điều gì? → Giới tính của thành viên trong hộ. Và quy ước mã hoá khi nhập liệu như sau: Nam thì nhập 1, Nữ thì nhập 2. → Sẽ có biến **m1ac2** chứa đựng thông tin về giới

- Trong Hình 3.2 (file dữ liệu thể hiện **Mục 1A** của bảng câu hỏi), bạn hãy thử xem **câu 5** là câu hỏi về điều gì? → Tuổi của thành viên → Sẽ có biến **m1ac5** cho biết thành viên của hộ ở tuổi nào.

- Theo bạn biến **m1ac2**, **m1ac5** là biến định tính hay biến định lượng? → Có lẽ bạn sẽ trả lời là **m1ac2** là biến định tính, còn **m1ac5** là biến định lượng.

- bạn hãy thử gõ lệnh **des m1ac2 m1ac5 m2ac6** xem điều gì xảy ra?

```
. des m1ac2 m1ac5 m2ac6
```

variable name	storage type	display format	value label	variable label
m1ac2	byte	%8.0g	M1AC2	2. Giới tính
m1ac5	int	%8.0g		5. Tuổi
m2ac6	byte	%8.0g	M2AC6	6.12 tháng qua có đi học

- **Xem thông tin sơ bộ về các biến (các giá trị của biến) - lệnh codebook?**

codebook m1ac2 m1ac5

```
m1ac2 2. Giới tính
```

```

      type:  numeric (byte)
      label:  M1AC2
      range:  [1,2]
unique values:  2
      units:  1
missing .:  0/38253

      tabulation:  Freq.    Numeric    Label
                   18810         1    Nam
                   19443         2    Nữ

```

```
m1ac5 5. Tuổi
```

```

      type:  numeric (int)
      range:  [0,103]
unique values:  102
      mean:    31.784
      std. dev: 20.6508
      units:  1
missing .:  0/38253

      percentiles:  10%    25%    50%    75%    90%
                   7      15      28      46      60

```

- **Đếm số quan sát trong bộ dữ liệu? → lệnh count**

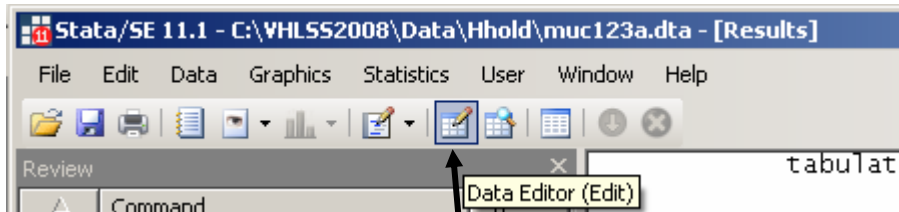
- Khi gõ lệnh **count** vào cửa sổ lệnh, bạn sẽ thấy thông tin sau trên cửa sổ kết quả

```
. count
```

```
38253
```

- **Xem dữ liệu? → Mở cửa sổ Data Editor**

Hình 3.4



Để xem dữ liệu, bạn có thể bấm **nút Data Editor**, hoặc gõ lệnh edit vào cửa sổ lệnh

Hình 3.5

Data Editor (Browse) - [muc123a.dta]

File Edit Data Tools

hoson[7] 15

	ting	huyen	xa	diaban	hoson	matv	m1ac2	m1ac3	m1ac4a	m1ac4b	m1ac5
1	101	1	3	1	13	1	N÷	Chñ hế	9	1934	73
2	101	1	3	1	13	2	N÷	Con	1	1969	39
3	101	1	3	1	13	3	N÷	Con	8	1974	33
4	101	1	3	1	14	1	Nam	Chñ hế	10	1943	64
5	101	1	3	1	14	2	N÷	Vĩ chàng	5	1955	53
6	101	1	3	1	14	3	Nam	Con	4	1986	22
7	101	1	3	1	15	1	Nam	Chñ hế	10	1946	61
8	101	1	3	1	15	2	N÷	Vĩ chàng	10	1947	60
9	101	1	9	19	15	1	Nam	Chñ hế	11	1957	50
10	101	1	9	19	15	2	N÷	Vĩ chàng	10	1966	41
11	101	1	9	19	19	1	N÷	Chñ hế	4	1944	64
12	101	1	9	19	19	2	Nam	Vĩ chàng	7	1946	61
13	101	1	9	19	19	3	N÷	Con	2	1985	23
14	101	1	9	19	20	1	Nam	Chñ hế	10	1957	50
15	101	1	9	19	20	2	N÷	Vĩ chàng	4	1957	51
16	101	1	9	19	20	3	N÷	Con	3	1989	19
17	101	1	15	50	13	1	Nam	Chñ hế	12	1972	35
18	101	1	15	50	13	2	N÷	Vĩ chàng	5	1974	34
19	101	1	15	50	13	3	Nam	Con	2	2002	6
20	101	1	15	50	13	4	Nam	Con	2	2002	6

Hình 3.6

hoson	matv	m1ac2	m1ac3	m1ac4a	m1ac4b
13	1	N÷			
13	2	N÷			
13	3	N÷			
14	1	Nam			
14	2	N÷			
14	3	Nam			
15	1	Nam			
15	2	N÷			
15	1	Nam			
15	2	N÷			
19	1	N÷			
19	2	Nam			
19	3	Nam			
20					
20					

Context menu options:

- Copy (Ctrl+C)
- Paste (Ctrl+V)
- Select All (Ctrl+A)
- Variable Properties...
- Replace Contents of Variable...
- Sort...
- Hide Selected Variables
- Show Only Selected Variables
- Show Entire Dataset
- Keep Only Selected Data
- Drop Selected Data
- Value Labels (selected)
- Preferences...
- Font...
- Select Value from Value Label
- Assign Value Label to Variable
- Manage Value Labels...
- Hide All Value Labels

Hình 3.7

	tỉnh	huyen	xa	diaban	hoso	matv	m1ac2	m1ac3	m1ac4a	m1ac4b	m1ac5
1	101	1	3	1	13	1	2	1	9	1934	73
2	101	1	3	1	13	2	2	3	1	1969	39
3	101	1	3	1	13	3	2	3	8	1974	33
4	101	1	3	1	14	1	1	1	10	1943	64
5	101	1	3	1	14	2	2	2	5	1955	53
6	101	1	3	1	14	3	1	3	4	1986	22
7	101	1	3	1	15	1	1	1	10	1946	61
8	101	1	3	1	15	2	2	2	10	1947	60
9	101	1	9	19	15	1	1	1	11	1957	50
10	101	1	9	19	15	2	2	2	10	1966	41
11	101	1	9	19	19	1	2	1	4	1944	64
12	101	1	9	19	19	2	1	2	7	1946	61
13	101	1	9	19	19	3	2	3	2	1985	23
14	101	1	9	19	20	1	1	1	10	1957	50
15	101	1	9	19	20	2	2	2	4	1957	51
16	101	1	9	19	20	3	2	3	3	1989	19

Từ Hình 3.6, nếu bạn muốn dữ liệu được thể hiện như Hình 3.7 thì **click chuột phải → chọn Value lable → Hide All Value Labels**

- Khi muốn gõ lệnh gì đó tiếp theo trong cửa sổ lệnh, bạn nên đóng cửa sổ Data Editor lại.

4. Tạo bảng tần số

- **Tạo bảng tần số một chiều?**

. tab m1ac2

2. Giới tính	Freq.	Percent	Cum.
Nam	18,810	49.17	49.17
Nữ	19,443	50.83	100.00
Total	38,253	100.00	

. tab m2ac1

1. Học hết lớp	Freq.	Percent	Cum.
Chưa hết lớp 1/chưa đi học	5,664	14.81	14.81
1	945	2.47	17.28
2	1,680	4.39	21.67
3	1,985	5.19	26.86
4	2,029	5.30	32.16
5	3,200	8.37	40.53
6	2,316	6.05	46.58
7	2,337	6.11	52.69
8	1,987	5.19	57.89
9	6,692	17.49	75.38
10	1,336	3.49	78.87
11	1,223	3.20	82.07
TN THPT	6,859	17.93	100.00
Total	38,253	100.00	

. tab m2ac6

6.12 tháng qua có đi học	Freq.	Percent	Cum.
Có	617	2.18	2.18
Không	27,695	97.82	100.00
Total	28,312	100.00	

```
. tab m2ac6, m
```

6.12 tháng qua có đi học	Freq.	Percent	Cum.
Có	617	1.61	1.61
Không	27,695	72.40	74.01
.	9,941	25.99	100.00
Total	38,253	100.00	

```
. tab m2ac9 if m2ac6==1
```

9.Loại trường	Freq.	Percent	Cum.
Công lập	514	83.31	83.31
Bán công	55	8.91	92.22
Dân lập	32	5.19	97.41
Tư thực	9	1.46	98.87
Khác	7	1.13	100.00
Total	617	100.00	

```
. tab m2ac9 if m2ac6==1, nol
```

9.Loại trường	Freq.	Percent	Cum.
1	514	83.31	83.31
2	55	8.91	92.22
3	32	5.19	97.41
4	9	1.46	98.87
5	7	1.13	100.00
Total	617	100.00	

Tạo bảng tần số về tình trạng hôn nhân phân theo nam và nữ?

```
. sort m1ac2
. by m1ac2: tab m1ac6
```

```
-----
-> m1ac2 = Nam
```

6. Hôn nhân	Freq.	Percent	Cum.
Chưa VC	5,535	36.81	36.81
Đang có VC	9,082	60.41	97.22
Goá	314	2.09	99.31
Ly hôn	60	0.40	99.71
Ly thân	44	0.29	100.00
Total	15,035	100.00	

```
-----
-> m1ac2 = Nữ
```

6. Hôn nhân	Freq.	Percent	Cum.
Chưa VC	4,568	28.78	28.78
Đang có VC	9,209	58.01	86.79
Goá	1,798	11.33	98.12
Ly hôn	205	1.29	99.41
Ly thân	94	0.59	100.00
Total	15,874	100.00	

Tạo bảng tần số hai chiều?

. tab m2ac1 m1ac2

1.Học hết lớp	2. Giới tính		Total
	Nam	Nữ	
Chưa hết lớp 1/chưa đ	2,430	3,234	5,664
1	430	515	945
2	691	989	1,680
3	862	1,123	1,985
4	948	1,081	2,029
5	1,505	1,695	3,200
6	1,179	1,137	2,316
7	1,166	1,171	2,337
8	1,036	951	1,987
9	3,393	3,299	6,692
10	714	622	1,336
11	701	522	1,223
TN THPT	3,755	3,104	6,859
Total	18,810	19,443	38,253

. tab m2ac1 m1ac2, col nof

1.Học hết lớp	2. Giới tính		Total
	Nam	Nữ	
Chưa hết lớp 1/chưa đ	12.92	16.63	14.81
1	2.29	2.65	2.47
2	3.67	5.09	4.39
3	4.58	5.78	5.19
4	5.04	5.56	5.30
5	8.00	8.72	8.37
6	6.27	5.85	6.05
7	6.20	6.02	6.11
8	5.51	4.89	5.19
9	18.04	16.97	17.49
10	3.80	3.20	3.49
11	3.73	2.68	3.20
TN THPT	19.96	15.96	17.93
Total	100.00	100.00	100.00

. tab m2ac2 m1ac2, col

2.Biết đọc, biết viết	2. Giới tính		Total
	Nam	Nữ	
Có	2,830	3,485	6,315
	52.79	50.20	51.33
Không	2,531	3,457	5,988
	47.21	49.80	48.67
Total	5,361	6,942	12,303
	100.00	100.00	100.00


```
. tab m2ac9 m1ac2 if m2ac6==1
```

9.Loại trường	2. Giới tính		
	Nam	Nữ	Total
Công lập	282	232	514
Bán công	32	23	55
Dân lập	19	13	32
Tư thực	7	2	9
Khác	5	2	7
Total	345	272	617

```
. tab m2ac9 m1ac2 if m2ac6==1, col
```

9.Loại trường	2. Giới tính		
	Nam	Nữ	Total
Công lập	282 81.74	232 85.29	514 83.31
Bán công	32 9.28	23 8.46	55 8.91
Dân lập	19 5.51	13 4.78	32 5.19
Tư thực	7 2.03	2 0.74	9 1.46
Khác	5 1.45	2 0.74	7 1.13
Total	345 100.00	272 100.00	617 100.00

```
. tab m2ac9 m1ac2 if m2ac6==1, col nof
```

9.Loại trường	2. Giới tính		
	Nam	Nữ	Total
Công lập	81.74	85.29	83.31
Bán công	9.28	8.46	8.91
Dân lập	5.51	4.78	5.19
Tư thực	2.03	0.74	1.46
Khác	1.45	0.74	1.13
Total	100.00	100.00	100.00

5. Tính các thống kê mô tả

- Tính thống kê mô tả của một biến định lượng?**

```
.sum m1ac5
```

Variable	Obs	Mean	Std. Dev.	Min	Max
m1ac5	38253	31.78399	20.65079	0	103

```
. sum m2ac13k
```

Variable	Obs	Mean	Std. Dev.	Min	Max
m2ac13k	10558	1608.373	2669.863	0	46160

```
. sum m1ac5 m2ac13k
```

Variable	Obs	Mean	Std. Dev.	Min	Max
m1ac5	38253	31.78399	20.65079	0	103
m2ac13k	10558	1608.373	2669.863	0	46160

```
. sum m1ac5 m2ac13k if m2ac6==1
```

Variable	Obs	Mean	Std. Dev.	Min	Max
m1ac5	617	17.87358	6.121887	0	49
m2ac13k	617	2244.407	2853.302	0	32000

- Tính thống kê mô tả của một biến định lượng phân theo một biến định tính?**

Cách 1

```
. tab m1ac2, sum(m1ac5)
```

2. Giới tính	Summary of 5. Tuổi		
tính	Mean	Std. Dev.	Freq.
Nam	30.419139	19.914699	18810
Nữ	33.104408	21.256024	19443
Total	31.783991	20.650785	38253

```
. tab m2ac9 if m2ac6==1, sum (m2ac13k)
```

9.Loại trường	Summary of 13k.Tổng số (a+b+...+i)		
trường	Mean	Std. Dev.	Freq.
Công lập	2245.072	2741.4057	514
Bán công	1838.8727	1083.6352	55
Dân lập	2423.1563	2293.5377	32
Tư thực	4997	10167.356	9
Khác	1025.7143	1711.2944	7
Total	2244.4068	2853.3015	617

Cách 2.

```
. by m1ac2: sum m1ac5
```

```
-> m1ac2 = Nam
```

Variable	Obs	Mean	Std. Dev.	Min	Max
m1ac5	18810	30.41914	19.9147	0	97

```
-> m1ac2 = Nữ
```

Variable	Obs	Mean	Std. Dev.	Min	Max
m1ac5	19443	33.10441	21.25602	0	103

• **Tính thống kê mô tả của một biến định lượng phân theo 2 biến định tính?**

```
. table m2ac8 m2ac9 if m2ac6==1, c( mean m2ac13k) format(%7.1f)
```

8.Hệ/cấp/bậc đang học	9.Loại trường				
	Công lập	Bán công	Dân lập	Tư thực	Khác
Nhà trẻ, MG	945.9	886.6	660.0	1966.7	
Tiểu học	336.4		240.0		
THCS	687.4				
THPT	1518.4	2087.1	2631.7	8928.3	555.0
Sơ cấp nghề	4931.7	2840.0	0.0		0.0
Trung cấp nghề	5081.5		2900.0	1680.0	0.0
TH CN	3561.2		2490.0		1410.0
Cao đẳng nghề	4153.3		430.0		
Cao đẳng	4071.6		1820.0		
Đại học	5145.9		3895.0		4660.0
Thạc sĩ	9930.0				
Tiến sĩ	21000.0				

```
. table m2ac8 m2ac9 if m2ac6==1, c(count m2ac13k) format(%7.1f)
```

8.Hệ/cấp/bậc đang học	9.Loại trường				
	Công lập	Bán công	Dân lập	Tư thực	Khác
Nhà trẻ, MG	26	12	2	3	
Tiểu học	21		1		
THCS	64				
THPT	262	42	18	4	2
Sơ cấp nghề	15	1	1		2
Trung cấp nghề	11		2	2	1
TH CN	32		2		1
Cao đẳng nghề	3		1		
Cao đẳng	12		1		
Đại học	65		4		1
Thạc sĩ	2				
Tiến sĩ	1				

```
. table m2ac8 if m2ac6==1, c(count m2ac13k mean m2ac13k) format(%7.1f)
```

8.Hệ/cấp/bậc đang học		
	N(m2ac13k)	mean(m2ac13k)
Nhà trẻ, MG	43	987.3
Tiểu học	22	332.0
THCS	64	687.4
THPT	328	1736.8
Sơ cấp nghề	19	4042.9
Trung cấp nghề	16	4066.0
TH CN	35	3438.5
Cao đẳng nghề	4	3222.5
Cao đẳng	13	3898.4
Đại học	70	5067.5
Thạc sĩ	2	9930.0
Tiến sĩ	1	21000.0

Trong tùy chọn của một số lệnh, Stata cho phép các loại thống kê được chỉ ra bởi các thống kê như sau:

<u>Cú pháp thống kê</u>	<u>Ý nghĩa</u>
mean	Trung bình mean
count	Đếm số quan sát
n	Giống như lệnh count (Đếm số quan sát)
sum	Tổng cộng
max	Giá trị lớn nhất
min	Giá trị nhỏ nhất
range	Biên độ = Giá trị lớn nhất - Giá trị nhỏ nhất
sd	Độ lệch chuẩn
sdmean	Độ lệch chuẩn của trung bình = Độ lệch chuẩn / {(Số quan sát) ^{0.5} }
skewness	Độ lệch của phân phối
kurtosis	Độ nhọn
median	Trung vị (Giống như p50)
p1	1% phân vị
p5	5% phân vị
p10	10% phân vị
p25	25% phân vị
p50	50% phân vị (trung vị)
p75	75% phân vị
p90	90% phân vị
p95	95% phân vị
p99	99% phân vị
iqr	p75 - p25
q	tương đương với "p25 p50 p75"

Ví dụ:

```

tabstat m1ac5, stats (mean median iqr sd)
tabstat m1ac5, stats (mean median min max range sd var cv skewness kurtosis)
table m2ac8 m2ac9 if m2ac5<=2, c( mean m2ac13k)
table m2ac8 m2ac9 if m2ac5<=2, c( mean m2ac13k) format(%7.2f)
table m2ac8 m2ac9 if (m2ac5<=2) & (m2ac9<=4), c( mean m2ac13k) format(%7.2f)
table m2ac8 m2ac9 if m2ac5<=2, c(count m2ac13k) format(%7.1f)
table m2ac8 if m2ac5<=2, c(count m2ac13k mean m2ac13k) format(%7.1f)
table m2ac8 if m2ac5<=2, c(count m2ac13k mean m2ac13k mean m1ac5 ) format(%7.1f)
tab m1ac2, sum( m1ac5) mean

```

```
tabstat m1ac5, stats (mean median iqr sd)
```

variable	mean	p50	iqr	sd
m1ac5	31.78399	28	31	20.65079

table m2ac8 m2ac9 if m2ac5 <= 2, c(mean m2ac13k) format(%7.2f)

8.Hệ/cấp/bậc đang học	Công lập	Bán công	9.Loại trường Dân lập	Tư thục	Khác	Missing
Nhà trẻ, MG	723.19	688.55	1154.58	3291.18	1158.29	0.00
Tiểu học	539.26	923.67	16515.43	12830.00	1398.80	952.50
THCS	824.08	1559.52	10880.78	699.33	425.00	
THPT	1453.13	2174.89	3196.20	5483.21	1167.00	
Sơ cấp nghề	3077.13	3034.50	3400.00	3625.00	3039.29	
Trung cấp nghề	4113.61	6291.67	3922.00	3100.00	13980.00	
TH CN	4214.58	6438.33	4585.47	3400.00	11830.00	
Cao đẳng nghề	4898.46	5122.50	4994.60	4973.33	2850.00	
Cao đẳng	4986.06	7755.50	4644.58	5626.67	11712.50	
Đại học	5892.89	8702.14	7845.68	1973.00	4773.42	
Thạc sĩ	11495.33					
Khác	640.00					

. tab m1ac2, sum(m1ac5) mean

2. Giới tính	Summary of 5. Tuổi Mean
Nam	30.419139
Nữ	33.104408
Total	31.783991

6. Sơ lược về hồi quy

- *Tạo biến giả?*

Cách thủ công nhất như sau:

. gen gioi= m1ac2

. replace gioi=0 if m1ac2==2
(19443 real changes made)

- *Ước lượng hàm hồi quy?*

. reg m2ac13k gioi m1ac5

Source	SS	df	MS	Number of obs =	10558
Model	1.3691e+10	2	6.8455e+09	F(2, 10555) =	1173.70
Residual	6.1561e+10	10555	5832410.47	Prob > F =	0.0000
Total	7.5252e+10	10557	7128167.96	R-squared =	0.1819
				Adj R-squared =	0.1818
				Root MSE =	2415

m2ac13k	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
gioi	75.52587	47.04069	1.61	0.108	-16.68276 167.7345
m1ac5	196.9525	4.070503	48.39	0.000	188.9736 204.9315
_cons	-989.2827	62.38184	-15.86	0.000	-1111.563 -867.0025

```
. reg m2ac13k gioi m1ac5 if m2ac5<=2
```

Source	SS	df	MS	Number of obs =	9941
Model	1.2967e+10	2	6.4834e+09	F(2, 9938) =	1130.28
Residual	5.7005e+10	9938	5736081.11	Prob > F =	0.0000
Total	6.9972e+10	9940	7039428.34	R-squared =	0.1853
				Adj R-squared =	0.1851
				Root MSE =	2395

m2ac13k	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
gioi	68.67071	48.06328	1.43	0.153	-25.54307 162.8845
m1ac5	203.2432	4.278268	47.51	0.000	194.8569 211.6295
_cons	-1045.171	64.07087	-16.31	0.000	-1170.763 -919.5791

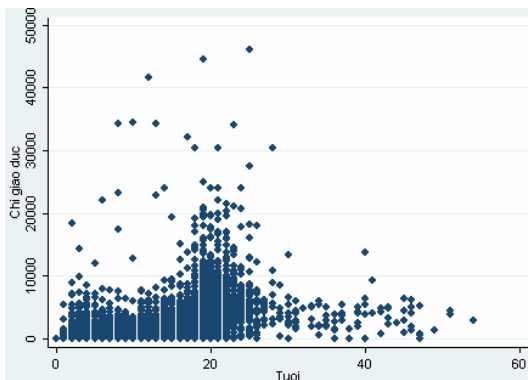
- Tính hệ số tương quan?**

```
. cor m2ac13k m1ac5 if m2ac6<=2
(obs=617)
```

	m2ac13k	m1ac5
m2ac13k	1.0000	
m1ac5	0.3382	1.0000

- Vẽ đồ thị phân tán?**

```
scatter m2ac13k m1ac5 if m2ac6==1
```



7. Nối 2 file dữ liệu bằng lệnh Merge

Giả sử, chẳng hạn bạn muốn nối file muc4a.dta vào muc123a.dta

Bước 1. Mở file using, sort, lưu lại tại một thư mục khác

```
. use "C:\VHLSS2008\Data\Hhold\muc4a.dta", clear

. count
35154

. sort  tinh huyen xa diaban hoso matv

. save "C:\VHLSS2008\muc4a_sorted.dta", replace
file C:\VHLSS2008\muc4a_sorted.dta saved
```

Bước 2. Mở file master, sort, dùng lệnh merge để nối

```
. use "C:\VHLSS2008\Data\Hhold\muc123a.dta", clear

. count
38253

. sort tinh huyen xa diaban hoso matv

. merge 1:1  tinh huyen xa diaban hoso matv using
"C:\VHLSS2008\muc4a_sorted.dta"
```

Result	# of obs.	
not matched	3,099	
from master	3,099	(<u>_merge</u> ==1)
from using	0	(<u>_merge</u> ==2)
matched	35,154	(<u>_merge</u> ==3)

Bước 3. Kiểm tra lại, xoá những quan sát không cần thiết, xoá biến _merge

```
. tab _merge
```

<u>_merge</u>	Freq.	Percent	Cum.
master only (1)	3,099	8.10	8.10
matched (3)	35,154	91.90	100.00
Total	38,253	100.00	

```
. keep if _merge==3
(3099 observations deleted)

. drop  _merge
```

Bạn hãy vào help để tìm hiểu thêm về lệnh merge trên stata 11!

Hình 7.1

	A	B	C	D	E	F	G	H	I
1	MỤC 4. THU NHẬP								
2	PHẦN 4A. TÌNH TRẠNG VIỆC LÀM								
4	HỎI VỀ TẤT CẢ CÁC THÀNH VIÊN CỦA HỘ TỪ 6 TUỔI TRỞ LÊN.								
6									
7	VIỆC LÀM CHIẾM NHIỀU THỜI GIAN NHẤT								
8	1 Trong 12 tháng qua, [ông/bà] có tham gia...			2 CÓ LÀM VIỆC?	3 Lý do ...[TÊN]... không làm việc trong 12 tháng qua?	4 Công việc nào chiếm nhiều thời gian nhất của ...[TÊN]... trong 12 tháng qua?		5 Công việc này	
9	M	a	b	c					
10	A	Đi làm để	Hoạt động sản	Hoạt động					
11		nhận tiền	xuất hoặc dịch	ngành nghề					
12	T	lượng, tiền	vụ về trồng trọt,	SXKD hoặc	(CÓ MÃ 1				
13	H	công?	chăn nuôi, làm	dịch vụ của	Ở CÂU 1)				
14	A		ng nghiệp và thu ý	hộ?					
15	N		sản cho hộ?						
16	H								
17		LÀM NHẬN	TỰ LÀM	TỰ SXKD,					
18	V	LƯƠNG,	NÔNG,	DV PHI					
19	I	CÔNG	LÀM, THỦY	NÔNG,					
20	E			LÀM, THỦY					
21	N	CÓ.....1	CÓ.....1	CÓ.....1					
22		KHÔNG.....2	KHÔNG.....2	KHÔNG.....2	(>>4)				
23									
24									
25									
26									
27									
28									
29									

Hình 7.2

```
. merge 1:1 pid time using filename
```

master + using = merged result

pid	time	x1
14	1	0
14	2	0
14	4	0
16	1	1
16	2	1
17	1	0

pid	time	x2
14	1	7
14	2	9
16	1	2
16	2	3
17	1	5
17	2	2

pid	time	x1	x2	_merge
14	1	0	7	3
14	2	0	9	3
14	4	0	.	1
16	1	1	2	3
16	2	1	3	3
17	1	0	5	3
17	2	.	2	2

This is a 1:1 merge because the combination of the values of **pid** and **time** uniquely identifies observations in both datasets.

Hình 7.4

```
. merge m:1 region using filename
```

master + using = merged result

id	region	a
1	2	26
2	1	29
3	2	22
4	3	21
5	1	24
6	5	20

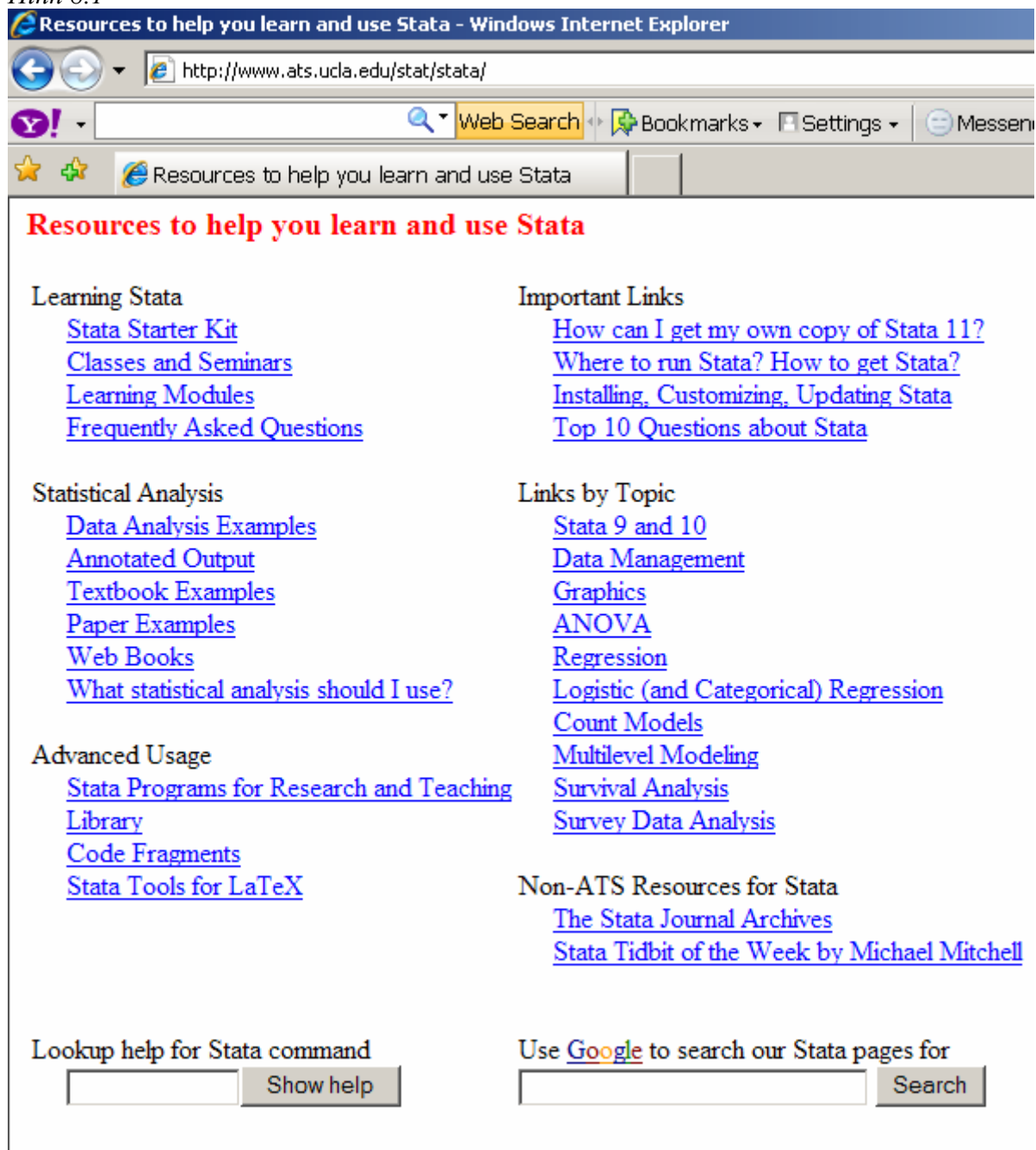
region	x
1	15
2	13
3	12
4	11

id	region	a	x	_merge
1	2	26	13	3
2	1	29	15	3
3	2	22	13	3
4	3	21	12	3
5	1	24	15	3
6	5	20	.	1
.	4	.	11	2

8. Trợ giúp

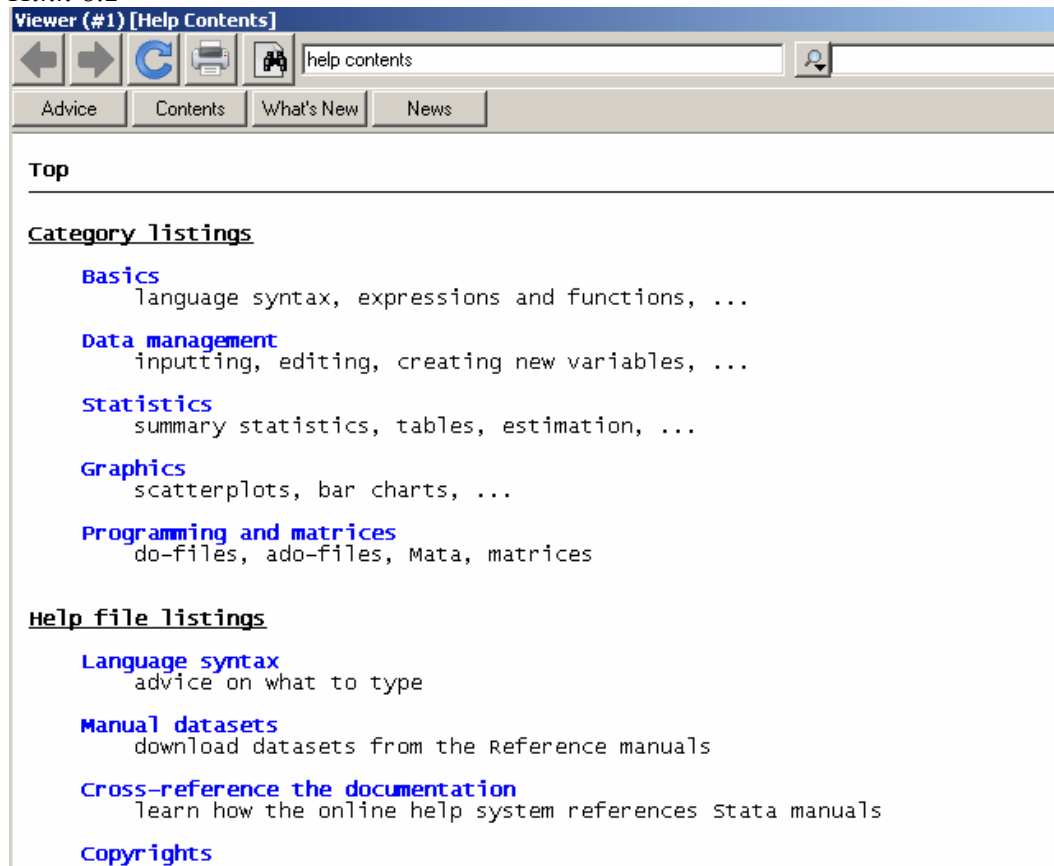
- Stata online: <http://www.ats.ucla.edu/stat/stata/> và rất nhiều trang khác!

Hình 8.1



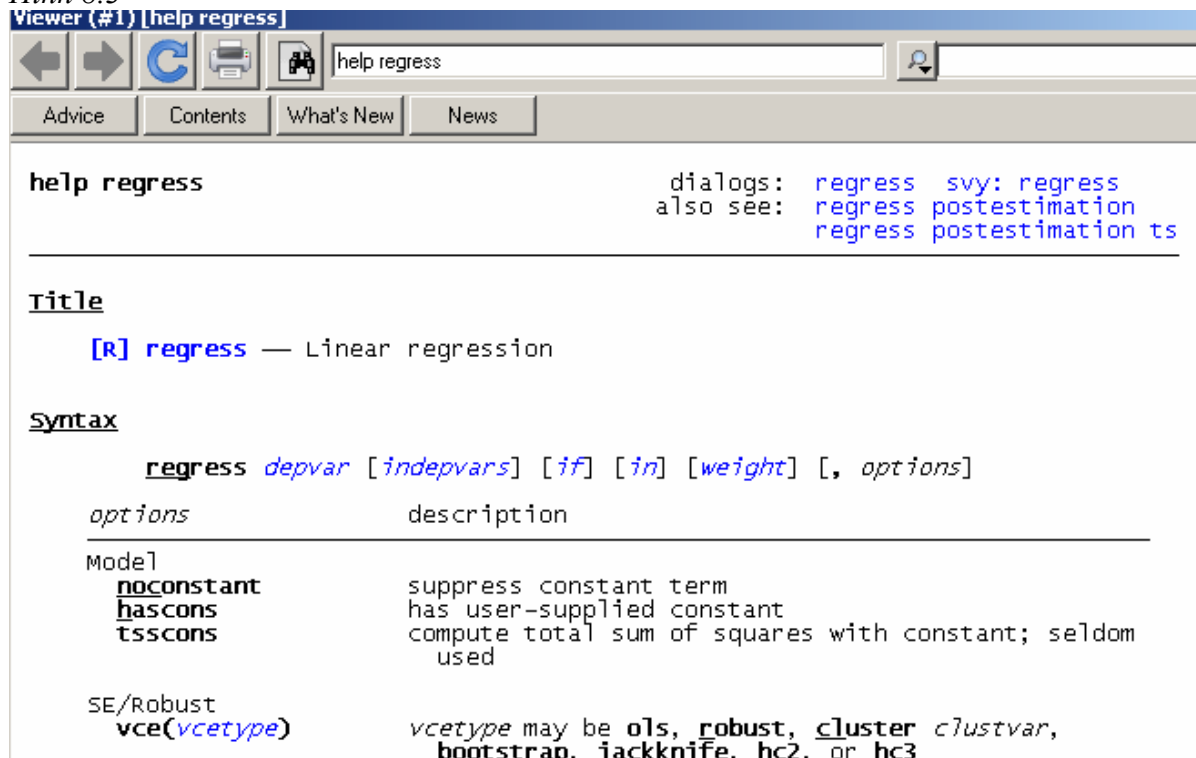
- Thư viện chương trình FETP
- Bạn có thể vào Mục Help\Contents của Stata để học tìm hiểu thêm về stata.

Hình 8.2



- Có thể tra cứu từng câu lệnh bằng cách **Help\Command**

Hình 8.3



- Các Sách, tài liệu, bài giảng ... mà giảng viên giới thiệu bạn
- Trao đổi với các chuyên gia trên diễn đàn thông tin phát triển Việt Nam: <http://divietnam.org>
- Và Google!

Phụ lục 1.
Hồi quy – những câu lệnh cơ bản trên Stata

Sử dụng file *muc1234a_hhexpe08.dta* mà bạn đã tạo ra để thực hiện các công việc (nối các file *muc123a.dta*, *muc4a.dta* và *hhexpe08.dta*)

1. Chuẩn bị dữ liệu

```
. egen chigd=rowtotal( m2ac13k m2ac16)

. gen gioi= m1ac2

. replace gioi=0 if m1ac2==2
(17952 real changes made)

. gen tuoi= m1ac5

. gen tuoi2= tuoi^2

. gen thanhthi= urban08

. recode thanhthi 2 =0
(thanhthi: 26301 changes made)

. tab reg8, gen(vung)
```

reg8	Freq.	Percent	Cum.
1	6,812	19.38	19.38
2	5,036	14.33	33.70
3	1,891	5.38	39.08
4	3,802	10.82	49.90
5	3,304	9.40	59.30
6	2,500	7.11	66.41
7	4,714	13.41	79.82
8	7,095	20.18	100.00
Total	35,154	100.00	

Câu 9. Thực hiện hàm hồi quy

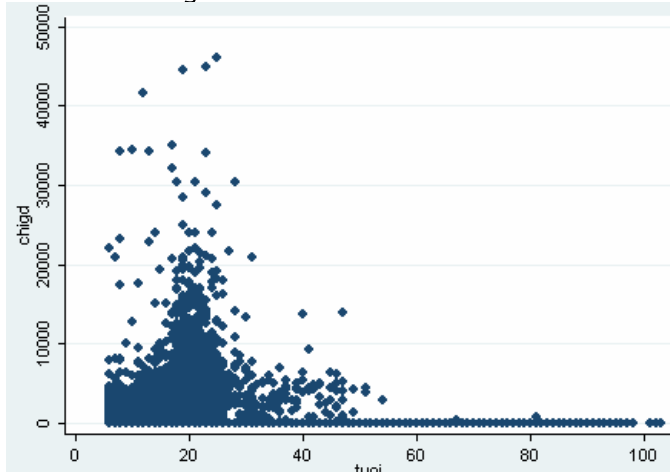
2. Tính hệ số tương quan

```
. pwcorr chigd tuoi hhsize, sig
```

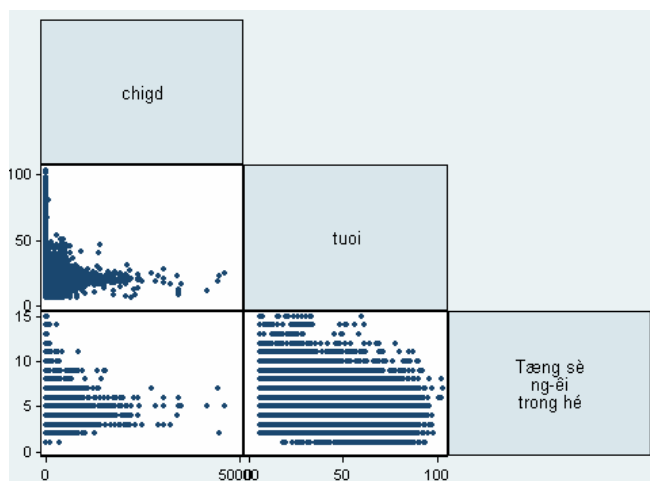
	chigd	tuoi	hhsize
chigd	1.0000		
tuoi	-0.2393	1.0000	
hhsize	-0.0177	-0.1926	1.0000

3. Vẽ đồ thị Scatter

scatter chigd tuoi



graph matrix chigd tuoi hhsiz, half



4. Ước lượng hàm hồi quy

```
reg chigd gioi tuoi bp thanhthi vung1 vung2 vung3 vung4 vung5 vung7 vung8
hhsiz if m2ac5<= 2
```

Mô hình 1

Source	SS	df	MS	Number of obs		
Model	2.0005e+10	12	1.6671e+09	F(12, 9006)	= 282.73	
Residual	5.3103e+10	9006	5896435.89	Prob > F	= 0.0000	
Total	7.3109e+10	9018	8106973.15	R-squared	= 0.2736	
				Adj R-squared	= 0.2727	
				Root MSE	= 2428.3	

chigd	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
gioi	52.54772	51.30177	1.02	0.306	-48.01542	153.1109
tuoi	347.9577	15.20595	22.88	0.000	318.1506	377.7648
tuoi ²	-3.115513	.4194744	-7.43	0.000	-3.937779	-2.293248
thanhthi	1081.325	61.32333	17.63	0.000	961.1169	1201.532
vung1	163.8409	104.2717	1.57	0.116	-40.55528	368.237
vung2	-273.0126	110.2669	-2.48	0.013	-489.1608	-56.86438
vung3	-655.3806	137.272	-4.77	0.000	-924.4648	-386.2963
vung4	-30.02082	110.9621	-0.27	0.787	-247.5317	187.49
vung5	77.38069	115.4724	0.67	0.503	-148.9715	303.7329
vung7	958.442	111.6069	8.59	0.000	739.6671	1177.217

```

vung8 |    21.29995    107.4682    0.20    0.843    -189.3622    231.9621
hhszise |   -76.84445    17.67135   -4.35    0.000   -111.4843   -42.20459
_cons |  -2364.973    185.9853   -12.72    0.000   -2729.547   -2000.4

```

4. Kiểm định Wald

Có người cho rằng quy mô hộ (hhszise) và thanhthi đều không ảnh hưởng đến chigd. Theo bạn, điều đó là đúng hay sai

```
. test thanhthi= hhszise=0
```

```

( 1)  thanhthi - hhszise = 0
( 2)  thanhthi = 0

```

```

F( 2, 9006) = 174.06
Prob > F = 0.0000

```

5. Kiểm định hiện tượng đa cộng tuyến

Bạn hãy kiểm định xem mô hình 1 có bị vi phạm hiện tượng đa cộng tuyến?

Sau khi ước lượng mô hình, bạn hãy gõ lệnh VIF

```
. vif
```

Variable	VIF	1/VIF
tuoi	9.20	0.108700
tuoi1p	9.18	0.108991
vung1	2.62	0.382354
vung8	2.30	0.435498
vung2	2.19	0.456844
vung4	2.13	0.468623
vung7	2.13	0.469353
vung5	1.97	0.508528
vung3	1.53	0.653950
hhszise	1.08	0.926313
thanhthi	1.07	0.931843
gioi	1.01	0.994199
Mean VIF	3.03	

6. Kiểm định hiện tượng phương sai thay đổi

Hãy kiểm định xem mô hình 1 có bị vi phạm hiện tượng phương sai thay đổi?

```
. hettest
```

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity

Ho: Constant variance

Variables: fitted values of chigd

```

chi2(1)      = 6006.98
Prob > chi2   = 0.0000

```

```
. imtest
```

Cameron & Trivedi's decomposition of IM-test

Source	chi2	df	p
Heteroskedasticity	393.87	59	0.0000
Skewness	104.13	12	0.0000
Kurtosis	13.30	1	0.0003
Total	511.31	72	0.0000

7. Sử dụng option robust sau lệnh reg để khắc phục hiện tượng phương sai thay đổi

Hãy ước lượng lại mô hình 1 mà có thể khắc phục hiện tượng phương sai thay đổi

```
reg chigd gioi tuoi tuoibp thanhthi vung1 vung2 vung3 vung4 vung5 vung7 vung8
hhsz if m2ac5<= 2, robust
```

Linear regression

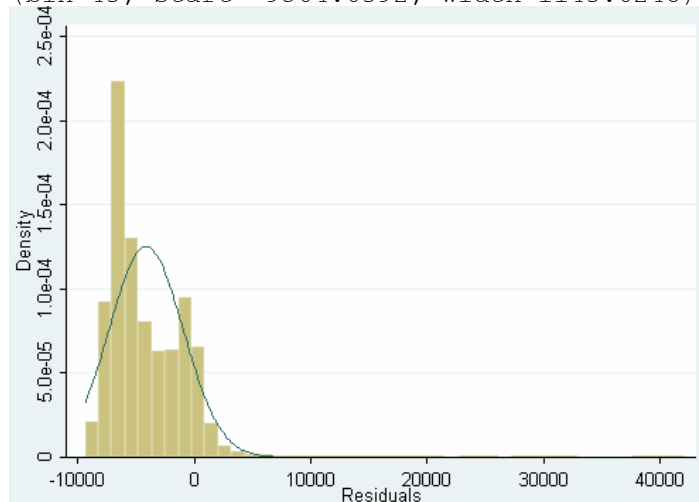
Number of obs = 9019
F(12, 9006) = 185.08
Prob > F = 0.0000
R-squared = 0.2736
Root MSE = 2428.3

chigd	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval]	
gioi	52.54772	51.17439	1.03	0.305	-47.76572	152.8612
tuoi	347.9577	21.66313	16.06	0.000	305.493	390.4224
tuoibp	-3.115513	.682129	-4.57	0.000	-4.452641	-1.778385
thanhthi	1081.325	74.50076	14.51	0.000	935.2862	1227.363
vung1	163.8409	87.9385	1.86	0.062	-8.53859	336.2203
vung2	-273.0126	87.90441	-3.11	0.002	-445.3252	-100.7
vung3	-655.3806	92.42965	-7.09	0.000	-836.5637	-474.1974
vung4	-30.02082	89.04394	-0.34	0.736	-204.5672	144.5256
vung5	77.38069	100.2281	0.77	0.440	-119.0891	273.8505
vung7	958.442	143.1455	6.70	0.000	677.8443	1239.04
vung8	21.29995	102.6027	0.21	0.836	-179.8247	222.4246
hhsz	-76.84445	15.91004	-4.83	0.000	-108.0317	-45.65715
_cons	-2364.973	202.2426	-11.69	0.000	-2761.415	-1968.532

8. Lưu lại phần dư của mô hình và kiểm định tính phân phối chuẩn của sai số

Lưu lại phần dư trong biến r, vẽ đồ thị phân phối của phần dư, tính thống kê skewness và kurtosis cho biến r

```
. predict r, resid
. histogram r, normal
(bin=45, start=-9364.6592, width=1145.6248)
```



```
. tabstat r, stat(skewness kurtosis)
```

variable	skewness	kurtosis
r	1.68251	13.18075

. sktest r

```

Skewness/Kurtosis tests for Normality
----- joint -----
Variable |      Obs   Pr(Skewness)   Pr(Kurtosis)   adj chi2(2)   Prob>chi2
-----+-----
r |    3.5e+04    0.0000        0.0000          .          .

```

. swilk r

```

Shapiro-Wilk W test for normal data
-----
Variable |      Obs      W      V      z      Prob>z
-----+-----
r |    35154    0.88757  1583.543  20.303  0.00000

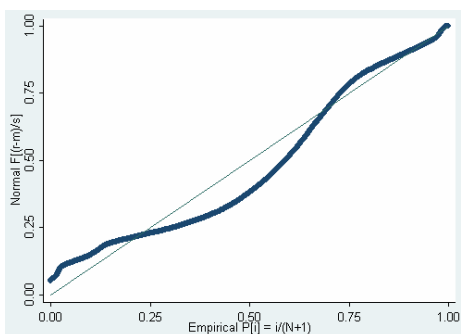
```

Ghi chú: Sau khi thực hiện hồi quy, gõ lệnh predict tên_biến kết hợp với các option sau có thể tạo ra những biến mới liên quan đến mô hình:

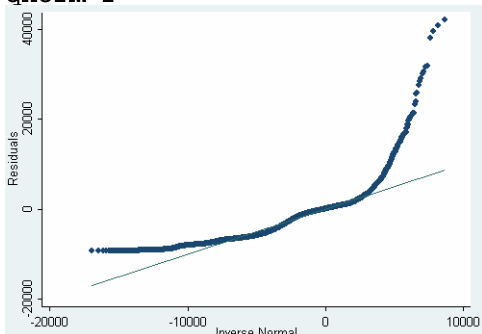
Để tạo ra	Thêm option sau lệnh predict
Giá trị dự báo của Y	Không cần option
Phần dư	resid
Phần dư chuẩn hoá	rstandard
Phần dư student hoá	Rstudent
Leverage	Lev hoặc hat
Sai số chuẩn của phần dư	Std_r
Cook's D	Cooksd
Sai số chuẩn của giá trị dự báo (cá biệt)	Std_f
Sai số chuẩn của giá trị dự báo (trung bình)	stdp

Có thể sử dụng đồ thị p, đồ thị q?

pnorm r



qnorm r



9. Ước lượng mô hình hồi quy – có sử dụng trọng số trong điều tra VHLSS

Hãy ước lượng lại mô hình trên, và chú ý đến vấn đề trọng số trong VHLSS

```
reg chigd gioi tuoi tuoi1p thanhthi vung1 vung2 vung3 vung4 vung5 vung7 vung8
hhszwt if m2ac5<=2 [pw= hhszwt], robust
(sum of wgt is 9.8756e+07)
```

Linear regression

Number of obs = 9019
F(12, 9006) = 122.22
Prob > F = 0.0000
R-squared = 0.2497
Root MSE = 2651.2

chigd	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval]	
gioi	35.68653	72.54715	0.49	0.623	-106.5224	177.8955
tuoi	341.5655	29.01901	11.77	0.000	284.6816	398.4493
tuoi1p	-3.148191	.888906	-3.54	0.000	-4.890649	-1.405733
thanhthi	1245.069	101.016	12.33	0.000	1047.055	1443.084
vung1	283.2034	90.92591	3.11	0.002	104.968	461.4389
vung2	-202.6547	86.93577	-2.33	0.020	-373.0686	-32.24081
vung3	-488.0537	97.7098	-4.99	0.000	-679.5871	-296.5203
vung4	132.6486	100.3422	1.32	0.186	-64.04499	329.3422
vung5	83.13013	94.97234	0.88	0.381	-103.0373	269.2975
vung7	1290.005	181.5838	7.10	0.000	934.0593	1645.95
vung8	118.2045	100.7079	1.17	0.241	-79.20587	315.6148
hhszwt	-80.93744	20.38689	-3.97	0.000	-120.9004	-40.9745
_cons	-2349.18	283.6993	-8.28	0.000	-2905.295	-1793.065

Bình thường, trong dữ liệu khảo sát thu nhập chi tiêu, đối với file dữ liệu cấp hộ, bạn sử dụng biến wt9 làm trọng số; với file cá nhân (thành viên), sử dụng biến hhszwt làm trọng số. (Trong các lệnh sum, tab ... khi muốn ước lượng trung bình, tỷ lệ tổng thể bạn sử dụng aw thay cho pw)

Những option mà thường sử dụng trong lệnh reg là

Noconstant : không ước lượng hằng số trong mô hình
level(#) : ước lượng khoảng tin cậy của hệ số hồi quy ở độ tin cậy #, mặc định là level(95)
Beta : tính toán thêm hệ số hồi quy đã chuẩn hoá

Bạn có thể sử dụng tiếp đầu ngữ trong câu lệnh?

stepwise [, options] : command

options	description
Model	
* pr(#)	significance level for removal from the model
* pe(#)	significance level for addition to the model


```
stepwise, pe(0.2): reg chigd gioi tuoi tuoibp thanhthi vung1 vung2 vung3 vung4 vung5 vung7
vung8 hhszwt if m2ac5<=2 [pw= hhszwt], robust
```

```
. stepwise, pe(0.2): reg chigd gioi tuoi tuoibp thanhthi vung1 vung2 vung3 vung4 vung5
vung7 vung8
```

```
> hhszwt if m2ac5<=2 [pw= hhszwt], robust
begin with empty model
```

```
p = 0.0000 < 0.2000 adding tuoi
p = 0.0000 < 0.2000 adding vung3
p = 0.0000 < 0.2000 adding thanhthi
p = 0.0000 < 0.2000 adding vung2
p = 0.0000 < 0.2000 adding vung7
p = 0.0000 < 0.2000 adding hhszwt
p = 0.0004 < 0.2000 adding tuoibp
p = 0.0032 < 0.2000 adding vung1
```

Linear regression

```
Number of obs = 9019
F( 8, 9010) = 164.56
Prob > F = 0.0000
R-squared = 0.2496
Root MSE = 2650.9
```

chigd	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval]	
tuoi	341.5594	29.06481	11.75	0.000	284.5857	398.533
vung3	-577.638	80.88375	-7.14	0.000	-736.1886	-419.0875
thanhthi	1239.731	100.5276	12.33	0.000	1042.675	1436.788
vung2	-298.6968	60.32118	-4.95	0.000	-416.94	-180.4536
vung7	1196.668	168.8873	7.09	0.000	865.6102	1527.725
hhszwt	-84.05998	20.19984	-4.16	0.000	-123.6562	-44.4637
tuoibp	-3.14264	.8913302	-3.53	0.000	-4.88985	-1.39543
vung1	186.0523	63.1691	2.95	0.003	62.2265	309.8781
_cons	-2219.553	269.1749	-8.25	0.000	-2747.197	-1691.909

Phụ lục 2

Bài tập về Quản lý dữ liệu – một số lệnh hay của Stata?

Câu 1. reshape wide

Với bảng hỏi: Mục 4B5.1 Thu thủy sản (*File Muc04_1B.xls-sheet m4b5(1)*) và File dữ liệu: muc4b51.dta

- bạn hãy tính trung bình số tiền thu được từ nuôi trồng và đánh bắt tôm (Chỉ tính cho những hộ có nuôi trồng hoặc đánh bắt tôm).
- Theo bạn, tỉnh nào có nhiều hộ nuôi trồng, đánh bắt tôm nhất? Và số lượng hộ nuôi trồng, đánh bắt tôm nhiều nhất ấy là bao nhiêu?

Gợi ý: dùng lệnh reshape wide

```
/* Bai tap quan ly du lieu - nang cao*/
*Cau 1
set mem 300m
use "C:\VHLSS2008\Data\Hhold\muc4b51.dta", clear
count
tab m4b51ma
tab m4b51ma, nol mis
sum m4b51c6b
keep tinh huyen xa diaban hoso m4b51ma m4b51c6b
rename m4b51c6b thuthuysan

tab m4b51ma, mis
recode m4b51ma . =4
reshape wide thuthuysan, i( tinh huyen xa diaban hoso) j( m4b51ma)
count

sum thuthuysan3 thuthuysan11 thuthuysan12 thuthuysan13 thuthuysan14 thuthuysan21
thuthuysan22 thuthuysan23 thuthuysan4

egen thutom=rowtotal( thuthuysan12 thuthuysan22) if ( thuthuysan12!=.)|(
thuthuysan22!=.)

sum thutom

gen cotom=1 if ( thuthuysan12!=.)|( thuthuysan22!=.)
recode cotom . =0

tab tinh cotom

save "C:\VHLSS2008\muc4b51_fileho_thuthuysan.dta", replace
```

Câu 2. egen

a. Từ file muc123a.dta bạn hãy tạo một biến cho biết thành viên trong hộ này có mấy người.

```
. use "C:\VHLSS2008\Data\Hhold\muc123a.dta", clear
. gen tam=1
. egen songuoi=sum(tam), by ( tinh huyen xa diaban hoso)
. edit
. list tinh huyen xa diaban hoso matv m1ac2 m1ac3 m1ac5 tam songuoi
in f/20
```

	tinh	huyen	xa	diaban	hoso	matv	m1ac2	m1ac3	m1ac5	tam	songuoi
1.	101	1	3	1	13	1	Nữ	Chủ hộ	73	1	3
2.	101	1	3	1	13	2	Nữ	Con	39	1	3
3.	101	1	3	1	13	3	Nữ	Con	33	1	3
4.	101	1	3	1	14	1	Nam	Chủ hộ	64	1	3
5.	101	1	3	1	14	2	Nữ	Vợ chồng	53	1	3
6.	101	1	3	1	14	3	Nam	Con	22	1	3
7.	101	1	3	1	15	1	Nam	Chủ hộ	61	1	2
8.	101	1	3	1	15	2	Nữ	Vợ chồng	60	1	2
9.	101	1	9	19	15	1	Nam	Chủ hộ	50	1	2
10.	101	1	9	19	15	2	Nữ	Vợ chồng	41	1	2
11.	101	1	9	19	19	1	Nữ	Chủ hộ	64	1	3
12.	101	1	9	19	19	2	Nam	Vợ chồng	61	1	3
13.	101	1	9	19	19	3	Nữ	Con	23	1	3
14.	101	1	9	19	20	1	Nam	Chủ hộ	50	1	3
15.	101	1	9	19	20	2	Nữ	Vợ chồng	51	1	3
16.	101	1	9	19	20	3	Nữ	Con	19	1	3
17.	101	1	15	50	13	1	Nam	Chủ hộ	35	1	4
18.	101	1	15	50	13	2	Nữ	Vợ chồng	34	1	4
19.	101	1	15	50	13	3	Nam	Con	6	1	4
20.	101	1	15	50	13	4	Nam	Con	6	1	4

b. Từ file muc123a.dta bạn hãy tạo một biến cho biết học vấn cao nhất của những người trong hộ

```
. egen hvmax =max( m2ac1) if m2ac1!=-1, by( tinh huyen xa diaban hoso)
. list tinh huyen xa diaban hoso matv m1ac3 m1ac5 m2ac1 hvmax in f/20
```

	tinh	huyen	xa	diaban	hoso	matv	m1ac3	m1ac5	m2ac1	hvmax
1.	101	1	3	1	13	1	Chủ hộ	73	4	12
2.	101	1	3	1	13	2	Con	39	TN THPT	12
3.	101	1	3	1	13	3	Con	33	TN THPT	12
4.	101	1	3	1	14	1	Chủ hộ	64	TN THPT	12
5.	101	1	3	1	14	2	Vợ chồng	53	TN THPT	12
6.	101	1	3	1	14	3	Con	22	TN THPT	12
7.	101	1	3	1	15	1	Chủ hộ	61	TN THPT	12
8.	101	1	3	1	15	2	Vợ chồng	60	TN THPT	12
9.	101	1	9	19	15	1	Chủ hộ	50	TN THPT	12
10.	101	1	9	19	15	2	Vợ chồng	41	TN THPT	12
11.	101	1	9	19	19	1	Chủ hộ	64	9	12
12.	101	1	9	19	19	2	Vợ chồng	61	TN THPT	12
13.	101	1	9	19	19	3	Con	23	TN THPT	12
14.	101	1	9	19	20	1	Chủ hộ	50	TN THPT	12
15.	101	1	9	19	20	2	Vợ chồng	51	TN THPT	12
16.	101	1	9	19	20	3	Con	19	TN THPT	12
17.	101	1	15	50	13	1	Chủ hộ	35	TN THPT	12
18.	101	1	15	50	13	2	Vợ chồng	34	TN THPT	12
19.	101	1	15	50	13	3	Con	6	Chưa hết lớp 1/chưa đi học	12
20.	101	1	15	50	13	4	Con	6	Chưa hết lớp 1/chưa đi học	12

Câu 3. egen – keep/drop và collapse

Từ file muc123a.dta là file cá nhân. Bạn hãy rút gọn file này ở cấp độ hộ. Và có 1 biến cho biết hộ có mấy người

- Cách 1. Làm như câu 2. Sau đó chỉ giữ lại người nào là chủ hộ
- Cách 2.

```
. gen quymoho=1
. collapse (sum) quymoho, by ( tinh huyen xa diaban hoso)
. list tinh huyen xa diaban hoso quymoho in f/10
```

	tinh	huyen	xa	diaban	hoso	quymoho
1.	101	1	3	1	13	3
2.	101	1	3	1	14	3
3.	101	1	3	1	15	2
4.	101	1	9	19	15	2
5.	101	1	9	19	19	3
6.	101	1	9	19	20	3
7.	101	1	15	50	13	4
8.	101	1	15	50	14	3
9.	101	1	15	50	15	4
10.	101	1	17	2	13	2

Câu 4. collapse

Bạn hãy dùng lệnh collapse để thu gọn lại dữ liệu muc123a.dta theo cấp hộ. Trong file hộ này, có biến cho biết tổng chi tiêu cho giáo dục của hộ, có biến đếm được số người hiện đang đi học hay đang nghỉ nghề của hộ, có biến cho biết học vấn cao nhất của người trong hộ, có biến cho biết tuổi trung bình của các thành viên trong hộ.

```
. egen chigd=rowtotal( m2ac13k m2ac16)
```

```
. list tỉnh huyện xa diaban hoso matv m2ac5 m2ac13k m2ac16 chigd in 1/40
```

	tỉnh	huyện	xa	diaban	hoso	matv	m2ac5	m2ac13k	m2ac16	chigd
1.	101	1	3	1	13	1	Không	.	0	0
2.	101	1	3	1	13	2	Không	.	0	0
3.	101	1	3	1	13	3	Không	.	0	0
4.	101	1	3	1	14	1	Không	.	0	0
5.	101	1	3	1	14	2	Không	.	0	0
6.	101	1	3	1	14	3	Nghỉ hè	2300	6000	8300
7.	101	1	3	1	15	1	Không	.	0	0
8.	101	1	3	1	15	2	Không	.	0	0
9.	101	1	9	19	15	1	Không	.	0	0
10.	101	1	9	19	15	2	Không	.	0	0
11.	101	1	9	19	19	1	Không	.	0	0
12.	101	1	9	19	19	2	Không	.	0	0
13.	101	1	9	19	19	3	Không	.	0	0
14.	101	1	9	19	20	1	Không	.	0	0
15.	101	1	9	19	20	2	Không	.	0	0
16.	101	1	9	19	20	3	Nghỉ hè	2370	0	2370
17.	101	1	15	50	13	1	Không	.	0	0
18.	101	1	15	50	13	2	Không	.	0	0
19.	101	1	15	50	13	3	Nghỉ hè	1900	0	1900
20.	101	1	15	50	13	4	Nghỉ hè	1900	0	1900
21.	101	1	15	50	14	1	Không	.	0	0
22.	101	1	15	50	14	2	Không	.	0	0
23.	101	1	15	50	14	3	Không	.	0	0
24.	101	1	15	50	15	1	Không	.	0	0
25.	101	1	15	50	15	2	Không	.	0	0
26.	101	1	15	50	15	3	Nghỉ hè	2900	3200	6100
27.	101	1	15	50	15	4	Nghỉ hè	2050	3000	5050
28.	101	1	17	2	13	1	Không	.	0	0
29.	101	1	17	2	13	2	Không	.	0	0
30.	101	1	17	2	14	1	Không	.	0	0
31.	101	1	17	2	14	2	Không	.	0	0
32.	101	1	17	2	14	3	Có	3474	0	3474
33.	101	1	17	2	19	1	Không	.	0	0
34.	101	1	17	2	19	2	Không	.	0	0
35.	101	1	21	36	13	1	Không	.	0	0
36.	101	1	21	36	13	2	Không	.	0	0
37.	101	1	21	36	13	3	Có	3000	0	3000
38.	101	1	21	36	15	1	Không	.	0	0
39.	101	1	21	36	15	2	Không	.	0	0
40.	101	1	21	36	15	3	Có	3050	0	3050

```
. collapse (mean) m1ac5 (sum) dihoc_nghihe (sum) chigd (max) m2ac1, by (
tinh huyen xa diaban hoso)
```

```
. list tinh huyen xa diaban hoso m1ac5 dihoc_nghihe chigd m2ac1 in 1/20
```

	tinh	huyen	xa	diaban	hoso	m1ac5	dihoc_~e	chigd	m2ac1
1.	101	1	3	1	13	48.3333	0	0	12
2.	101	1	3	1	14	46.3333	1	8300	12
3.	101	1	3	1	15	60.5	0	0	12
4.	101	1	9	19	15	45.5	0	0	12
5.	101	1	9	19	19	49.3333	0	0	12
6.	101	1	9	19	20	40	1	2370	12
7.	101	1	15	50	13	20.25	2	3800	12
8.	101	1	15	50	14	18.3333	0	0	12
9.	101	1	15	50	15	35	2	11150	12
10.	101	1	17	2	13	33.5	0	0	12
11.	101	1	17	2	14	43.3333	1	3474	12
12.	101	1	17	2	19	47	0	0	12
13.	101	1	21	36	13	27.3333	1	3000	11
14.	101	1	21	36	15	43.3333	1	3050	12
15.	101	1	21	36	19	39.2	1	8220	12
16.	101	1	23	18	13	49.5	0	0	12
17.	101	1	23	18	24	29.5	0	0	12
18.	101	1	23	18	29	20	0	0	12
19.	101	3	3	23	13	28.75	2	8350	12
20.	101	3	3	23	15	24.25	1	4950	12

Câu 5. append using

Bạn hãy tạo một dữ liệu gộp ở file hhexpe06.dta và file hhexpe08.dta

```
. use "C:\VHLSS2008\Data\Hhold\hhexpe08.dta", clear
(Household expenditures: 2008 VHLSS)

. keep tinh huyen xa diaban hoso hhszsize hhexplrl wt9 urban08 reg8

. save "C:\VHLSS2008\hhexpe08_sua.dta"
file C:\VHLSS2008\hhexpe08_sua.dta saved

. gen nam=2008

. count
9189

. save "C:\VHLSS2008\hhexpe08_sua.dta", replace
file C:\VHLSS2008\hhexpe08_sua.dta saved

. use "C:\VHLSS2006\VHLSS2006\Data\hhold\hhexpe06.dta", clear
(Household expenditures: 2006 VHLSS)

. keep tinh huyen xa diaban hoso reg8 urban06 hhszsize wt9 hhexplrl

. sort tinh huyen xa diaban hoso

. gen nam=2006

. save "C:\VHLSS2008\hhexpe06_sua.dta"

. count
9189
```

```
. save "C:\VHLSS2008\hhexpe06_sua.dta", replace
file C:\VHLSS2008\hhexpe06_sua.dta saved

. use "C:\VHLSS2008\hhexpe08_sua.dta", clear
(Household expenditures: 2008 VHLSS)

. append using "C:\VHLSS2008\hhexpe06_sua.dta"

. count
18378

.. list tỉnh huyện xa diaban hoso reg8 hhsiz hhexplrl nam urban08 urban06 in
9180/9210

. list tỉnh huyện xa diaban hoso reg8 hhsiz hhexplrl nam urban08 urban06 in
9180/9210, nol
```

	tỉnh	huyện	xa	diaban	hoso	reg8	hhsiz	hhexplrl	nam	urban08	urban06
9180.	823	13	6	29	19	8	4	37240.17	2008	2	.
9181.	823	13	11	4	13	8	3	47041.98	2008	2	.
9182.	823	13	11	4	14	8	4	26782.1	2008	2	.
9183.	823	13	11	4	15	8	6	50781.51	2008	2	.
9184.	823	13	12	25	13	8	5	13658.05	2008	2	.
9185.	823	13	12	25	15	8	6	17925.5	2008	2	.
9186.	823	13	12	25	19	8	4	19404.83	2008	2	.
9187.	823	13	17	1	13	8	3	8687.102	2008	2	.
9188.	823	13	17	1	14	8	5	18933.47	2008	2	.
9189.	823	13	17	1	20	8	4	19210.81	2008	2	.
9190.	101	1	3	14	15	1	4	50885.41	2006	.	1
9191.	101	1	3	14	19	1	4	70989.08	2006	.	1
9192.	101	1	3	14	24	1	4	47160.71	2006	.	1
9193.	101	1	9	19	13	1	2	22773.67	2006	.	1
9194.	101	1	9	19	15	1	3	34226.87	2006	.	1
9195.	101	1	9	19	19	1	3	32712.68	2006	.	1
9196.	101	1	15	27	13	1	3	15950.77	2006	.	1
9197.	101	1	15	27	14	1	1	72100.53	2006	.	1
9198.	101	1	15	27	15	1	4	30387.58	2006	.	1
9199.	101	1	17	2	13	1	3	23695.05	2006	.	1
9200.	101	1	17	2	14	1	3	34148.03	2006	.	1
9201.	101	1	17	2	19	1	3	26933.93	2006	.	1
9202.	101	1	21	24	13	1	4	117633.9	2006	.	1
9203.	101	1	21	24	14	1	2	23580.65	2006	.	1
9204.	101	1	21	24	19	1	4	82198.54	2006	.	1
9205.	101	1	23	18	13	1	4	52940.96	2006	.	1
9206.	101	1	23	18	24	1	4	45195.64	2006	.	1
9207.	101	1	23	18	25	1	3	39613.94	2006	.	1
9208.	101	3	3	14	13	1	4	85210.34	2006	.	1
9209.	101	3	3	14	14	1	5	41575.87	2006	.	1
9210.	101	3	3	14	15	1	3	29575.11	2006	.	1

Câu 6. Lệnh Merge

Bạn đã từng Merge 2 file cấp hộ – hộ, cá nhân – cá nhân, cá nhân – hộ. Bây giờ bạn thử suy nghĩ xem ghép 1 file cấp xã/phường vào file cấp hộ thì sẽ ra sao? Trong nhiều biến, những hộ ở cùng một xã/phường sẽ có cùng những thông tin của xã/phường phải không?

Phụ lục 3. Mô hình Logit**1. Dữ liệu**

Dữ liệu được trích từ VHLSS2006 về các thành viên trong hộ từ 15 tuổi trở lên. Trong dữ liệu có một số biến sau:

Tên biến	Giải thích	Ghi chú
covieclam	Có việc làm trong 12 tháng qua	1: Có 0: Không
thanhthi	Thành thị/nông thôn	1: Thành thị 0: Nông thôn
Daihoc, caodang, thpt, thes, tieuhoc	Các biến giả phản ánh bằng cấp cao nhất đã đạt được	- Các biến này được mã hoá lại từ biến m2ac3a - Biến giả tham chiếu là không có bằng cấp
kinhhoa	Dân tộc kinh, hoac	1: Dân tộc Kinh, hoặc Hoa 0: Dân tộc khác
quymoho	Quy mô hộ	Số người trong hộ (người)
Tuoi	Tuổi	
thunhapbq	Thu nhập bình quân/người/hộ/tháng	Ngàn đ
Kv1v5	Khu vực làm việc 5	1: Tự làm cho gia đình 2: Làm cho hộ khác 3: Kinh tế nhà nước – tập thể 4: Kinh tế tư nhân 5: Kinh tế có vốn đầu tư nước ngoài
Kv1v3	Khu vực làm việc 3	1: Tự làm cho gia đình, làm cho hộ khác, hoặc kinh tế tư nhân 2: Kinh tế nhà nước – tập thể 3: Kinh tế có vốn đầu tư nước ngoài
Vung1, Vung2, ... Vung8	Các biến giả mã hoá các vùng trong cả nước	Các biến này được mã hoá lại từ biến Reg8
bac, trung, nam	Các biến giả phản thể hiện miền Bắc, miền Trung&tây nguyên, miền nam	Các biến này được mã hoá lại từ biến Reg8
hhszwt	Trọng số cá nhân	

. count /* Tổng số quan sát*/
29360

Mô tả một số biến**. tab thanhthi**

thanh thi:			
1 thanh			
thi, 0 nong			
thon	Freq.	Percent	Cum.
0	21,846	74.41	74.41
1	7,514	25.59	100.00
Total	29,360	100.00	

. tab m2ac3a

3. Bằng cấp			
cao nhất -			
GDPT	Freq.	Percent	Cum.
K0 bằng cấp	188	0.83	0.83
Tiểu học	7,266	32.20	33.03
THCS	8,983	39.81	72.84
THPT	4,954	21.95	94.80
Cao đẳng	345	1.53	96.33
Đại học	801	3.55	99.88
Thạc sĩ	22	0.10	99.97
Tiến sĩ	6	0.03	100.00
Total	22,565	100.00	

. tab kinhhoa

kinhhwa	Freq.	Percent	Cum.
0	4,798	16.34	16.34
1	24,562	83.66	100.00
Total	29,360	100.00	

. tab vung

vung	Freq.	Percent	Cum.
1	5,866	19.98	19.98
2	4,300	14.65	34.63
3	1,468	5.00	39.63
4	3,143	10.71	50.33
5	2,734	9.31	59.64
6	1,869	6.37	66.01
7	3,912	13.32	79.33
8	6,068	20.67	100.00
Total	29,360	100.00	

. tab gioi

gioi 1.nam			
0 nu	Freq.	Percent	Cum.
0	15,121	51.50	51.50
1	14,239	48.50	100.00
Total	29,360	100.00	

. tab thanhthi

thanh thi:				
1 thanh				
thi, 0 nong				
thon	Freq.	Percent		Cum.
0	21,846	74.41		74.41
1	7,514	25.59		100.00
Total	29,360	100.00		

```
. mean tuoi quymoho thunhapbq
```

Mean estimation Number of obs = 29360

	Mean	Std. Err.	[95% Conf. Interval]	
tuoi	38.50232	.1031738	38.30009	38.70454
quymoho	4.798774	.0105579	4.77808	4.819468
thunhapbq	693.8959	4.09364	685.8722	701.9196

```
. tab kv1v5
```

kv1v5	Freq.	Percent	Cum.
1	15,350	68.63	68.63
2	3,280	14.67	83.30
3	2,390	10.69	93.99
4	978	4.37	98.36
5	367	1.64	100.00
Total	22,365	100.00	

```
. tab kv1v3
```

kvlv3	Freq.	Percent	Cum.
1	19,608	87.67	87.67
2	2,390	10.69	98.36
3	367	1.64	100.00
Total	22,365	100.00	

2. Ước lượng một số mô hình Logit

Mô hình 1. Các yếu tố ảnh hưởng đến xác suất có việc làm

a. Mô hình không tính đến trọng số

```
. logit covieclam gioi tuoi kinhhoa daihoc caodang thpt thcs tieuhoc if thanhthi==1
```

```
Iteration 0: log likelihood = -4681.615
Iteration 1: log likelihood = -4501.1785
Iteration 2: log likelihood = -4496.9115
Iteration 3: log likelihood = -4496.8878
Iteration 4: log likelihood = -4496.8878
```

```
Logistic regression               Number of obs   =       7514
                                LR chi2(8)          =       369.45
                                Prob > chi2          =       0.0000
                                Pseudo R2           =       0.0395
```

	covieclam	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
	gioi	.2749431	.0516756	5.32	0.000	.1736608 .3762254
	tuoi	.0016395	.0015832	1.04	0.300	-.0014636 .0047426
	kinhhwa	-.4514802	.1271611	-3.55	0.000	-.7007114 -.202249
	daihoc	1.647956	.1372616	12.01	0.000	1.378929 1.916984
	caodang	1.613607	.223183	7.23	0.000	1.176177 2.051038
	thpt	.3663942	.0857136	4.27	0.000	.1983986 .5343898
	thcs	.3866646	.0846816	4.57	0.000	.2206916 .5526376
	tieuhoc	1.064819	.0914975	11.64	0.000	.8854877 1.244151
	_cons	.4565806	.1548386	2.95	0.003	.1531025 .7600586

b. Mô hình có tính đến trọng số

```
. logit covieclam gioi tuoi kinhhoa daihoc caodang thpt thcs tieuhoc if thanhthi==1
> [pwhhszwt]
```

```
(sum of wgt is 8.1676e+07)
Iteration 0: log pseudolikelihood = -4726.9682
Iteration 1: log pseudolikelihood = -4496.5079
Iteration 2: log pseudolikelihood = -4490.0821
Iteration 3: log pseudolikelihood = -4490.022
Iteration 4: log pseudolikelihood = -4490.022
```

```
Logistic regression               Number of obs   =       7514
                                Wald chi2(8)        =       325.64
                                Prob > chi2         =       0.0000
                                Pseudo R2           =       0.0501
```

	covieclam	Coef.	Robust Std. Err.	z	P> z	[95% Conf. Interval]
	gioi	.3760212	.0597878	6.29	0.000	.2588393 .493203
	tuoi	-.0026357	.0020915	-1.26	0.208	-.006735 .0014636
	kinhhwa	-.5075088	.1535001	-3.31	0.001	-.8083635 -.206654
	daihoc	1.800663	.1590845	11.32	0.000	1.488863 2.112463
	caodang	1.869919	.2621885	7.13	0.000	1.356039 2.383799
	thpt	.35563	.0987501	3.60	0.000	.1620834 .5491766
	thcs	.3508384	.0968012	3.62	0.000	.1611115 .5405653
	tieuhoc	1.090695	.1079582	10.10	0.000	.8791004 1.302289
	_cons	.5925872	.1868671	3.17	0.002	.2263344 .95884

Mô hình 2. Các yếu tố ảnh hưởng đến xác suất có việc làm**a. Mô hình không tính đến trọng số**

```
. logit covieclam gioi thanhthi tuoi tuoibp kinhhoa daihoc caodang thpt thcs tieuhoc
> kinhhoa
```

note: kinhhoa dropped because of collinearity

Iteration 0: log likelihood = -16120.212

Iteration 1: log likelihood = -11270.986

Iteration 2: log likelihood = -10853.121

Iteration 3: log likelihood = -10829.292

Iteration 4: log likelihood = -10829.156

Iteration 5: log likelihood = -10829.156

```
Logistic regression               Number of obs   =       29360
                                LR chi2(10)        =      10582.11
                                Prob > chi2         =       0.0000
                                Pseudo R2           =       0.3282

Log likelihood = -10829.156
```

covieclam	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
gioi	.4357346	.0350008	12.45	0.000	.3671343 .504335
thanhthi	-.8299765	.039965	-20.77	0.000	-.9083065 -.7516464
tuoi	.4144831	.005585	74.21	0.000	.4035367 .4254296
tuoibp	-.0047818	.0000652	-73.29	0.000	-.0049097 -.0046539
kinhhwa	-.5583331	.0525236	-10.63	0.000	-.6612775 -.4553888
daihoc	.2838882	.1264527	2.25	0.025	.0360455 .5317309
caodang	.4035841	.187377	2.15	0.031	.036332 .7708362
thpt	-.7995127	.0608258	-13.14	0.000	-.9187292 -.6802963
thcs	-.3051377	.0561067	-5.44	0.000	-.4151048 -.1951706
tieuhoc	.4442288	.0599536	7.41	0.000	.3267219 .5617357
_cons	-5.143131	.1065803	-48.26	0.000	-5.352025 -4.934238

b. Mô hình có tính đến trọng số

```
. logit covieclam gioi thanhthi tuoi tuoibp kinhhoa daihoc caodang thpt thcs tieuhoc
> kinhhoa [pw=hhszwt]
```

(sum of wgt is 2.9721e+08)

note: kinhhoa dropped because of collinearity

Iteration 0: log pseudolikelihood = -16413.351

Iteration 1: log pseudolikelihood = -11501.346

Iteration 2: log pseudolikelihood = -11068.455

Iteration 3: log pseudolikelihood = -11042.538

Iteration 4: log pseudolikelihood = -11042.382

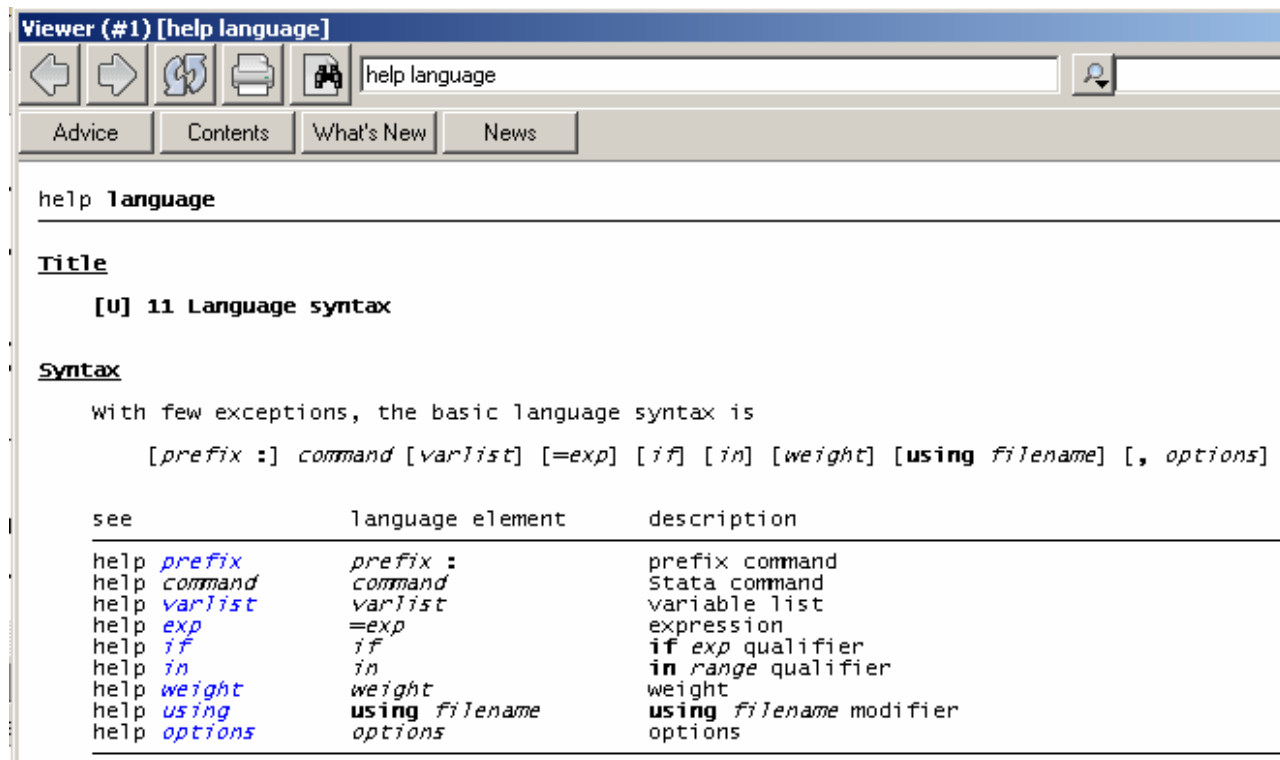
```
Logistic regression               Number of obs   =       29360
                                Wald chi2(10)        =      3963.25
                                Prob > chi2         =       0.0000
                                Pseudo R2           =       0.3272

Log pseudolikelihood = -11042.382
```

covieclam	Coef.	Robust Std. Err.	z	P> z	[95% Conf. Interval]
gioi	.4702713	.0386441	12.17	0.000	.3945302 .5460124
thanhthi	-.8711648	.0454963	-19.15	0.000	-.960336 -.7819937
tuoi	.422557	.0072647	58.17	0.000	.4083184 .4367956
tuoibp	-.0049484	.000089	-55.58	0.000	-.0051229 -.0047739
kinhhwa	-.5832457	.0610073	-9.56	0.000	-.7028177 -.4636737
daihoc	.5034056	.1468915	3.43	0.001	.2155035 .7913077
caodang	.6192097	.2236963	2.77	0.006	.1807731 1.057646
thpt	-.6663316	.073634	-9.05	0.000	-.8106516 -.5220116
thcs	-.2068623	.0676326	-3.06	0.002	-.3394197 -.0743049
tieuhoc	.5462198	.0774623	7.05	0.000	.3943964 .6980432
_cons	-5.298014	.1266096	-41.85	0.000	-5.546164 -5.049864

Phụ lục 4. Cấu trúc lệnh trong Stata, vấn đề trọng số trong phân tích dữ liệu VHLSS¹

Hình 4.1



[lệnh prefix:] **cú pháp lệnh** [danh sách biến] [biểu thức] [điều kiện] [phạm vi] [trọng số] [using tên file] [,tùy chọn]

Trong cấu trúc lệnh, nếu mục nào đặt trong 2 dấu ngoặc vuông [] tức là không bắt buộc phải có mục này

Có những hướng dẫn, có mục đặt trong dấu 2 dấu ngoặc nhọn <>, mục này bắt buộc phải có khi gõ lệnh.

- **Prefix:**

Một **lệnh prefix** mà bạn đã biết đến và thường sử dụng là **by**. Bạn còn nhớ không?

- **Command:** gõ lệnh mà bạn cần thực hiện. Một số lệnh stata cho phép viết tắt. Ví dụ, lệnh sum mà bạn đã sử dụng là viết tắt của lệnh summarize. Bạn cũng có thể gõ tắt lệnh này bằng chữ **su**

=exp (biểu thức)

Ví dụ, bạn cần tạo biến tuoi, và tuoibp. Biết rằng **gen** là lệnh để tạo một biến mới

```
. gen tuoi= m1ac5
```

```
. gen tuoibp= m1ac5^2
```

¹ Những kết quả ở phụ lục 4 được tính toán trên VHLSS 2006

help summarize

dialog: **sum****Title****[R] summarize** — Summary statistics**Syntax****summarize** [*varlist*] [*if*] [*in*] [*weight*] [, *options*]

<i>options</i>	description
Main	
detail	display additional statistics
meanonly	suppress the display; calculate only the mean; programmer's option
format	use variable's display format
separator(#)	draw separator line after every # variables; default is separator(5)

varlist may contain time-series operators; see **tsvarlist**.**by** is allowed; see **[D] by**.**aweight**s, **fweight**s, and **iweight**s are allowed. However, **iweight**s may not be used with the **detail** option; see **weight**.

- **Varlist** (danh sách biến): chỉ ra danh sách biến chịu tác động của câu lệnh. Nhưng nếu không có biến nào được chỉ ra thì lệnh Stata sẽ có tác động lên tất cả các biến.
- **If** (điều kiện)

Stata chỉ thực hiện câu lệnh đối với các quan sát mà có kết quả của biểu thức so sánh trong điều kiện if là đúng.

Ví dụ: đếm số người ở TPHCM; đếm số người ở Đà Nẵng và TPHCM và

```
. count if tinh==701
1257
```

```
. count if tinh==501 | tinh==701
1759
```

Ví dụ: Tạo bảng tần số cho biến loại trường học

```
. tab m2ac4 if urban06==1
```

4.Loại			
trường đã			
TN	Freq.	Percent	Cum.
-----+-----			
Công lập	6,919	95.49	95.49
Bán công	187	2.58	98.07
Dân lập	76	1.05	99.12
Tư thực	52	0.72	99.83
Khác	12	0.17	100.00
-----+-----			
Total	7,246	100.00	

- Chú ý rằng khi so sánh bằng, chúng ta sử dụng 2 dấu =, tức là == (sau lệnh if). Còn ở mục trên, khi tạo biến tuổi, trong phép gán, chúng ta gõ **gen tuoi= m1ac5**

- **using filename**

Trong lệnh merge, bạn đã từng sử dụng mục [using filename]

- **In range (phạm vi)**

Chỉ ra phạm vi các quan sát chịu tác động bởi câu lệnh²

. tab m4ac1a

1A. Làm nhận lương công			
	Freq.	Percent	Cum.
Có	9,447	26.11	26.11
Không	26,728	73.89	100.00
Total	36,175	100.00	

. tab m4ac1a in 100 /*tạo bảng tần số cho biến m4ac1a cho quan sát thứ 100, chính bằng giá trị của biến này tại quan sát thứ 100*/

1A. Làm nhận lương công			
	Freq.	Percent	Cum.
Không	1	100.00	100.00
Total	1	100.00	

. tab m4ac1a in 100/1000 /*tạo bảng tần số cho biến m4ac1a cho các quan sát từ thứ 100 đến 1000 */

1A. Làm nhận lương công			
	Freq.	Percent	Cum.
Có	281	34.10	34.10
Không	543	65.90	100.00
Total	824	100.00	

. tab m4ac1a in f/100 /*tạo bảng tần số cho biến m4ac1a cho các quan sát từ thứ 1 đến 100 */

1A. Làm nhận lương công			
	Freq.	Percent	Cum.
Có	42	45.16	45.16
Không	51	54.84	100.00
Total	93	100.00	

. tab m4ac1a in 100/1 /*tạo bảng tần số cho biến m4ac1a cho các quan sát từ thứ 100 đến quan sát cuối cùng */

1A. Làm nhận lương công			
	Freq.	Percent	Cum.
Có	9,405	26.06	26.06
Không	26,678	73.94	100.00
Total	36,083	100.00	

² Xem thêm Phụ lục 2

- **Weight (trọng số)**

Cho phép các phép phân tích có sử dụng đến trọng số (hay quyền số)

Khi phân tích VHLSS cần sử dụng trọng số nếu bạn muốn ước lượng các tham số thống kê cho tổng thể. Trong VHLSS2006, và tương tự ở VHLSS2008 có 2 biến lưu trọng số.

wt9: trọng số hộ (khi sử dụng dữ liệu mẫu khảo sát thu nhập và chi tiêu với cỡ mẫu 9189 hộ)

hhszwt: trọng số cá nhân

Hai biến trên có quan hệ như sau: $hhszwt = hhszwt * wt9$

Với hhszwt là tổng số người trong hộ

Hình 4.2

	wt9	hhszwt	hhszwt	m
1	3107.318	12429.27	4	
2	3107.318	12429.27	4	
3	3107.318	12429.27	4	
4	3107.318	12429.27	4	
5	3107.318	12429.27	4	
6	3107.318	12429.27	4	
7	3107.318	12429.27	4	
8	3107.318	12429.27	4	
9	3107.318	12429.27	4	
10	3107.318	12429.27	4	
11	3107.318	12429.27	4	
12	3107.318	12429.27	4	
13	3092.521	6185.043	2	
14	3092.521	6185.043	2	
15	3092.521	9277.564	3	
16	3092.521	9277.564	3	
17	3092.521	9277.564	3	
18	3092.521	9277.564	3	
19	3092.521	9277.564	3	
20	3092.521	9277.564	3	
21	3085.123	9255.369	3	
22	3085.123	9255.369	3	

Ví dụ: Biến reg8 lưu trữ thông tin về vùng. Có 8 vùng trong cả nước. Theo bạn làm sao biết được reg8=1 là tương ứng với Vùng nào? (Hãy xem trong sheet tính của file excel Muc1.xls)

tab reg8

reg8	Freq.	Percent	Cum.
1	7,433	19.02	19.02
2	5,698	14.58	33.61
3	2,163	5.54	39.14
4	4,337	11.10	50.24
5	3,634	9.30	59.55
6	2,848	7.29	66.83
7	5,134	13.14	79.97
8	7,824	20.03	100.00
Total	39,071	100.00	

tab reg8 [aw= hhszwt]

reg8	Freq.	Percent	Cum.
1	7,544.17409	19.31	19.31
2	4,442.9992	11.37	30.68
3	1,437.84886	3.68	34.36
4	5,212.3526	13.34	47.70
5	3,285.1965	8.41	56.11
6	2,708.8742	6.93	63.04
7	6,573.3225	16.82	79.87
8	7,866.2321	20.13	100.00

. sum m4ac11

Variable	Obs	Mean	Std. Dev.	Min	Max
m4ac11	7091	11426.16	10756.02	180	480000

. sum m4ac11 [aw=hhszwt]

Variable	Obs	Weight	Mean	Std. Dev.	Min	Max
m4ac11	7091	75166467.1	11798.71	10781.87	180	480000

Khi phân tích VHLSS, với lệnh hồi quy, bạn dùng chữ **pw** thay cho chữ **aw** để khai báo trọng số

- **options (Các tùy chọn)**

Nhiều câu lệnh trong STATA cho phép có các tùy chọn riêng, các tùy chọn này chỉ được chỉ được chỉ ra sau dấu phẩy (dấu ,).

Ví dụ: tùy chọn detail của lệnh sum

```
sum m4ac11 [aw=hhszwt], detail
```

11. Tiền lương, tiền công

	Percentiles	Smallest		
1%	900	180		
5%	2000	210		
10%	3000	210	Obs	7091
25%	5500	225	Sum of Wgt.	75166467.1
50%	9600		Mean	11798.71
		Largest	Std. Dev.	10781.87
75%	15000	120000		
90%	22583	150000	Variance	1.16e+08
95%	30000	156000	Skewness	12.99603
99%	44400	480000	Kurtosis	498.5566

Phụ lục 5. Kiểu dữ liệu, một số lệnh, hàm toán học, toán tử thường dùng*Hàm toán học (Mathematic Functions)*

Câu lệnh	Diễn giải
abs(x)	Giá trị tuyệt đối (Absolute value)
sin(x), cos(x), tan(x)	Sin, cos, tg
int(x), round(x)	Lấy số nguyên/làm tròn số
exp(x)	Hàm mũ Exponential function
ln(x)	Logarit tự nhiên (Natural logarithm)
logit(x), invlogit(x)	Log của tỷ lệ odd và nghịch đảo của nó
max(x), min(x)	GT lớn nhất và nhỏ nhất
sqrt(x)	Căn bậc (Square root)
sum(x)	Tổng cộng

Các lệnh thông dụng về quản lý dữ liệu (Data Management)

Chú ý:	Trong biểu thức dấu == được dùng cho việc kiểm định biểu thức so sánh, thường được dùng sau lệnh if . Còn dấu = được dùng cho phép gán, ví dụ trong lệnh tạo biến mới
des, save, edit	Mô tả biến, Lưu trữ, chỉnh sửa dữ liệu
gen, xtile, replace, recode	Tạo biến mới, tạo biến phân nhóm cho một biến nào đó theo phân vị, thay thế giá trị, mã hoá lại biến
keep, drop	Giữ lại/ xoá biến hay các quan sát
label, format	Tạo nhãn cho biến, tạo định dạng dữ liệu của biến
append, merge	Nối các quan sát, nối các biến từ những file khác nhau
rename	Đổi tên biến
sort, order, move	Sắp xếp các quan sát theo thứ tự, sắp xếp biến, di chuyển biến
egen; collapse	Tạo biến mới; thu gọn dữ liệu

Phân tích hồi quy (Regression Analysis)

Câu lệnh	Diễn giải
correlate , regress	Tương quan, hồi quy với OLS
Logit, Mlogit	Mô hình Binary logistic (mô hình logit), mô hình Mlogit

Các toán tử trong Stata

Ký hiệu	ý nghĩa
<u>Số học</u>	
+	Cộng
-	Trừ
*	Nhân
/	Chia
^	Luỹ thừa
<u>Quan hệ</u>	
>	Lớn hơn
<	Nhỏ hơn
>=	Lớn hơn hoặc bằng
<=	Nhỏ hơn hoặc bằng
==	Bằng
~=	Không bằng (khác)
!=	Không bằng (khác)
<u>Lôgíc</u>	
~	Không
	Hoặc
&	Và

Chú ý:

Trong biểu thức dấu == được dùng cho việc kiểm định biểu thức so sánh, thường được dùng sau lệnh **if**. Còn dấu = được dùng cho phép gán, ví dụ trong lệnh tạo biến mới

Kiểu dữ liệu (Data Types)

Dạng	Hình thức	Diễn giải
float	Số thực	-1.7×10^{38} đến 1.7×10^{36}
double	Số thực	-8.9×10^{307} đến 8.9×10^{307}
byte	Số nguyên	-127 ~ 100
int	Số nguyên	-32767 ~ 32740
long	Số nguyên	-2,147,483,647 ~ 2,147,483,620
str#	Chuỗi (dạng text)	str1 đến str244