An Introduction to statistical methods with Stata

Robert A. Yaffee, Ph.D.

NYU Silver
Silver School *of* Social Work

Jan 14, 2013 th Timberlake Consultants
Union NJ
London, UK

Source of some of the quotes:
http://math.furman.edu/~mwoodard/mqs/ascquotg.html

- "[Statistics are] the only tools by which an opening can be cut through the formidable thicket of difficulties that bars the path of those who pursue the Science of Man."

Reportedly from

Pearson, The Life and Labours of Francis Galton, 1914. downloaded from the above source on 30 June 2009.

Introduction to Stata I

- Invocation of Stata
- Why Stata?
  - Best bang for the buck
  - Easier than R or S+
  - Cheaper than SAS
  - Example datasets are free and included
  - Updated regularly from the web
  - SSC program archive
  - Can handle panel data
  - Can handle complex sample analysis
  - Can handle advanced models
  - Web interface
  - Stata list server
  - Stata Journal
  - Users Group meetings
- Approaches to learning Stata
  - Menus for novices
  - Batch for professionals

Grand outline II

- Introduction to Stata --- continued
  - Configuration of Stata (adding your own editor)
  - Free data sources
  - Variable construction (including date and time variables, etc.)
  - Variable transformations (recoding, replacing, functional, and power)
  - Missing value management (single and multiple imputation)
  - Codebook construction
  - Dataset construction: cross-sectional, longitudinal, time series, panel, survival
  - File management (appending and merging, wide-long conversion)
- Data cleaning
  - Range and consistency checks
  - file comparison
- Exploratory graphical visualization Edward Tufte’s contribution
  - Histograms , Bar graphs, Line graphs, matrix scatterplots, Pie charts, Panel graphs, and Annotation

Grand Outline-III

- Research Project planning concerns
  - Power and sample size analysis   Jacob Cohen's contribution
  - Sampling (simple random, stratified, clustered, stratified -clustered)
  - Attrition and censoring in longitudinal studies
  - Hypothesis testing
- Item analysis and scale construction
  - Reliability and validity analysis
- Summary statistics for sample description
- Categorical data analysis   Leo Goodman's contribution
  - Tabulations
  - Cross-tabulations
  - Statistical tests
- T-tests   William Gossett's contribution
  - One-sample
  - Two independent samples
  - Paired

Grand outline IV

- ANOVA contribution of R.A. Fisher
  - Assumptions and tests for them
  - One-way ANOVA
  - Two-way ANOVA
    - Random, Fixed, and Mixed models
  - Repeated Measures WSANOVA
- Regression analysis   contributions from Gauss and Legendre
  - Univariate
    - Assumptions and tests for them
    - Modeling strategies and critiques
    - General-to-specific (David F. Hendry Jean Francois Richard) Hierarchical, All possible subsets
    - Robust regression (Halbert White and Huber and others)
      - Heteroscedastically consistent estimation
      - Outlier down-weighting
- Bootstrapping regression models   Brad Efron's contribution

Grand Outline V if time permits

- Regression analysis with Limited Dependent Variables
  - Poisson count models
  - Logistic and Probit models for binary dependent variables
  - Skewed logistic models
  - Ordered Logistic and Ordered Probit regression models for ordinal dependent variables
  - Multinomial logistic regression models for categorical dependent variables

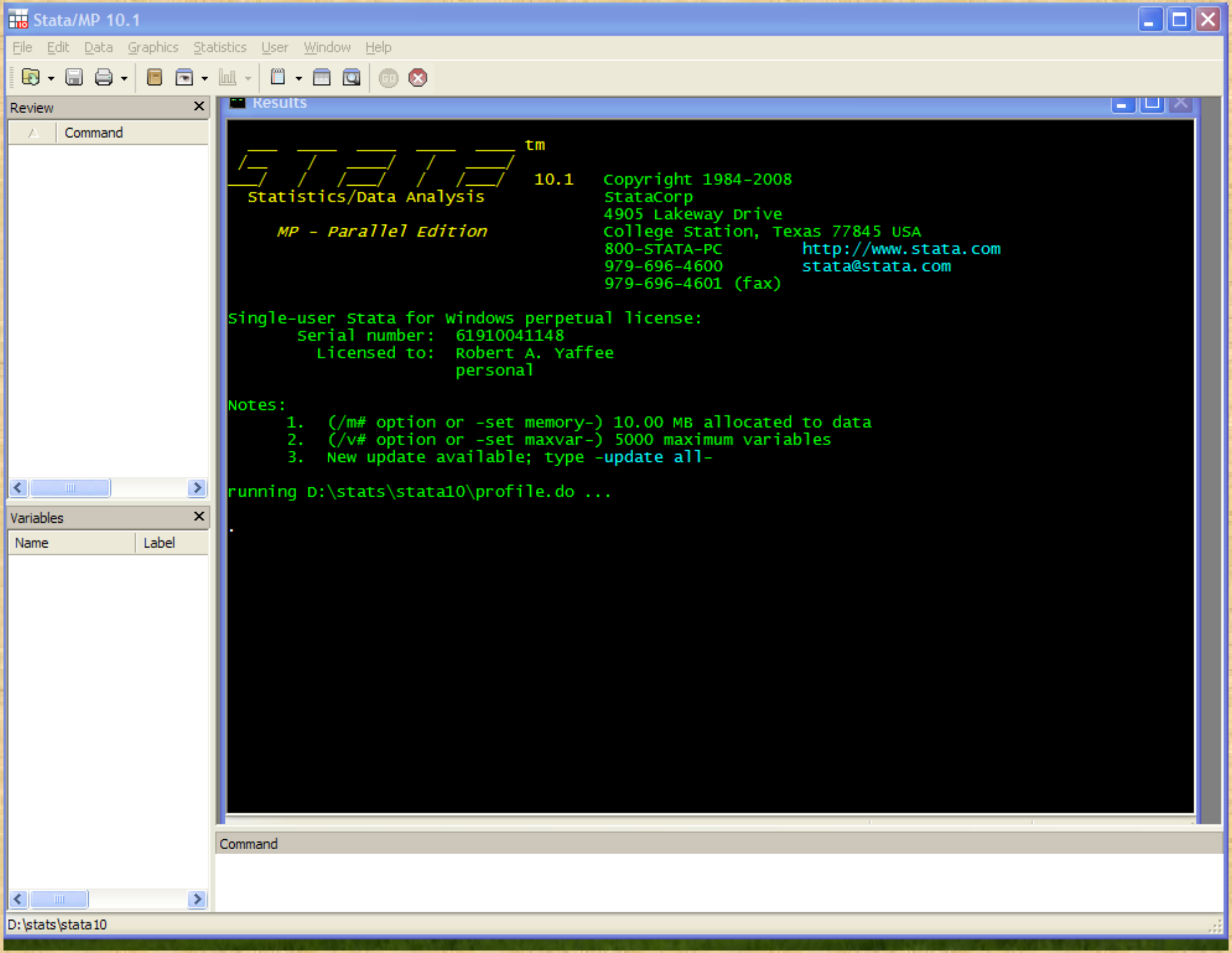
Configuration, logging, execution, and output

- Configuring your Stata
  - Preferences
  - profile.do command file
  - Logging your own work
  - smcl files
  - Translate command
  - Saving graphs
- Running Stata
  - Saving output
  - Printing output

Configuring Stata:

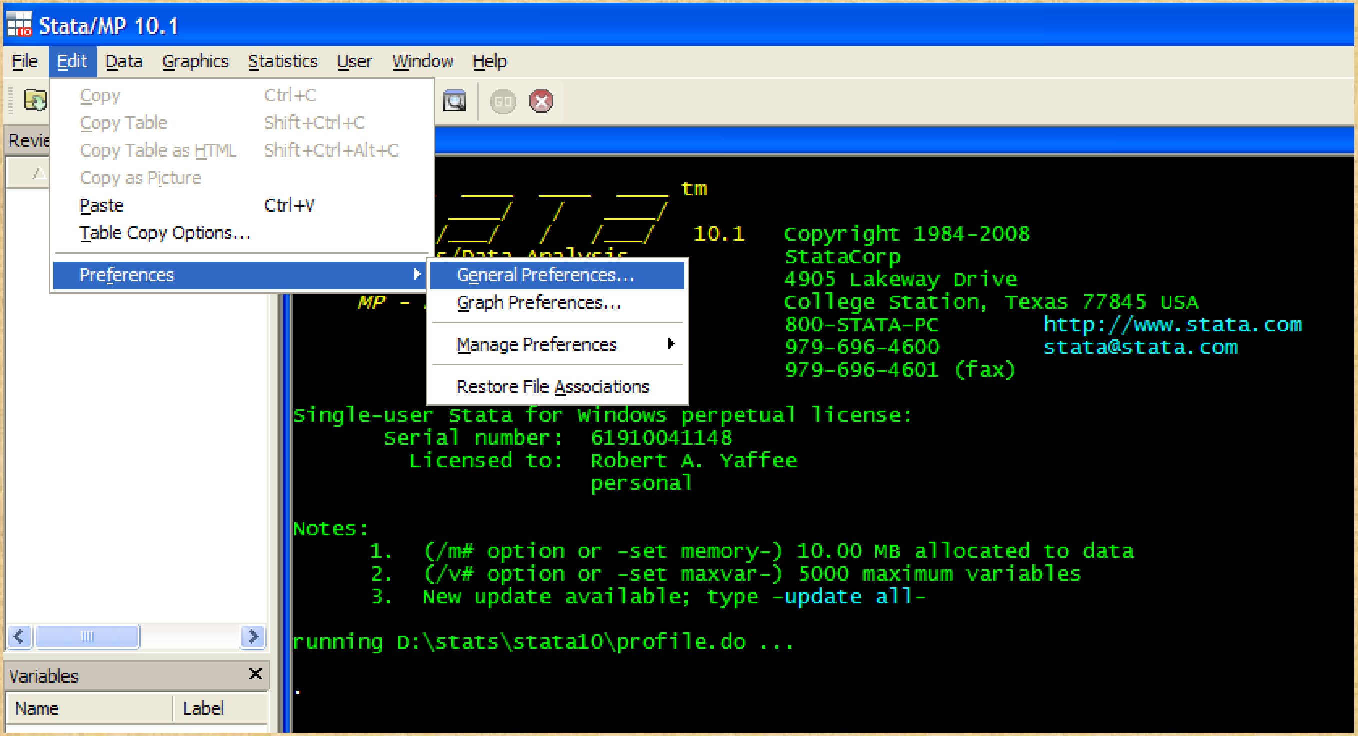
Double click on Stata icon

Stata platform appears



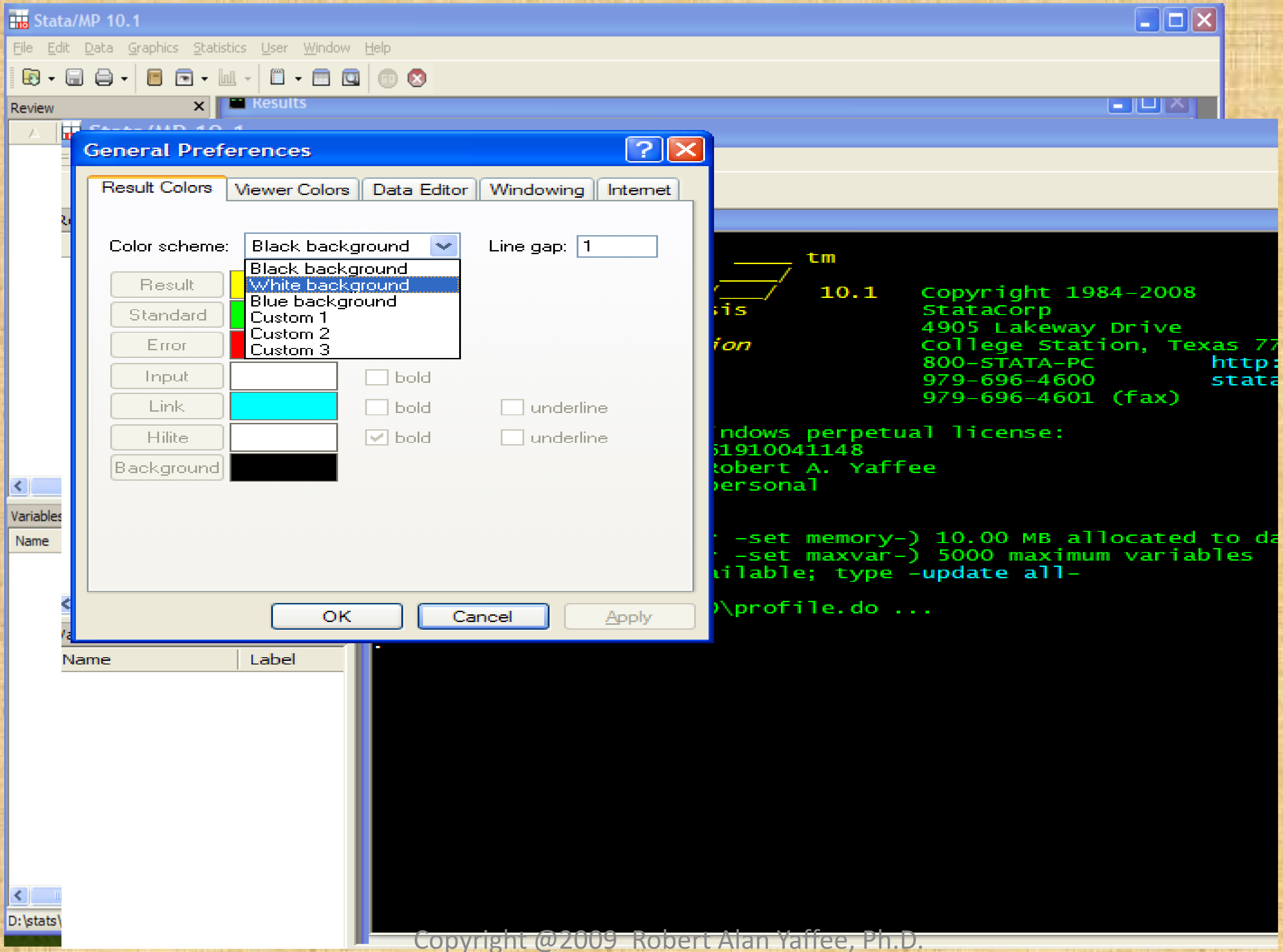
```
  ___  ____  ____  ____  ____ tm
 /__    /   ____/   /   ____/
___/   /   /___/   /   /___/   10.1   Copyright 1984-2008
  Statistics/Data Analysis            StataCorp
                                      4905 Lakeway Drive
     MP - Parallel Edition            College Station, Texas 77845 USA
                                      800-STATA-PC        http://www.stata.com
                                      979-696-4600        stata@stata.com
                                      979-696-4601 (fax)

Single-user Stata for Windows perpetual license:
       Serial number:  61910041148
         Licensed to:  Robert A. Yaffee
                       personal

Notes:
      1.  (/m# option or -set memory-) 10.00 MB allocated to data
      2.  (/v# option or -set maxvar-) 5000 maximum variables
      3.  New update available; type -update all-

running D:\stats\stata10\profile.do ...

.
```

Click Edit, preferences, General Preferences



```
                                                   tm
                   10.1   Copyright 1984-2008
                          StataCorp
                          4905 Lakeway Drive
                          College Station, Texas 77845 USA
                          800-STATA-PC        http://www.stata.com
                          979-696-4600        stata@stata.com
                          979-696-4601 (fax)

Single-user Stata for Windows perpetual license:
       Serial number:  61910041148
         Licensed to:  Robert A. Yaffee
                       personal

Notes:
      1.  (/m# option or -set memory-) 10.00 MB allocated to data
      2.  (/v# option or -set maxvar-) 5000 maximum variables
      3.  New update available; type -update all-

running D:\stats\stata10\profile.do ...

.
```

Select White background and click “OK”



Color scheme: Black background

Line gap: 1

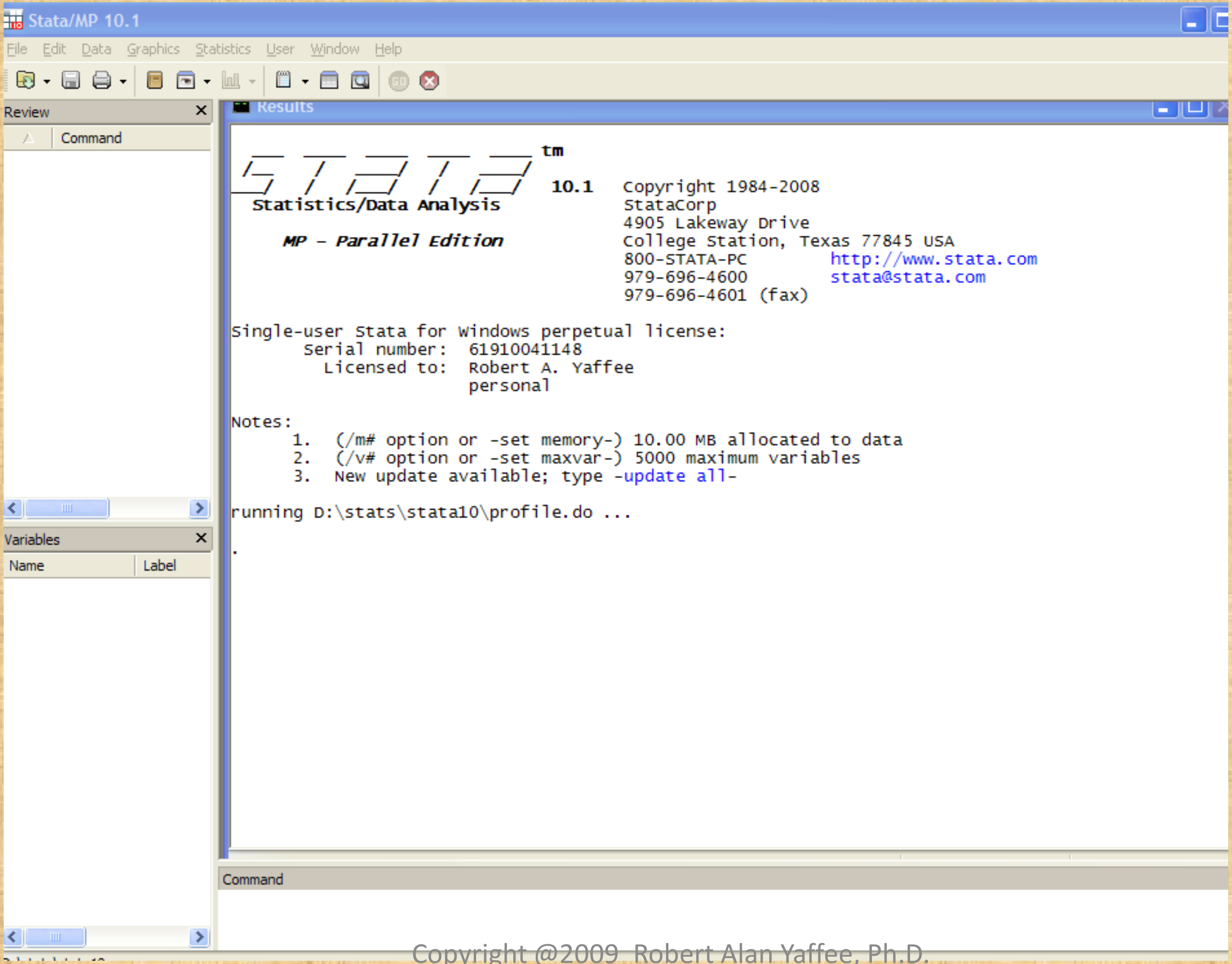
☐ bold

☐ bold ☐ underline

☑ bold ☐ underline

```
  ____   tm
/____/    10.1    Copyright 1984-2008
is                StataCorp
                  4905 Lakeway Drive
ion               College Station, Texas 77
                  800-STATA-PC        http:
                  979-696-4600        stata
                  979-696-4601 (fax)

ndows perpetual license:
51910041148
Robert A. Yaffee
personal

' -set memory-) 10.00 MB allocated to da
' -set maxvar-) 5000 maximum variables
ilable; type -update all-

)\profile.do ...
```

The Background color is now white



```
  ___  ____  ____  ____  ____ tm
 /__    /   ____/   /   ____/
___/   /   /___/   /   /___/   10.1   Copyright 1984-2008
  Statistics/Data Analysis            StataCorp
                                      4905 Lakeway Drive
     MP - Parallel Edition            College Station, Texas 77845 USA
                                      800-STATA-PC        http://www.stata.com
                                      979-696-4600        stata@stata.com
                                      979-696-4601 (fax)

Single-user Stata for Windows perpetual license:
       Serial number:  61910041148
         Licensed to:  Robert A. Yaffee
                       personal

Notes:
      1.  (/m# option or -set memory-) 10.00 MB allocated to data
      2.  (/v# option or -set maxvar-) 5000 maximum variables
      3.  New update available; type -update all-

running D:\stats\stata10\profile.do ...

.
```

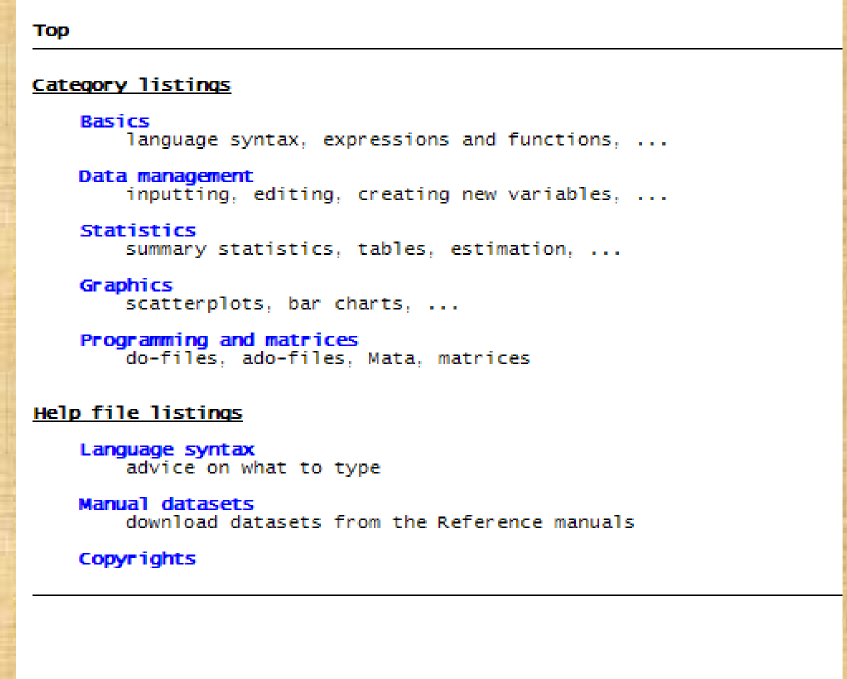
Changing font size in any window

- You may for presentation or personal display, right click on any window, and alter the font size.
- This can make the output easier to read for those who are viewing the output.

Getting help with Stata

- F1 is help
- You can type: help command,
  - where command is any command you need help with for the proper syntax
  - you can type: find keyword
on the command line where a keyword will help the search progress among the Stata help files and on the internet
  - You can google a Stata command and get help on the internet

Type: help or help keyword



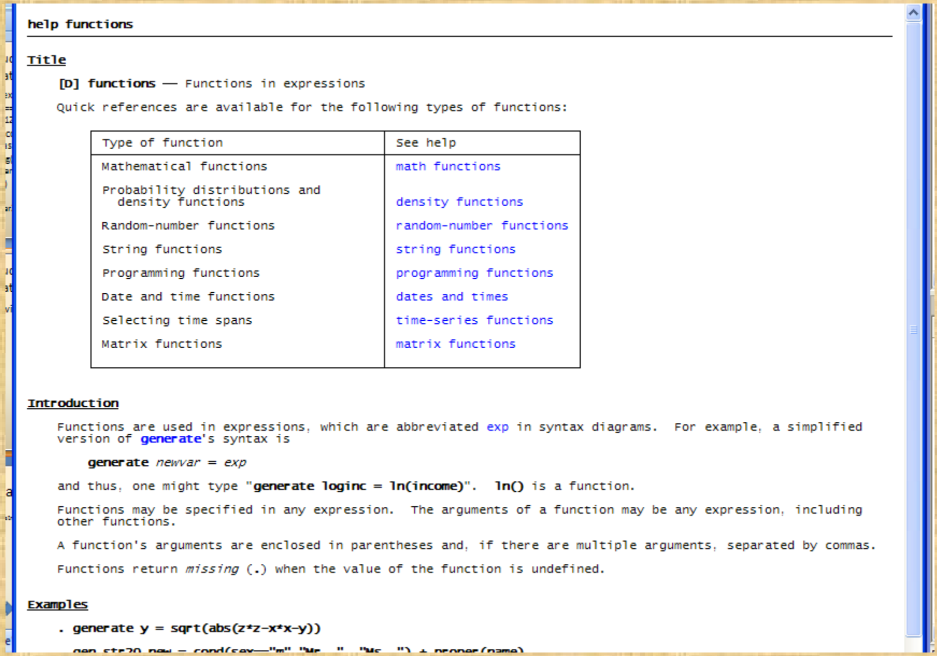
Top

Category listings

- **Basics**
  language syntax, expressions and functions, ...
- **Data management**
  inputting, editing, creating new variables, ...
- **Statistics**
  summary statistics, tables, estimation, ...
- **Graphics**
  scatterplots, bar charts, ...
- **Programming and matrices**
  do-files, ado-files, Mata, matrices

Help file listings

- **Language syntax**
  advice on what to type
- **Manual datasets**
  download datasets from the Reference manuals
- **Copyrights**



help functions

Title

[D] functions — Functions in expressions

Quick references are available for the following types of functions:

Type of function	See help
Mathematical functions	math functions
Probability distributions and density functions	density functions
Random-number functions	random-number functions
String functions	string functions
Programming functions	programming functions
Date and time functions	dates and times
Selecting time spans	time-series functions
Matrix functions	matrix functions

Introduction

Functions are used in expressions, which are abbreviated exp in syntax diagrams. For example, a simplified version of **generate**'s syntax is

generate *newvar* = *exp*

and thus, one might type "**generate loginc = ln(income)**". **ln()** is a function.

Functions may be specified in any expression. The arguments of a function may be any expression, including other functions.

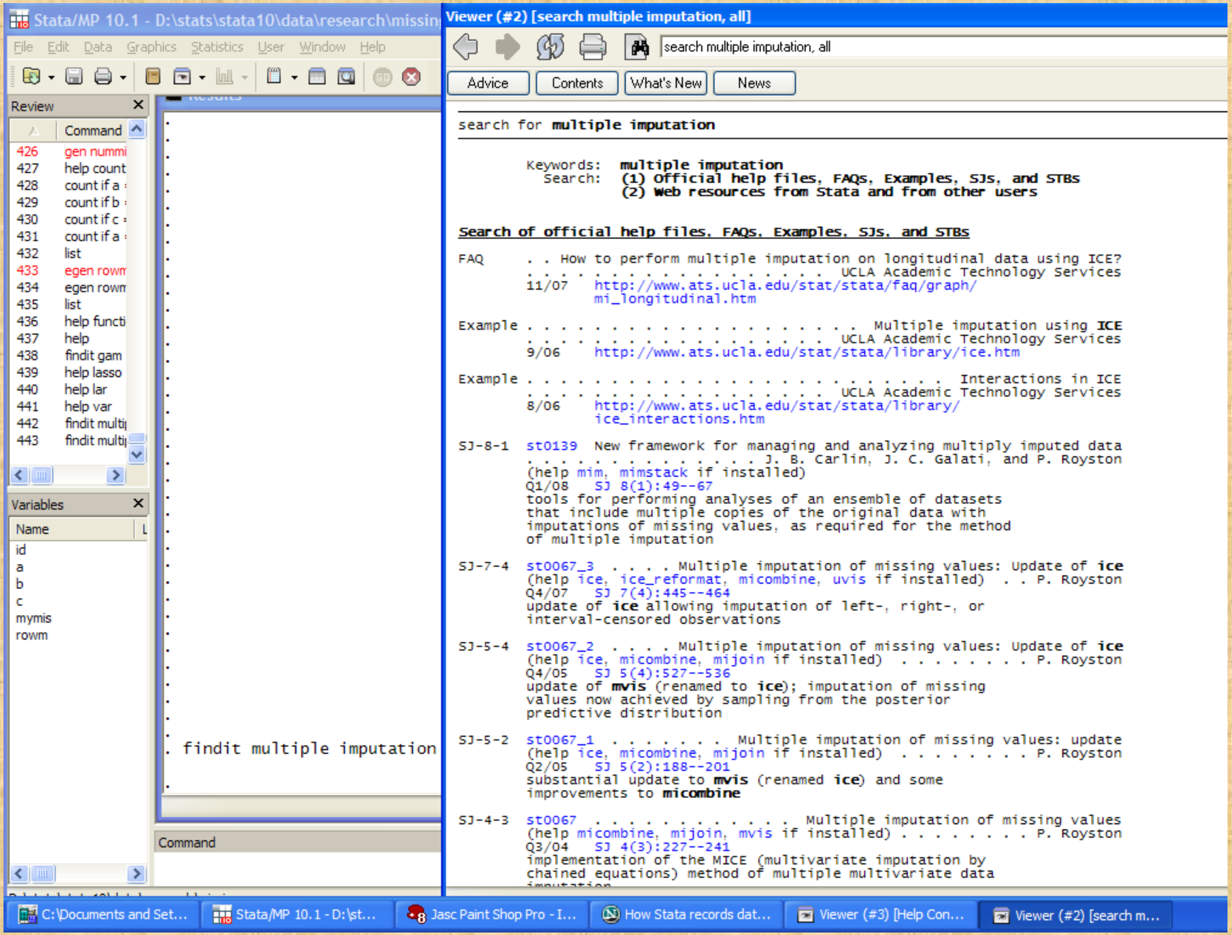
A function's arguments are enclosed in parentheses and, if there are multiple arguments, separated by commas.

Functions return *missing* (.) when the value of the function is undefined.

Examples

```
. generate y = sqrt(abs(z*z-x*x-y))
```

Type: findit keyword



```
. findit multiple imputation
```

```
search multiple imputation, all
```

```
search for multiple imputation

        Keywords:  multiple imputation
          Search:  (1) Official help files, FAQs, Examples, SJs, and STBs
                   (2) Web resources from Stata and from other users

Search of official help files, FAQs, Examples, SJs, and STBs

FAQ     . . How to perform multiple imputation on longitudinal data using ICE?
        . . . . . . . . . . . . . . . . . .  UCLA Academic Technology Services
        11/07   http://www.ats.ucla.edu/stat/stata/faq/graph/
                mi_longitudinal.htm

Example . . . . . . . . . . . . . . . . . . . .  Multiple imputation using ICE
        . . . . . . . . . . . . . . . . . .  UCLA Academic Technology Services
        9/06    http://www.ats.ucla.edu/stat/stata/library/ice.htm

Example . . . . . . . . . . . . . . . . . . . . . . . . .  Interactions in ICE
        . . . . . . . . . . . . . . . . . .  UCLA Academic Technology Services
        8/06    http://www.ats.ucla.edu/stat/stata/library/
                ice_interactions.htm

SJ-8-1  st0139  New framework for managing and analyzing multiply imputed data
        . . . . . . . . . . . . . .  J. B. Carlin, J. C. Galati, and P. Royston
        (help mim, mimstack if installed)
        Q1/08   SJ 8(1):49--67
        tools for performing analyses of an ensemble of datasets
        that include multiple copies of the original data with
        imputations of missing values, as required for the method
        of multiple imputation

SJ-7-4  st0067_3  . . . . Multiple imputation of missing values: Update of ice
        (help ice, ice_reformat, micombine, uvis if installed)  . . P. Royston
        Q4/07   SJ 7(4):445--464
        update of ice allowing imputation of left-, right-, or
        interval-censored observations

SJ-5-4  st0067_2  . . . . Multiple imputation of missing values: Update of ice
        (help ice, micombine, mijoin if installed)  . . . . . . . . P. Royston
        Q4/05   SJ 5(4):527--536
        update of mvis (renamed to ice); imputation of missing
        values now achieved by sampling from the posterior
        predictive distribution

SJ-5-2  st0067_1  . . . . . . .  Multiple imputation of missing values: update
        (help ice, micombine, mijoin if installed)  . . . . . . . . P. Royston
        Q2/05   SJ 5(2):188--201
        substantial update to mvis (renamed ice) and some
        improvements to micombine

SJ-4-3  st0067  . . . . . . . . . . . .  Multiple imputation of missing values
        (help micombine, mijoin, mvis if installed) . . . . . . . . P. Royston
        Q3/04   SJ 4(3):227--241
        implementation of the MICE (multivariate imputation by
        chained equations) method of multiple multivariate data
```

Important Stata Resources

- Stata has excellent manuals
- Stata offers first rate technical support
- Stata can download from the web
- UCLA ATS has excellent Stata help
- It has movies which teach Stata for those who need or wish visual instruction
- Stata Press publishes texts dealing with Stata commands
- It has a list of command examples
- Type: findit keyword on command line while connected to web. Keyword is any name in which you are interested.
- FRED St. Louis Federal Reserve Economic Database : freduse command
- Yahoo
- Economic report to the President

SSC archive

- Be sure you are connected to the www
- Type: ssc describe a
- Type: ssc describe b

Installing from the web

- Suppose you wish to download the datasets and do files from Regression models and categorical variables using Stata by J. Scott Long and Jeremy Freese. You could use the following commands:

- net search spost
- if you are using version 9, you can execute the following commands:
- net get rmcdvs
- net from http://www.indiana.edu/~jslsoc/data
- net get spost9_do
- net install spost9_ado

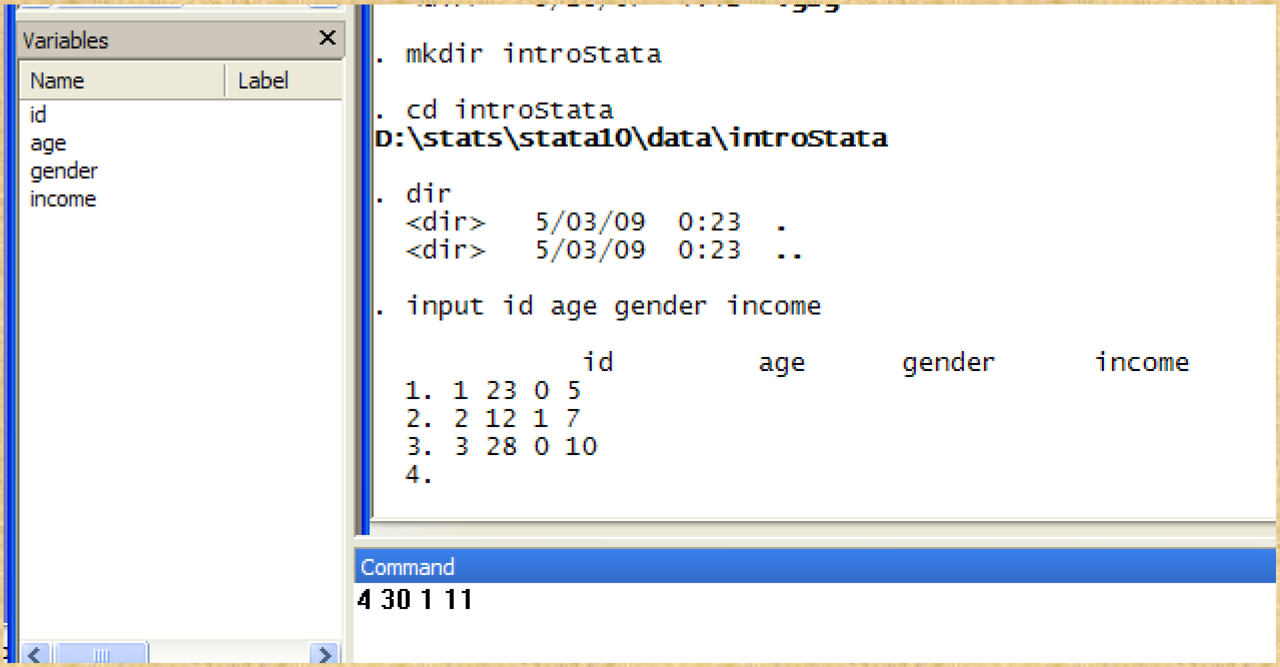
File construction and data definition

- File construction
  - Input command
    - Id variables
      - For rectangular datasets
      - For hierarchical datasets
    - Date variables
      - For time series datasets
    - Panel variables and date variables
      - For panel datasets
    - Variable definition
      - Numeric
      - String
      - dates
    - Variable labels
    - Formats
    - Linking formats

Data definition-continued

- missing data management
- Wide files
- Long files
- Save
- Saveold
- Do files
- Ado files

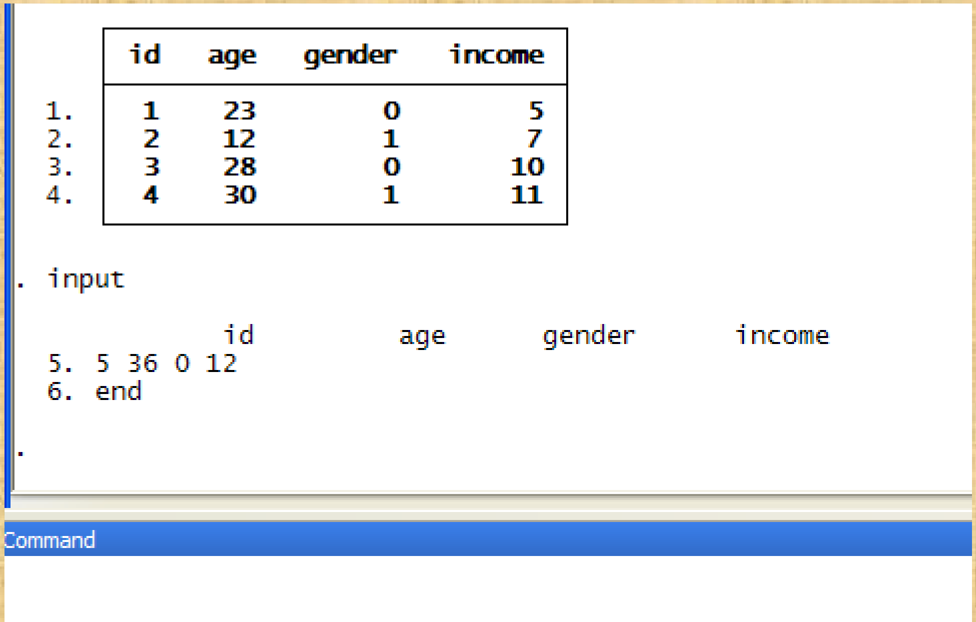
File Construction: for raw data input, the input command can be used



```
. mkdir introStata

. cd introStata
D:\stats\stata10\data\introStata

. dir
  <dir>   5/03/09  0:23  .
  <dir>   5/03/09  0:23  ..

. input id age gender income

                id          age       gender        income
  1. 1 23 0 5
  2. 2 12 1 7
  3. 3 28 0 10
  4.
```

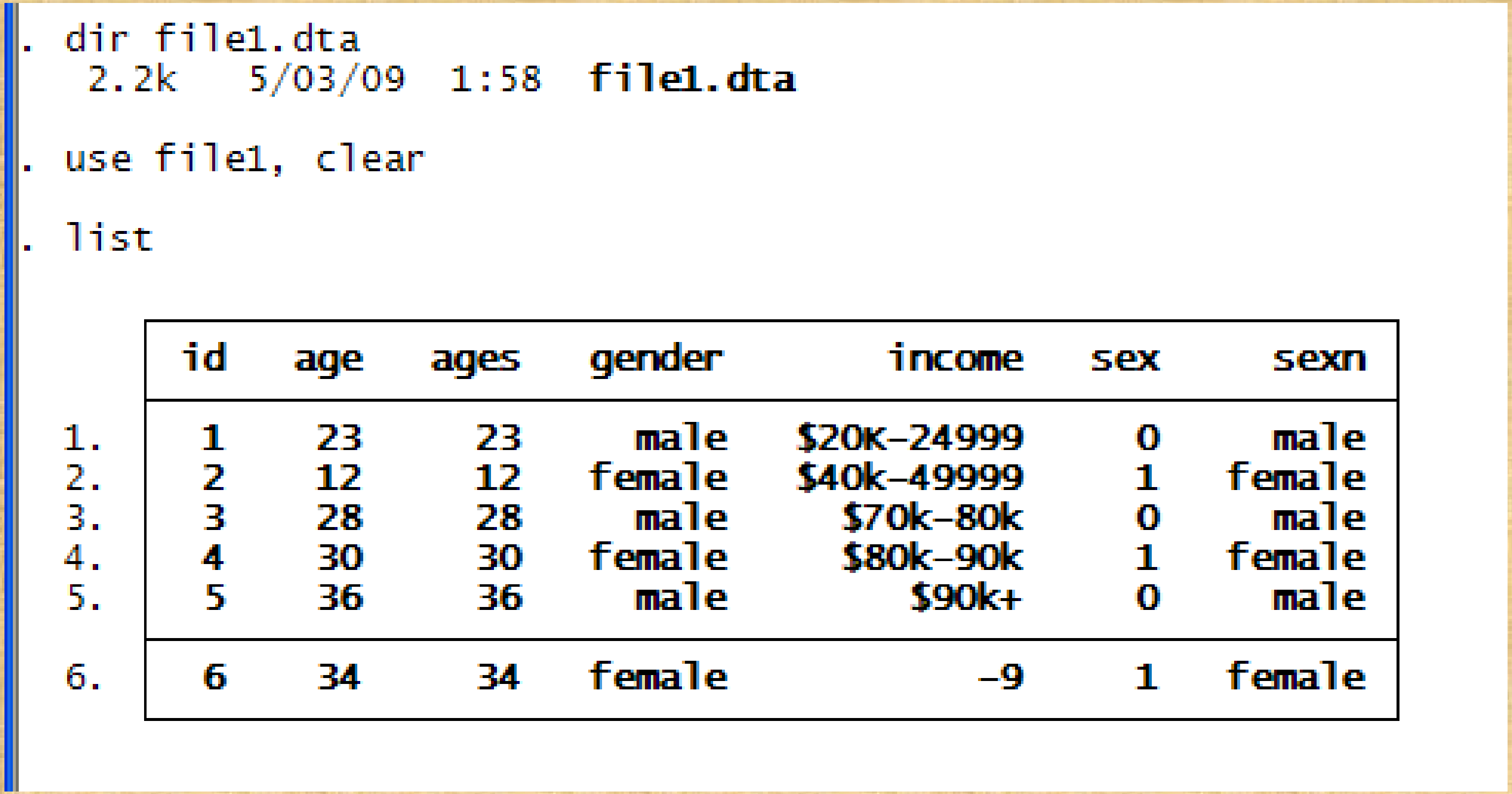
```
4 30 1 11
```

Type “end” to complete data input



```
     +---------------------------------+
     |  id   age   gender   income     |
     |---------------------------------|
  1. |   1    23        0        5     |
  2. |   2    12        1        7     |
  3. |   3    28        0       10     |
  4. |   4    30        1       11     |
     +---------------------------------+

. input

            id          age        gender        income
  5. 5 36 0 12
  6. end

.
```

Accessing a Stata dataset file1.dta with the use command

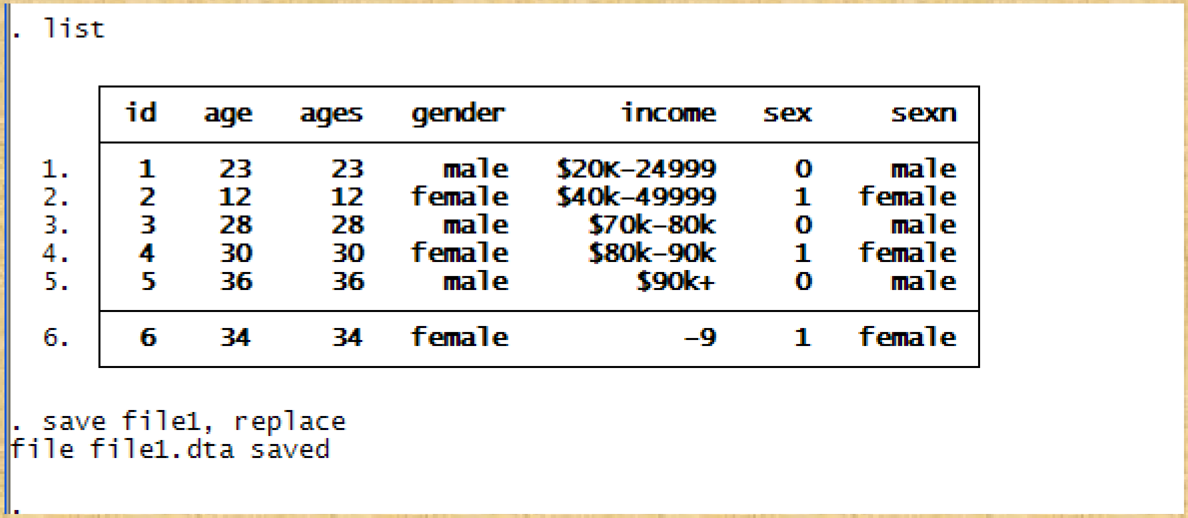


```
. dir file1.dta
   2.2k   5/03/09  1:58   file1.dta

. use file1, clear

. list

     +------------------------------------------------------------+
     | id   age   ages   gender        income   sex          sexn |
     |------------------------------------------------------------|
  1. |  1    23     23     male   $20K-24999     0          male |
  2. |  2    12     12   female   $40k-49999     1        female |
  3. |  3    28     28     male     $70k-80k     0          male |
  4. |  4    30     30   female     $80k-90k     1        female |
  5. |  5    36     36     male       $90k+     0          male |
     |------------------------------------------------------------|
  6. |  6    34     34   female           -9     1        female |
     +------------------------------------------------------------+
```

Saving a Stata dataset

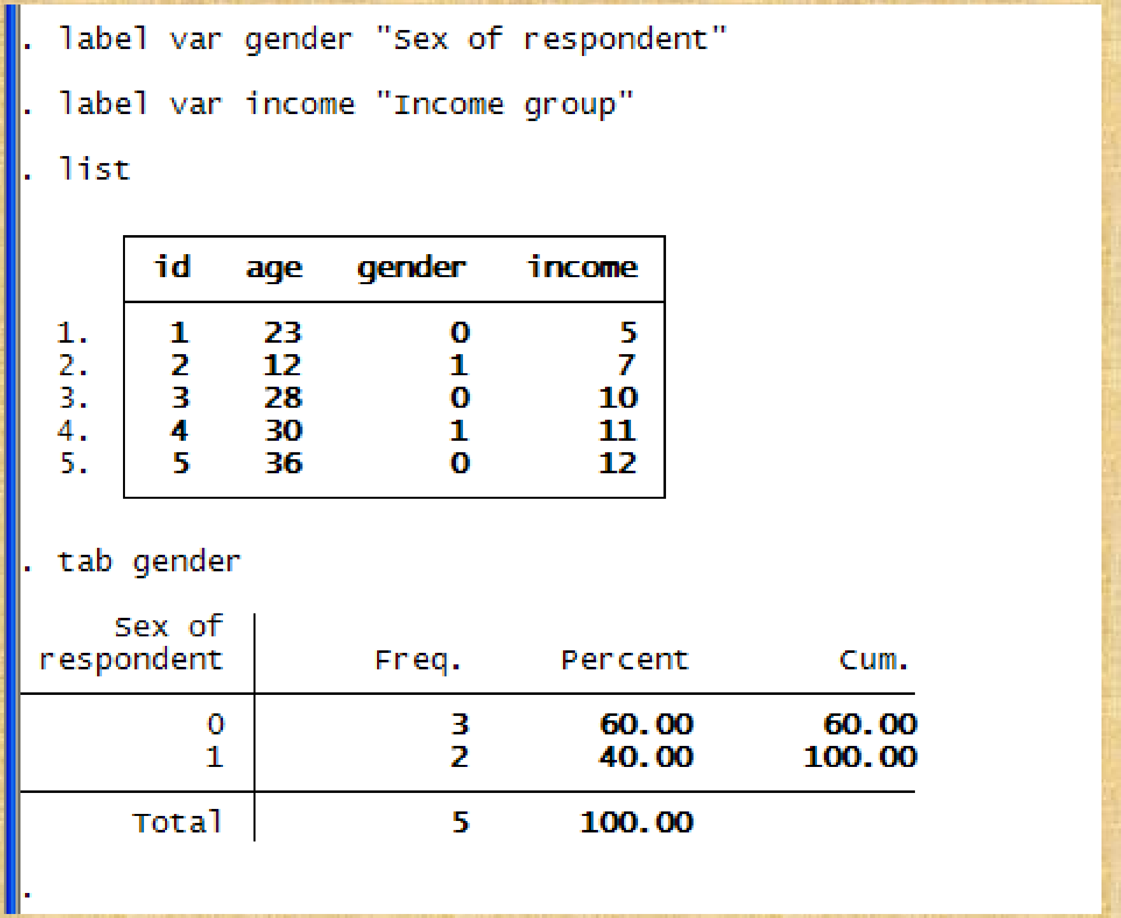
- You can type: save filename
  - If this is the first time you are saving it.
- You can type: save filename, replace
  - If you are replacing an earlier version with a newer one.
- You can type: saveold filename
  - If you wish to save it in Stata9 format

Saving a Stata dataset



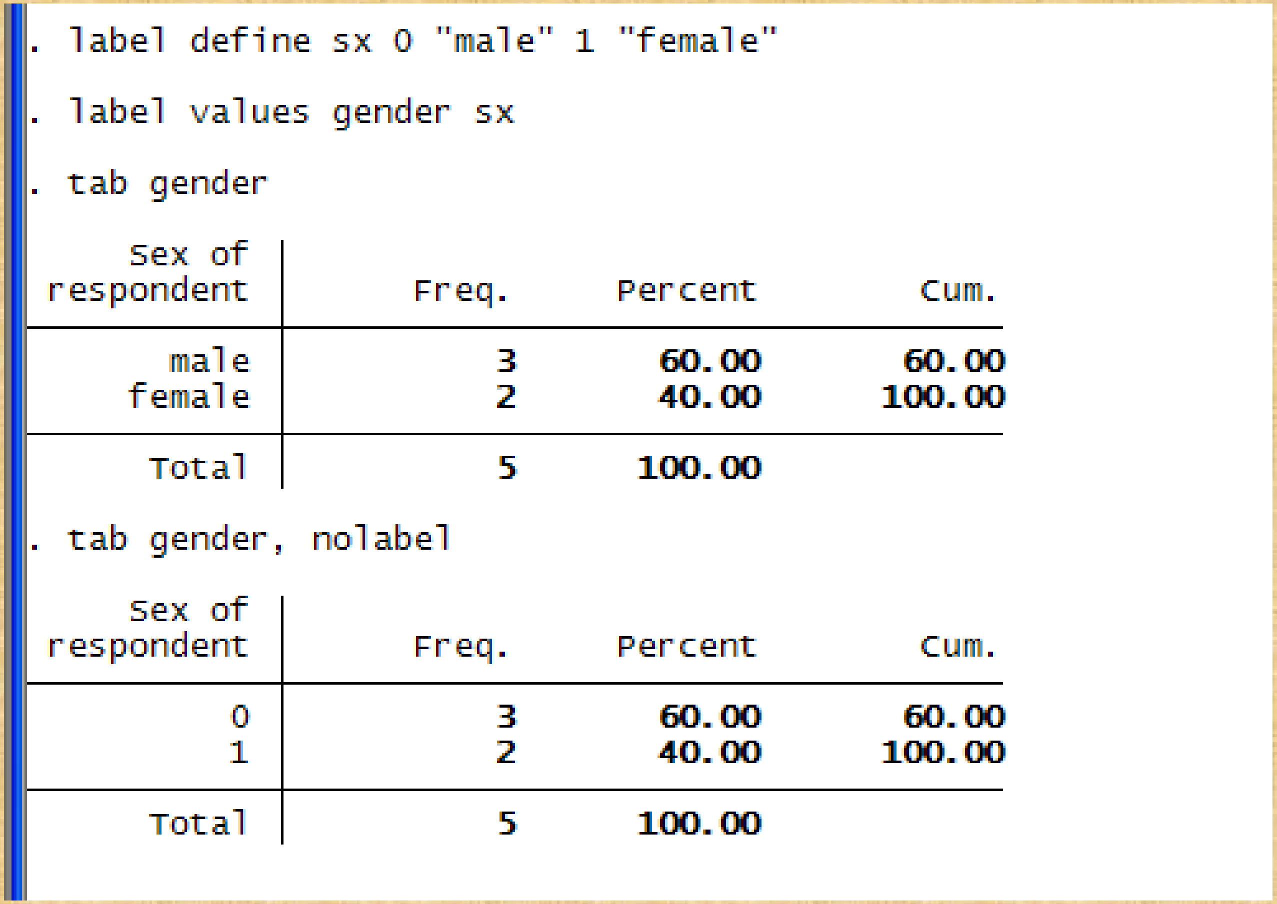
```
. list

     +---------------------------------------------------------+
     | id   age   ages   gender        income   sex      sexn |
     |---------------------------------------------------------|
  1. |  1    23     23     male   $20K-24999     0      male |
  2. |  2    12     12   female   $40k-49999     1    female |
  3. |  3    28     28     male     $70k-80k     0      male |
  4. |  4    30     30   female     $80k-90k     1    female |
  5. |  5    36     36     male        $90k+     0      male |
     |---------------------------------------------------------|
  6. |  6    34     34   female           -9     1    female |
     +---------------------------------------------------------+

. save file1, replace
file file1.dta saved

.
```

Data definition
Variable labels



```
. label var gender "Sex of respondent"

. label var income "Income group"

. list

     +----------------------------+
     | id   age   gender   income |
     |----------------------------|
  1. |  1    23        0        5 |
  2. |  2    12        1        7 |
  3. |  3    28        0       10 |
  4. |  4    30        1       11 |
  5. |  5    36        0       12 |
     +----------------------------+

. tab gender

     Sex of |
 respondent |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |          3       60.00       60.00
          1 |          2       40.00      100.00
------------+-----------------------------------
      Total |          5      100.00

.
```

Data definition
Value labels or formats

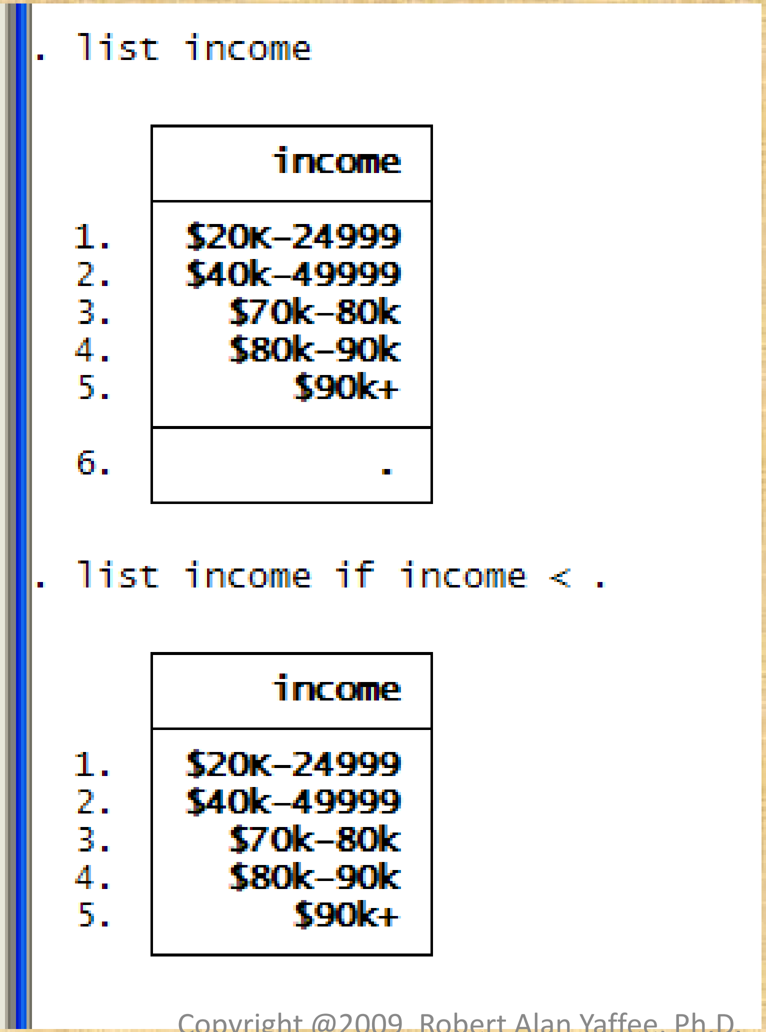


```
. label define sx 0 "male" 1 "female"

. label values gender sx

. tab gender

     Sex of |
 respondent |      Freq.     Percent        Cum.
------------+-----------------------------------
       male |          3       60.00       60.00
     female |          2       40.00      100.00
------------+-----------------------------------
      Total |          5      100.00

. tab gender, nolabel

     Sex of |
 respondent |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |          3       60.00       60.00
          1 |          2       40.00      100.00
------------+-----------------------------------
      Total |          5      100.00
```

Missing values in Stata

- Missing values in Stata are treated as large positive numbers
- They may be system missing and represented by a .
- They may be 26 other codes from .a to .z
  - For missing values analysis.
- Therefore, when executing operations in Stata, you might want to qualify your requests for estimations with the condition if not equal to missing, for example
- list income, if income < .

Stata will omit these system missing values from computations



```
. list income

     +-----------+
     |    income |
     |-----------|
  1. | $20K-24999 |
  2. | $40k-49999 |
  3. |   $70k-80k |
  4. |   $80k-90k |
  5. |     $90k+ |
     |-----------|
  6. |         . |
     +-----------+

. list income if income < .

     +-----------+
     |    income |
     |-----------|
  1. | $20K-24999 |
  2. | $40k-49999 |
  3. |   $70k-80k |
  4. |   $80k-90k |
  5. |     $90k+ |
     +-----------+
```

Variable construction: with generate

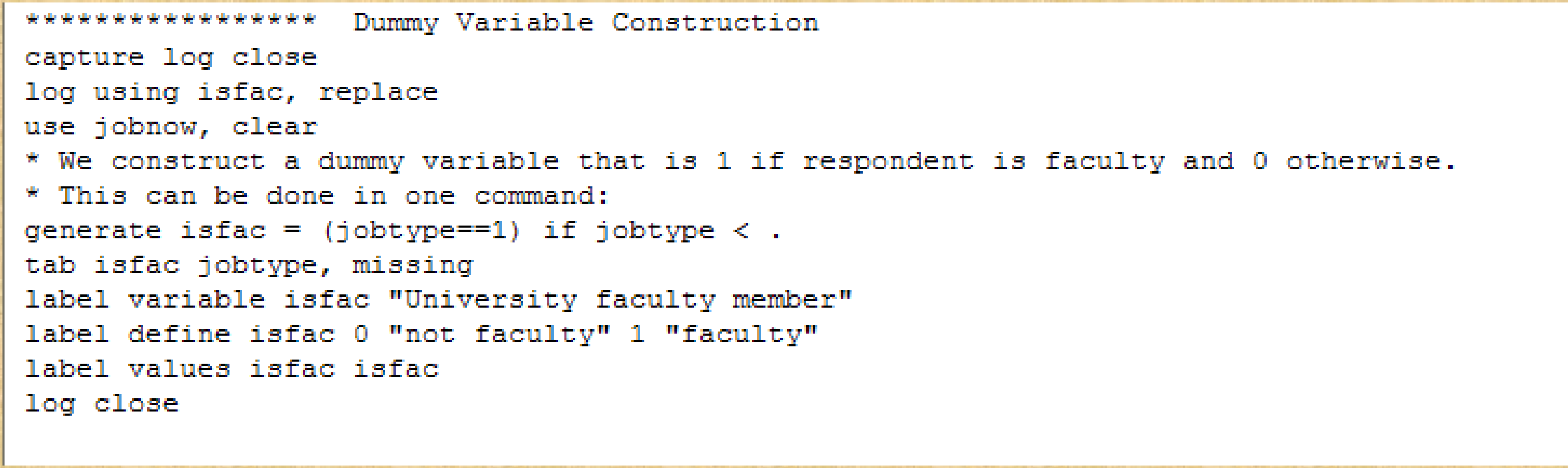
When constructing variables, be sure you don’t recode the missing into 0 by using an if income < .

generate wealthy = 0 if income < .

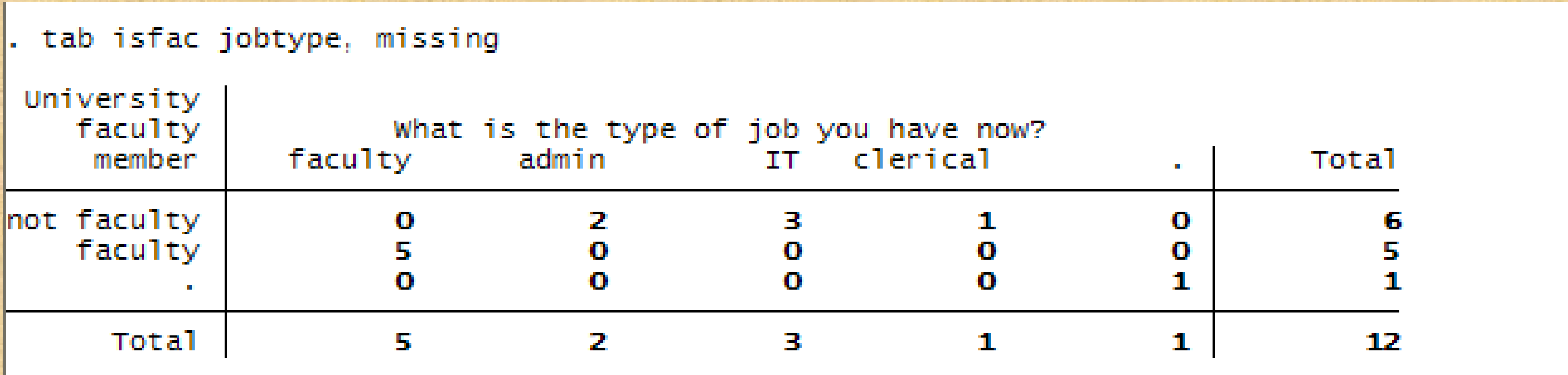
replace wealthy = 1 if income < . & income > 7

Dummy Variable construction

Long and Freese, op cit, 68-70.



```
*****************  Dummy Variable Construction
capture log close
log using isfac, replace
use jobnow, clear
* We construct a dummy variable that is 1 if respondent is faculty and 0 otherwise.
* This can be done in one command:
generate isfac = (jobtype==1) if jobtype < .
tab isfac jobtype, missing
label variable isfac "University faculty member"
label define isfac 0 "not faculty" 1 "faculty"
label values isfac isfac
log close
```



```
. tab isfac jobtype, missing

 University |
    faculty |          What is the type of job you have now?
     member |   faculty      admin         IT   clerical          . |     Total
------------+-------------------------------------------------------+----------
not faculty |         0          2          3          1          0 |         6
    faculty |         5          0          0          0          0 |         5
          . |         0          0          0          0          1 |         1
------------+-------------------------------------------------------+----------
      Total |         5          2          3          1          1 |        12
```

Stata egen functions

The egen rowwise functions all ignore missing values . They will only return a missing if all components are missing. For example:

egen x123max= rowmax(x1,x2,x3) computes row maximum of the three variables specified.

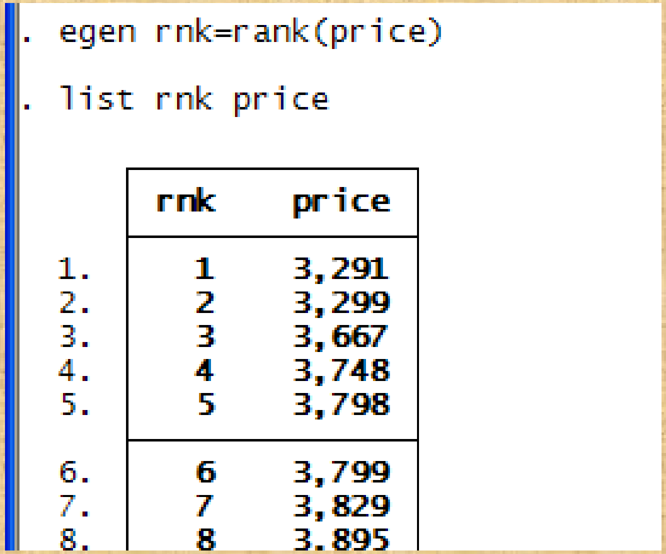
egen x123mean=rowmean(x1,x2,x3) computes row mean of x1,x2, and x3.

egen x123total=rowtotal(x1,x2,x3) computes rowtotal of x1,x2, and x3.

egen rowmss = rowmiss(x1,x2,x3,x4) indicates number of missing values in the row of x1 through x4 variables.

Other Stata egen functions

- egen rnk=rank(v1)
  - Will rank the cases according to variable v1



```
. egen rnk=rank(price)

. list rnk price

     +-------------+
     | rnk   price |
     |-------------|
  1. |   1   3,291 |
  2. |   2   3,299 |
  3. |   3   3,667 |
  4. |   4   3,748 |
  5. |   5   3,798 |
     |-------------|
  6. |   6   3,799 |
  7. |   7   3,829 |
  8. |   8   3,895 |
```

Anycount(varlist), values(numlist)
Anyvalue (varlist), values(integer numlist)
Mean(varlist)
Median(varlist)
mode(“)

creates a constant in a list containing this statistic

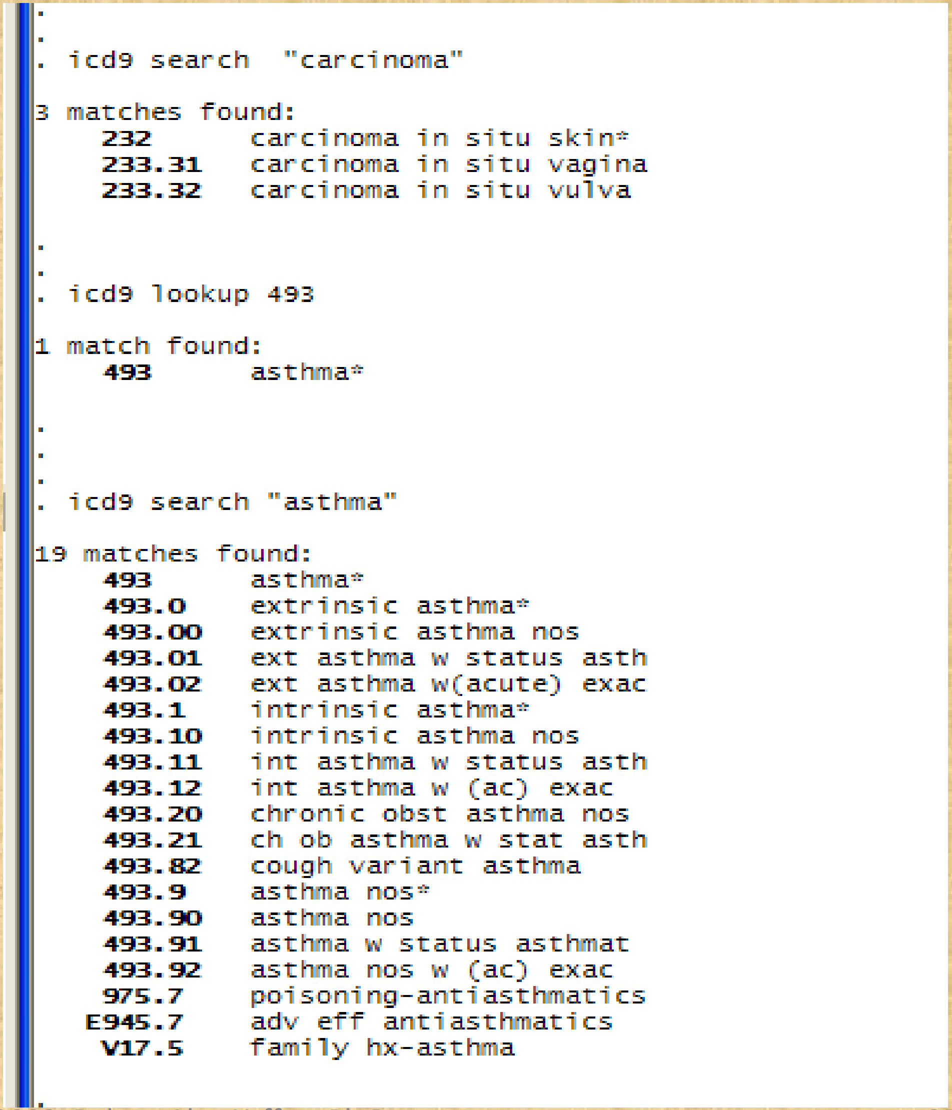
ICD9 codes are stored within

Medical researchers use the international statistical codes for diseases and related health problems

Stata has them built in.

They are regularly updated

You can generate new variables with them or search old variables for elaborated definitions.



```
.
.
. icd9 search  "carcinoma"

3 matches found:
    232        carcinoma in situ skin*
    233.31     carcinoma in situ vagina
    233.32     carcinoma in situ vulva

.
.
. icd9 lookup 493

1 match found:
    493        asthma*

.
.
.
. icd9 search "asthma"

19 matches found:
    493        asthma*
    493.0      extrinsic asthma*
    493.00     extrinsic asthma nos
    493.01     ext asthma w status asth
    493.02     ext asthma w(acute) exac
    493.1      intrinsic asthma*
    493.10     intrinsic asthma nos
    493.11     int asthma w status asth
    493.12     int asthma w (ac) exac
    493.20     chronic obst asthma nos
    493.21     ch ob asthma w stat asth
    493.82     cough variant asthma
    493.9      asthma nos*
    493.90     asthma nos
    493.91     asthma w status asthmat
    493.92     asthma nos w (ac) exac
    975.7      poisoning-antiasthmatics
   E945.7      adv eff antiasthmatics
    V17.5      family hx-asthma

.
```

Standardization of variables
Long and Freese, op. cit., p.96

- X standardized coefficients

Suppose you have a regression formula,

x-standardization

$$y = a + b_1 x_1 + b_2 x_2 + \cdots + b_k x_k + e$$

where

$e =$ *error, disturbance, innovation, shock*

we divide each x_k *by* σ_k *and multiply the* b_k *by that quantity:*

$$y = a + \sigma_1 b_1 \frac{x_1}{\sigma_1} + \sigma_2 b_2 \frac{x_2}{\sigma_2} + \cdots + \sigma_k b_k \frac{x_k}{\sigma_k} + e, \quad \text{so the } x-\text{stdzed}$$

coefficient $=$

$$\beta_1^s = \sigma_1 b_1 \frac{x_1}{\sigma_1}$$

Interpretation of x-standardization

- For a continuous variable, for an increase in one standard deviation of x, the amount of change in the dependent variable, y, holding all other x variables constant, associated with this increase in x is :
- This amount = $\beta^s = \sigma b$

y standardization

- When we divide a continuous dependent variable by its standard deviation, we have to divide the whole equation by the same amount. This is called y standardization.

Y standardization

$$y\text{-}standardization$$

$$y = a + b_1 x_1 + b_2 x_2 + \cdots + b_k x_k + e$$

where

$e = error,\ disturbance,\ innovation,\ shock$

and $\sigma_y = $ *standard deviation of the dependent variable.*

we divide y *and each* b_k *by* σ_y :

$$\frac{y}{\sigma_y} = \frac{a}{\sigma_y} + \frac{b_1}{\sigma_y} x_1 + \frac{b_2}{\sigma_y} x_2 + \cdots + \frac{b_k}{\sigma_y} x_k + \frac{e}{\sigma_y}, \quad \text{so the } y - \text{stdzed}$$

coefficient $=$

$$\beta_k^{s_y} = \frac{b_k}{\sigma_y}$$

Interpretation of Y standardization

Long and Feeze, op. cit., 97

- For an increase in one unit of x_k, the amount of change in Y associated with that change is β^{sy} standard deviations, holding all other variables constant.
- For a dummy variable having characteristic x as opposed to not having it, the amount of change in Y is β^{sy} standard deviations, holding all other variables constant.

Y standardization with latent variable y* Long and Freese, op. cit.,97

- We divide the whole equation by the standard deviation of y. It is assumed that the variance of the error in a probit model=1.
- To estimate the variance of the latent variable y*, we find that
- $\text{Var}(y^*) = \beta \text{Var}(x) \beta + \text{Var}(e)$ so that
- $\text{Var}(y^*)\ {}^*) = \beta \text{Var}(x) \beta + 1$
- Where Var(x)=Covariance matrix of xs from the real data.

Full Standardization

$$Full\text{-}standardization$$

$$y = a + b_1 x_1 + b_2 x_2 + \cdots + b_k x_k + e$$

$$where$$

$$e = error,\ disturbance,\ innovation,\ shock$$

$$we\ divide\ each\ x_k\ by\ \sigma_k\ and\ multiply\ the\ b_k\ by\ that\ quantity:$$

$$\frac{y}{\sigma_y} = \frac{a}{\sigma_y} + \sigma_1 b_1 \frac{x_1}{\sigma_y} + \sigma_2 b_2 \frac{x_2}{\sigma_y} + \cdots + \sigma_k b_k \frac{x_k}{\sigma_y} + \frac{e}{\sigma_y},\quad so\ the\ x - stdzed$$

$$coefficient =$$

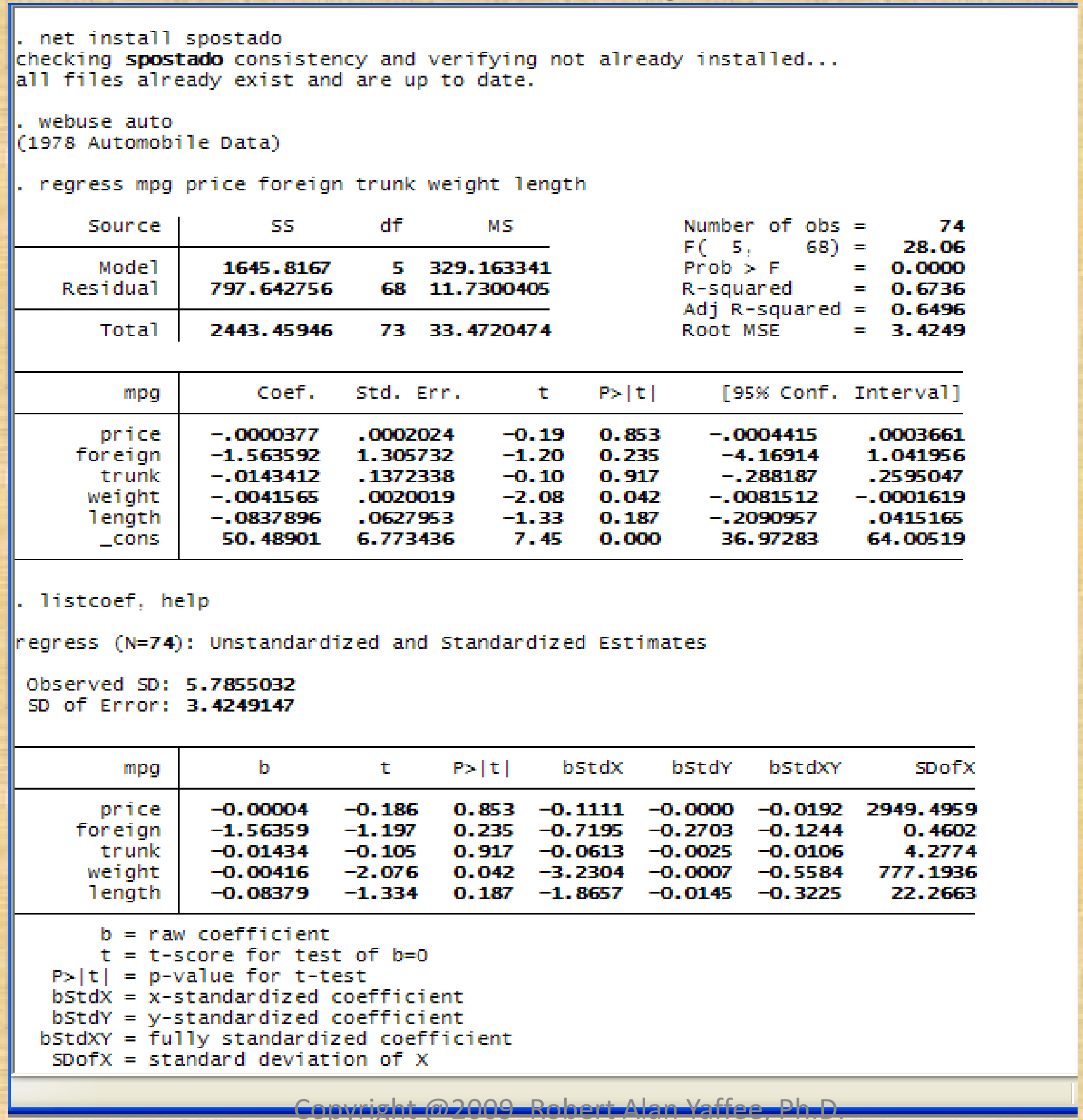
$$\beta_1^s = \sigma_1 b_1 \frac{x_1}{\sigma_y}$$

Interpretation of Full Standardization

Ibid.

- After full standardization, the interpretation of change of the regression coefficient in such a model is:
- “ For a standard deviation increase in xi, y is expected to change by β_i^s standard deviations, while holding all other variables constant.”
- Type: listcoef after running the regression analysis using OLS.

net install spostado



```
. net install spostado
checking spostado consistency and verifying not already installed...
all files already exist and are up to date.

. webuse auto
(1978 Automobile Data)

. regress mpg price foreign trunk weight length

      Source |       SS       df       MS              Number of obs =      74
-------------+------------------------------           F(  5,    68) =   28.06
       Model |  1645.8167     5  329.163341           Prob > F      =  0.0000
    Residual |  797.642756    68  11.7300405           R-squared     =  0.6736
-------------+------------------------------           Adj R-squared =  0.6496
       Total |  2443.45946    73  33.4720474           Root MSE      =  3.4249

------------------------------------------------------------------------------
         mpg |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       price |  -.0000377   .0002024    -0.19   0.853    -.0004415    .0003661
     foreign |  -1.563592   1.305732    -1.20   0.235     -4.16914    1.041956
       trunk |  -.0143412   .1372338    -0.10   0.917     -.288187    .2595047
      weight |  -.0041565   .0020019    -2.08   0.042    -.0081512   -.0001619
      length |  -.0837896   .0627953    -1.33   0.187    -.2090957    .0415165
       _cons |   50.48901   6.773436     7.45   0.000     36.97283    64.00519
------------------------------------------------------------------------------

. listcoef, help

regress (N=74): Unstandardized and Standardized Estimates

 Observed SD: 5.7855032
 SD of Error: 3.4249147

-------------------------------------------------------------------------------
         mpg |      b         t     P>|t|    bStdX    bStdY   bStdXY      SDofX
-------------+-----------------------------------------------------------------
       price |  -0.00004   -0.186   0.853  -0.1111  -0.0000  -0.0192  2949.4959
     foreign |  -1.56359   -1.197   0.235  -0.7195  -0.2703  -0.1244     0.4602
       trunk |  -0.01434   -0.105   0.917  -0.0613  -0.0025  -0.0106     4.2774
      weight |  -0.00416   -2.076   0.042  -3.2304  -0.0007  -0.5584   777.1936
      length |  -0.08379   -1.334   0.187  -1.8657  -0.0145  -0.3225    22.2663
-------------------------------------------------------------------------------
       b = raw coefficient
       t = t-score for test of b=0
   P>|t| = p-value for t-test
   bStdX = x-standardized coefficient
   bStdY = y-standardized coefficient
  bStdXY = fully standardized coefficient
   SDofX = standard deviation of X
```

Covariance

$$Covariance = \frac{\sum_{i=1}^{n}\left(x_i - \overline{x}\right)(y_i - \overline{y})}{n}$$

$$Cov(Xa) = 0$$

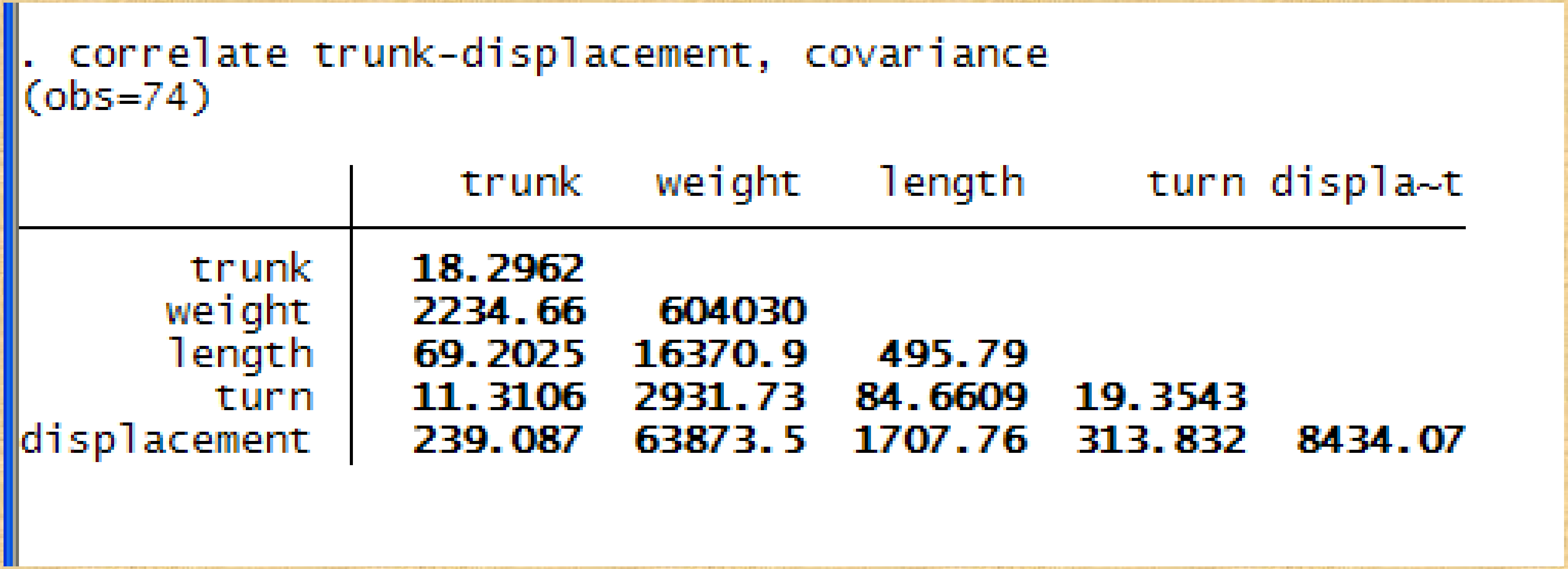
$$Cov(X'X) = VAR(X)$$

$$COV(X'Y) = E(XY) = E\left(x_i - \overline{x}\right)(y_i - \overline{y})$$

$$= E(x_i y_i) - E(\overline{x} y_i) - E(x_i \overline{y}) + E(\overline{xy})$$

$$= E(x_i y_i)$$

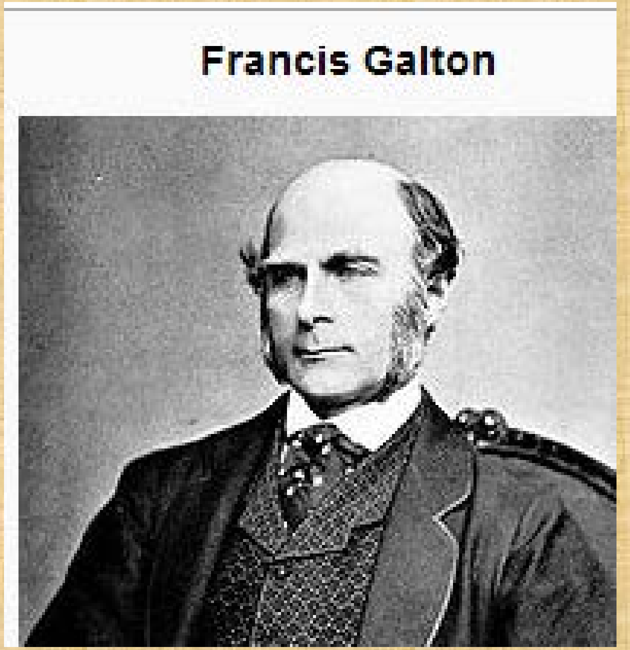
Covariances in Stata



```
. correlate trunk-displacement, covariance
(obs=74)

             |    trunk   weight   length     turn displa~t
-------------+---------------------------------------------
       trunk |  18.2962
      weight |  2234.66   604030
      length |  69.2025  16370.9   495.79
        turn |  11.3106  2931.73  84.6609  19.3543
displacement |  239.087  63873.5  1707.76  313.832  8434.07
```

Francis Galton

- Invented the correlation coefficient and laid the groundwork for regression analysis.



Karl Pearson

He read mathematics at Cambridge University in the latter 19th Century. His name is attached to the Chi-square goodness of fit test and the Pearson correlation coefficient. Francis Galton invented the correlation coefficient, but it was named after Karl Pearson.

Pearson product-moment correlations

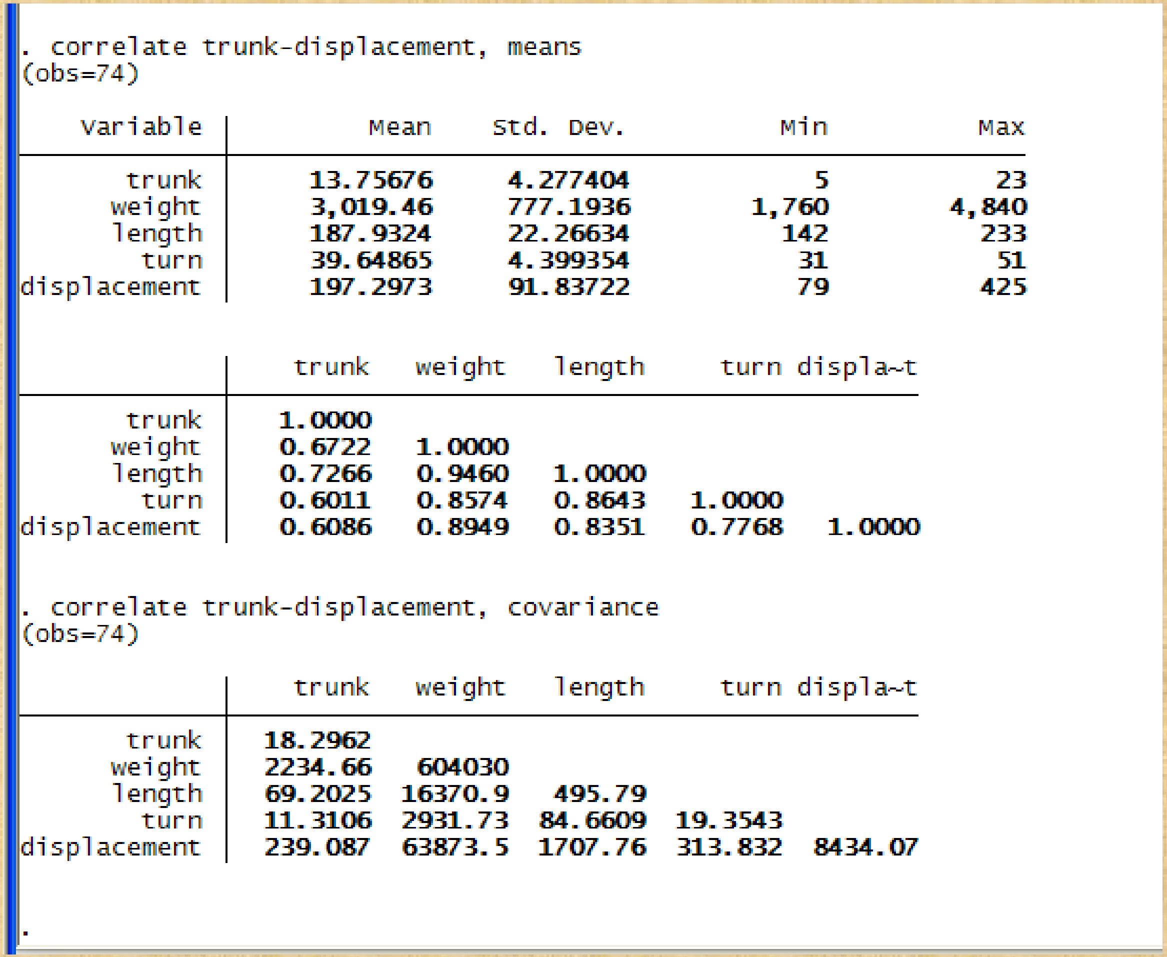
- Used when both variables are continuous or highly ordinal (with 15 or more levels)

$$r_{pearson} = \frac{\dfrac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{n}}{\sqrt{\dfrac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n}\dfrac{\sum(y_i-\bar{y})^2}{n}}} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2\sum(y_i-\bar{y})^2}}$$

$$Corr(XY) = \frac{Cov(X'Y)}{S(X)S(Y)} = \frac{E(X'Y)}{\sqrt{E(X'X)E(Y'Y)}}$$

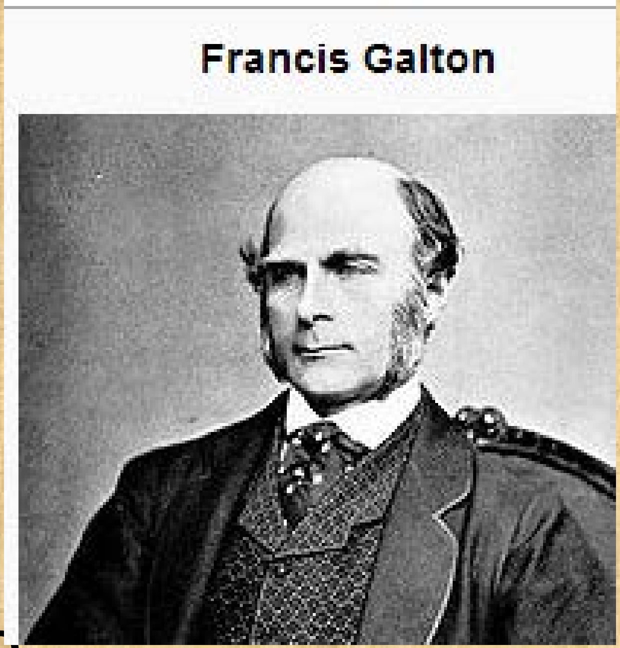
$$\therefore Correlations\ are\ standardized\ covariances$$

Covariances and Correlations



```
. correlate trunk-displacement, means
(obs=74)

    Variable |       Mean    Std. Dev.          Min          Max
-------------+----------------------------------------------------
       trunk |   13.75676    4.277404            5           23
      weight |   3,019.46    777.1936        1,760        4,840
      length |   187.9324    22.26634          142          233
        turn |   39.64865    4.399354           31           51
displacement |   197.2973    91.83722           79          425


             |    trunk   weight   length     turn displa~t
-------------+---------------------------------------------
       trunk |   1.0000
      weight |   0.6722   1.0000
      length |   0.7266   0.9460   1.0000
        turn |   0.6011   0.8574   0.8643   1.0000
displacement |   0.6086   0.8949   0.8351   0.7768   1.0000

. correlate trunk-displacement, covariance
(obs=74)

             |    trunk   weight   length     turn displa~t
-------------+---------------------------------------------
       trunk |  18.2962
      weight |  2234.66   604030
      length |  69.2025  16370.9   495.79
        turn |  11.3106  2931.73  84.6609  19.3543
displacement |  239.087  63873.5  1707.76  313.832  8434.07


.
```

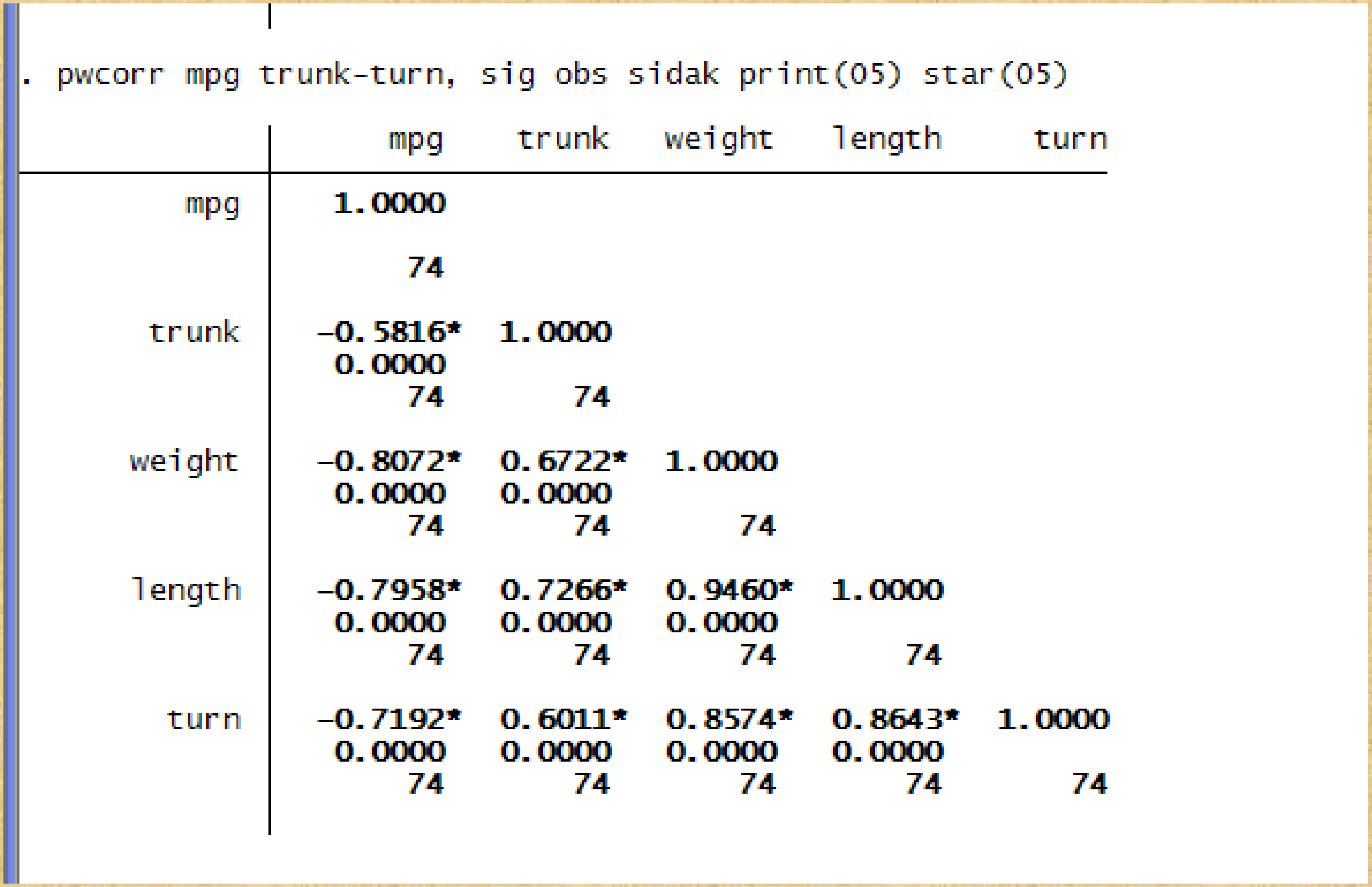
Francis Galton: the father of Correlations



correlate computes listwise correlations

pwcorr computes pairwise correlations, though there is a listwise option. You can also get nobs and sig as options for this command.

Pairwise correlations



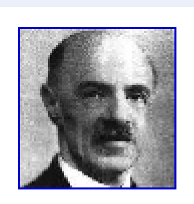
```
. pwcorr mpg trunk-turn, sig obs sidak print(05) star(05)

             |      mpg    trunk   weight   length     turn
-------------+---------------------------------------------
         mpg |   1.0000 
             |
             |       74
             |
       trunk |  -0.5816*  1.0000 
             |   0.0000
             |       74       74
             |
      weight |  -0.8072*  0.6722*  1.0000 
             |   0.0000   0.0000
             |       74       74       74
             |
      length |  -0.7958*  0.7266*  0.9460*  1.0000 
             |   0.0000   0.0000   0.0000
             |       74       74       74       74
             |
        turn |  -0.7192*  0.6011*  0.8574*  0.8643*  1.0000 
             |   0.0000   0.0000   0.0000   0.0000
             |       74       74       74       74       74
             |
```

Properties of Pearson Correlations

- They measure only the significance, direction, and strength of linear relationships. They are not designed to work with binary or ordinal variables.
- If the relationship is quadratic or mostly nonlinear, these correlations may not detect them.
- Therefore, do scattergrams between the two variables first.
- Then do a lowess plot to detect nonlinearity in the relationship.

Charles Spearman’s ρ correlation for ordinal variables

Stata Release 10 Reference Manual Q-Z, (2007). College Station, Tx: StataCorp, 321.



- Spearman’s rho was named after Charles Spearman, who used ranks to compute the correlation formula and handled ties with average ranks of the ordinal variables.

$$Spearman's\ \rho_{yx} = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$

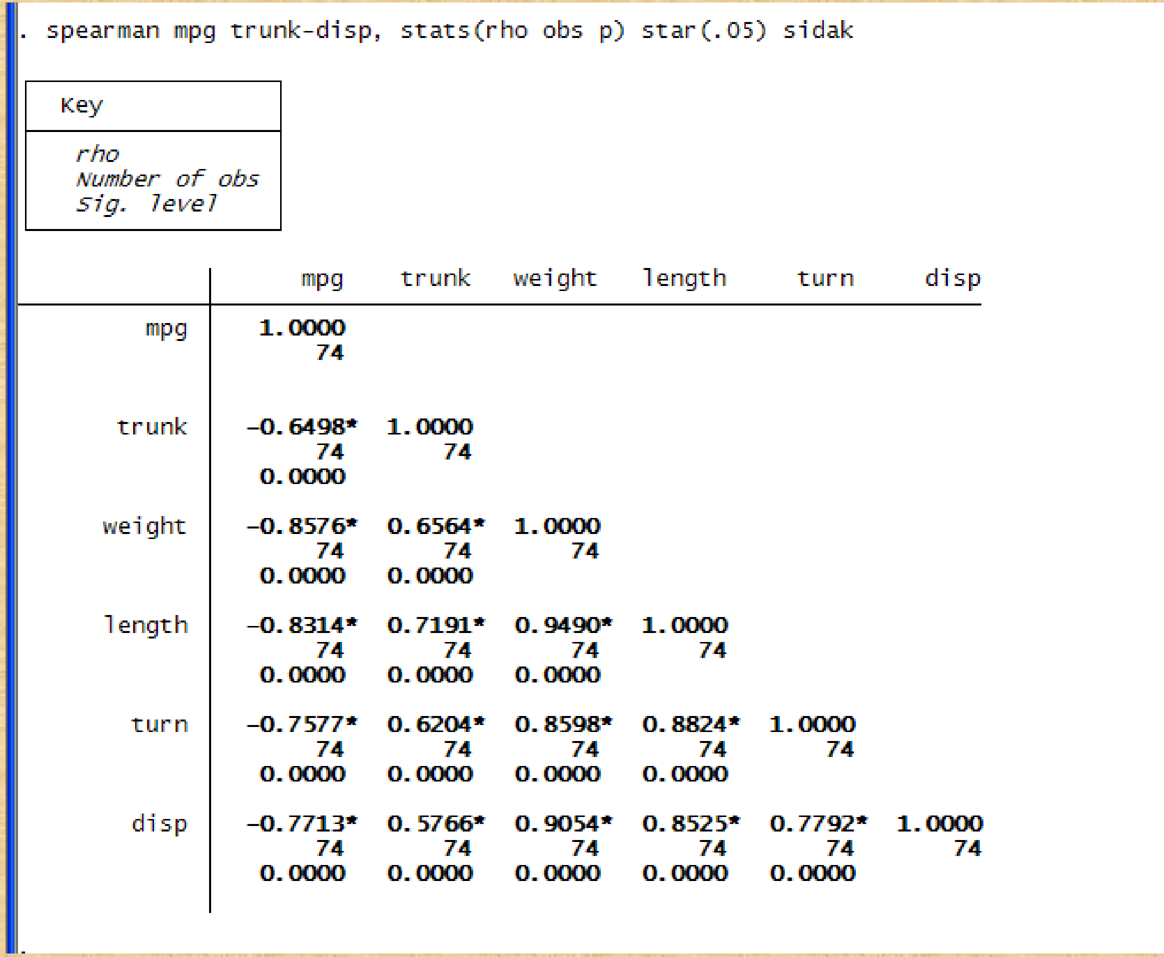
$$\begin{aligned}&where\\&d_i = difference\ between\ ranks\ of\ corresponding\\&\qquad values\ of\ X_i\ and\ Y_i\end{aligned}$$

Significance testing.

Significance tested with

$$p = 2 * ttail\left(n-2, \frac{|\hat{\rho}|\sqrt{n-2}}{\sqrt{1-\hat{\rho}^2}}\right)$$

Spearman correlations for ordinal variables



```
. spearman mpg trunk-disp, stats(rho obs p) star(.05) sidak

+-----------------+
|   Key           |
|-----------------|
|     rho         |
|     Number of obs |
|     Sig. level  |
+-----------------+

             |      mpg    trunk   weight   length     turn     disp
-------------+------------------------------------------------------
         mpg |   1.0000
             |       74
             |
             |
       trunk |  -0.6498*  1.0000
             |       74       74
             |   0.0000
             |
      weight |  -0.8576*  0.6564*  1.0000
             |       74       74       74
             |   0.0000   0.0000
             |
      length |  -0.8314*  0.7191*  0.9490*  1.0000
             |       74       74       74       74
             |   0.0000   0.0000   0.0000
             |
        turn |  -0.7577*  0.6204*  0.8598*  0.8824*  1.0000
             |       74       74       74       74       74
             |   0.0000   0.0000   0.0000   0.0000
             |
        disp |  -0.7713*  0.5766*  0.9054*  0.8525*  0.7792*  1.0000
             |       74       74       74       74       74       74
             |   0.0000   0.0000   0.0000   0.0000   0.0000
             |

.
```

Sir Maurice George Kendall's rank correlations : Tau a and Tau b

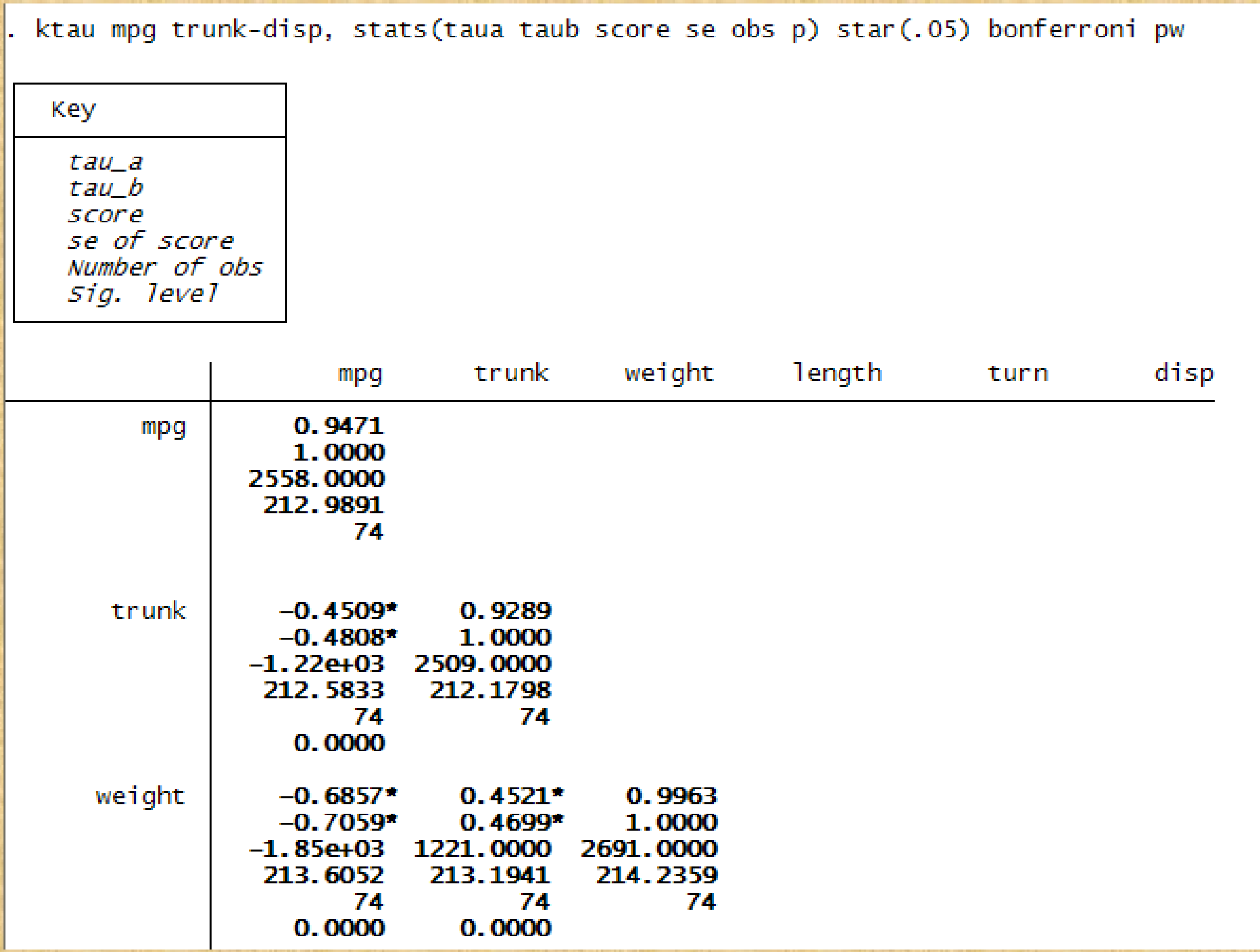
- Stata Reference Release 10, Manual R-Z, StataCorp, College Station, Tx: 321-322.

$$\tau_a = \frac{\sum (C-D)}{n(n-2)/2}$$

$$\tau_b = \frac{\sum (C-D)}{\sqrt{N-U}\sqrt{N-V}}$$

$$\begin{aligned}
&where\\
&N = n(n-2)/2\\
&U = \sum_{i=1}^{N_1} u_i(u_i - 1)/2 \ \ with\ \ N_1 = \#sets\ of\ \ tied\ \ x\,values\\
&\qquad\qquad\qquad\qquad u_i = \#tied\ \ x\,values\,in\ the\ ith\ set\\
&V = \sum_{j=1}^{N_2} v_j(v_i - 1)/2 \ \ with\ \ N_2 = \#sets\ of\ \ tied\ \ y\ values\\
&\qquad\qquad\qquad\qquad v_j = \#\ tied\ \ y\ values\ in\ the\ \ jth\ set.
\end{aligned}$$

Kendall’s correlation for ordinal variables (*Ibid*)



```
. ktau mpg trunk-disp, stats(taua taub score se obs p) star(.05) bonferroni pw

+-------------------+
|  Key              |
|-------------------|
|    tau_a          |
|    tau_b          |
|    score          |
|    se of score    |
|    Number of obs  |
|    Sig. level     |
+-------------------+

             |        mpg      trunk     weight     length       turn       disp
-------------+------------------------------------------------------------------
         mpg |     0.9471
             |     1.0000
             |  2558.0000
             |   212.9891
             |         74
             |
             |
       trunk |    -0.4509*    0.9289
             |    -0.4808*    1.0000
             |  -1.22e+03  2509.0000
             |   212.5833   212.1798
             |         74         74
             |     0.0000
             |
      weight |    -0.6857*    0.4521*    0.9963
             |    -0.7059*    0.4699*    1.0000
             |  -1.85e+03  1221.0000  2691.0000
             |   213.6052   213.1941   214.2359
             |         74         74         74
             |     0.0000     0.0000
```

Kendall’s Corr significance tests

- Stata Base Release 10 , 2007, Reference Manual, Q-Z, StataCorp, College Station, Tx, 322.

$$Assume \ \ S = \sum C - \sum D$$

$$z = \frac{|S|}{\sqrt{Var(S)}} \quad or \quad z = \frac{|S| - 1}{\sqrt{Var(S)}} \quad if \ \ a \ \ continuity \ \ correction \ \ is \ \ desired$$

$$where$$

$$Var(S) = \frac{1}{18}\left\{ n(n-1)(2n+5) - \sum_{i=1}^{N_1} u_i(u_i - 1)(2u_i + 5) - \sum_{i=1}^{N_2} v_i(v_i - 1)(2v_i + 5) \right\}$$

$$+ \frac{1}{9n(n-1)(n-2}\left\{ \sum_{i=1}^{N_1} u_i(u_i - 1)(u_i - 2) - \sum_{i=1}^{N_2} v_i(v_i - 1)(v_i - 2) \right\}$$

$$+ \frac{1}{2n(n-1)}\left\{ \left\{ \sum_{i=1}^{N_1} u_i(u_i - 1) \right\} \left\{ \sum_{i=1}^{N_2} v_i(v_i - 1) \right\} \right\}$$

Tetrachoric Correlations for Binary Variables.

Stata Release 10 Base Reference Manual (2007). College Station, Tx.: StataCorp, 480.

$$\rho_{tetrachoric} = \frac{\alpha - 1}{\alpha + 1}$$

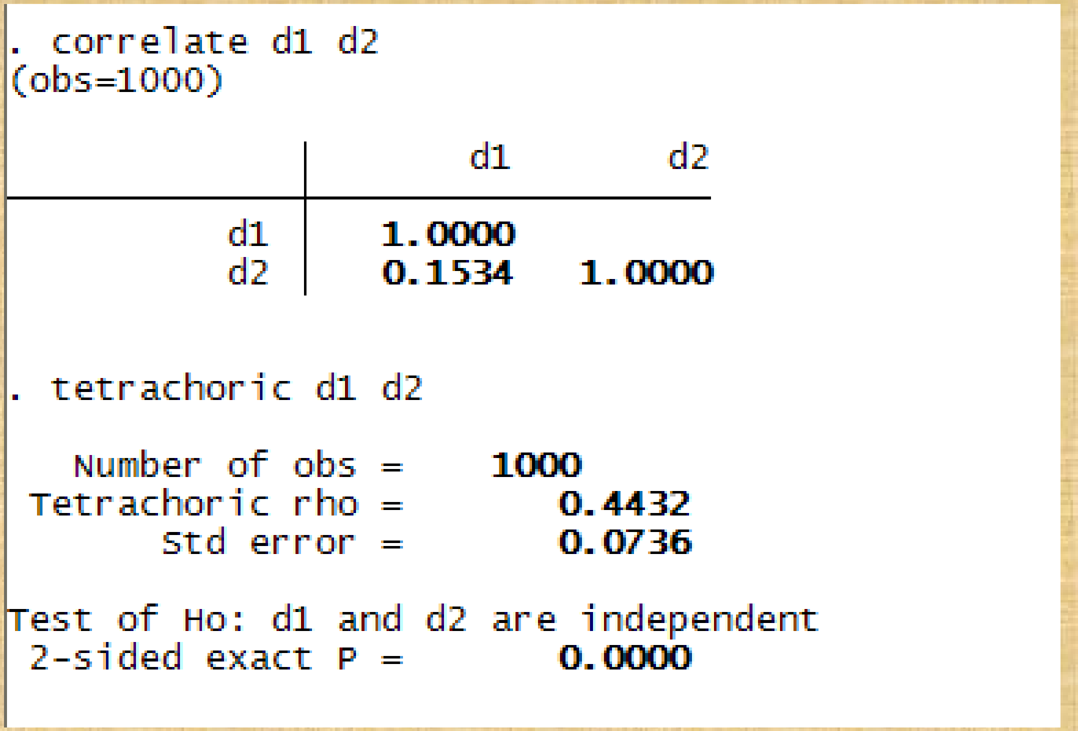
where

$$\alpha = \left(\frac{n_{00}\, n_{11}}{n_{01} n_{10}} \right)^{\pi/4}$$

$$avar(\hat{\rho}) = \left(\frac{\pi\alpha}{2(1+\alpha)^2} \right)^2 \left(\frac{1}{n_{00}} + \frac{1}{n_{01}} + \frac{1}{n_{10}} + \frac{1}{n_{11}} \right)$$

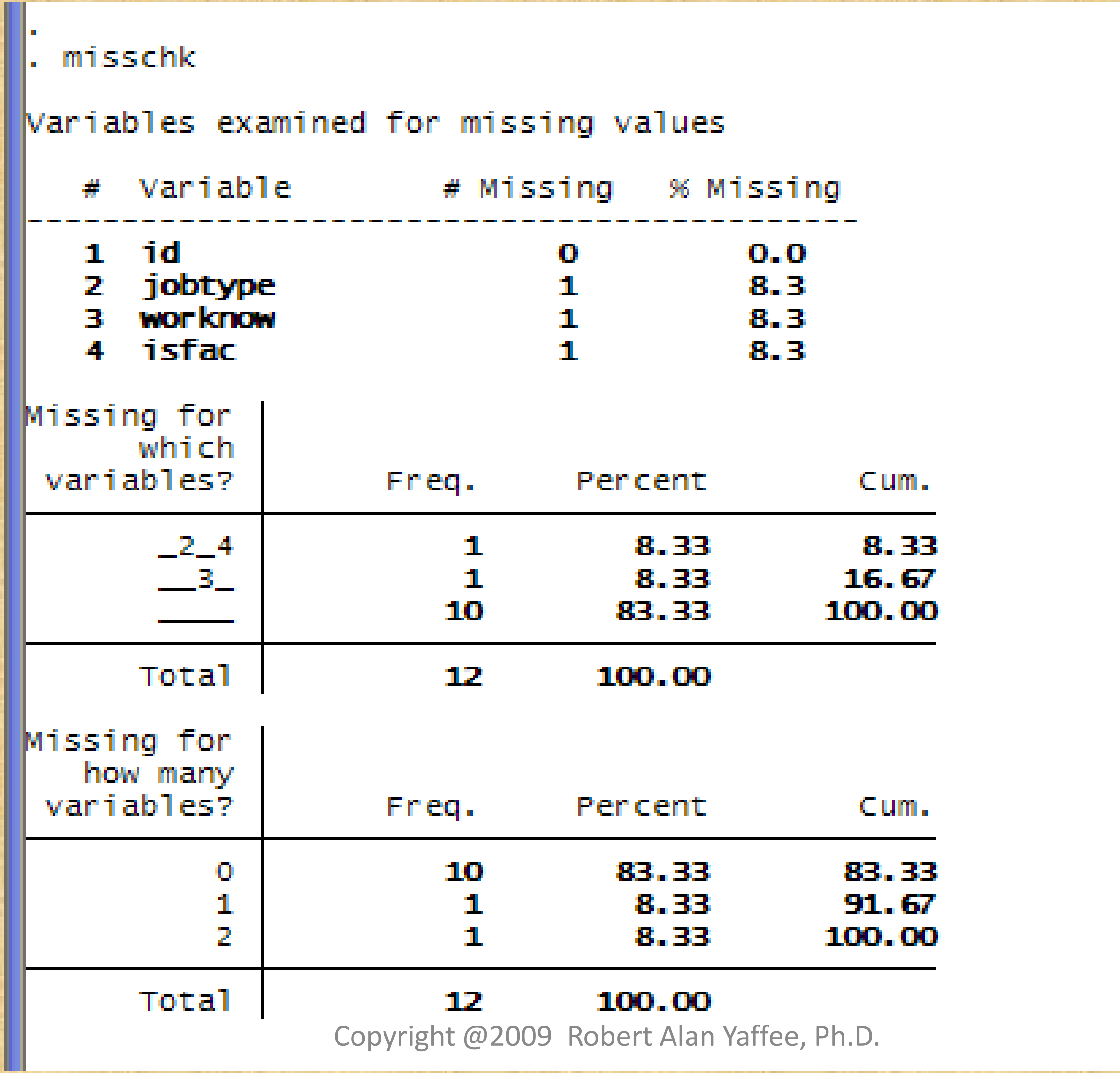
$$all\ \ n_{i\,j} > 0$$

Tetrachoric correlations for binary variables



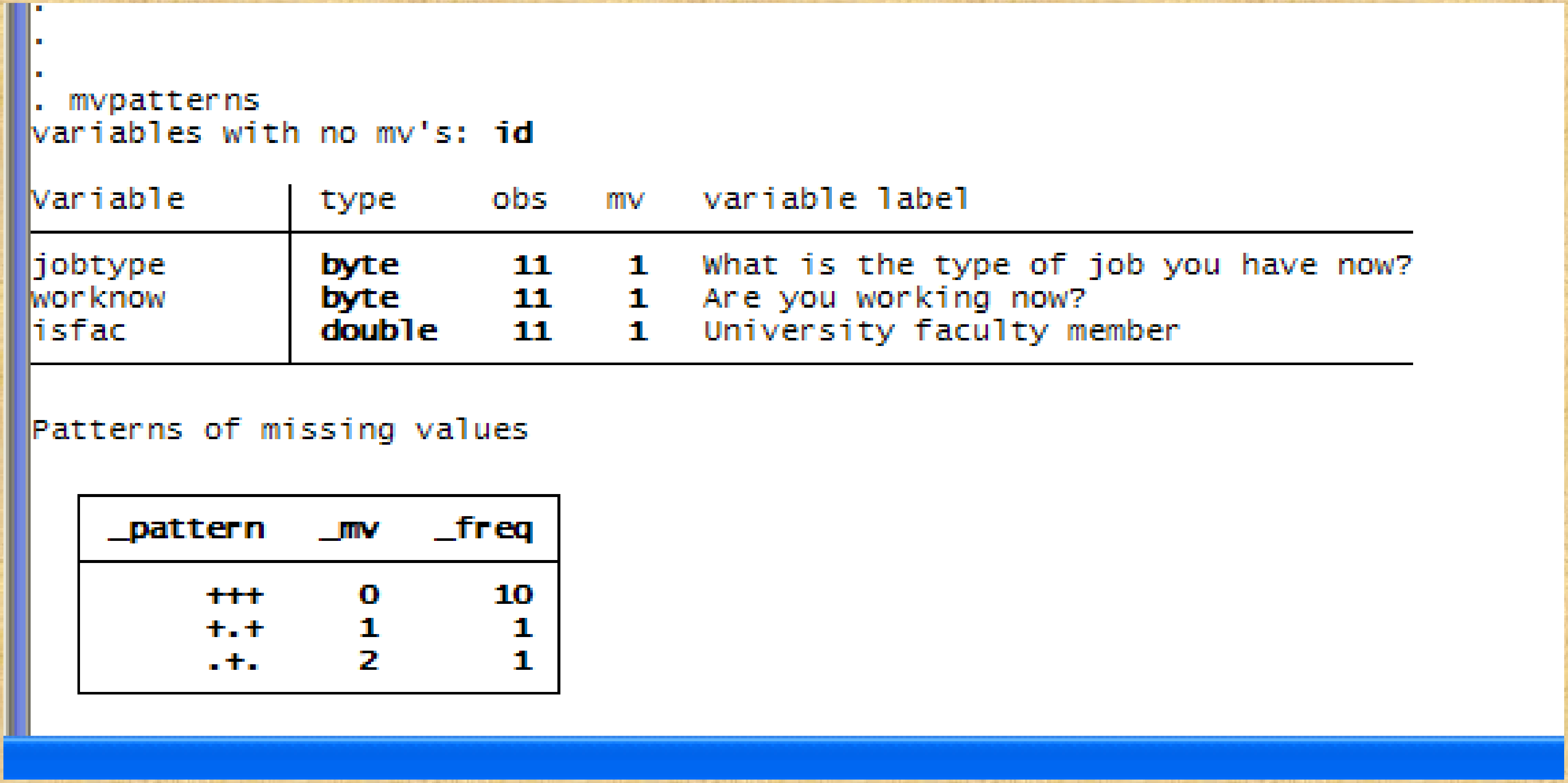
```
. correlate d1 d2
(obs=1000)

             |       d1       d2
-------------+------------------
          d1 |   1.0000
          d2 |   0.1534   1.0000


. tetrachoric d1 d2

   Number of obs =     1000
 Tetrachoric rho =        0.4432
       Std error =        0.0736

Test of Ho: d1 and d2 are independent
 2-sided exact P =        0.0000
```

Checking the dataset for missing values



```
.
. misschk

Variables examined for missing values

   #  Variable        # Missing   % Missing
--------------------------------------------
   1  id                   0         0.0
   2  jobtype              1         8.3
   3  worknow              1         8.3
   4  isfac                1         8.3

Missing for |
      which |
 variables? |      Freq.     Percent        Cum.
------------+-----------------------------------
       _2_4 |          1        8.33        8.33
       __3_ |          1        8.33       16.67
       ____ |         10       83.33      100.00
------------+-----------------------------------
      Total |         12      100.00

Missing for |
   how many |
 variables? |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |         10       83.33       83.33
          1 |          1        8.33       91.67
          2 |          1        8.33      100.00
------------+-----------------------------------
      Total |         12      100.00
```

Detecting missing value patterns

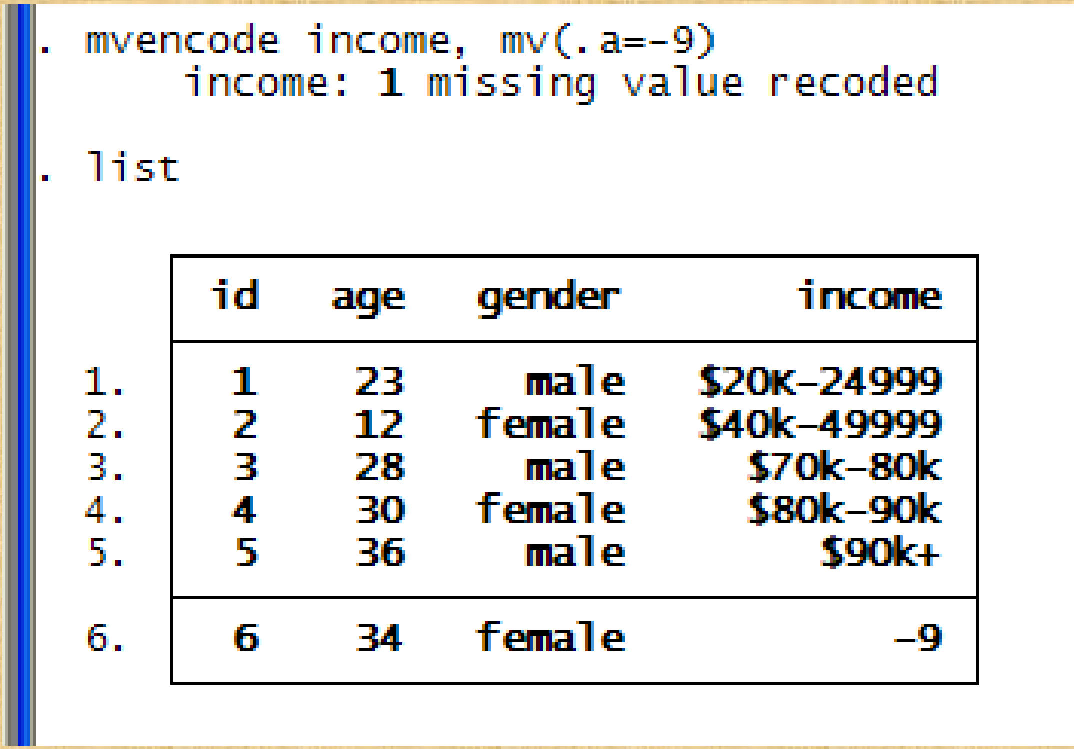


```
.
.
.
. mvpatterns
variables with no mv's: id

Variable     | type      obs   mv   variable label
-------------+---------------------------------------------------------
jobtype      | byte       11    1   What is the type of job you have now?
worknow      | byte       11    1   Are you working now?
isfac        | double     11    1   University faculty member
-----------------------------------------------------------------------

Patterns of missing values

  +------------------------+
  | _pattern   _mv   _freq |
  |------------------------|
  |      +++     0      10 |
  |      +.+     1       1 |
  |      .+.     2       1 |
  +------------------------+
```

Recoding missing values

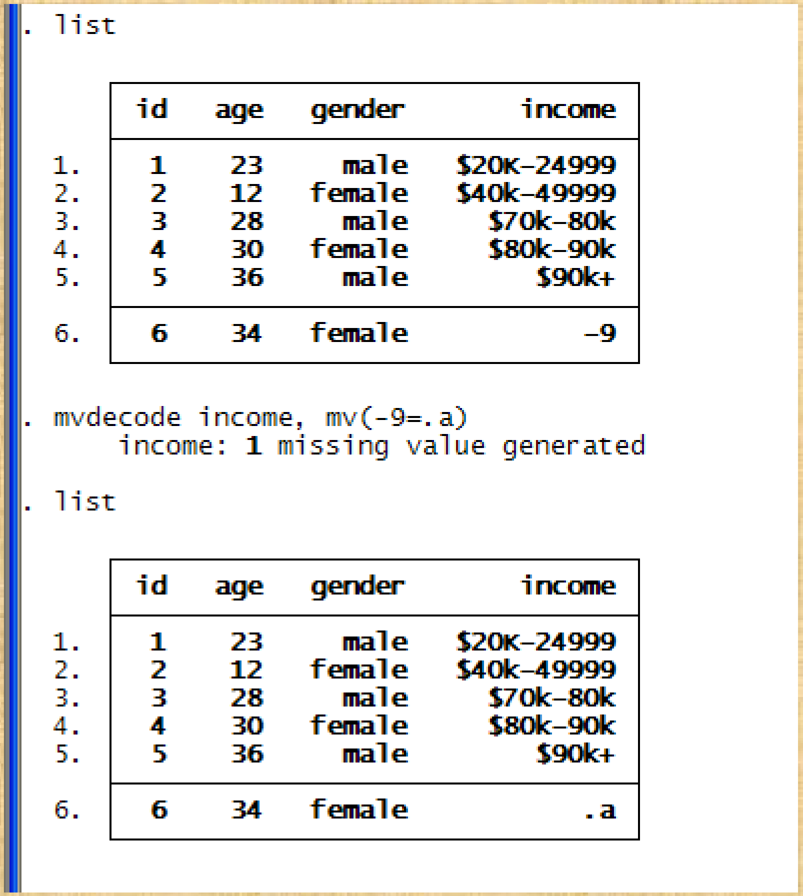
- mvdecode and mvencode commands
- mvdecode permits you to recode various values of a variable to missing.
- mvencode permits you to recode missing values to a nonmissing value. For example: mvencode income, mv(-9=.a)

mvencode: converting from special to numeric missing value codes



```
. mvencode income, mv(.a=-9)
     income: 1 missing value recoded

. list

     +----------------------------------+
     | id   age   gender         income |
     |----------------------------------|
  1. |  1    23     male   $20K-24999   |
  2. |  2    12   female   $40k-49999   |
  3. |  3    28     male     $70k-80k   |
  4. |  4    30   female     $80k-90k   |
  5. |  5    36     male        $90k+   |
     |----------------------------------|
  6. |  6    34   female           -9   |
     +----------------------------------+
```

Mvdecoding: converting from one to another missing value codes



```
. list

     +---------------------------------------+
     | id   age   gender              income |
     |---------------------------------------|
  1. |  1    23     male        $20K-24999 |
  2. |  2    12   female        $40k-49999 |
  3. |  3    28     male          $70k-80k |
  4. |  4    30   female          $80k-90k |
  5. |  5    36     male             $90k+ |
     |---------------------------------------|
  6. |  6    34   female                -9 |
     +---------------------------------------+

. mvdecode income, mv(-9=.a)
      income: 1 missing value generated

. list

     +---------------------------------------+
     | id   age   gender              income |
     |---------------------------------------|
  1. |  1    23     male        $20K-24999 |
  2. |  2    12   female        $40k-49999 |
  3. |  3    28     male          $70k-80k |
  4. |  4    30   female          $80k-90k |
  5. |  5    36     male             $90k+ |
     |---------------------------------------|
  6. |  6    34   female                .a |
     +---------------------------------------+
```

Missing value replacement

- Stata can perform multiple imputation the its mice procedure developed by Patrick Royston.
- It is available as a free download from Stata Software Components archive
- ssc install mice can be typed on the command line.

Variable transformation: Recoding variables

```
recode income (1/3=1)(4/6=2)(7/12=3)
```

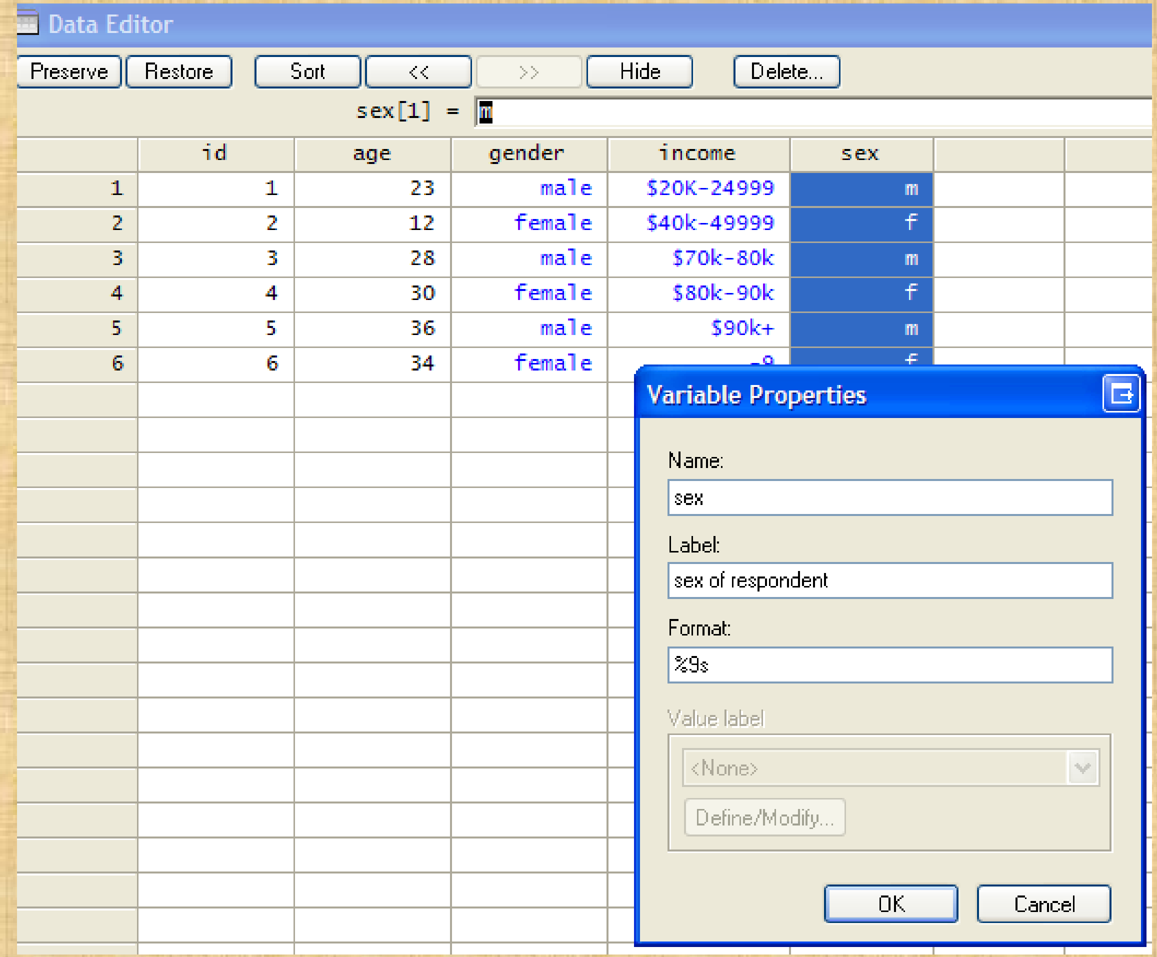
or

```
generate incgrp = 1
replace incgrp=2 if income > 3 | income < 8
replace incgrp = 3 if income > 6 & income < .
```

Variable formats

- String: alpha %9s
- Numeric: numeric %8.2g
- Date : day %td, week %tw, month %tm, quarter %tq, year %ty
- Panel: it where i=group and t = date

Variable format conversion: from string to numeric



sex[1] = m

	id	age	gender	income	sex
1	1	23	male	$20K-24999	m
2	2	12	female	$40k-49999	f
3	3	28	male	$70k-80k	m
4	4	30	female	$80k-90k	f
5	5	36	male	$90k+	m
6	6	34	female	-9	f

Variable Properties

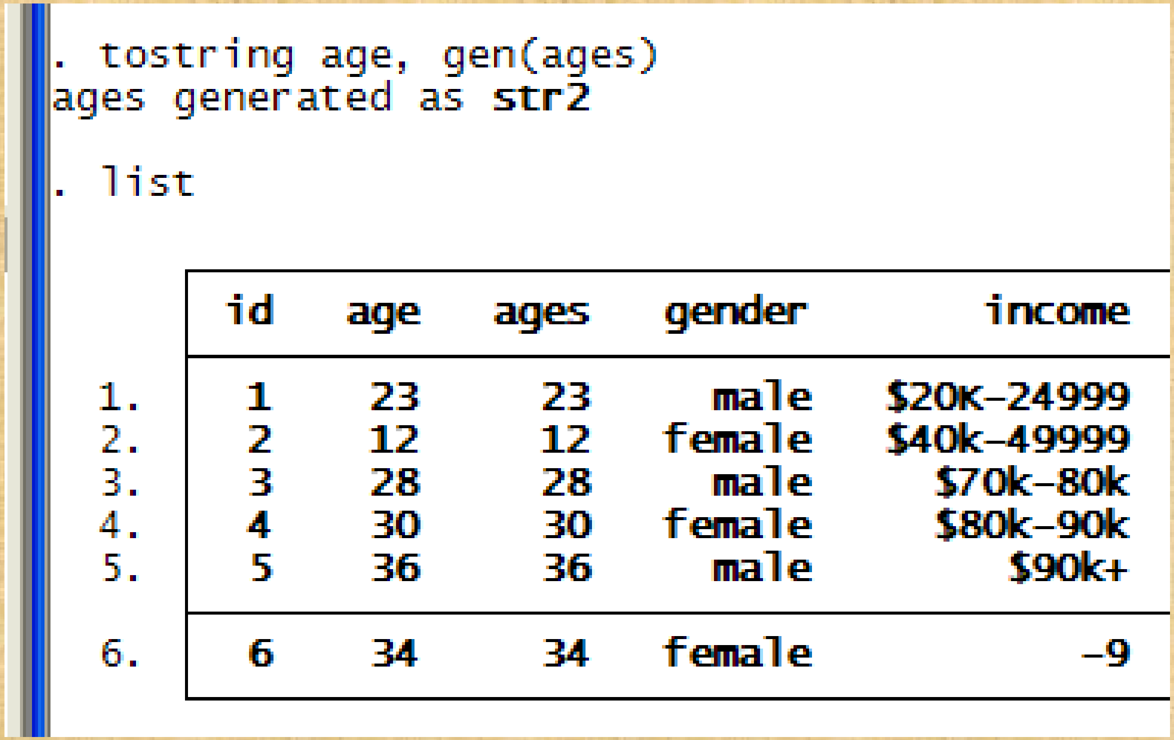
Name: sex

Label: sex of respondent

Format: %9s

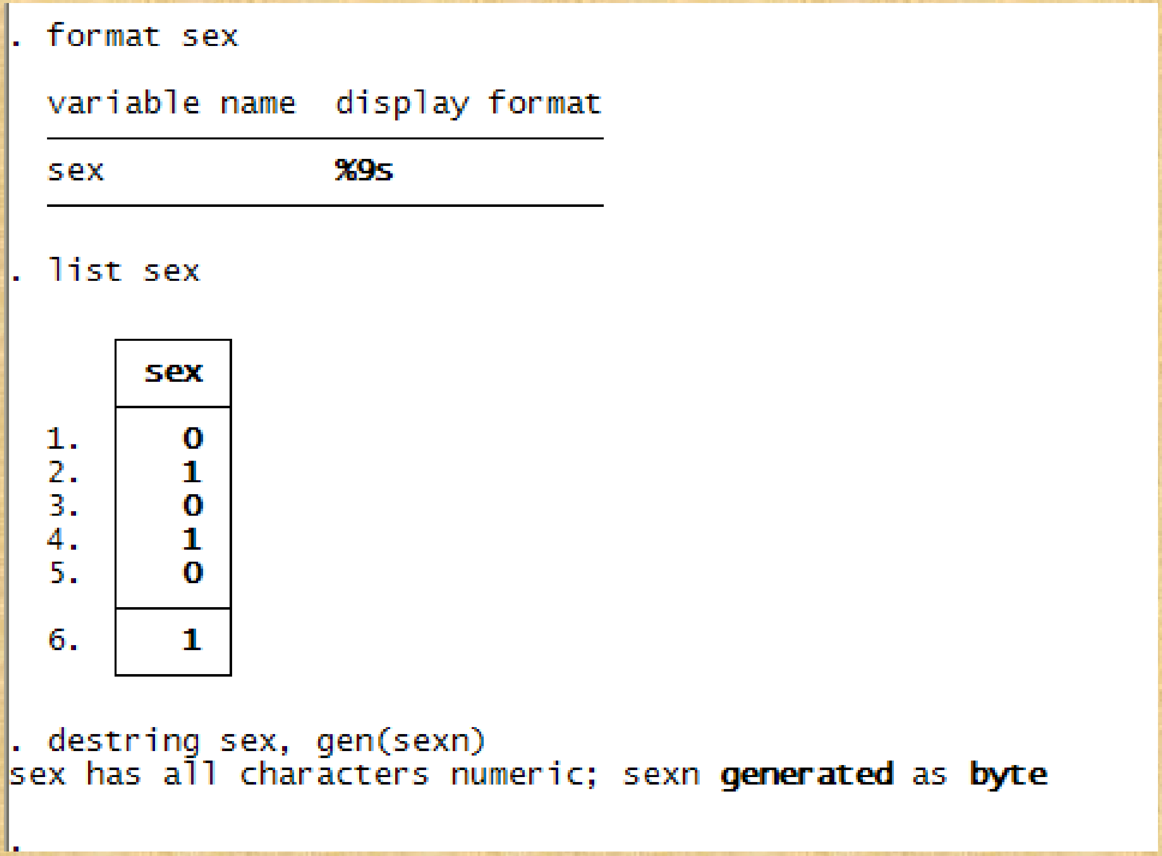
Value label: <None>

Converting a numeric to a string variable



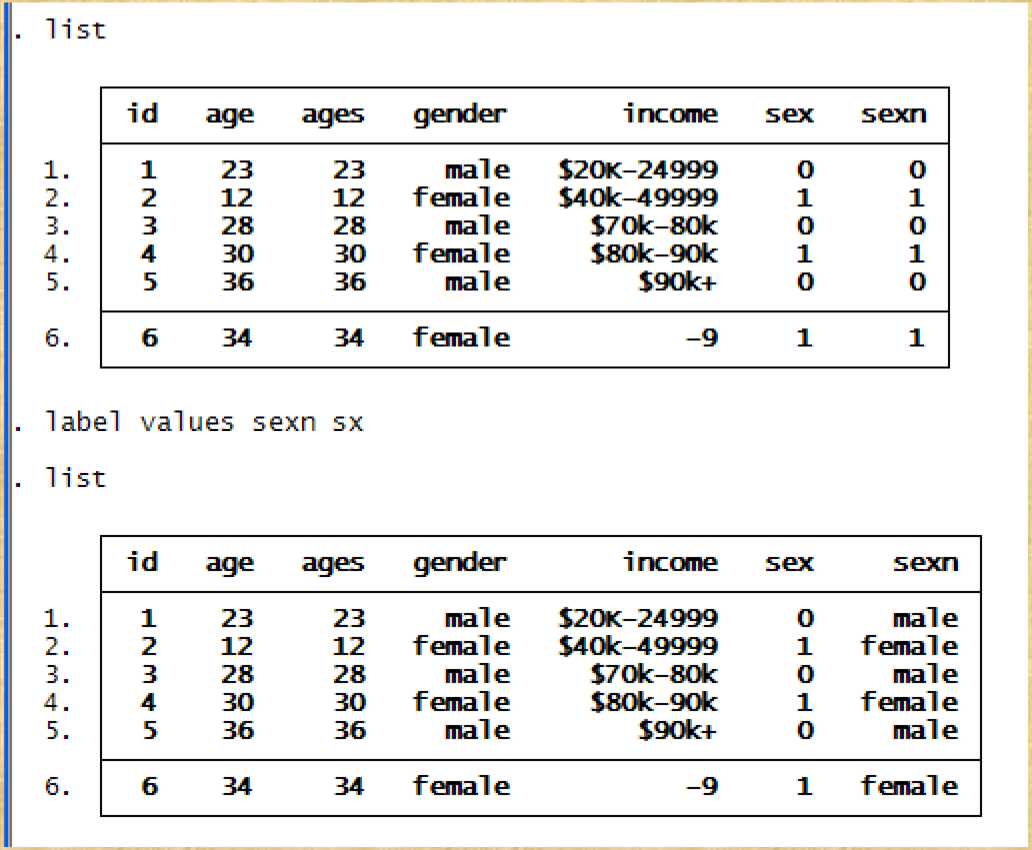
```
. tostring age, gen(ages)
ages generated as str2

. list

     +---------------------------------------------
     | id   age   ages   gender         income
     |---------------------------------------------
  1. |  1    23     23     male     $20K-24999
  2. |  2    12     12   female     $40k-49999
  3. |  3    28     28     male       $70k-80k
  4. |  4    30     30   female       $80k-90k
  5. |  5    36     36     male          $90k+
     |---------------------------------------------
  6. |  6    34     34   female             -9
     +---------------------------------------------
```

Variable format conversion: converting numeric to string



```
. format sex

  variable name   display format
  ------------------------------
  sex             %9s
  ------------------------------

. list sex

       +-----+
       | sex |
       |-----|
    1. |   0 |
    2. |   1 |
    3. |   0 |
    4. |   1 |
    5. |   0 |
       |-----|
    6. |   1 |
       +-----+

. destring sex, gen(sexn)
sex has all characters numeric; sexn generated as byte

.
```

Variable transformation: converting string to numeric variables



```
. list

     ┌──────────────────────────────────────────────────────────┐
     │ id   age   ages   gender        income   sex   sexn     │
     ├──────────────────────────────────────────────────────────┤
  1. │  1    23     23     male   $20K-24999     0      0      │
  2. │  2    12     12   female   $40k-49999     1      1      │
  3. │  3    28     28     male     $70k-80k     0      0      │
  4. │  4    30     30   female     $80k-90k     1      1      │
  5. │  5    36     36     male        $90k+     0      0      │
     ├──────────────────────────────────────────────────────────┤
  6. │  6    34     34   female           -9     1      1      │
     └──────────────────────────────────────────────────────────┘

. label values sexn sx

. list

     ┌─────────────────────────────────────────────────────────────┐
     │ id   age   ages   gender        income   sex      sexn      │
     ├─────────────────────────────────────────────────────────────┤
  1. │  1    23     23     male   $20K-24999     0      male      │
  2. │  2    12     12   female   $40k-49999     1    female      │
  3. │  3    28     28     male     $70k-80k     0      male      │
  4. │  4    30     30   female     $80k-90k     1    female      │
  5. │  5    36     36     male        $90k+     0      male      │
     ├─────────────────────────────────────────────────────────────┤
  6. │  6    34     34   female           -9     1    female      │
     └─────────────────────────────────────────────────────────────┘
```

Exercises 1

1. Get help on the egen command within Stata
2. Use the findit command to obtain help on a keyword of interest
3. Get help on datasets available
4. Download from the web the lifeexp.dta dataset
5. Describe the dataset
6. Use the inspect command to check for missing values
7. Examine the variables for missing values
8. Give the variable safewater a variable label.
9. Do a frequencies analysis on the variable, safewater
10. List the countries with 100% safewater

Exercises 1 continued

1. List countries with less than 75% safewater
2. List countries with more than 75% and less than 96% safewater
3. List countries with less than 10% population growth
4. Download bpwide.dta from web
5. Crosstabulate sex and agegroup (show counts, row and column percentages)
6. What is the Pearson correlation between the blood pressure before and after?
7. Is this a significant correlation?
8. Is this a linear relationship?
9. Construct your own dataset with 3 discrete variables and 2 continuous variables with 5 observations. Label the variables and the values of the discrete variables. Tabulate the variables. Crosstabulate two of the discrete variables and obtain a chi-square test for significance between them. If they are ordinal obtain a Gamma and a Kendall's tau a correlation between them.
10. Construct a variable that gives the row average of three of your variables in that dataset. Be sure that this variable does not use missing values.

A comment by Hadamard

- Hadamard, Jacques

- The shortest path between two truths in the real domain passes through the complex domain.
- Quoted in The Mathematical Intelligencer, v. 13, no. 1, Winter 1991.

File management

- Codebook construction
    - inspection
    - recoded variables
    - tabulations
    - summaries
    - missing value analysis
    - basic histograms and boxplots
- File merging
- File appending
- File conversion
    - from wide to long
    - from long to wide
- Construction of special files
    - Time series datasets
    - Panel datasets
    - Survival datasets
    - complex survey analysis
- Do Files
- Ado Files

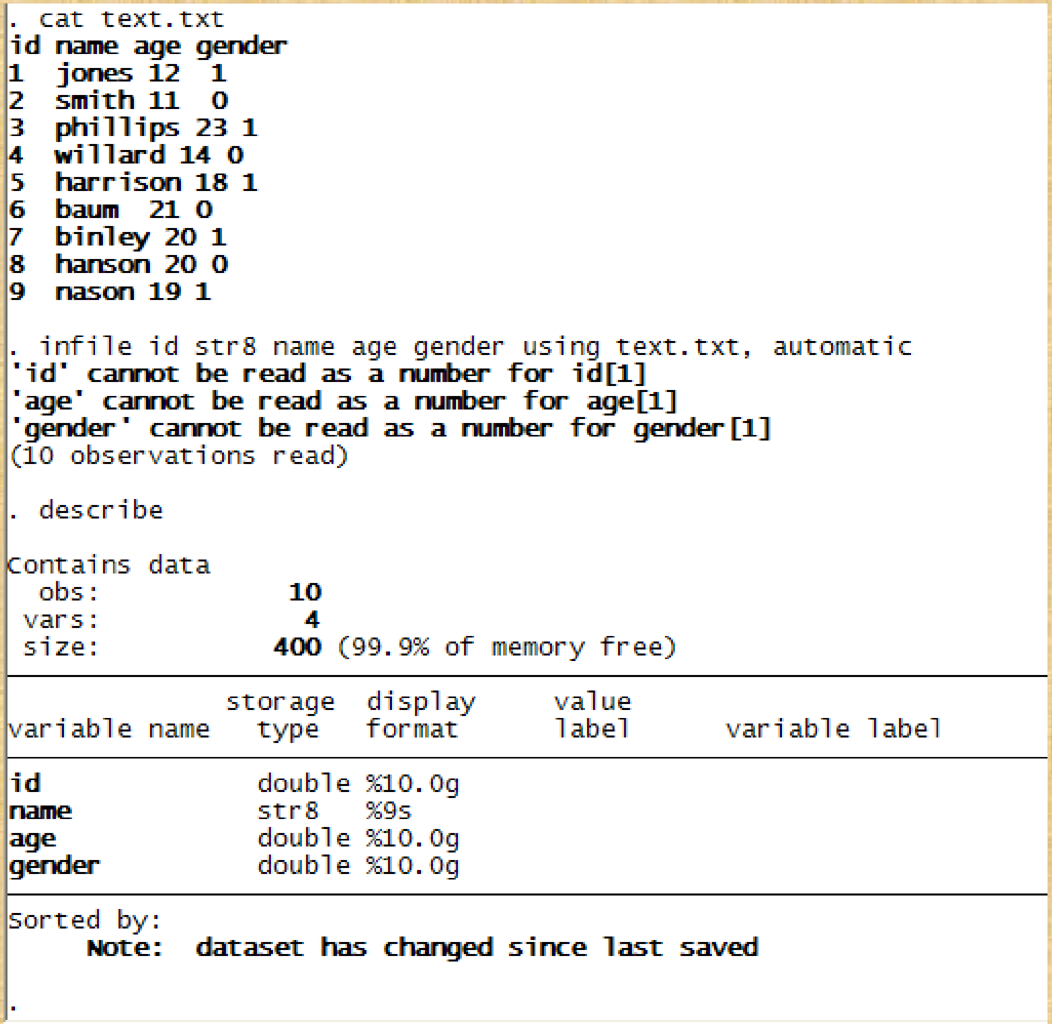
Import- export

- Importing data
  - Transferring data from excel
  - Insheet command with raw files
  - Statransfer
  - DBMSCopy
- Exporting data
  - Saveasas
  - Save as excel
  - Save as access
  - Statransfer
  - DBMSCopy

Importing data

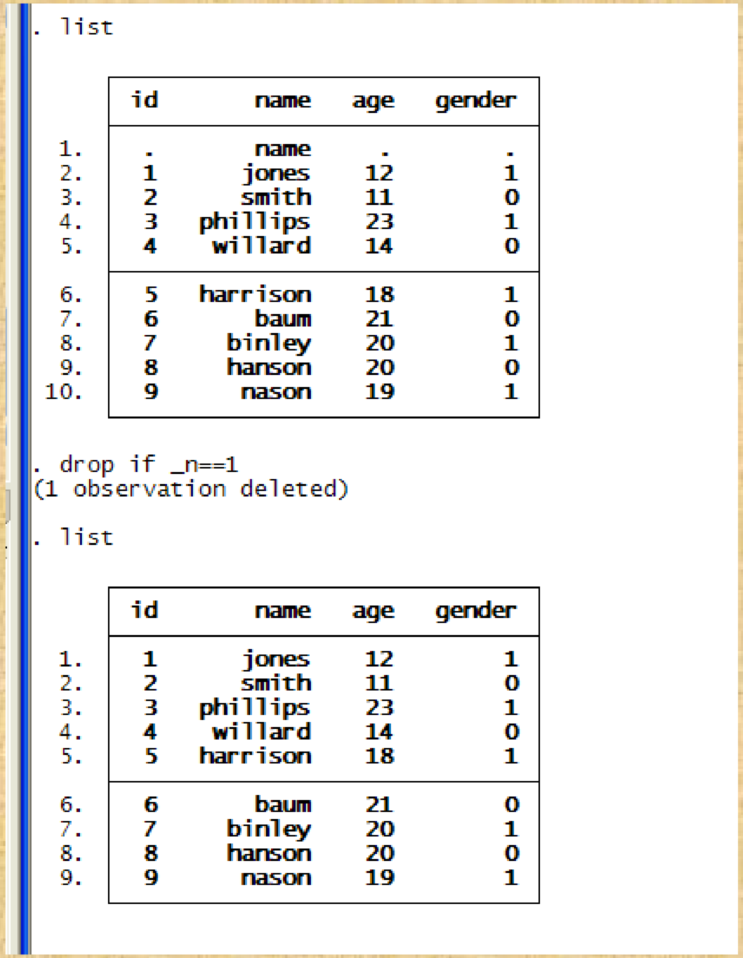
- From ascii text
- From spreadsheet files
- from other statistical packages
  - With stat transfer
  - With dbmscopy

Importing data from ascii text files



```
. cat text.txt
id name age gender
1  jones 12  1
2  smith 11  0
3  phillips 23 1
4  willard 14 0
5  harrison 18 1
6  baum  21 0
7  binley 20 1
8  hanson 20 0
9  nason 19 1

. infile id str8 name age gender using text.txt, automatic
'id' cannot be read as a number for id[1]
'age' cannot be read as a number for age[1]
'gender' cannot be read as a number for gender[1]
(10 observations read)

. describe

Contains data
  obs:            10
 vars:             4
 size:           400 (99.9% of memory free)
-------------------------------------------------------------------
              storage  display     value
variable name   type   format      label      variable label
-------------------------------------------------------------------
id              double %10.0g
name            str8   %9s
age             double %10.0g
gender          double %10.0g
-------------------------------------------------------------------
Sorted by:
     Note:  dataset has changed since last saved

.
```

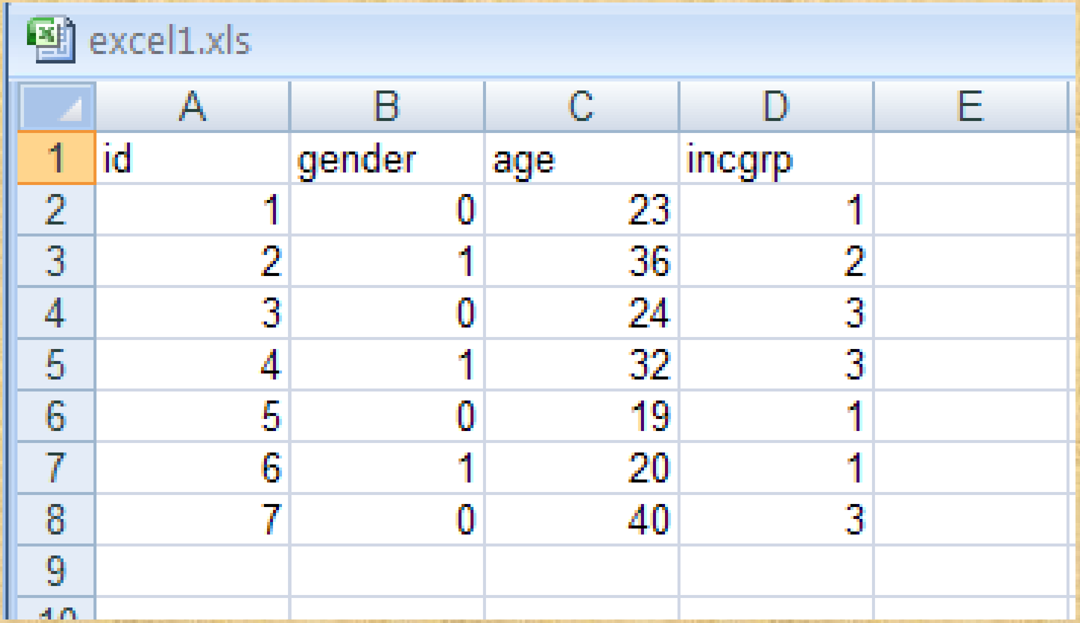
Post-importation refinement



```
. list

     +---------------------------------+
     | id       name   age      gender |
     |---------------------------------|
  1. |  .       name     .           . |
  2. |  1      jones    12           1 |
  3. |  2      smith    11           0 |
  4. |  3   phillips    23           1 |
  5. |  4    willard    14           0 |
     |---------------------------------|
  6. |  5   harrison    18           1 |
  7. |  6       baum    21           0 |
  8. |  7     binley    20           1 |
  9. |  8     hanson    20           0 |
 10. |  9      nason    19           1 |
     +---------------------------------+

. drop if _n==1
(1 observation deleted)

. list

     +---------------------------------+
     | id       name   age      gender |
     |---------------------------------|
  1. |  1      jones    12           1 |
  2. |  2      smith    11           0 |
  3. |  3   phillips    23           1 |
  4. |  4    willard    14           0 |
  5. |  5   harrison    18           1 |
     |---------------------------------|
  6. |  6       baum    21           0 |
  7. |  7     binley    20           1 |
  8. |  8     hanson    20           0 |
  9. |  9      nason    19           1 |
     +---------------------------------+
```

Transferring from Excel to Stata

- This is performed with a copy and paste operation. Suppose we have an excel worksheet 97-2003 file: excel1.xls

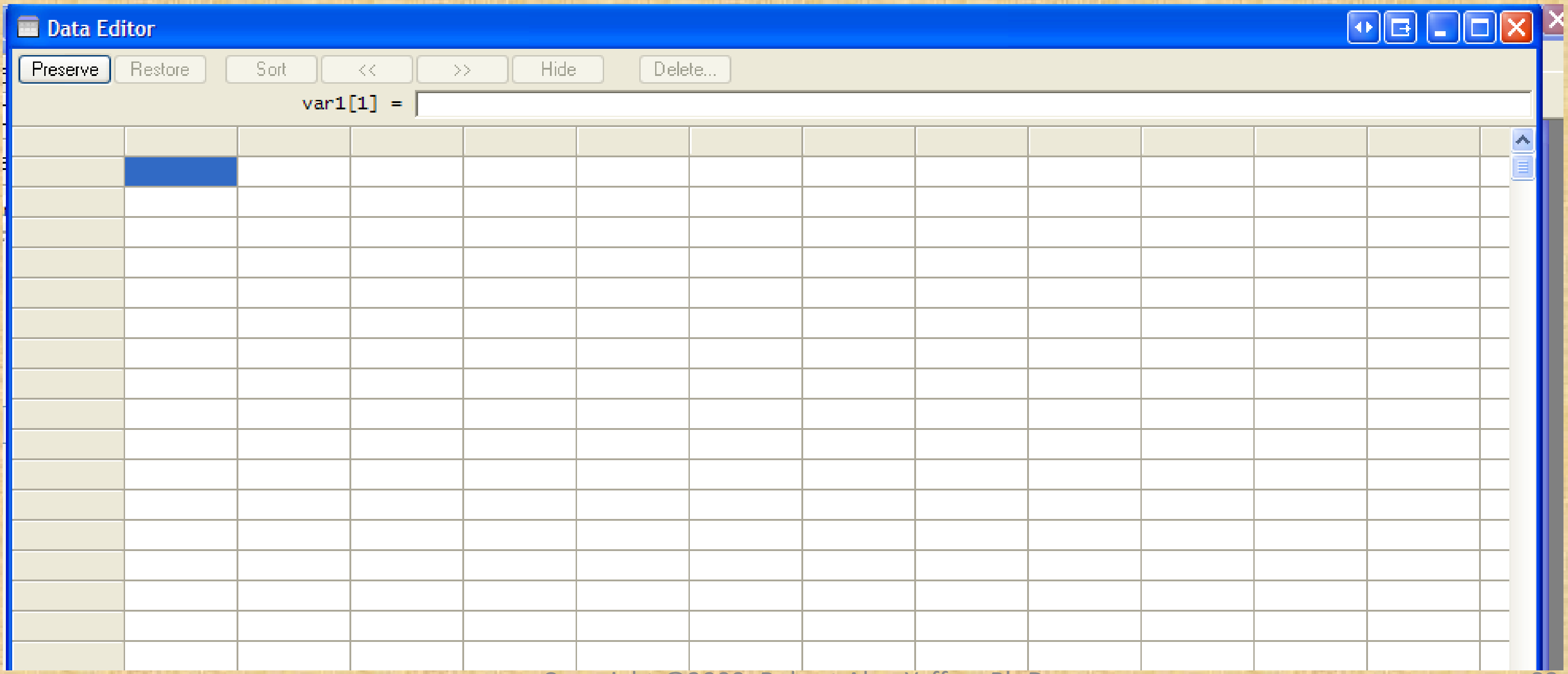


excel1.xls

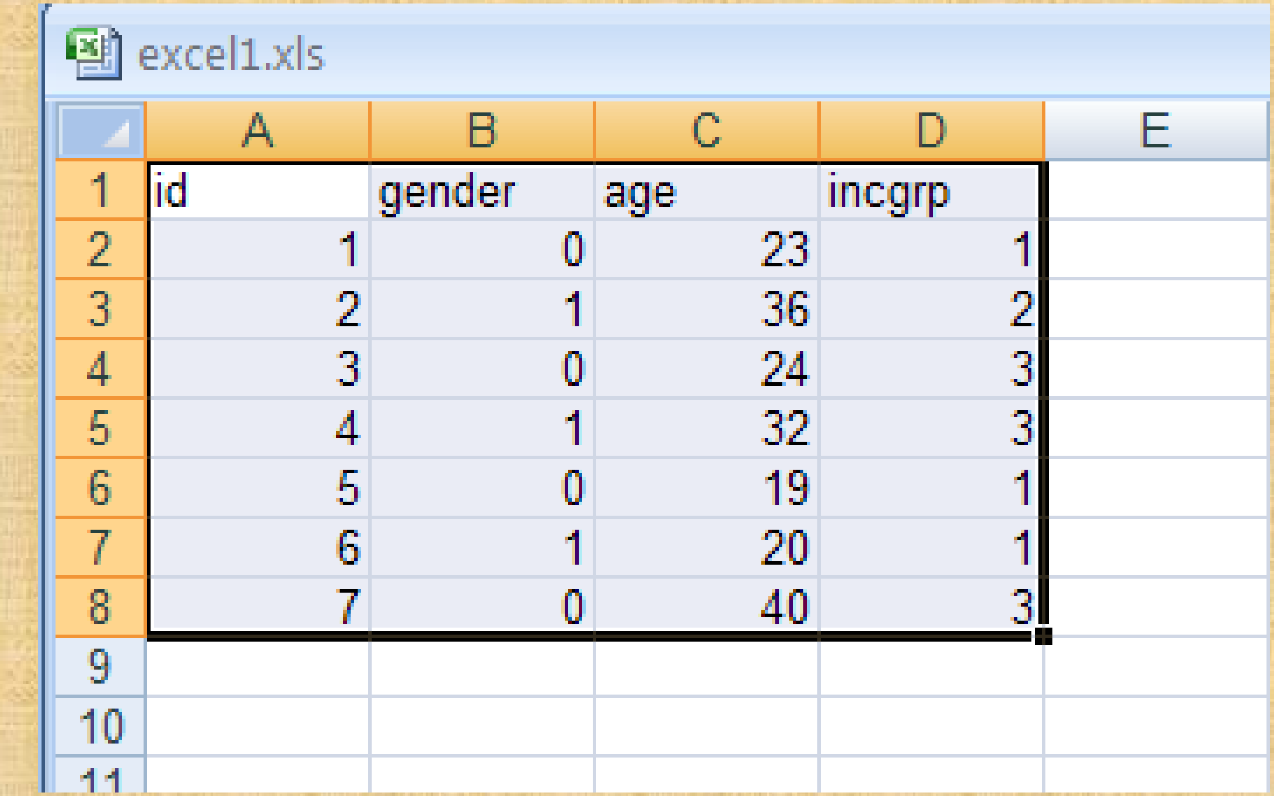
	A	B	C	D	E
1	id	gender	age	incgrp	
2	1	0	23	1	
3	2	1	36	2	
4	3	0	24	3	
5	4	1	32	3	
6	5	0	19	1	
7	6	1	20	1	
8	7	0	40	3	
9					

We can select all and paste it into a Stata datasheet

Be sure your data are cleared out.
On the command line, type: edit
A data editor opens below

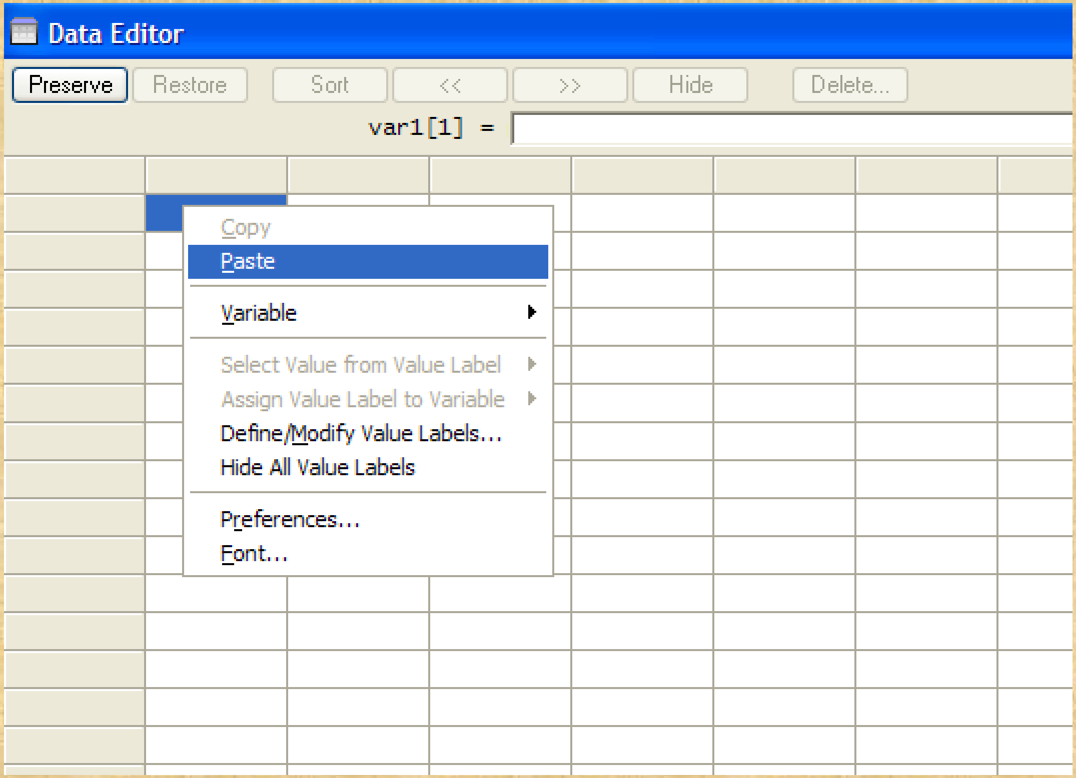


Select your data including the first line on which the variable names are contained. Right click on copy:

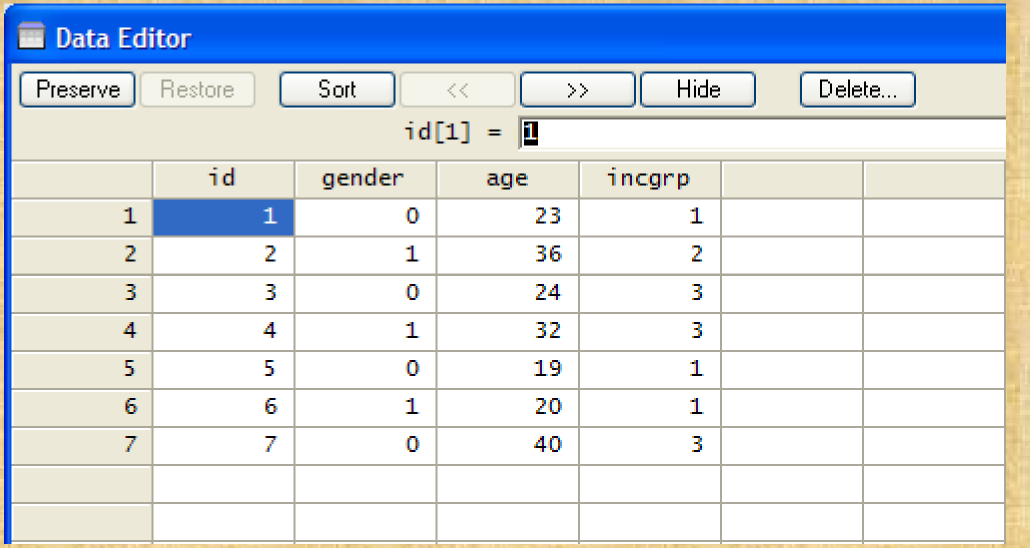


	A	B	C	D	E
1	id	gender	age	incgrp	
2	1	0	23	1	
3	2	1	36	2	
4	3	0	24	3	
5	4	1	32	3	
6	5	0	19	1	
7	6	1	20	1	
8	7	0	40	3	
9					
10					
11					

Paste in row1 column1 the selected data



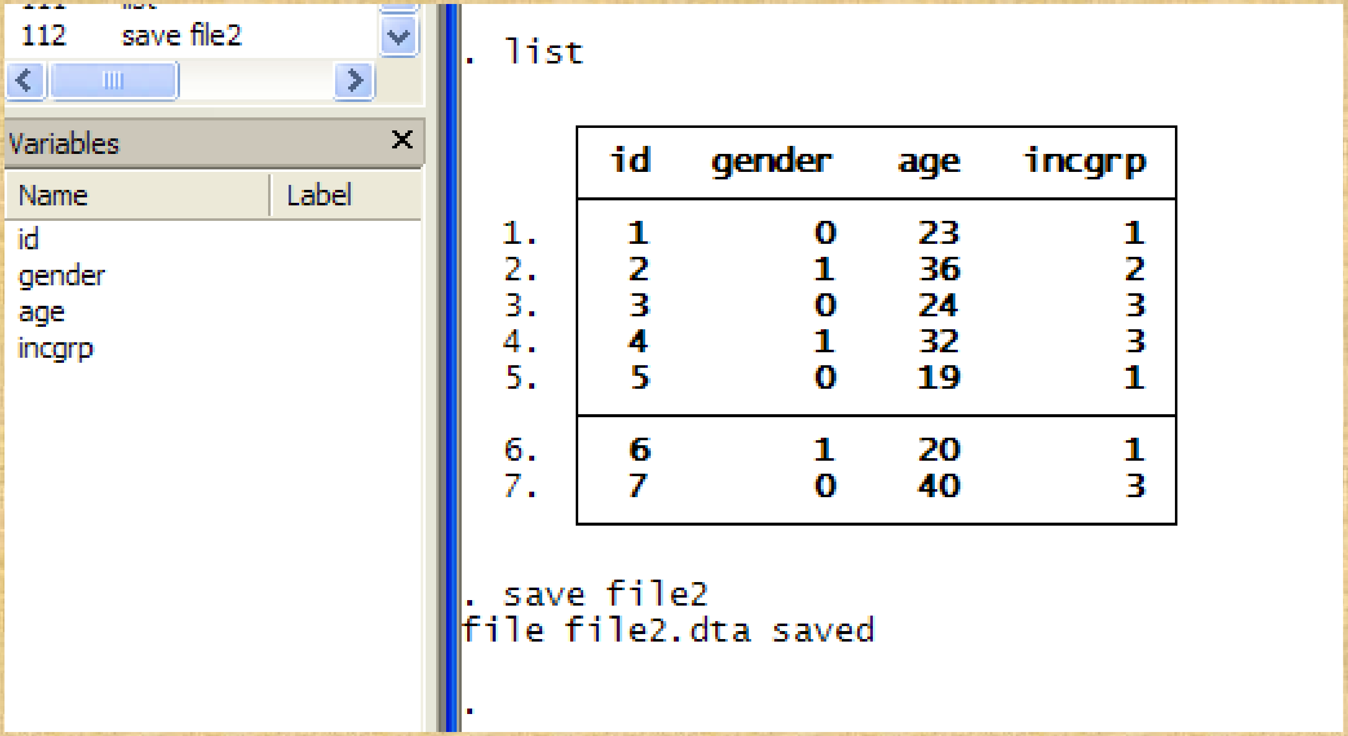
The data are pasted into the Stata data editor, click on preserve, and x out



id[1] = 1

	id	gender	age	incgrp
1	1	0	23	1
2	2	1	36	2
3	3	0	24	3
4	4	1	32	3
5	5	0	19	1
6	6	1	20	1
7	7	0	40	3

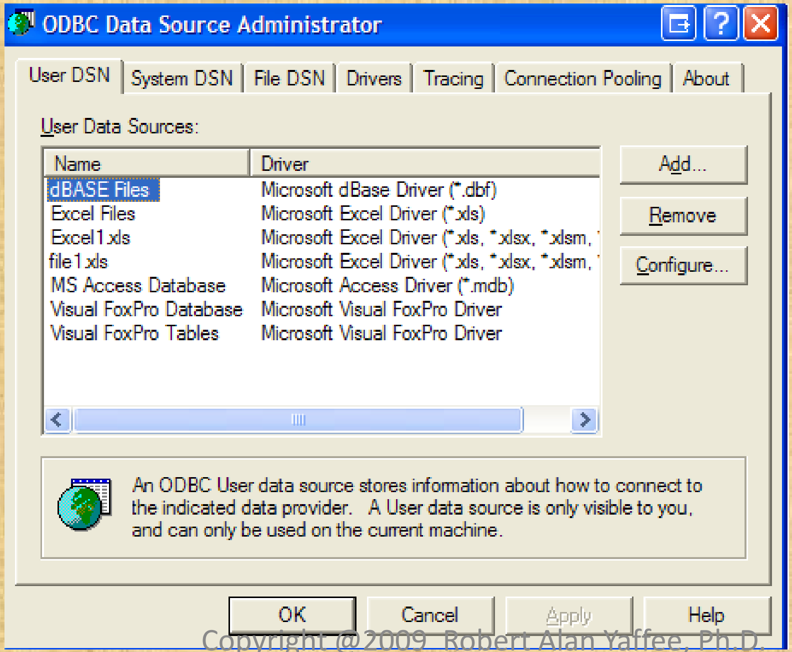
The data set is preserved in Stata. Save the file with the save filename command.



```
. list

     +---------------------------+
     | id   gender   age  incgrp |
     |---------------------------|
  1. |  1        0    23       1 |
  2. |  2        1    36       2 |
  3. |  3        0    24       3 |
  4. |  4        1    32       3 |
  5. |  5        0    19       1 |
     |---------------------------|
  6. |  6        1    20       1 |
  7. |  7        0    40       3 |
     +---------------------------+

. save file2
file file2.dta saved

.
```

Importing an Excel file with ODBC MS open database connectivity

- Save the excel file as file1.xls
- Go to administrative tools in the control panel and select the odbc options
- Setup the odbc dsn options in the control panel to include file1.xls



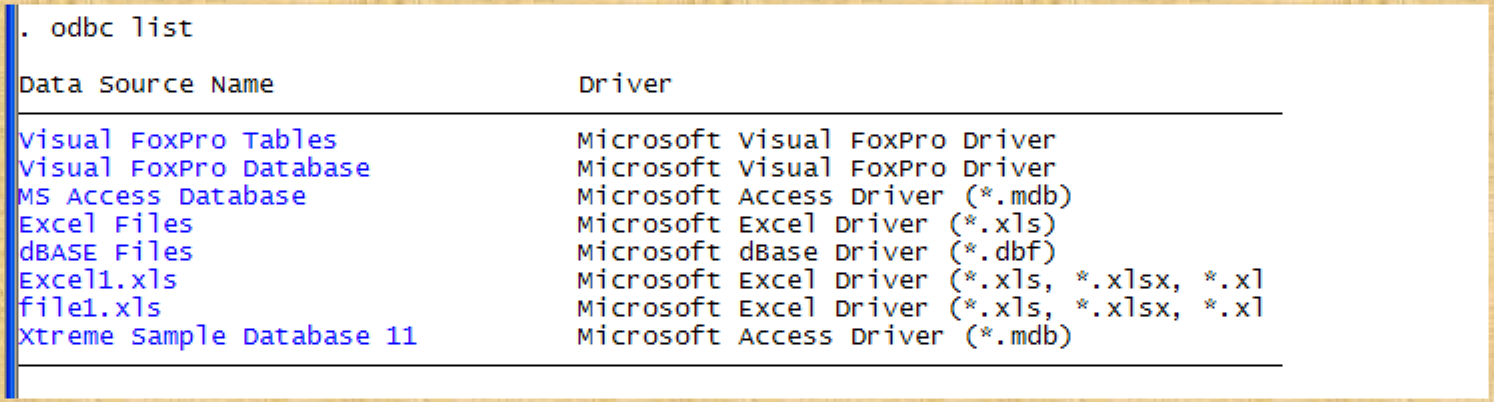
User Data Sources:

Name	Driver
dBASE Files	Microsoft dBase Driver (*.dbf)
Excel Files	Microsoft Excel Driver (*.xls)
Excel1.xls	Microsoft Excel Driver (*.xls, *.xlsx, *.xlsm,
file1.xls	Microsoft Excel Driver (*.xls, *.xlsx, *.xlsm,
MS Access Database	Microsoft Access Driver (*.mdb)
Visual FoxPro Database	Microsoft Visual FoxPro Driver
Visual FoxPro Tables	Microsoft Visual FoxPro Driver

An ODBC User data source stores information about how to connect to the indicated data provider. A User data source is only visible to you, and can only be used on the current machine.

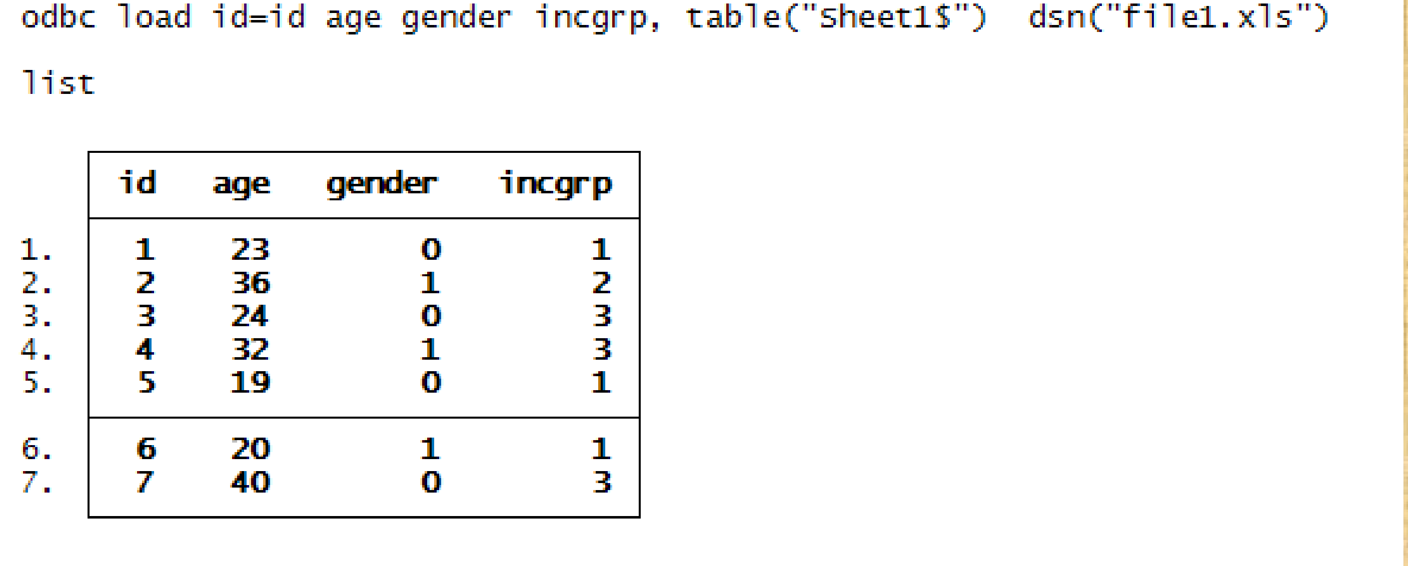
In Stata, confirm listing

- Confirm this by going back into Stata and typing:
- odbc list and being able to see your file in the list.



```
. odbc list

Data Source Name                   Driver
-------------------------------------------------------------------------------
Visual FoxPro Tables               Microsoft Visual FoxPro Driver
Visual FoxPro Database             Microsoft Visual FoxPro Driver
MS Access Database                 Microsoft Access Driver (*.mdb)
Excel Files                        Microsoft Excel Driver (*.xls)
dBASE Files                        Microsoft dBase Driver (*.dbf)
Excel1.xls                         Microsoft Excel Driver (*.xls, *.xlsx, *.xl
file1.xls                          Microsoft Excel Driver (*.xls, *.xlsx, *.xl
Xtreme Sample Database 11          Microsoft Access Driver (*.mdb)
-------------------------------------------------------------------------------
```

This will import the file to Stata

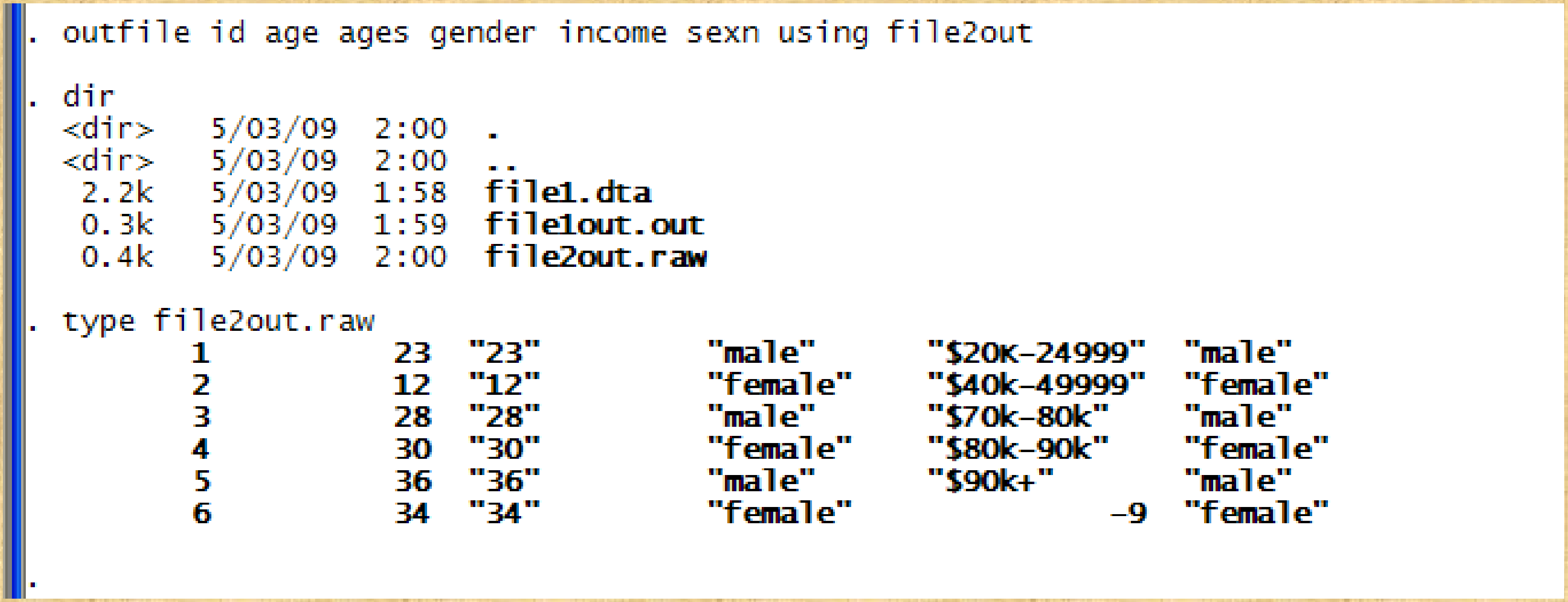


```
odbc load id=id age gender incgrp, table("Sheet1$")  dsn("file1.xls")

list

     +-----------------------------+
     | id   age   gender   incgrp  |
     |-----------------------------|
  1. |  1    23        0        1  |
  2. |  2    36        1        2  |
  3. |  3    24        0        3  |
  4. |  4    32        1        3  |
  5. |  5    19        0        1  |
     |-----------------------------|
  6. |  6    20        1        1  |
  7. |  7    40        0        3  |
     +-----------------------------+
```

Exporting data files

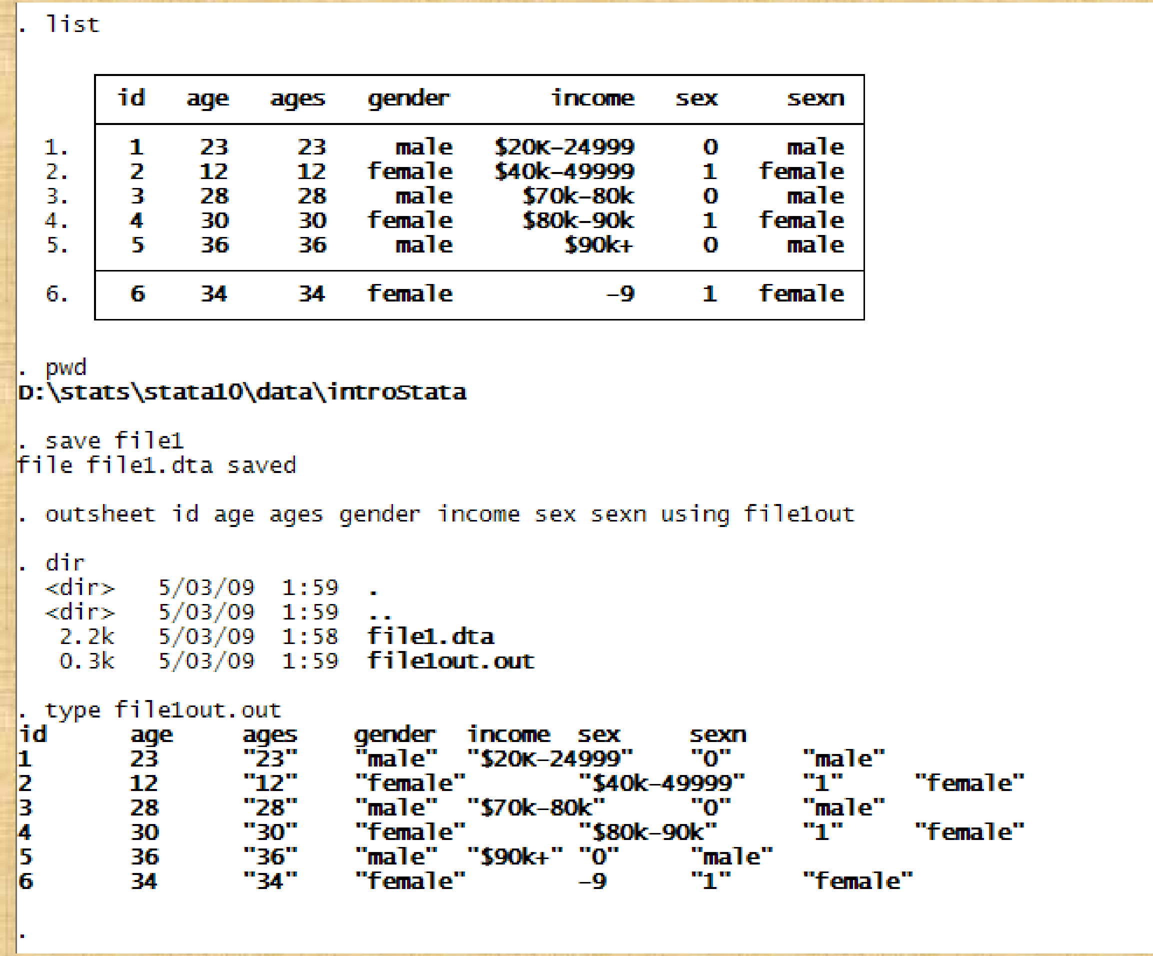
- To other statistical packages
  - With DBMScopy
  - With Statransfer
- Raw data files

Exporting a raw data file



```
. outfile id age ages gender income sexn using file2out

. dir
  <dir>   5/03/09  2:00  .
  <dir>   5/03/09  2:00  ..
   2.2k   5/03/09  1:58  file1.dta
   0.3k   5/03/09  1:59  file1out.out
   0.4k   5/03/09  2:00  file2out.raw

. type file2out.raw
         1          23  "23"         "male"      "$20k-24999"  "male"
         2          12  "12"         "female"    "$40k-49999"  "female"
         3          28  "28"         "male"      "$70k-80k"    "male"
         4          30  "30"         "female"    "$80k-90k"    "female"
         5          36  "36"         "male"      "$90k+"       "male"
         6          34  "34"         "female"              -9  "female"

.
```

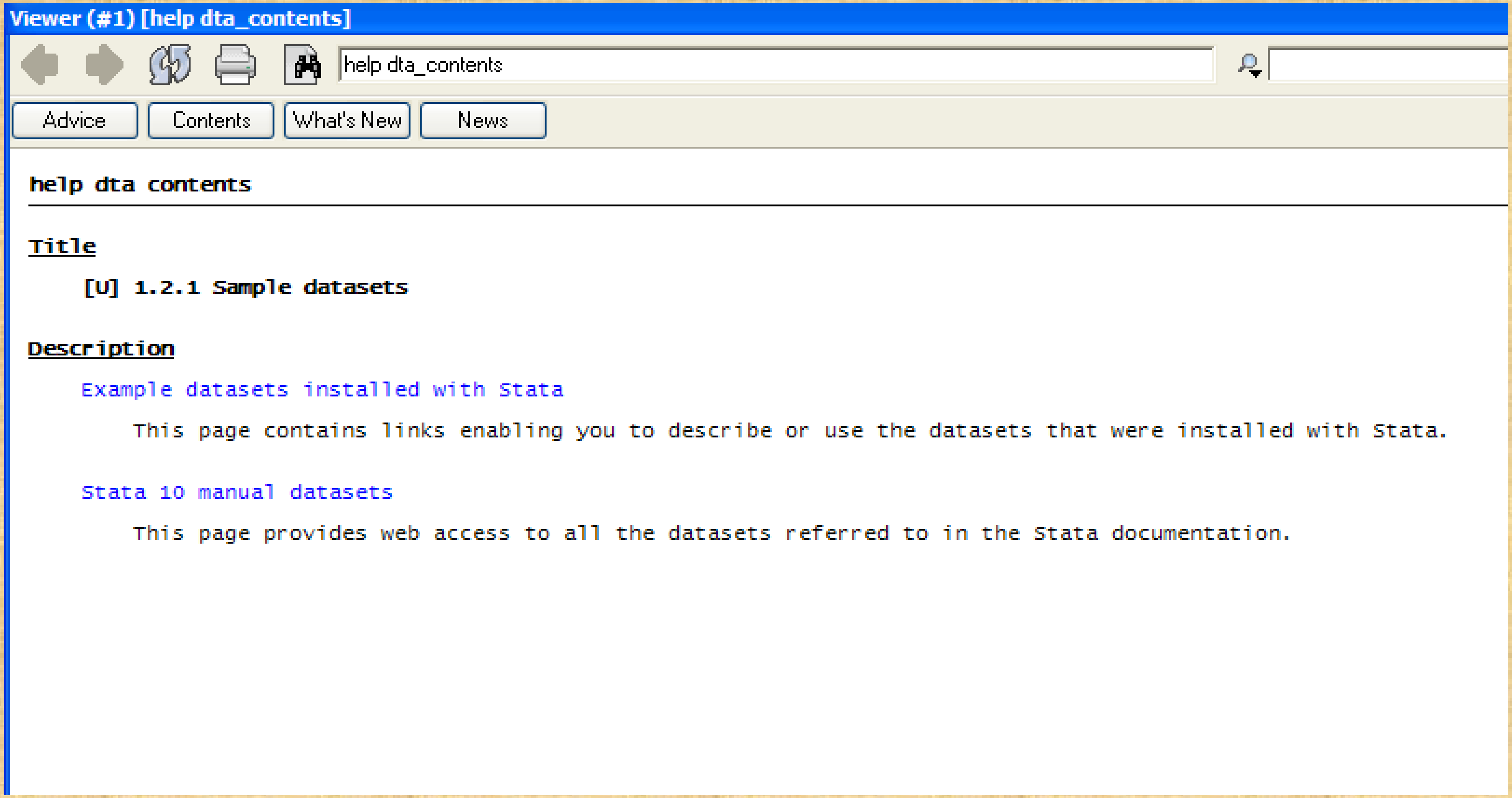
outsheet using file1out



```
. list

     +-------------------------------------------------------+
     | id   age   ages   gender        income   sex     sexn |
     |-------------------------------------------------------|
  1. |  1    23     23     male   $20K-24999     0     male |
  2. |  2    12     12   female   $40k-49999     1   female |
  3. |  3    28     28     male     $70k-80k     0     male |
  4. |  4    30     30   female     $80k-90k     1   female |
  5. |  5    36     36     male        $90k+     0     male |
     |-------------------------------------------------------|
  6. |  6    34     34   female           -9     1   female |
     +-------------------------------------------------------+

. pwd
D:\stats\stata10\data\introStata

. save file1
file file1.dta saved

. outsheet id age ages gender income sex sexn using file1out

. dir
  <dir>   5/03/09  1:59  .
  <dir>   5/03/09  1:59  ..
   2.2k   5/03/09  1:58  file1.dta
   0.3k   5/03/09  1:59  file1out.out

. type file1out.out
id      age     ages    gender  income  sex     sexn
1       23      "23"    "male"  "$20K-24999"    "0"     "male"
2       12      "12"    "female"        "$40k-49999"    "1"     "female"
3       28      "28"    "male"  "$70k-80k"      "0"     "male"
4       30      "30"    "female"        "$80k-90k"      "1"     "female"
5       36      "36"    "male"  "$90k+" "0"     "male"
6       34      "34"    "female"        -9      "1"     "female"

.
```

Accessing example datasets

- Type: help datasets



help dta contents

Title

[U] 1.2.1 Sample datasets

Description

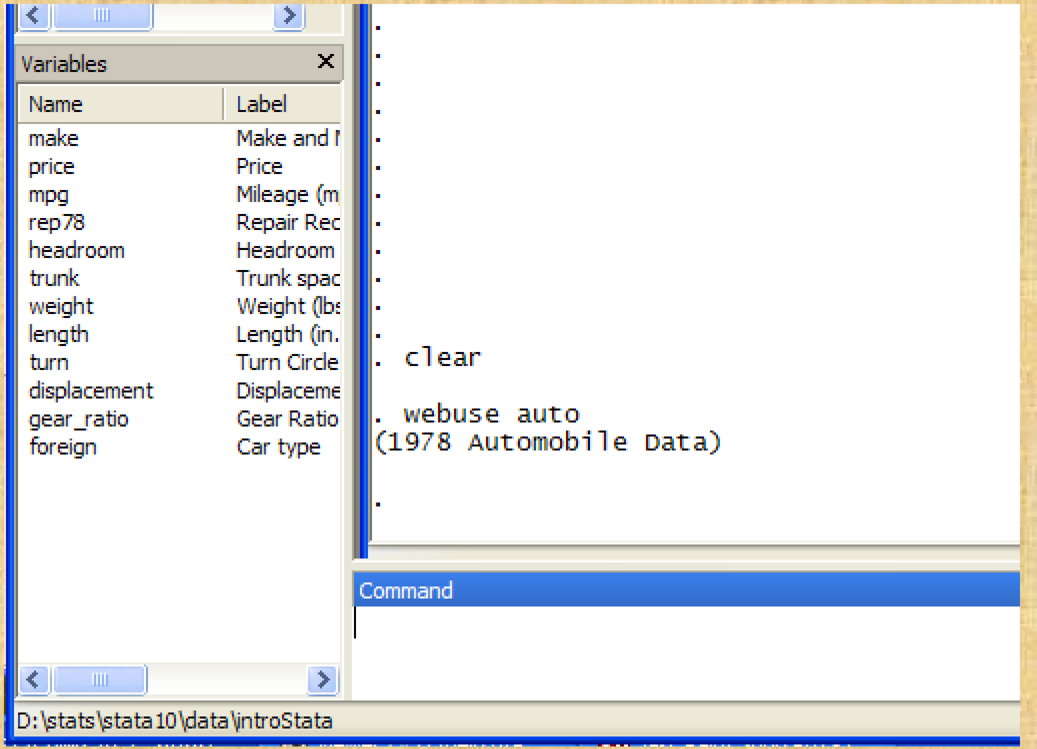
Example datasets installed with Stata

This page contains links enabling you to describe or use the datasets that were installed with Stata.

Stata 10 manual datasets

This page provides web access to all the datasets referred to in the Stata documentation.

On command line: type: webuse auto



Name	Label
make	Make and I
price	Price
mpg	Mileage (m
rep78	Repair Rec
headroom	Headroom
trunk	Trunk spac
weight	Weight (lbs
length	Length (in.
turn	Turn Circle
displacement	Displaceme
gear_ratio	Gear Ratio
foreign	Car type

```
.
.
.
.
.
.
.
.
.
.
.
.
. clear

. webuse auto
(1978 Automobile Data)

.
```

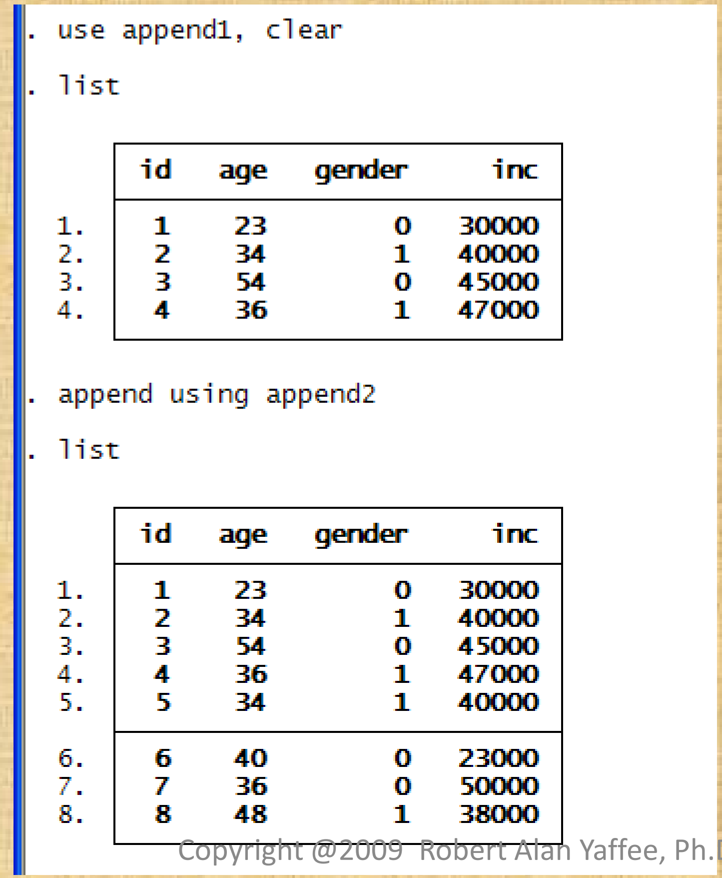
D:\stats\stata10\data\introStata

Combining datasets

- Appending: adding cases to the same variables
- Merging: adding variables to the same cases
- Mixtures
- Caveats: be sure that the missing values are coded the same and designated missing
- Sort both datasets by the same variables before combining.

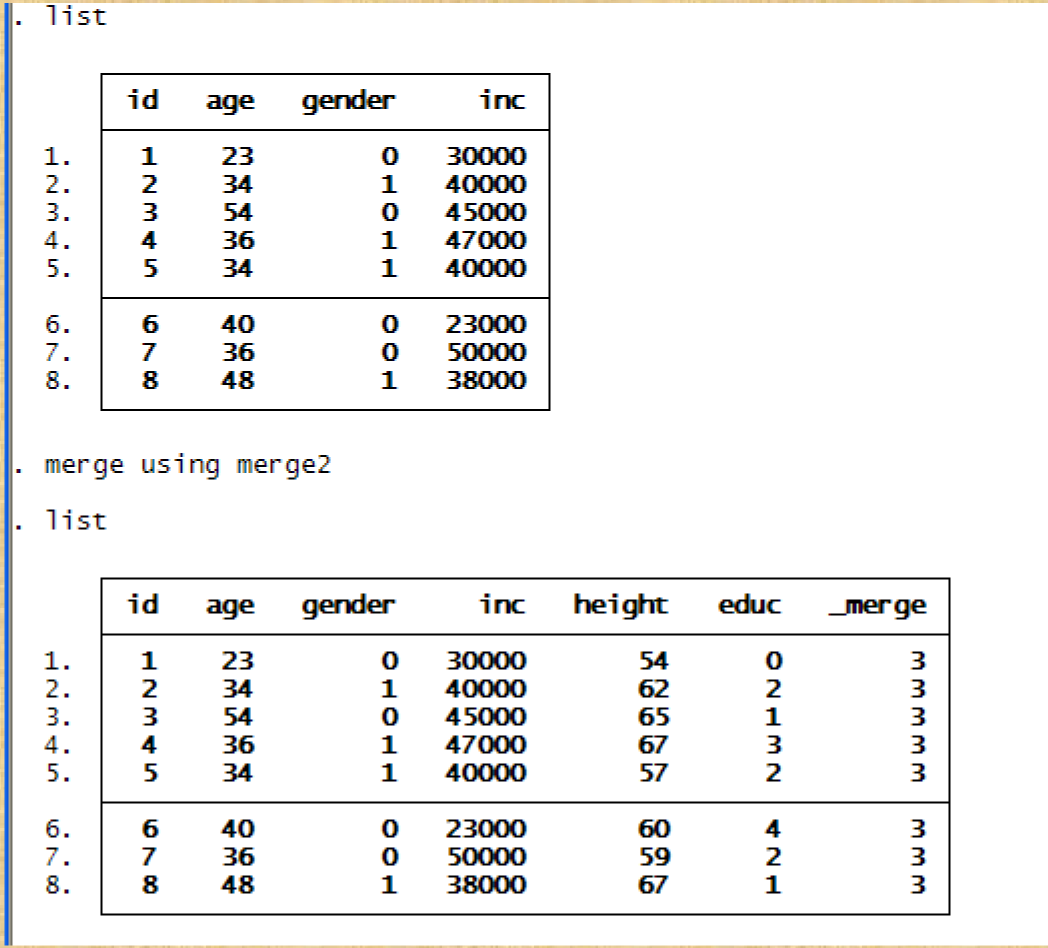
Appending datasets

- Adding cases to the same variables can be done with the append command. This concatenates the data.



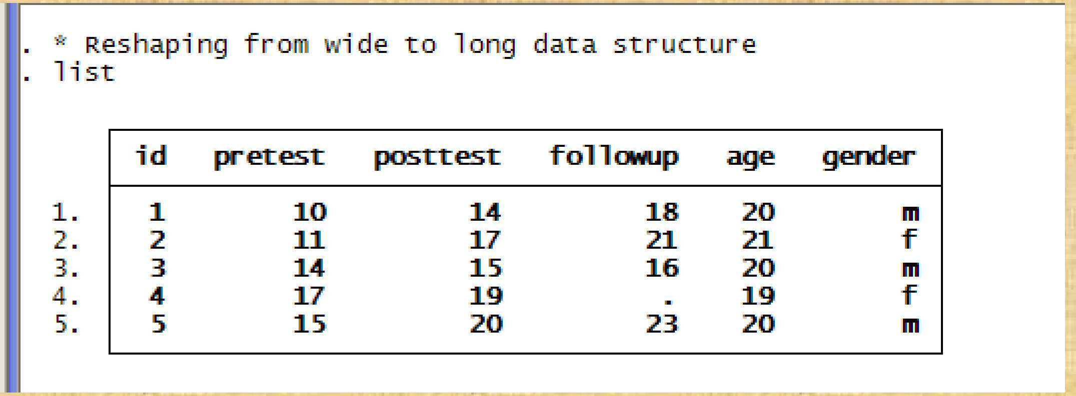
```
. use append1, clear

. list

     +---------------------------+
     | id   age   gender     inc |
     |---------------------------|
  1. |  1    23        0   30000 |
  2. |  2    34        1   40000 |
  3. |  3    54        0   45000 |
  4. |  4    36        1   47000 |
     +---------------------------+

. append using append2

. list

     +---------------------------+
     | id   age   gender     inc |
     |---------------------------|
  1. |  1    23        0   30000 |
  2. |  2    34        1   40000 |
  3. |  3    54        0   45000 |
  4. |  4    36        1   47000 |
  5. |  5    34        1   40000 |
     |---------------------------|
  6. |  6    40        0   23000 |
  7. |  7    36        0   50000 |
  8. |  8    48        1   38000 |
     +---------------------------+
```

Merging datasets

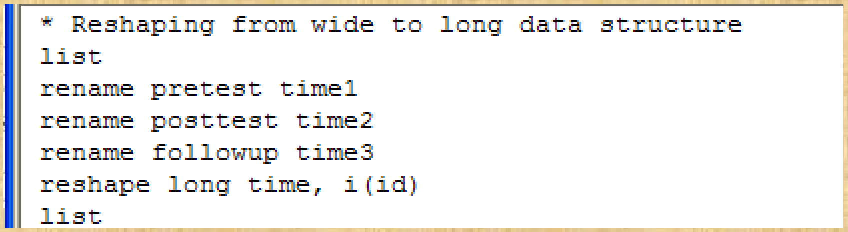


```
. list

     +--------------------------------+
     | id   age   gender          inc |
     |--------------------------------|
  1. |  1    23        0        30000 |
  2. |  2    34        1        40000 |
  3. |  3    54        0        45000 |
  4. |  4    36        1        47000 |
  5. |  5    34        1        40000 |
     |--------------------------------|
  6. |  6    40        0        23000 |
  7. |  7    36        0        50000 |
  8. |  8    48        1        38000 |
     +--------------------------------+

. merge using merge2

. list

     +-------------------------------------------------------------+
     | id   age   gender          inc   height   educ       _merge |
     |-------------------------------------------------------------|
  1. |  1    23        0        30000       54      0            3 |
  2. |  2    34        1        40000       62      2            3 |
  3. |  3    54        0        45000       65      1            3 |
  4. |  4    36        1        47000       67      3            3 |
  5. |  5    34        1        40000       57      2            3 |
     |-------------------------------------------------------------|
  6. |  6    40        0        23000       60      4            3 |
  7. |  7    36        0        50000       59      2            3 |
  8. |  8    48        1        38000       67      1            3 |
     +-------------------------------------------------------------+
```

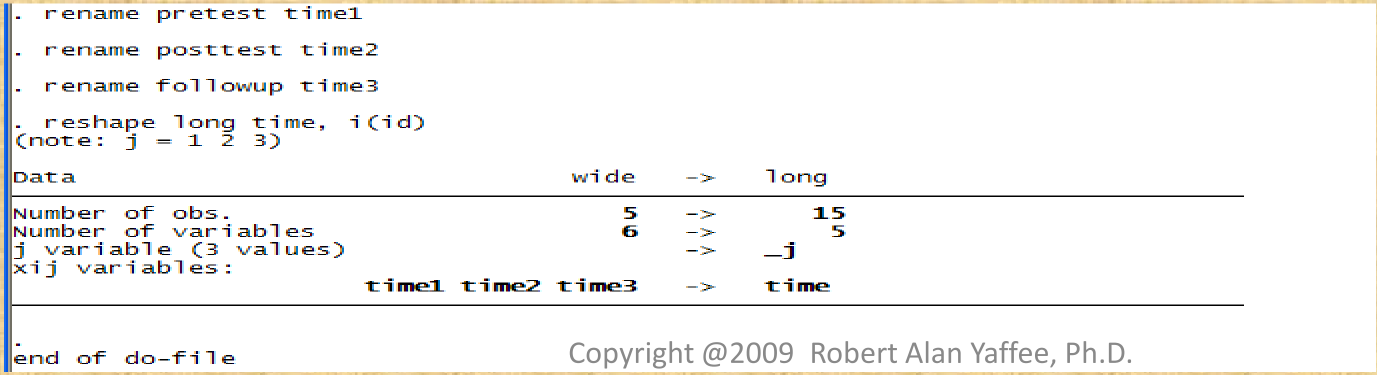
Reshaping files:
Wide to long and Long to Wide



```
. * Reshaping from wide to long data structure
. list

     +------------------------------------------------------+
     | id   pretest   posttest   followup   age   gender   |
     |------------------------------------------------------|
  1. |  1        10         14         18    20        m   |
  2. |  2        11         17         21    21        f   |
  3. |  3        14         15         16    20        m   |
  4. |  4        17         19          .    19        f   |
  5. |  5        15         20         23    20        m   |
     +------------------------------------------------------+
```

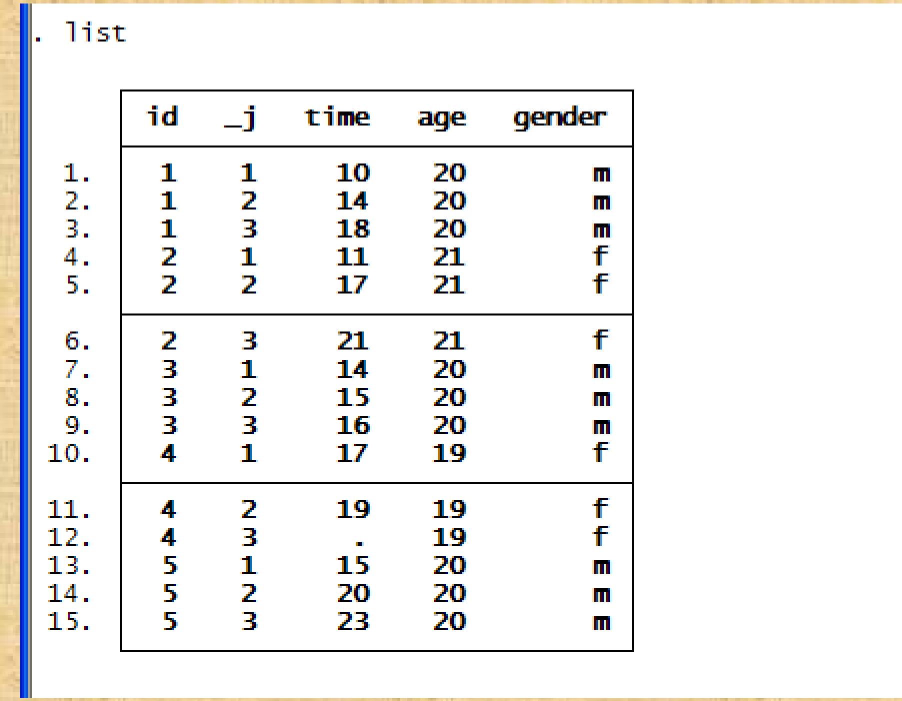


```
* Reshaping from wide to long data structure
list
rename pretest time1
rename posttest time2
rename followup time3
reshape long time, i(id)
list
```



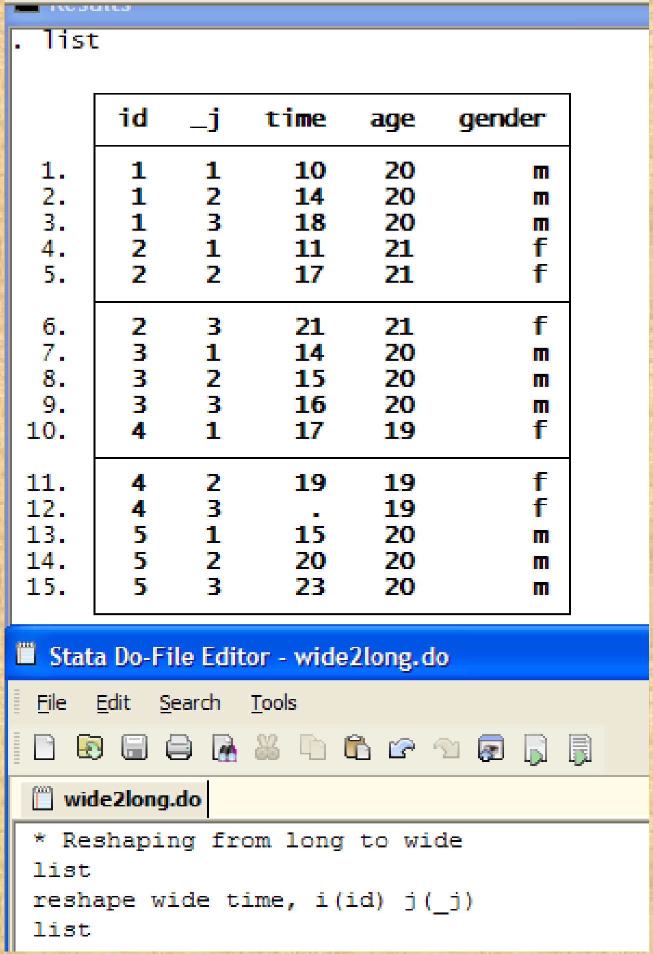
```
. rename pretest time1

. rename posttest time2

. rename followup time3

. reshape long time, i(id)
(note: j = 1 2 3)

Data                               wide   ->   long
-----------------------------------------------------------------------------
Number of obs.                        5   ->      15
Number of variables                   6   ->       5
j variable (3 values)                     ->   _j
xij variables:
                      time1 time2 time3   ->   time
-----------------------------------------------------------------------------

.
end of do-file
```

Output of reshape from wide to long (person-period dataset)



```
. list

     +-----------------------------------+
     | id   _j   time   age      gender |
     |-----------------------------------|
  1. |  1    1     10    20           m |
  2. |  1    2     14    20           m |
  3. |  1    3     18    20           m |
  4. |  2    1     11    21           f |
  5. |  2    2     17    21           f |
     |-----------------------------------|
  6. |  2    3     21    21           f |
  7. |  3    1     14    20           m |
  8. |  3    2     15    20           m |
  9. |  3    3     16    20           m |
 10. |  4    1     17    19           f |
     |-----------------------------------|
 11. |  4    2     19    19           f |
 12. |  4    3      .    19           f |
 13. |  5    1     15    20           m |
 14. |  5    2     20    20           m |
 15. |  5    3     23    20           m |
     +-----------------------------------+
```

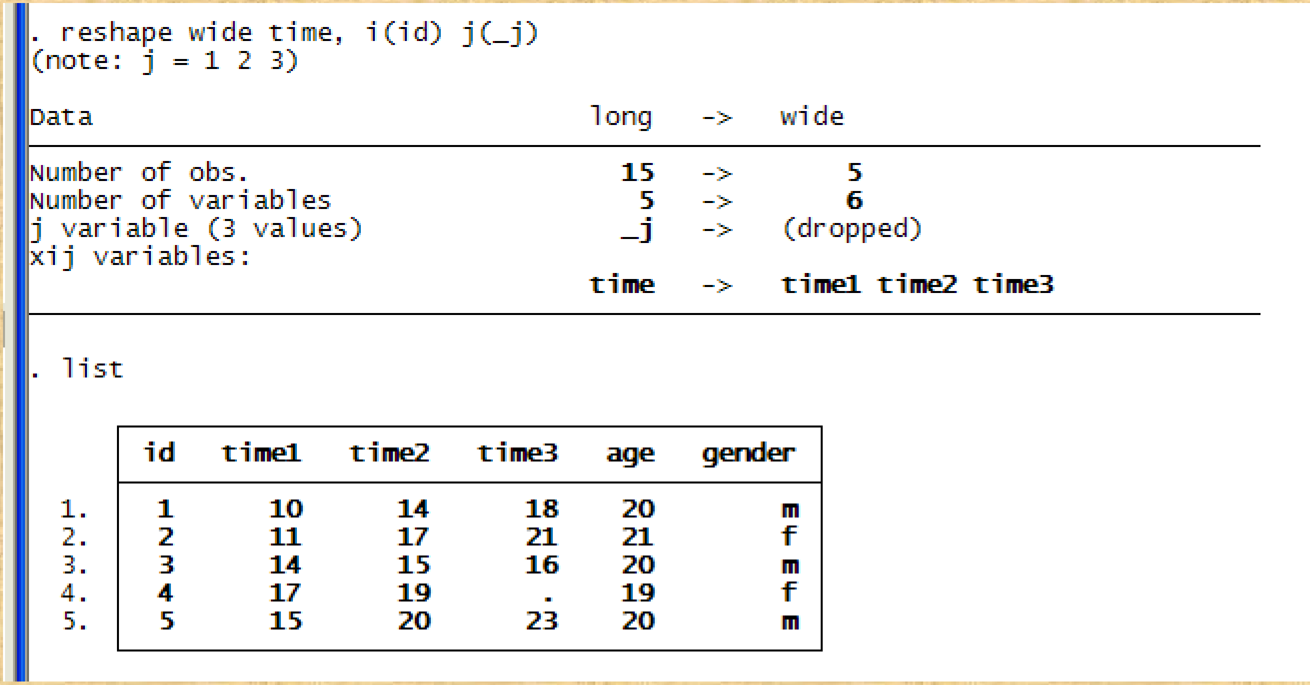
Reshape from long to wide



```
. list

     +-------------------------------+
     | id   _j   time   age   gender |
     |-------------------------------|
  1. |  1    1     10    20        m |
  2. |  1    2     14    20        m |
  3. |  1    3     18    20        m |
  4. |  2    1     11    21        f |
  5. |  2    2     17    21        f |
     |-------------------------------|
  6. |  2    3     21    21        f |
  7. |  3    1     14    20        m |
  8. |  3    2     15    20        m |
  9. |  3    3     16    20        m |
 10. |  4    1     17    19        f |
     |-------------------------------|
 11. |  4    2     19    19        f |
 12. |  4    3      .    19        f |
 13. |  5    1     15    20        m |
 14. |  5    2     20    20        m |
 15. |  5    3     23    20        m |
     +-------------------------------+
```

```
* Reshaping from long to wide
list
reshape wide time, i(id) j(_j)
list
```

Output from reshaping from long to wide



```
. reshape wide time, i(id) j(_j)
(note: j = 1 2 3)

Data                               long   ->   wide
-----------------------------------------------------------------------------
Number of obs.                       15   ->       5
Number of variables                   5   ->       6
j variable (3 values)                _j   ->   (dropped)
xij variables:
                                   time   ->   time1 time2 time3
-----------------------------------------------------------------------------

. list

     +-----------------------------------------------+
     | id   time1   time2   time3   age   gender |
     |-----------------------------------------------|
  1. |  1      10      14      18    20        m |
  2. |  2      11      17      21    21        f |
  3. |  3      14      15      16    20        m |
  4. |  4      17      19       .    19        f |
  5. |  5      15      20      23    20        m |
     +-----------------------------------------------+
```

sort var1 var2

- You may reorganize your data with a sort command. You may sort by a series of variables.

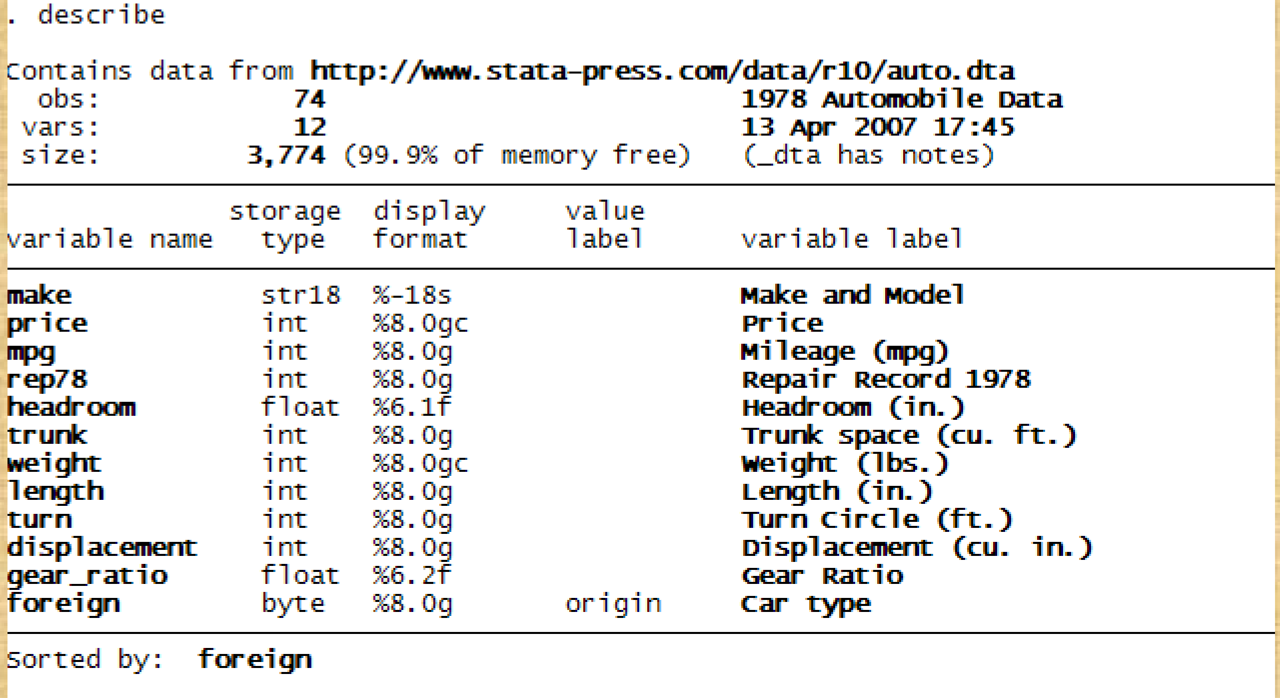
Data management

- Data management
  Data cleaning range checks with tabulate
  summary statistics with summarize
  consistency checks with pwcorr
  file comparison utilities

  - Codebook maintenance
    - Summary statistics
    - Recoded variables
    - New variables
    - Multiple sorts
    - Basic graphs
    - Missing values
  - Variable transformations
    - Rename
    - Recode
    - Replace if
    - Generate if
    - Egen
    - List if
  - Missing value management
    - Storage as very large numbers
    - Mvdecode
    - Mvencode
    - Drop
    - Keep
    - Imputation
      - Single
      - Multiple with mice
  - Log files contain time and date
    - Headers
    - Why use log files?
    - Need to keep record and log of work

We wish to examine the dataset

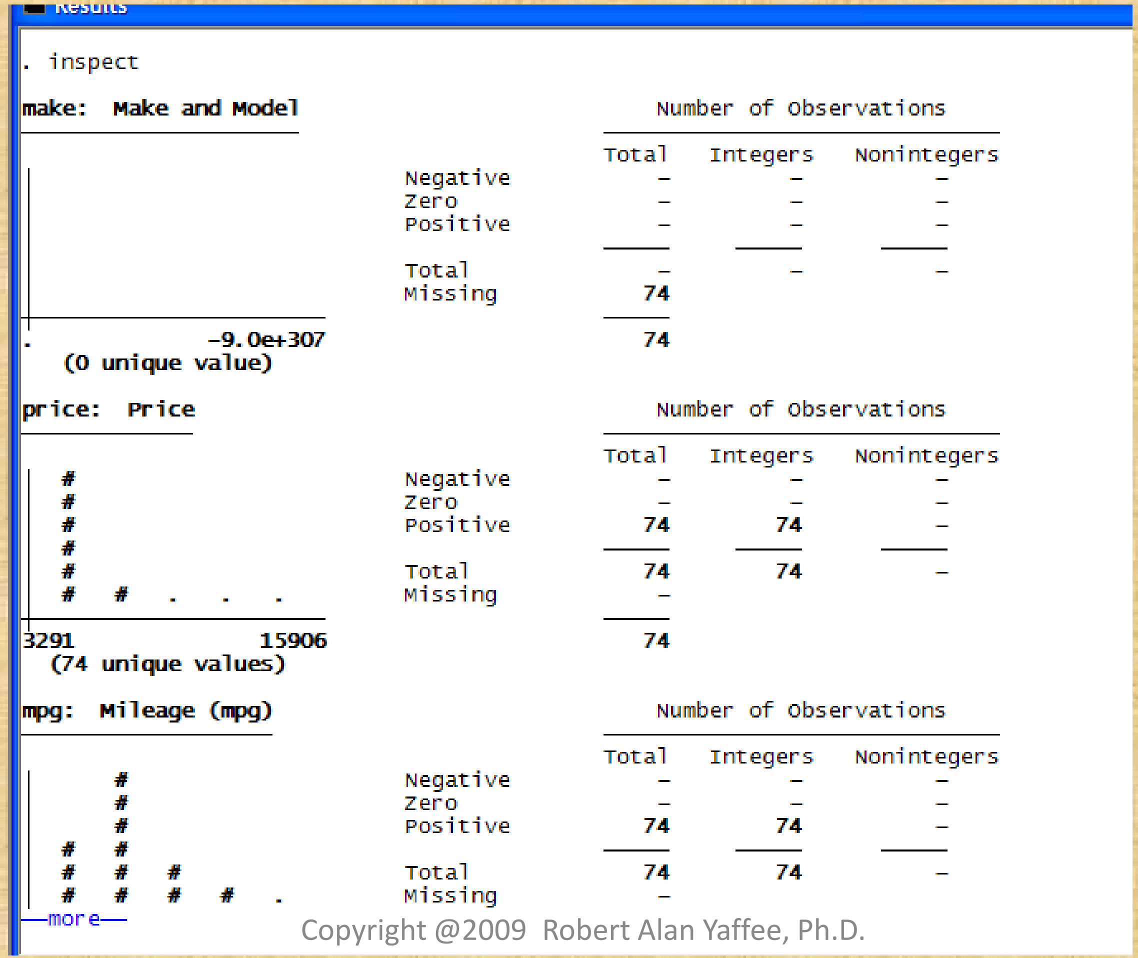
- Type: describe



```
. describe

Contains data from http://www.stata-press.com/data/r10/auto.dta
  obs:            74                          1978 Automobile Data
 vars:            12                          13 Apr 2007 17:45
 size:         3,774 (99.9% of memory free)   (_dta has notes)
-------------------------------------------------------------------------------
              storage  display     value
variable name   type   format      label      variable label
-------------------------------------------------------------------------------
make            str18  %-18s                  Make and Model
price           int    %8.0gc                 Price
mpg             int    %8.0g                  Mileage (mpg)
rep78           int    %8.0g                  Repair Record 1978
headroom        float  %6.1f                  Headroom (in.)
trunk           int    %8.0g                  Trunk space (cu. ft.)
weight          int    %8.0gc                 Weight (lbs.)
length          int    %8.0g                  Length (in.)
turn            int    %8.0g                  Turn Circle (ft.)
displacement    int    %8.0g                  Displacement (cu. in.)
gear_ratio      float  %6.2f                  Gear Ratio
foreign         byte   %8.0g       origin     Car type
-------------------------------------------------------------------------------
Sorted by:  foreign
```

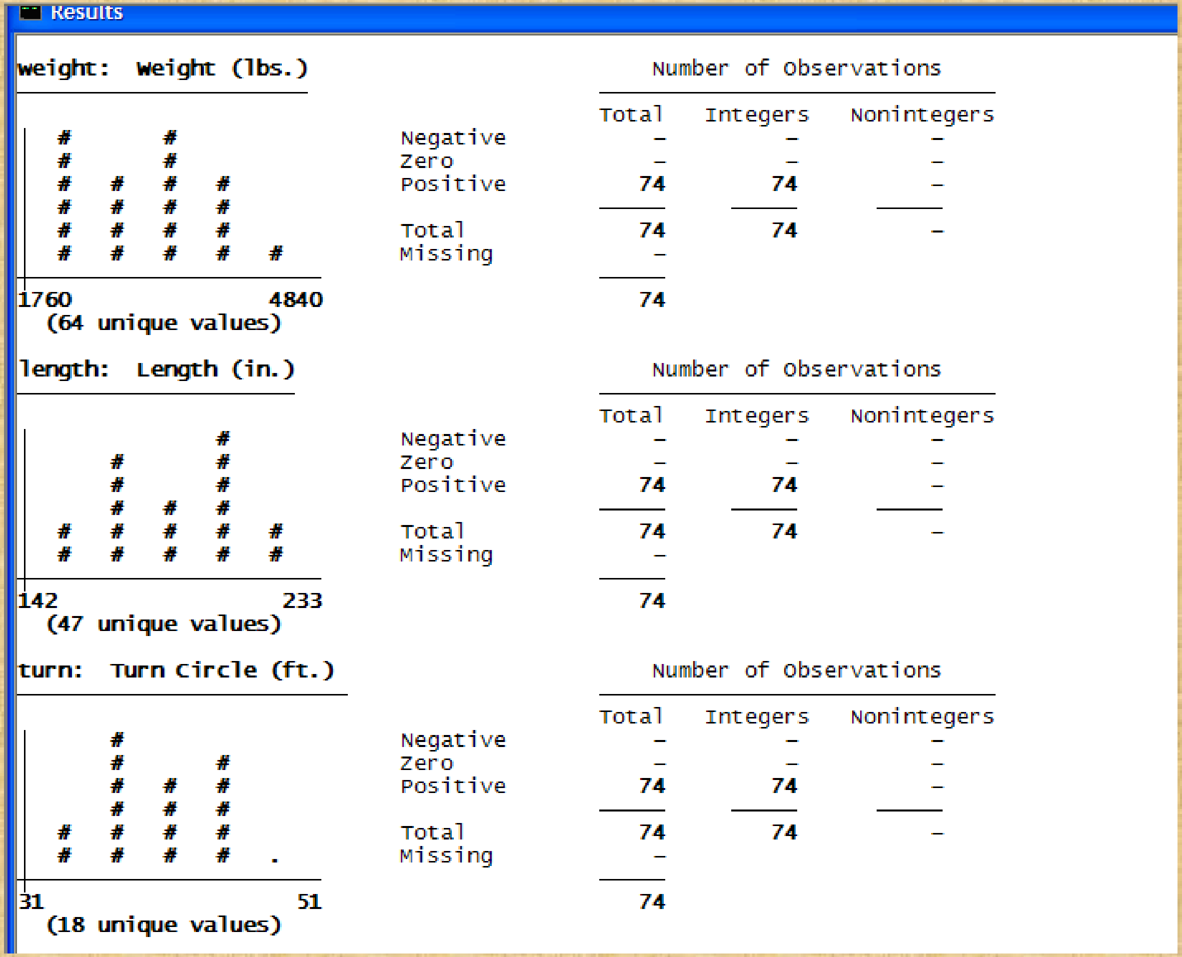
Data Scan: reveals histograms and missing data.

Type: inspect on the command line



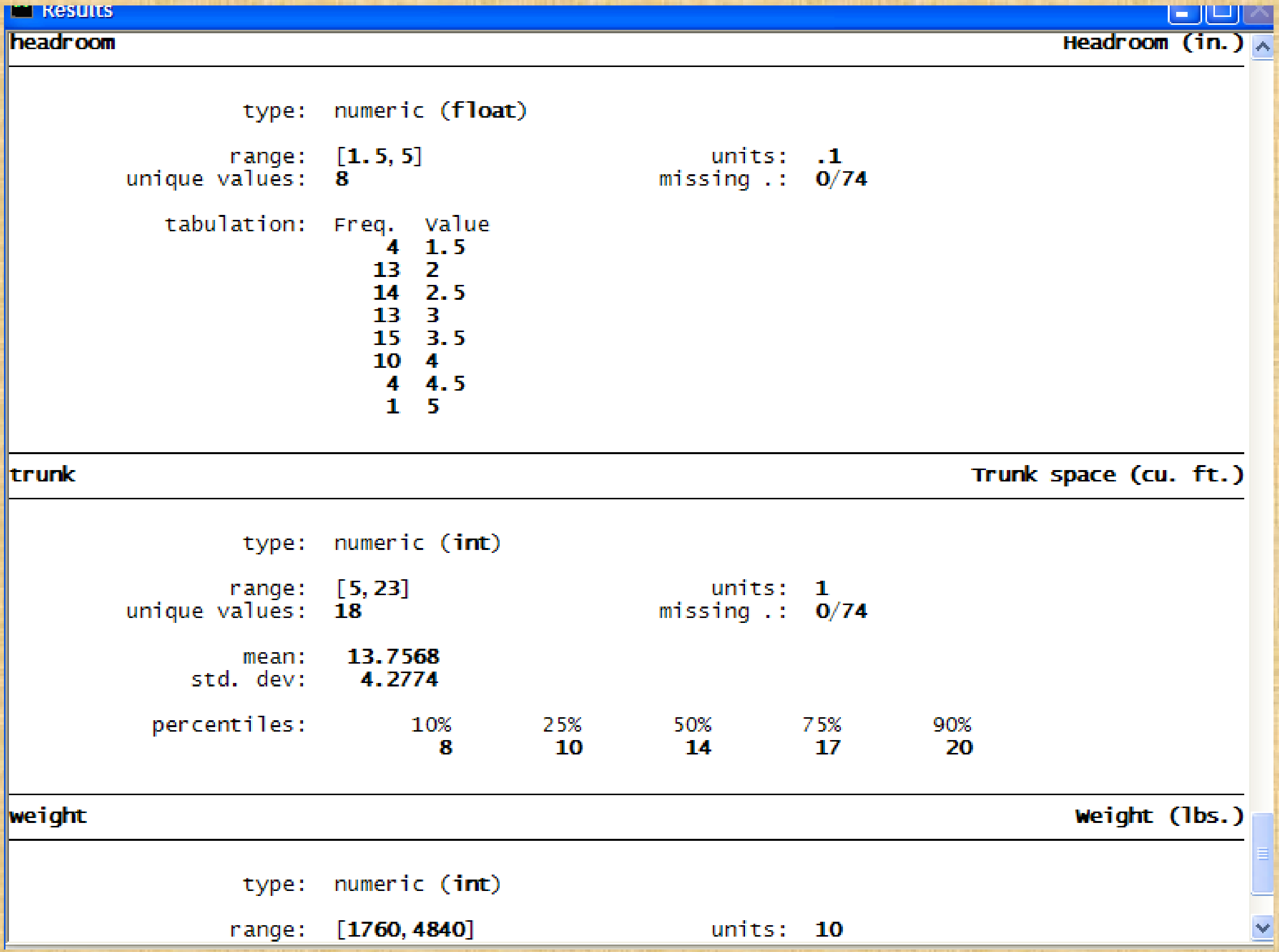
```
. inspect

make:  Make and Model                           Number of Observations
---------------------                       ------------------------------
                                            Total   Integers   Nonintegers
|                            Negative           -         -          -
|                            Zero               -         -          -
|                            Positive           -         -          -
|                                           -----     -----      -----
|
|                            Total              -         -          -
|                            Missing           74
+----------------------                     -----
.             -9.0e+307                        74
   (0 unique value)

price:  Price                                   Number of Observations
-------------                               ------------------------------
                                            Total   Integers   Nonintegers
|   #                        Negative           -         -          -
|   #                        Zero               -         -          -
|   #                        Positive          74        74          -
|   #                                       -----     -----      -----
|   #                        Total             74        74          -
|   #   #   .   .   .        Missing            -
+----------------------                     -----
3291              15906                        74
  (74 unique values)

mpg:  Mileage (mpg)                             Number of Observations
-------------------                         ------------------------------
                                            Total   Integers   Nonintegers
|      #                     Negative           -         -          -
|      #                     Zero               -         -          -
|      #                     Positive          74        74          -
|  #   #                                    -----     -----      -----
|  #   #   #                 Total             74        74          -
|  #   #   #   #   .         Missing            -
--more--
```

Inspect-continued



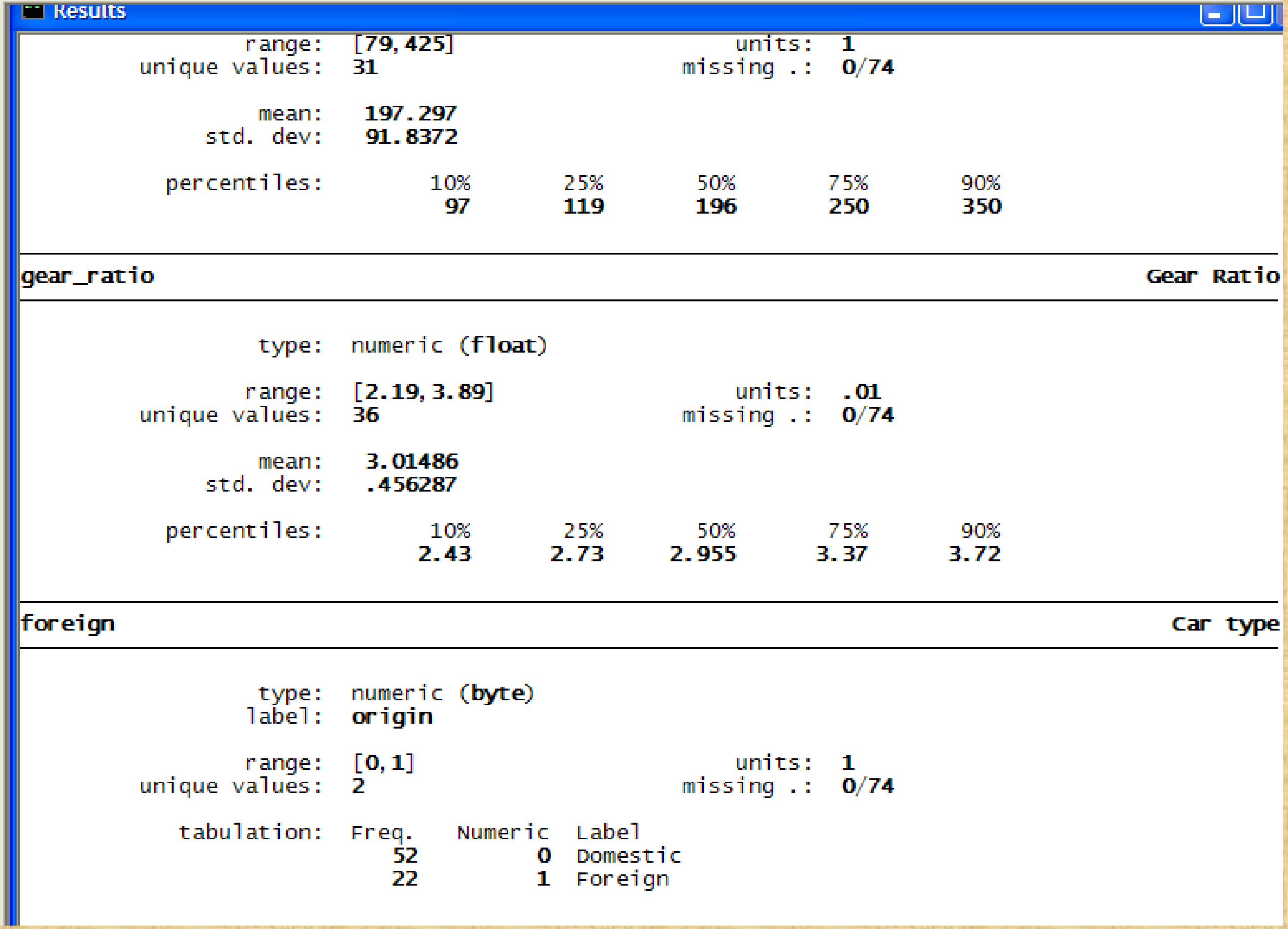
```
weight:  Weight (lbs.)                           Number of Observations
-----------------------                     -------------------------------
                                             Total   Integers   Nonintegers
|  #     #                  Negative            -          -            -
|  #     #                  Zero                -          -            -
|  #  #  #  #               Positive           74         74            -
|  #  #  #  #                                -----      -----        -----
|  #  #  #  #               Total              74         74            -
|  #  #  #  #  #            Missing             -
+-----------------------                     -----
1760              4840                         74
  (64 unique values)

length:  Length (in.)                            Number of Observations
-----------------------                     -------------------------------
                                             Total   Integers   Nonintegers
|           #               Negative            -          -            -
|     #     #               Zero                -          -            -
|     #     #               Positive           74         74            -
|     #  #  #                                -----      -----        -----
|  #  #  #  #  #            Total              74         74            -
|  #  #  #  #  #            Missing             -
+-----------------------                     -----
142                233                         74
  (47 unique values)

turn:  Turn Circle (ft.)                         Number of Observations
-------------------------                   -------------------------------
                                             Total   Integers   Nonintegers
|     #                     Negative            -          -            -
|     #     #               Zero                -          -            -
|     #  #  #               Positive           74         74            -
|     #  #  #                                -----      -----        -----
|  #  #  #  #               Total              74         74            -
|  #  #  #  #  .            Missing             -
+-----------------------                     -----
31                  51                         74
  (18 unique values)
```

codebook command



```
headroom                                                        Headroom (in.)

                  type:  numeric (float)

                 range:  [1.5,5]                      units:  .1
         unique values:  8                        missing .:  0/74

            tabulation:  Freq.   Value
                             4   1.5
                            13   2
                            14   2.5
                            13   3
                            15   3.5
                            10   4
                             4   4.5
                             1   5

trunk                                                      Trunk space (cu. ft.)

                  type:  numeric (int)

                 range:  [5,23]                       units:  1
         unique values:  18                       missing .:  0/74

                  mean:   13.7568
             std. dev:    4.2774

           percentiles:        10%       25%       50%       75%       90%
                                 8        10        14        17        20

weight                                                            Weight (lbs.)

                  type:  numeric (int)

                 range:  [1760,4840]                  units:  10
```

Codebook command-continued

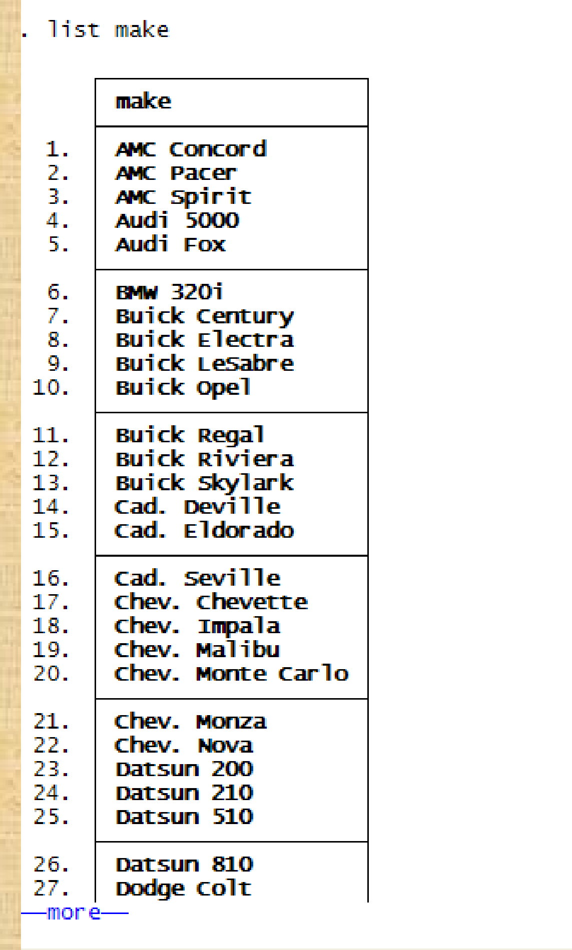


```
                 range:  [79,425]                     units:  1
         unique values:  31                       missing .:  0/74

                  mean:   197.297
              std. dev:   91.8372

           percentiles:        10%       25%       50%       75%       90%
                                97       119       196       250       350

------------------------------------------------------------------------------
gear_ratio                                                          Gear Ratio
------------------------------------------------------------------------------

                  type:  numeric (float)

                 range:  [2.19,3.89]                  units:  .01
         unique values:  36                       missing .:  0/74

                  mean:   3.01486
              std. dev:   .456287

           percentiles:        10%       25%       50%       75%       90%
                              2.43      2.73     2.955      3.37      3.72

------------------------------------------------------------------------------
foreign                                                               Car type
------------------------------------------------------------------------------

                  type:  numeric (byte)
                 label:  origin

                 range:  [0,1]                        units:  1
         unique values:  2                        missing .:  0/74

            tabulation:  Freq.   Numeric  Label
                            52         0  Domestic
                            22         1  Foreign
```

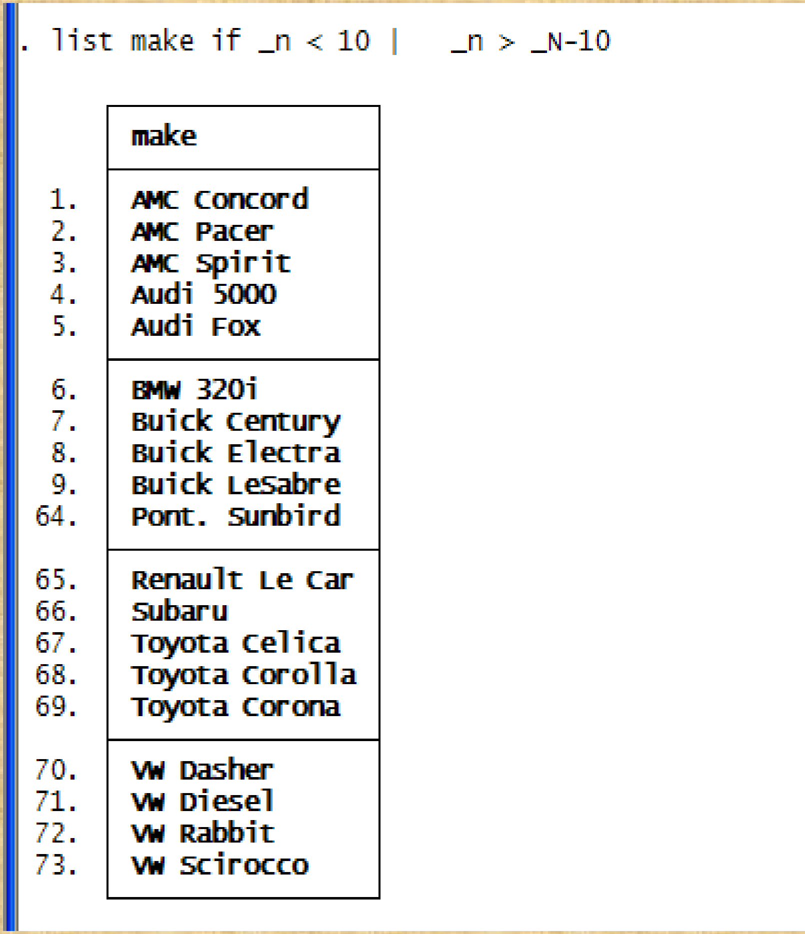
Data cleaning

- codebook
- inspect
- list
- assert
- count
- extremes
- duplicates
- format
- Missing functions
- Range checks with tabulate
- Consistency checks with correlate or pwcorr
- File comparison utilities

We can list out data



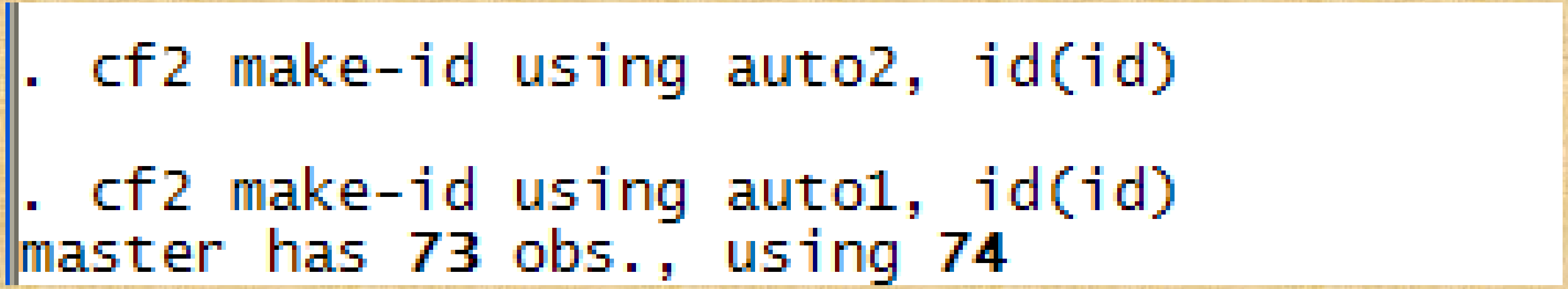
```
. list make

     +----------------+
     | make           |
     |----------------|
  1. | AMC Concord    |
  2. | AMC Pacer      |
  3. | AMC Spirit     |
  4. | Audi 5000      |
  5. | Audi Fox       |
     |----------------|
  6. | BMW 320i       |
  7. | Buick Century  |
  8. | Buick Electra  |
  9. | Buick LeSabre  |
 10. | Buick Opel     |
     |----------------|
 11. | Buick Regal    |
 12. | Buick Riviera  |
 13. | Buick Skylark  |
 14. | Cad. Deville   |
 15. | Cad. Eldorado  |
     |----------------|
 16. | Cad. Seville   |
 17. | Chev. Chevette |
 18. | Chev. Impala   |
 19. | Chev. Malibu   |
 20. | Chev. Monte Carlo |
     |----------------|
 21. | Chev. Monza    |
 22. | Chev. Nova     |
 23. | Datsun 200     |
 24. | Datsun 210     |
 25. | Datsun 510     |
     |----------------|
 26. | Datsun 810     |
 27. | Dodge Colt     |
—more—
```



```
. list make if _n < 10 |   _n > _N-10

     +-----------------+
     | make            |
     |-----------------|
  1. | AMC Concord     |
  2. | AMC Pacer       |
  3. | AMC Spirit      |
  4. | Audi 5000       |
  5. | Audi Fox        |
     |-----------------|
  6. | BMW 320i        |
  7. | Buick Century   |
  8. | Buick Electra   |
  9. | Buick LeSabre   |
 64. | Pont. Sunbird   |
     |-----------------|
 65. | Renault Le Car  |
 66. | Subaru          |
 67. | Toyota Celica   |
 68. | Toyota Corolla  |
 69. | Toyota Corona   |
     |-----------------|
 70. | VW Dasher       |
 71. | VW Diesel       |
 72. | VW Rabbit       |
 73. | VW Scirocco     |
     +-----------------+
```

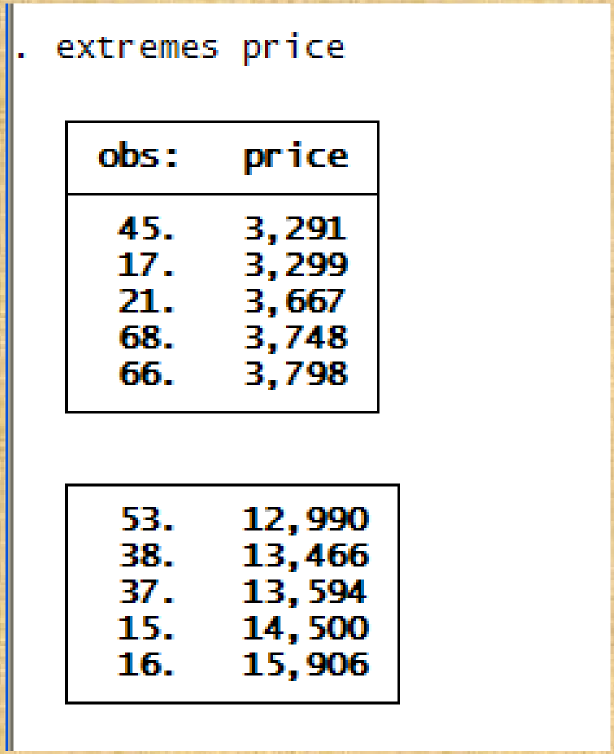
Double data entry and file comparison

- We can have 2 people enter the data and then compare the files to see if they differ.



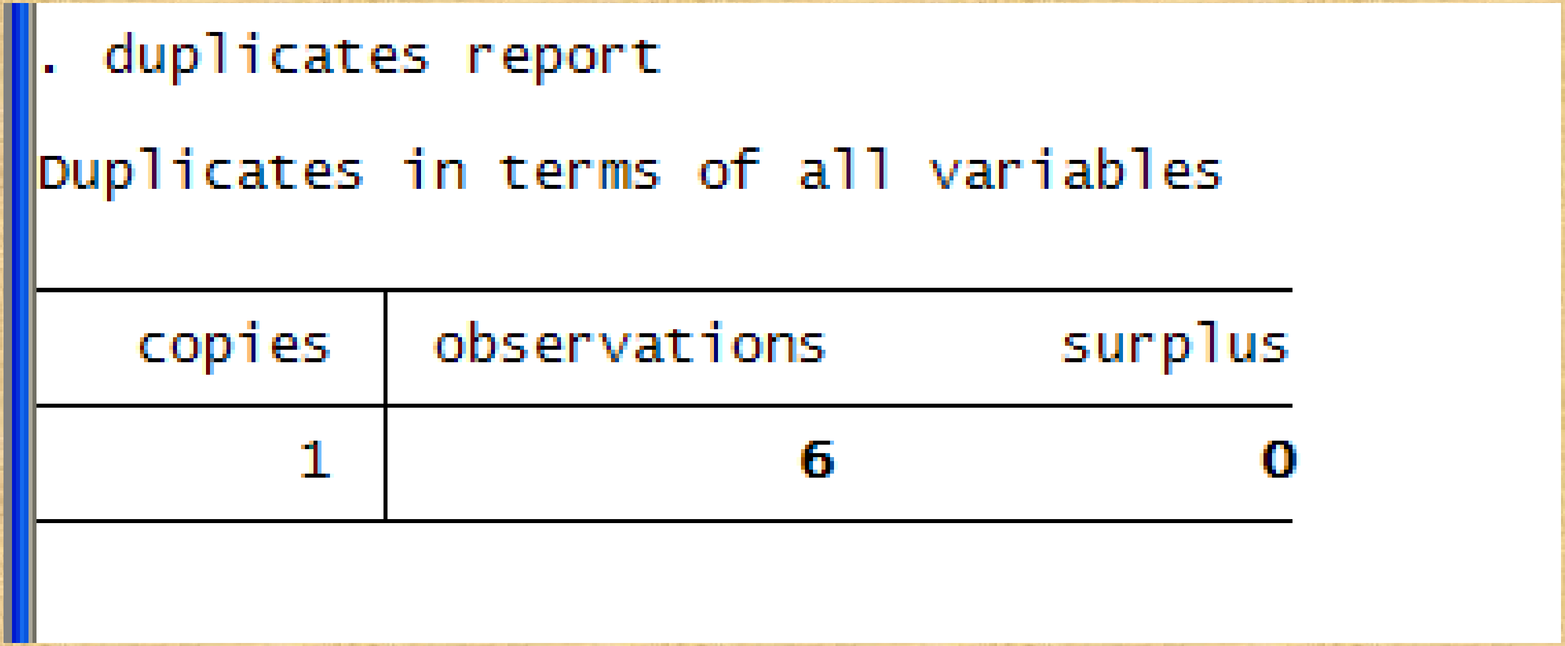
```
. cf2 make-id using auto2, id(id)

. cf2 make-id using auto1, id(id)
master has 73 obs., using 74
```

We can check for extreme values of a variable



```
. extremes price

  +----------------+
  | obs:     price |
  |----------------|
  |  45.     3,291 |
  |  17.     3,299 |
  |  21.     3,667 |
  |  68.     3,748 |
  |  66.     3,798 |
  +----------------+

  +-----------------+
  |  53.     12,990 |
  |  38.     13,466 |
  |  37.     13,594 |
  |  15.     14,500 |
  |  16.     15,906 |
  +-----------------+
```

Checking for duplication of observations: duplicates report

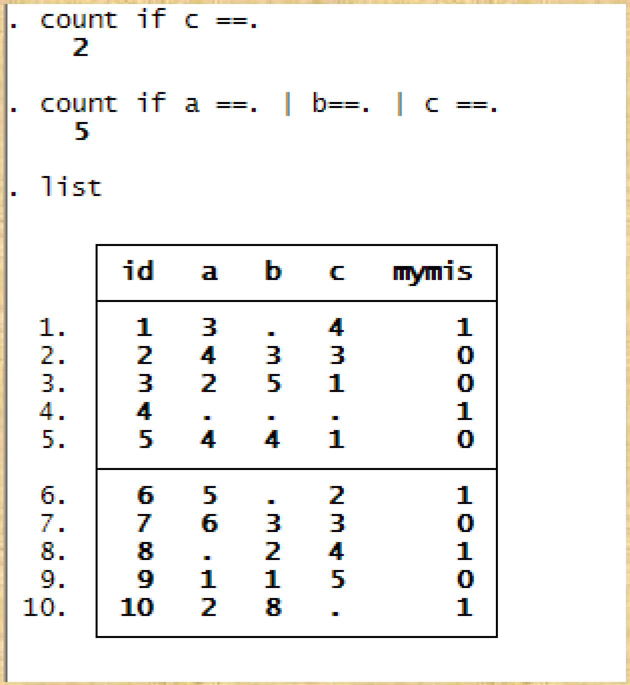


```
. duplicates report

Duplicates in terms of all variables

--------------------------------------
   copies | observations       surplus
----------+---------------------------
        1 |            6             0
--------------------------------------
```

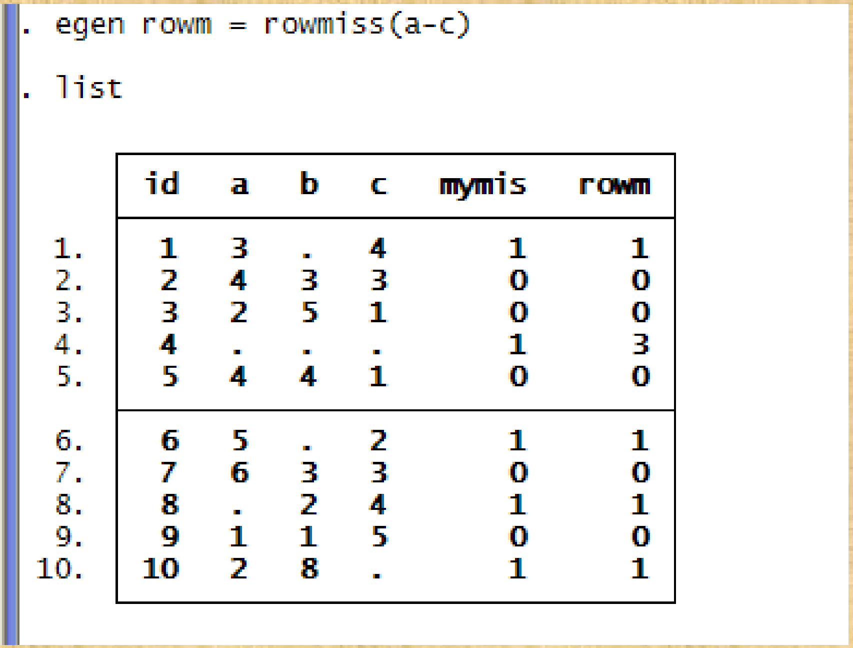
Missing values Review

- Be sure missing values are properly coded for your purposes.
- gen mymis= missing(var1-var10) constructs variable, mymis, which is coded 1 for a line on which any of these variables has a missing value and 0 for a case with no missing values.
- egen rowm = rowmis(var1-var10)
- count if a==.| b==. | c==.

Missing value functions-continued



```
. count if c ==.
  2

. count if a ==. | b==. | c ==.
  5

. list

     +-------------------------+
     | id   a   b   c    mymis |
     |-------------------------|
  1. |  1   3   .   4        1 |
  2. |  2   4   3   3        0 |
  3. |  3   2   5   1        0 |
  4. |  4   .   .   .        1 |
  5. |  5   4   4   1        0 |
     |-------------------------|
  6. |  6   5   .   2        1 |
  7. |  7   6   3   3        0 |
  8. |  8   .   2   4        1 |
  9. |  9   1   1   5        0 |
 10. | 10   2   8   .        1 |
     +-------------------------+
```

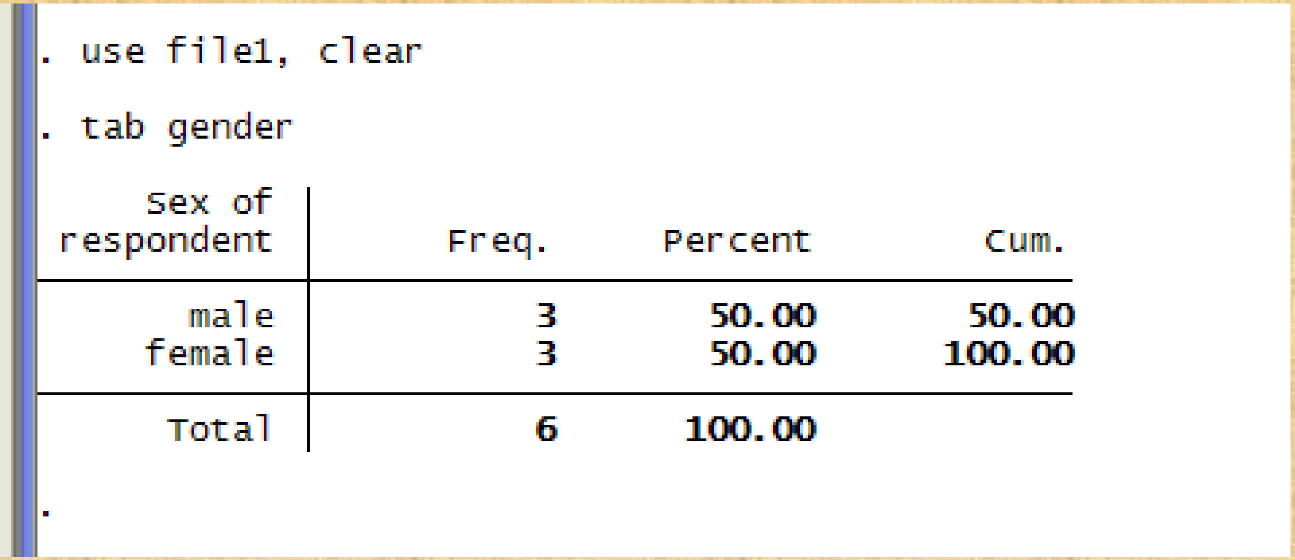


```
. egen rowm = rowmiss(a-c)

. list

     +--------------------------------+
     | id   a   b   c   mymis   rowm |
     |--------------------------------|
  1. |  1   3   .   4       1      1 |
  2. |  2   4   3   3       0      0 |
  3. |  3   2   5   1       0      0 |
  4. |  4   .   .   .       1      3 |
  5. |  5   4   4   1       0      0 |
     |--------------------------------|
  6. |  6   5   .   2       1      1 |
  7. |  7   6   3   3       0      0 |
  8. |  8   .   2   4       1      1 |
  9. |  9   1   1   5       0      0 |
 10. | 10   2   8   .       1      1 |
     +--------------------------------+
```

Inspect

- This will indicate the proportion of missing values and the numbers of them for each variable in the dataset.

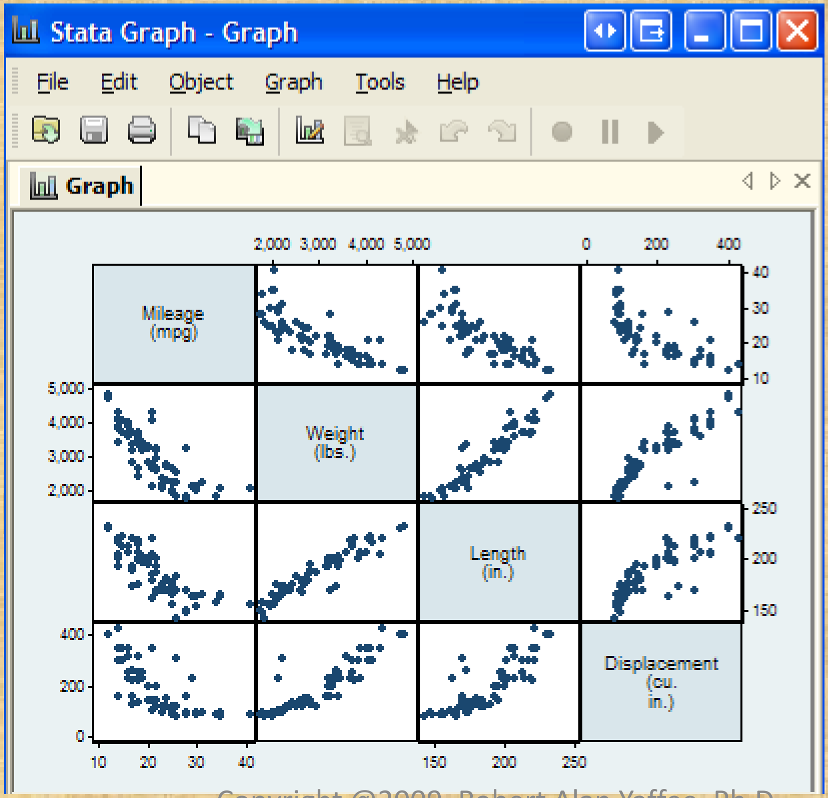
Range checks (with the tab command)



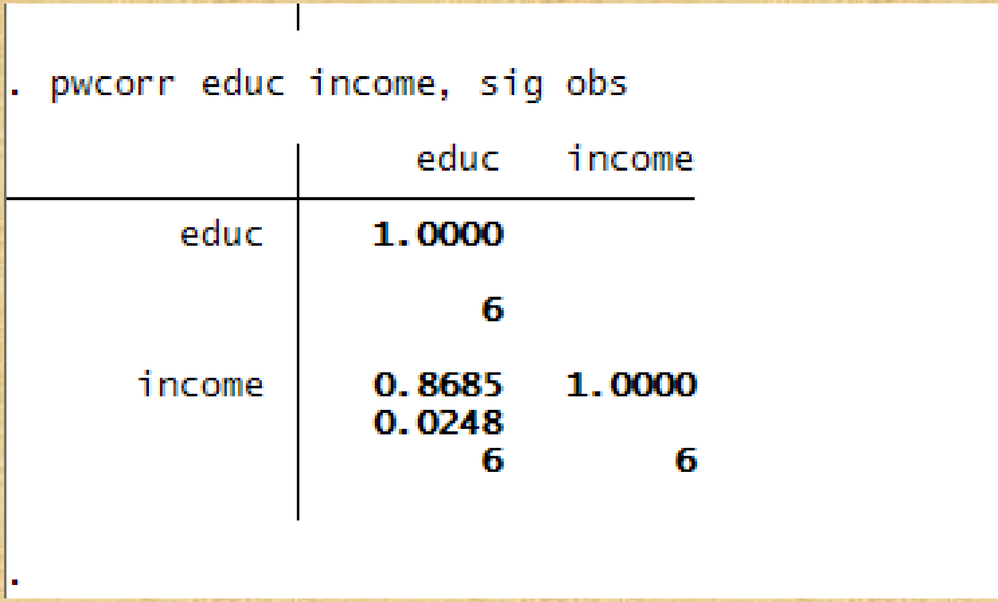
```
. use file1, clear

. tab gender

     Sex of |
 respondent |      Freq.     Percent        Cum.
------------+-----------------------------------
       male |          3       50.00       50.00
     female |          3       50.00      100.00
------------+-----------------------------------
      Total |          6      100.00

.
```

Graphical consistency checks with a matrix plot

```
. graph matrix mpg weight length displacement
```

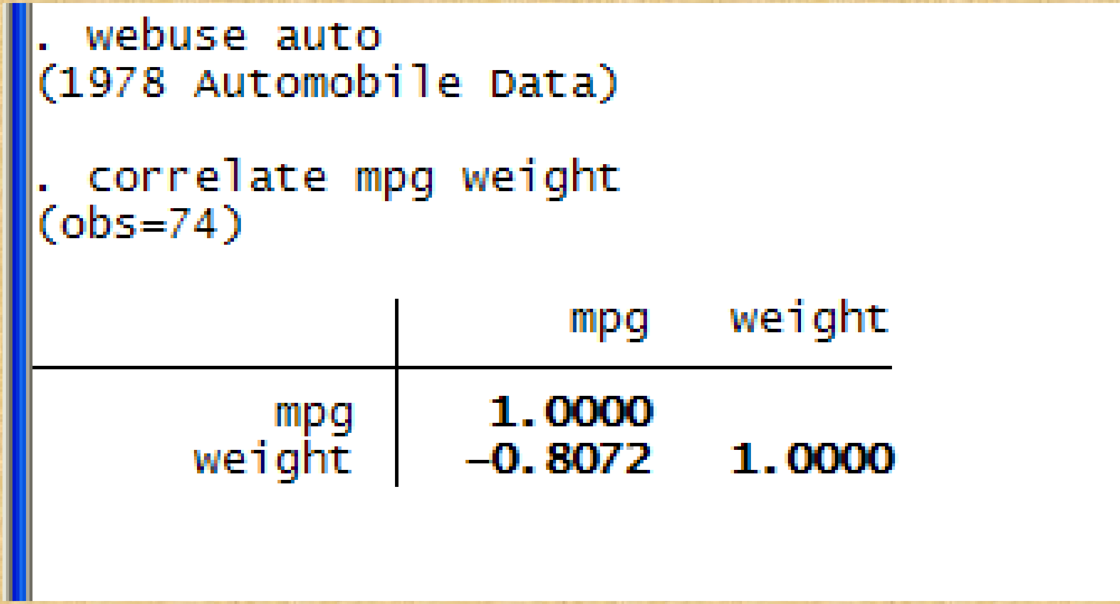


Consistency checks with pwcorr (pairwise Pearson correlations)



```
             |

. pwcorr educ income, sig obs

             |     educ   income
-------------+------------------
        educ |   1.0000
             |
             |        6
             |
      income |   0.8685   1.0000
             |   0.0248
             |        6        6
             |

.
```

Consistency checks with listwise correlate command



```
. webuse auto
(1978 Automobile Data)

. correlate mpg weight
(obs=74)

             |      mpg   weight
-------------+------------------
         mpg |   1.0000
      weight |  -0.8072   1.0000
```

Problems with bivariate correlations

- They are dependent on the levels of measurement of the variables to which they are applied.
- They do not detect nonlinear correlations.
- They do not detect influence of intervening variables.
- They do not detect the influence of antecedent variables.
- “All the world is multivariate (Edward Tufte)”
- They are not sufficient statistics. They are not adequate for an analysis.

Variable construction

1. Subset(conditional) if is used to qualify commands:

```
summarize if _n < 100, detail
```

2. Generate is used to create new variables

```
generate newvar=oldvar + 1
generate  dummy=0 if oldvar ~=.
```

Variable construction and transformation

1. Replace is used to recode existing variables

```
replace newvar = -9 if newva r==.
replace dummy = 1 if oldvar < 12 & oldvar ~=.
```

2. Egen command is used to construct variables across the rows of the dataset.

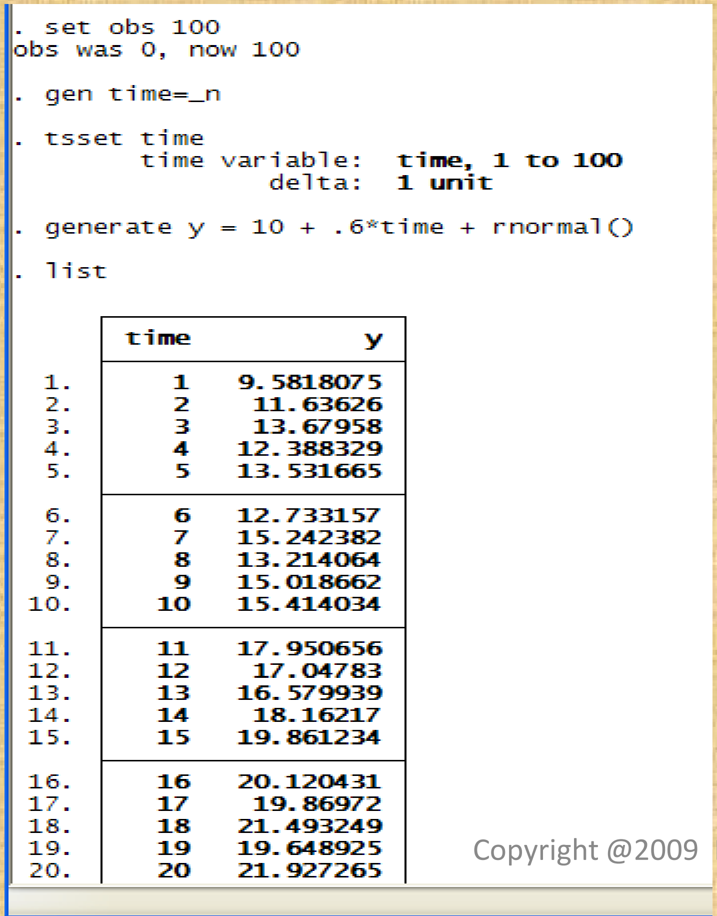
```
egen rownmis = rownonmissing(var1,var2,var3)
egen meanr= rowmean(var2,var4,var7)
egen maxr=rowmax(var4-var6)
egen sdr = rowsd(var5-var9)
egen rowtot = rowtotal(var1,var2,var3)
```

Variable construction and transformation

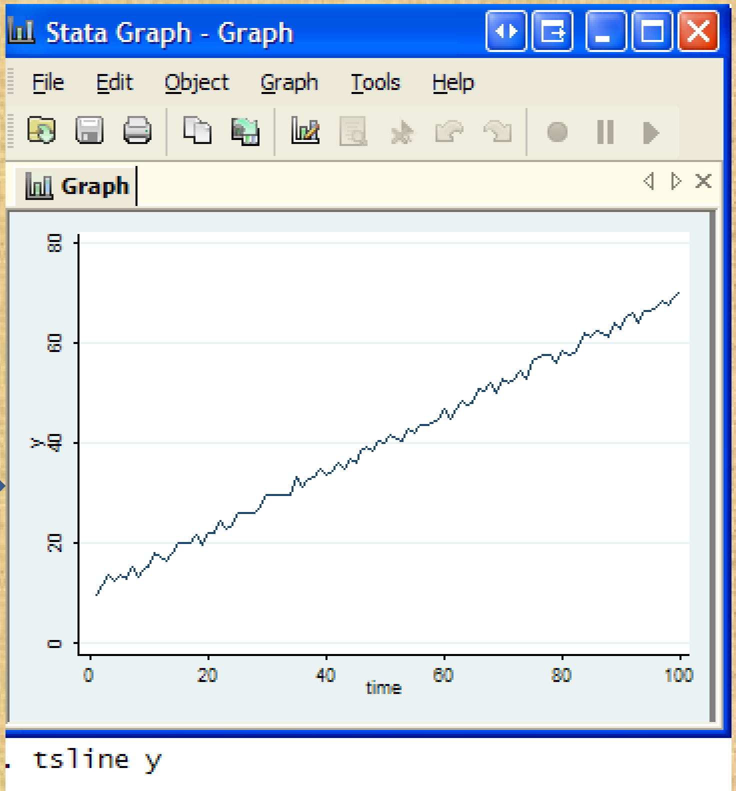
- Observation numbering with _n and _N

Indexing date – time variables for time series analysis

- You can index a dataset by time and construct a date (time) variable in order to perform time series analysis.

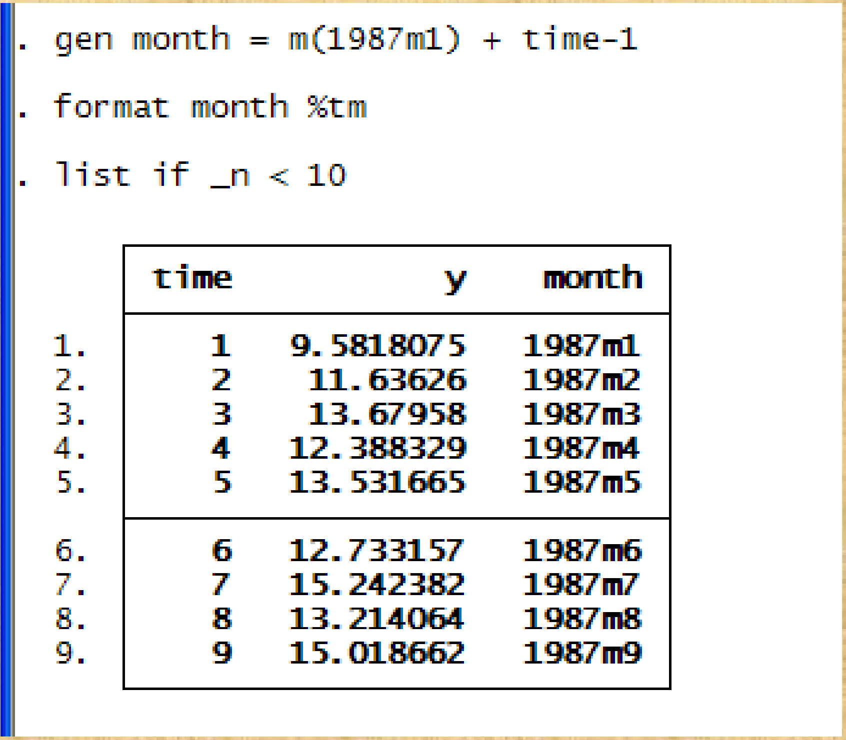


```
. set obs 100
obs was 0, now 100

. gen time=_n

. tsset time
        time variable:  time, 1 to 100
                delta:  1 unit

. generate y = 10 + .6*time + rnormal()

. list

     +------------------+
     | time           y |
     |------------------|
  1. |    1   9.5818075 |
  2. |    2    11.63626 |
  3. |    3    13.67958 |
  4. |    4   12.388329 |
  5. |    5   13.531665 |
     |------------------|
  6. |    6   12.733157 |
  7. |    7   15.242382 |
  8. |    8   13.214064 |
  9. |    9   15.018662 |
 10. |   10   15.414034 |
     |------------------|
 11. |   11   17.950656 |
 12. |   12    17.04783 |
 13. |   13   16.579939 |
 14. |   14    18.16217 |
 15. |   15   19.861234 |
     |------------------|
 16. |   16   20.120431 |
 17. |   17    19.86972 |
 18. |   18   21.493249 |
 19. |   19   19.648925 |
 20. |   20   21.927265 |
```

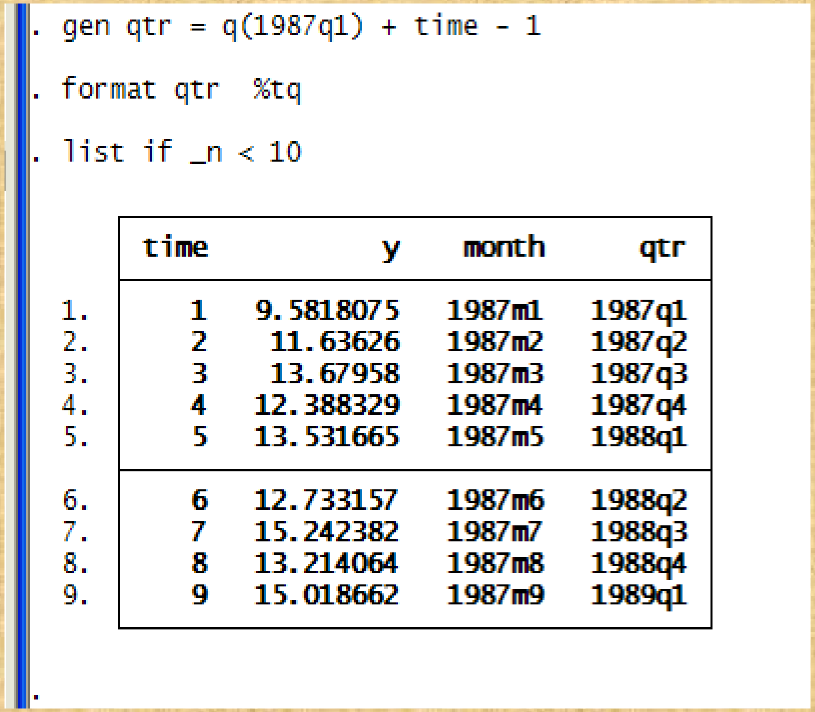


```
. tsline y
```

Time variable formats

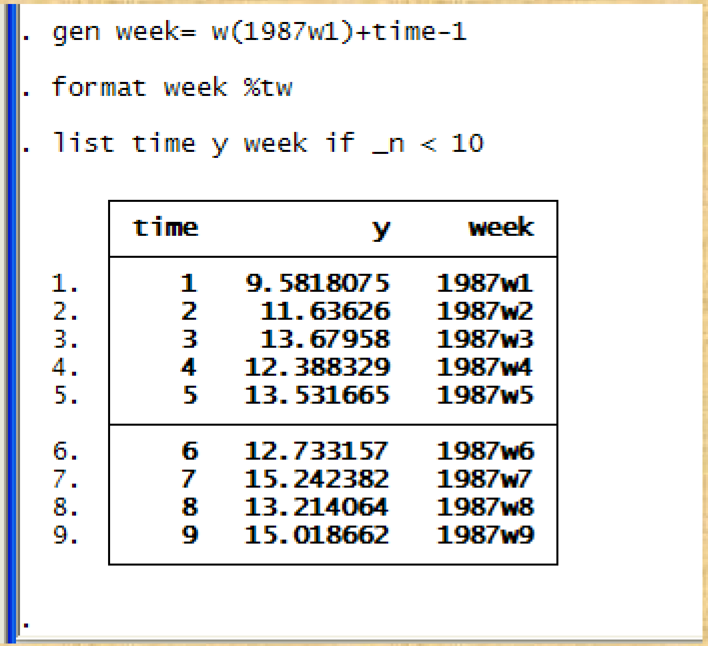


```
. gen month = m(1987m1) + time-1

. format month %tm

. list if _n < 10

     +-----------------------------+
     | time           y     month  |
     |-----------------------------|
  1. |    1   9.5818075    1987m1  |
  2. |    2    11.63626    1987m2  |
  3. |    3    13.67958    1987m3  |
  4. |    4   12.388329    1987m4  |
  5. |    5   13.531665    1987m5  |
     |-----------------------------|
  6. |    6   12.733157    1987m6  |
  7. |    7   15.242382    1987m7  |
  8. |    8   13.214064    1987m8  |
  9. |    9   15.018662    1987m9  |
     +-----------------------------+
```

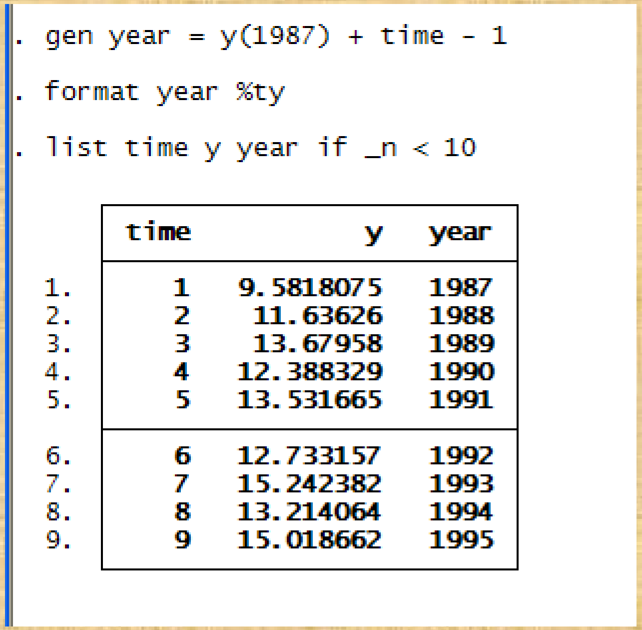


```
. gen qtr = q(1987q1) + time - 1

. format qtr  %tq

. list if _n < 10

     +--------------------------------------+
     | time           y    month       qtr  |
     |--------------------------------------|
  1. |    1   9.5818075   1987m1    1987q1  |
  2. |    2    11.63626   1987m2    1987q2  |
  3. |    3    13.67958   1987m3    1987q3  |
  4. |    4   12.388329   1987m4    1987q4  |
  5. |    5   13.531665   1987m5    1988q1  |
     |--------------------------------------|
  6. |    6   12.733157   1987m6    1988q2  |
  7. |    7   15.242382   1987m7    1988q3  |
  8. |    8   13.214064   1987m8    1988q4  |
  9. |    9   15.018662   1987m9    1989q1  |
     +--------------------------------------+

.
```

More time variable formats

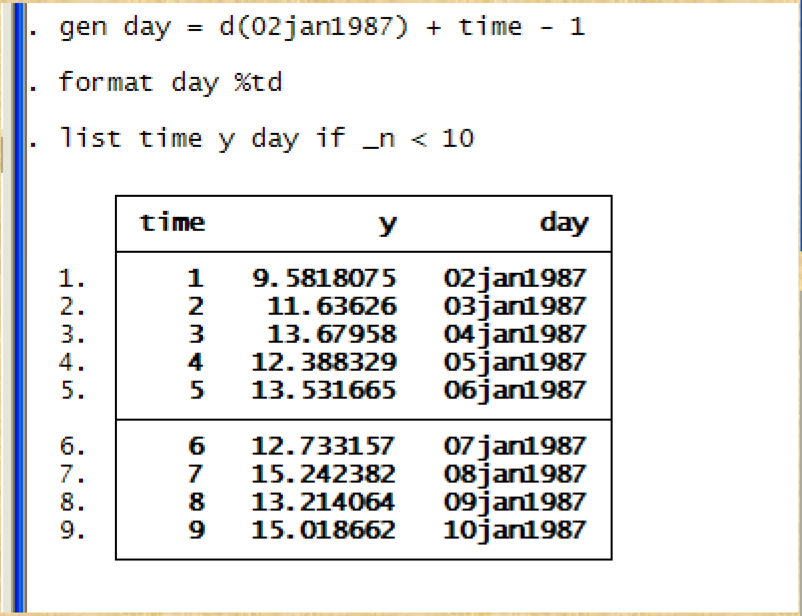


```
. gen week= w(1987w1)+time-1

. format week %tw

. list time y week if _n < 10

     +------------------------------+
     | time           y      week   |
     |------------------------------|
  1. |    1   9.5818075    1987w1   |
  2. |    2    11.63626    1987w2   |
  3. |    3    13.67958    1987w3   |
  4. |    4   12.388329    1987w4   |
  5. |    5   13.531665    1987w5   |
     |------------------------------|
  6. |    6   12.733157    1987w6   |
  7. |    7   15.242382    1987w7   |
  8. |    8   13.214064    1987w8   |
  9. |    9   15.018662    1987w9   |
     +------------------------------+

.
```

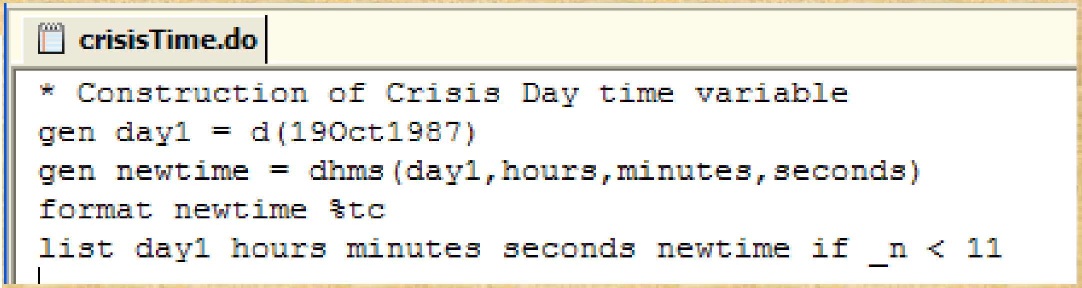


```
. gen year = y(1987) + time - 1

. format year %ty

. list time y year if _n < 10

     +----------------------------+
     | time           y    year   |
     |----------------------------|
  1. |    1   9.5818075    1987   |
  2. |    2    11.63626    1988   |
  3. |    3    13.67958    1989   |
  4. |    4   12.388329    1990   |
  5. |    5   13.531665    1991   |
     |----------------------------|
  6. |    6   12.733157    1992   |
  7. |    7   15.242382    1993   |
  8. |    8   13.214064    1994   |
  9. |    9   15.018662    1995   |
     +----------------------------+
```

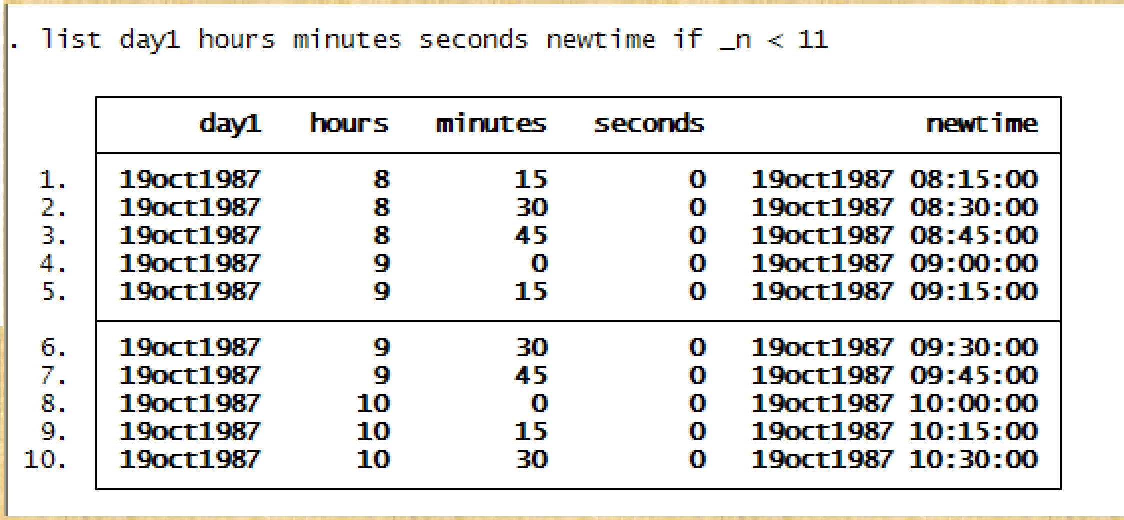
Time variables



```
. gen day = d(02jan1987) + time - 1

. format day %td

. list time y day if _n < 10

     +-----------------------------+
     | time          y         day |
     |-----------------------------|
  1. |    1   9.5818075   02jan1987 |
  2. |    2    11.63626   03jan1987 |
  3. |    3    13.67958   04jan1987 |
  4. |    4   12.388329   05jan1987 |
  5. |    5   13.531665   06jan1987 |
     |-----------------------------|
  6. |    6   12.733157   07jan1987 |
  7. |    7   15.242382   08jan1987 |
  8. |    8   13.214064   09jan1987 |
  9. |    9   15.018662   10jan1987 |
     +-----------------------------+
```



```
* Construction of Crisis Day time variable
gen day1 = d(19Oct1987)
gen newtime = dhms(day1,hours,minutes,seconds)
format newtime %tc
list day1 hours minutes seconds newtime if _n < 11
```

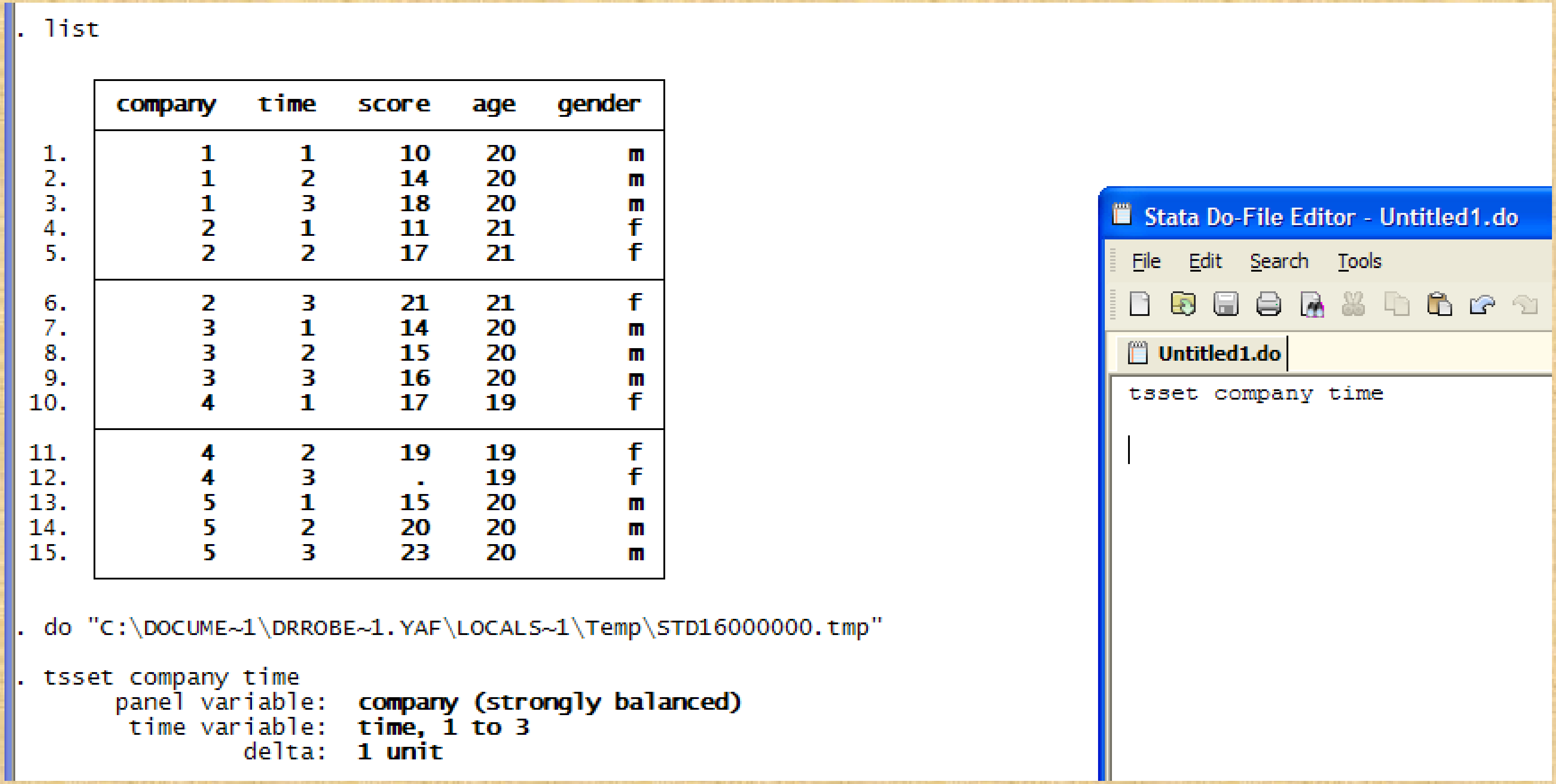


```
. list day1 hours minutes seconds newtime if _n < 11

     +-------------------------------------------------------------------+
     |      day1   hours   minutes   seconds                    newtime |
     |-------------------------------------------------------------------|
  1. | 19oct1987       8        15         0   19oct1987 08:15:00 |
  2. | 19oct1987       8        30         0   19oct1987 08:30:00 |
  3. | 19oct1987       8        45         0   19oct1987 08:45:00 |
  4. | 19oct1987       9         0         0   19oct1987 09:00:00 |
  5. | 19oct1987       9        15         0   19oct1987 09:15:00 |
     |-------------------------------------------------------------------|
  6. | 19oct1987       9        30         0   19oct1987 09:30:00 |
  7. | 19oct1987       9        45         0   19oct1987 09:45:00 |
  8. | 19oct1987      10         0         0   19oct1987 10:00:00 |
  9. | 19oct1987      10        15         0   19oct1987 10:15:00 |
 10. | 19oct1987      10        30         0   19oct1987 10:30:00 |
     +-------------------------------------------------------------------+
```

Indexing the observations by time

- After the time variable is formatted,
- Type: tsset ‘name of time variable’   (tip: don’t use the quotes)
- Then type: tsline ‘name of variable to analyze’
- Add a title to the graph
- tsline gdp, title(time plot of GDP)

Indexing Panel datasets



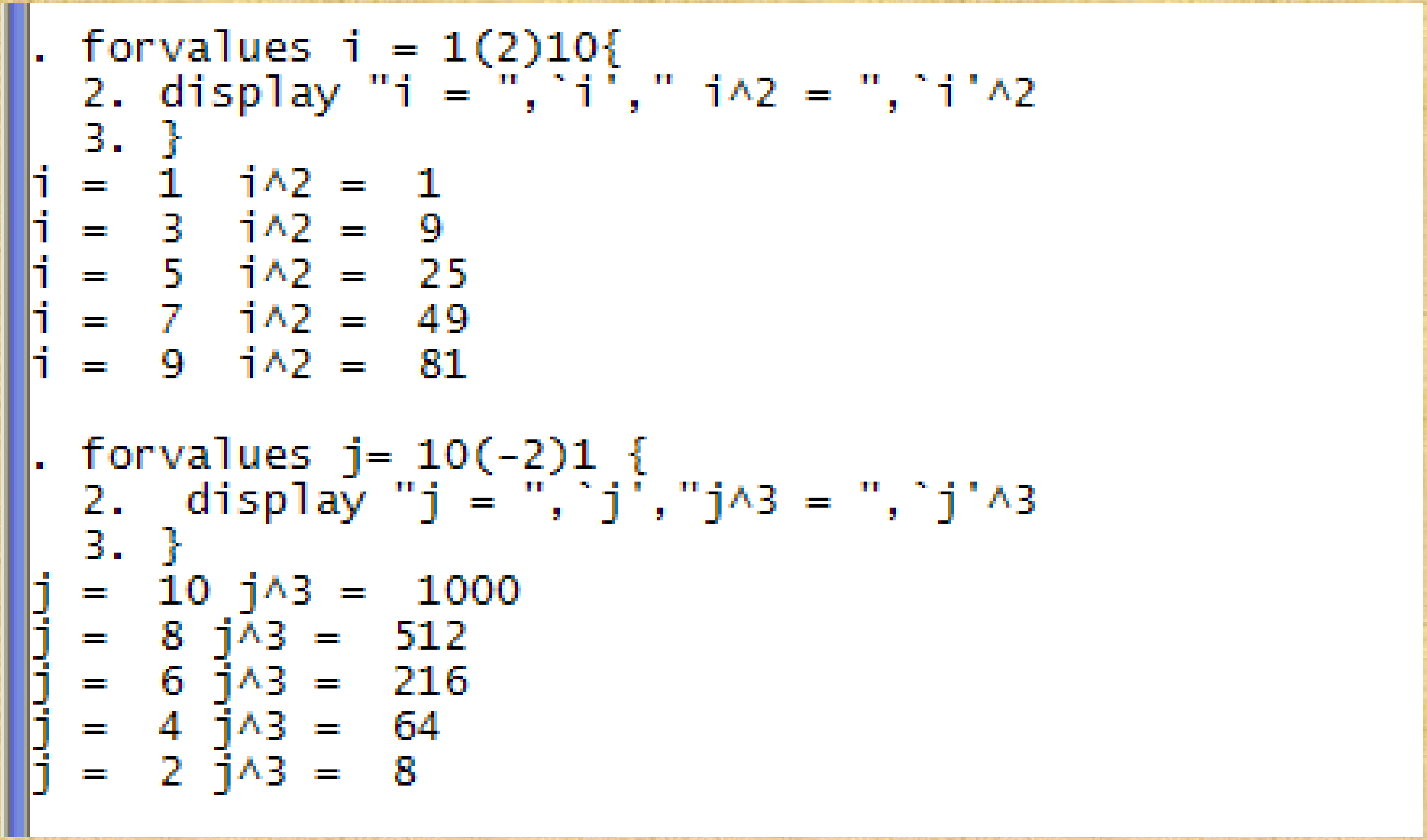
```
. list

     +-----------------------------------------+
     | company   time   score   age   gender |
     |-----------------------------------------|
  1. |       1      1      10    20        m |
  2. |       1      2      14    20        m |
  3. |       1      3      18    20        m |
  4. |       2      1      11    21        f |
  5. |       2      2      17    21        f |
     |-----------------------------------------|
  6. |       2      3      21    21        f |
  7. |       3      1      14    20        m |
  8. |       3      2      15    20        m |
  9. |       3      3      16    20        m |
 10. |       4      1      17    19        f |
     |-----------------------------------------|
 11. |       4      2      19    19        f |
 12. |       4      3       .    19        f |
 13. |       5      1      15    20        m |
 14. |       5      2      20    20        m |
 15. |       5      3      23    20        m |
     +-----------------------------------------+

. do "C:\DOCUME~1\DRROBE~1.YAF\LOCALS~1\Temp\STD16000000.tmp"

. tsset company time
       panel variable:  company (strongly balanced)
        time variable:  time, 1 to 3
                delta:  1 unit
```

```
tsset company time
```

Loop programming

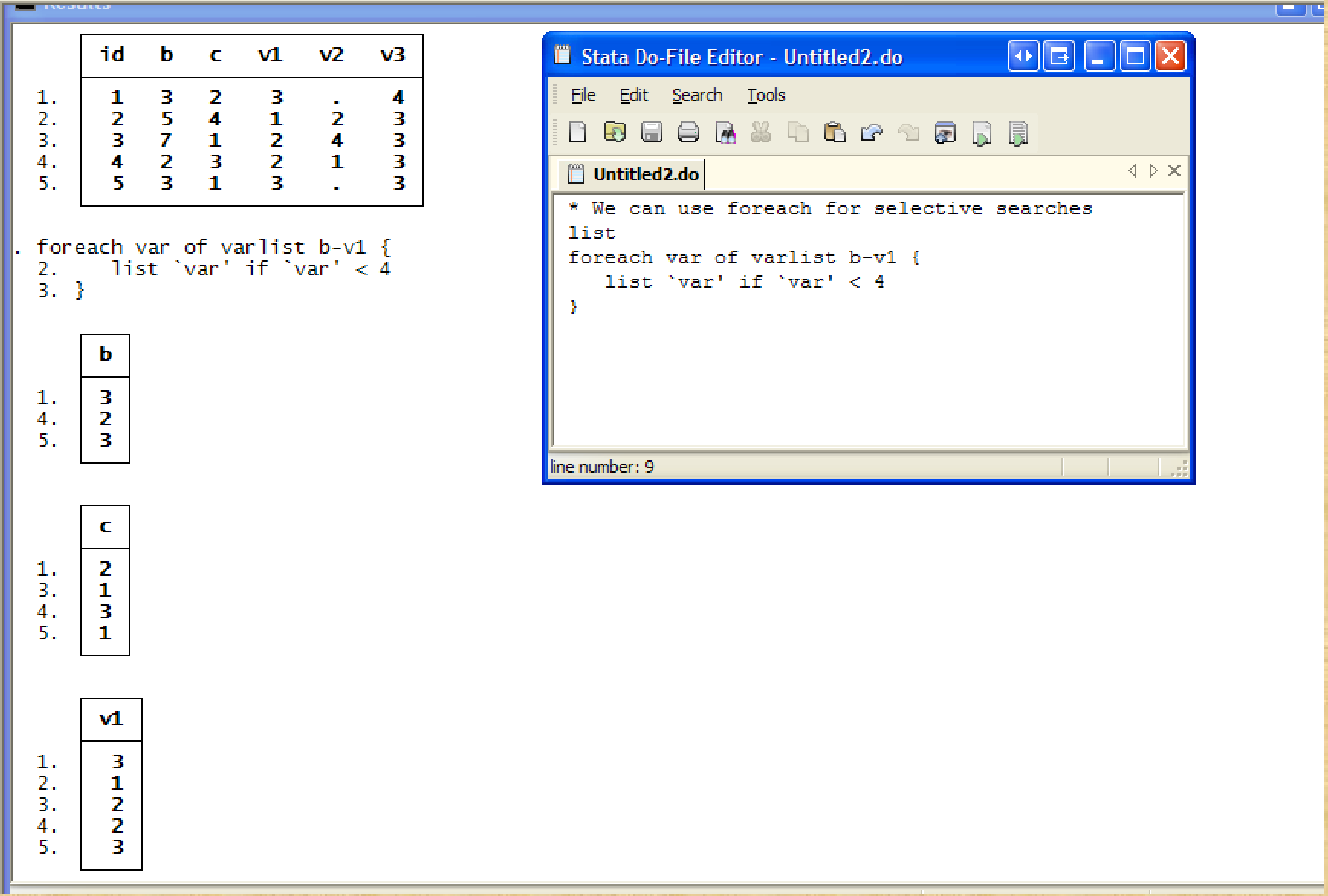
- For recodes
- For aggregation
- For simulation

The forvalues loop
for looping over consecutive values



```
. forvalues i = 1(2)10{
  2. display "i = ",`i'," i^2 = ",`i'^2
  3. }
i =  1  i^2 =  1
i =  3  i^2 =  9
i =  5  i^2 =  25
i =  7  i^2 =  49
i =  9  i^2 =  81

. forvalues j= 10(-2)1 {
  2.  display "j = ",`j',"j^3 = ",`j'^3
  3. }
j =  10 j^3 =  1000
j =  8 j^3 =  512
j =  6 j^3 =  216
j =  4 j^3 =  64
j =  2 j^3 =  8
```

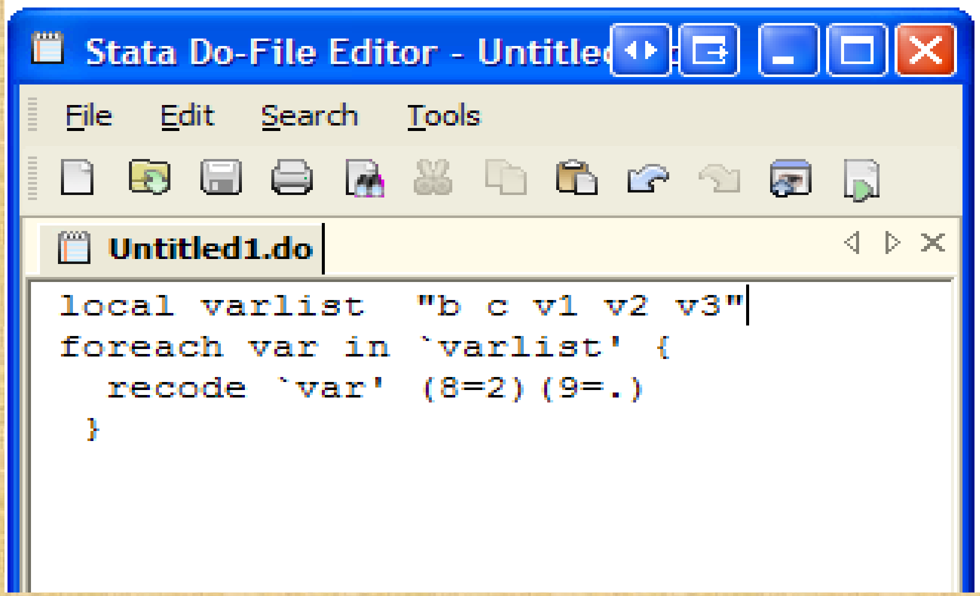
The foreach loop



```
     id   b   c   v1   v2   v3

 1.   1   3   2    3    .    4
 2.   2   5   4    1    2    3
 3.   3   7   1    2    4    3
 4.   4   2   3    2    1    3
 5.   5   3   1    3    .    3

. foreach var of varlist b-v1 {
  2.      list `var' if `var' < 4
  3. }

      b

 1.   3
 4.   2
 5.   3

      c

 1.   2
 3.   1
 4.   3
 5.   1

      v1

 1.    3
 2.    1
 3.    2
 4.    2
 5.    3
```

```
* We can use foreach for selective searches
list
foreach var of varlist b-v1 {
   list `var' if `var' < 4
}
```

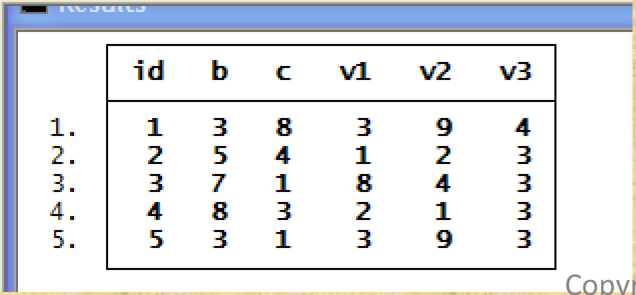
The foreach Loop



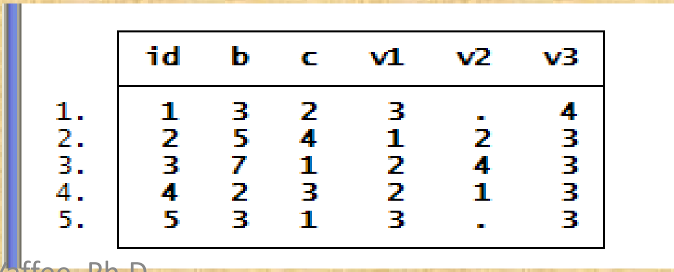
```
local varlist  "b c v1 v2 v3"
foreach var in `varlist' {
  recode `var' (8=2)(9=.)
 }
```

Loop.dta

Converts the data set on the right to that on the left.

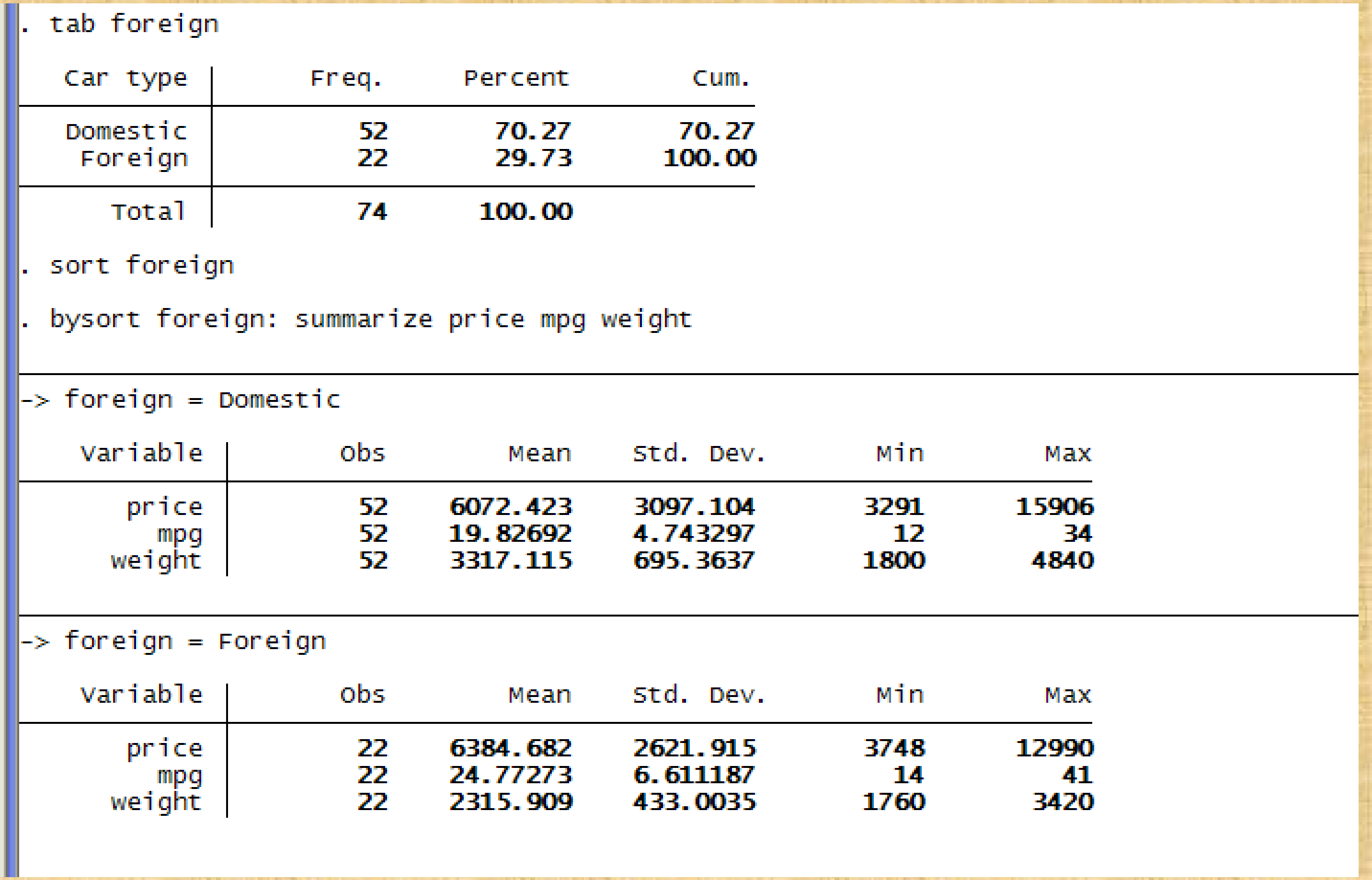


	id	b	c	v1	v2	v3
1.	1	3	8	3	9	4
2.	2	5	4	1	2	3
3.	3	7	1	8	4	3
4.	4	8	3	2	1	3
5.	5	3	1	3	9	3



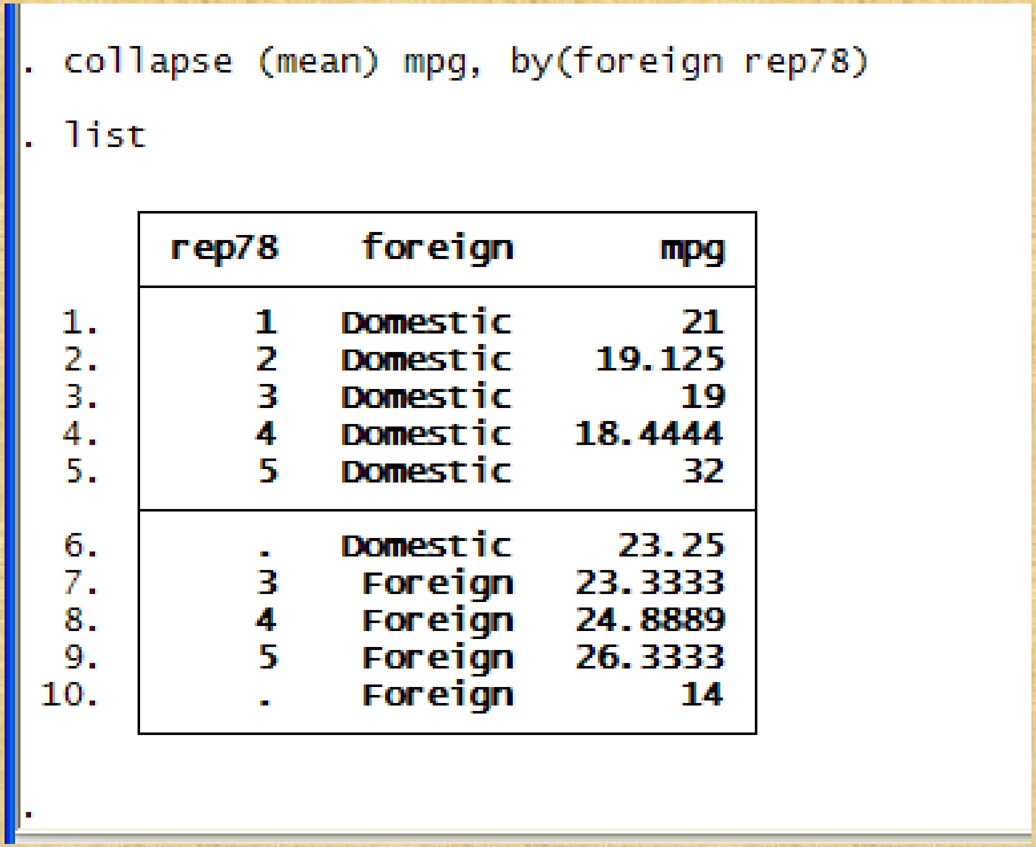
	id	b	c	v1	v2	v3
1.	1	3	2	3	.	4
2.	2	5	4	1	2	3
3.	3	7	1	2	4	3
4.	4	2	3	2	1	3
5.	5	3	1	3	.	3

conditional statistics: statistics by groups



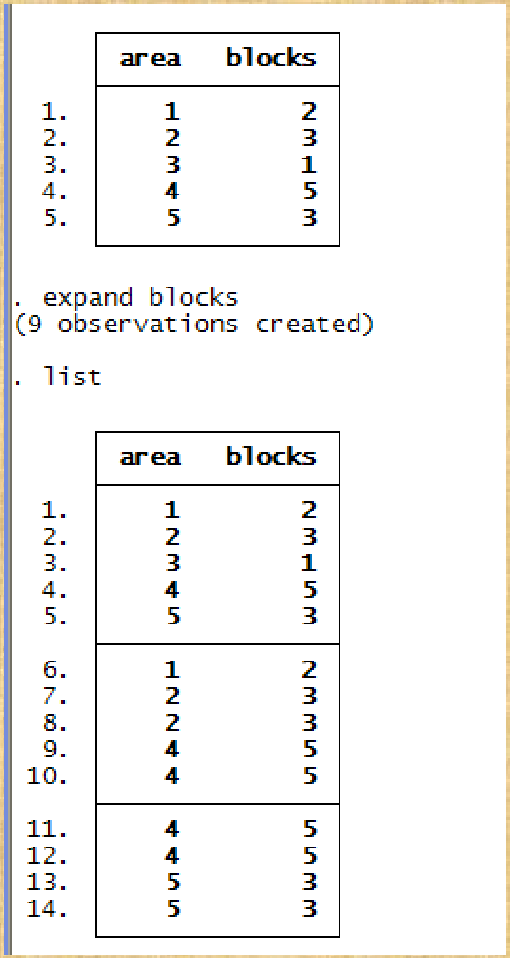
```
. tab foreign

   Car type |      Freq.     Percent        Cum.
------------+-----------------------------------
   Domestic |         52       70.27       70.27
    Foreign |         22       29.73      100.00
------------+-----------------------------------
      Total |         74      100.00

. sort foreign

. bysort foreign: summarize price mpg weight

-------------------------------------------------------------------------------------------
-> foreign = Domestic

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
       price |        52    6072.423    3097.104       3291      15906
         mpg |        52    19.82692    4.743297         12         34
      weight |        52    3317.115    695.3637       1800       4840

-------------------------------------------------------------------------------------------
-> foreign = Foreign

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
       price |        22    6384.682    2621.915       3748      12990
         mpg |        22    24.77273    6.611187         14         41
      weight |        22    2315.909    433.0035       1760       3420
```

Collapse command for aggregating datasets

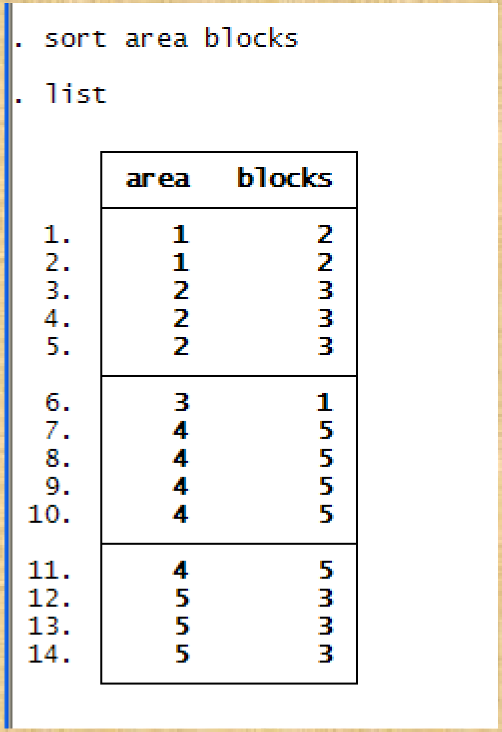


```
. collapse (mean) mpg, by(foreign rep78)

. list

     +----------------------------+
     | rep78    foreign       mpg |
     |----------------------------|
  1. |     1   Domestic        21 |
  2. |     2   Domestic    19.125 |
  3. |     3   Domestic        19 |
  4. |     4   Domestic   18.4444 |
  5. |     5   Domestic        32 |
     |----------------------------|
  6. |     .   Domestic     23.25 |
  7. |     3    Foreign   23.3333 |
  8. |     4    Foreign   24.8889 |
  9. |     5    Foreign   26.3333 |
 10. |     .    Foreign        14 |
     +----------------------------+

.
```

Expand for elaboration by a subdivision

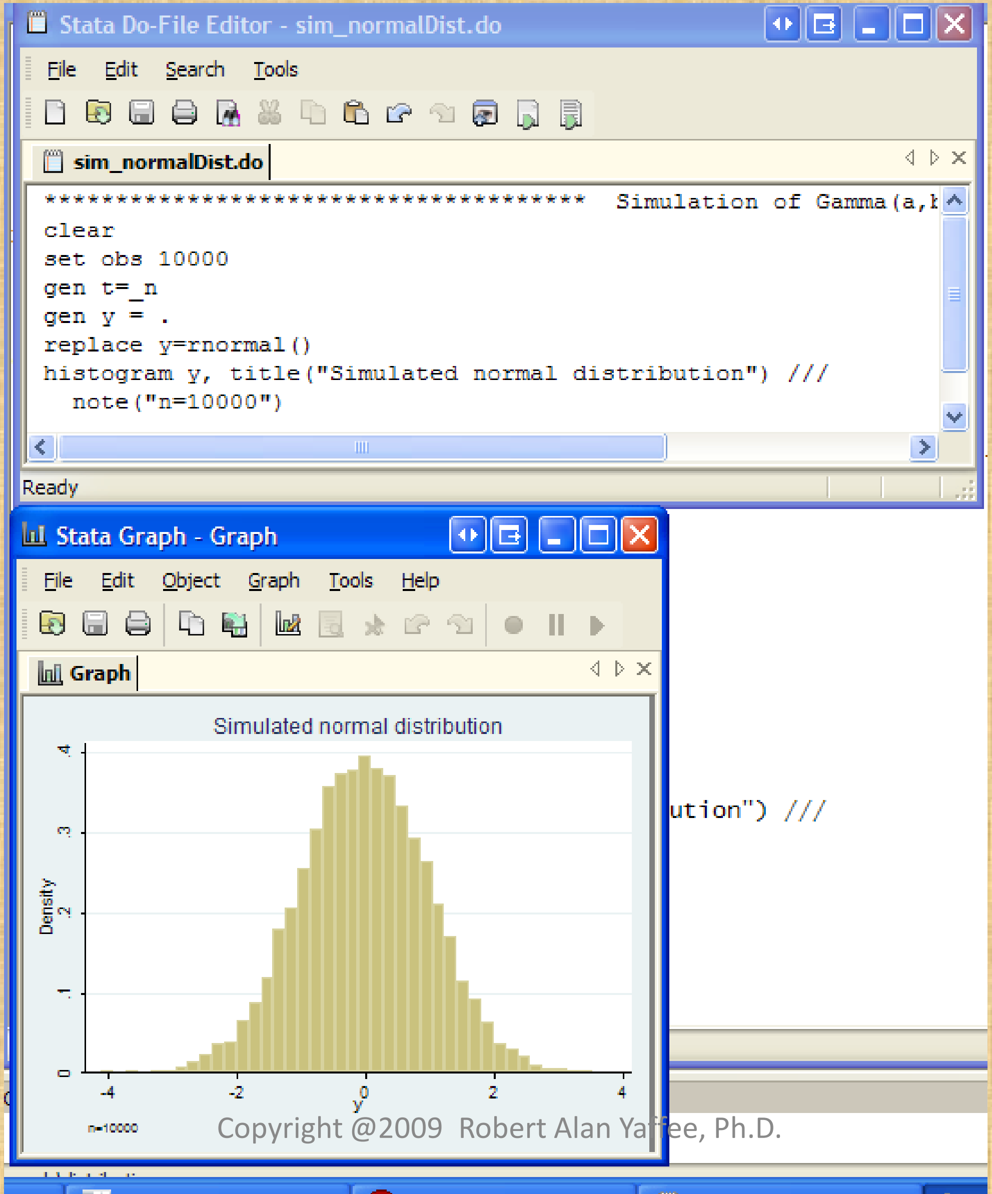


```
     +----------------+
     | area    blocks |
     |----------------|
  1. |    1         2 |
  2. |    2         3 |
  3. |    3         1 |
  4. |    4         5 |
  5. |    5         3 |
     +----------------+

. expand blocks
(9 observations created)

. list

     +----------------+
     | area    blocks |
     |----------------|
  1. |    1         2 |
  2. |    2         3 |
  3. |    3         1 |
  4. |    4         5 |
  5. |    5         3 |
     |----------------|
  6. |    1         2 |
  7. |    2         3 |
  8. |    2         3 |
  9. |    4         5 |
 10. |    4         5 |
     |----------------|
 11. |    4         5 |
 12. |    4         5 |
 13. |    5         3 |
 14. |    5         3 |
     +----------------+
```



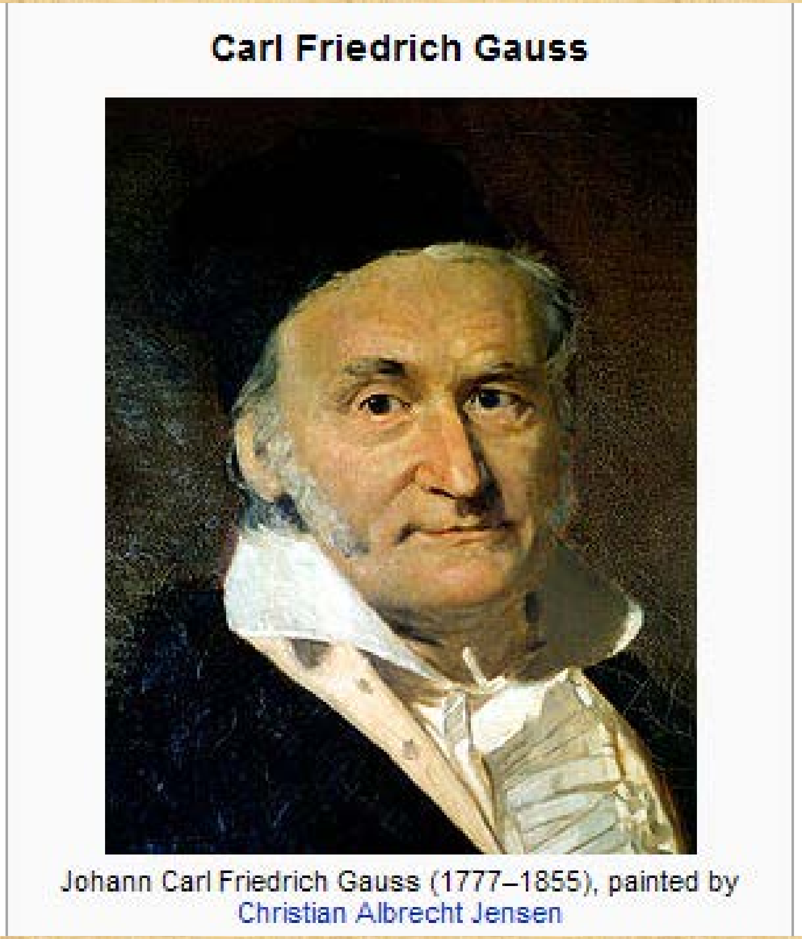
```
. sort area blocks

. list

     +----------------+
     | area    blocks |
     |----------------|
  1. |    1         2 |
  2. |    1         2 |
  3. |    2         3 |
  4. |    2         3 |
  5. |    2         3 |
     |----------------|
  6. |    3         1 |
  7. |    4         5 |
  8. |    4         5 |
  9. |    4         5 |
 10. |    4         5 |
     |----------------|
 11. |    4         5 |
 12. |    5         3 |
 13. |    5         3 |
 14. |    5         3 |
     +----------------+
```

Simulation of distributions



```
************************************** Simulation of Gamma(a,
clear
set obs 10000
gen t=_n
gen y = .
replace y=rnormal()
histogram y, title("Simulated normal distribution") ///
  note("n=10000")
```

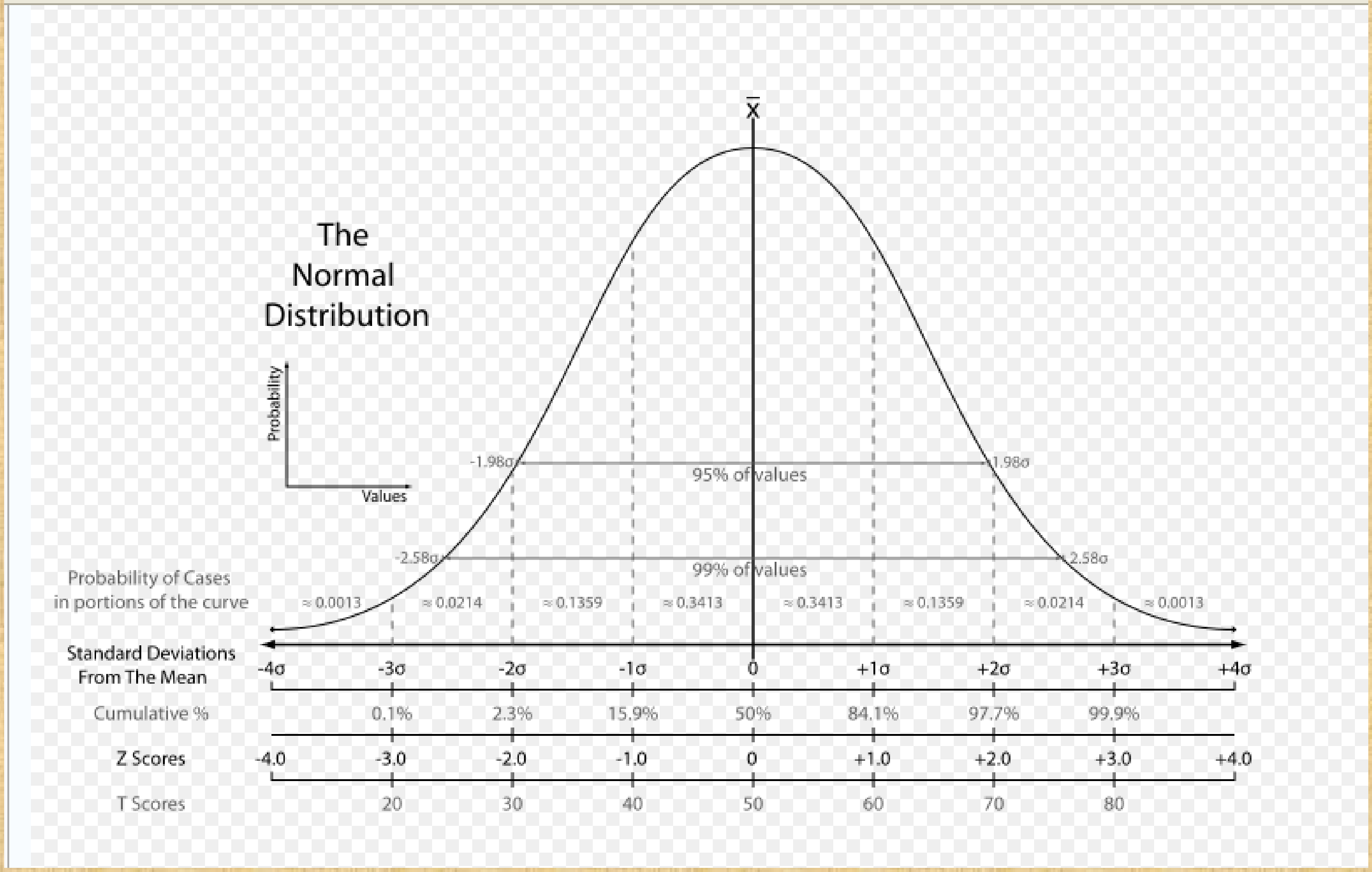
The Father of the Gaussian Distribution



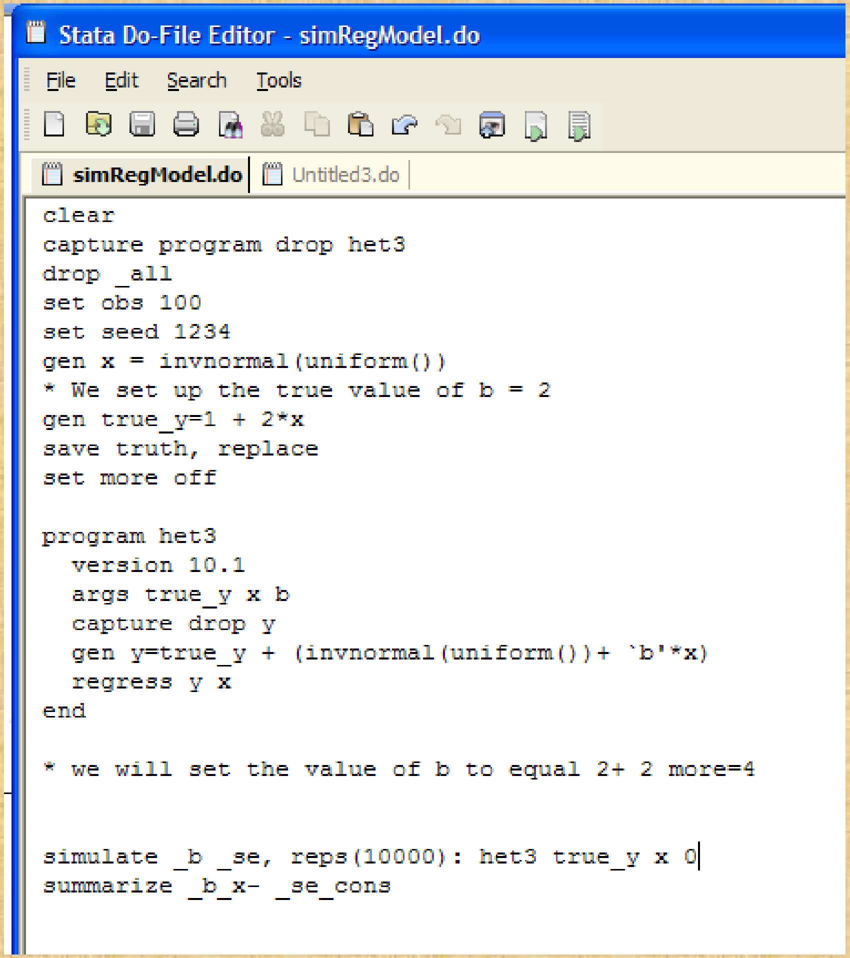
Carl Friedrich Gauss

Johann Carl Friedrich Gauss (1777–1855), painted by Christian Albrecht Jensen

Normal Distribution
Gaussian Distribution per C. F. Gauss)



Monte Carlo Simulations



```
clear
capture program drop het3
drop _all
set obs 100
set seed 1234
gen x = invnormal(uniform())
* We set up the true value of b = 2
gen true_y=1 + 2*x
save truth, replace
set more off

program het3
  version 10.1
  args true_y x b
  capture drop y
  gen y=true_y + (invnormal(uniform())+ `b'*x)
  regress y x
end

* we will set the value of b to equal 2+ 2 more=4


simulate _b _se, reps(10000): het3 true_y x 0
summarize _b_x- _se_cons
```

Exercises 2

1. Construct two toy datasets, and merge them by a common id variable.
2. Concatenate those two datasets.
3. Download a dataset from ucla.ats from yahoo
4. Convert the dataset to a Stata dataset
5. Graph the series
6. Using the loop.dta dataset and a foreach loop in Stata, convert all values > 7 to missing

Exercises 2

7. Construct a log of your work and then
8. Convert the blood pressure dataset, bpwide.dta to a long dataset
9. Convert the wide dataset to a long one,
10. Close the log file
11. Convert the smcl file to a txt file
12. Reshape the reshapeW.dta file to a reshape long form.

Exercises 2

12. Use the dataset auto1.dta. Obtain the means of the mpg and weight by foreign to obtain aggregate statistics.
13. Plot a histogram of the mpg variable.
14. List the extreme values of that variable.
15. What is the mode of the mpg of all the cars in the dataset? What is the mean? What is the range?
16. Construct a table of means by car type (foreign).
17. What is the Pearson correlation between mpg and weight?
18. What is the correlation between mpg and foreign?
19. Is the relationship between repair record 1978 and car type significant? What is the Gamma correlation of that relationship?
20. How do we show whether this relationship is statistically significant?

Statistical and project planning

- Size matters
  - Power and sample size analysis: A priori versus
  - Post-hoc.
- Sampling planning
  - Probability sampling, clinical trials, and other respectable methods of data collection
  - Stata is wonderful for complex samples
- Respondent protection
  - Informed consent
  - Confidentiality
  - Anonymity
  - Protection of health related information by law
- Pilot studies
  - Proper size
  - Control groups
  - Random selection and random assignment
  - Matching
- Data security
  - Storage
  - Off-site storage
  - Masking of id
- Longitudinal analysis
  - TIME SERIES DATA
  - PANEL DATA
  - SPATIAL DATA
  - For longitudinal studies, censoring and sample attrition must be estimated and planned for. Comparison of pretest scores.

Power and sample size analysis

- Conventional statistics are asymptotic. They work when the same size becomes large (and often do not work with smaller samples).
- The question becomes how large a sample is large enough?
- Power and sample size analyses usually indicate the sufficiency of the sample size.
- To properly plan a research project, we must determine how many subjects or respondents we must interview or question.

Statistical Power Analysis for the Behavioral Sciences (1988) was at NYU

Jacob Cohen, PhD

Power analysis

- There are 3 types of errors that can be made. The type 1 error is rejecting a true null hypothesis. The probability of this type of error is called alpha, α. This is a false negative.
- The type 2 error is accepting a null hypothesis when it is false and should be rejected. The probability of this type of error is called beta,β, and is not to be confused with a standardized regression coefficient, also called beta.

Type 3 errors

- Not asking the correct question in the first place ☺

What significance level should be used?

- The level of significance to be used depends on the consequences of making a mistake.
- For social sciences, alphas of .05 are generally used by convention.   A scholar's reputation may be at stake here.
- For medical and toxicological studies, much more stringent standards are required because the consequences of making a mistake may be life-threatening. Alphas of 0.01 to 0.001 are often used in these cases.

Power and Sample size analysis

- The power of a statistical test is defined as $1 - \beta$.
- The power to reject a false positive depends on the ability to detect an effect of that size.
- Jacob Cohen (1988) Statistical Power Analysis for the Behavioral Sciences, Lawrence Erlbaum Associates: Hillsdale, NJ has formulated conventional (small, medium, and large) effect sizes for basic statistical tests.

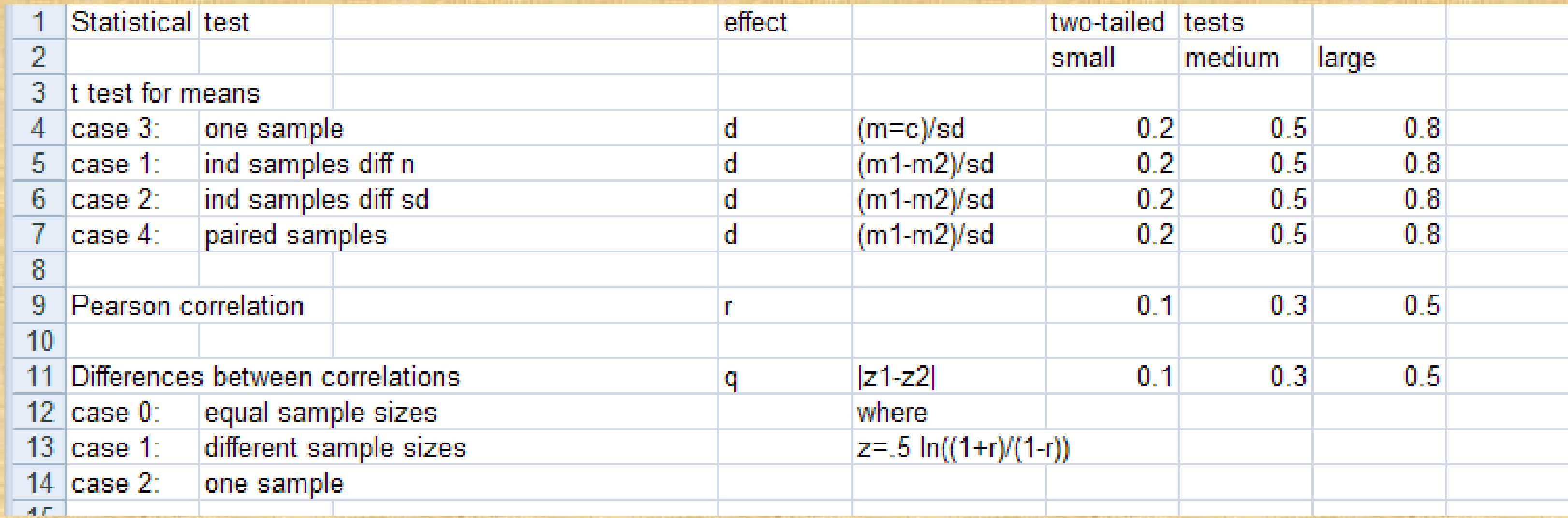
Tables given the n needed are supplied.

The conventional standard is that the project director should enough respondents or subjects to have a sample large enough to detect a medium or small effect size with a power of at least 0.80.

If performing a t-test, small, medium, and large effect sizes are d=2.,.5,.8., where

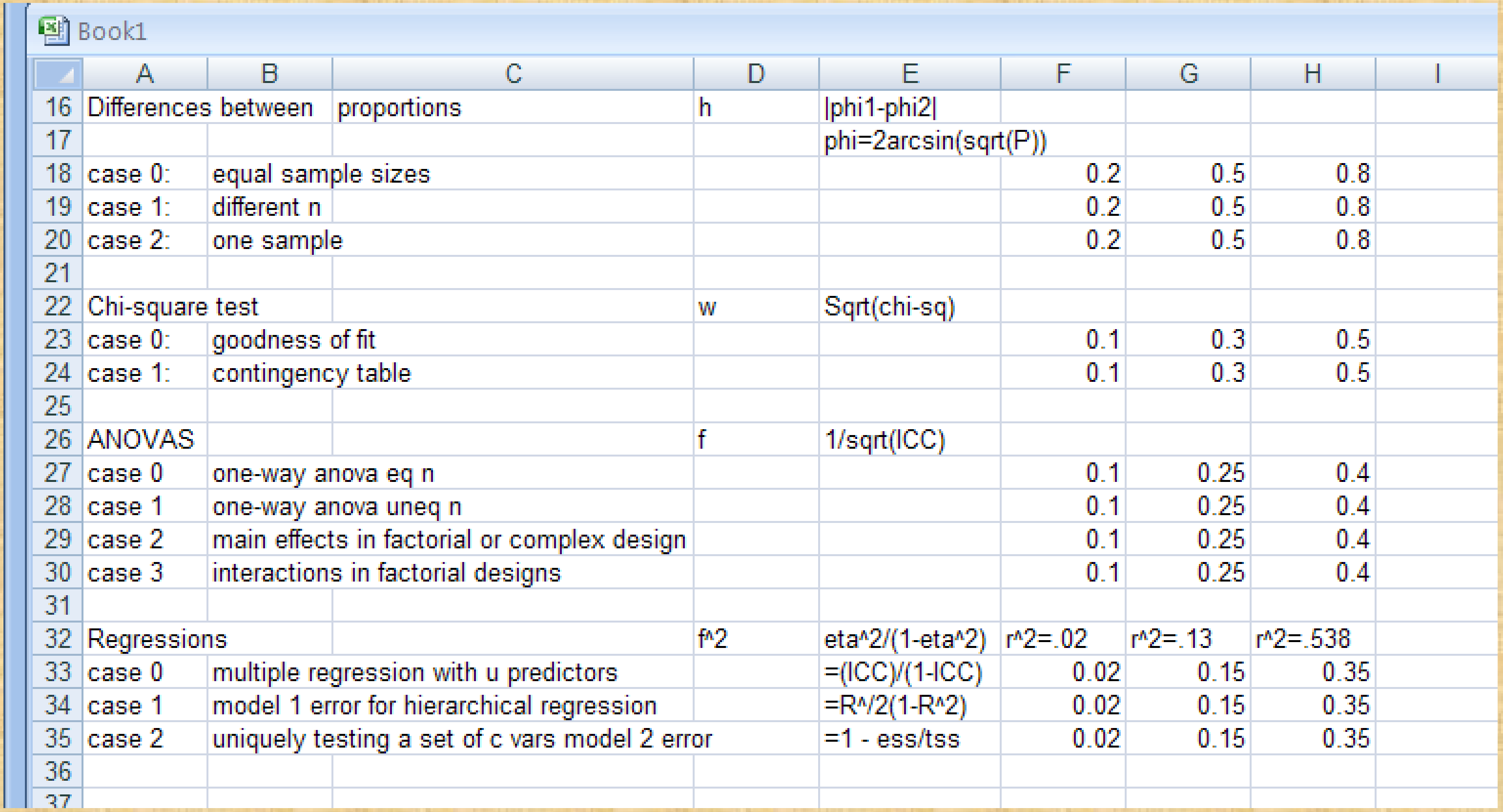
d = (m1-m2)/(stdev)

Cohen’s effect sizes



1	Statistical	test		effect		two-tailed	tests	
2						small	medium	large
3	t test for means							
4	case 3:	one sample		d	(m=c)/sd	0.2	0.5	0.8
5	case 1:	ind samples diff n		d	(m1-m2)/sd	0.2	0.5	0.8
6	case 2:	ind samples diff sd		d	(m1-m2)/sd	0.2	0.5	0.8
7	case 4:	paired samples		d	(m1-m2)/sd	0.2	0.5	0.8
8								
9	Pearson correlation			r		0.1	0.3	0.5
10								
11	Differences between correlations			q	|z1-z2|	0.1	0.3	0.5
12	case 0:	equal sample sizes			where			
13	case 1:	different sample sizes			z=.5 ln((1+r)/(1-r))			
14	case 2:	one sample						

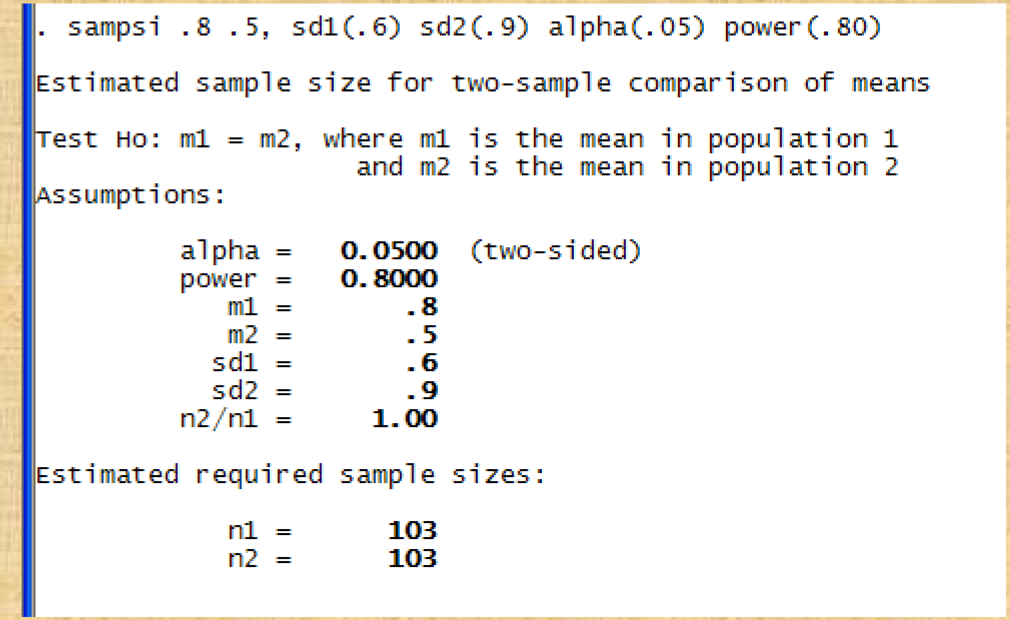
Cohen’s effect sizes



	A	B	C	D	E	F	G	H	I
16	Differences between		proportions	h	|phi1-phi2|				
17					phi=2arcsin(sqrt(P))				
18	case 0:	equal sample sizes				0.2	0.5	0.8	
19	case 1:	different n				0.2	0.5	0.8	
20	case 2:	one sample				0.2	0.5	0.8	
21									
22	Chi-square test			w	Sqrt(chi-sq)				
23	case 0:	goodness of fit				0.1	0.3	0.5	
24	case 1:	contingency table				0.1	0.3	0.5	
25									
26	ANOVAS			f	1/sqrt(ICC)				
27	case 0	one-way anova eq n				0.1	0.25	0.4	
28	case 1	one-way anova uneq n				0.1	0.25	0.4	
29	case 2	main effects in factorial or complex design				0.1	0.25	0.4	
30	case 3	interactions in factorial designs				0.1	0.25	0.4	
31									
32	Regressions			f^2	eta^2/(1-eta^2)	r^2=.02	r^2=.13	r^2=.538	
33	case 0	multiple regression with u predictors			=(ICC)/(1-ICC)	0.02	0.15	0.35	
34	case 1	model 1 error for hierarchical regression			=R^2(1-R^2)	0.02	0.15	0.35	
35	case 2	uniquely testing a set of c vars model 2 error			=1 - ess/tss	0.02	0.15	0.35	
36									
37									

Stata can compute
post-hoc power and sample size

- For t-tests and proportions



```
. sampsi .8 .5, sd1(.6) sd2(.9) alpha(.05) power(.80)

Estimated sample size for two-sample comparison of means

Test Ho: m1 = m2, where m1 is the mean in population 1
                    and m2 is the mean in population 2
Assumptions:

         alpha =   0.0500  (two-sided)
         power =   0.8000
            m1 =       .8
            m2 =       .5
           sd1 =       .6
           sd2 =       .9
         n2/n1 =     1.00

Estimated required sample sizes:

            n1 =      103
            n2 =      103
```

For repeated measures contrasts
For survival analysis problems

Attrition is to be compensated for in the planning of the sample size.

- A pilot study will indicate the rate of attrition if it is representatively sampled and collected.
- If there is five percent attrition, then 105% of the sample size should be collected. Sample size = Number to be collected/.95 = 211 (rounded to nearest integer)
- If the margin of error is 5 more %, then 110% of the needed sample size should be collected for the larger sample. Sample size to be collected = 211/.95 = 223 (rounded to nearest integer).
- If the pilot study indicates that 10 percent of the items will on the average not be answered for those who remain in the study, then add another 10% to the 110% of the needed sample size that must be collected. 120% of the needed sample size must be obtained in the planning stage. Sample size to be collected = 223/.90 = 248.
- If you are studying a hard to reach minority, increase your safety margin. If you are conducting a longitudinal study in an area that is politically unstable, be careful, focus on your primary objective and avoid unnecessary entanglements or distractions.

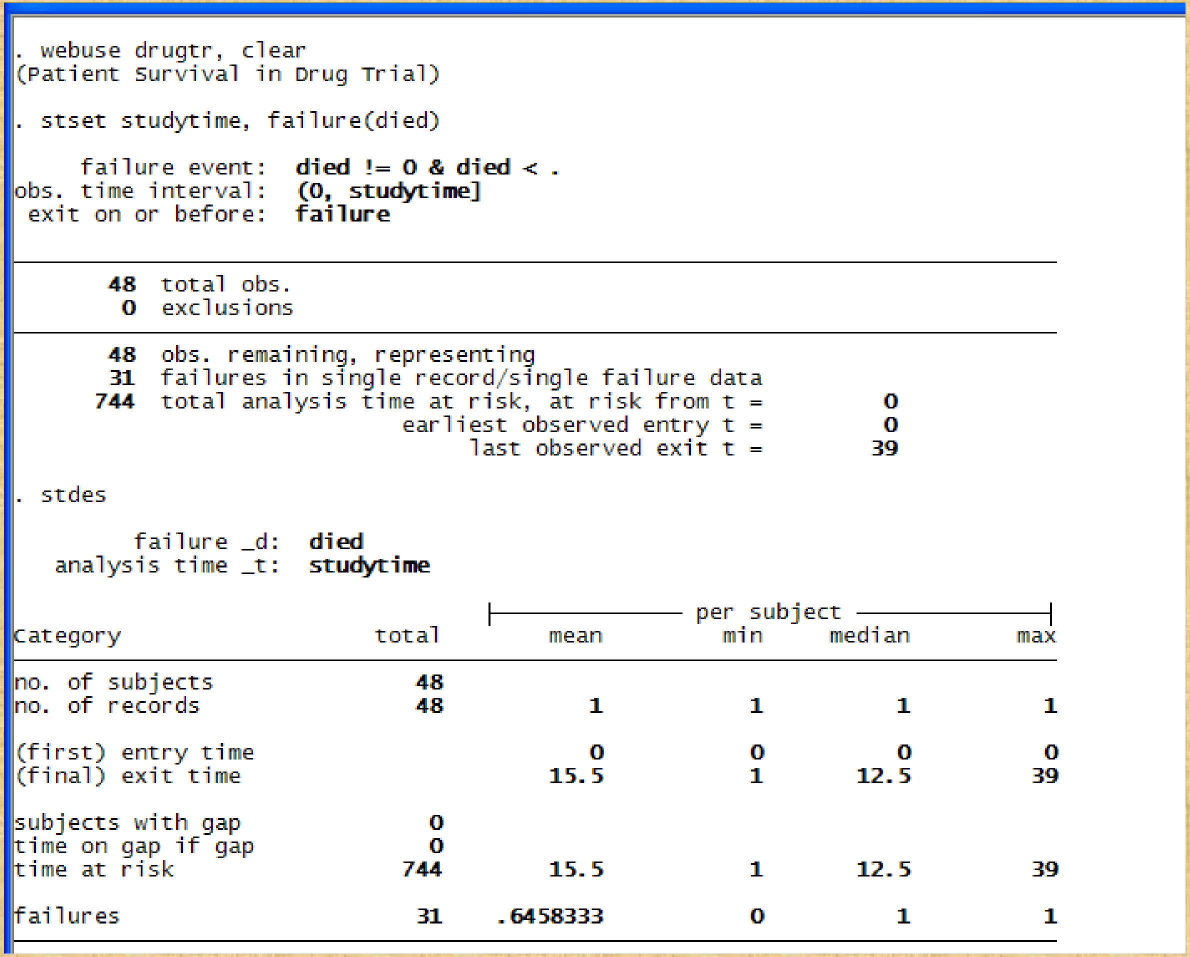
Attrition and censoring in longitudinal studies

- Attrition is accounted for by censoring. There is right censoring when a person is lost to followup.
- There is right censoring when the event has not occurred prior to the end of the study.
- There is interval censoring when the patient is in jail for 2 weeks and cannot attend his midterm interview.

Sample size reduction

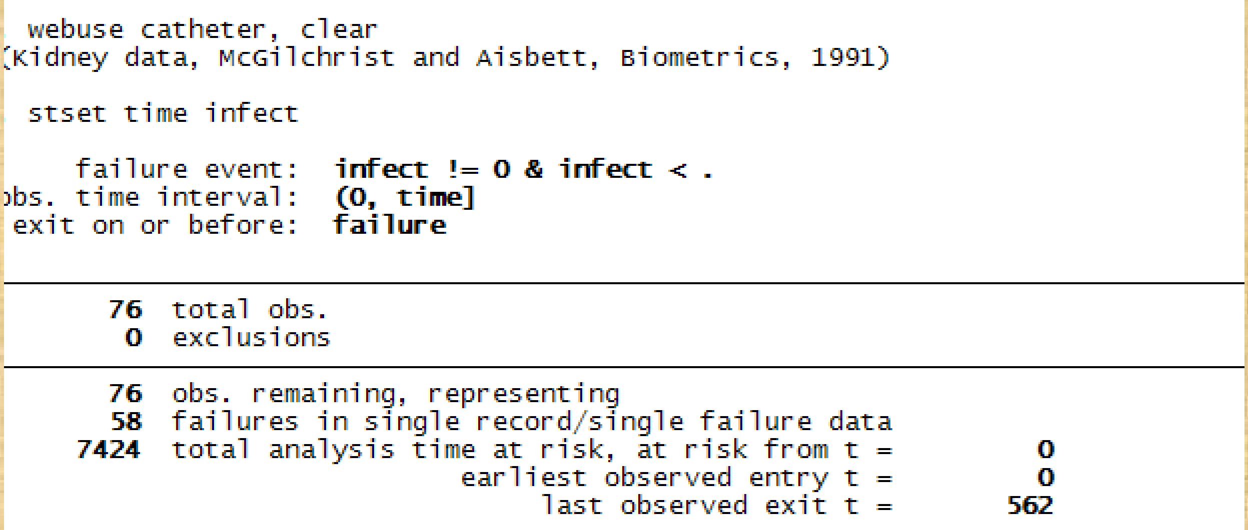
- Bubble sheets cannot always be read clearly if the survey is on both sides of the paper. Machines make many mistakes in scanning such sheets.
- Bubble sheets cannot always be read clearly if the answers are not closed ended. Avoid open-ended questions
- Transmission over the web must be double checked to be sure that there was not information corruption in transmission of data.
- Do not allow people not to answer if the answer is negative. This breeds confusion and uncertainty.

Indexing survival-time data for bio-statistical analysis



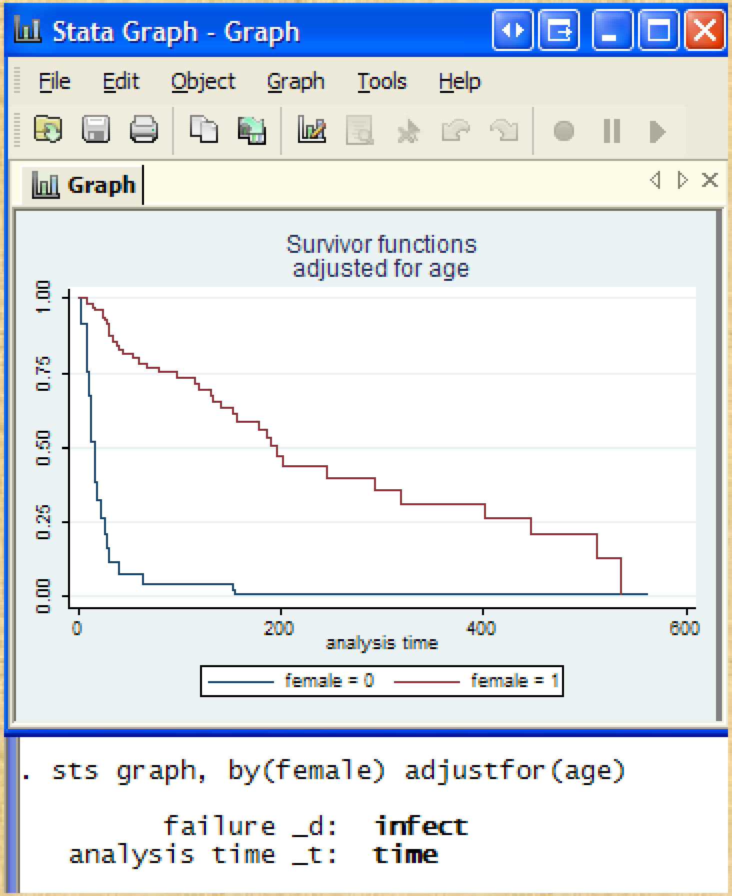
```
. webuse drugtr, clear
(Patient Survival in Drug Trial)

. stset studytime, failure(died)

     failure event:  died != 0 & died < .
obs. time interval:  (0, studytime]
 exit on or before:  failure

------------------------------------------------------------------------------
       48  total obs.
        0  exclusions
------------------------------------------------------------------------------
       48  obs. remaining, representing
       31  failures in single record/single failure data
      744  total analysis time at risk, at risk from t =         0
                             earliest observed entry t =         0
                                  last observed exit t =        39

. stdes

         failure _d:  died
   analysis time _t:  studytime

                                   |-------------- per subject --------------|
Category                   total        mean         min     median        max
------------------------------------------------------------------------------
no. of subjects               48
no. of records                48           1           1          1          1

(first) entry time                         0           0          0          0
(final) exit time                       15.5           1       12.5         39

subjects with gap              0
time on gap if gap             0
time at risk                 744        15.5           1       12.5         39

failures                      31    .6458333           0          1          1
------------------------------------------------------------------------------
```

Indexing survival time data for prostate cancer



```
. webuse catheter, clear
(Kidney data, McGilchrist and Aisbett, Biometrics, 1991)

. stset time infect

     failure event:  infect != 0 & infect < .
bs. time interval:  (0, time]
 exit on or before:  failure

------------------------------------------------------------------------------
       76  total obs.
        0  exclusions
------------------------------------------------------------------------------
       76  obs. remaining, representing
       58  failures in single record/single failure data
     7424  total analysis time at risk, at risk from t =         0
                             earliest observed entry t =         0
                                  last observed exit t =       562
```

Kaplan Meier Survival curves by gender adjusted for age



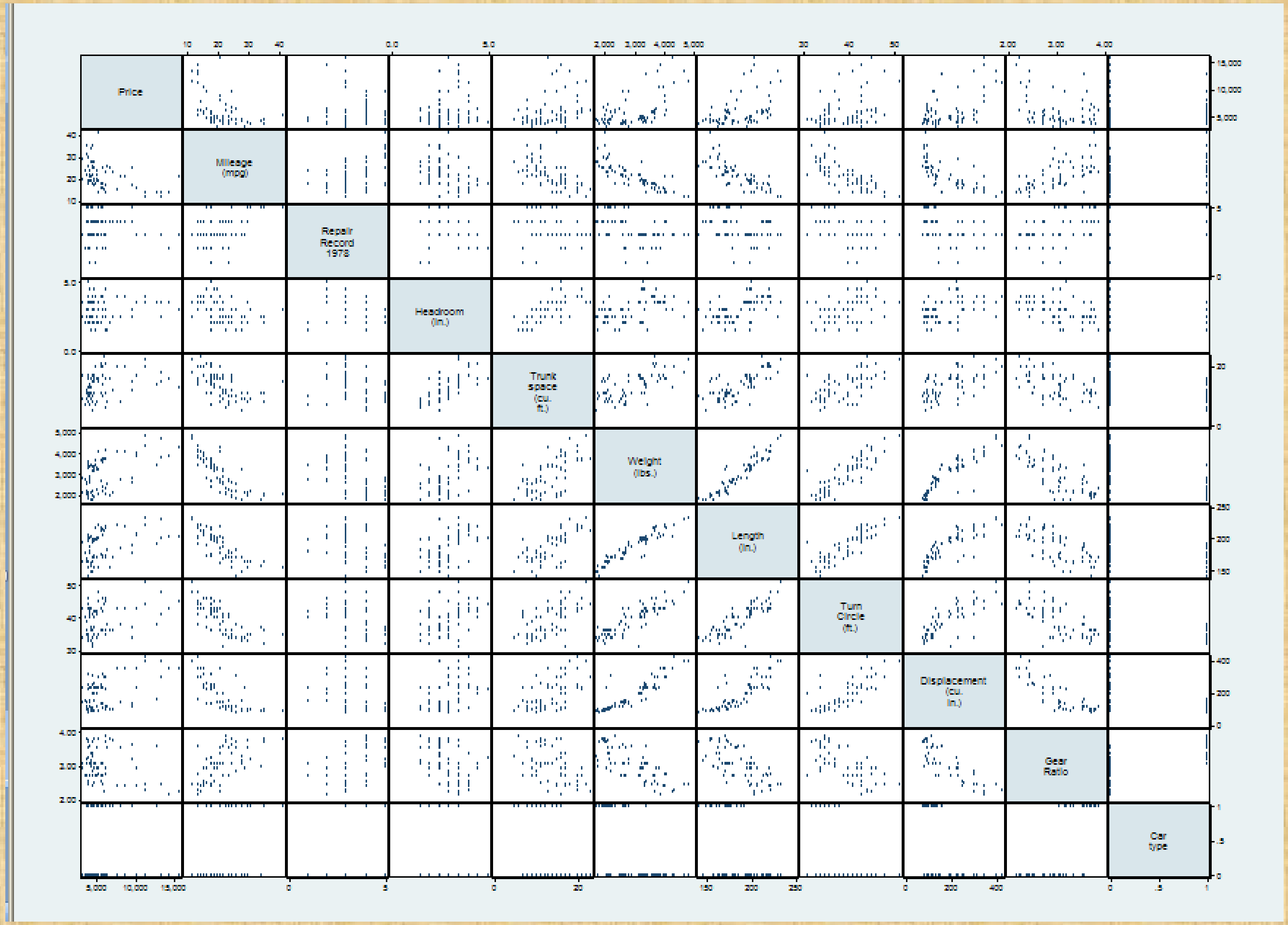
```
. sts graph, by(female) adjustfor(age)

         failure _d:  infect
   analysis time _t:  time
```

Visualization for exploratory data analysis and model diagnosis

- 1-dimensional Univariate
  - Histograms
  - Box plots
  - Stem and leaf plots
  - Quantile plots
  - Bar graphs
  - Pie charts
- 2-dimensional Scatterplot matrices
  - Scatterplots
  - Time series plots
- Multi-dimensional plots
  - Panel plots
  - 3-D scatterplots
- Graphics editor

Exploratory Data Analysis
Edward Tufte (Princeton Univ)

Matrix scatterplots for exploring functional form of relationships

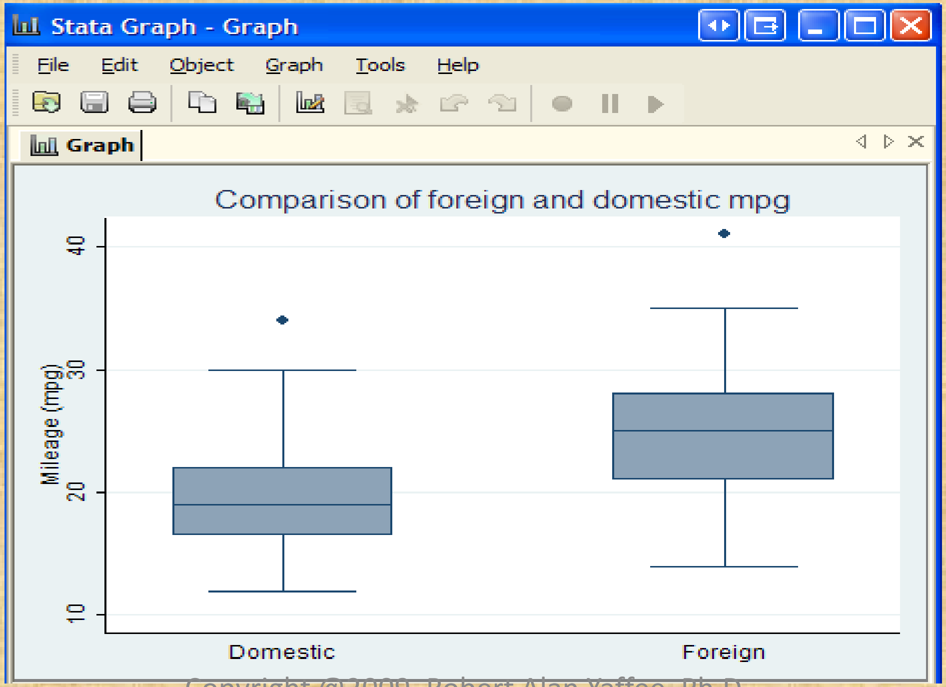


graph matrix mpg-foreign

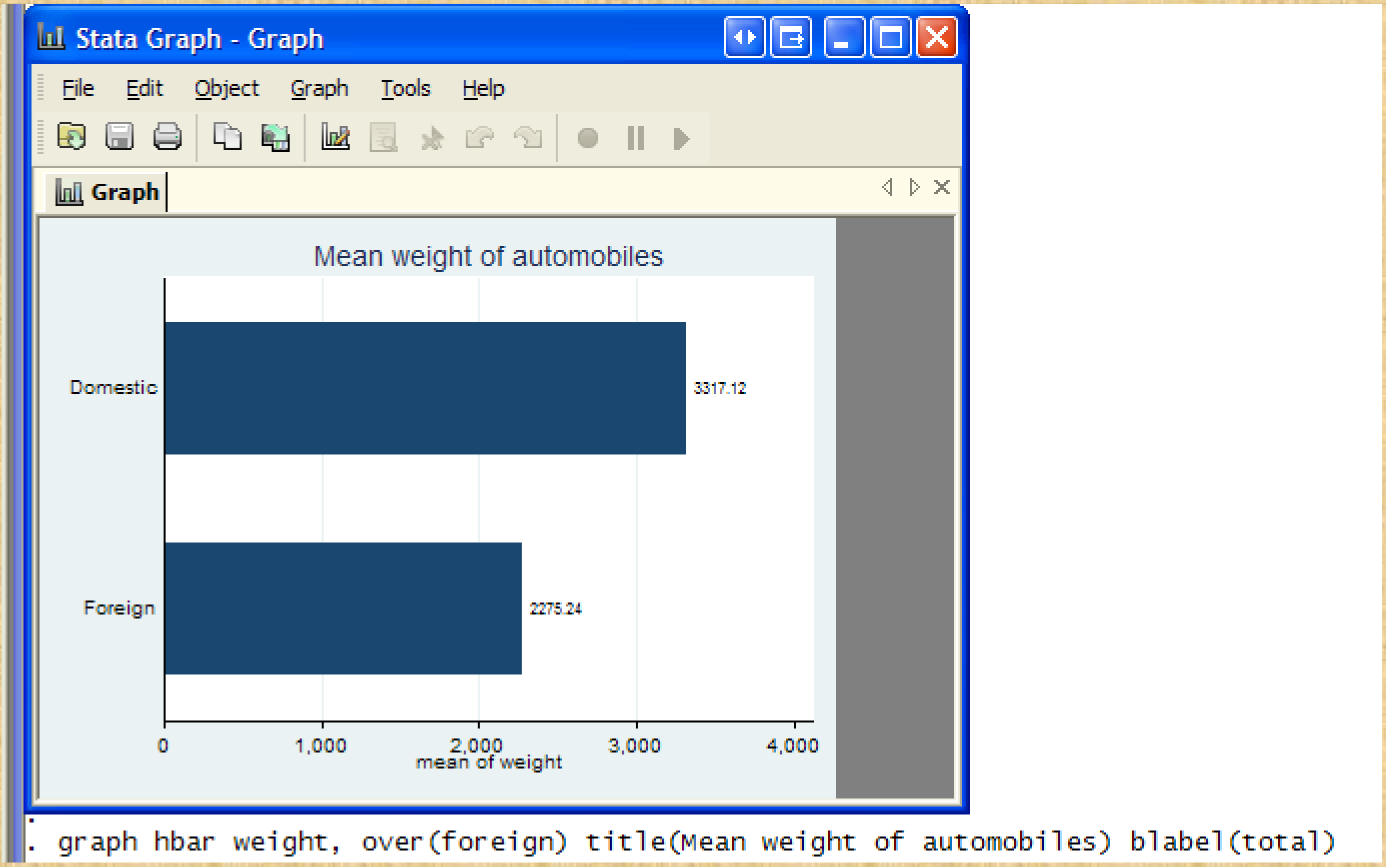
Exploratory data analysis

sort foreign

graph box mpg, over(foreign) title(Comparison of box plots)

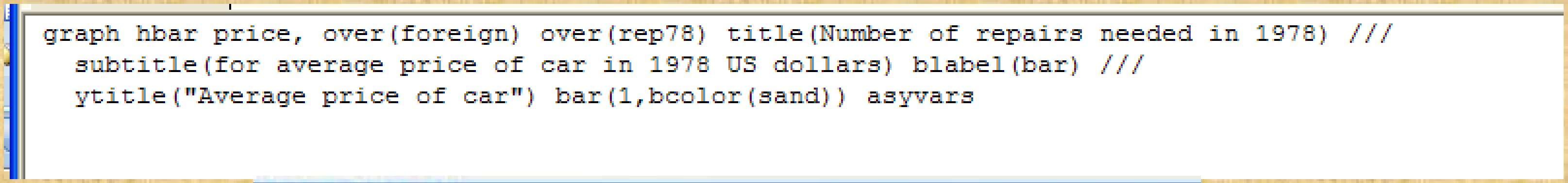


Horizontal bar charts

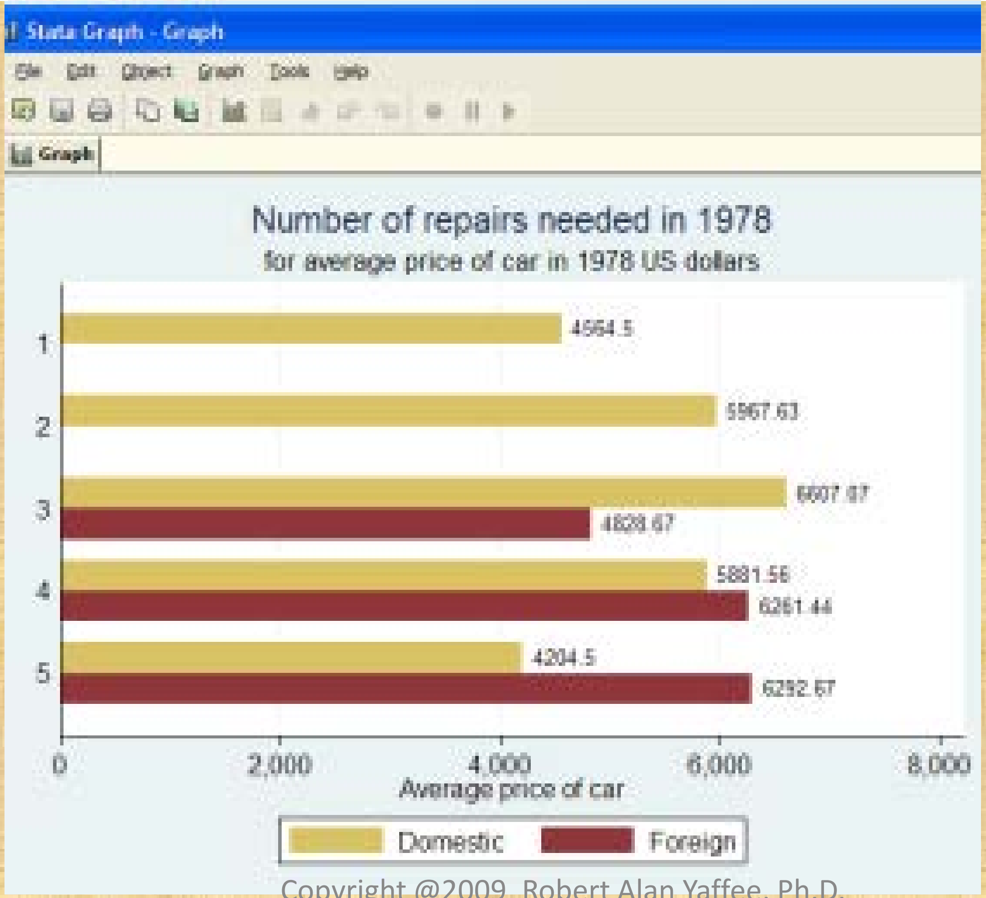


```
. graph hbar weight, over(foreign) title(Mean weight of automobiles) blabel(total)
```

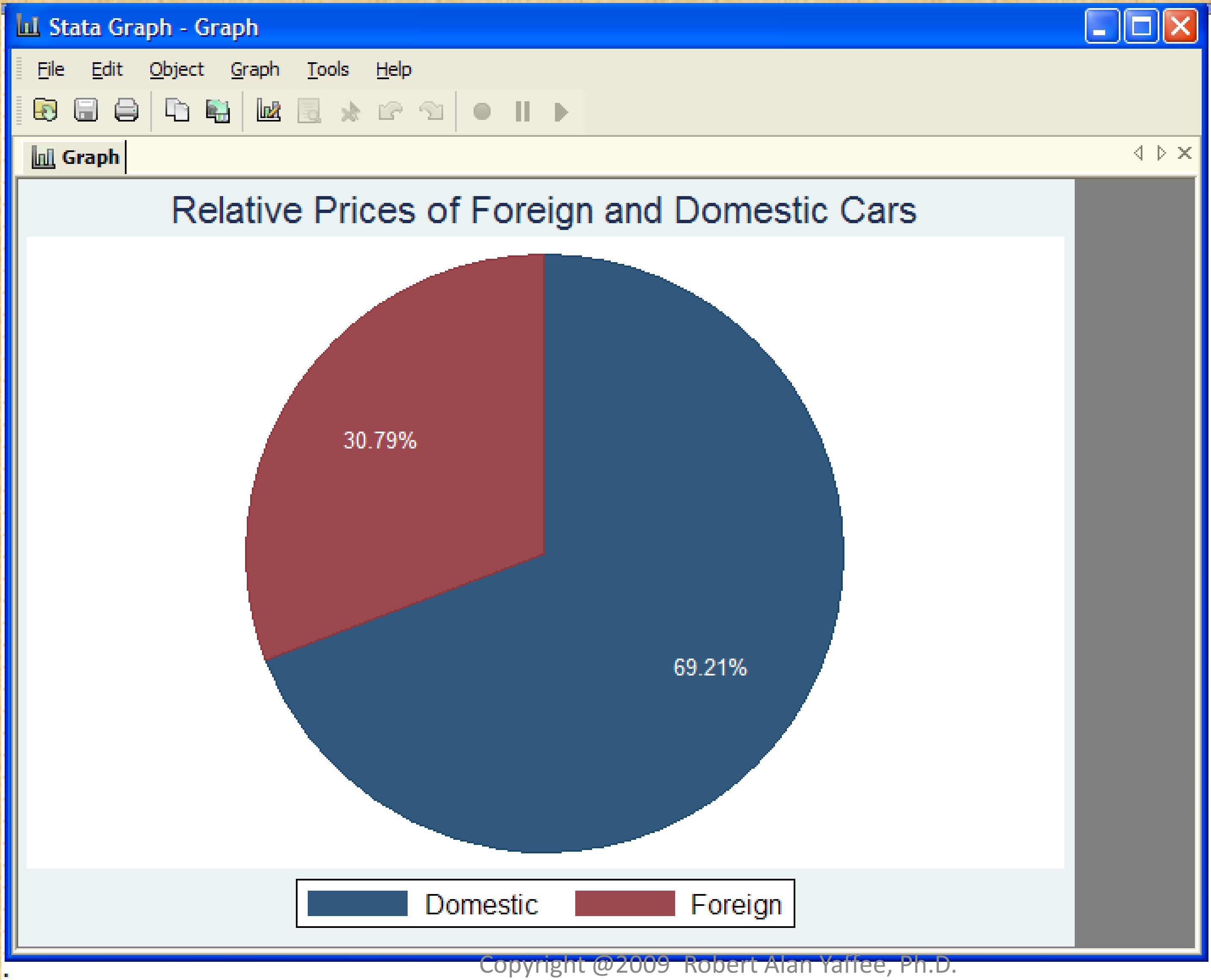
Nesting horizontal bar charts



```
graph hbar price, over(foreign) over(rep78) title(Number of repairs needed in 1978) ///
  subtitle(for average price of car in 1978 US dollars) blabel(bar) ///
  ytitle("Average price of car") bar(1,bcolor(sand)) asyvars
```

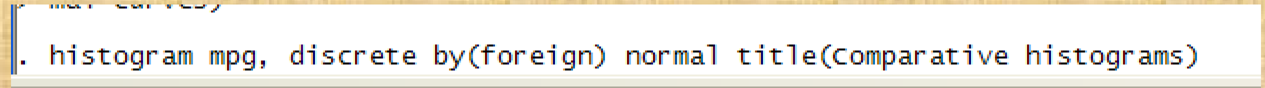


Pie Charts

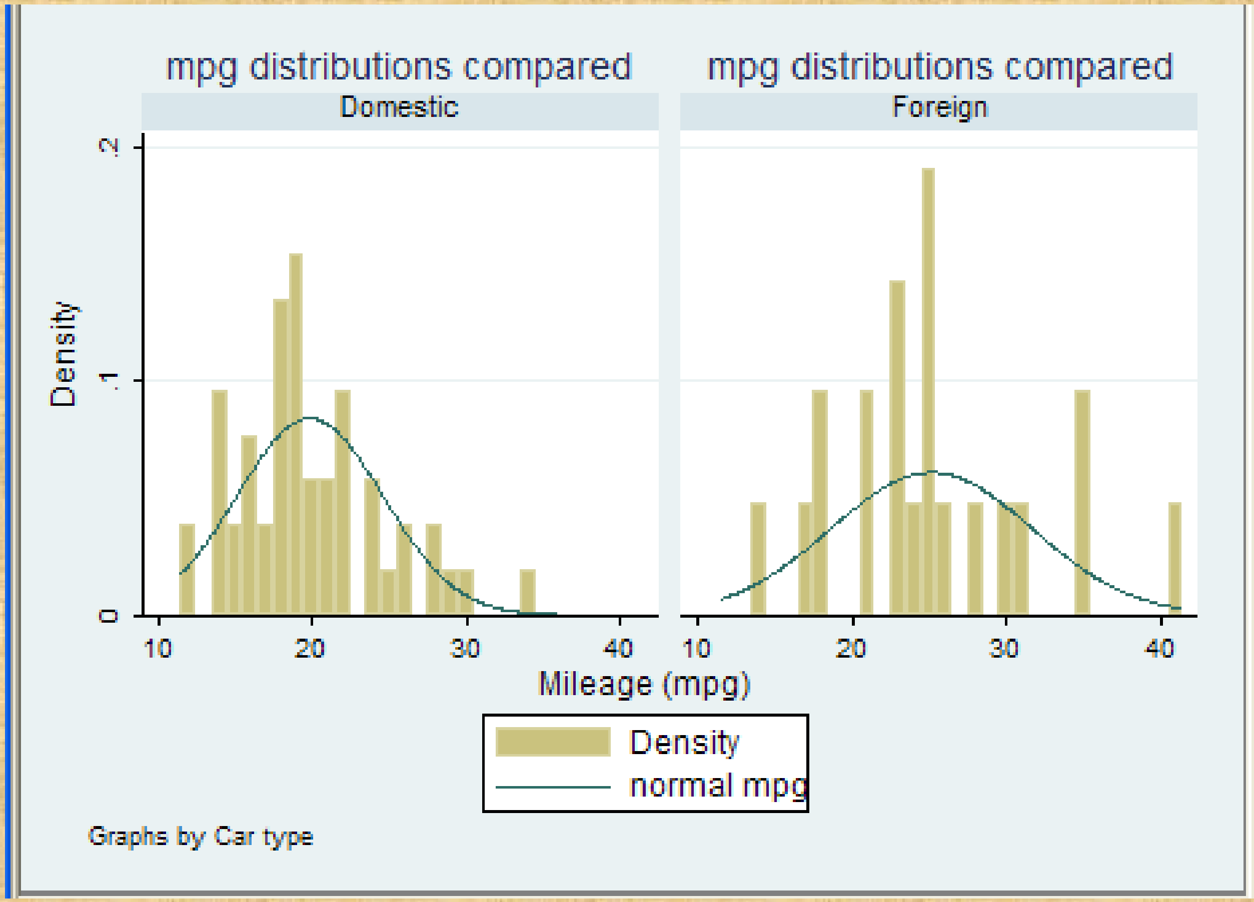


```
.
. graph pie price, over(foreign) title(Relative Prices of Foreign and Domestic Cars) plabel(_all percent,color(white))
```

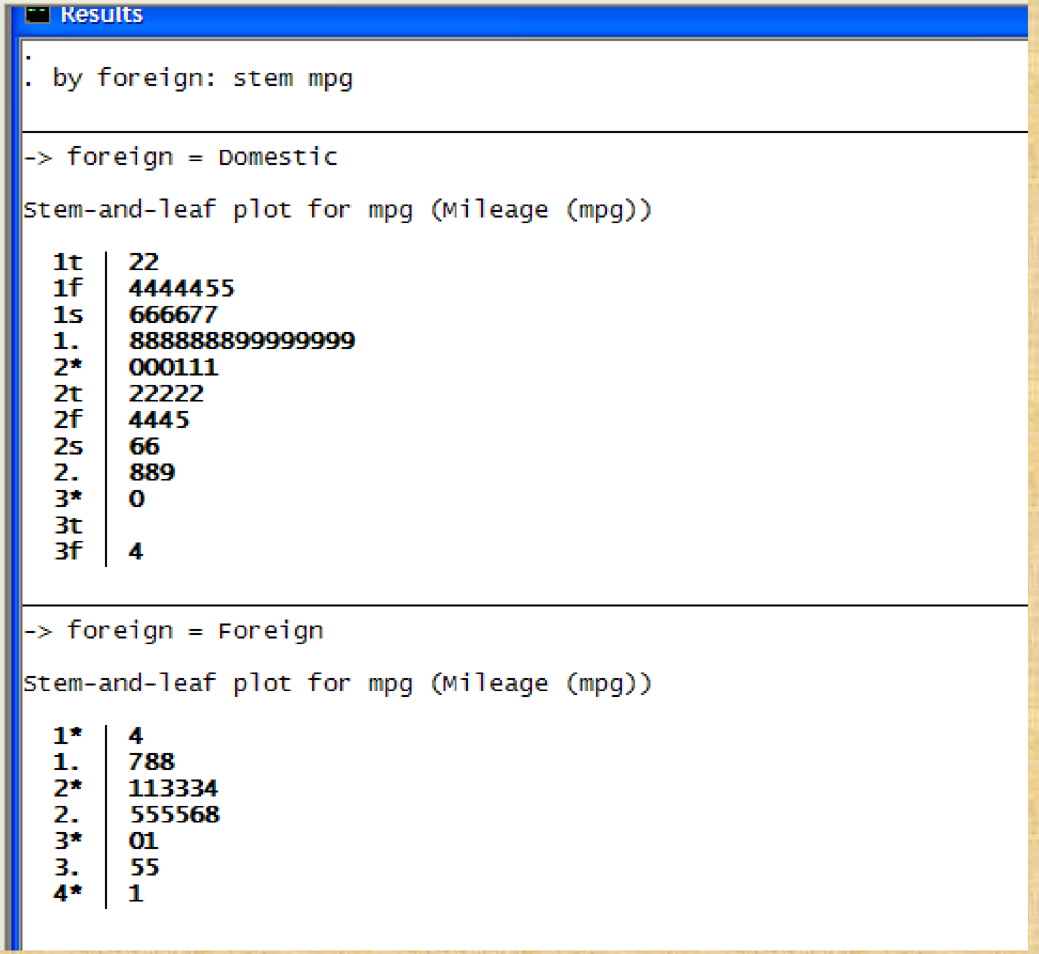
Comparative histograms



```
. histogram mpg, discrete by(foreign) normal title(Comparative histograms)
```

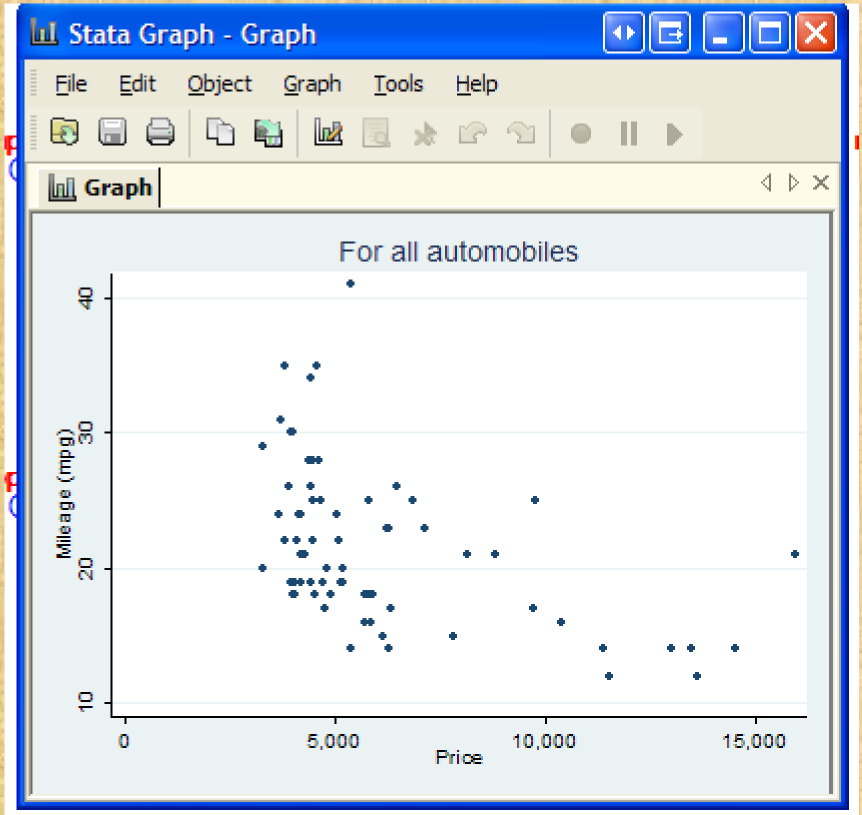


Comparative stem and leaf plots



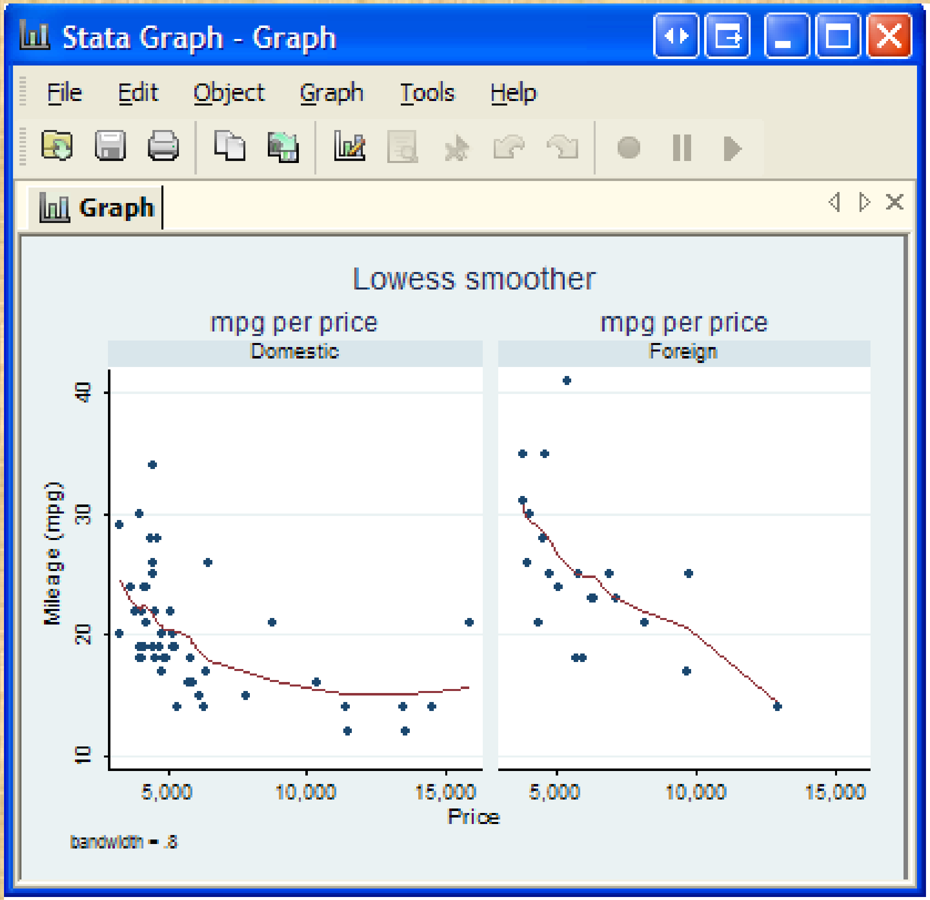
```
.
. by foreign: stem mpg

-> foreign = Domestic

Stem-and-leaf plot for mpg (Mileage (mpg))

  1t | 22
  1f | 4444455
  1s | 666677
  1. | 8888888899999999
  2* | 000111
  2t | 22222
  2f | 4445
  2s | 66
  2. | 889
  3* | 0
  3t |
  3f | 4

-> foreign = Foreign

Stem-and-leaf plot for mpg (Mileage (mpg))

  1* | 4
  1. | 788
  2* | 113334
  2. | 555568
  3* | 01
  3. | 55
  4* | 1
```

The relationship between fuel economy and luxury in auto purchases



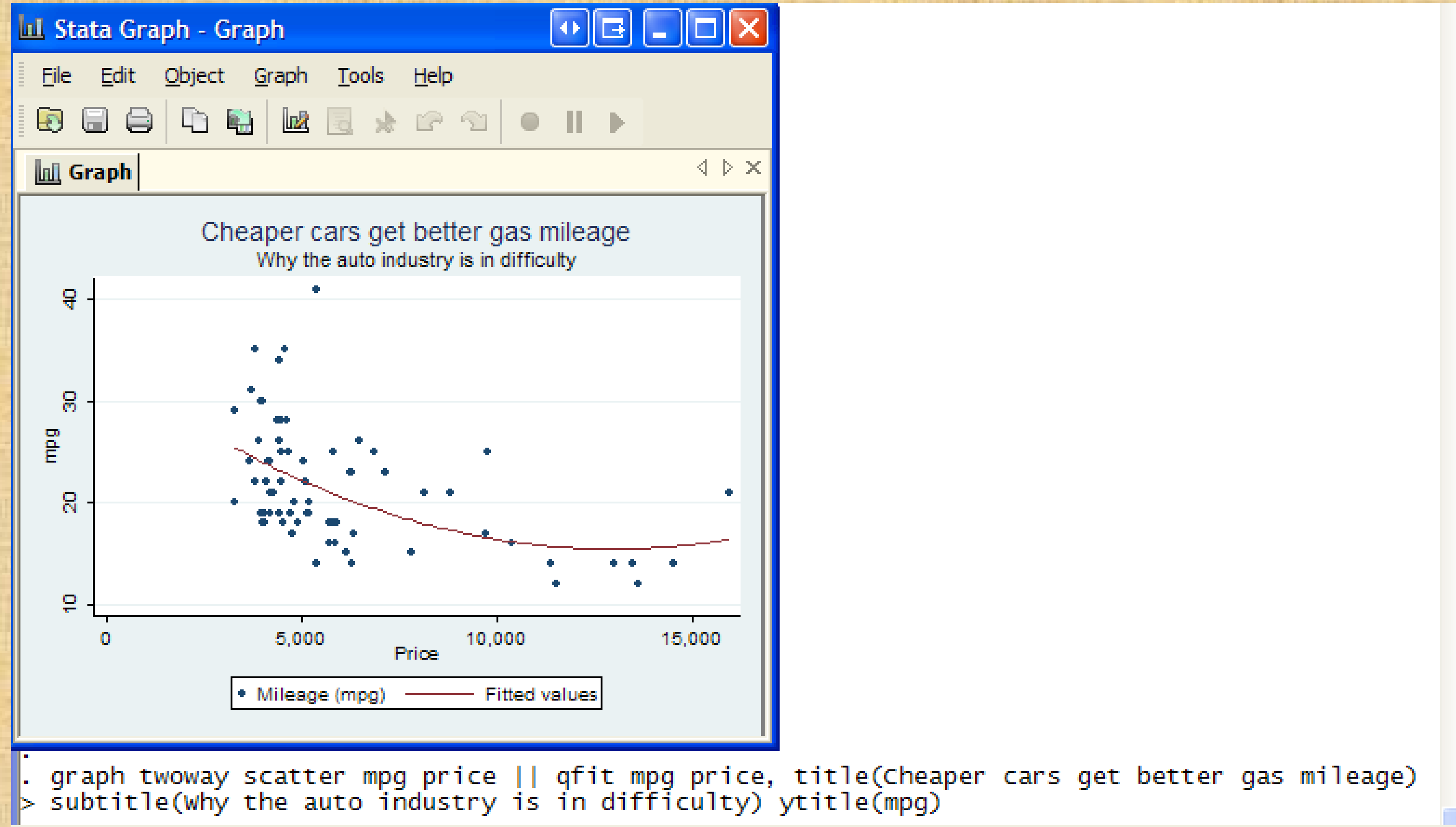
```
scatter mpg price, title(For all automobiles)
```

Comparison of mpg per price between foreign and domestic cars



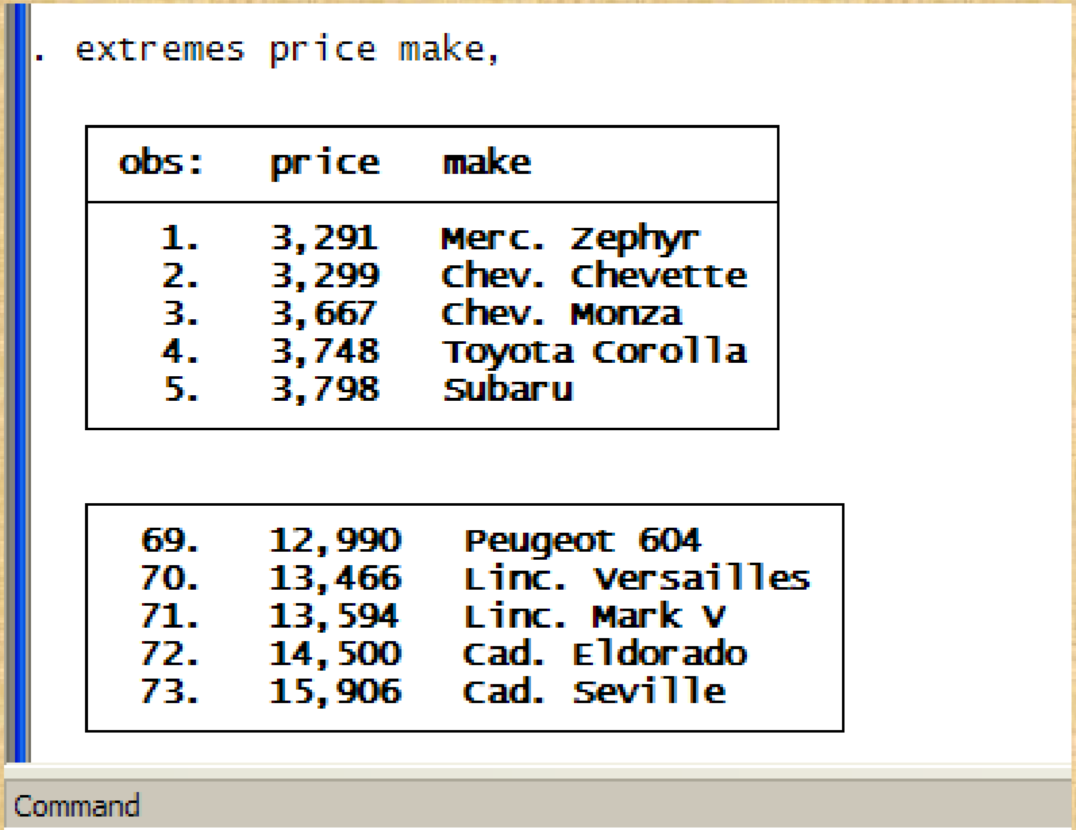
```
lowess mpg price, by(foreign) title(mpg per price)
```

Nonlinear fit between mpg and price



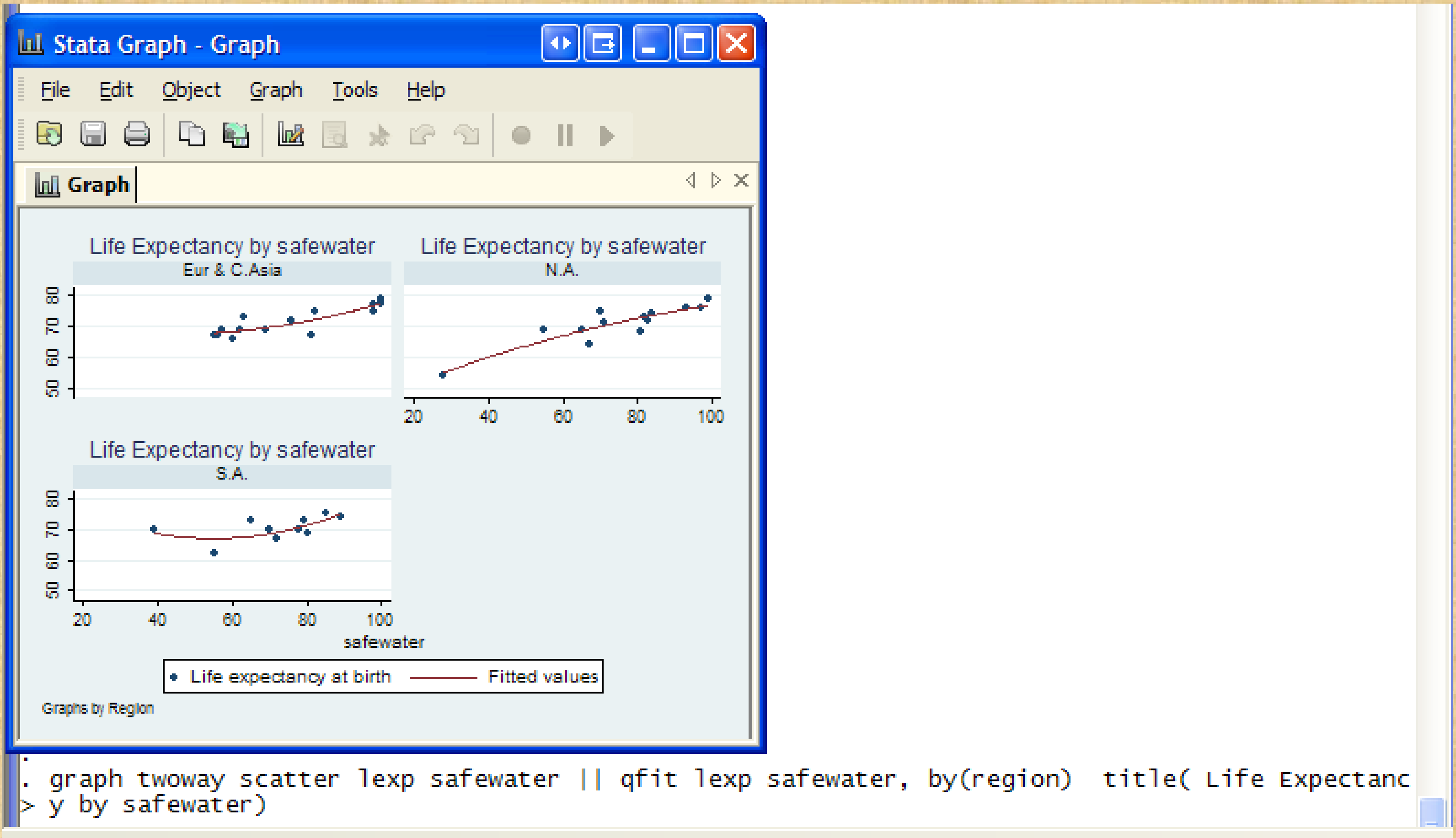
```
.
. graph twoway scatter mpg price || qfit mpg price, title(Cheaper cars get better gas mileage)
> subtitle(Why the auto industry is in difficulty) ytitle(mpg)
```

Identifying the most and least expensive cars



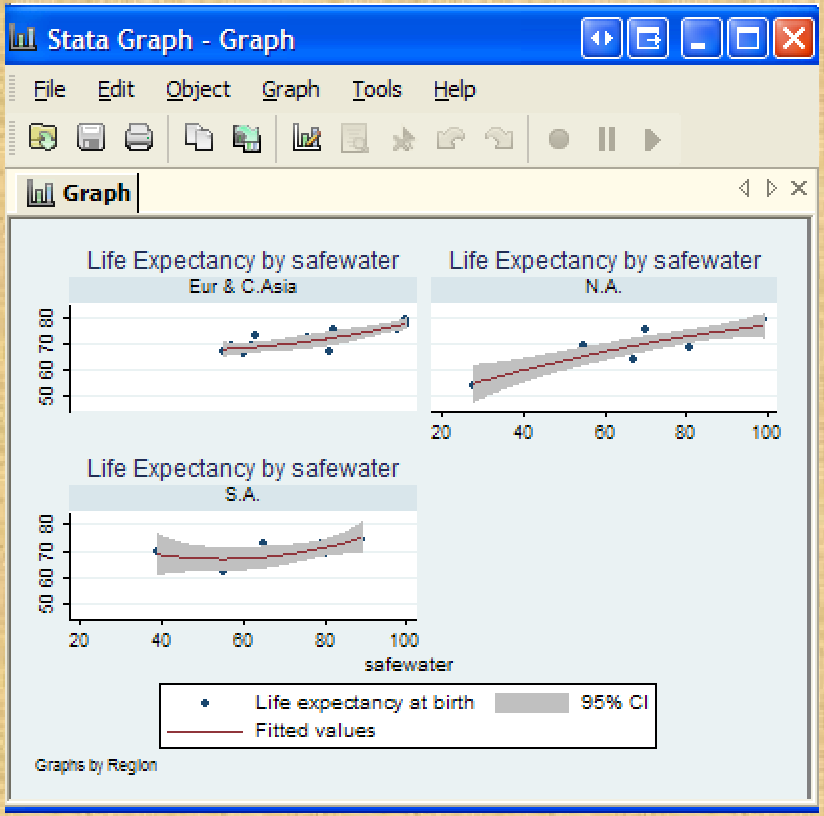
```
. extremes price make,

  +------------------------------------+
  | obs:    price    make              |
  |------------------------------------|
  |   1.    3,291    Merc. Zephyr      |
  |   2.    3,299    Chev. Chevette    |
  |   3.    3,667    Chev. Monza       |
  |   4.    3,748    Toyota Corolla    |
  |   5.    3,798    Subaru            |
  +------------------------------------+

  +-----------------------------------------+
  |  69.    12,990    Peugeot 604           |
  |  70.    13,466    Linc. Versailles      |
  |  71.    13,594    Linc. Mark V          |
  |  72.    14,500    Cad. Eldorado         |
  |  73.    15,906    Cad. Seville          |
  +-----------------------------------------+
```

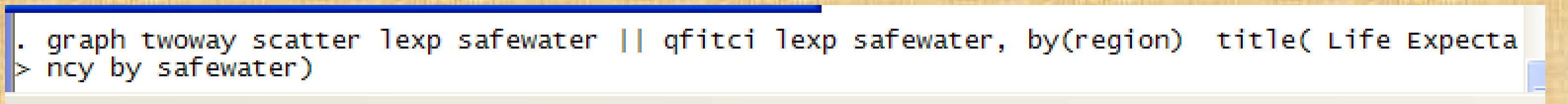
Paneled graphs



```
.
. graph twoway scatter lexp safewater || qfit lexp safewater, by(region)  title( Life Expectanc
> y by safewater)
```

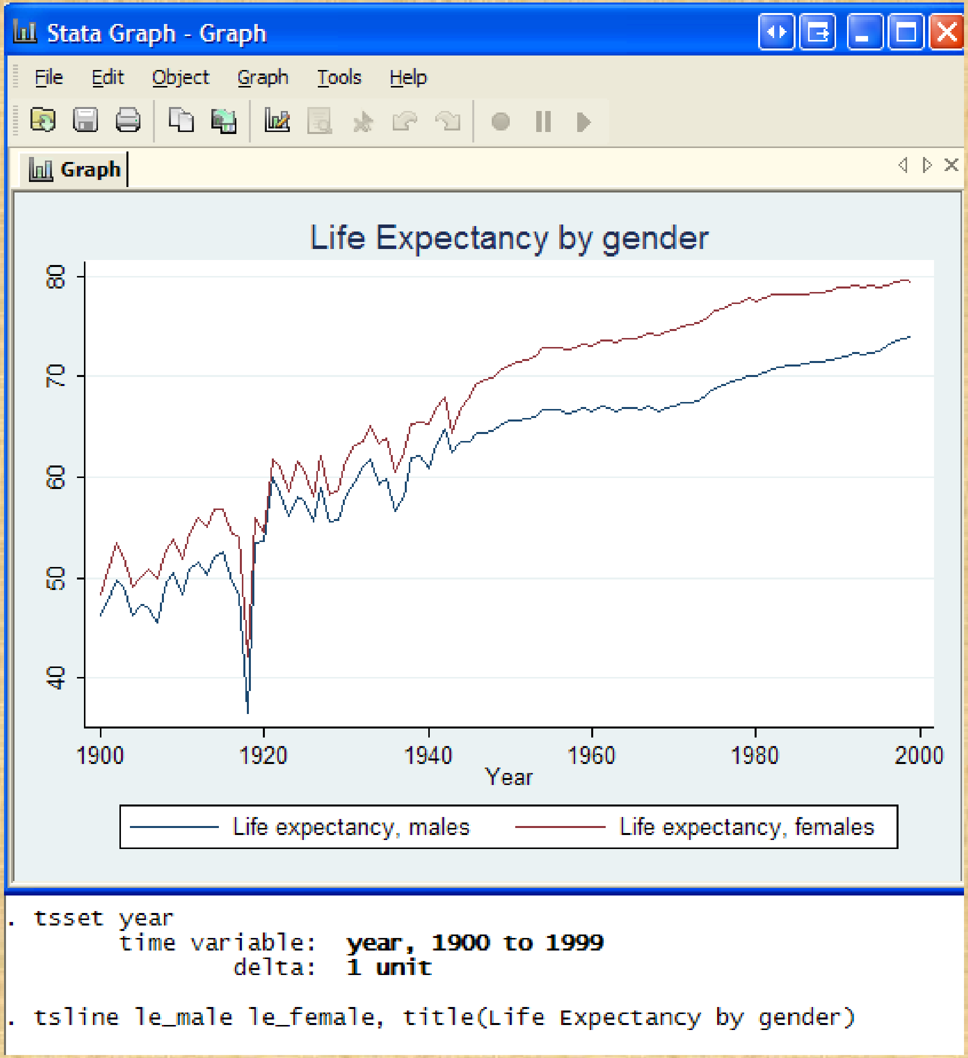
Fit and confidence intervals





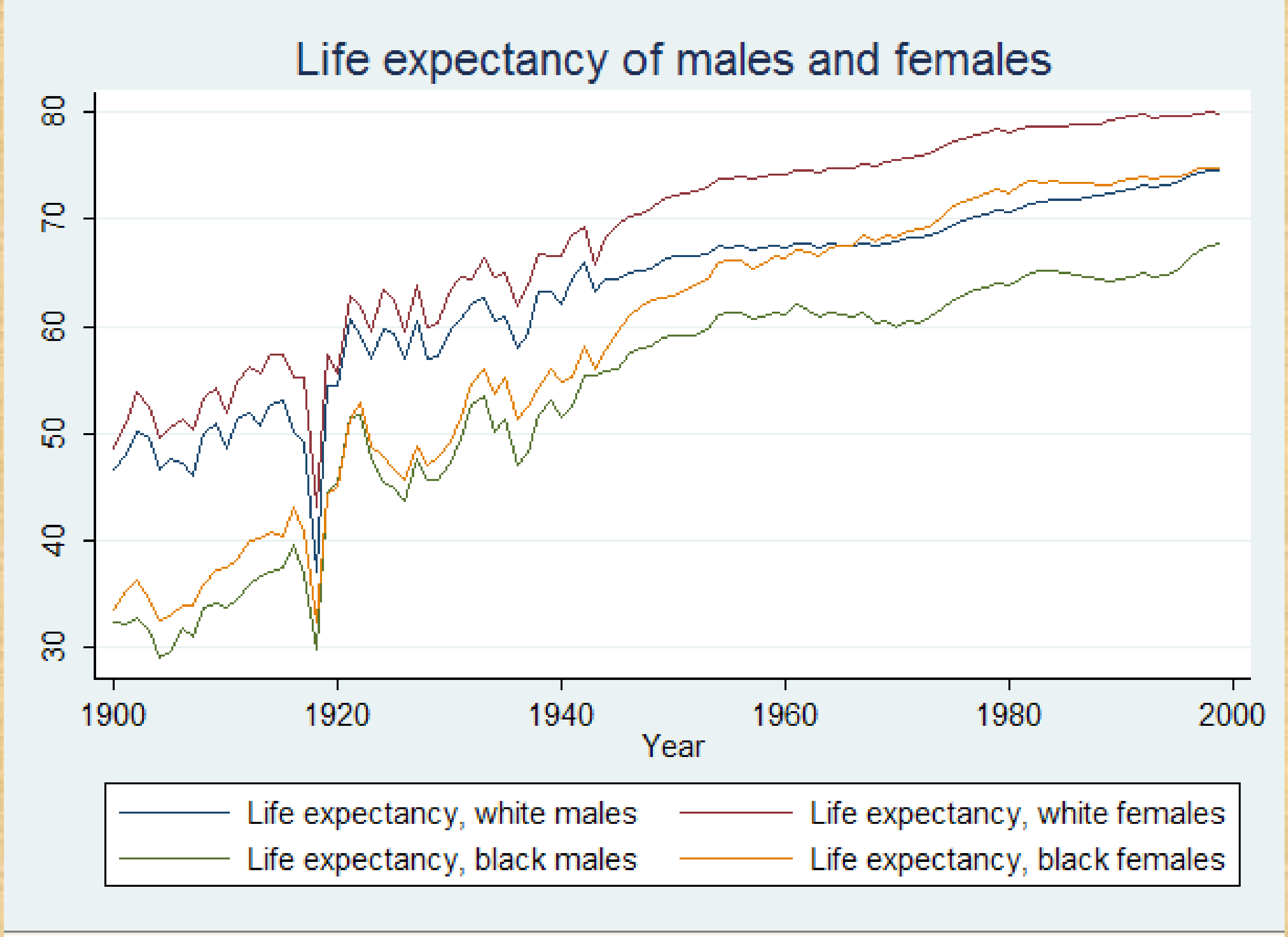
```
. graph twoway scatter lexp safewater || qfitci lexp safewater, by(region)  title( Life Expecta
> ncy by safewater)
```

Time series plot of life expectancy by gender

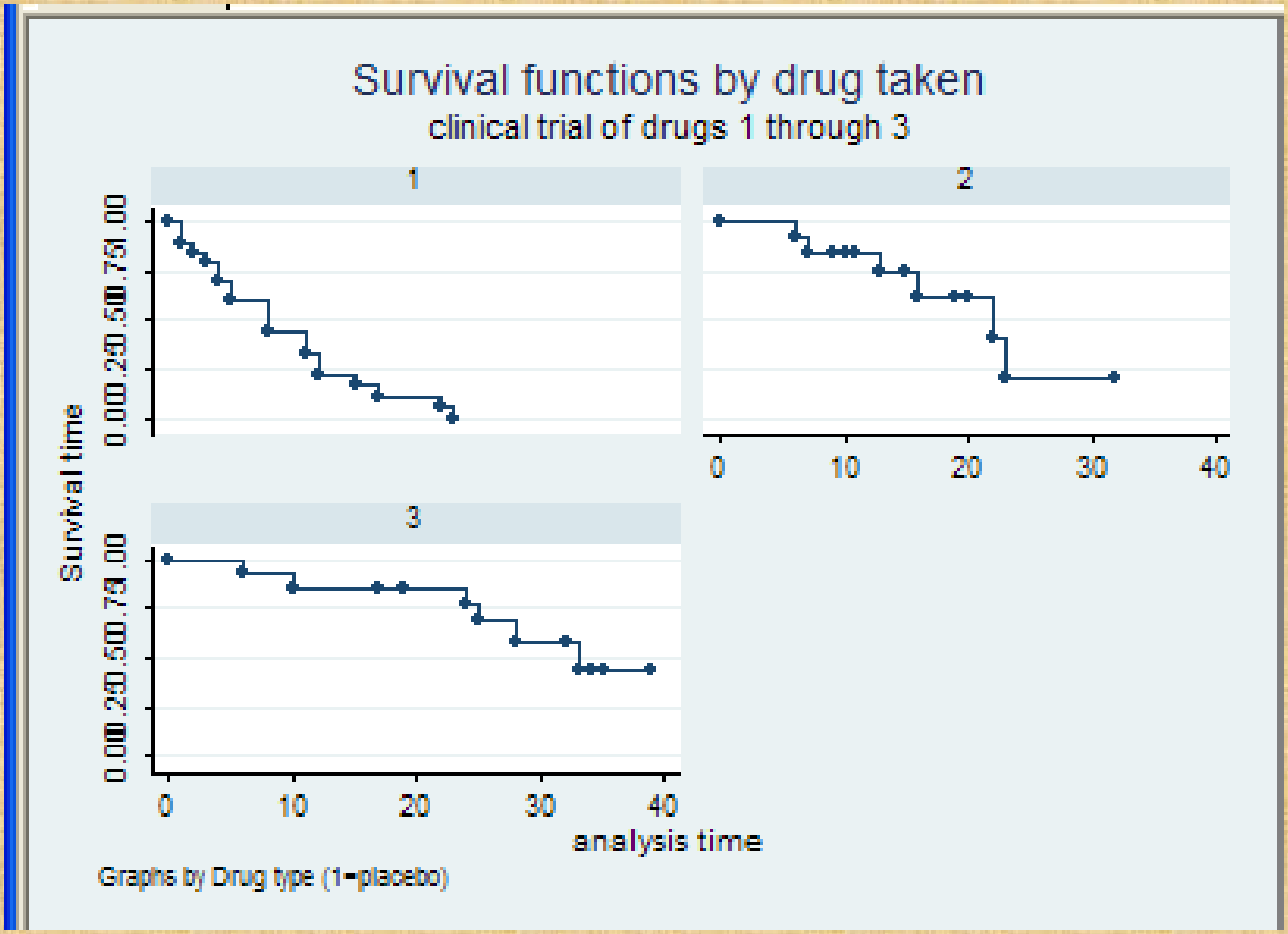


```
. tsset year
        time variable:  year, 1900 to 1999
                delta:  1 unit

. tsline le_male le_female, title(Life Expectancy by gender)
```

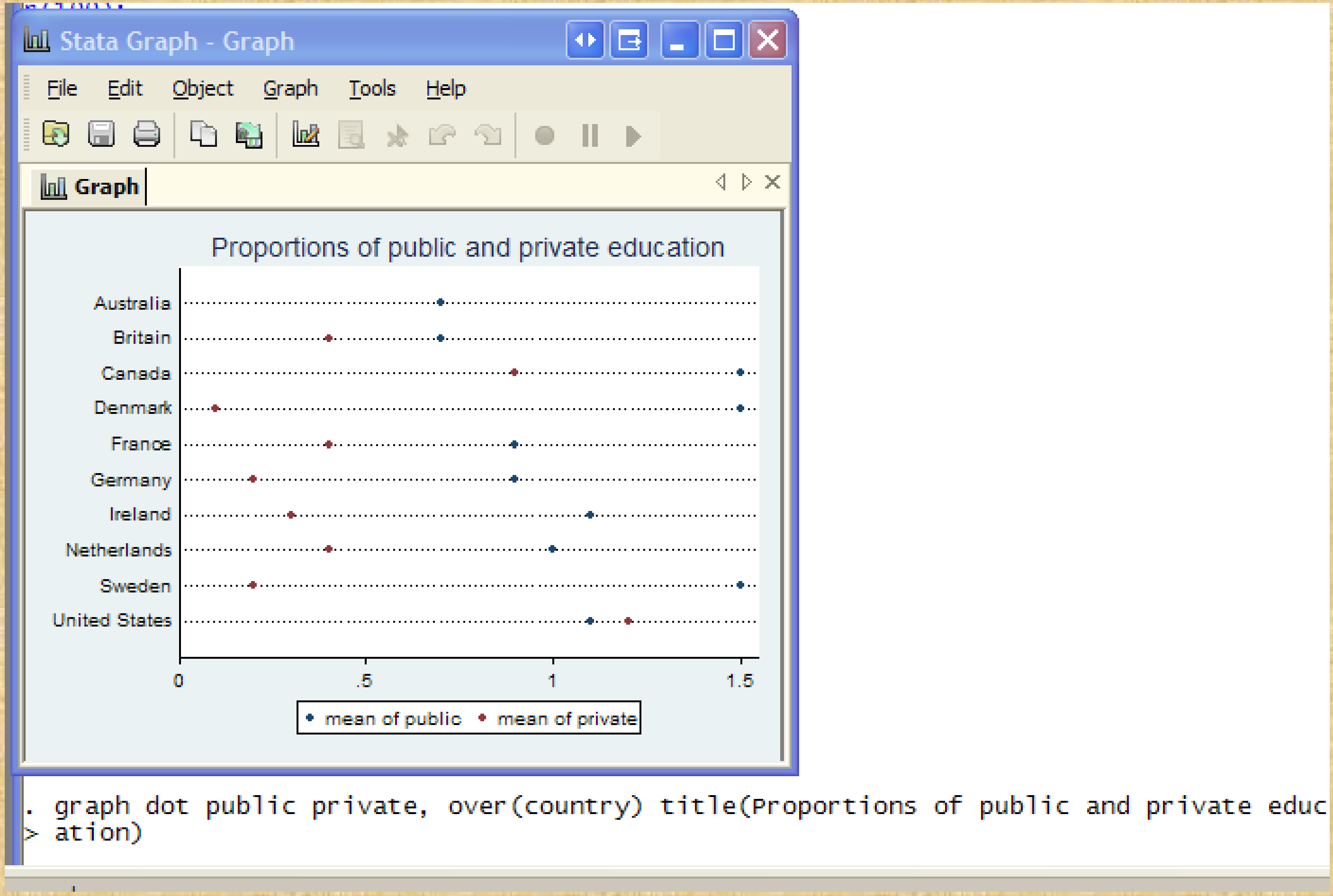
Life expectancy by sex and race over time



Survival Analysis

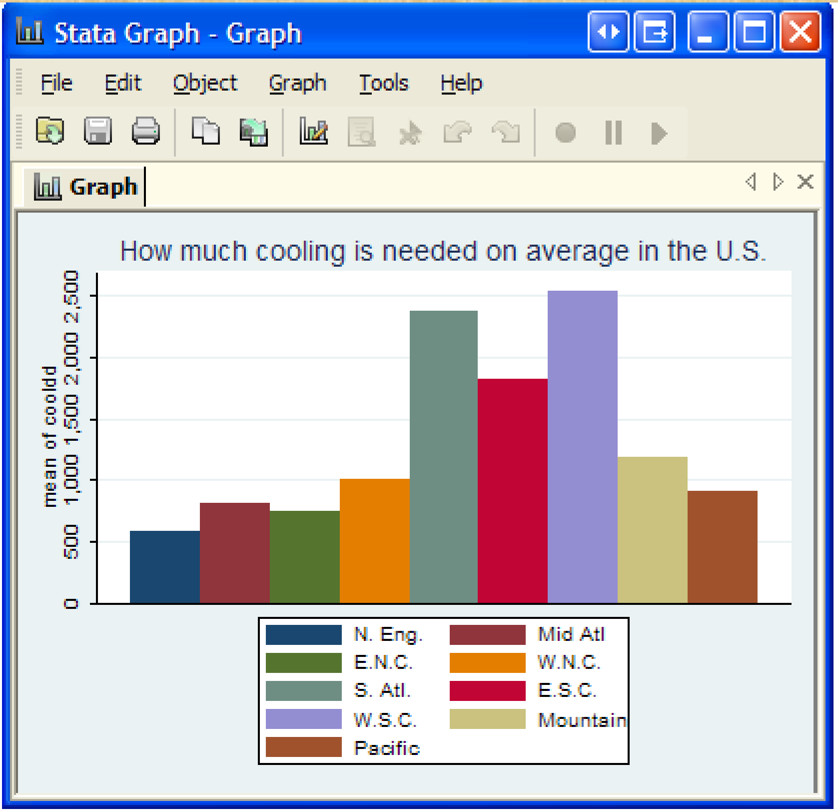


Dot plot of public and private education by country



```
. graph dot public private, over(country) title(Proportions of public and private educ
> ation)
```

How much air conditioning is needed on average in the U.S. each year?



```
graph bar (mean) cooldd, over(division) title(How much cooling is needed on average in the U.S.) asyvars
```

Saving and Exporting graphs

You can save a Stata graph with the command:

```
graph save ch1fig1.gph
```

- The gph suffix indicates that it is a Stata graph.
- If you wish to resave this graph later, attach the replace option after the graph name.

You can export a Stata graph with the command:

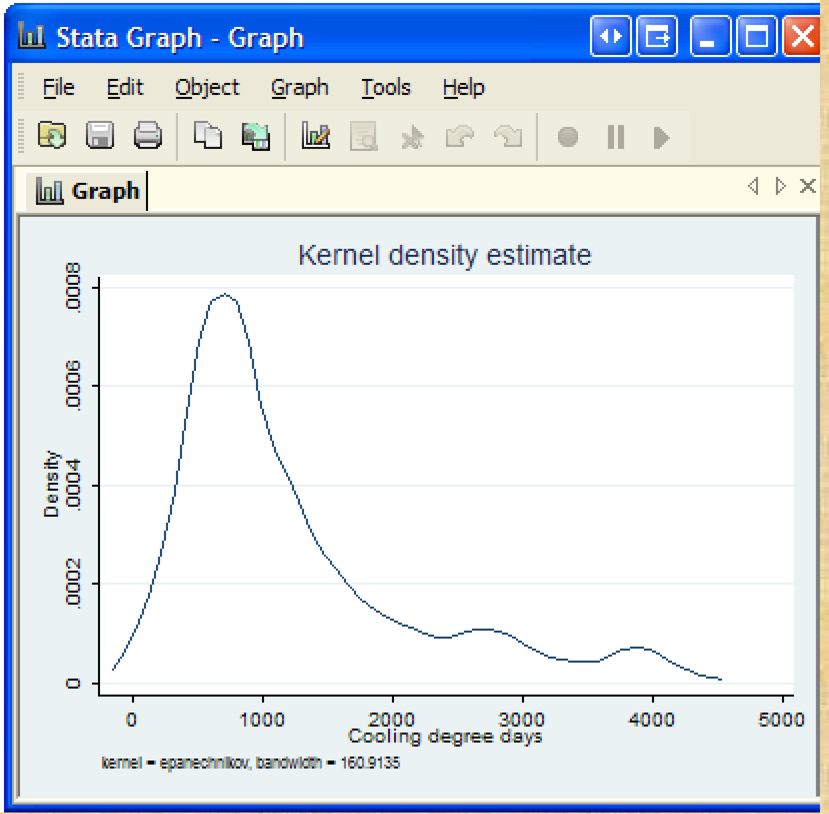
```
graph export ch1fig1.wmf, replace
graph export ch1fig1.emf, replace
graph export ch1fig1.eps, replace
graph export.ch1fig1.tif
graph export.ch1fig1.pdf
```

Distributional analysis

- Simulation with random number generators of normal, poisson, chi square, binomial, gamma, hypergeometric , and other distributions
- Kernel density plots (distributional structure)
- Histograms (with superimposed normal curves)
- Lowess plots (linearity and functional form)
- Quantile plots
- Stem and leaf plots

Kernel density plots

- Nonparametric density plots



$$f_{kernel}(x) = \frac{\mathbf{1}}{nh_j}\sum_{i=\mathbf{1}}^{n} K_j\left(\frac{(x_{ij} - x_j)}{h_j}\right)$$

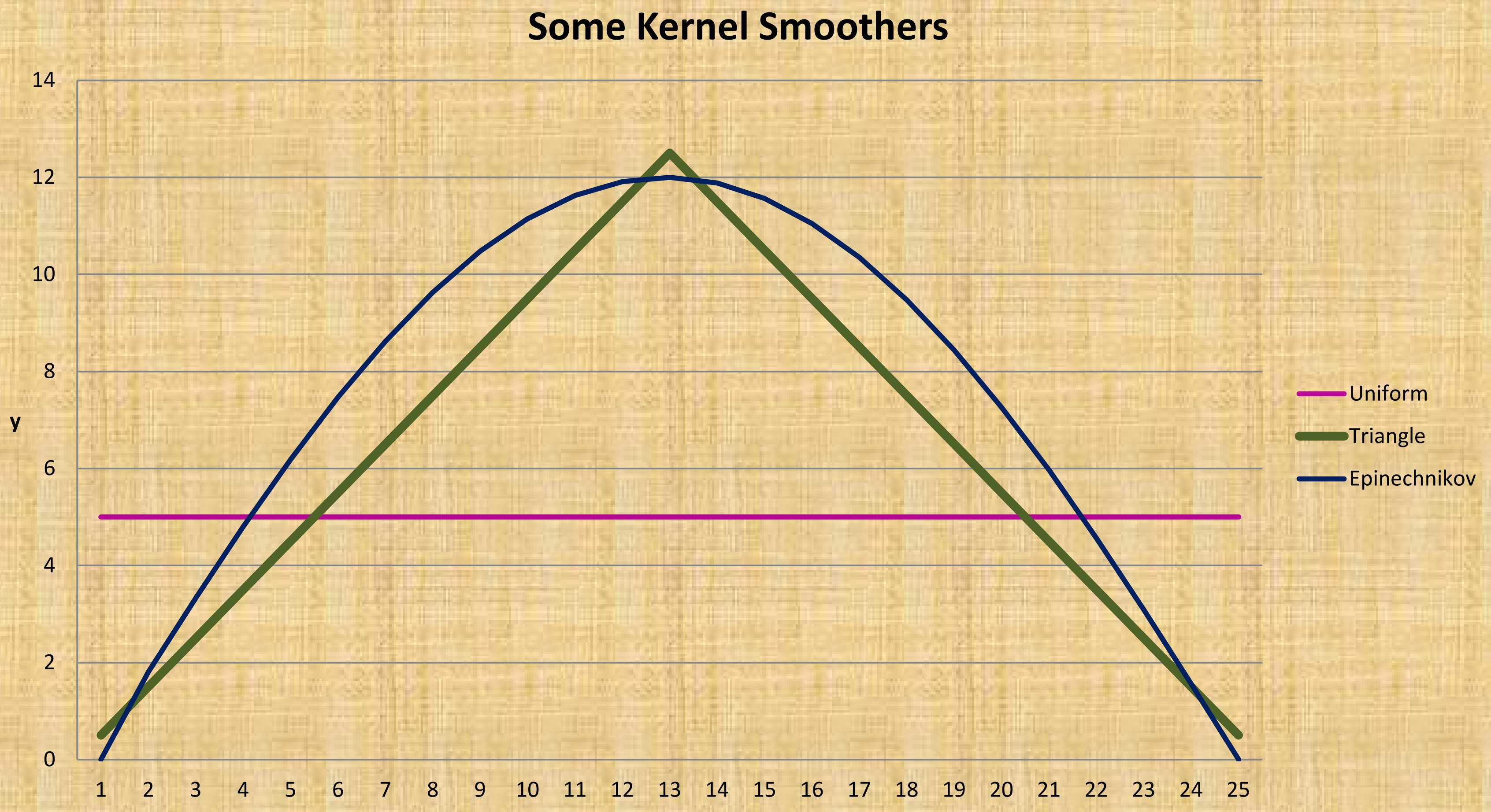
where

$f(.) = \text{kernel density estimator}$

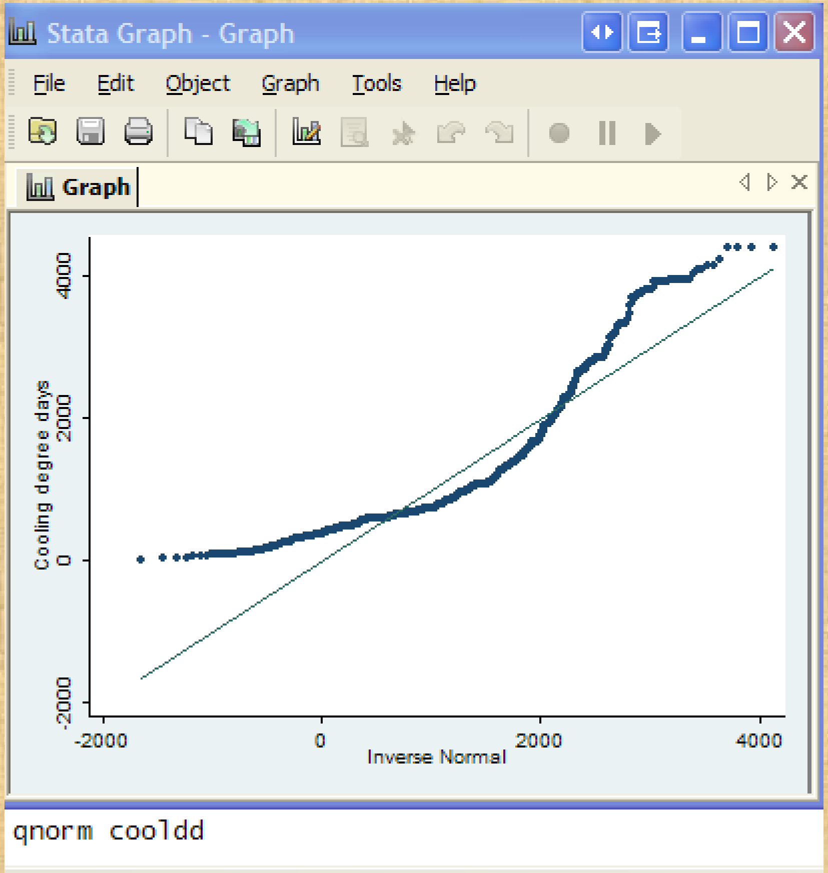
$K = \text{kernel function symmetrically weights observations}$

$h_j = \text{bandwidth for local smoothing of data}$

Some Kernel functions

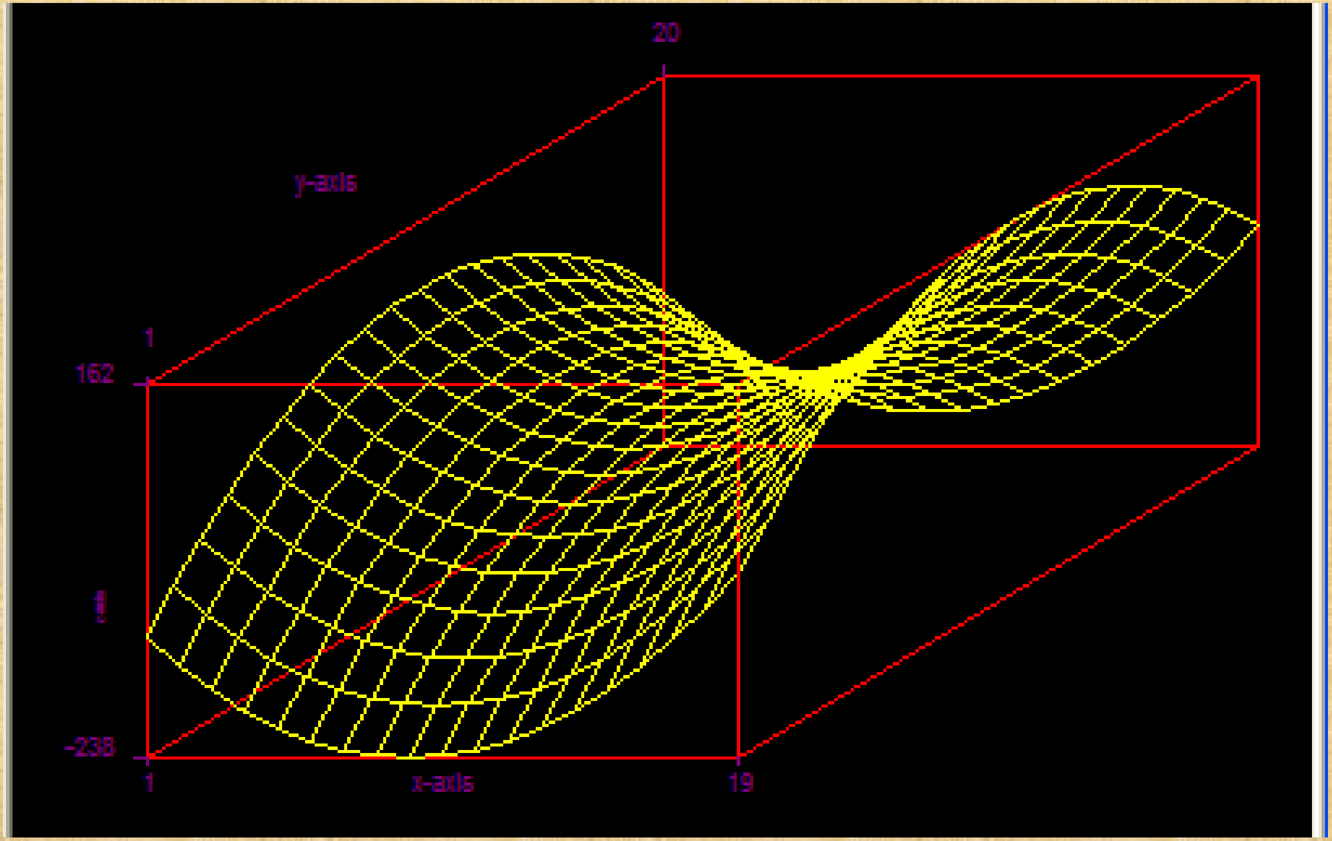


Quantile normal plots



```
qnorm cooldd
```

3d Graphs can be generated with some user effort



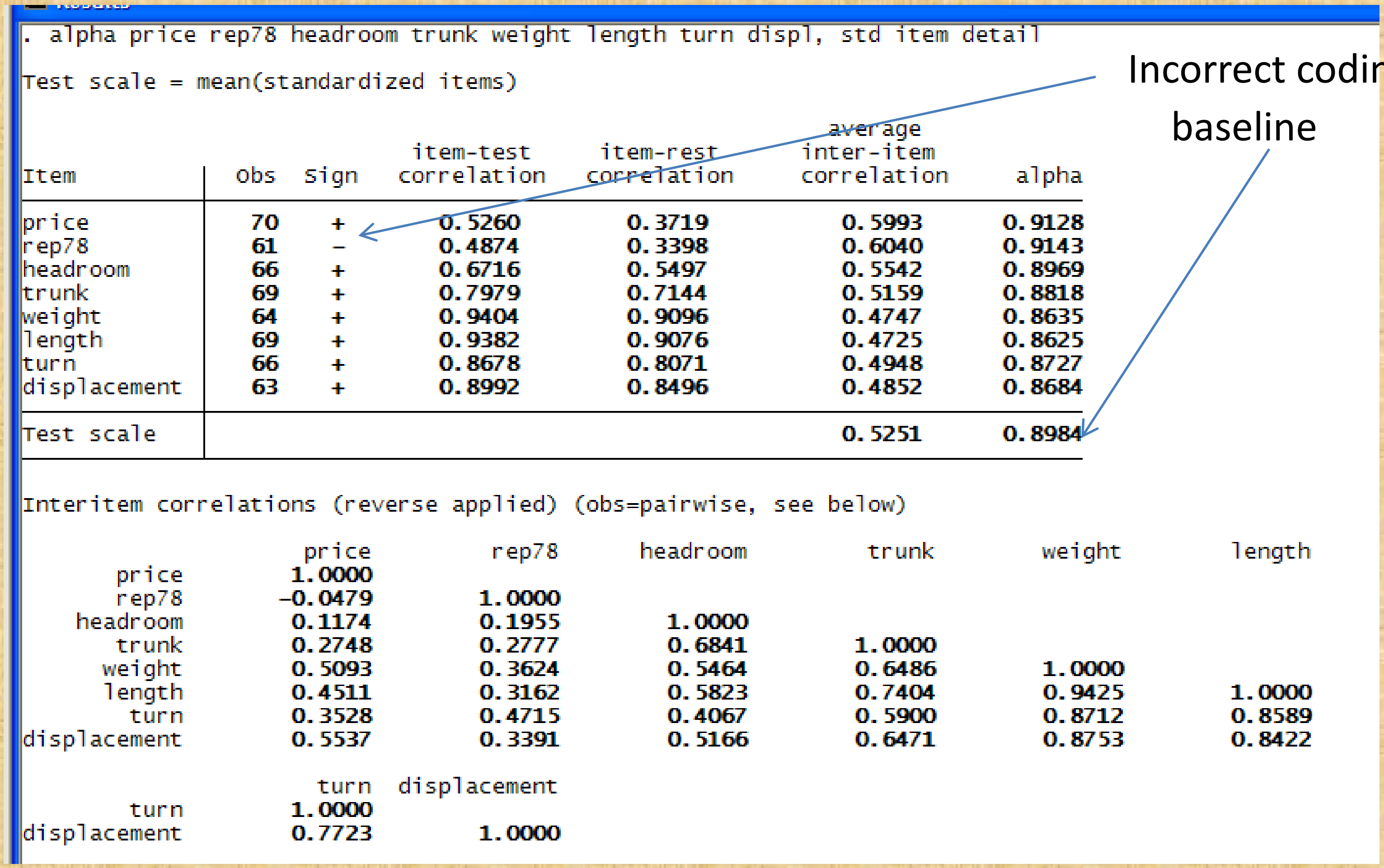
Item and Scale analysis

- Scale construction
- Alpha reliability
- Kappa reliability
- ICC23 reliability is also possible but will not be shown here. You have to download icc23 from the ssc archive.

Exercises 3

1. Plot a matrix scatterplot of headroom to weight in the dataset auto1.dta
2. Plot a lowess graph between mpg and weight
3. Use a horizontal bar chart to show the mpg of foreign and domestic cars. Put a main and axis titles in it. Put in a note or caption describing it.
4. Generate a stem-leaf plot of weight by foreign.
5. Generate a dot plot of make by mpg.
6. Generate a kernel density plot of mpg.
7. Generate a time plot of GDP downloaded from FRED.
8. Generate an overlay time plot of CPI and GDP over the same range of time, downloading both from FRED.

Cronbach Alpha reliability (internal consistency of scale items)



```
. alpha price rep78 headroom trunk weight length turn displ, std item detail

Test scale = mean(standardized items)

                                                          average
                             item-test     item-rest     inter-item
Item         |  Obs  Sign   correlation   correlation   correlation     alpha
-------------+-----------------------------------------------------------------
price        |   70    +      0.5260        0.3719        0.5993       0.9128
rep78        |   61    -      0.4874        0.3398        0.6040       0.9143
headroom     |   66    +      0.6716        0.5497        0.5542       0.8969
trunk        |   69    +      0.7979        0.7144        0.5159       0.8818
weight       |   64    +      0.9404        0.9096        0.4747       0.8635
length       |   69    +      0.9382        0.9076        0.4725       0.8625
turn         |   66    +      0.8678        0.8071        0.4948       0.8727
displacement |   63    +      0.8992        0.8496        0.4852       0.8684
-------------+-----------------------------------------------------------------
Test scale   |                                            0.5251       0.8984
-------------+-----------------------------------------------------------------

Interitem correlations (reverse applied) (obs=pairwise, see below)

                 price        rep78     headroom        trunk       weight       length
       price    1.0000
       rep78   -0.0479       1.0000
    headroom    0.1174       0.1955       1.0000
       trunk    0.2748       0.2777       0.6841       1.0000
      weight    0.5093       0.3624       0.5464       0.6486       1.0000
      length    0.4511       0.3162       0.5823       0.7404       0.9425       1.0000
        turn    0.3528       0.4715       0.4067       0.5900       0.8712       0.8589
displacement    0.5537       0.3391       0.5166       0.6471       0.8753       0.8422

                  turn  displacement
        turn    1.0000
displacement    0.7723       1.0000
```

Incorrect coding

baseline

Do we reverse code?

- If (-0.20 <= correlation => .20), we can reverse code if this improves scale alpha.
- Otherwise, we delete the item.
- We iterate until scale alpha is greater than 0.70. If scale alpha < 0.70, we use individual items instead of scale.


Jacob Cohen, PhD

Cohen’s Kappa Reliability

- Kappa reliability is a form of interrater agreement that is evidence of independent corroboration of concurrence of interpretation. The higher this agreement, the more there appears to be a consensus about the meaning of the object of evaluation.
- Kappa is designed to correct for chance agreement.

Cohen’s kappa (1960)

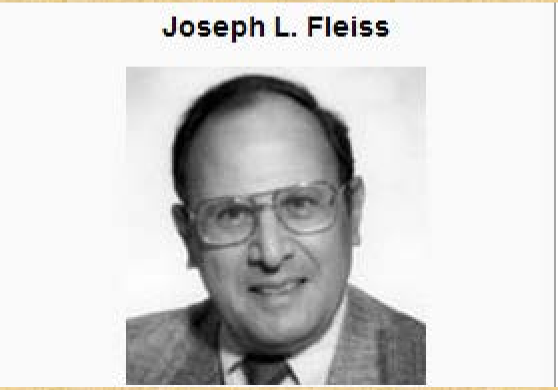
for two raters classifying n items into C categories

- The denominator in the ratio corrects for chance agreement

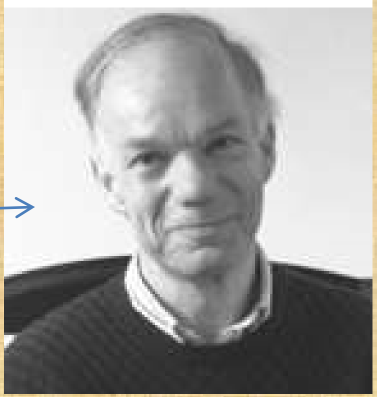
$$\kappa_{Cohen} = \frac{Pr(observed\ \ agreement) - Pr(expected\ [by\ chance]\, agreement)}{1 - Pr(expected\ [by\ chance]\ agreement)}$$

0 = no agreement
0-.20 very low agreement
.21-.40 low agreement
.41-.60 moderate agreement
.61-.8 full agreement
.81-1.00 almost perfect agreement

Joe Fleiss (Columbia)and Jack Cohen (NYU) came up with the weighted kappa



Fleiss developed the modern Intra-Class correlation coefficient with Pat Shrout (formerly of Columbia and now at NYU)



- Fleiss and Cohen(1973), "The Equivalence of the weighted Kappa and the intraclass correlation coefficient as measures of reliability" in Educational and Psychological Measurement, Vol. 33, pp. 257-268 wrote that the weighted Kappa was equivalent to the intraclass correlation coefficient as a measure of reliability.

Fleiss's Kappa (1981)

- Joe Fleiss's kappa

$$\kappa_{Fleiss} = \frac{\overline{P} - \overline{P}_e}{\boldsymbol{1} - \overline{P}_e}$$

where

the numerator accounts for actual agreement above chance,

the denominator accounts for extent of possible agreement

above chance.

Intra-class correlation Coefficient as a measure of reliability

Winer, Brown, and Michaels (1991) Statistical Principles of Experimental Design, McGraw Hill: New York, 127-129.

- If the model is a two-way ANOVA layout, are the judges fixed or random? The targets are deemed random. If the judges are fixed, the model is a two-way mixed effects ANOVA. If they are random, the model is a two-way random (randomized block design) effects ANOVA. Another effect to be controlled for is he error variance.

$$\textit{Types of Intra} - \textit{class correlation} = \textit{Cohen's multi} - \textit{rater kappa}:$$

$$\textit{When treatments are random}:$$

$$ICC(consistency) = \frac{Variance(rating) - variance(residual)}{[Variance(rating) - Avg(variance(resid)] + Variance(ratings) - Avg[variance(resid)]}$$

$$ICC(absolute\ agreement) = \frac{[Variance(rating\ of\ targets) - (variance(error)]}{Variance(error) + Average(Variance(rating\ of\ targets) - average(Variance(error))}$$

$$\textit{If targets are fixed}, ICC = \omega^2 (omega - squared)$$

Intra-class correlation in a nutshell

- It is the proportion of agreement to the total amount of variation (from agreement, disagreement, possible interaction, and error).
- There are more than 5 ways of computing this ICC.

For fixed treatments ω^2

- Stata can compute omega squared:

$$\omega^2 = \frac{\frac{\sum \tau_j^2}{k}}{\left(\frac{\sum \tau_j^2}{k} + \sigma_e^2\right)}$$

$= proportion\ of\ population\ variance\ accounted$

$for\ by\ agreement$

$where$

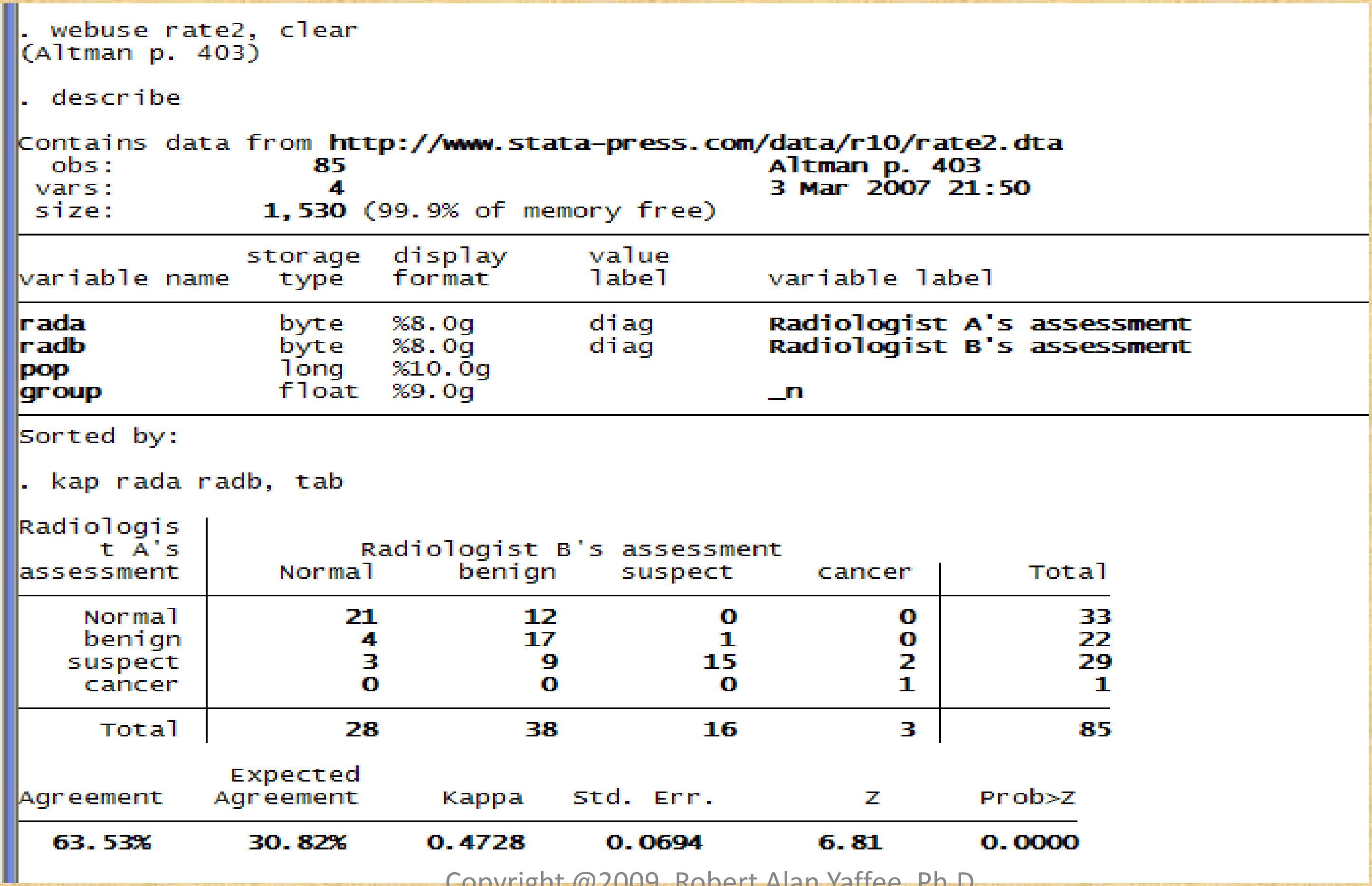
$k = number\ of\ treatment\ groups\ or\ rating\ categories$

$\tau_j^2 = Sum\ of\ squares\ of\ treatment\ or\ ratings$

$\sigma_e^2 = error\ variance$

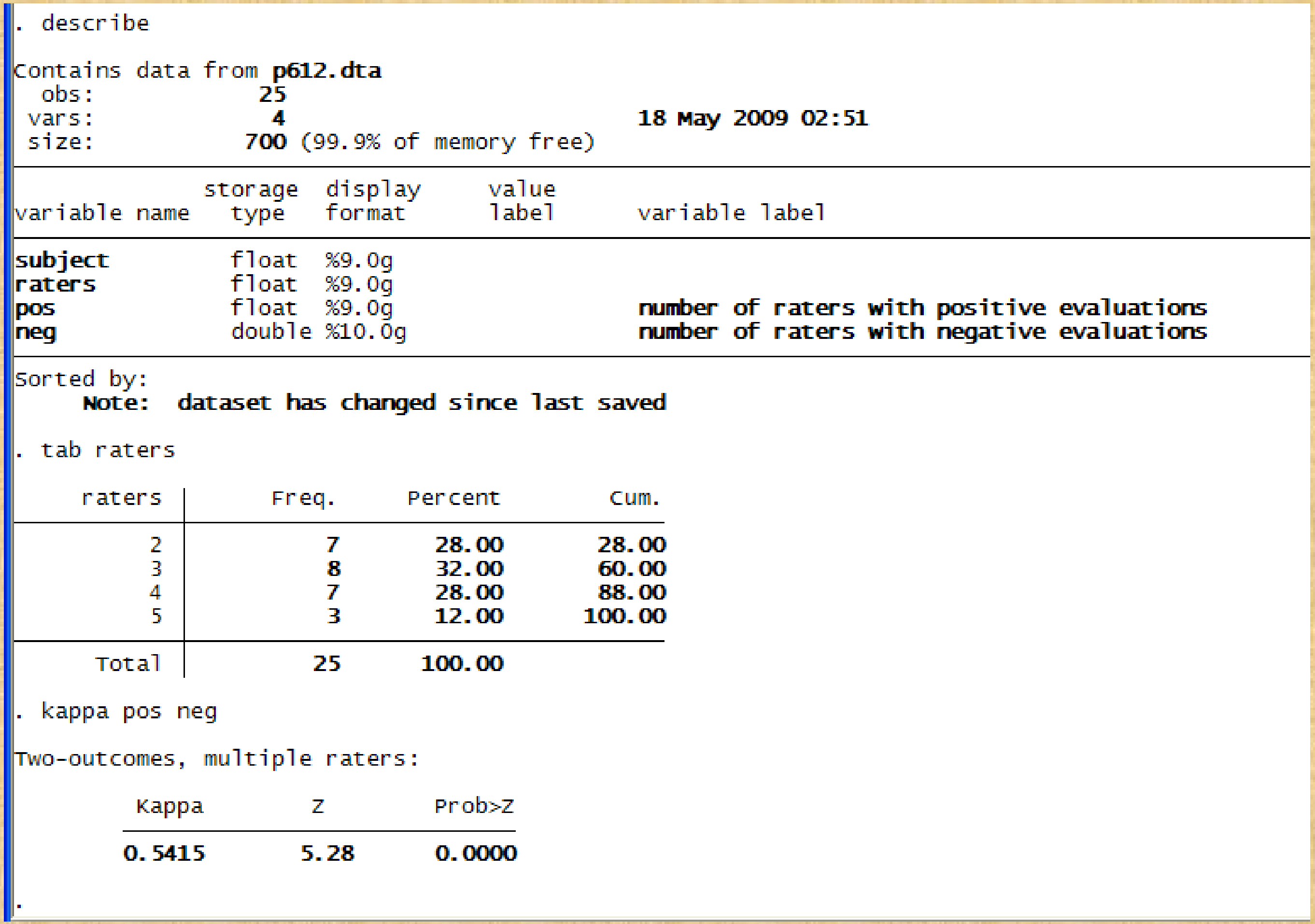
Kappa reliability

- Corrects for chance and applicable with multiple raters.



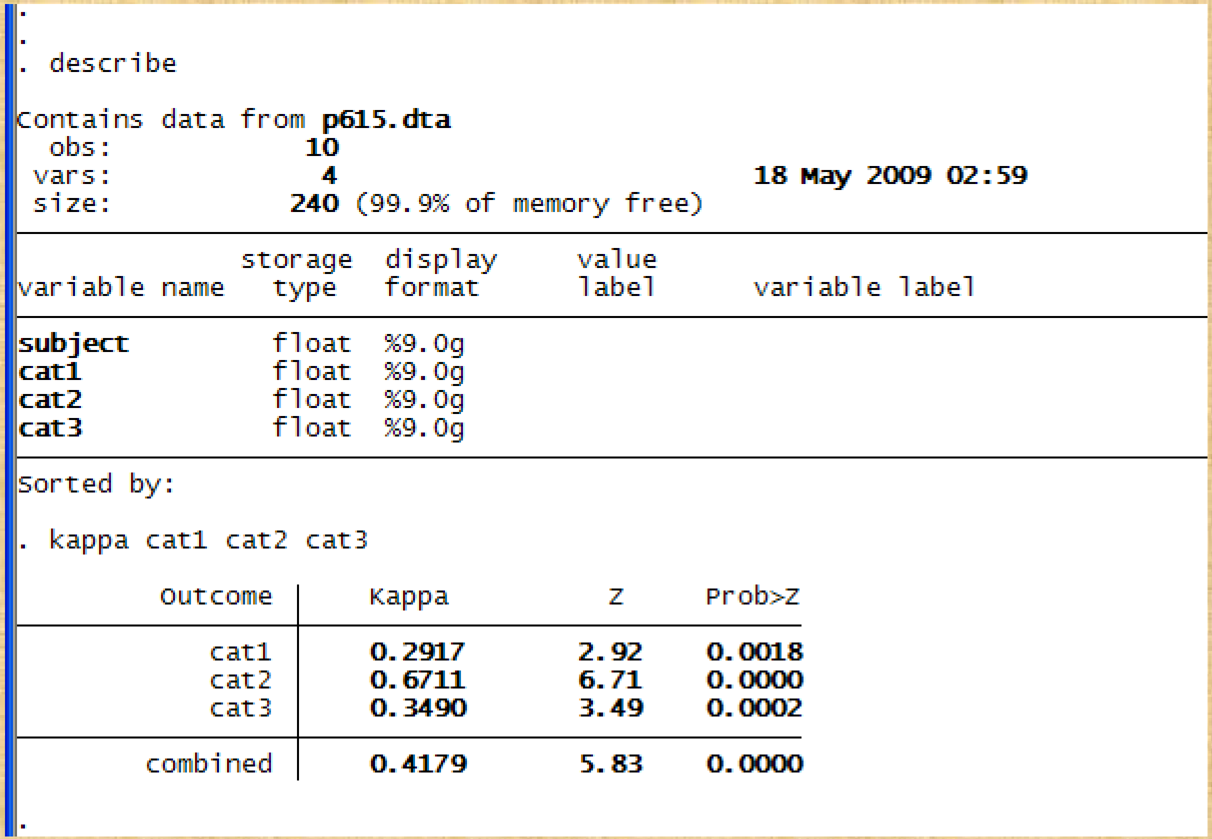
```
. webuse rate2, clear
(Altman p. 403)

. describe

Contains data from http://www.stata-press.com/data/r10/rate2.dta
  obs:            85                           Altman p. 403
 vars:             4                           3 Mar 2007 21:50
 size:         1,530 (99.9% of memory free)
-------------------------------------------------------------------------------
              storage  display     value
variable name   type   format      label      variable label
-------------------------------------------------------------------------------
rada            byte   %8.0g       diag       Radiologist A's assessment
radb            byte   %8.0g       diag       Radiologist B's assessment
pop             long   %10.0g
group           float  %9.0g                  _n
-------------------------------------------------------------------------------
Sorted by:

. kap rada radb, tab

Radiologis |
   t A's   |         Radiologist B's assessment
assessment |   Normal     benign    suspect     cancer |     Total
-----------+--------------------------------------------+----------
    Normal |       21         12          0          0 |        33
    benign |        4         17          1          0 |        22
   suspect |        3          9         15          2 |        29
    cancer |        0          0          0          1 |         1
-----------+--------------------------------------------+----------
     Total |       28         38         16          3 |        85

             Expected
Agreement   Agreement     Kappa   Std. Err.         Z      Prob>Z
-----------------------------------------------------------------
  63.53%      30.82%     0.4728     0.0694       6.81      0.0000
```

Multi-rater kappa κ



```
. describe

Contains data from p612.dta
  obs:            25
 vars:             4                          18 May 2009 02:51
 size:           700 (99.9% of memory free)
-----------------------------------------------------------------------------------------------
              storage  display     value
variable name   type   format      label      variable label
-----------------------------------------------------------------------------------------------
subject         float  %9.0g
raters          float  %9.0g
pos             float  %9.0g                  number of raters with positive evaluations
neg             double %10.0g                 number of raters with negative evaluations
-----------------------------------------------------------------------------------------------
Sorted by:
     Note:  dataset has changed since last saved

. tab raters

     raters |      Freq.     Percent        Cum.
------------+-----------------------------------
          2 |          7       28.00       28.00
          3 |          8       32.00       60.00
          4 |          7       28.00       88.00
          5 |          3       12.00      100.00
------------+-----------------------------------
      Total |         25      100.00

. kappa pos neg

Two-outcomes, multiple raters:

         Kappa        Z        Prob>Z
        -----------------------------
        0.5415      5.28       0.0000

.
```

Multi-rater multi-category fixed number of raters kappa



```
.
.
. describe

Contains data from p615.dta
  obs:            10
 vars:             4                          18 May 2009 02:59
 size:           240 (99.9% of memory free)
-------------------------------------------------------------------------
              storage  display     value
variable name   type   format      label      variable label
-------------------------------------------------------------------------
subject         float  %9.0g
cat1            float  %9.0g
cat2            float  %9.0g
cat3            float  %9.0g
-------------------------------------------------------------------------
Sorted by:

. kappa cat1 cat2 cat3

           Outcome |    Kappa          Z     Prob>Z
-------------------+-----------------------------------
              cat1 |    0.2917       2.92    0.0018
              cat2 |    0.6711       6.71    0.0000
              cat3 |    0.3490       3.49    0.0002
-------------------+-----------------------------------
          combined |    0.4179       5.83    0.0000

.
```

There are IntraClass Correlations available (types 2 and 3)

- Download from SSC archive
- ssc install icc23

Summary statistics including measure of central tendency.

- Summarize   mean range, std deviation
- Summarize, detail
- Tabstat
- Means
- Group or aggregation statistics with statsby

Enligtenment in a taxi!

- **Hardy, Godfrey H. (1877 - 1947)**
- [On Ramanujan]
  I remember once going to see him when he was lying ill at Putney. I had ridden in taxi cab number 1729 and remarked that the number seemed to me rather a dull one, and that I hoped it was not an unfavorable omen. "No," he replied, "it is a very interesting number; it is the smallest number expressible as the sum of two cubes in two different ways."
  Ramanujan, London: Cambridge University Press, 1940.

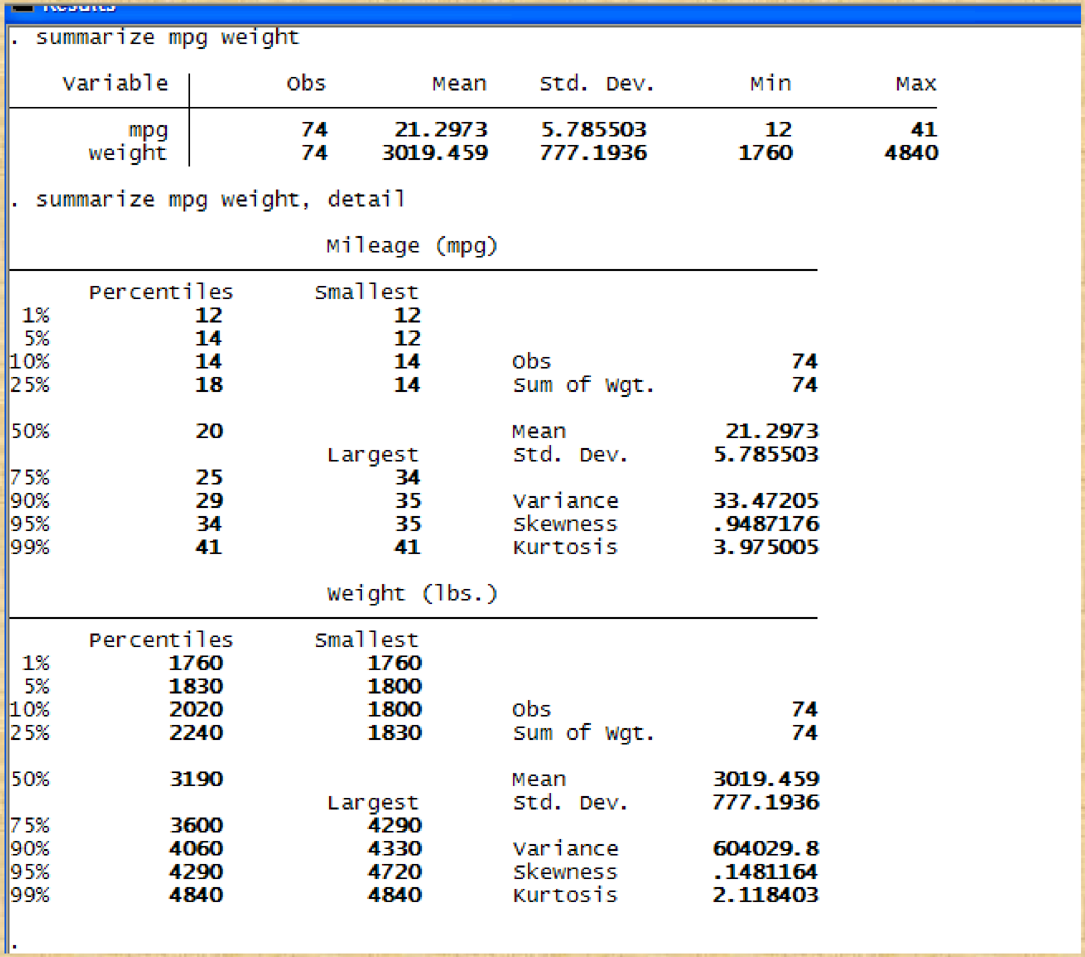
Variable transformations

- To transform or not to transform
  - When to
  - When not to
- Retransformation
- Normalizing transformations
- Variance stabilizing transformations
- To log or not to log
  - Naturally
  - By another base

Henri Poincare 1854-1912

- Later mathematicians will regard set theory as a disease from which one has recovered.

Summary univariate statistics

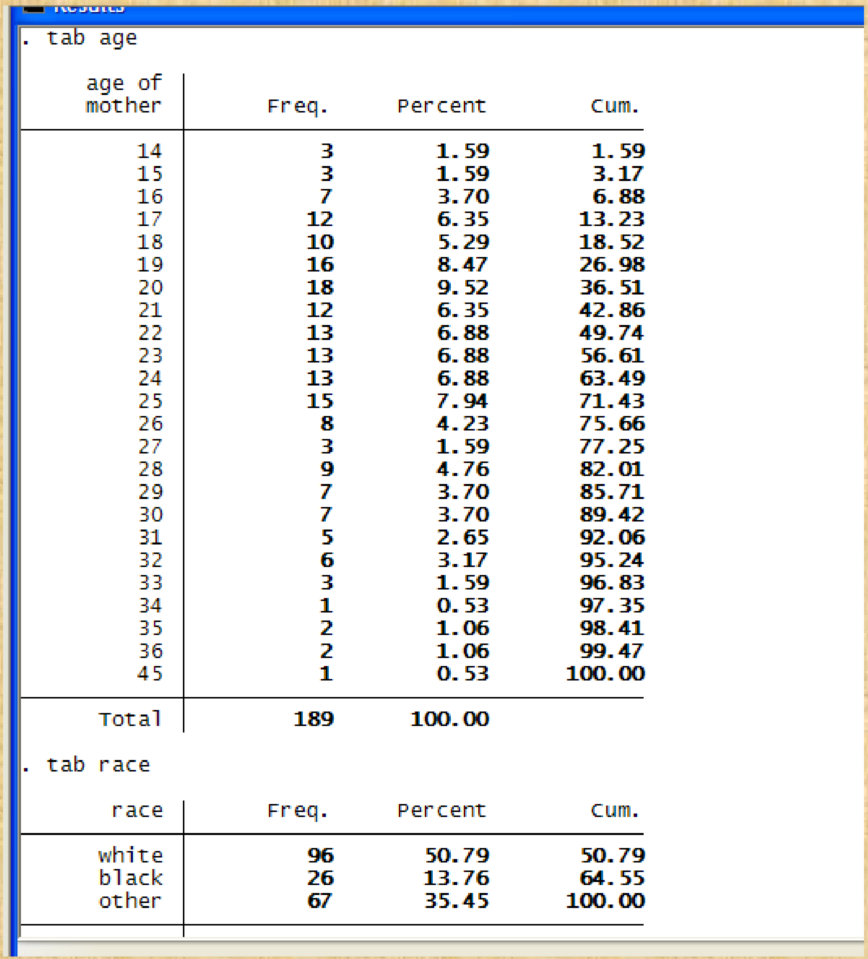


```
. summarize mpg weight

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
         mpg |        74     21.2973    5.785503         12         41
      weight |        74    3019.459    777.1936       1760       4840

. summarize mpg weight, detail

                      Mileage (mpg)
-------------------------------------------------------------
      Percentiles      Smallest
 1%           12             12
 5%           14             12
10%           14             14       Obs                  74
25%           18             14       Sum of Wgt.          74

50%           20                      Mean            21.2973
                        Largest       Std. Dev.      5.785503
75%           25             34
90%           29             35       Variance       33.47205
95%           34             35       Skewness       .9487176
99%           41             41       Kurtosis       3.975005

                      Weight (lbs.)
-------------------------------------------------------------
      Percentiles      Smallest
 1%         1760           1760
 5%         1830           1800
10%         2020           1800       Obs                  74
25%         2240           1830       Sum of Wgt.          74

50%         3190                      Mean           3019.459
                        Largest       Std. Dev.      777.1936
75%         3600           4290
90%         4060           4330       Variance       604029.8
95%         4290           4720       Skewness       .1481164
99%         4840           4840       Kurtosis       2.118403

.
```

Basic categorical data analysis

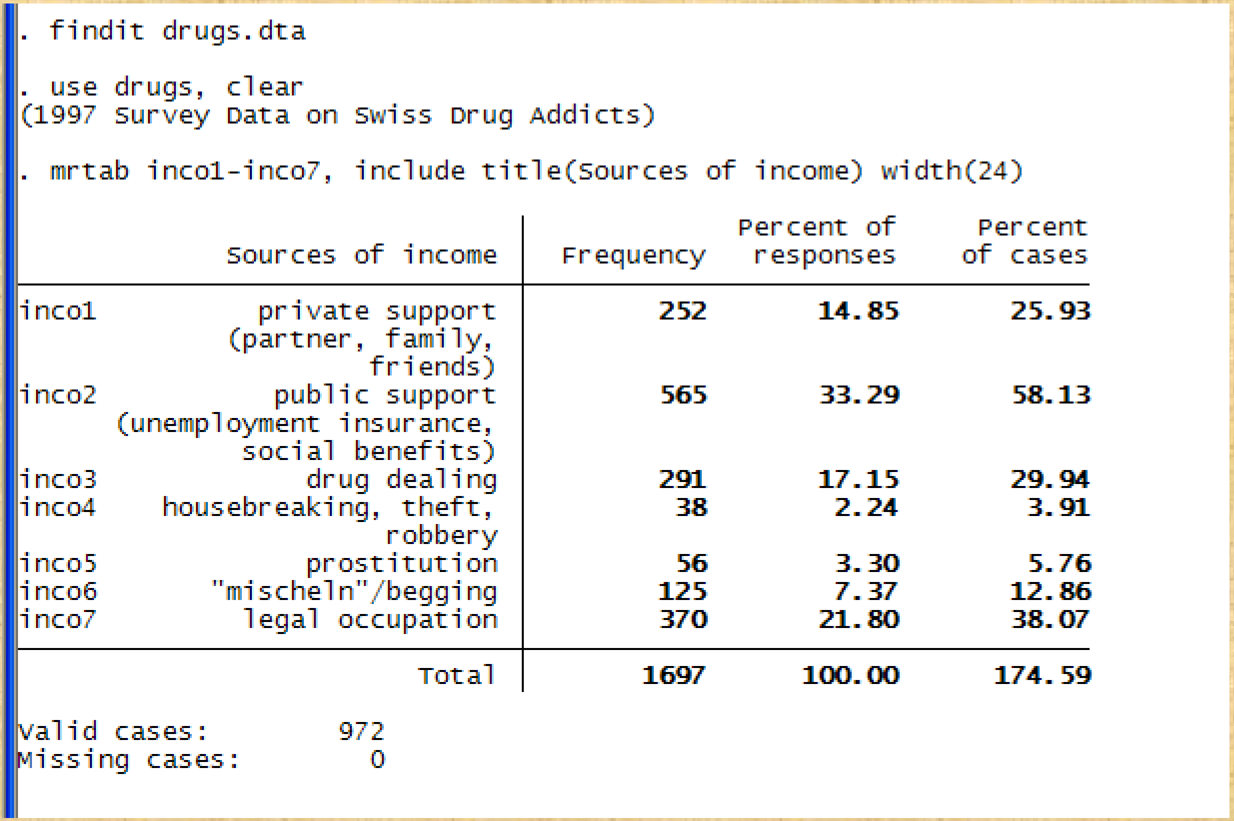
- Tabulate Tables and Crosstabulations
  - With labels
  - Without labels
  - Inference with
    - Chi-square χ^2
    - Likelihood ratio chi-square LR χ^2
    - Gamma γ
    - Kendalls τ
- Tabstat

One-way tabulations (Frequencies analysis)



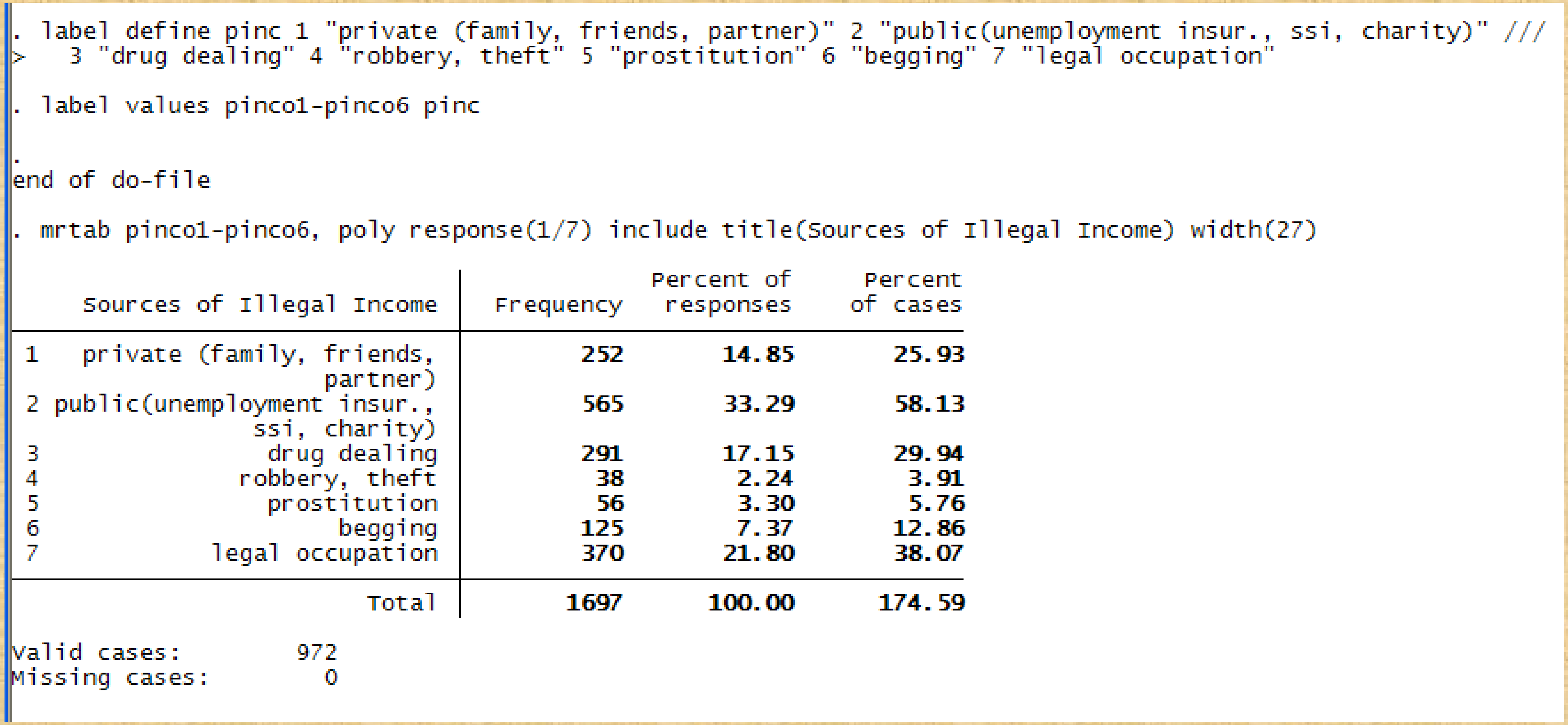
```
. tab age

     age of |
     mother |      Freq.     Percent        Cum.
------------+-----------------------------------
         14 |          3        1.59        1.59
         15 |          3        1.59        3.17
         16 |          7        3.70        6.88
         17 |         12        6.35       13.23
         18 |         10        5.29       18.52
         19 |         16        8.47       26.98
         20 |         18        9.52       36.51
         21 |         12        6.35       42.86
         22 |         13        6.88       49.74
         23 |         13        6.88       56.61
         24 |         13        6.88       63.49
         25 |         15        7.94       71.43
         26 |          8        4.23       75.66
         27 |          3        1.59       77.25
         28 |          9        4.76       82.01
         29 |          7        3.70       85.71
         30 |          7        3.70       89.42
         31 |          5        2.65       92.06
         32 |          6        3.17       95.24
         33 |          3        1.59       96.83
         34 |          1        0.53       97.35
         35 |          2        1.06       98.41
         36 |          2        1.06       99.47
         45 |          1        0.53      100.00
------------+-----------------------------------
      Total |        189      100.00

. tab race

       race |      Freq.     Percent        Cum.
------------+-----------------------------------
      white |         96       50.79       50.79
      black |         26       13.76       64.55
      other |         67       35.45      100.00
------------+-----------------------------------
```

Multiple response using (dummy indicators) courtesy of Ben Jann ETH



```
. findit drugs.dta

. use drugs, clear
(1997 Survey Data on Swiss Drug Addicts)

. mrtab inco1-inco7, include title(Sources of income) width(24)

                               |                Percent of      Percent
             Sources of income |  Frequency      responses     of cases
-------------------------------+-----------------------------------------
inco1          private support |        252          14.85        25.93
            (partner, family,  |
                      friends) |
inco2           public support |        565          33.29        58.13
      (unemployment insurance, |
              social benefits) |
inco3             drug dealing |        291          17.15        29.94
inco4     housebreaking, theft,|         38           2.24         3.91
                       robbery |
inco5             prostitution |         56           3.30         5.76
inco6        "mischeln"/begging|        125           7.37        12.86
inco7          legal occupation|        370          21.80        38.07
-------------------------------+-----------------------------------------
                         Total |       1697         100.00       174.59

Valid cases:         972
Missing cases:         0
```

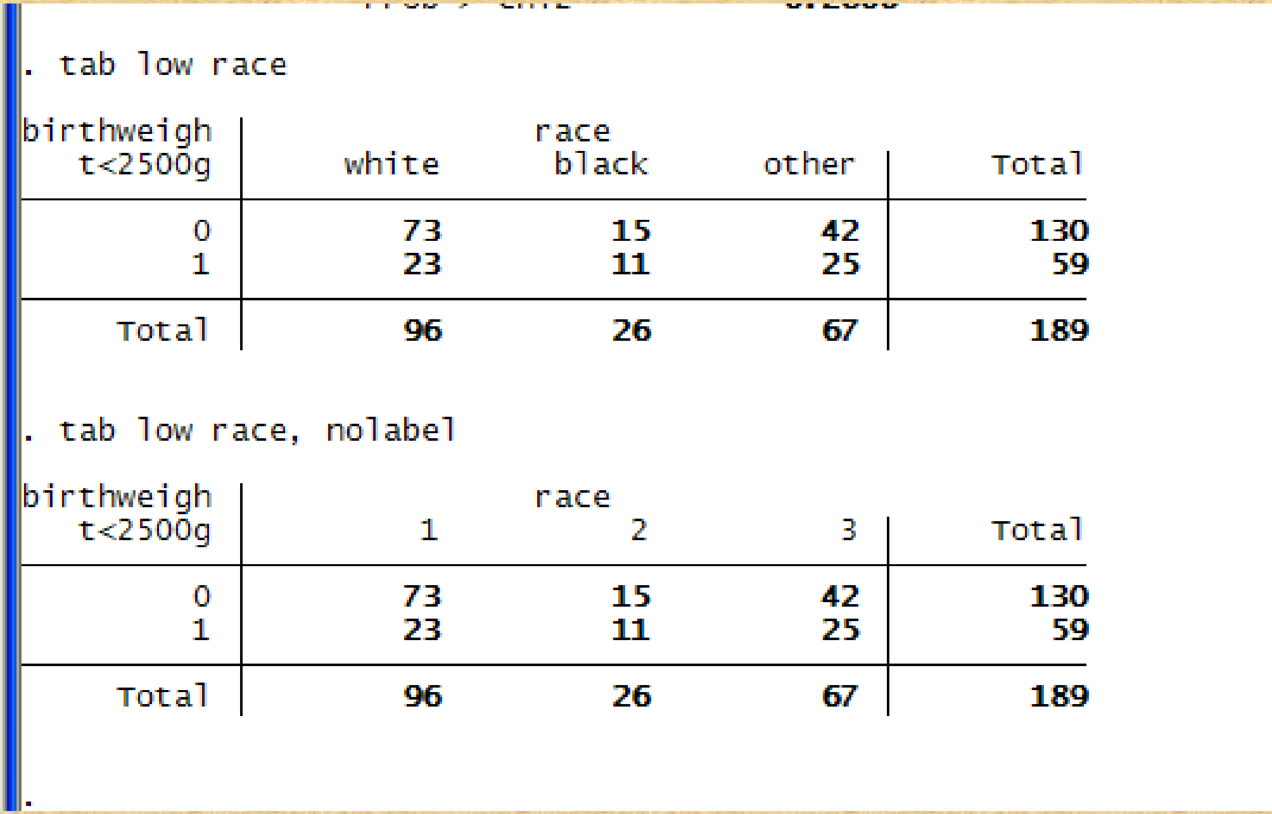
Multiple response (polytomous categories) courtesy of Ben Jann (ETH)



```
. label define pinc 1 "private (family, friends, partner)" 2 "public(unemployment insur., ssi, charity)" ///
>    3 "drug dealing" 4 "robbery, theft" 5 "prostitution" 6 "begging" 7 "legal occupation"

. label values pinco1-pinco6 pinc

.
end of do-file

. mrtab pinco1-pinco6, poly response(1/7) include title(Sources of Illegal Income) width(27)

                              |               Percent of       Percent
     Sources of Illegal Income |  Frequency    responses      of cases
------------------------------+----------------------------------------
1   private (family, friends, |        252        14.85         25.93
                     partner) |
2 public(unemployment insur., |        565        33.29         58.13
                  ssi, charity) |
3                drug dealing |        291        17.15         29.94
4              robbery, theft |         38         2.24          3.91
5                prostitution |         56         3.30          5.76
6                     begging |        125         7.37         12.86
7            legal occupation |        370        21.80         38.07
------------------------------+----------------------------------------
                        Total |       1697       100.00        174.59

Valid cases:        972
Missing cases:        0
```

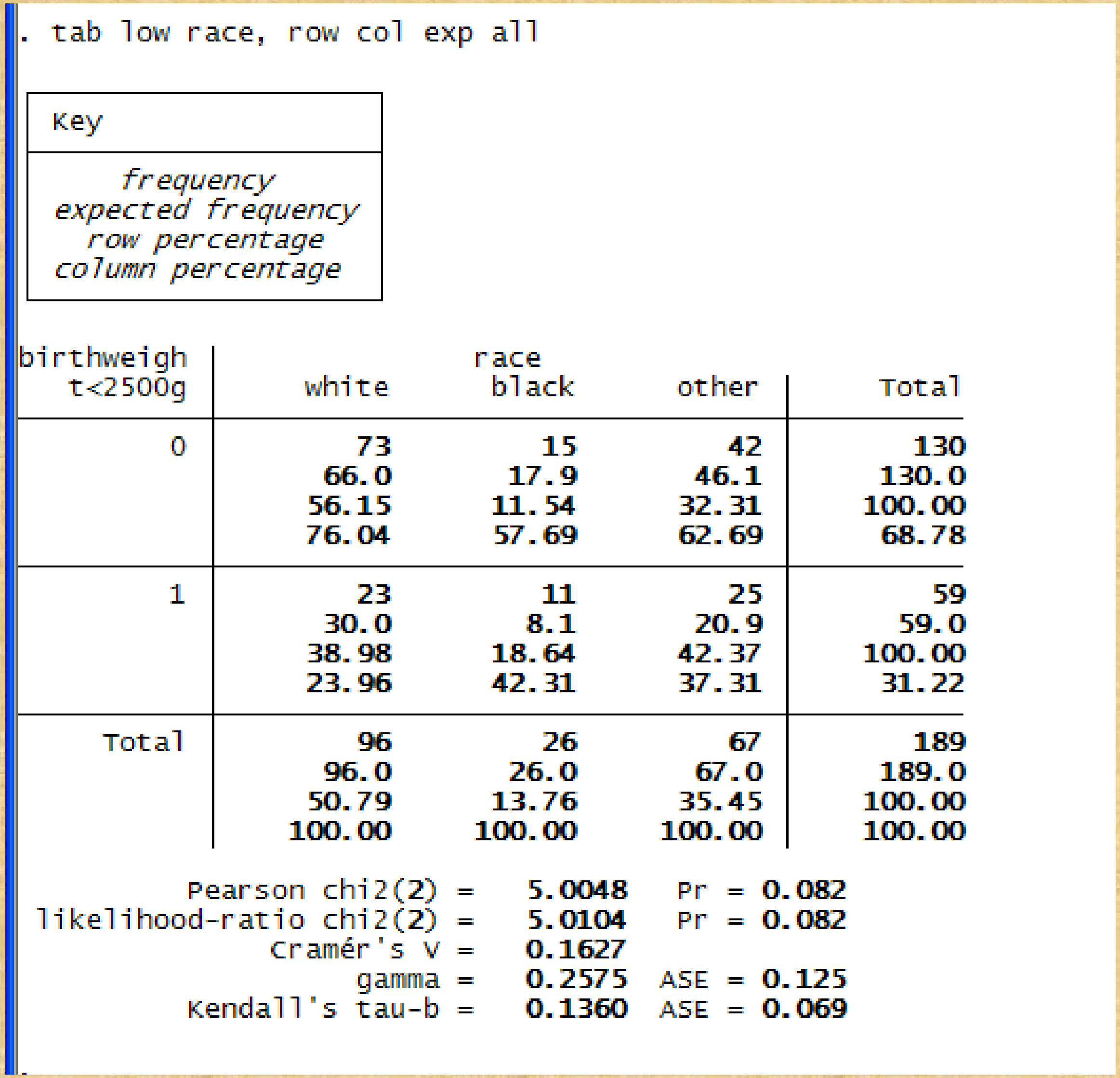
Two-way Tabulations with and without labels

Leo Goodman developed much categorical data analysis.



```
. tab low race

birthweigh |                 race
   t<2500g |     white      black      other |     Total
-----------+---------------------------------+----------
         0 |        73         15         42 |       130
         1 |        23         11         25 |        59
-----------+---------------------------------+----------
     Total |        96         26         67 |       189

. tab low race, nolabel

birthweigh |                 race
   t<2500g |         1          2          3 |     Total
-----------+---------------------------------+----------
         0 |        73         15         42 |       130
         1 |        23         11         25 |        59
-----------+---------------------------------+----------
     Total |        96         26         67 |       189


.
```

Bivariate tabulation inference



```
. tab low race, row col exp all

+--------------------+
| Key                |
|--------------------|
|     frequency      |
| expected frequency |
|   row percentage   |
| column percentage  |
+--------------------+

birthweigh |               race
 t<2500g   |     white      black      other |     Total
-----------+---------------------------------+----------
         0 |        73         15         42 |       130
           |      66.0       17.9       46.1 |     130.0
           |     56.15      11.54      32.31 |    100.00
           |     76.04      57.69      62.69 |     68.78
-----------+---------------------------------+----------
         1 |        23         11         25 |        59
           |      30.0        8.1       20.9 |      59.0
           |     38.98      18.64      42.37 |    100.00
           |     23.96      42.31      37.31 |     31.22
-----------+---------------------------------+----------
     Total |        96         26         67 |       189
           |      96.0       26.0       67.0 |     189.0
           |     50.79      13.76      35.45 |    100.00
           |    100.00     100.00     100.00 |    100.00

          Pearson chi2(2) =   5.0048   Pr = 0.082
 likelihood-ratio chi2(2) =   5.0104   Pr = 0.082
               Cramér's V =   0.1627
                    gamma =   0.2575  ASE = 0.125
          Kendall's tau-b =   0.1360  ASE = 0.069

.
```

Pearson Chi-square

- Named after Karl Pearson

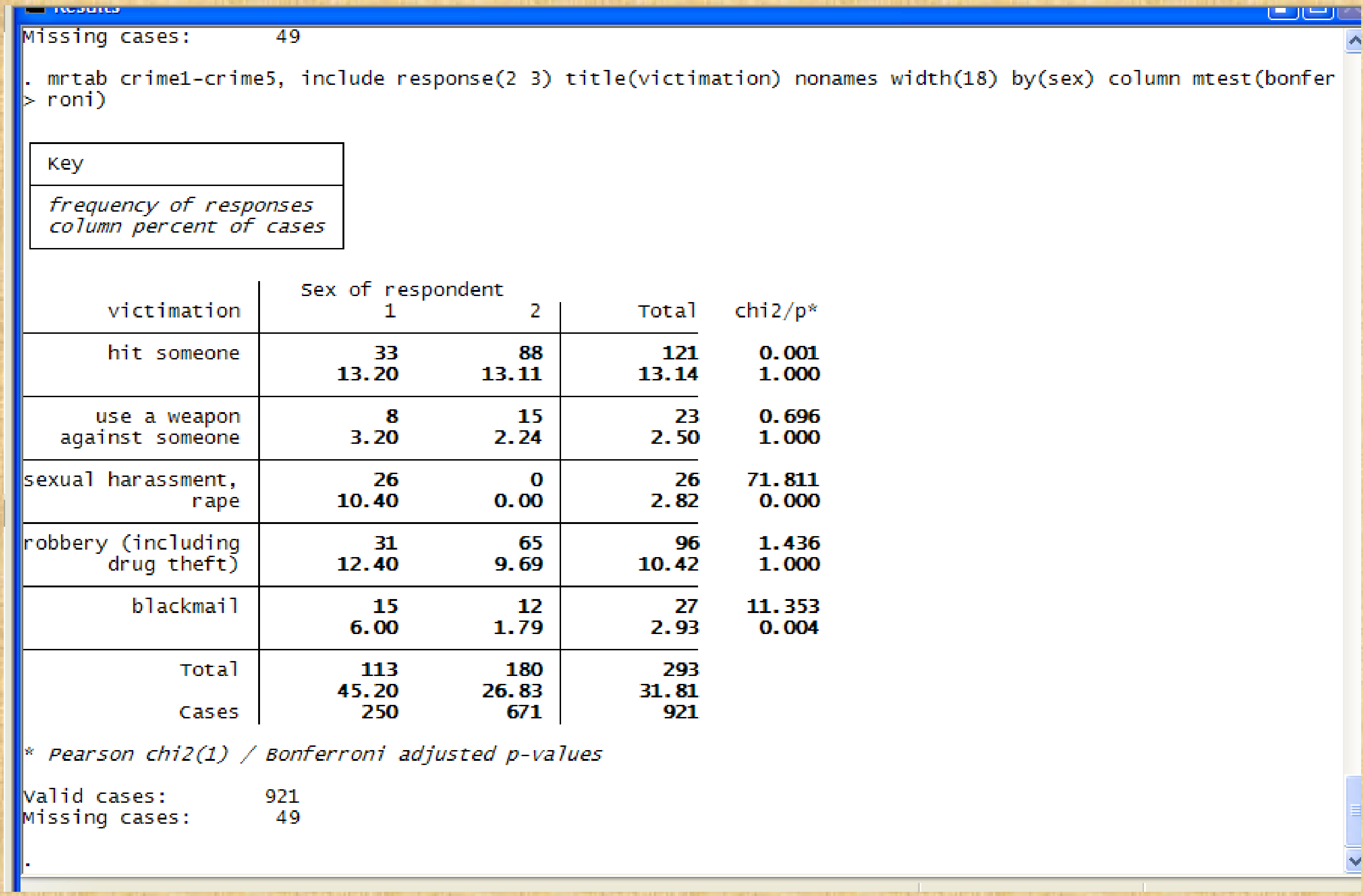
$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\left(\frac{(o_{ij}-e_{ij})^2}{e_{ij}}\right)$$

$$e_{ij} = \exp ected \;\; frequency = \frac{(row\ total_i * column\ total_j)}{Grand\ \ total_{ij}}$$

$$o_{ij} = observed\ \ count\ in\ cell\ of\ row\ i$$

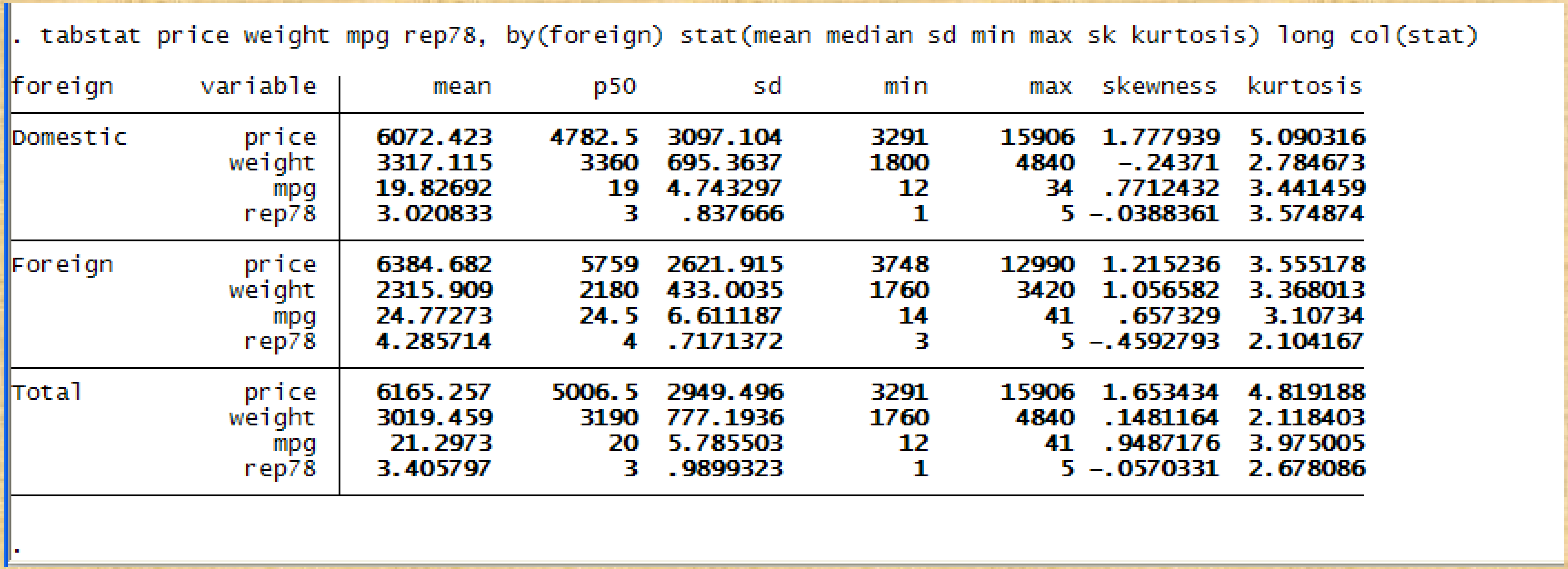
$$and\ \ column\ \ j$$

Multiple Response courtesy of Ben Jann, ETH



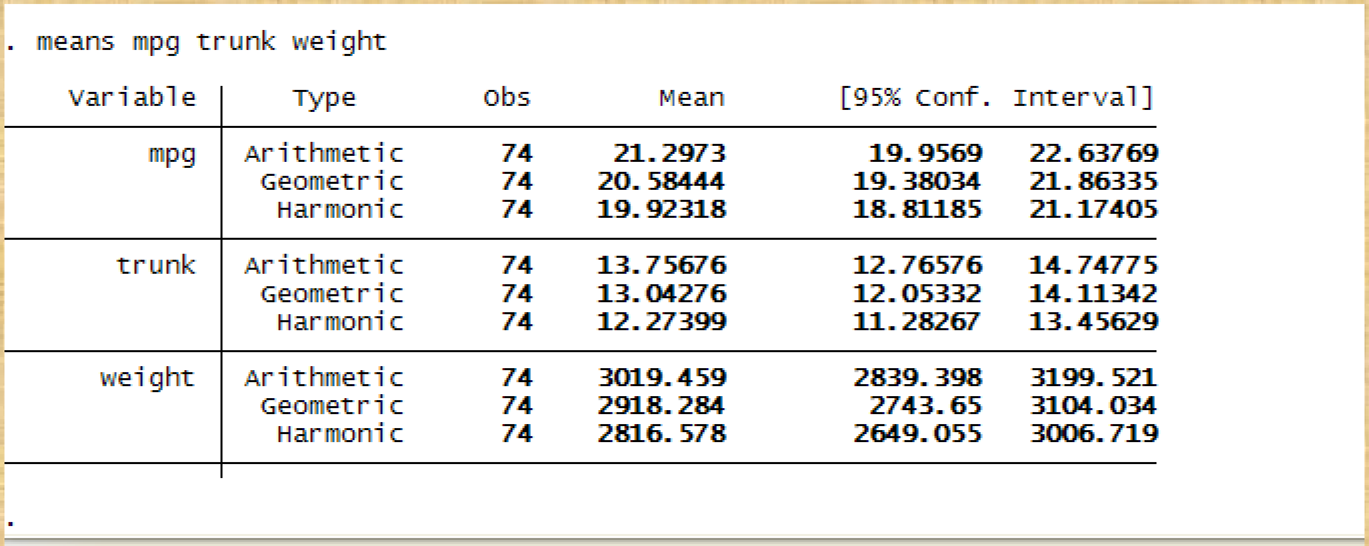
```
Missing cases:       49

. mrtab crime1-crime5, include response(2 3) title(victimation) nonames width(18) by(sex) column mtest(bonfer
> roni)

  +-------------------------+
  | Key                     |
  |-------------------------|
  | frequency of responses  |
  | column percent of cases |
  +-------------------------+

                   |   Sex of respondent
       victimation |        1           2 |      Total    chi2/p*
-------------------+-----------------------+---------------------
       hit someone |       33          88 |        121      0.001
                   |    13.20       13.11 |      13.14      1.000
-------------------+-----------------------+---------------------
      use a weapon |        8          15 |         23      0.696
   against someone |     3.20        2.24 |       2.50      1.000
-------------------+-----------------------+---------------------
sexual harassment, |       26           0 |         26     71.811
              rape |    10.40        0.00 |       2.82      0.000
-------------------+-----------------------+---------------------
robbery (including |       31          65 |         96      1.436
       drug theft) |    12.40        9.69 |      10.42      1.000
-------------------+-----------------------+---------------------
         blackmail |       15          12 |         27     11.353
                   |     6.00        1.79 |       2.93      0.004
-------------------+-----------------------+---------------------
             Total |      113         180 |        293
                   |    45.20       26.83 |      31.81
             Cases |      250         671 |        921

* Pearson chi2(1) / Bonferroni adjusted p-values

Valid cases:        921
Missing cases:       49

.
```

Customized tables



```
. tabstat price weight mpg rep78, by(foreign) stat(mean median sd min max sk kurtosis) long col(stat)

foreign     variable |      mean       p50        sd       min       max  skewness  kurtosis
---------------------+----------------------------------------------------------------------
Domestic       price |  6072.423    4782.5  3097.104      3291     15906  1.777939  5.090316
              weight |  3317.115      3360  695.3637      1800      4840   -.24371  2.784673
                 mpg |  19.82692        19  4.743297        12        34  .7712432  3.441459
               rep78 |  3.020833         3   .837666         1         5 -.0388361  3.574874
---------------------+----------------------------------------------------------------------
Foreign        price |  6384.682      5759  2621.915      3748     12990  1.215236  3.555178
              weight |  2315.909      2180  433.0035      1760      3420  1.056582  3.368013
                 mpg |  24.77273      24.5  6.611187        14        41   .657329   3.10734
               rep78 |  4.285714         4  .7171372         3         5 -.4592793  2.104167
---------------------+----------------------------------------------------------------------
Total          price |  6165.257    5006.5  2949.496      3291     15906  1.653434  4.819188
              weight |  3019.459      3190  777.1936      1760      4840  .1481164  2.118403
                 mpg |   21.2973        20  5.785503        12        41  .9487176  3.975005
               rep78 |  3.405797         3  .9899323         1         5 -.0570331  2.678086
--------------------------------------------------------------------------------------------

.
```

Means



```
. means mpg trunk weight

    Variable |    Type             Obs        Mean       [95% Conf. Interval]
-------------+---------------------------------------------------------------
         mpg | Arithmetic           74     21.2973        19.9569   22.63769
             |  Geometric           74    20.58444       19.38034   21.86335
             |   Harmonic           74    19.92318       18.81185   21.17405
-------------+---------------------------------------------------------------
       trunk | Arithmetic           74    13.75676       12.76576   14.74775
             |  Geometric           74    13.04276       12.05332   14.11342
             |   Harmonic           74    12.27399       11.28267   13.45629
-------------+---------------------------------------------------------------
      weight | Arithmetic           74    3019.459       2839.398   3199.521
             |  Geometric           74    2918.284        2743.65   3104.034
             |   Harmonic           74    2816.578       2649.055   3006.719
-------------+---------------------------------------------------------------

.
```

Comparison of two means

- Parametric t-tests
  - Assumptions
    - Observations are i.i.d.
    - Variances may be equal or corrected for nonequality
  - One sample
  - Two independent sample
  - Paired
- Alternative Nonparametric rank tests
  - Man-Whitney U test
  - Wilcoxon signrank

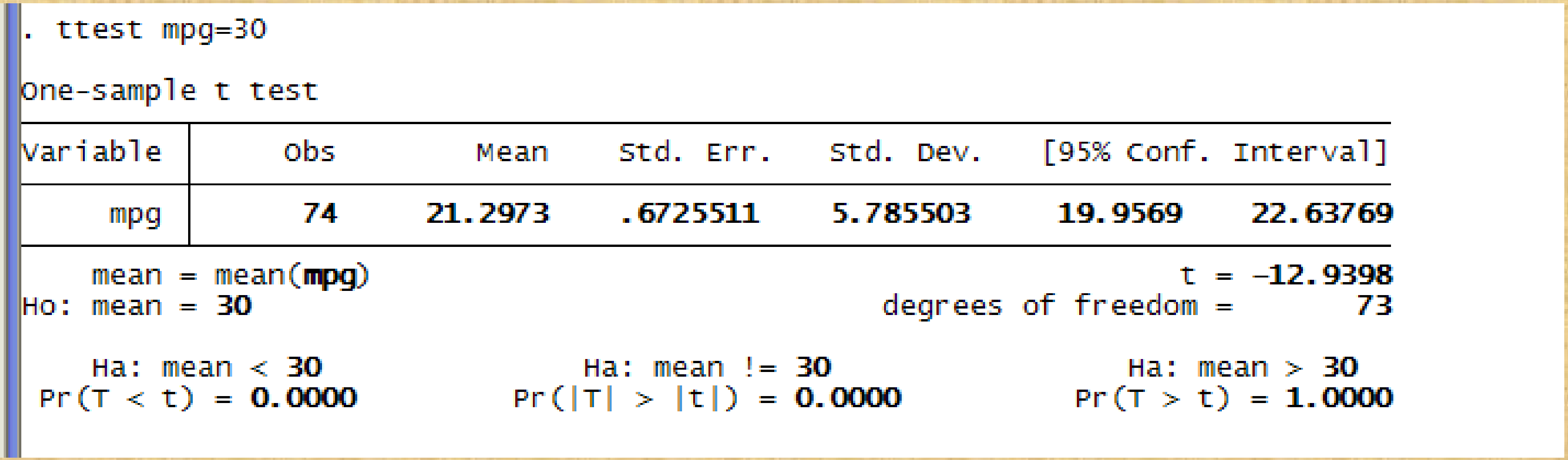
William S. Gosset
(a.k.a. Student)

- Worked at Guiness's brewery in Dublin and developed the t tests and t distribution to solve problems he encountered there.

William Sealy Gosset

Student in 1908

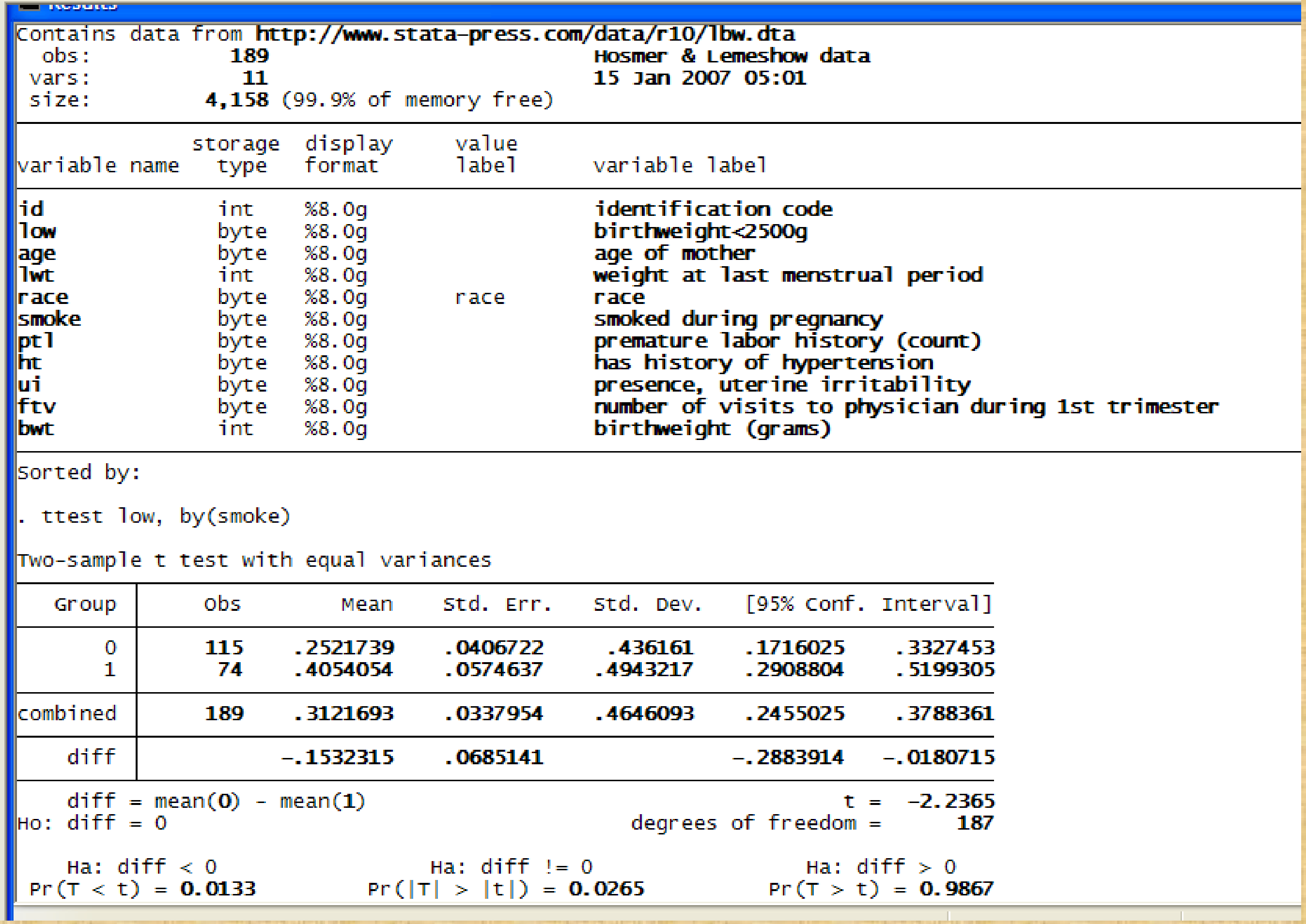
One sample t-test



```
. ttest mpg=30

One-sample t test
------------------------------------------------------------------------------
Variable |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
     mpg |      74     21.2973    .6725511    5.785503     19.9569    22.63769
------------------------------------------------------------------------------
    mean = mean(mpg)                                              t = -12.9398
Ho: mean = 30                                    degrees of freedom =       73

    Ha: mean < 30               Ha: mean != 30                Ha: mean > 30
 Pr(T < t) = 0.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 1.0000
```

$$One\ sample\ t = \frac{\overline{x}_{2i} - \mu_0}{\frac{sd}{\sqrt{n}}} \qquad df = n - 1$$

Independent samples t-test



```
Contains data from http://www.stata-press.com/data/r10/lbw.dta
  obs:           189                          Hosmer & Lemeshow data
 vars:            11                          15 Jan 2007 05:01
 size:         4,158 (99.9% of memory free)

              storage  display     value
variable name   type   format      label      variable label

id              int    %8.0g                  identification code
low             byte   %8.0g                  birthweight<2500g
age             byte   %8.0g                  age of mother
lwt             int    %8.0g                  weight at last menstrual period
race            byte   %8.0g       race       race
smoke           byte   %8.0g                  smoked during pregnancy
ptl             byte   %8.0g                  premature labor history (count)
ht              byte   %8.0g                  has history of hypertension
ui              byte   %8.0g                  presence, uterine irritability
ftv             byte   %8.0g                  number of visits to physician during 1st trimester
bwt             int    %8.0g                  birthweight (grams)

Sorted by:

. ttest low, by(smoke)

Two-sample t test with equal variances

   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       0 |     115    .2521739    .0406722     .436161    .1716025    .3327453
       1 |      74    .4054054    .0574637    .4943217    .2908804    .5199305
---------+--------------------------------------------------------------------
combined |     189    .3121693    .0337954    .4646093    .2455025    .3788361
---------+--------------------------------------------------------------------
    diff |           -.1532315    .0685141               -.2883914   -.0180715
------------------------------------------------------------------------------
    diff = mean(0) - mean(1)                                      t =  -2.2365
Ho: diff = 0                                     degrees of freedom =      187

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.0133         Pr(|T| > |t|) = 0.0265          Pr(T > t) = 0.9867
```

Independent sample t-test

- Separate sample t test:

$$t = \frac{\bar{y} - \bar{x}}{\left(\frac{(n_y - 1)s_x^2 + (n_x - 1)s_y^2}{n_x + n_y - 2} \right)^{1/2} \left(\frac{1}{n_x} + \frac{1}{n_y} \right)^{1/2}}$$

$$df = n_x + n_y - 2$$

Satterthwaite (1946) and Welch (1997) df corrections for unequal variances

Stata Release 10 Reference Guide Q-Z (2007). StataCorp: College Station, Tx: 539.

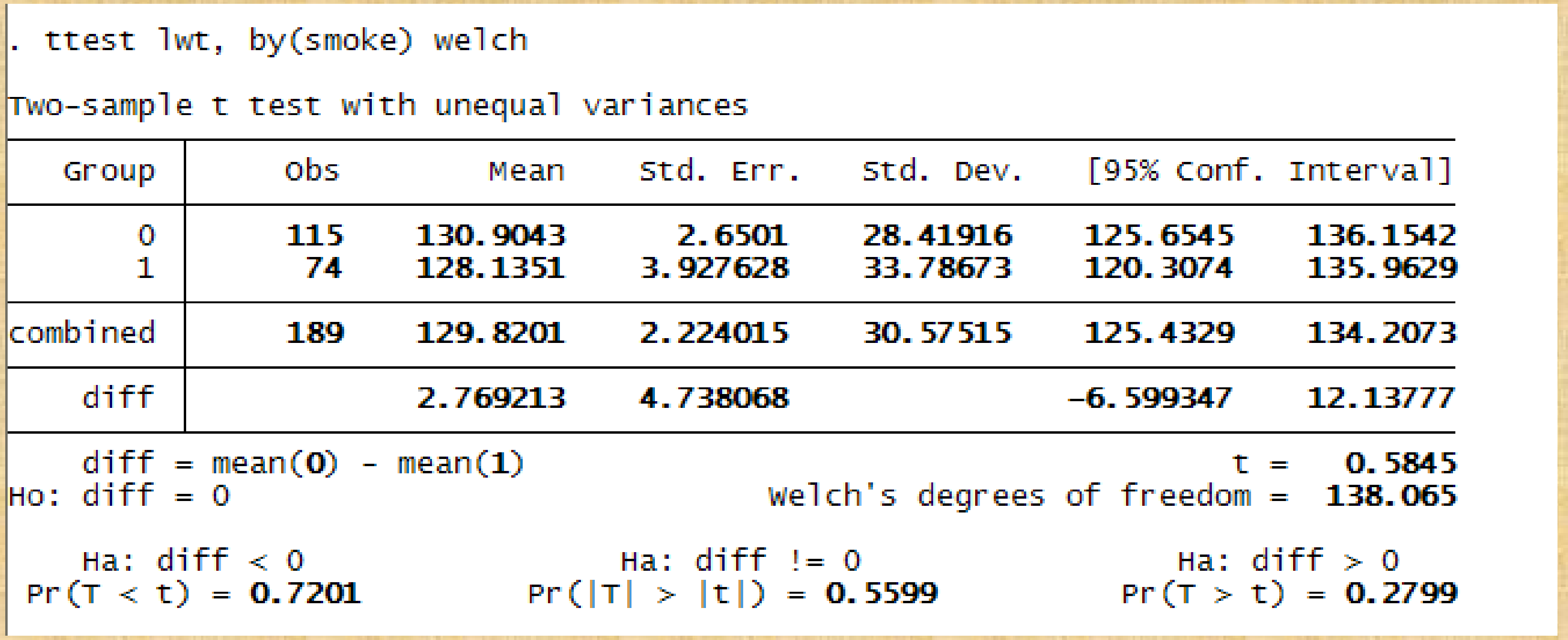
$$Satterthwaite's\ df\,(\,for\ unequal\ variances) = v$$

$$Welch's\ (\mathbf{1997})\ df\ = w$$

$$where$$

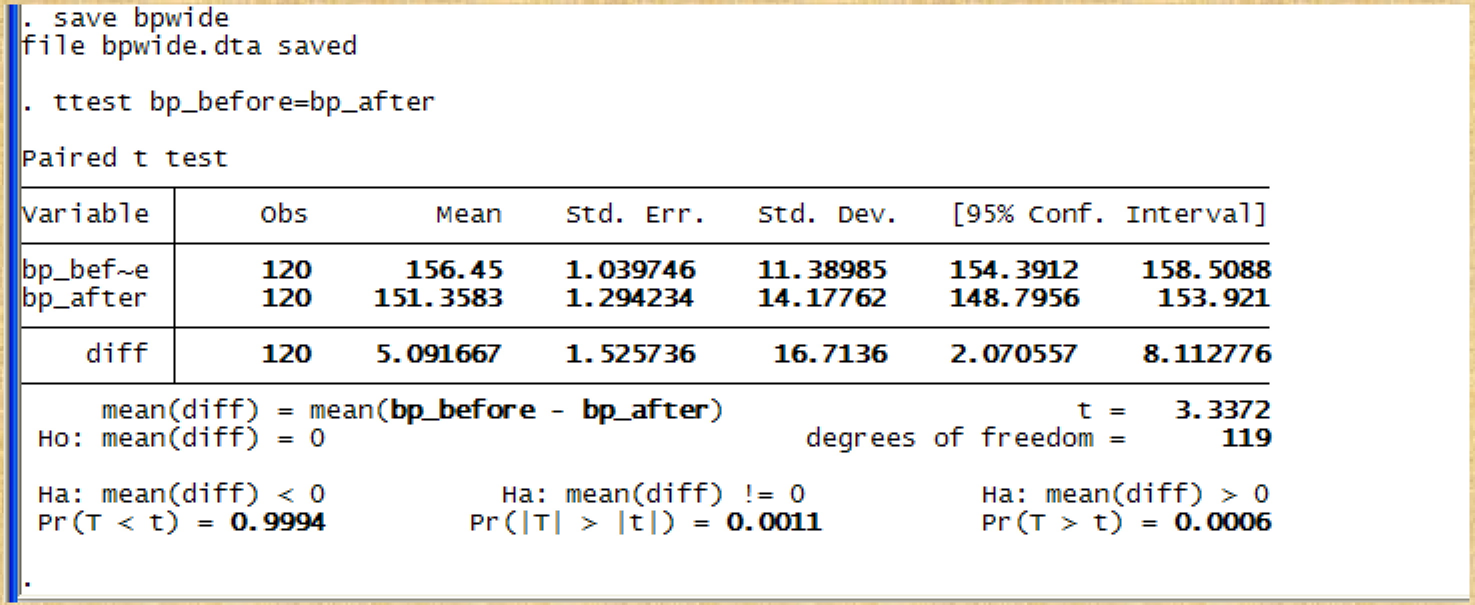
$$v = \frac{\left(\dfrac{s_x^2}{n_x} + \dfrac{s_y^2}{n_y}\right)^2}{\dfrac{\left(s_x^2\right)^2}{n_x - 1} + \dfrac{\left(s_y^2\right)^2}{n_y - 1}} \qquad w = -2 + \frac{\left(\dfrac{s_x^2}{n_x} + \dfrac{s_y^2}{n_y}\right)^2}{\dfrac{\left(s_x^2\right)^2}{n_x + 1} + \dfrac{\left(s_y^2\right)^2}{n_y + 1}}$$

Welch’s correction for unequal variances



```
. ttest lwt, by(smoke) welch

Two-sample t test with unequal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       0 |     115    130.9043      2.6501    28.41916    125.6545    136.1542
       1 |      74    128.1351    3.927628    33.78673    120.3074    135.9629
---------+--------------------------------------------------------------------
combined |     189    129.8201    2.224015    30.57515    125.4329    134.2073
---------+--------------------------------------------------------------------
    diff |            2.769213    4.738068               -6.599347    12.13777
------------------------------------------------------------------------------
    diff = mean(0) - mean(1)                                      t =   0.5845
Ho: diff = 0                             Welch's degrees of freedom =  138.065

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.7201         Pr(|T| > |t|) = 0.5599          Pr(T > t) = 0.2799
```

Paired t-test



```
. save bpwide
file bpwide.dta saved

. ttest bp_before=bp_after

Paired t test
------------------------------------------------------------------------------
Variable |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
bp_bef~e |     120      156.45    1.039746    11.38985    154.3912    158.5088
bp_after |     120    151.3583    1.294234    14.17762    148.7956     153.921
---------+--------------------------------------------------------------------
    diff |     120    5.091667    1.525736     16.7136    2.070557    8.112776
------------------------------------------------------------------------------
     mean(diff) = mean(bp_before - bp_after)                      t =   3.3372
 Ho: mean(diff) = 0                              degrees of freedom =      119

 Ha: mean(diff) < 0           Ha: mean(diff) != 0           Ha: mean(diff) > 0
 Pr(T < t) = 0.9994         Pr(|T| > |t|) = 0.0011          Pr(T > t) = 0.0006

.
```

$$paired\ t = \frac{\overline{x}_{2i} - \overline{x}_{1i}}{\frac{sd}{\sqrt{n}}} \qquad df = n - 1$$

ANOVAs for comparisons of more than two means

- Anova
  - Main effects
    - Fixed
    - Random
    - Mixed
  - Interactions
    - Proper specification
    - Plots
    - Tests of
  - Repeated measures models
- Anova postestimation
  - Contrasts
  - Post-hoc tests with multiple comparison adjustments
  - Assumptions
    - Linearity
    - iid observations
    - Residual diagnostics
      - Homogeneity tests
      - Normality tests
      - Outlier detection(developed by R. A. Fisher)

R. A. Fisher

Sir Ronald Aylmer Fisher (1890-1962)

Decomposition of Sum of Squares

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\bar{y}_i - \bar{\bar{y}})^2 + \sum_{i=1}^{n}(y_i - \bar{y}_i)^2$$

by their respective df

$$dft = n - 1 \qquad dfm = k - 1 \qquad dfe = n - k$$

gives

$$MStotal \quad = \quad MS\, betweenGroups + \; MS\; WithinGroups(error)$$

$$\sum_{i=1}^{n}\frac{(y_i - \bar{\bar{y}})^2}{n-1} = \sum_{i=1}^{n}\frac{(\bar{y}_i - \bar{\bar{y}})^2}{k-1} \qquad + \qquad \sum_{i=1}^{n}\frac{(y_i - \bar{y}_i)^2}{n-k}$$

Total variance = Model variance + error variance

The {Omnibus} F test (named after R.A. Fisher)

$$F(mdf, edf) = \frac{Anova\ model\ variance}{error\ variance}$$

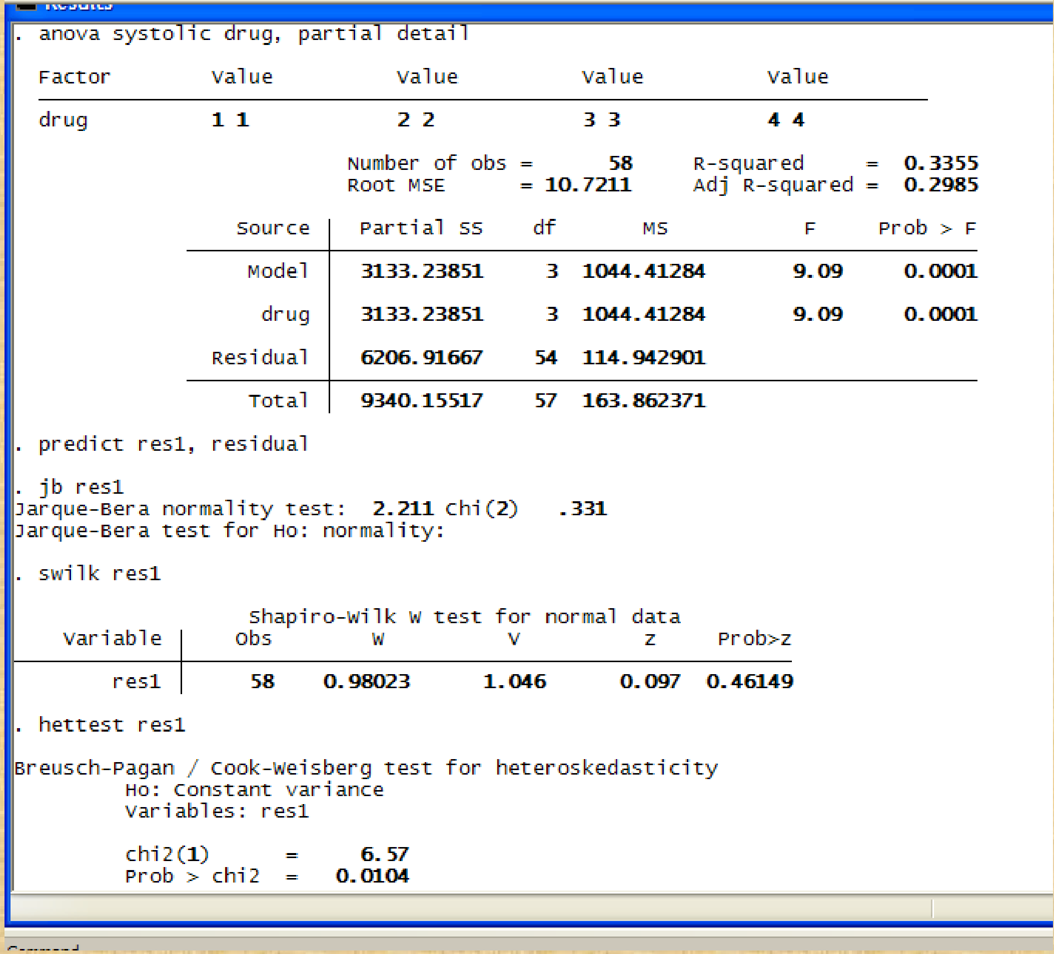
$$F(k, n-k) = \frac{\dfrac{R^2}{k-\mathbf{1}}}{\dfrac{\mathbf{1}-R^2}{n-k}}$$

Interaction terms

Sum of squares	df	Variance
SSx	# x levels – 1	SSx/(xlev-1)
SSy	# y levels – 1	SSy /(ylev-1)
SSx * SSy	(x-1)(y-1)	Ssxy/(ylev-1)(xlev-1)

- Proper specification
- X Y and x*y must all be in the model

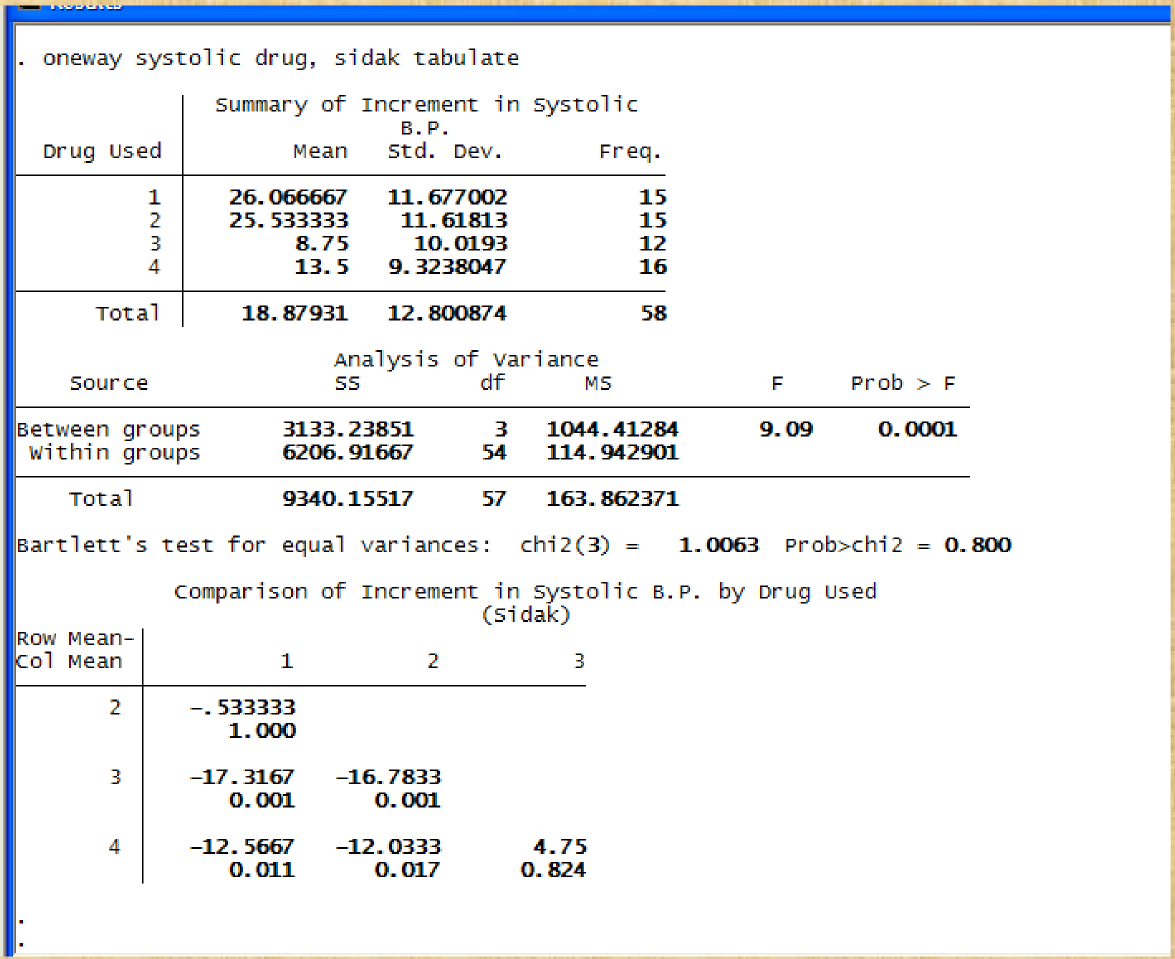
One-Way ANOVA
with residual diagnostics



```
. anova systolic drug, partial detail

  Factor             Value          Value          Value          Value
  ---------------------------------------------------------------------
  drug               1 1            2 2            3 3            4 4

                           Number of obs =      58     R-squared     =  0.3355
                           Root MSE      = 10.7211     Adj R-squared =  0.2985

                  Source |  Partial SS    df       MS           F     Prob > F
              -----------+----------------------------------------------------
                   Model |  3133.23851     3   1044.41284      9.09     0.0001
                         |
                    drug |  3133.23851     3   1044.41284      9.09     0.0001
                         |
                Residual |  6206.91667    54   114.942901
              -----------+----------------------------------------------------
                   Total |  9340.15517    57   163.862371

. predict res1, residual

. jb res1
Jarque-Bera normality test:  2.211 Chi(2)   .331
Jarque-Bera test for Ho: normality:

. swilk res1

                   Shapiro-Wilk W test for normal data
    Variable |    Obs         W           V          z      Prob>z
-------------+-----------------------------------------------------
        res1 |     58    0.98023      1.046      0.097    0.46149

. hettest res1

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: res1

         chi2(1)      =     6.57
         Prob > chi2  =   0.0104
```

Heteroscedasticity problem

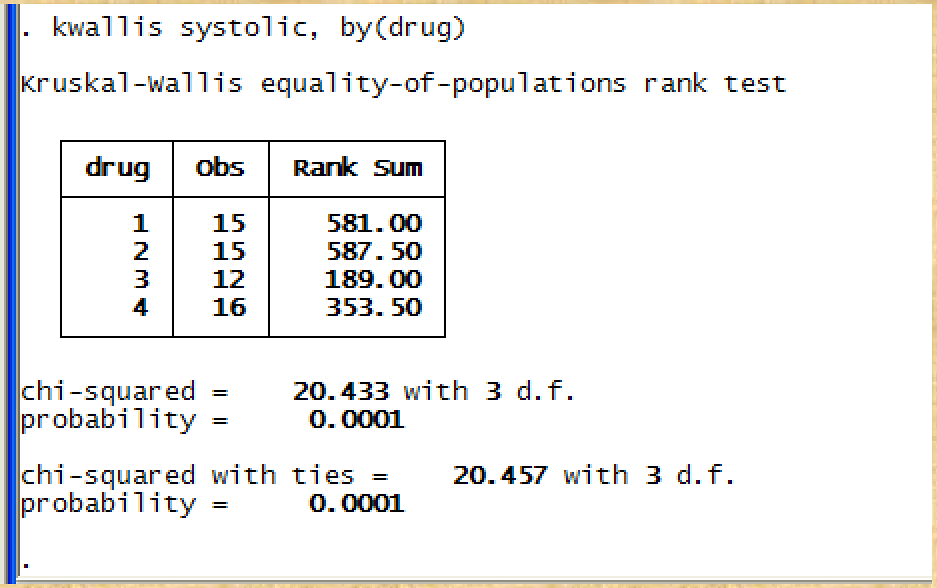
One-way ANOVA with post-hoc tests



```
. oneway systolic drug, sidak tabulate

            |  Summary of Increment in Systolic
            |                B.P.
  Drug Used |        Mean   Std. Dev.       Freq.
------------+------------------------------------
          1 |   26.066667   11.677002          15
          2 |   25.533333    11.61813          15
          3 |        8.75    10.0193           12
          4 |        13.5   9.3238047          16
------------+------------------------------------
      Total |    18.87931   12.800874          58

                        Analysis of Variance
    Source              SS         df      MS            F     Prob > F
------------------------------------------------------------------------
Between groups      3133.23851      3   1044.41284      9.09     0.0001
 Within groups      6206.91667     54   114.942901
------------------------------------------------------------------------
    Total           9340.15517     57   163.862371

Bartlett's test for equal variances:  chi2(3) =   1.0063  Prob>chi2 = 0.800

            Comparison of Increment in Systolic B.P. by Drug Used
                                (Sidak)
Row Mean-|
Col Mean |          1          2          3
---------+---------------------------------
       2 |   -.533333
         |      1.000
         |
       3 |   -17.3167   -16.7833
         |      0.001      0.001
         |
       4 |   -12.5667   -12.0333       4.75
         |      0.011      0.017      0.824

.
.
```

Nonparametric one-way alternative

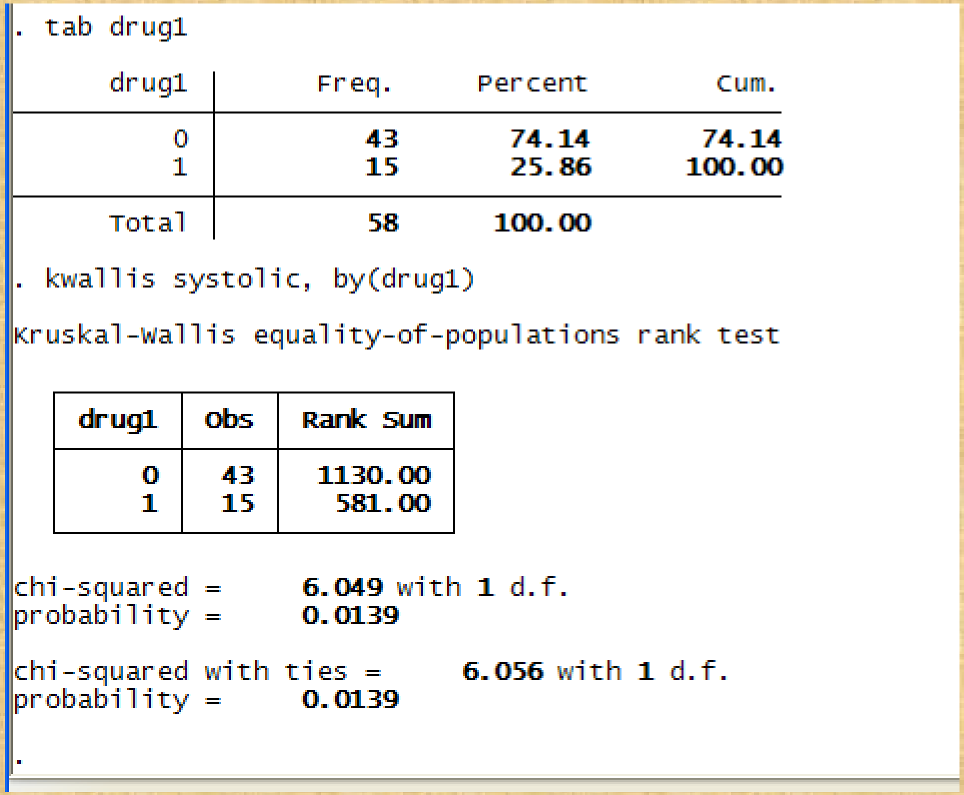
- Kruskall-Wallis one way nonparametric ANOVA



```
. kwallis systolic, by(drug)

Kruskal-Wallis equality-of-populations rank test

  +------------------------+
  | drug | Obs | Rank Sum |
  |------+-----+----------|
  |    1 |  15 |   581.00 |
  |    2 |  15 |   587.50 |
  |    3 |  12 |   189.00 |
  |    4 |  16 |   353.50 |
  +------------------------+

chi-squared =    20.433 with 3 d.f.
probability =     0.0001

chi-squared with ties =    20.457 with 3 d.f.
probability =     0.0001

.
```

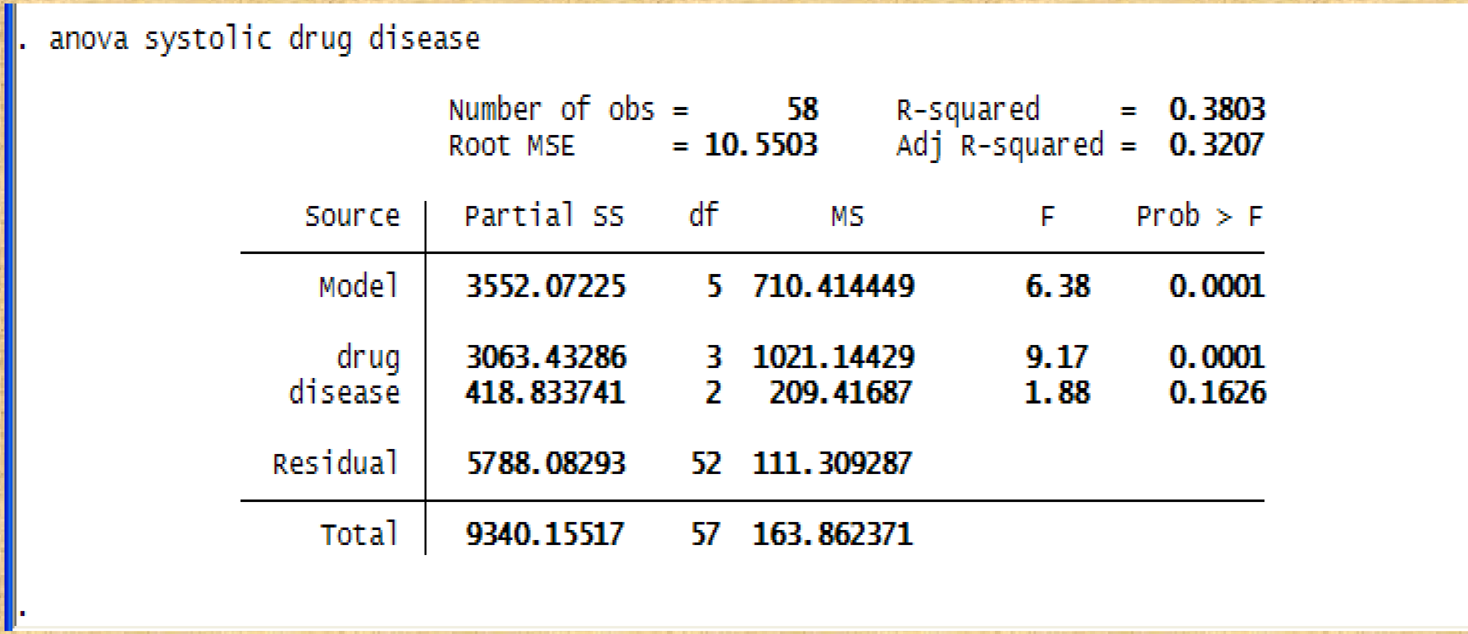
Distribution free one-way nonparametric ANOVA by rank sum

Kruskall -Wallis nonparametric multiple comparisons



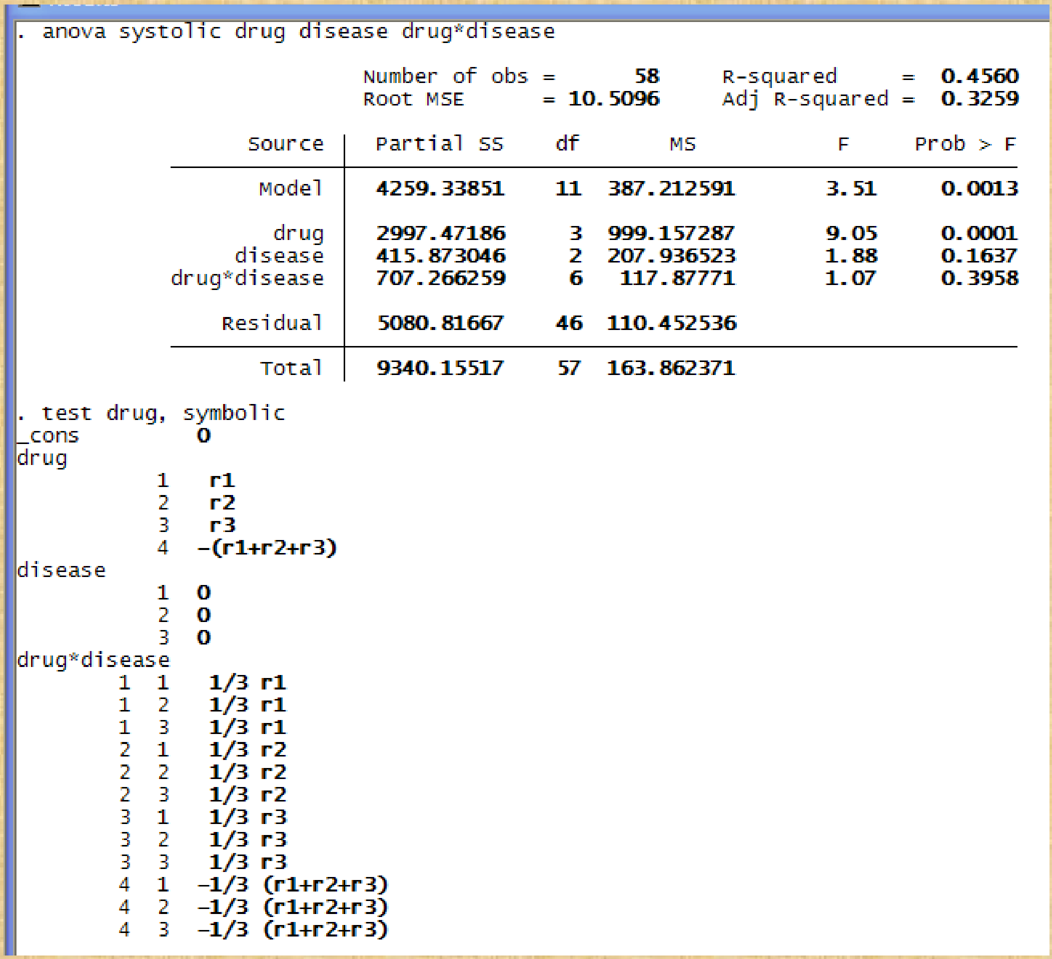
```
. tab drug1

      drug1 |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |         43       74.14       74.14
          1 |         15       25.86      100.00
------------+-----------------------------------
      Total |         58      100.00

. kwallis systolic, by(drug1)

Kruskal-Wallis equality-of-populations rank test

  +------------------------+
  | drug1 | Obs | Rank Sum |
  |-------+-----+----------|
  |     0 |  43 |  1130.00 |
  |     1 |  15 |   581.00 |
  +------------------------+

chi-squared =     6.049 with 1 d.f.
probability =     0.0139

chi-squared with ties =     6.056 with 1 d.f.
probability =     0.0139

.
```

Main effects ANOVA



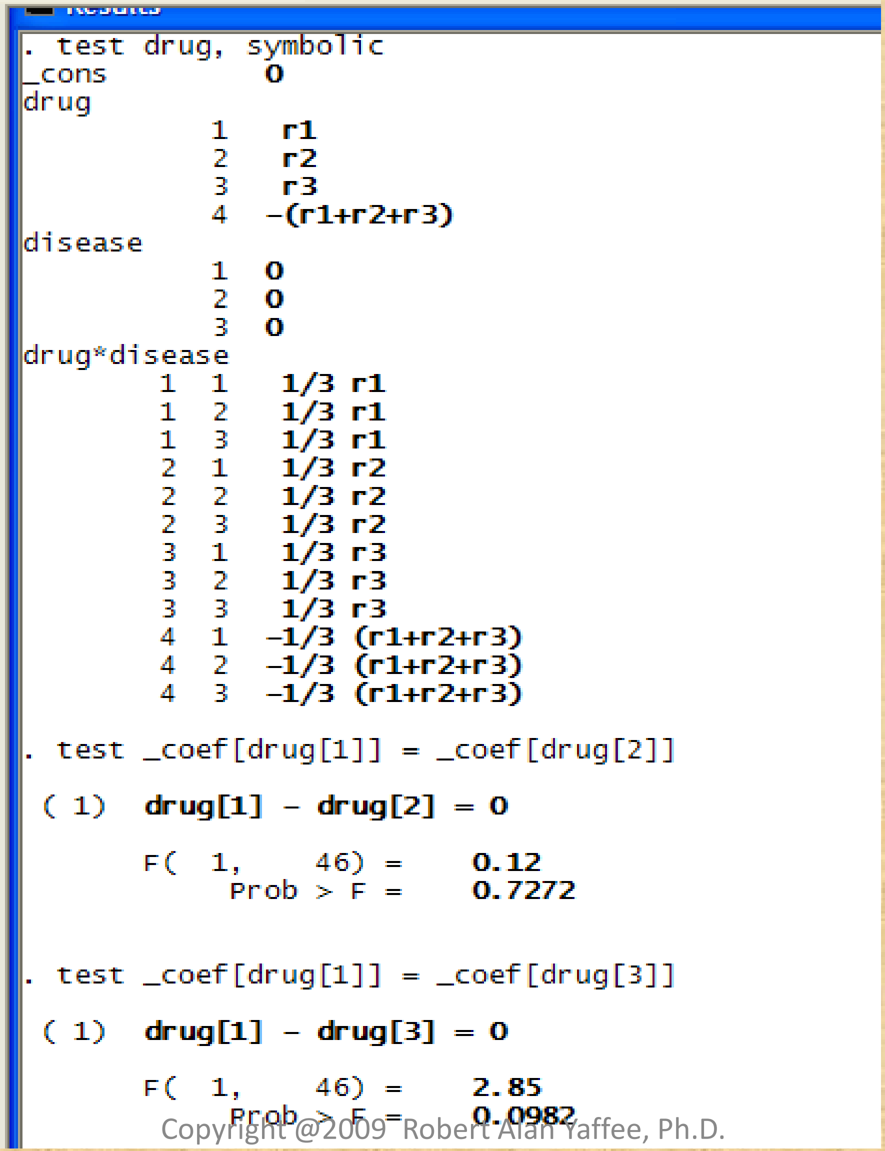
```
. anova systolic drug disease

                           Number of obs =      58     R-squared     =  0.3803
                           Root MSE      = 10.5503     Adj R-squared =  0.3207

                  Source |  Partial SS    df       MS           F     Prob > F
              -----------+----------------------------------------------------
                   Model |  3552.07225     5  710.414449       6.38     0.0001
                         |
                    drug |  3063.43286     3  1021.14429       9.17     0.0001
                 disease |  418.833741     2   209.41687       1.88     0.1626
                         |
                Residual |  5788.08293    52  111.309287
              -----------+----------------------------------------------------
                   Total |  9340.15517    57  163.862371

.
```

Factorial Anova



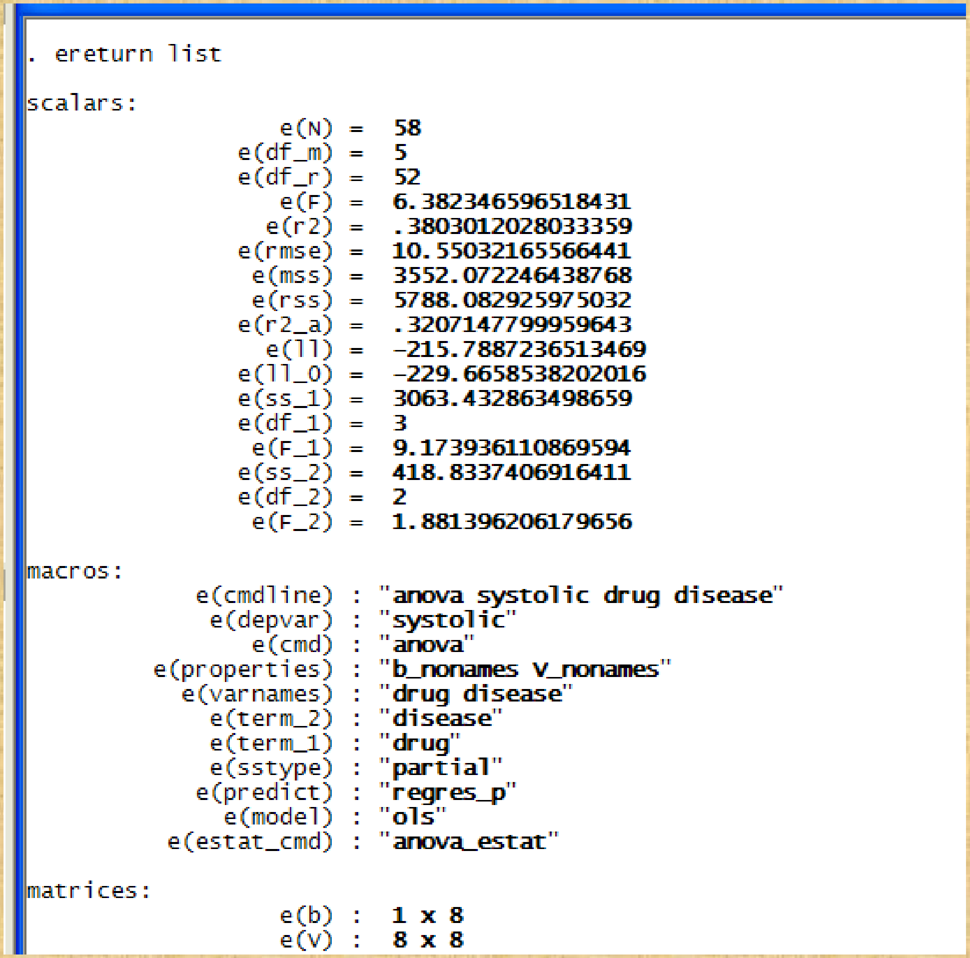
```
. anova systolic drug disease drug*disease

                           Number of obs =      58     R-squared     =  0.4560
                           Root MSE      = 10.5096     Adj R-squared =  0.3259

                  Source |  Partial SS    df       MS           F     Prob > F
              -----------+----------------------------------------------------
                   Model |  4259.33851    11  387.212591       3.51     0.0013
                         |
                    drug |  2997.47186     3  999.157287       9.05     0.0001
                 disease |  415.873046     2  207.936523       1.88     0.1637
            drug*disease |  707.266259     6   117.87771       1.07     0.3958
                         |
                Residual |  5080.81667    46  110.452536
              -----------+----------------------------------------------------
                   Total |  9340.15517    57  163.862371

. test drug, symbolic
_cons               0
drug
             1     r1
             2     r2
             3     r3
             4    -(r1+r2+r3)
disease
             1    0
             2    0
             3    0
drug*disease
          1  1     1/3 r1
          1  2     1/3 r1
          1  3     1/3 r1
          2  1     1/3 r2
          2  2     1/3 r2
          2  3     1/3 r2
          3  1     1/3 r3
          3  2     1/3 r3
          3  3     1/3 r3
          4  1    -1/3 (r1+r2+r3)
          4  2    -1/3 (r1+r2+r3)
          4  3    -1/3 (r1+r2+r3)
```

Anova Contrasts



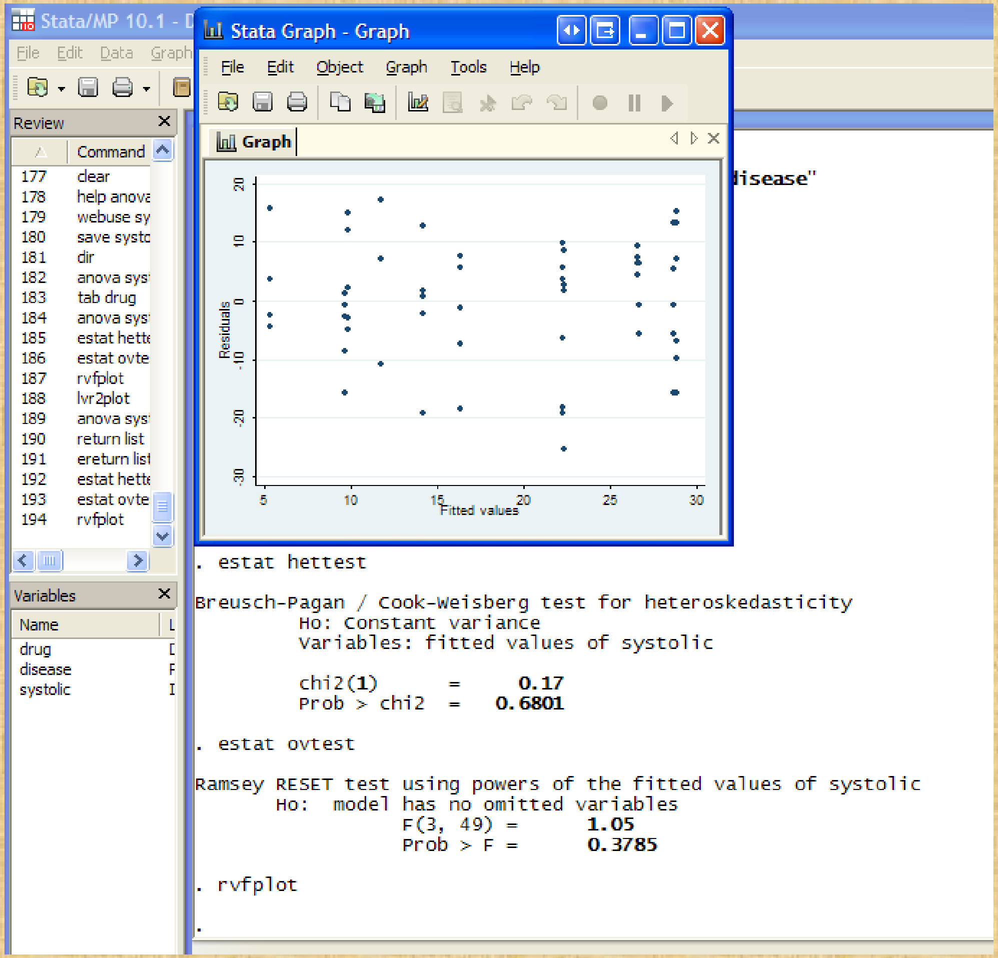
```
. test drug, symbolic
_cons          0
drug
          1    r1
          2    r2
          3    r3
          4   -(r1+r2+r3)
disease
          1   0
          2   0
          3   0
drug*disease
       1  1    1/3 r1
       1  2    1/3 r1
       1  3    1/3 r1
       2  1    1/3 r2
       2  2    1/3 r2
       2  3    1/3 r2
       3  1    1/3 r3
       3  2    1/3 r3
       3  3    1/3 r3
       4  1   -1/3 (r1+r2+r3)
       4  2   -1/3 (r1+r2+r3)
       4  3   -1/3 (r1+r2+r3)

. test _coef[drug[1]] = _coef[drug[2]]

 ( 1)  drug[1] - drug[2] = 0

       F(  1,    46) =    0.12
            Prob > F =    0.7272


. test _coef[drug[1]] = _coef[drug[3]]

 ( 1)  drug[1] - drug[3] = 0

       F(  1,    46) =    2.85
            Prob > F =    0.0982
```

Arguments to pass on



```
. ereturn list

scalars:
                  e(N) =  58
               e(df_m) =  5
               e(df_r) =  52
                  e(F) =  6.382346596518431
                 e(r2) =  .3803012028033359
               e(rmse) =  10.55032165566441
                e(mss) =  3552.072246438768
                e(rss) =  5788.082925975032
               e(r2_a) =  .3207147799959643
                 e(ll) =  -215.7887236513469
               e(ll_0) =  -229.6658538202016
               e(ss_1) =  3063.432863498659
               e(df_1) =  3
                e(F_1) =  9.173936110869594
               e(ss_2) =  418.8337406916411
               e(df_2) =  2
                e(F_2) =  1.881396206179656

macros:
            e(cmdline) : "anova systolic drug disease"
             e(depvar) : "systolic"
                e(cmd) : "anova"
         e(properties) : "b_nonames V_nonames"
           e(varnames) : "drug disease"
             e(term_2) : "disease"
             e(term_1) : "drug"
             e(sstype) : "partial"
            e(predict) : "regres_p"
              e(model) : "ols"
          e(estat_cmd) : "anova_estat"

matrices:
                  e(b) :  1 x 8
                  e(V) :  8 x 8
```

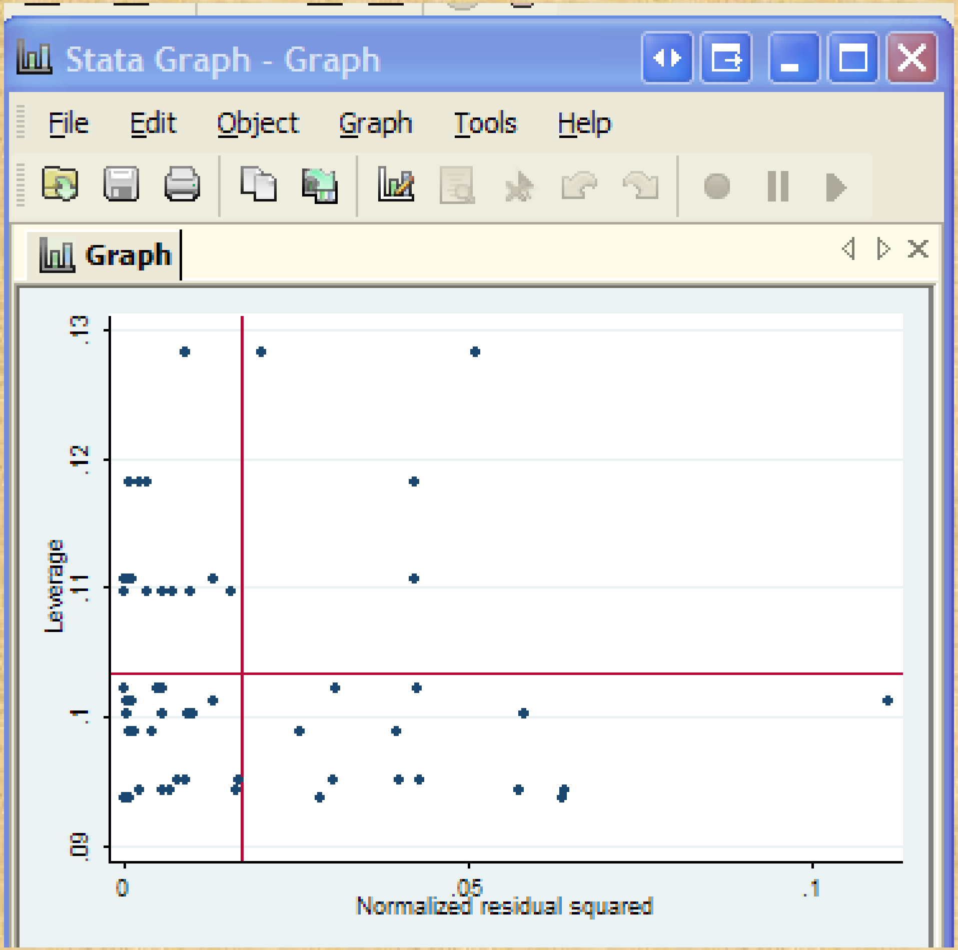
ANOVA Postestimation



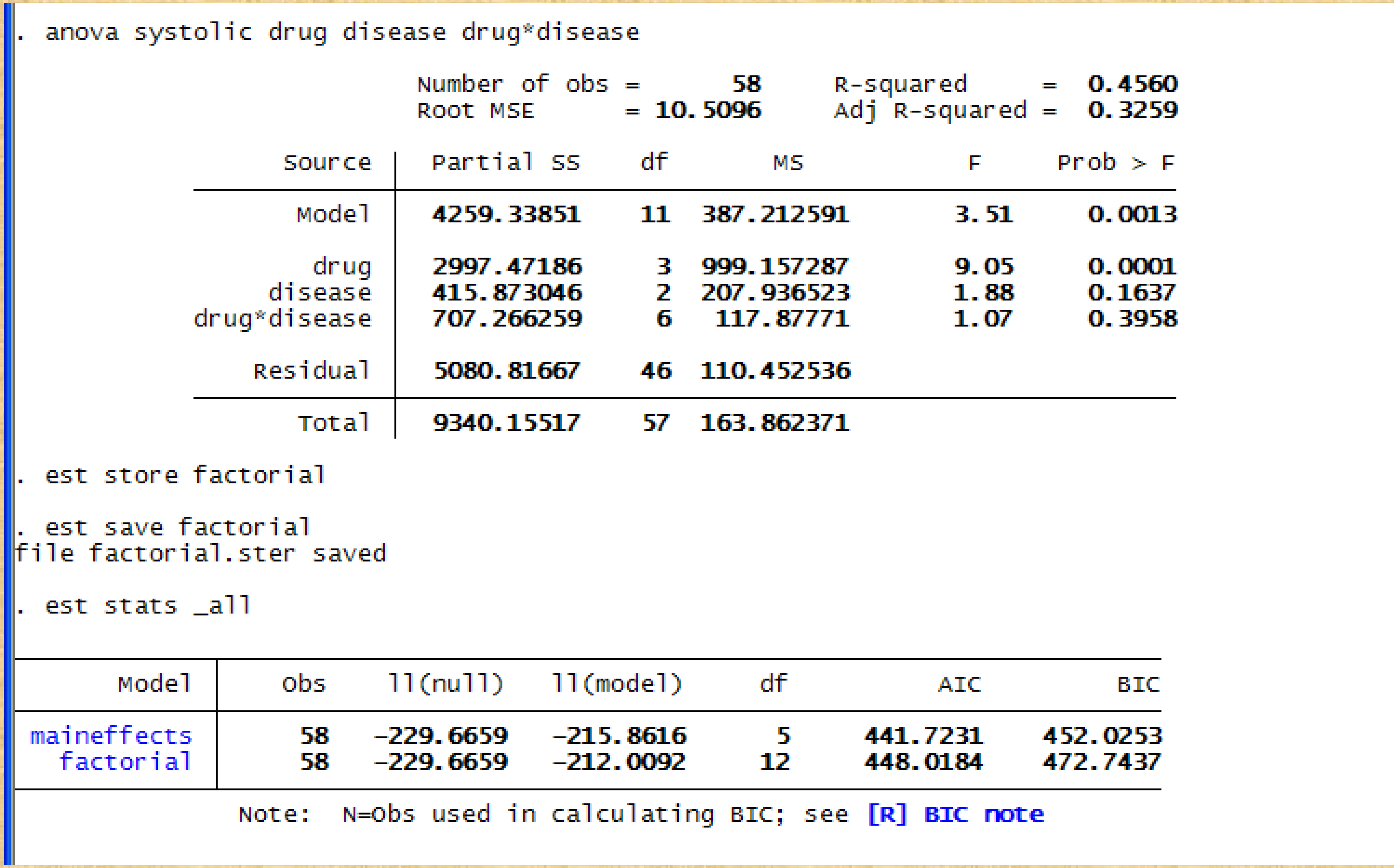
	Command
177	clear
178	help anova
179	webuse sy
180	save systo
181	dir
182	anova sys
183	tab drug
184	anova sys
185	estat hette
186	estat ovte
187	rvfplot
188	lvr2plot
189	anova sys
190	return list
191	ereturn list
192	estat hette
193	estat ovte
194	rvfplot

```
isease"

. estat hettest

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: fitted values of systolic

         chi2(1)      =     0.17
         Prob > chi2  =   0.6801

. estat ovtest

Ramsey RESET test using powers of the fitted values of systolic
       Ho:  model has no omitted variables
                 F(3, 49) =      1.05
                 Prob > F =      0.3785

. rvfplot

.
```

lvr2plot

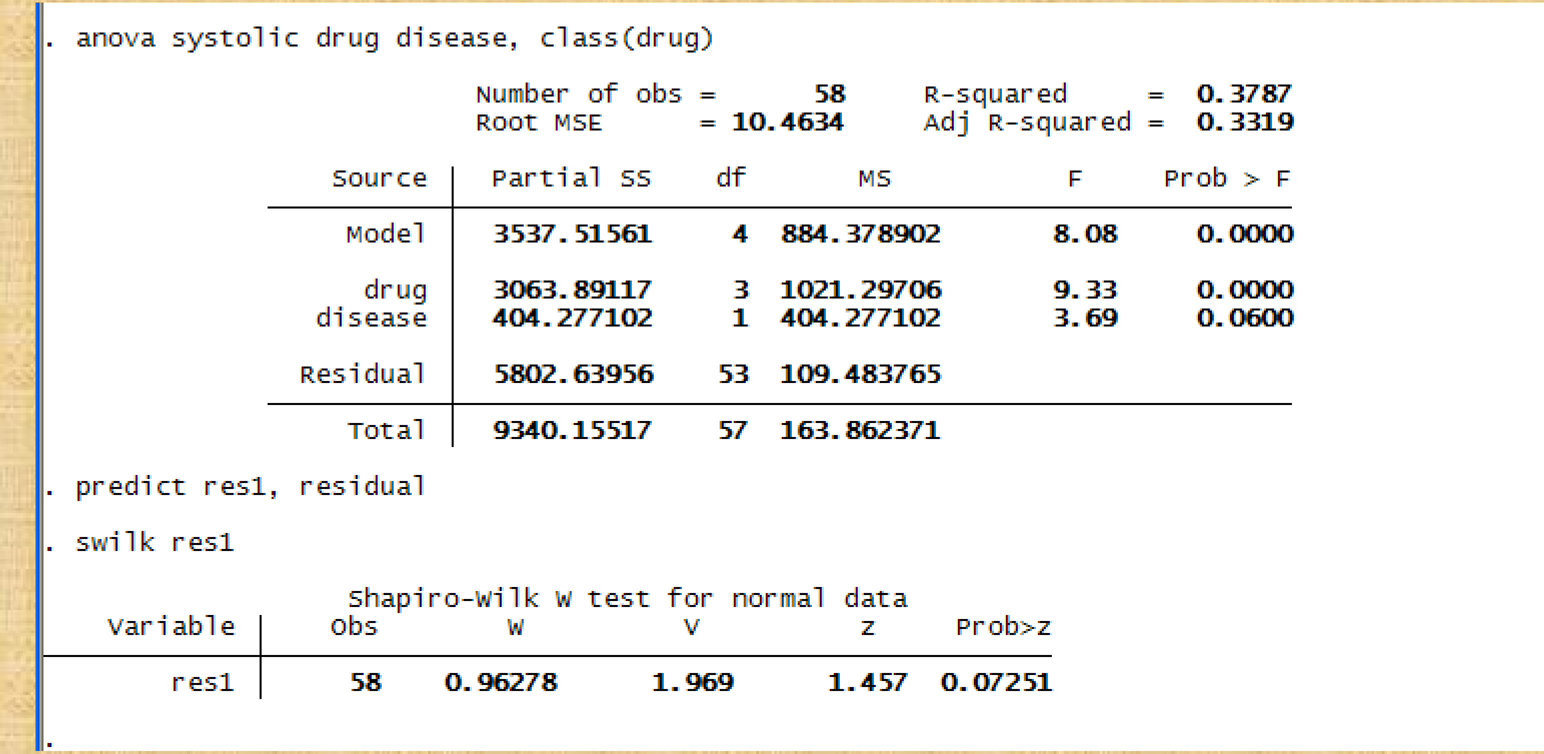

Stata Graph - Graph
File
Edit
Object
Graph
Tools
Help
Graph
Leverage
.09
.1
.11
.12
.13
0
.05
.1
Normalized residual squared

Full-factorial model and model comparison



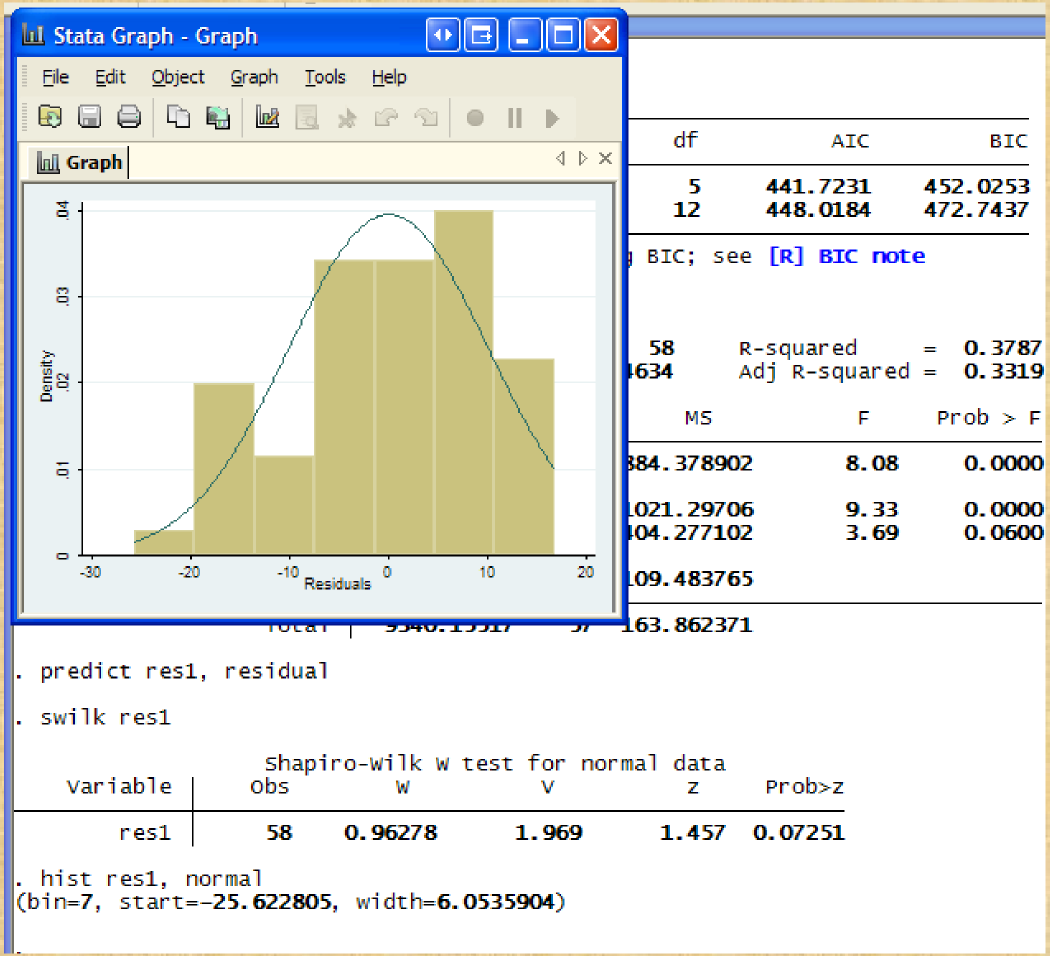
```
. anova systolic drug disease drug*disease

                           Number of obs =      58     R-squared     =  0.4560
                           Root MSE      = 10.5096     Adj R-squared =  0.3259

                  Source |  Partial SS    df       MS           F     Prob > F
              -----------+----------------------------------------------------
                   Model |  4259.33851    11   387.212591      3.51     0.0013
                         |
                    drug |  2997.47186     3   999.157287      9.05     0.0001
                 disease |  415.873046     2   207.936523      1.88     0.1637
            drug*disease |  707.266259     6    117.87771      1.07     0.3958
                         |
                Residual |  5080.81667    46   110.452536
              -----------+----------------------------------------------------
                   Total |  9340.15517    57   163.862371

. est store factorial

. est save factorial
file factorial.ster saved

. est stats _all

-----------------------------------------------------------------------------
       Model |    Obs    ll(null)   ll(model)     df          AIC         BIC
-------------+---------------------------------------------------------------
 maineffects |     58   -229.6659   -215.8616      5     441.7231    452.0253
   factorial |     58   -229.6659   -212.0092     12     448.0184    472.7437
-----------------------------------------------------------------------------
               Note:  N=Obs used in calculating BIC; see [R] BIC note
```

Final model with residual normality diagnosis



```
. anova systolic drug disease, class(drug)

                           Number of obs =      58     R-squared     =  0.3787
                           Root MSE      = 10.4634     Adj R-squared =  0.3319

                  Source |  Partial SS    df       MS           F     Prob > F
              -----------+----------------------------------------------------
                   Model |  3537.51561     4   884.378902      8.08     0.0000
                         |
                    drug |  3063.89117     3   1021.29706      9.33     0.0000
                 disease |  404.277102     1   404.277102      3.69     0.0600
                         |
                Residual |  5802.63956    53   109.483765
              -----------+----------------------------------------------------
                   Total |  9340.15517    57   163.862371

. predict res1, residual

. swilk res1

                   Shapiro-Wilk W test for normal data
    Variable |    Obs       W           V         z       Prob>z
-------------+--------------------------------------------------
        res1 |     58    0.96278      1.969     1.457    0.07251

.
```

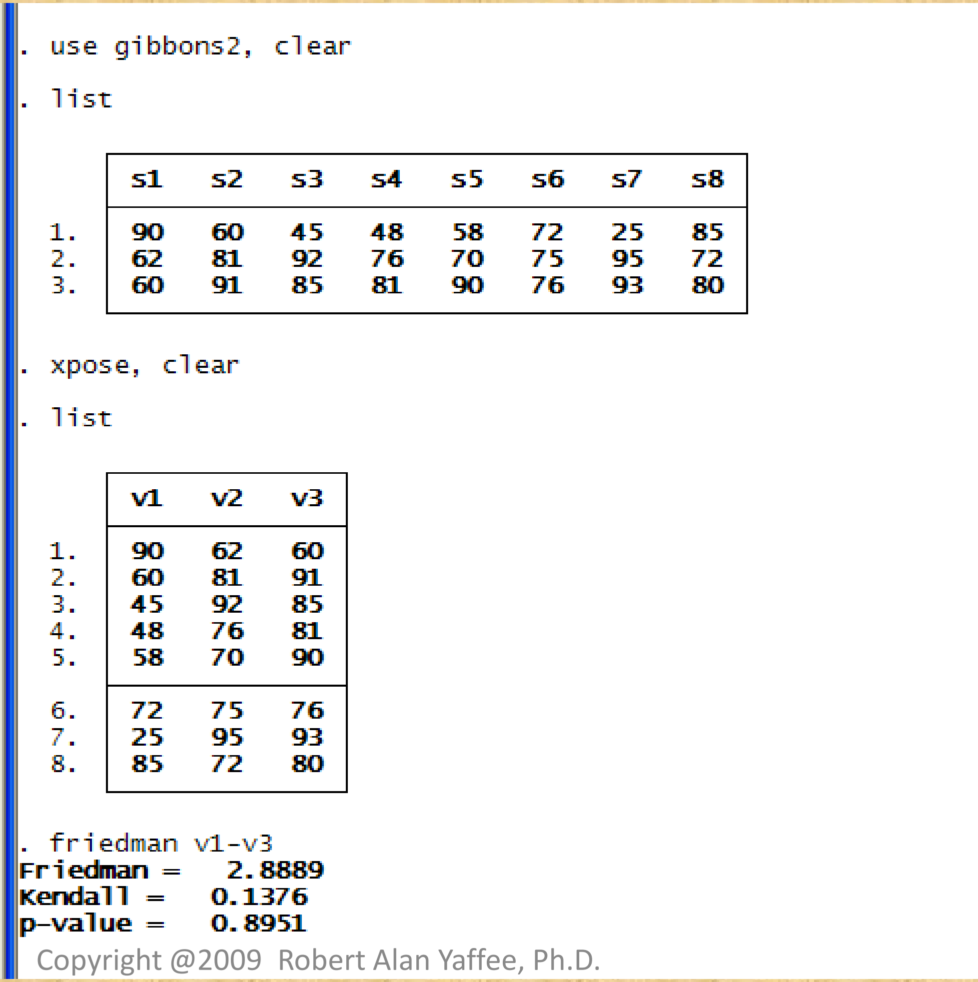
Graphical review of residuals



```
     df           AIC          BIC

      5     441.7231     452.0253
     12     448.0184     472.7437

 BIC; see [R] BIC note


     58      R-squared      =   0.3787
    634      Adj R-squared =   0.3319

      MS                F     Prob > F

 84.378902           8.08      0.0000

 021.29706           9.33      0.0000
 04.277102           3.69      0.0600

 09.483765

 163.862371

. predict res1, residual

. swilk res1

                   Shapiro-Wilk W test for normal data
    Variable |       Obs          W           V          z     Prob>z
-------------+-------------------------------------------------
        res1 |        58    0.96278       1.969      1.457   0.07251

. hist res1, normal
(bin=7, start=-25.622805, width=6.0535904)
```

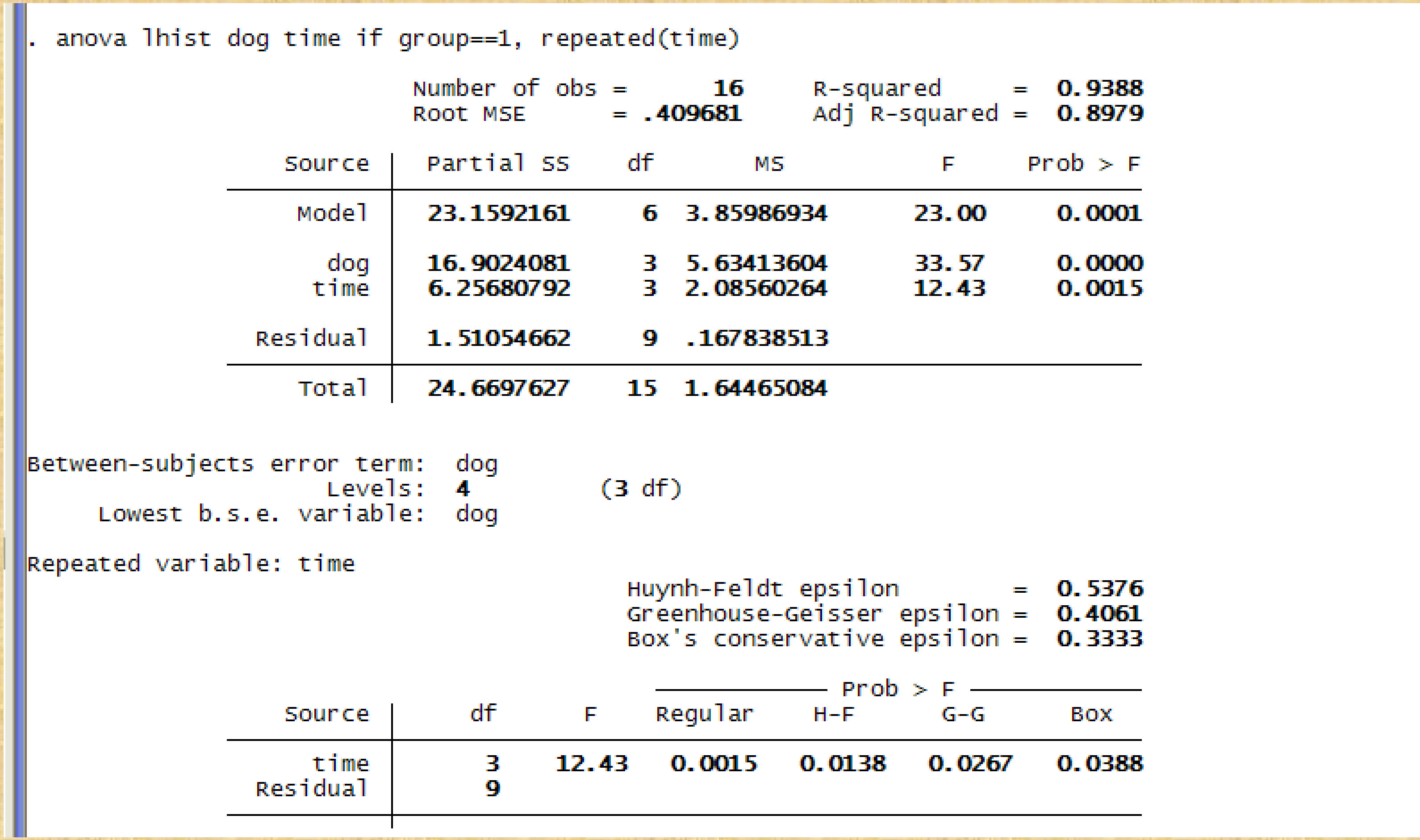
Nonparametric Friedman 2-way ANOVA written in Stata by Richard Goldstein

- Type: findit friedman
- Install snp-1



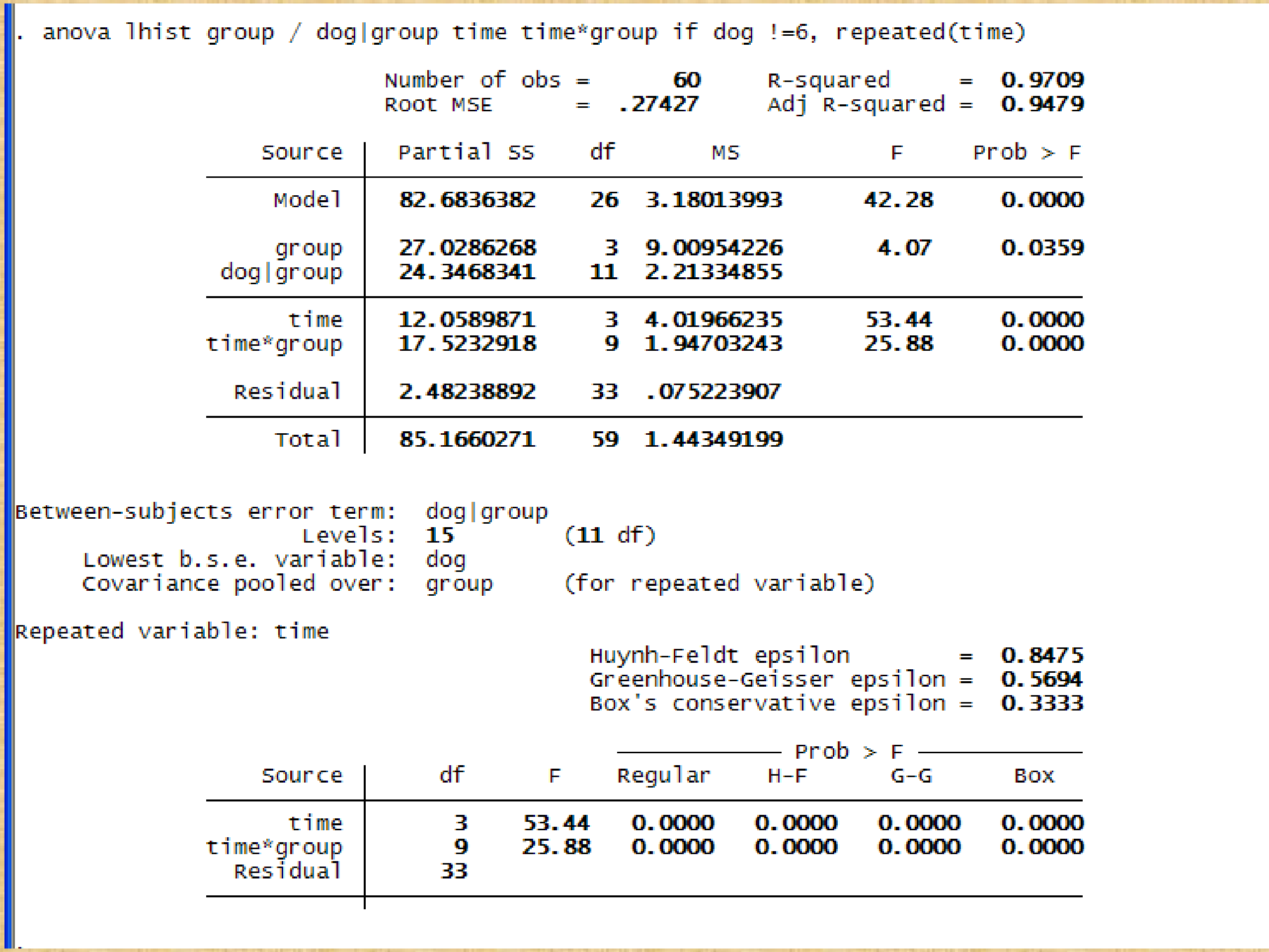
```
. use gibbons2, clear

. list

     +---------------------------------------+
     | s1   s2   s3   s4   s5   s6   s7   s8 |
     |---------------------------------------|
  1. | 90   60   45   48   58   72   25   85 |
  2. | 62   81   92   76   70   75   95   72 |
  3. | 60   91   85   81   90   76   93   80 |
     +---------------------------------------+

. xpose, clear

. list

     +--------------+
     | v1   v2   v3 |
     |--------------|
  1. | 90   62   60 |
  2. | 60   81   91 |
  3. | 45   92   85 |
  4. | 48   76   81 |
  5. | 58   70   90 |
     |--------------|
  6. | 72   75   76 |
  7. | 25   95   93 |
  8. | 85   72   80 |
     +--------------+

. friedman v1-v3
Friedman =    2.8889
Kendall =    0.1376
p-value =    0.8951
```

Repeated measures ANOVAs



```
. anova lhist dog time if group==1, repeated(time)

                           Number of obs =      16     R-squared     =  0.9388
                           Root MSE      = .409681     Adj R-squared =  0.8979

                  Source |  Partial SS    df       MS           F     Prob > F
              -----------+----------------------------------------------------
                   Model |  23.1592161     6  3.85986934      23.00     0.0001
                         |
                     dog |  16.9024081     3  5.63413604      33.57     0.0000
                    time |  6.25680792     3  2.08560264      12.43     0.0015
                         |
                Residual |  1.51054662     9  .167838513
              -----------+----------------------------------------------------
                   Total |  24.6697627    15  1.64465084


Between-subjects error term:  dog
                     Levels:  4        (3 df)
     Lowest b.s.e. variable:  dog

Repeated variable: time
                                          Huynh-Feldt epsilon        =  0.5376
                                          Greenhouse-Geisser epsilon =  0.4061
                                          Box's conservative epsilon =  0.3333

                                            ------------ Prob > F -----------
                  Source |     df      F    Regular    H-F      G-G      Box
              -----------+----------------------------------------------------
                    time |      3    12.43   0.0015   0.0138   0.0267   0.0388
                Residual |      9
              -----------+----------------------------------------------------
```

Repeated measures ANOVAs



```
. anova lhist group / dog|group time time*group if dog !=6, repeated(time)

                           Number of obs =      60     R-squared     =  0.9709
                           Root MSE      =  .27427     Adj R-squared =  0.9479

                  Source |  Partial SS    df       MS           F     Prob > F
              -----------+----------------------------------------------------
                   Model |  82.6836382    26  3.18013993      42.28     0.0000
                         |
                   group |  27.0286268     3  9.00954226       4.07     0.0359
               dog|group |  24.3468341    11  2.21334855
              -----------+----------------------------------------------------
                    time |  12.0589871     3  4.01966235      53.44     0.0000
              time*group |  17.5232918     9  1.94703243      25.88     0.0000
                         |
                Residual |  2.48238892    33  .075223907
              -----------+----------------------------------------------------
                   Total |  85.1660271    59  1.44349199


Between-subjects error term:  dog|group
                     Levels:  15         (11 df)
     Lowest b.s.e. variable:  dog
     Covariance pooled over:  group      (for repeated variable)

Repeated variable: time
                                          Huynh-Feldt epsilon        =  0.8475
                                          Greenhouse-Geisser epsilon =  0.5694
                                          Box's conservative epsilon =  0.3333

                                            ------------ Prob > F ------------
                  Source |     df      F    Regular    H-F      G-G      Box
              -----------+----------------------------------------------------
                    time |      3    53.44   0.0000   0.0000   0.0000   0.0000
              time*group |      9    25.88   0.0000   0.0000   0.0000   0.0000
                Residual |     33
              -----------+----------------------------------------------------

.
```

Fixed, random, and mixed effects models

- Fixed effects are clearly specified with all levels being sampled-e.g., gender.
- Random effects are those which are supposedly randomly sampled with only some of the levels included in the study: e.g., subjects.
- Mixed effects models have both fixed and random effects in the model.

The error variance for such effects differ and therefore must be clearly identified.

- F tests have to be properly constructed with these different effects.

Expected mean squares (Variances)

Expected Mean Squares for Different Designs

Source: Michaels, Brown, & Winer, 1993, *Statistical Principles of Experimental Design, 304*

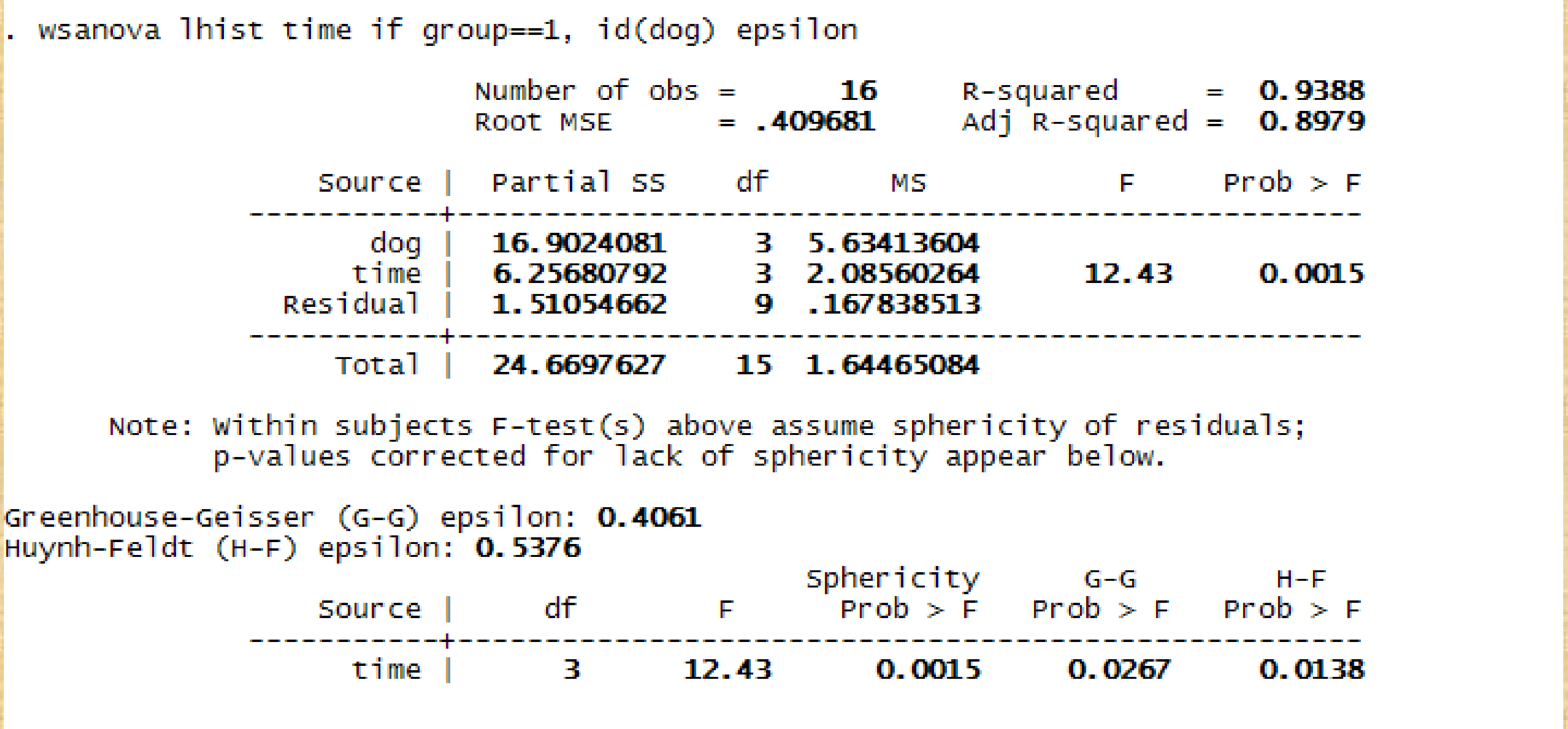
	Case 1: Fixed **a fixed, b fixed**	**Case 2: Mixed** **a fixed b random**	**Case 3: Random** **a random b random**
MS_a	$\sigma_e^2 + nq\sigma_a^2$	$\sigma_e^2 + n\sigma_{ab}^2 + nq\sigma_a^2$	$\sigma_e^2 + n\sigma_{ab}^2 + nq\sigma_a^2$
Ms_b	$\sigma_e^2 + nq\sigma_b^2$	$\sigma_e^2 + np\sigma_b^2$	$\sigma_e^2 + n\sigma_{ab}^2 + np\sigma_b^2$
MS_{ab}	$\sigma_e^2 + n\sigma_{ab}^2$	$\sigma_e^2 + n\sigma_{ab}^2$	$\sigma_e^2 + n\sigma_{ab}^2$
MS_{error}	σ_e^2	σ_e^2	σ_e^2

F test for fixed effect = MS_a/MS_{error}

F test for random effect = MS_b/MS_{error}

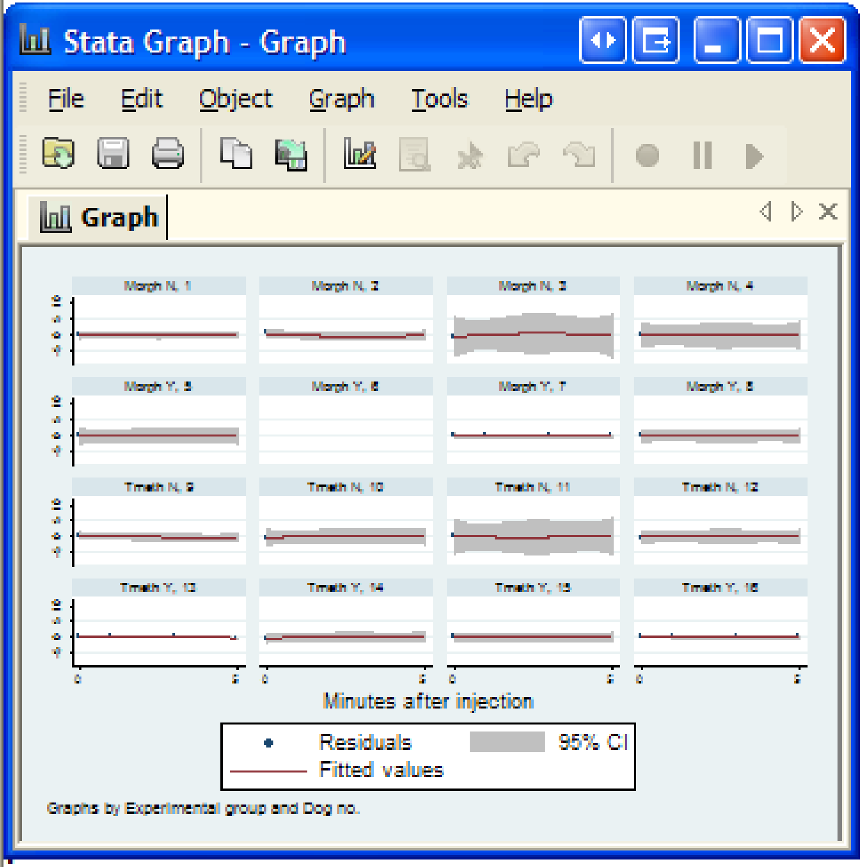
F-test for mixed effect : fixed $=MS_a/Ms_{ab}$ Random$=MS_b/MS_{error}$

Repeated measures with wsanova



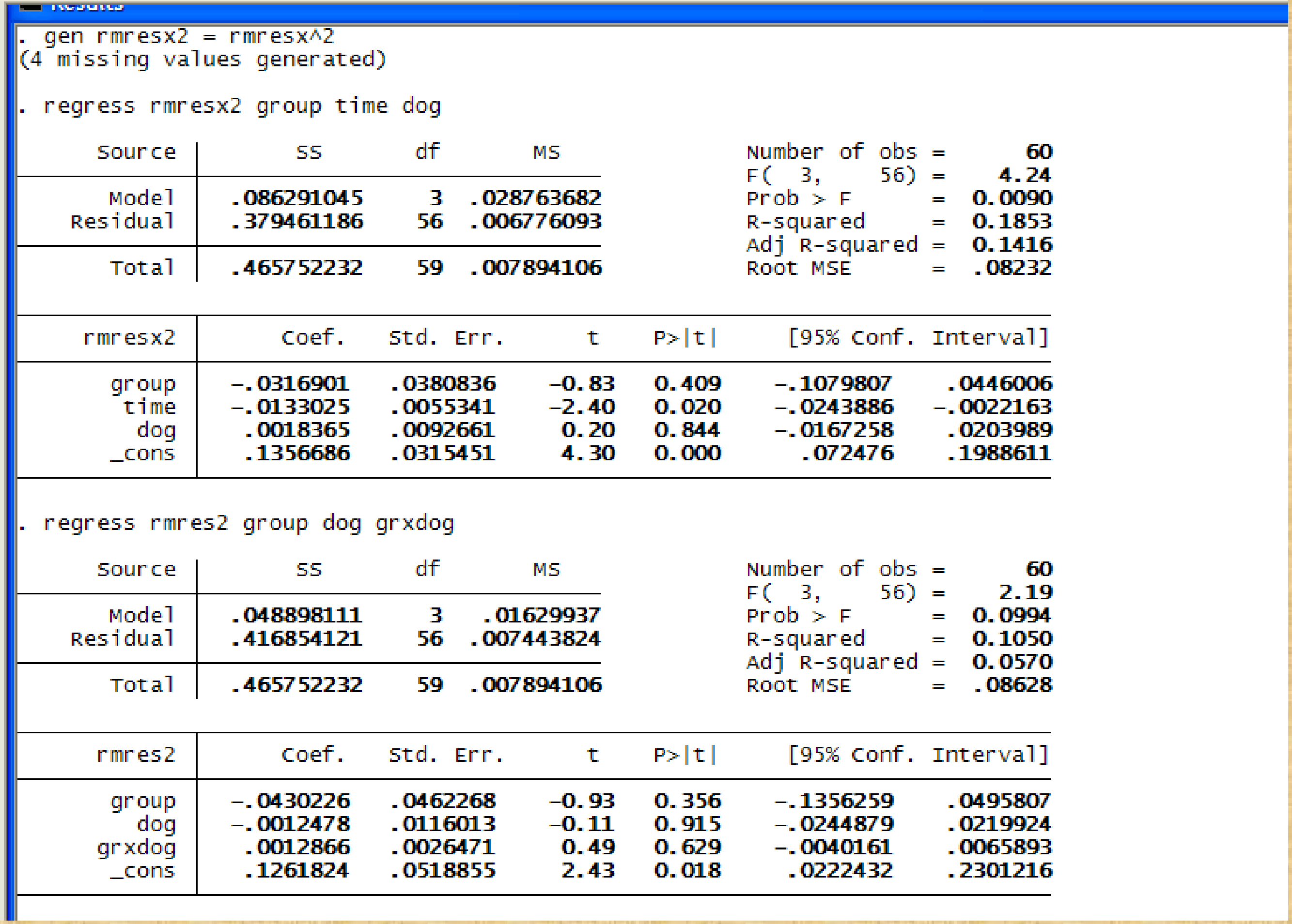
```
. wsanova lhist time if group==1, id(dog) epsilon

                           Number of obs =      16     R-squared     =  0.9388
                           Root MSE      = .409681     Adj R-squared =  0.8979

                  Source |  Partial SS    df       MS           F     Prob > F
              -----------+----------------------------------------------------
                     dog |  16.9024081     3  5.63413604
                    time |  6.25680792     3  2.08560264      12.43     0.0015
                Residual |  1.51054662     9  .167838513
              -----------+----------------------------------------------------
                   Total |  24.6697627    15  1.64465084

      Note: Within subjects F-test(s) above assume sphericity of residuals;
            p-values corrected for lack of sphericity appear below.

Greenhouse-Geisser (G-G) epsilon: 0.4061
Huynh-Feldt (H-F) epsilon: 0.5376
                                              Sphericity     G-G         H-F
                  Source |     df          F     Prob > F   Prob > F    Prob > F
              -----------+----------------------------------------------------
                    time |      3      12.43       0.0015     0.0267      0.0138
```

Residual diagnostics



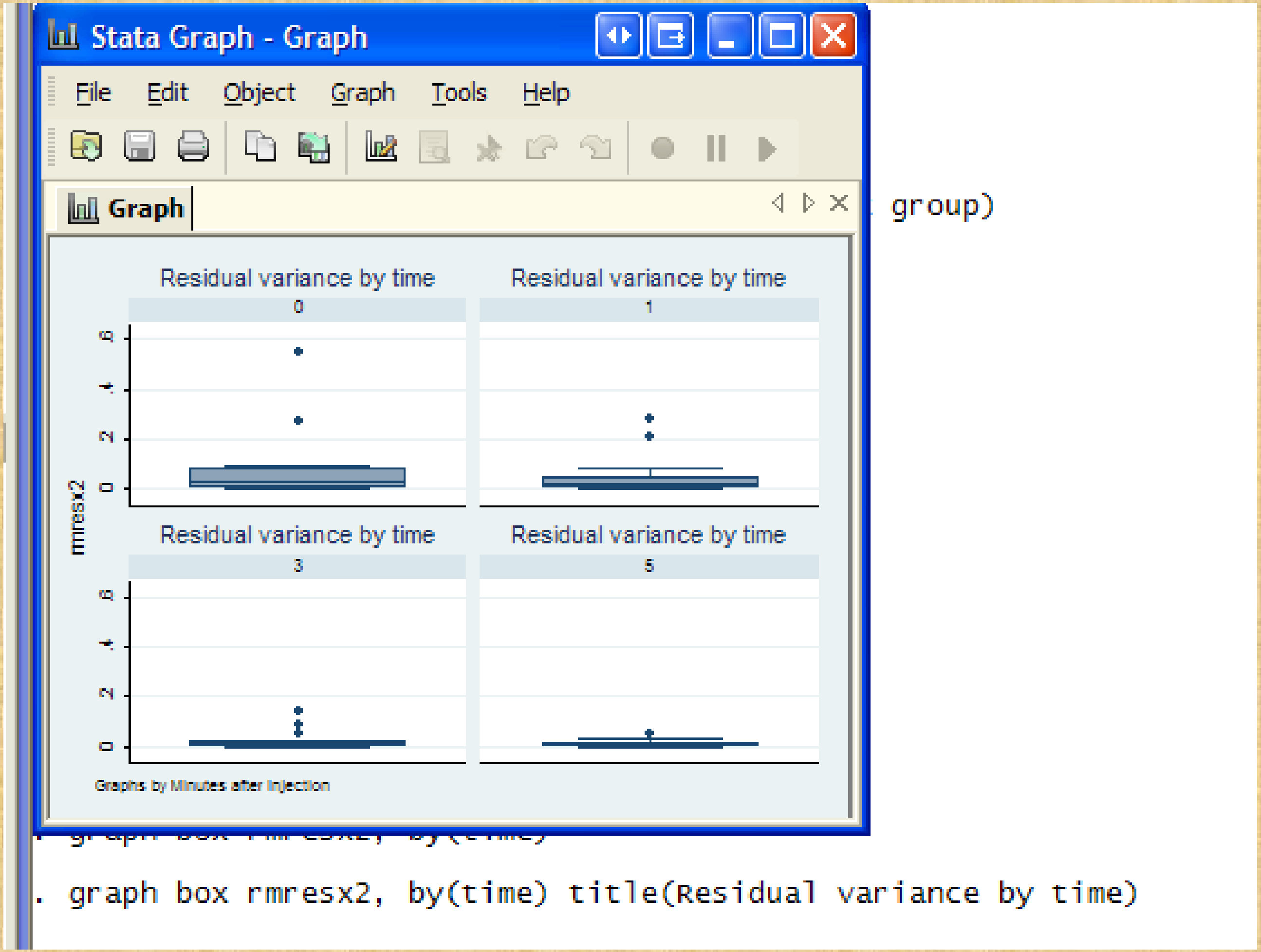
```
. twoway (scatter rmres time) (qfitci rmres time), by(group dog)
(note:  regress could not fit model)
```

Within-subject residual serial correlation confirmed



```
. gen rmresx2 = rmresx^2
(4 missing values generated)

. regress rmresx2 group time dog

      Source |       SS       df       MS              Number of obs =      60
-------------+------------------------------           F(  3,    56) =    4.24
       Model |  .086291045     3  .028763682           Prob > F      =  0.0090
    Residual |  .379461186    56  .006776093           R-squared     =  0.1853
-------------+------------------------------           Adj R-squared =  0.1416
       Total |  .465752232    59  .007894106           Root MSE      =  .08232

------------------------------------------------------------------------------
     rmresx2 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       group |  -.0316901   .0380836    -0.83   0.409    -.1079807    .0446006
        time |  -.0133025   .0055341    -2.40   0.020    -.0243886   -.0022163
         dog |   .0018365   .0092661     0.20   0.844    -.0167258    .0203989
       _cons |   .1356686   .0315451     4.30   0.000      .072476    .1988611
------------------------------------------------------------------------------

. regress rmres2 group dog grxdog

      Source |       SS       df       MS              Number of obs =      60
-------------+------------------------------           F(  3,    56) =    2.19
       Model |  .048898111     3   .01629937           Prob > F      =  0.0994
    Residual |  .416854121    56  .007443824           R-squared     =  0.1050
-------------+------------------------------           Adj R-squared =  0.0570
       Total |  .465752232    59  .007894106           Root MSE      =  .08628

------------------------------------------------------------------------------
      rmres2 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       group |  -.0430226   .0462268    -0.93   0.356    -.1356259    .0495807
         dog |  -.0012478   .0116013    -0.11   0.915    -.0244879    .0219924
      grxdog |   .0012866   .0026471     0.49   0.629    -.0040161    .0065893
       _cons |   .1261824   .0518855     2.43   0.018     .0222432    .2301216
------------------------------------------------------------------------------
```

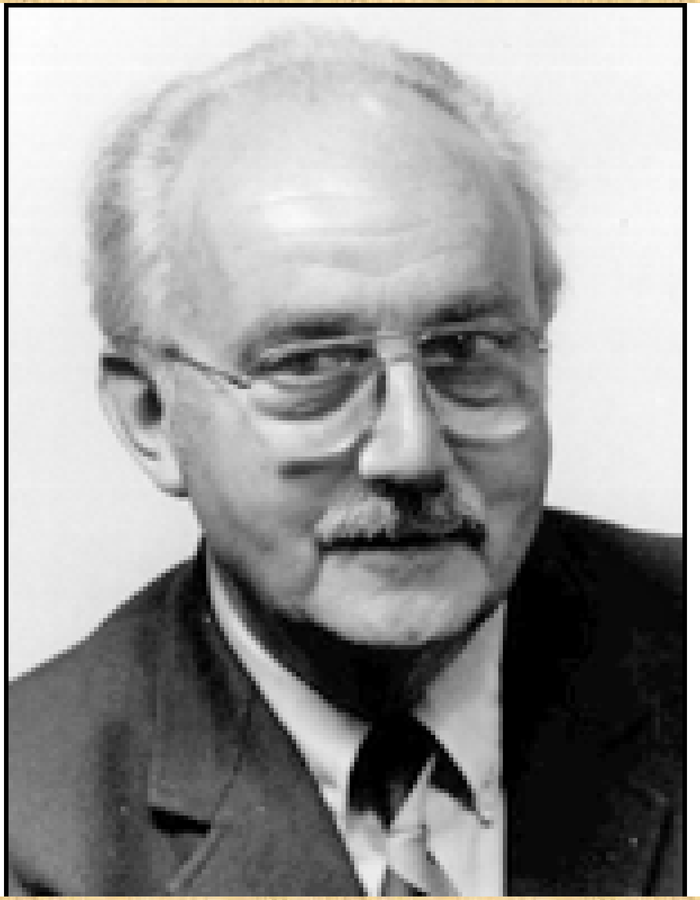
Residual diagnostics of heteroscedasticity



```
group)
. graph box rmresx2, by(time)

. graph box rmresx2, by(time) title(Residual variance by time)
```

George E. P. Box

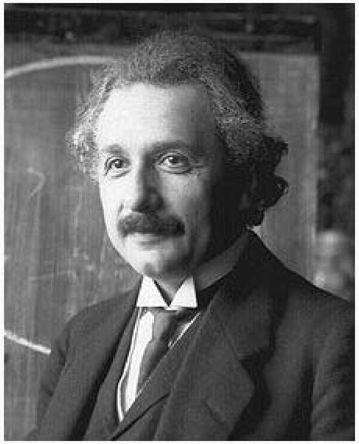
(“All models are wrong, but some happen to be useful.”



Albert Einstein

Institute of Advanced Studies Princeton, NJ

Albert Einstein



Albert Einstein, 1921

- Formulated the principle of parsimony: Keep it as simple as possible, but not simpler.
- So far as the theories of mathematics are about reality, they are not certain; so far as they are certain, they are not about reality.
- Do not worry about your difficulties in mathematics, I assure you that mine are greater.

OLS regression analysis
Adrien-Marie Legendre and C. F. Gauss

Adrien-Marie Legendre
Mathematician (1752-1833)

Regression models

- Basic theory
- Graph the data first (graph matrix of dependent with candidate independent variables). Search for possible good relationships. (p. 105)
- Ask if transformations to linearity are needed? Power transformations? Regression splines for piecewise models?

Assumptions of Ordinary Least Squares (OLS) (classical) regression analysis

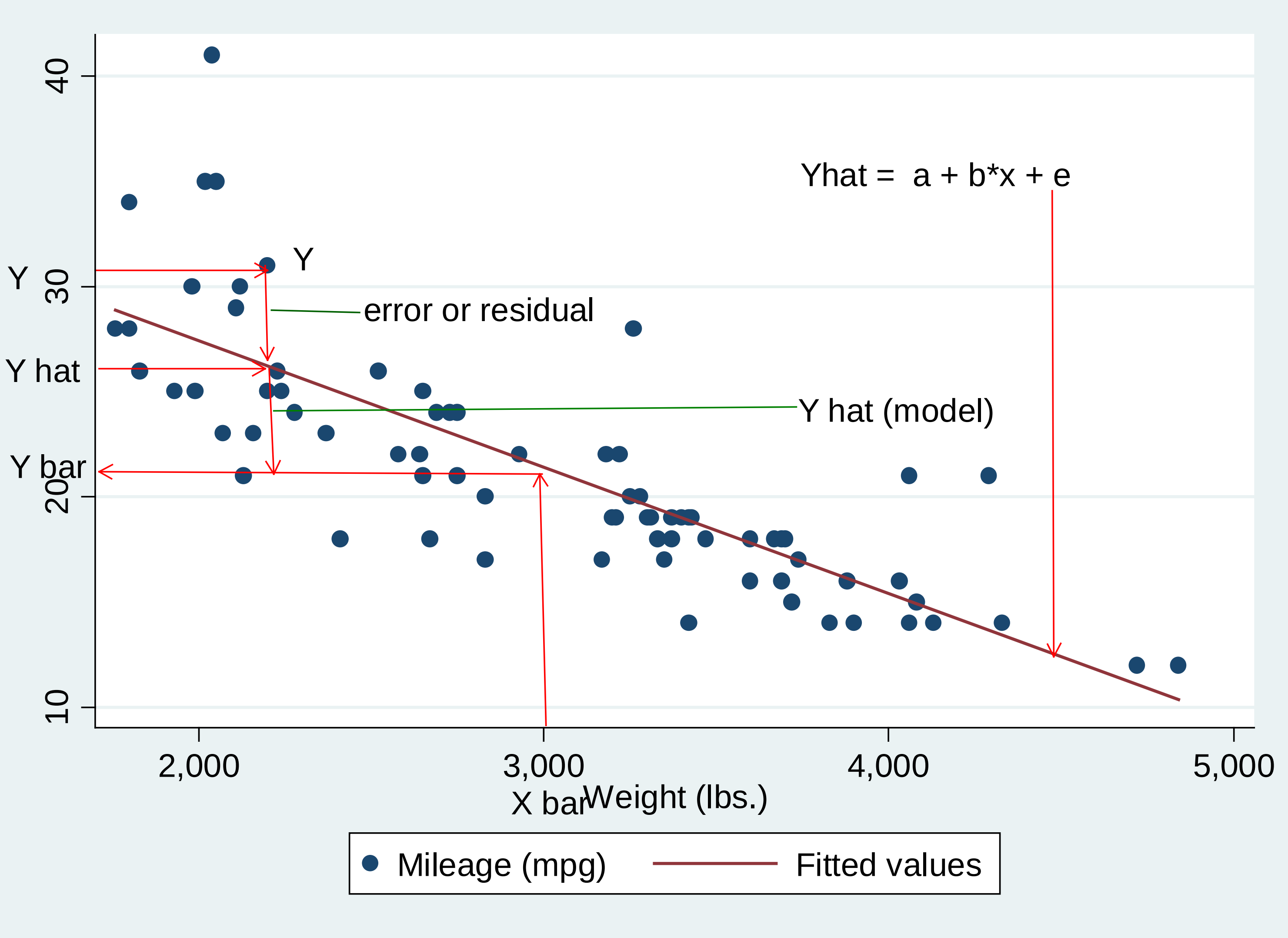
- Linear functional form
- Normality of residuals
- Homogeneity of variance
- Observations are iid. Errors are not correlated with the predictor variables.
- No outliers distorting the mean
- No multicollinearity
- Predictors are fixed or deterministic
  - If they are stochastic due to measurement error that could bias the model.

Regression analysis

developed by C.F. Gauss and Adrian Marie LeGendre

- Simple OLS theory if the dependent var is continuous
  - If assumptions are fulfilled
  - Polynomial regression
  - All possible subsets regression
- Problems with stepwise regression
- Regression postestimation
  - For normality
  - For heterogeneity of residuals
  - For multicollinearity
  - For functional form

Basic Regression model theory



$$total \quad = \quad model \quad + \quad error$$

$$(y_i - \bar{y}) = (\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i)$$

we square these

$$(y_i - \bar{y})^2 = (\hat{y}_i - \bar{y})^2 + (y_i - \hat{y}_i)^2$$

we add all of them up to obtain

$$total\ SS \quad = \quad model\ SS \quad + \quad error\ SS$$

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

$$Dividing\ \ the\ \ SS$$

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

$$by\ \ their\ \ respective\ \ df$$

$$dft = n - 1 \qquad dfm = k \qquad dfe = n - k - 1$$

$$gives$$

$$MStotal \quad = \quad MS\,regression + \ MS\ \ error$$

$$\sum_{i=1}^{n}\frac{(y_i - \bar{y})^2}{n-1} = \sum_{i=1}^{n}\frac{(\hat{y}_i - \bar{y})^2}{k} + \sum_{i=1}^{n}\frac{(y_i - \hat{y}_i)^2}{n-k-1}$$

$$Total\ \ variance = \ \ Model\ \ variance + error\ \ variance$$

Omnibus F test

$$F(mdf, edf) = \frac{regression\ model\ variance}{error\ variance}$$

$$F(k, n-k-\mathbf{1}) = \frac{\dfrac{R^2}{k}}{\dfrac{\mathbf{1}-R^2}{n-k-\mathbf{1}}}$$

We can solve for b

$$e_i = \hat{y}_i - y_i$$

$$e_i^2 = (\hat{y}_i - y_i)^2$$

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (\hat{y}_i - y_i)^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - a - bx_i)^2$$

$$\frac{\partial \sum_{i=1}^{n} e_i^2}{\partial b} = \frac{\partial \sum_{i=1}^{n} (y_i - a - bx_i)^2}{\partial b}$$

$$0 = 2\sum xy - 2b\sum x^2$$

$$b = \frac{\sum_{i=1}^{n} xy}{\sum_{i=1}^{n} x^2}$$

We can also solve for a

$$y_i = a + bx_i$$

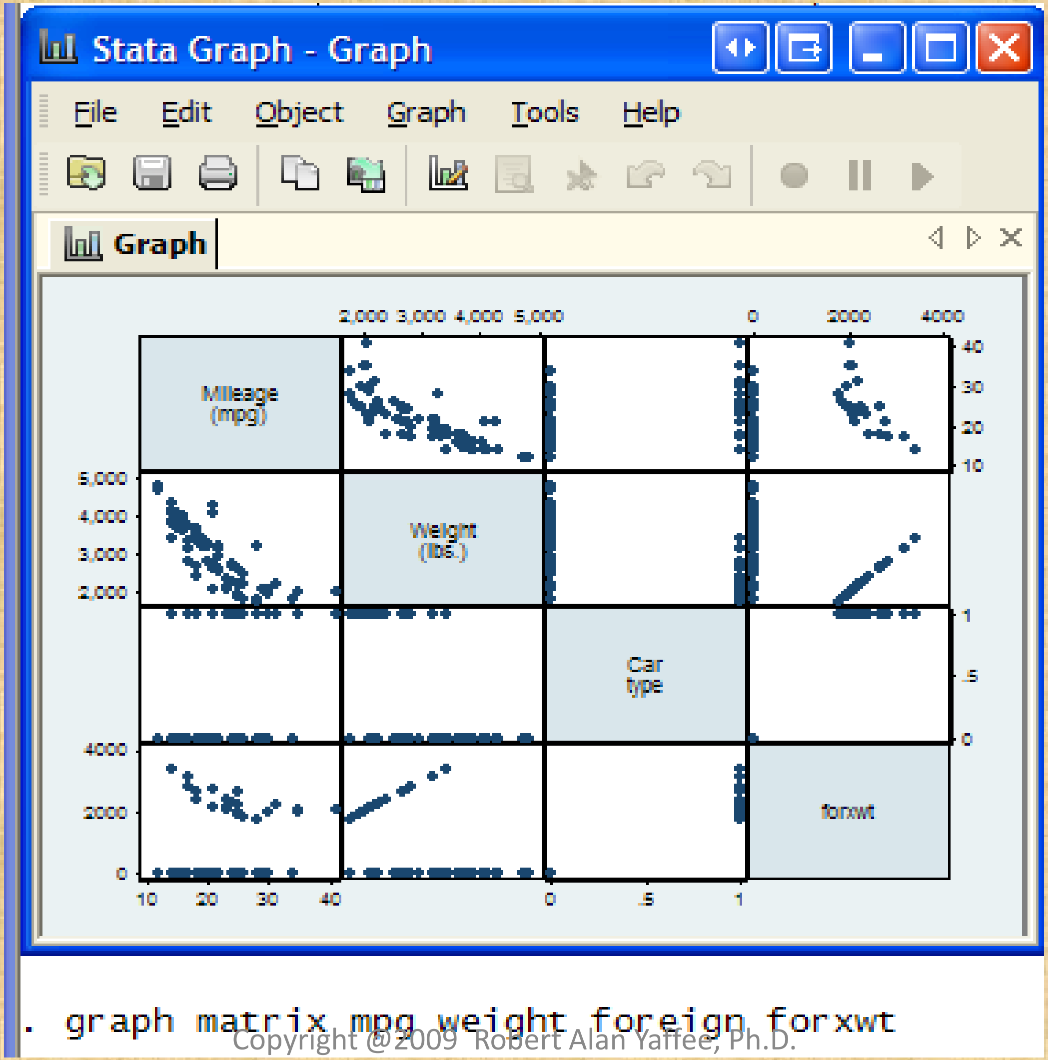
$$\text{for each person}$$

$$\text{sum for the whole sample:}$$

$$\sum_{i=1}^{n} y_i = n * a + b \sum_{i}^{n} x_i$$

$$a = \overline{y} - b\overline{x}$$

Diagnosing functional form with a matrix graph



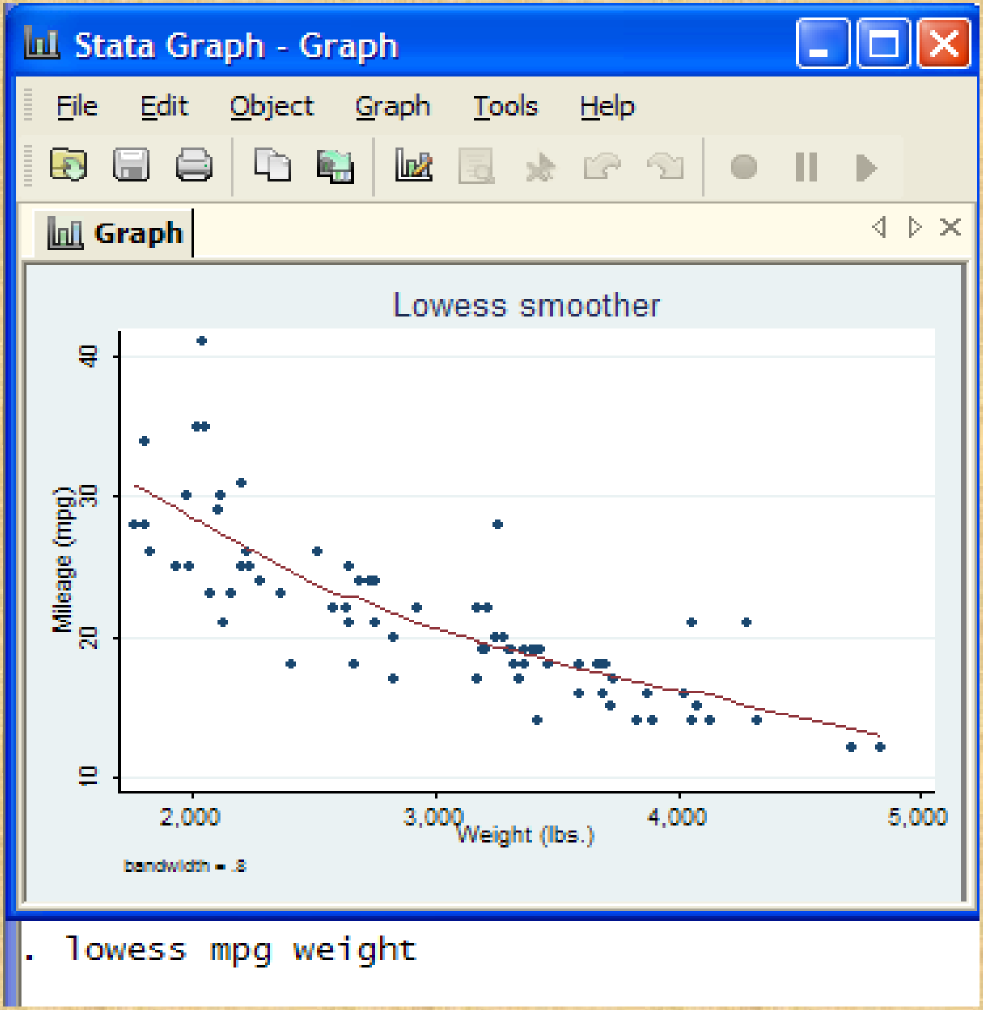
```
. graph matrix mpg weight foreign forxwt
```

Functional form

- Are any of the foregoing plots indicative of possible nonlinear relationships?
- Which ones?
- Mpg and weight?
- Mpg and forxwt?

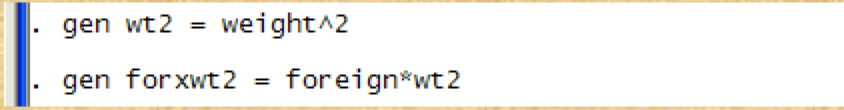
Frank E. Harrell Jr. (2001) Regression Modeling Strategies, Springer: New York. advocated using lowess and/or splines to model the nonlinearity found in most n relationships. Chapter 2.

A lowess plot

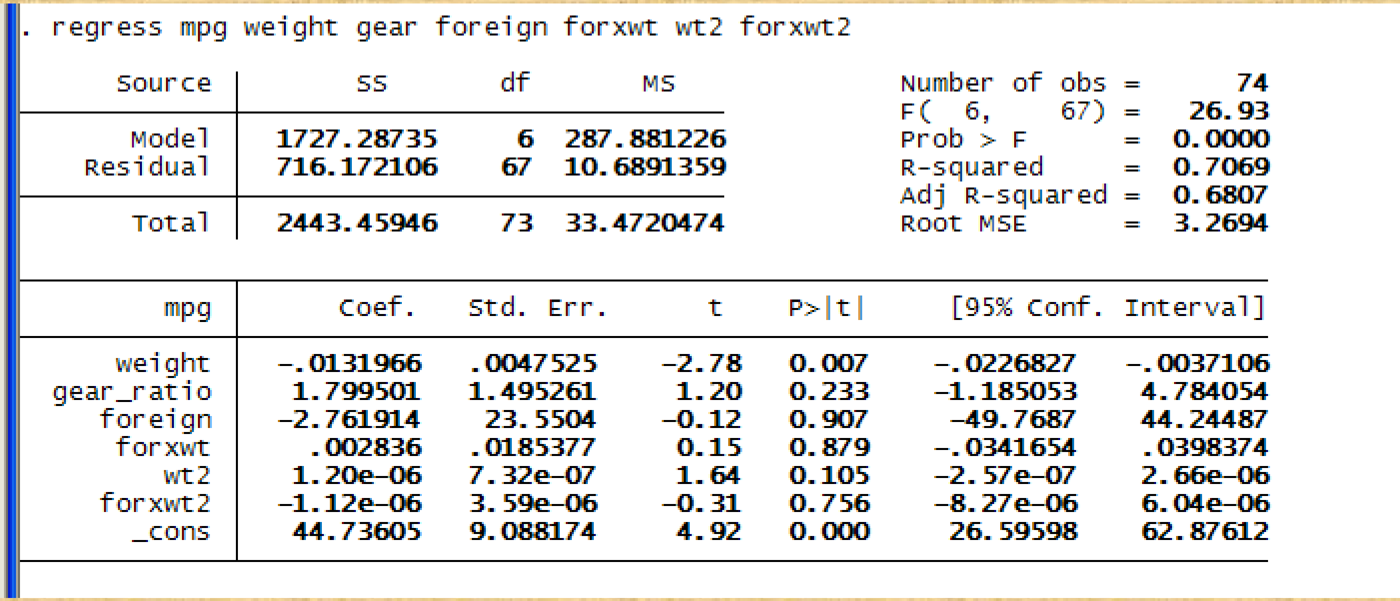


```
. lowess mpg weight
```

Polynomial regression



```
. gen wt2 = weight^2

. gen forxwt2 = foreign*wt2
```

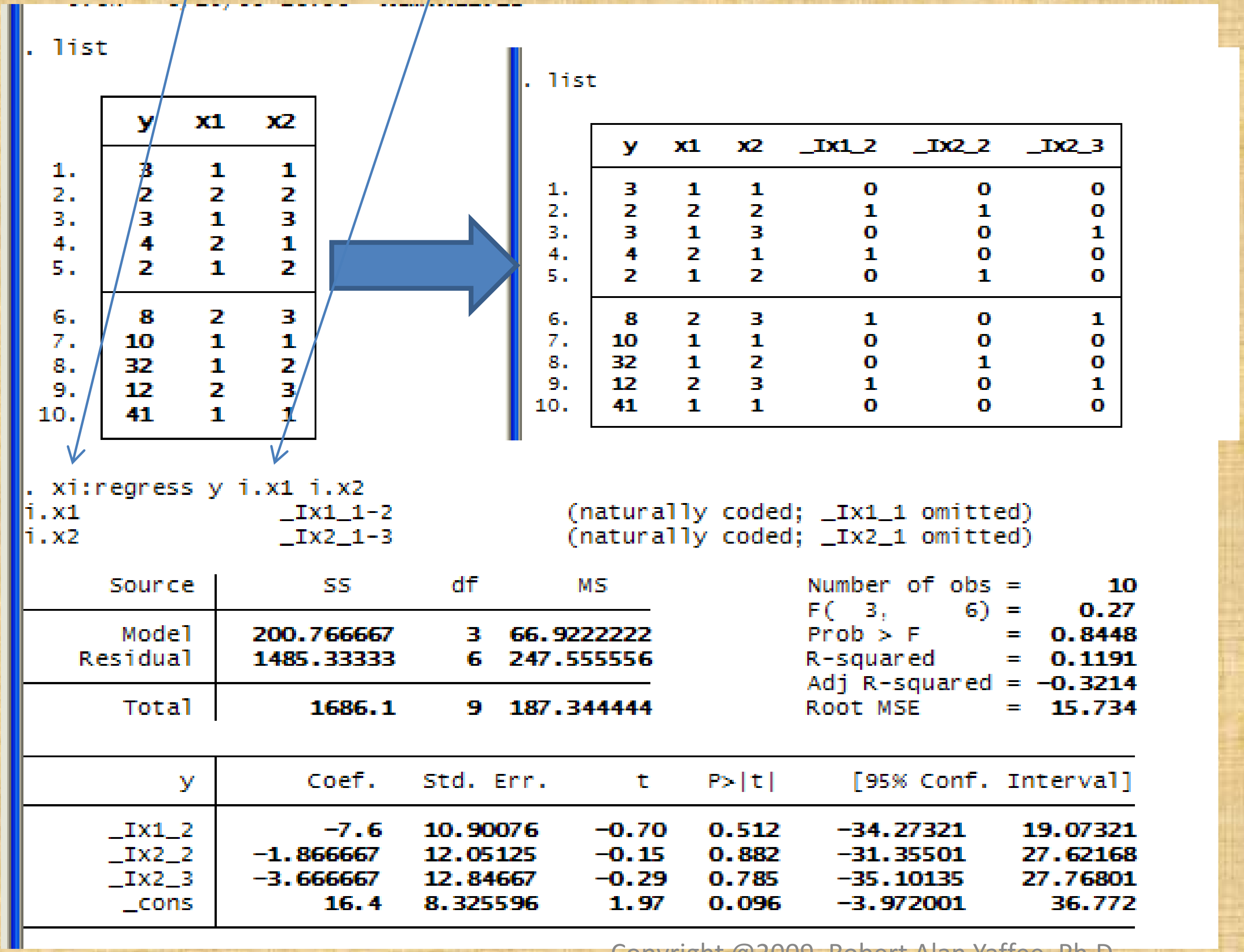


```
. regress mpg weight gear foreign forxwt wt2 forxwt2

      Source |       SS       df       MS              Number of obs =      74
-------------+------------------------------           F(  6,    67) =   26.93
       Model |  1727.28735     6  287.881226           Prob > F      =  0.0000
    Residual |  716.172106    67  10.6891359           R-squared     =  0.7069
-------------+------------------------------           Adj R-squared =  0.6807
       Total |  2443.45946    73  33.4720474           Root MSE      =  3.2694

------------------------------------------------------------------------------
         mpg |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      weight |  -.0131966   .0047525    -2.78   0.007    -.0226827   -.0037106
  gear_ratio |   1.799501   1.495261     1.20   0.233    -1.185053    4.784054
     foreign |  -2.761914    23.5504    -0.12   0.907     -49.7687    44.24487
      forxwt |    .002836   .0185377     0.15   0.879    -.0341654    .0398374
         wt2 |   1.20e-06   7.32e-07     1.64   0.105    -2.57e-07    2.66e-06
     forxwt2 |  -1.12e-06   3.59e-06    -0.31   0.756    -8.27e-06    6.04e-06
       _cons |   44.73605   9.088174     4.92   0.000     26.59598    62.87612
------------------------------------------------------------------------------
```

xi: Interaction analysis

- Macro for converting categorical data to dummy variables for analysis.
- The macro will also construct all of the main effects and first-order interaction terms for such an analysis.

xi: and i. prefixes for dummy coding categorical variables with main effects

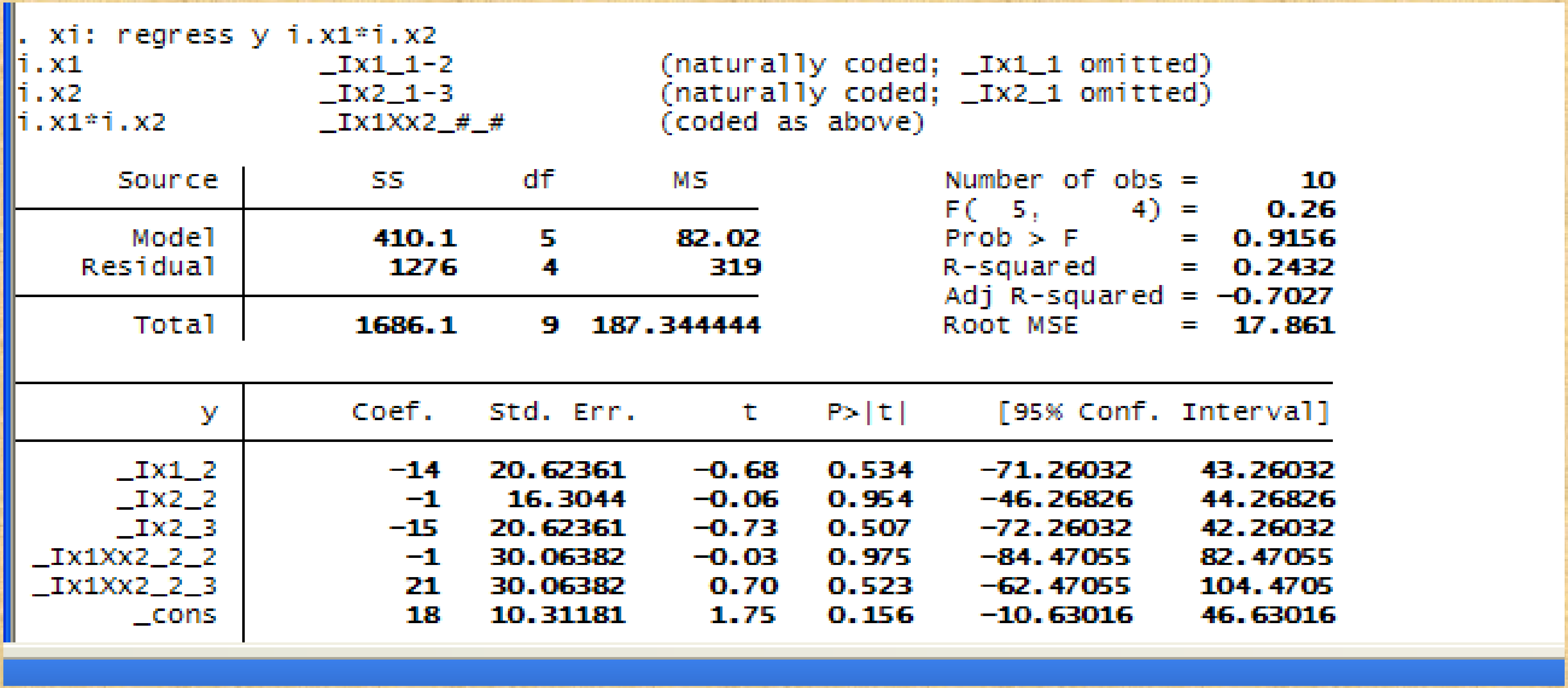


```
. list

     +--------------+
     |  y   x1   x2 |
     |--------------|
  1. |  3    1    1 |
  2. |  2    2    2 |
  3. |  3    1    3 |
  4. |  4    2    1 |
  5. |  2    1    2 |
     |--------------|
  6. |  8    2    3 |
  7. | 10    1    1 |
  8. | 32    1    2 |
  9. | 12    2    3 |
 10. | 41    1    1 |
     +--------------+
```

```
. list

     +-----------------------------------------------+
     |  y   x1   x2   _Ix1_2   _Ix2_2   _Ix2_3 |
     |-----------------------------------------------|
  1. |  3    1    1        0        0        0 |
  2. |  2    2    2        1        1        0 |
  3. |  3    1    3        0        0        1 |
  4. |  4    2    1        1        0        0 |
  5. |  2    1    2        0        1        0 |
     |-----------------------------------------------|
  6. |  8    2    3        1        0        1 |
  7. | 10    1    1        0        0        0 |
  8. | 32    1    2        0        1        0 |
  9. | 12    2    3        1        0        1 |
 10. | 41    1    1        0        0        0 |
     +-----------------------------------------------+
```

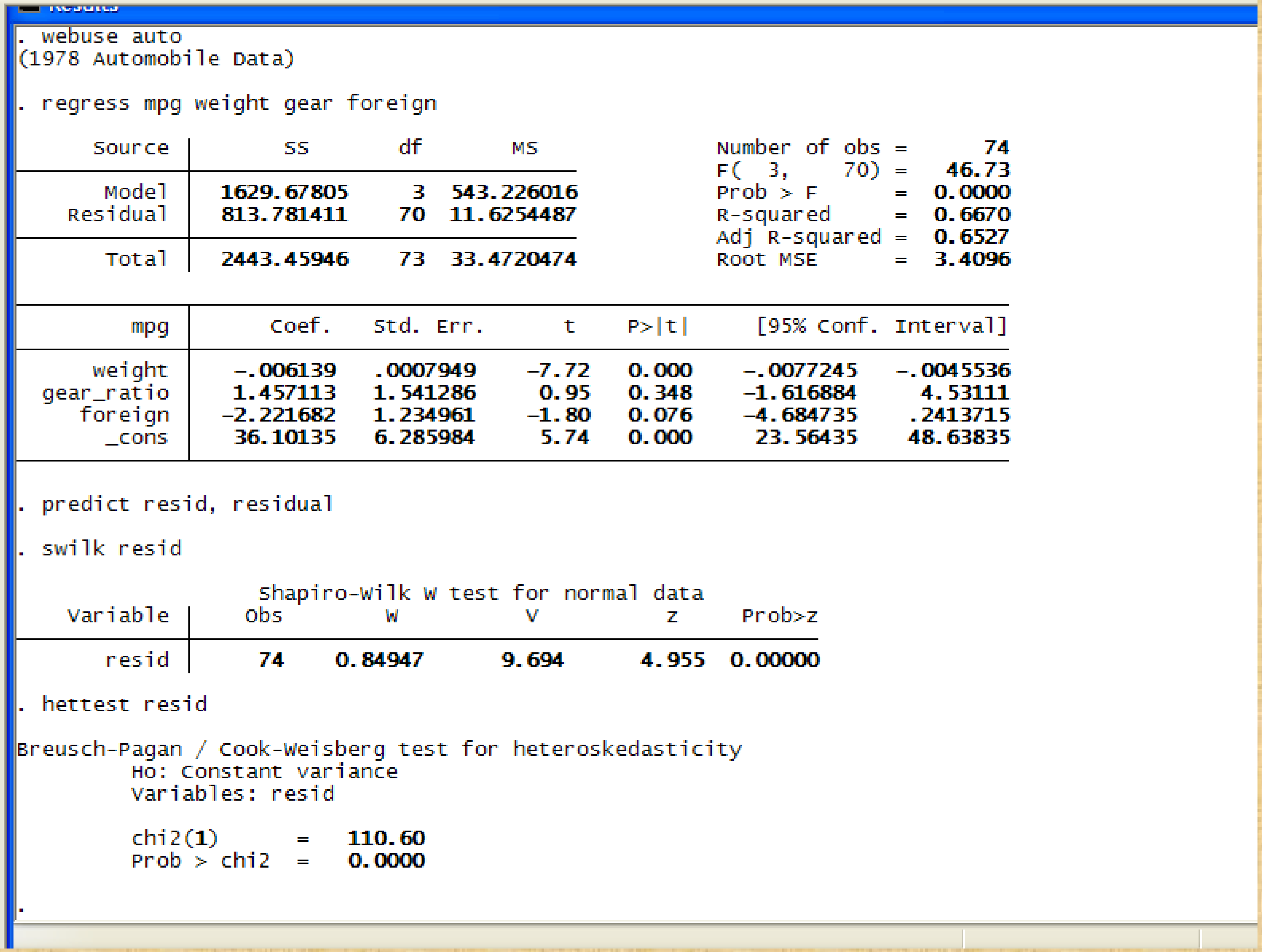
```
. xi:regress y i.x1 i.x2
i.x1              _Ix1_1-2            (naturally coded; _Ix1_1 omitted)
i.x2              _Ix2_1-3            (naturally coded; _Ix2_1 omitted)

      Source |       SS       df       MS              Number of obs =      10
-------------+------------------------------           F(  3,     6) =    0.27
       Model |  200.766667     3  66.9222222           Prob > F      =  0.8448
    Residual |  1485.33333     6  247.555556           R-squared     =  0.1191
-------------+------------------------------           Adj R-squared = -0.3214
       Total |      1686.1     9  187.344444           Root MSE      =  15.734

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      _Ix1_2 |       -7.6   10.90076    -0.70   0.512    -34.27321    19.07321
      _Ix2_2 |  -1.866667   12.05125    -0.15   0.882    -31.35501    27.62168
      _Ix2_3 |  -3.666667   12.84667    -0.29   0.785    -35.10135    27.76801
       _cons |       16.4   8.325596     1.97   0.096    -3.972001      36.772
------------------------------------------------------------------------------
```

xi: and i.x1*i.x2 construct dummy variables for all main effects and interactions for the model



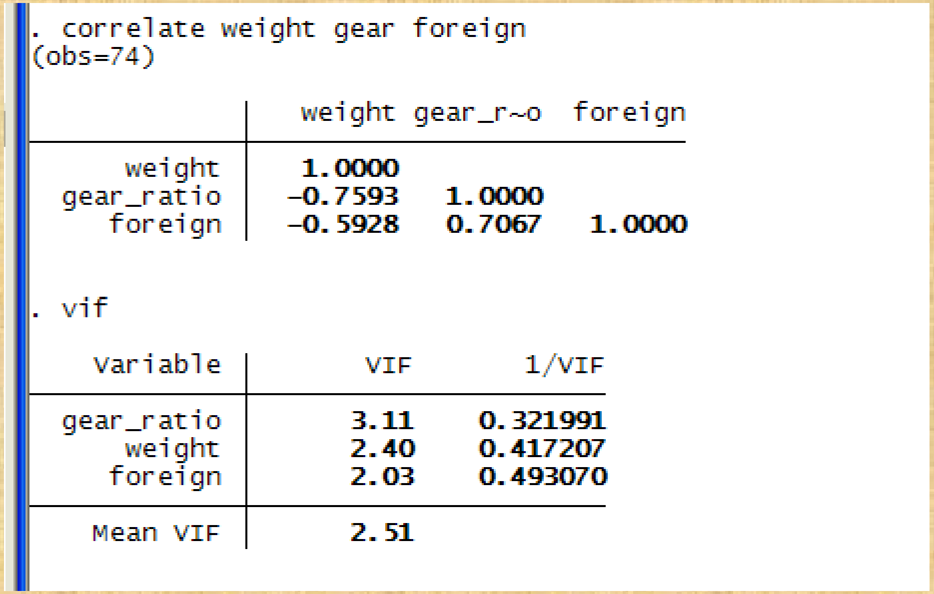
```
. xi: regress y i.x1*i.x2
i.x1              _Ix1_1-2            (naturally coded; _Ix1_1 omitted)
i.x2              _Ix2_1-3            (naturally coded; _Ix2_1 omitted)
i.x1*i.x2         _Ix1Xx2_#_#         (coded as above)

      Source |       SS       df       MS              Number of obs =      10
-------------+------------------------------           F(  5,     4) =    0.26
       Model |      410.1      5       82.02           Prob > F      =  0.9156
    Residual |       1276      4         319           R-squared     =  0.2432
-------------+------------------------------           Adj R-squared = -0.7027
       Total |     1686.1      9  187.344444           Root MSE      =  17.861

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      _Ix1_2 |        -14   20.62361    -0.68   0.534    -71.26032    43.26032
      _Ix2_2 |         -1    16.3044    -0.06   0.954    -46.26826    44.26826
      _Ix2_3 |        -15   20.62361    -0.73   0.507    -72.26032    42.26032
 _Ix1Xx2_2_2 |         -1   30.06382    -0.03   0.975    -84.47055    82.47055
 _Ix1Xx2_2_3 |         21   30.06382     0.70   0.523    -62.47055    104.4705
       _cons |         18   10.31181     1.75   0.156    -10.63016    46.63016
```

OLS regression with some residual diagnostics



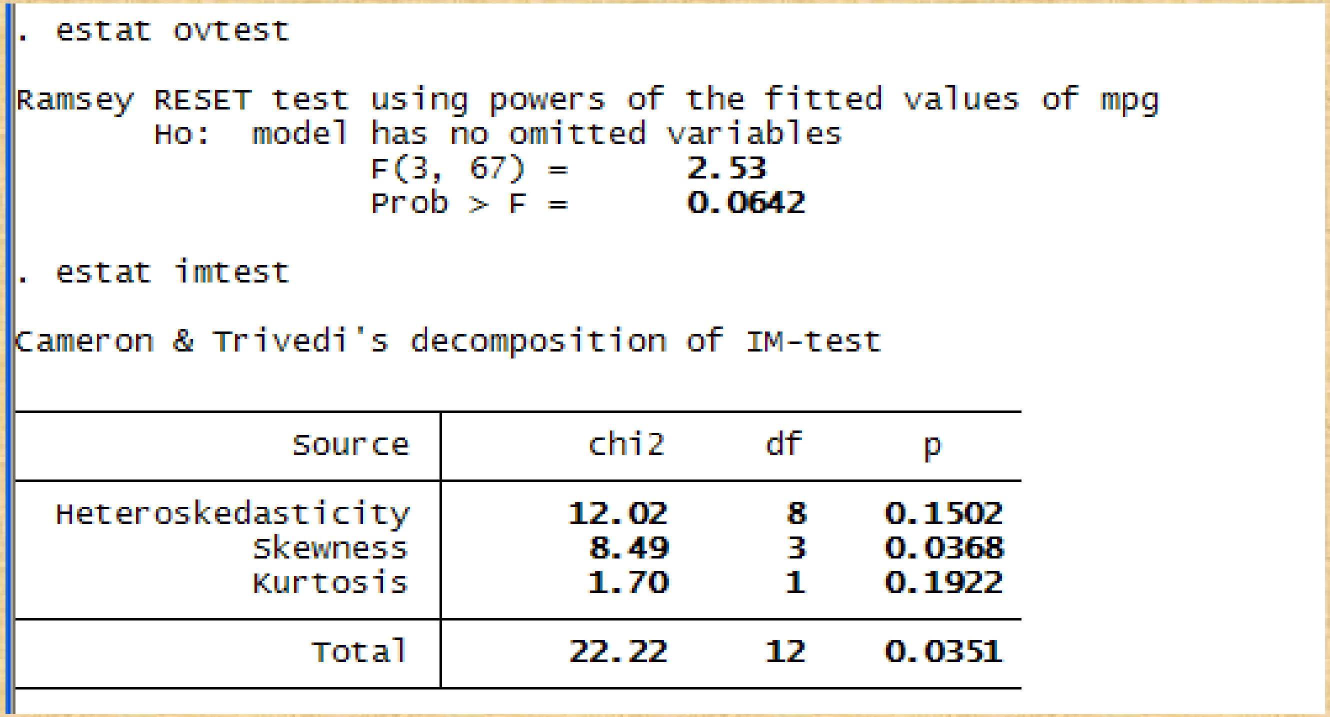
```
. webuse auto
(1978 Automobile Data)

. regress mpg weight gear foreign

      Source |       SS       df       MS              Number of obs =      74
-------------+------------------------------           F(  3,    70) =   46.73
       Model |  1629.67805     3  543.226016           Prob > F      =  0.0000
    Residual |  813.781411    70  11.6254487           R-squared     =  0.6670
-------------+------------------------------           Adj R-squared =  0.6527
       Total |  2443.45946    73  33.4720474           Root MSE      =  3.4096

------------------------------------------------------------------------------
         mpg |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      weight |   -.006139   .0007949    -7.72   0.000    -.0077245   -.0045536
  gear_ratio |   1.457113   1.541286     0.95   0.348    -1.616884     4.53111
     foreign |  -2.221682   1.234961    -1.80   0.076    -4.684735    .2413715
       _cons |   36.10135   6.285984     5.74   0.000     23.56435    48.63835
------------------------------------------------------------------------------

. predict resid, residual

. swilk resid

                   Shapiro-Wilk W test for normal data

    Variable |    Obs       W           V         z       Prob>z
-------------+-------------------------------------------------
       resid |     74    0.84947      9.694     4.955    0.00000

. hettest resid

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: resid

         chi2(1)      =   110.60
         Prob > chi2  =   0.0000

.
```

More residual diagnostics



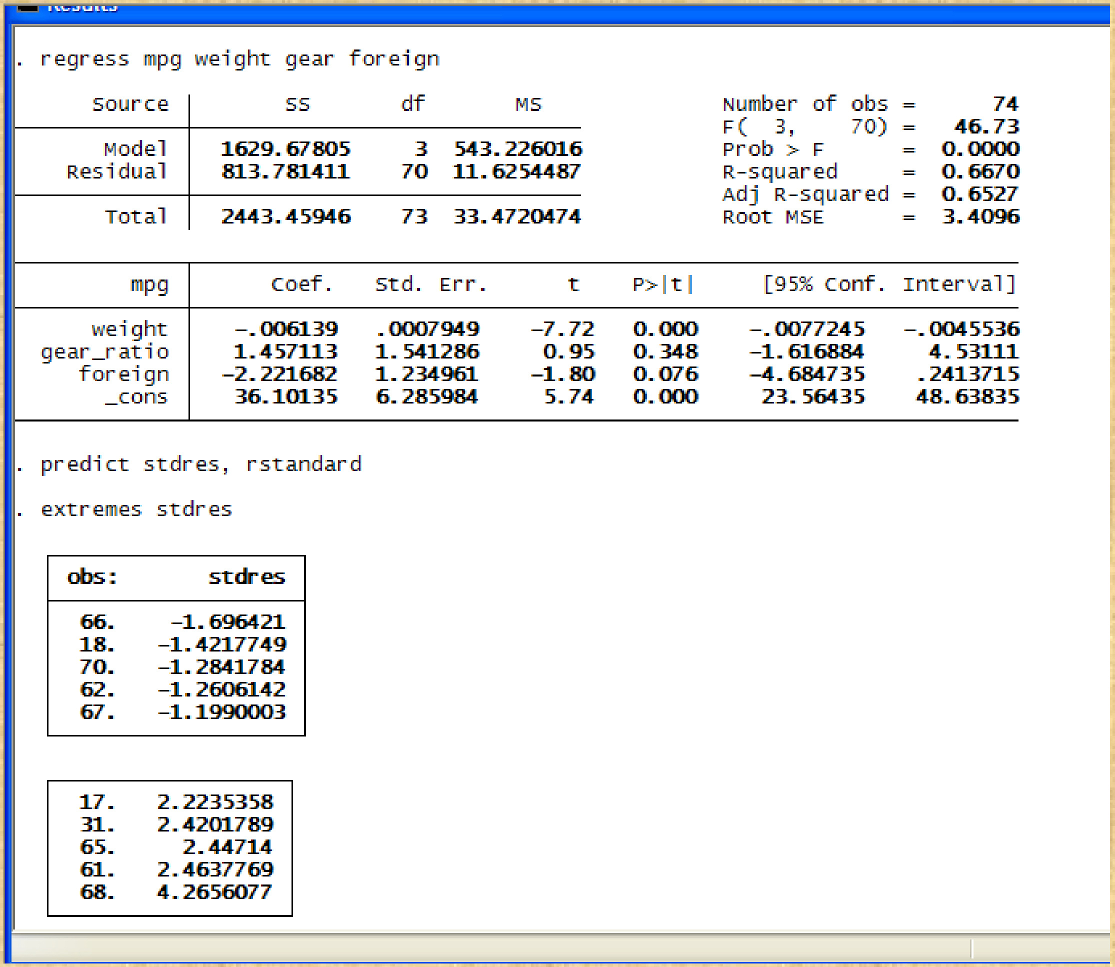
```
. correlate weight gear foreign
(obs=74)

             |   weight gear_r~o  foreign
-------------+---------------------------
      weight |   1.0000
  gear_ratio |  -0.7593   1.0000
     foreign |  -0.5928   0.7067   1.0000


. vif

    Variable |       VIF       1/VIF
-------------+----------------------
  gear_ratio |      3.11    0.321991
      weight |      2.40    0.417207
     foreign |      2.03    0.493070
-------------+----------------------
    Mean VIF |      2.51
```

Other residual diagnostics



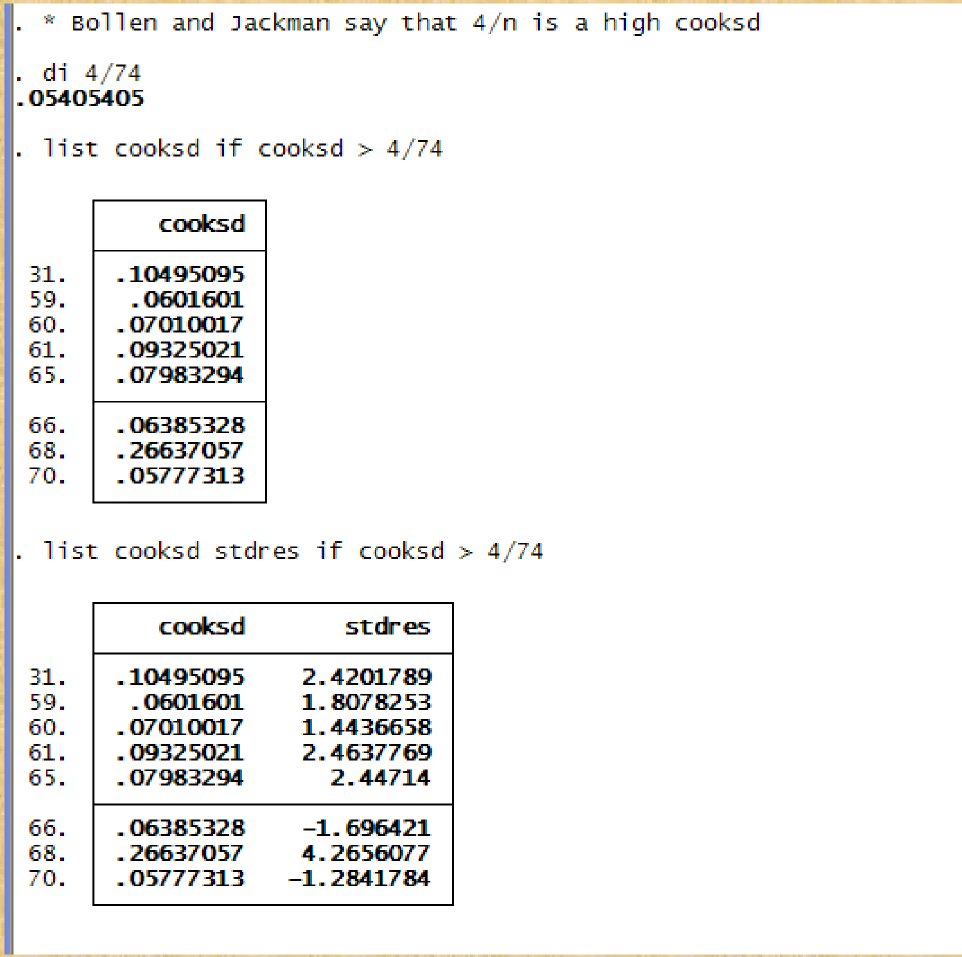
```
. estat ovtest

Ramsey RESET test using powers of the fitted values of mpg
       Ho:  model has no omitted variables
                 F(3, 67) =      2.53
                 Prob > F =      0.0642

. estat imtest

Cameron & Trivedi's decomposition of IM-test

---------------------------------------------------
              Source |       chi2     df      p
---------------------+-----------------------------
  Heteroskedasticity |      12.02      8    0.1502
            Skewness |       8.49      3    0.0368
            Kurtosis |       1.70      1    0.1922
---------------------+-----------------------------
               Total |      22.22     12    0.0351
---------------------------------------------------
```

Outlier diagnosis
(residuals larger than 3 std errors)



```
. regress mpg weight gear foreign

      Source |       SS       df       MS              Number of obs =      74
-------------+------------------------------           F(  3,    70) =   46.73
       Model |  1629.67805     3  543.226016           Prob > F      =  0.0000
    Residual |  813.781411    70  11.6254487           R-squared     =  0.6670
-------------+------------------------------           Adj R-squared =  0.6527
       Total |  2443.45946    73  33.4720474           Root MSE      =  3.4096

------------------------------------------------------------------------------
         mpg |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      weight |   -.006139   .0007949    -7.72   0.000    -.0077245   -.0045536
  gear_ratio |   1.457113   1.541286     0.95   0.348    -1.616884     4.53111
     foreign |  -2.221682   1.234961    -1.80   0.076    -4.684735    .2413715
       _cons |   36.10135   6.285984     5.74   0.000     23.56435    48.63835
------------------------------------------------------------------------------

. predict stdres, rstandard

. extremes stdres

  +------------------+
  | obs:      stdres |
  |------------------|
  |  66.    -1.696421 |
  |  18.   -1.4217749 |
  |  70.   -1.2841784 |
  |  62.   -1.2606142 |
  |  67.   -1.1990003 |
  +------------------+

  +-----------------+
  |  17.   2.2235358 |
  |  31.   2.4201789 |
  |  65.     2.44714 |
  |  61.   2.4637769 |
  |  68.   4.2656077 |
  +-----------------+
```

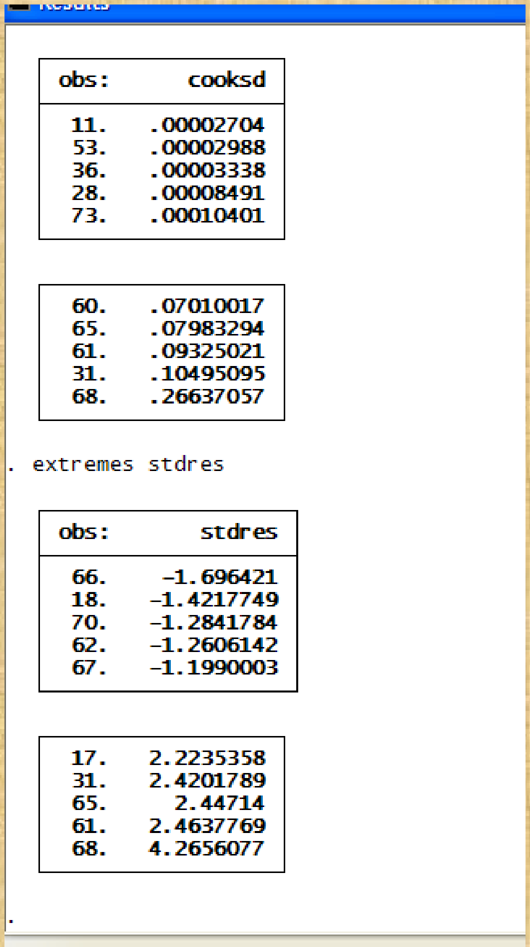
Testing for influential outliers (Bollen, K. and Jackman, R.W., 1990)



```
. * Bollen and Jackman say that 4/n is a high cooksd

. di 4/74
.05405405

. list cooksd if cooksd > 4/74


     +-----------+
     |    cooksd |
     |-----------|
 31. |  .10495095 |
 59. |   .0601601 |
 60. |  .07010017 |
 61. |  .09325021 |
 65. |  .07983294 |
     |-----------|
 66. |  .06385328 |
 68. |  .26637057 |
 70. |  .05777313 |
     +-----------+

. list cooksd stdres if cooksd > 4/74


     +-------------------------+
     |    cooksd       stdres |
     |-------------------------|
 31. |  .10495095    2.4201789 |
 59. |   .0601601    1.8078253 |
 60. |  .07010017    1.4436658 |
 61. |  .09325021    2.4637769 |
 65. |  .07983294      2.44714 |
     |-------------------------|
 66. |  .06385328    -1.696421 |
 68. |  .26637057    4.2656077 |
 70. |  .05777313   -1.2841784 |
     +-------------------------+
```

Extremes cooksd

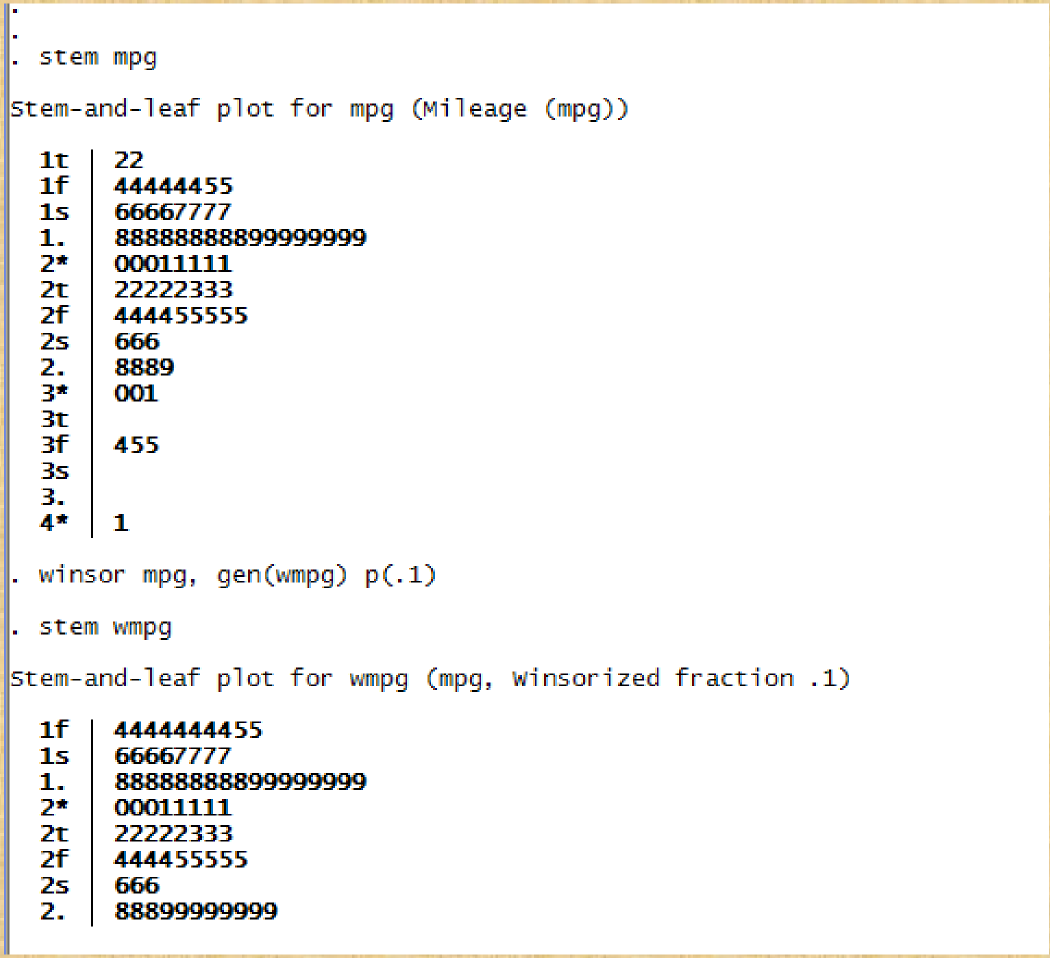
When Cook's distance > n/4 then it may be a problem



```
  +-------------------+
  | obs:       cooksd |
  |-------------------|
  |  11.    .00002704 |
  |  53.    .00002988 |
  |  36.    .00003338 |
  |  28.    .00008491 |
  |  73.    .00010401 |
  +-------------------+

  +-------------------+
  |  60.    .07010017 |
  |  65.    .07983294 |
  |  61.    .09325021 |
  |  31.    .10495095 |
  |  68.    .26637057 |
  +-------------------+

. extremes stdres

  +--------------------+
  | obs:        stdres |
  |--------------------|
  |  66.     -1.696421 |
  |  18.    -1.4217749 |
  |  70.    -1.2841784 |
  |  62.    -1.2606142 |
  |  67.    -1.1990003 |
  +--------------------+

  +-------------------+
  |  17.    2.2235358 |
  |  31.    2.4201789 |
  |  65.      2.44714 |
  |  61.    2.4637769 |
  |  68.    4.2656077 |
  +-------------------+

.
```

How can you deal with the extreme values? Winsorizing

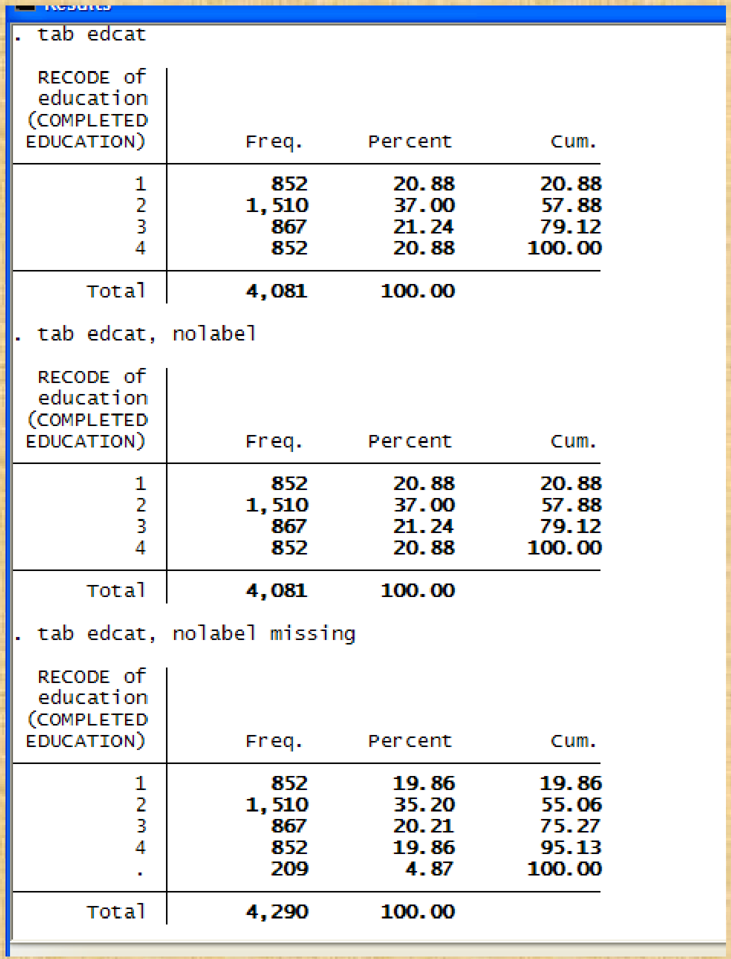
Taking the extreme non-missing ordered values of x and sets equal to the next highest and lowest values.



```
.
.
. stem mpg

Stem-and-leaf plot for mpg (Mileage (mpg))

  1t | 22
  1f | 44444455
  1s | 66667777
  1. | 88888888899999999
  2* | 00011111
  2t | 22222333
  2f | 444455555
  2s | 666
  2. | 8889
  3* | 001
  3t |
  3f | 455
  3s |
  3. |
  4* | 1

. winsor mpg, gen(wmpg) p(.1)

. stem wmpg

Stem-and-leaf plot for wmpg (mpg, Winsorized fraction .1)

  1f | 4444444455
  1s | 66667777
  1. | 88888888899999999
  2* | 00011111
  2t | 22222333
  2f | 444455555
  2s | 666
  2. | 88899999999
```

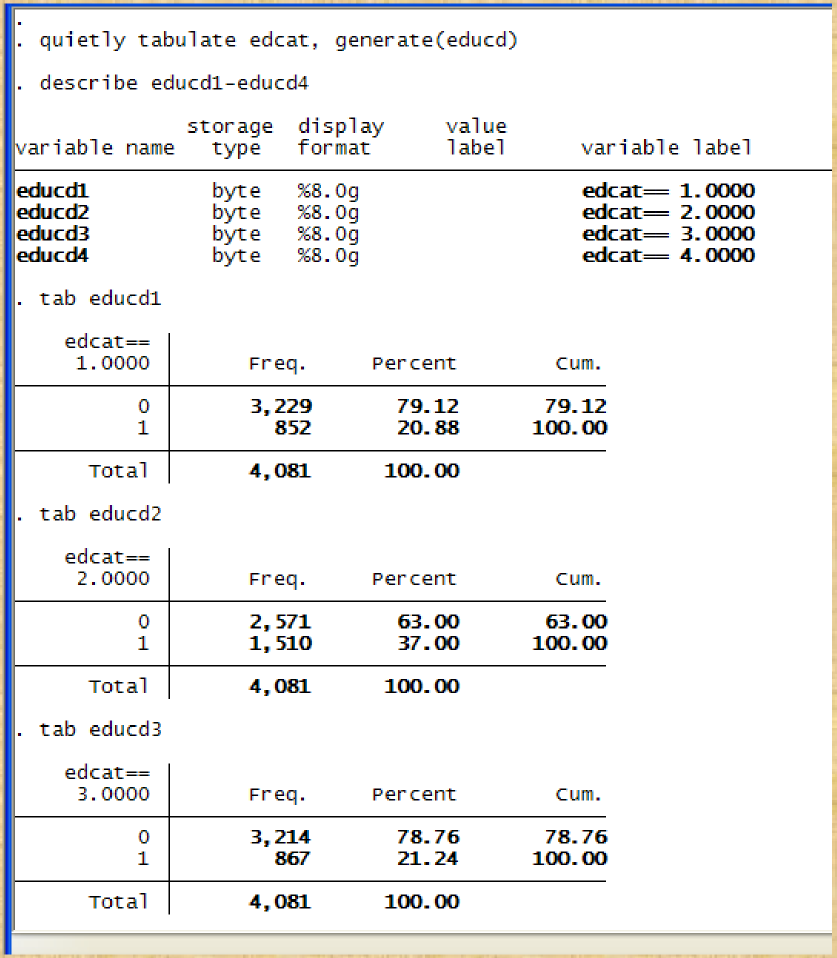
Automatic Interaction construction

First check the variables for missing values



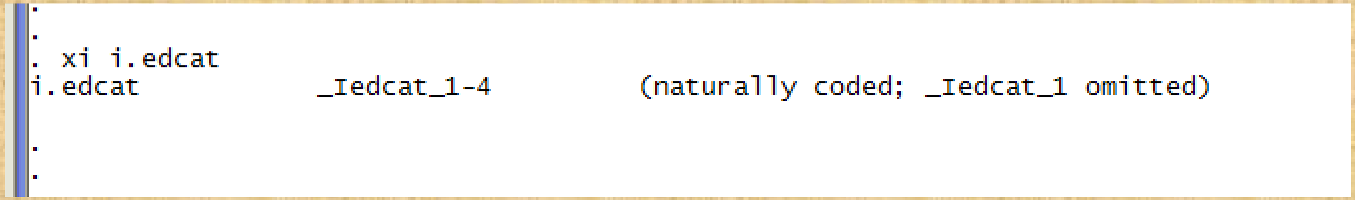
```
. tab edcat

  RECODE of |
  education |
 (COMPLETED |
 EDUCATION) |      Freq.     Percent        Cum.
------------+-----------------------------------
          1 |        852       20.88       20.88
          2 |      1,510       37.00       57.88
          3 |        867       21.24       79.12
          4 |        852       20.88      100.00
------------+-----------------------------------
      Total |      4,081      100.00

. tab edcat, nolabel

  RECODE of |
  education |
 (COMPLETED |
 EDUCATION) |      Freq.     Percent        Cum.
------------+-----------------------------------
          1 |        852       20.88       20.88
          2 |      1,510       37.00       57.88
          3 |        867       21.24       79.12
          4 |        852       20.88      100.00
------------+-----------------------------------
      Total |      4,081      100.00

. tab edcat, nolabel missing

  RECODE of |
  education |
 (COMPLETED |
 EDUCATION) |      Freq.     Percent        Cum.
------------+-----------------------------------
          1 |        852       19.86       19.86
          2 |      1,510       35.20       55.06
          3 |        867       20.21       75.27
          4 |        852       19.86       95.13
          . |        209        4.87      100.00
------------+-----------------------------------
      Total |      4,290      100.00
```

Construct dummy variables

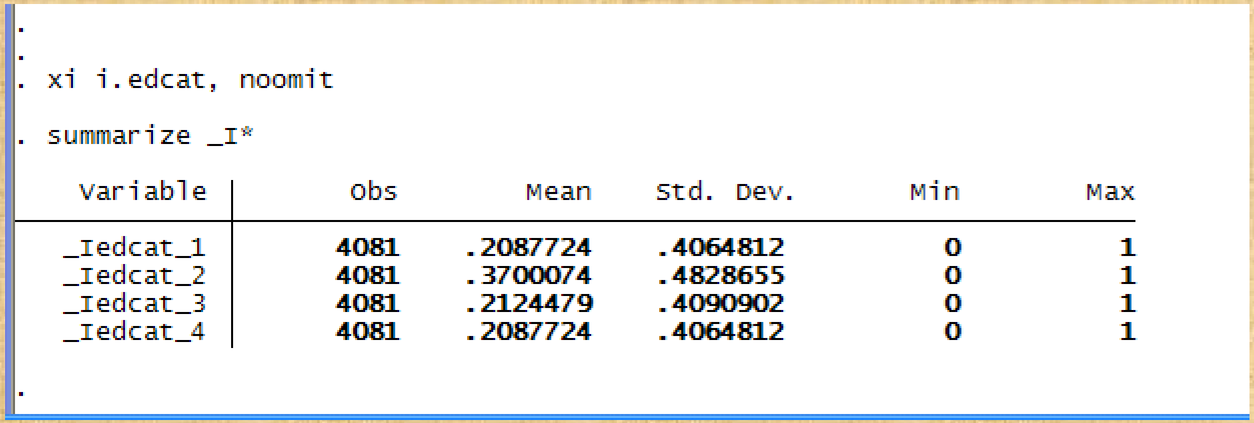


```
.
. quietly tabulate edcat, generate(educd)

. describe educd1-educd4

              storage  display     value
variable name   type   format      label      variable label
-------------------------------------------------------------------------------
educd1          byte   %8.0g                  edcat== 1.0000
educd2          byte   %8.0g                  edcat== 2.0000
educd3          byte   %8.0g                  edcat== 3.0000
educd4          byte   %8.0g                  edcat== 4.0000

. tab educd1

    edcat== |
     1.0000 |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |      3,229       79.12       79.12
          1 |        852       20.88      100.00
------------+-----------------------------------
      Total |      4,081      100.00

. tab educd2

    edcat== |
     2.0000 |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |      2,571       63.00       63.00
          1 |      1,510       37.00      100.00
------------+-----------------------------------
      Total |      4,081      100.00

. tab educd3

    edcat== |
     3.0000 |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |      3,214       78.76       78.76
          1 |        867       21.24      100.00
------------+-----------------------------------
      Total |      4,081      100.00
```

Construct indicator variables with xi



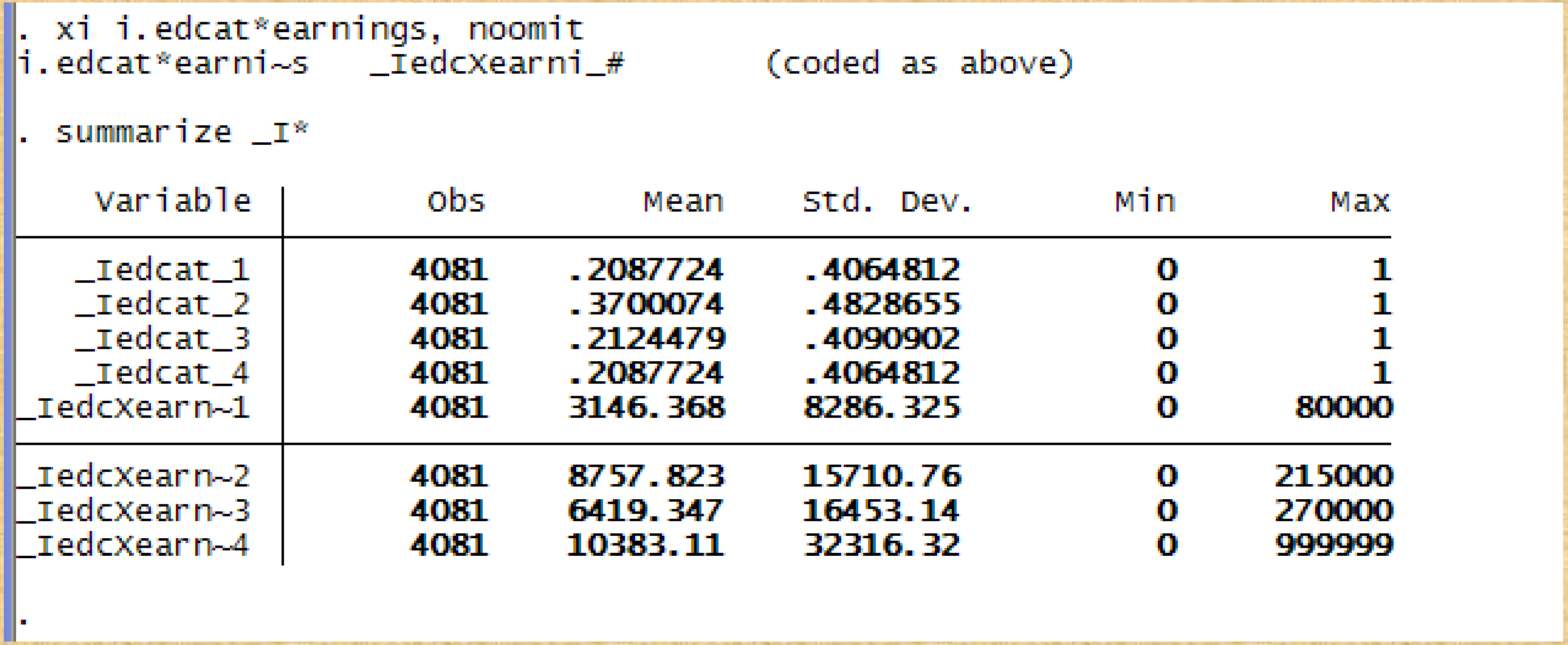
```
.
.
. xi i.edcat
i.edcat           _Iedcat_1-4         (naturally coded; _Iedcat_1 omitted)

.
.
```



```
.
.
. xi i.edcat, noomit

. summarize _I*

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
   _Iedcat_1 |      4081    .2087724    .4064812          0          1
   _Iedcat_2 |      4081    .3700074    .4828655          0          1
   _Iedcat_3 |      4081    .2124479    .4090902          0          1
   _Iedcat_4 |      4081    .2087724    .4064812          0          1

.
```

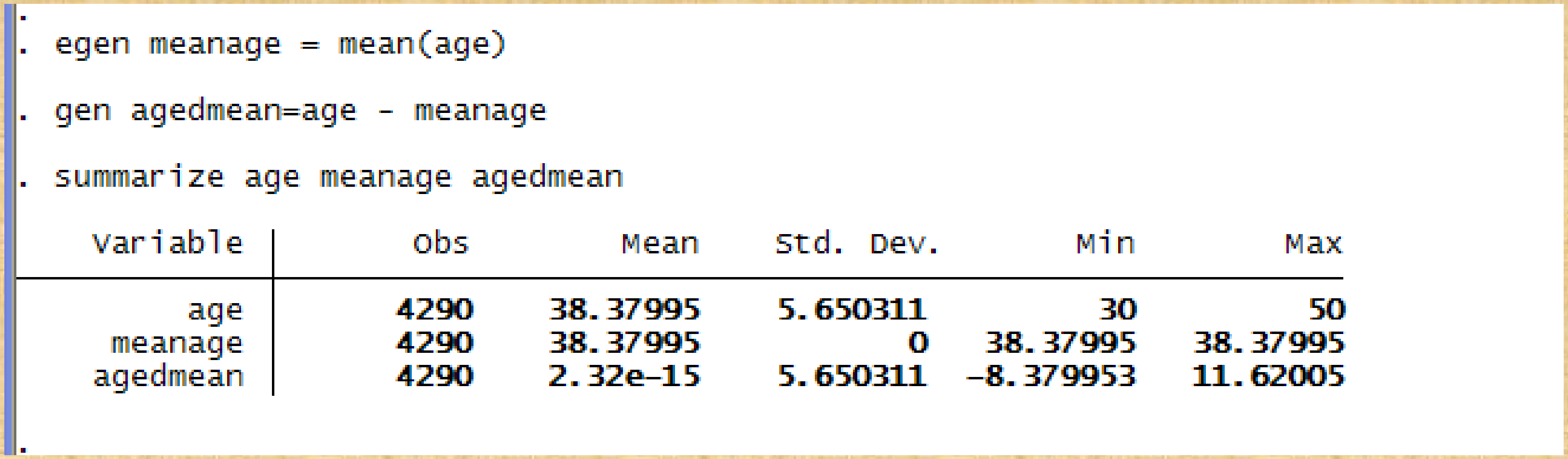
Construct interactions with xi

Cameron and Trivedi, op cit, p49 •



```
. xi i.edcat*earnings, noomit
i.edcat*earni~s   _IedcXearni_#       (coded as above)

. summarize _I*

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
   _Iedcat_1 |      4081    .2087724    .4064812          0          1
   _Iedcat_2 |      4081    .3700074    .4828655          0          1
   _Iedcat_3 |      4081    .2124479    .4090902          0          1
   _Iedcat_4 |      4081    .2087724    .4064812          0          1
_IedcXearn~1 |      4081    3146.368    8286.325          0      80000
-------------+--------------------------------------------------------
_IedcXearn~2 |      4081    8757.823    15710.76          0     215000
_IedcXearn~3 |      4081    6419.347    16453.14          0     270000
_IedcXearn~4 |      4081    10383.11    32316.32          0     999999

.
```

Demeaning variables



```
.
. egen meanage = mean(age)

. gen agedmean=age - meanage

. summarize age meanage agedmean

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
         age |      4290    38.37995    5.650311         30         50
     meanage |      4290    38.37995           0   38.37995   38.37995
    agedmean |      4290    2.32e-15    5.650311  -8.379953   11.62005

.
```

Modeling and Graphing Interactions

- Interactions are defined as the joint effect over and above the main effects.
- Therefore, both main effects must be in the model whether or not they are significant, to properly specify an interaction term.

When one variable is dummy-coded

$$y = a + b_1 x_1 + b_2 x_2 + b_3 x_1 x_2$$

$$y = 65 + 1.7 x_1 + 970 x_2 + -.5 x_1 x_2$$

$$Case\ 1: \ assume\ \ x_2 = dummy\ variable$$

$$for\ example: gender\ is\ coded\quad 0 = male\quad 1 = \ female$$

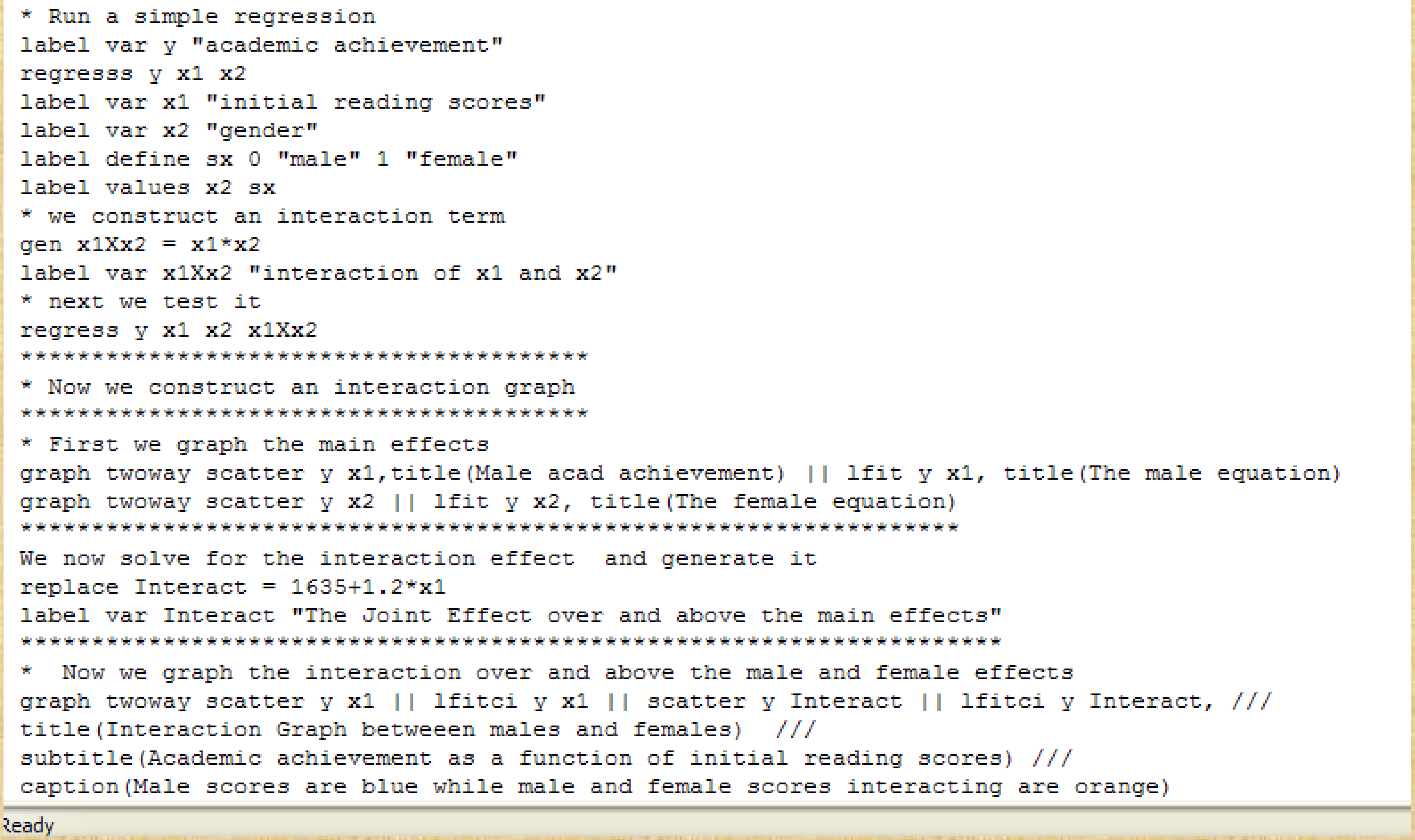
$$male\ equation: \ y = 65 + 1.7 x_1 + 970 * 0 - .5 * 0$$

$$= 65 + 1.7 x_1$$

$$female\ equation: \ y = (65 + 970 * 1) + (1.7 - .5 * 1) x_1$$

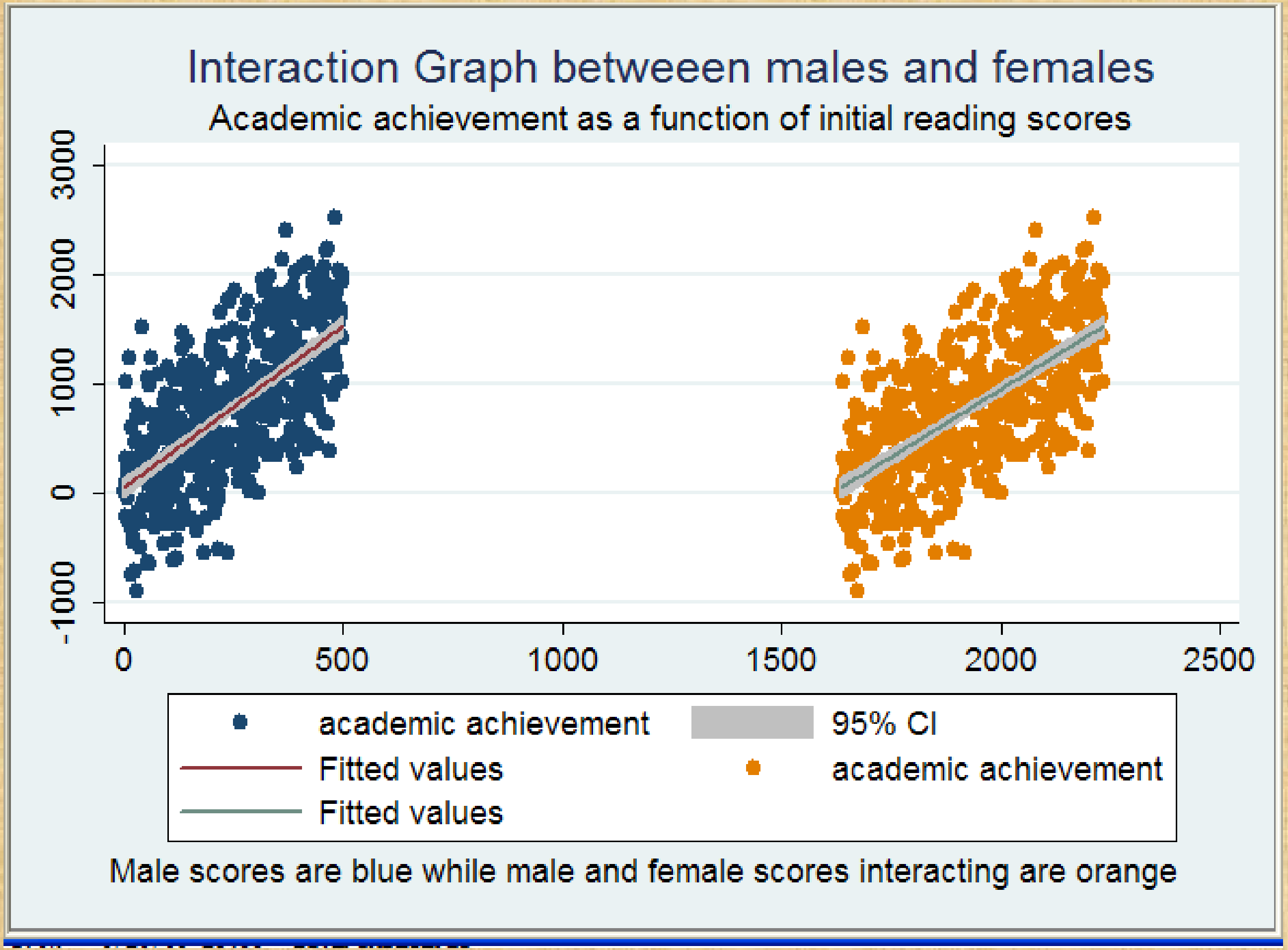
$$= 1635 + 1.2 x_1$$

Stata commands for plotting the interaction



```
* Run a simple regression
label var y "academic achievement"
regresss y x1 x2
label var x1 "initial reading scores"
label var x2 "gender"
label define sx 0 "male" 1 "female"
label values x2 sx
* we construct an interaction term
gen x1Xx2 = x1*x2
label var x1Xx2 "interaction of x1 and x2"
* next we test it
regress y x1 x2 x1Xx2
****************************************
* Now we construct an interaction graph
****************************************
* First we graph the main effects
graph twoway scatter y x1,title(Male acad achievement) || lfit y x1, title(The male equation)
graph twoway scatter y x2 || lfit y x2, title(The female equation)
******************************************************************
We now solve for the interaction effect  and generate it
replace Interact = 1635+1.2*x1
label var Interact "The Joint Effect over and above the main effects"
*********************************************************************
*  Now we graph the interaction over and above the male and female effects
graph twoway scatter y x1 || lfitci y x1 || scatter y Interact || lfitci y Interact, ///
title(Interaction Graph betweeen males and females)  ///
subtitle(Academic achievement as a function of initial reading scores) ///
caption(Male scores are blue while male and female scores interacting are orange)
```

Interaction graph
a non-crossed interaction

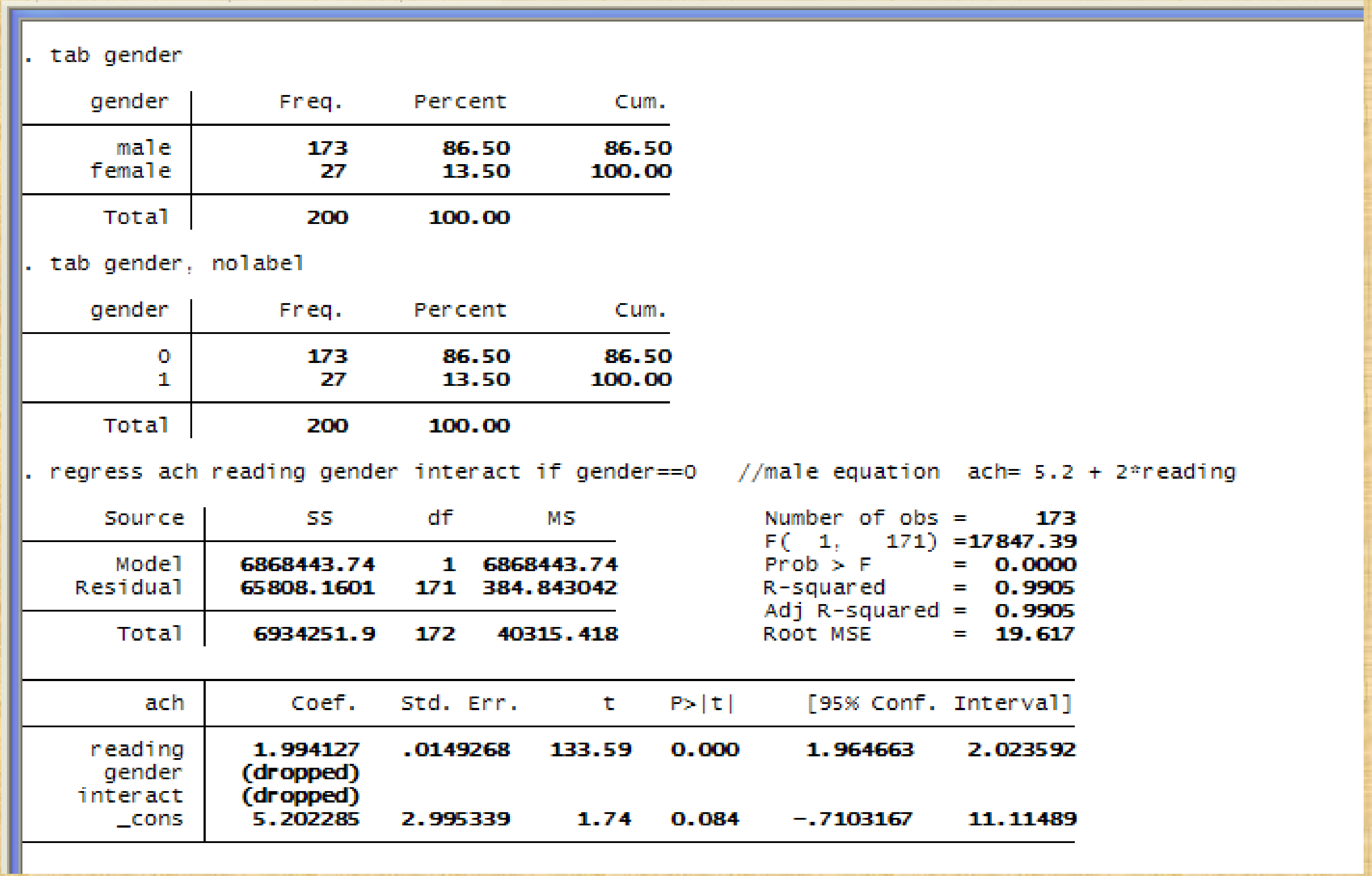


Interaction Graph betweeen males and females

Academic achievement as a function of initial reading scores

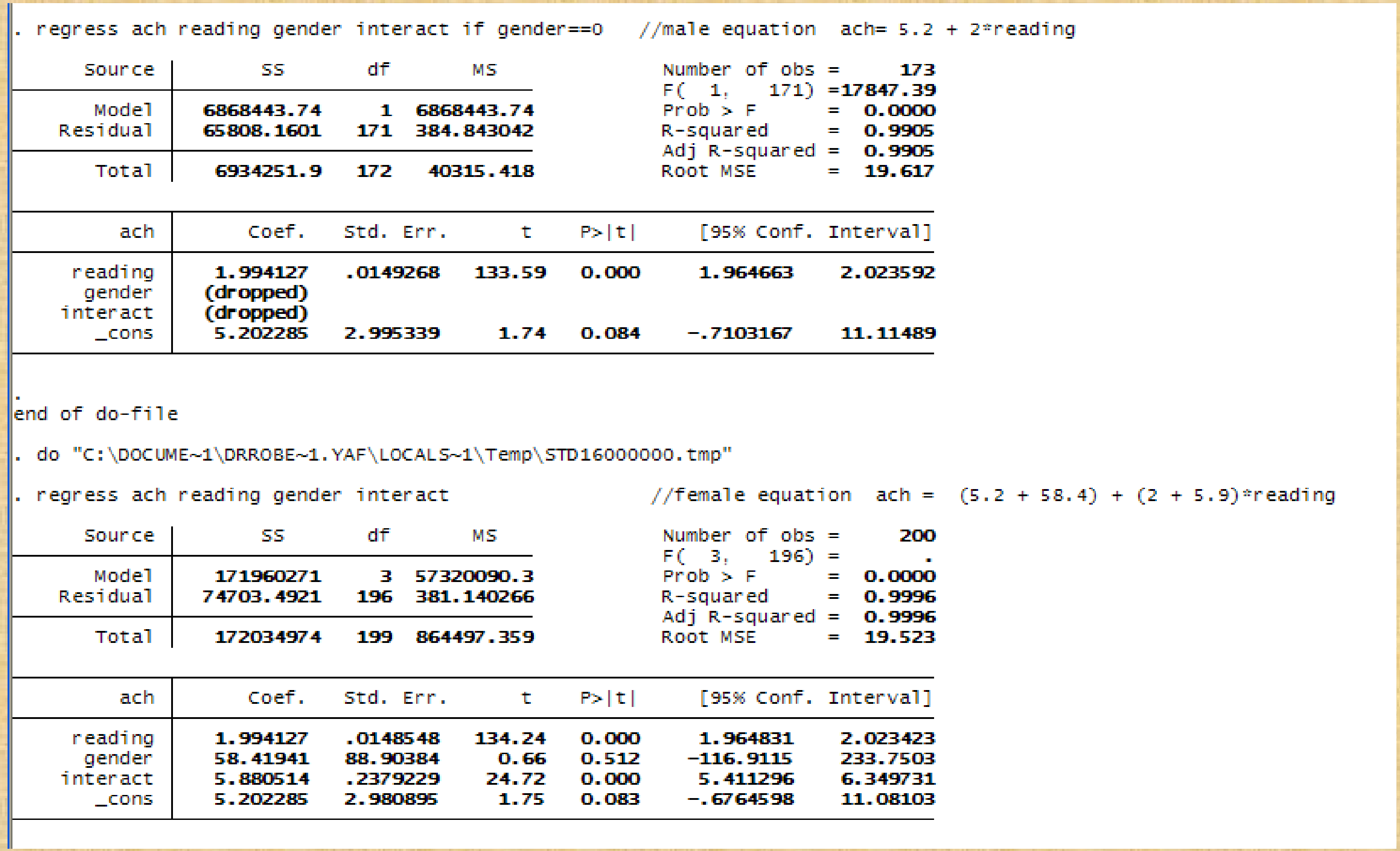
Male scores are blue while male and female scores interacting are orange

Graphing the interaction



```
. tab gender

     gender |      Freq.     Percent        Cum.
------------+-----------------------------------
       male |        173       86.50       86.50
     female |         27       13.50      100.00
------------+-----------------------------------
      Total |        200      100.00

. tab gender, nolabel

     gender |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |        173       86.50       86.50
          1 |         27       13.50      100.00
------------+-----------------------------------
      Total |        200      100.00

. regress ach reading gender interact if gender==0   //male equation  ach= 5.2 + 2*reading

      Source |       SS       df       MS              Number of obs =     173
-------------+------------------------------           F(  1,   171) =17847.39
       Model |  6868443.74     1  6868443.74           Prob > F      =  0.0000
    Residual |  65808.1601   171  384.843042           R-squared     =  0.9905
-------------+------------------------------           Adj R-squared =  0.9905
       Total |   6934251.9   172   40315.418           Root MSE      =  19.617

------------------------------------------------------------------------------
         ach |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     reading |   1.994127   .0149268   133.59   0.000     1.964663    2.023592
      gender |  (dropped)
    interact |  (dropped)
       _cons |   5.202285   2.995339     1.74   0.084    -.7103167    11.11489
------------------------------------------------------------------------------
```

The male and female equations



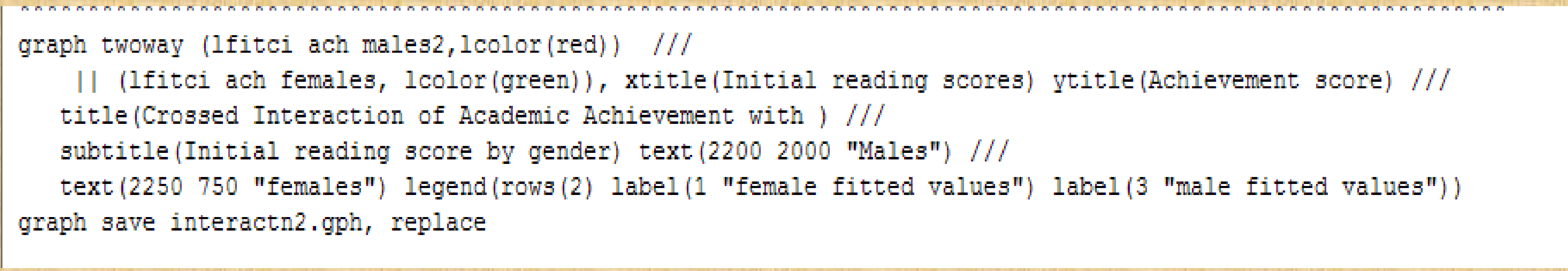
```
. regress ach reading gender interact if gender==0   //male equation  ach= 5.2 + 2*reading

      Source |       SS       df       MS              Number of obs =     173
-------------+------------------------------           F(  1,   171) =17847.39
       Model |  6868443.74     1  6868443.74           Prob > F      =  0.0000
    Residual |  65808.1601   171  384.843042           R-squared     =  0.9905
-------------+------------------------------           Adj R-squared =  0.9905
       Total |   6934251.9   172   40315.418           Root MSE      =  19.617

------------------------------------------------------------------------------
         ach |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     reading |   1.994127   .0149268   133.59   0.000     1.964663    2.023592
      gender |  (dropped)
    interact |  (dropped)
       _cons |   5.202285   2.995339     1.74   0.084    -.7103167    11.11489
------------------------------------------------------------------------------

.
end of do-file

. do "C:\DOCUME~1\DRROBE~1.YAF\LOCALS~1\Temp\STD16000000.tmp"

. regress ach reading gender interact                  //female equation  ach =  (5.2 + 58.4) + (2 + 5.9)*reading

      Source |       SS       df       MS              Number of obs =     200
-------------+------------------------------           F(  3,   196) =       .
       Model |   171960271     3  57320090.3           Prob > F      =  0.0000
    Residual |  74703.4921   196  381.140266           R-squared     =  0.9996
-------------+------------------------------           Adj R-squared =  0.9996
       Total |   172034974   199  864497.359           Root MSE      =  19.523

------------------------------------------------------------------------------
         ach |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     reading |   1.994127   .0148548   134.24   0.000     1.964831    2.023423
      gender |   58.41941   88.90384     0.66   0.512    -116.9115    233.7503
    interact |   5.880514   .2379229    24.72   0.000     5.411296    6.349731
       _cons |   5.202285   2.980895     1.75   0.083    -.6764598    11.08103
------------------------------------------------------------------------------
```

Modeling the gender effects and its interaction

- Both equations can be inferred from the interaction model with its main effects included.

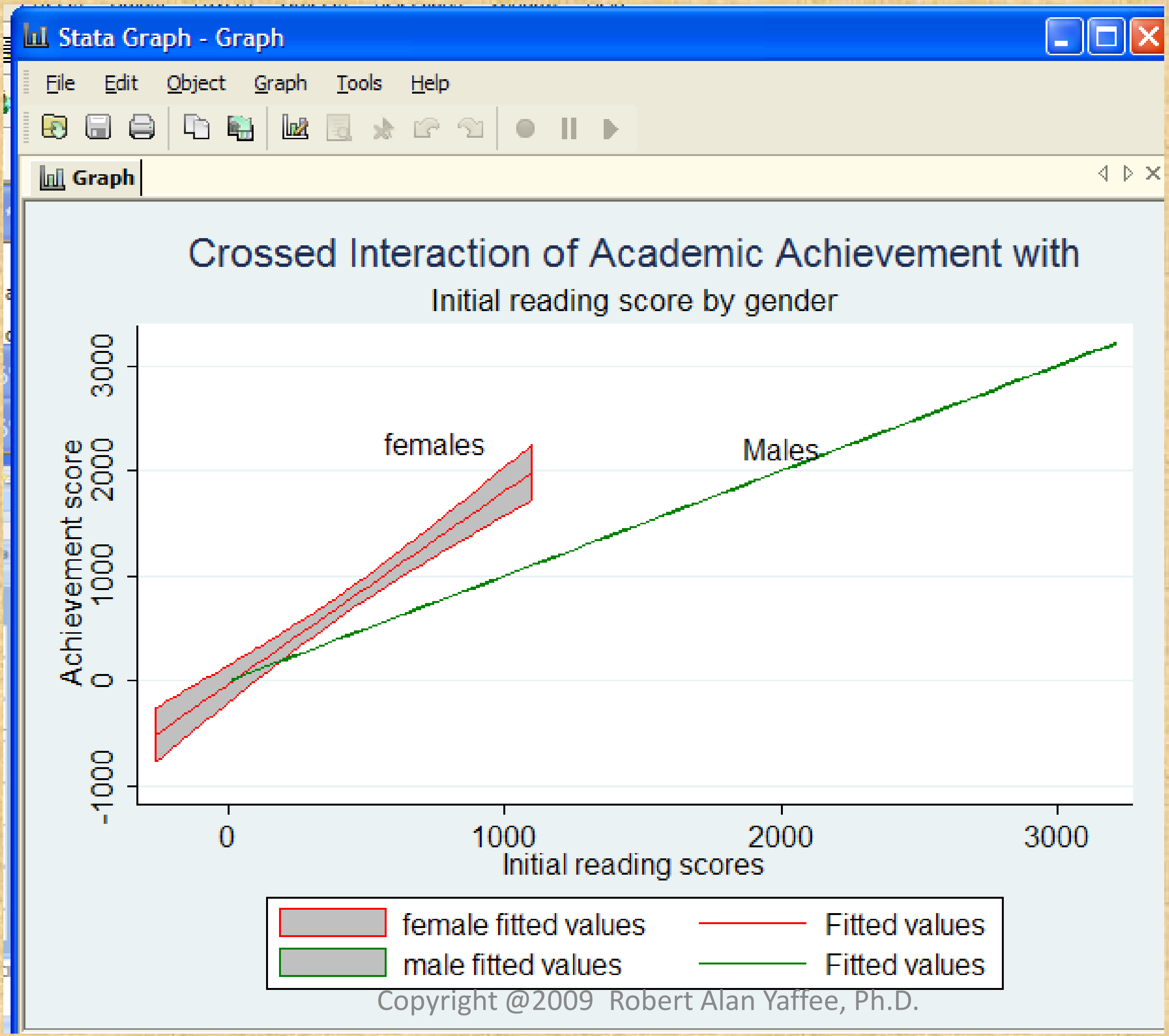
$$\begin{aligned}
&\textit{Male equation}:\\
&Achievement = \mathbf{5.2} + \mathbf{2} * reading + gender + reading * gender \ \ if \ \ gender == \mathbf{0}\\
&\qquad\qquad\qquad = \mathbf{5.2} + \mathbf{2} * reading + \ \mathbf{58.4} * \mathbf{0} + \mathbf{5.88} * reading * \mathbf{0}\\
&\qquad\qquad\qquad = \mathbf{5.2} + \mathbf{2} * reading\\
&\textit{Female equation}:\\
&Achievement \ = \mathbf{5.2} \ + \mathbf{2} * reading + \mathbf{58.4} * \mathbf{1} \ + \mathbf{5.88} * reading * \mathbf{1} \ if \ gender == \mathbf{1}\\
&\qquad\qquad\qquad = (\mathbf{5.2} + \mathbf{58.4}) \ + \mathbf{8} * reading
\end{aligned}$$

Commands for generating the first order linear interaction graph



```
graph twoway (lfitci ach males2,lcolor(red))  ///
    || (lfitci ach females, lcolor(green)), xtitle(Initial reading scores) ytitle(Achievement score) ///
   title(Crossed Interaction of Academic Achievement with ) ///
   subtitle(Initial reading score by gender) text(2200 2000 "Males") ///
   text(2250 750 "females") legend(rows(2) label(1 "female fitted values") label(3 "male fitted values"))
graph save interactn2.gph, replace
```

Graph of the gender by reading interaction for academic achievement



Copyright @2009 Robert Alan Yaffee, Ph.D.

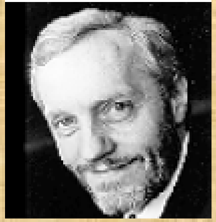
When 2 variables are continuous

- Split one at the mean.
- Cut it off 1 sd above and 1 sd below the mean.
- Run the regression for all three portions.
- You will get a different regression line for each
- Then plot those regression lines

Modeling strategies

- Hierarchical regression (Jack and Pat Cohen popularized this approach sequential set inclusion, not multilevel modeling)
  - From specific to general
  - Two levels of analysis

Sir David F. Hendry



- Stepwise regression
  - Problems with it.
- General-to-specific modeling
  - Specification error can bias results more than
    - multicollinearity
  - Avoidance of specification error

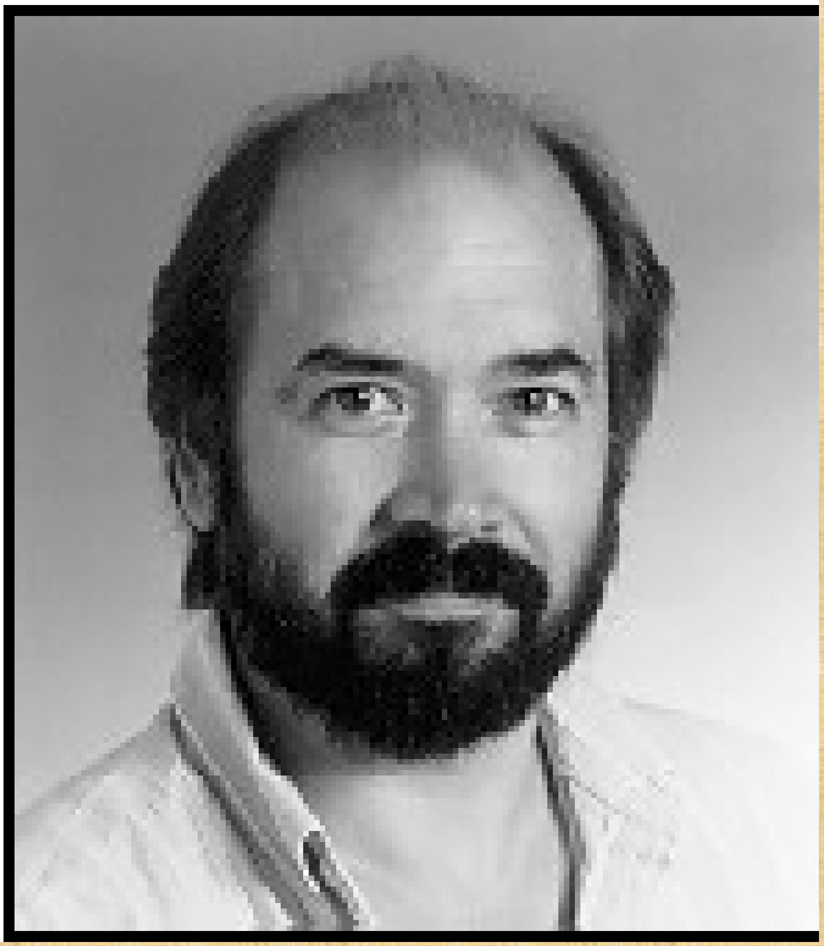
Robust regression

- Outlier diagnosis
  - Outlier downweighting
- White estimators
- Weighted Least Squares for heteroscedastic correction
- Median regression
- Quantile regression
- Bootstrapped regression for empirical standard errors

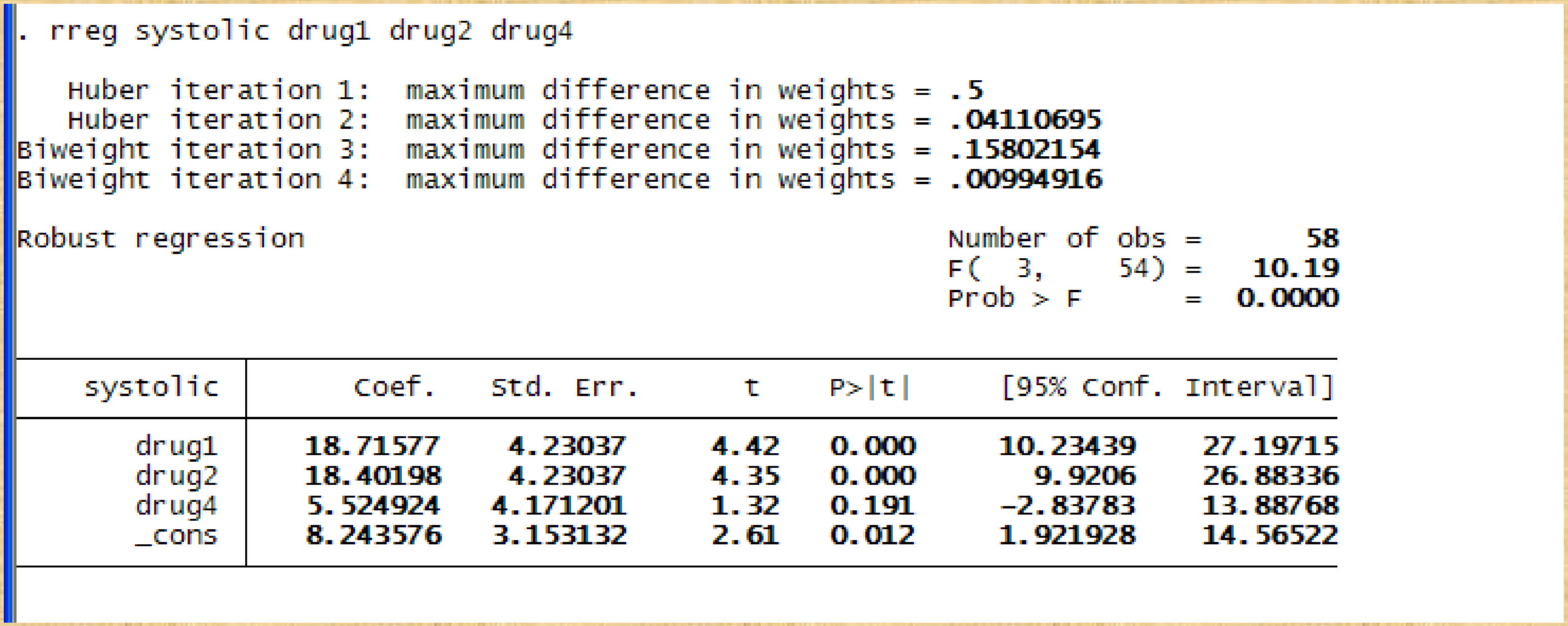
Halbert White

Father of the Sandwich Variance (White) estimator

This variance estimator is robust to moderate violations of heteroscedasticity when the sample gets large.

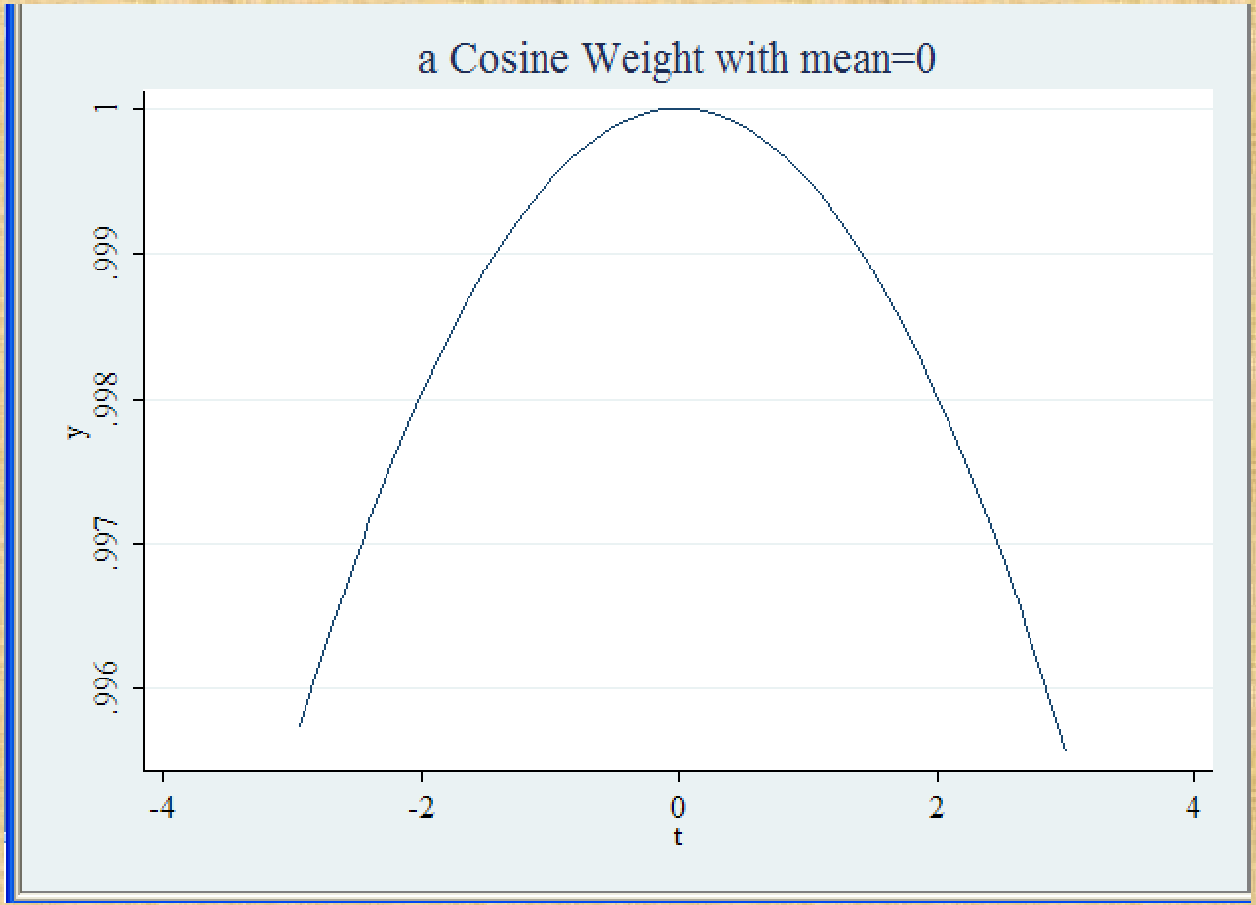


Robust regression with outlier downweighting

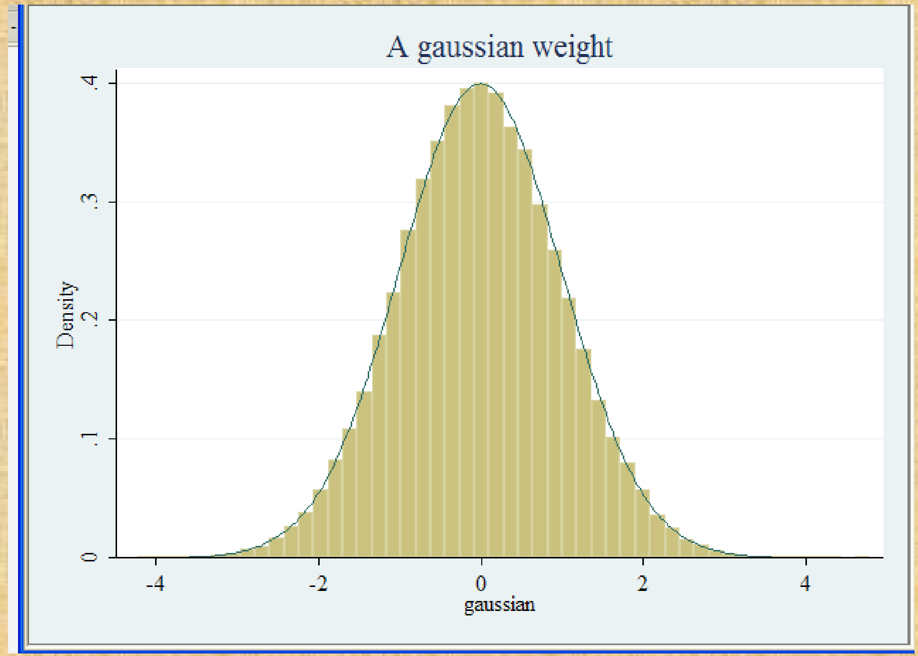


```
. rreg systolic drug1 drug2 drug4

   Huber iteration 1:  maximum difference in weights = .5
   Huber iteration 2:  maximum difference in weights = .04110695
Biweight iteration 3:  maximum difference in weights = .15802154
Biweight iteration 4:  maximum difference in weights = .00994916

Robust regression                                      Number of obs =      58
                                                       F(  3,     54) =   10.19
                                                       Prob > F      =  0.0000

------------------------------------------------------------------------------
    systolic |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       drug1 |   18.71577    4.23037     4.42   0.000     10.23439    27.19715
       drug2 |   18.40198    4.23037     4.35   0.000       9.9206    26.88336
       drug4 |   5.524924   4.171201     1.32   0.191     -2.83783    13.88768
       _cons |   8.243576   3.153132     2.61   0.012     1.921928    14.56522
------------------------------------------------------------------------------
```

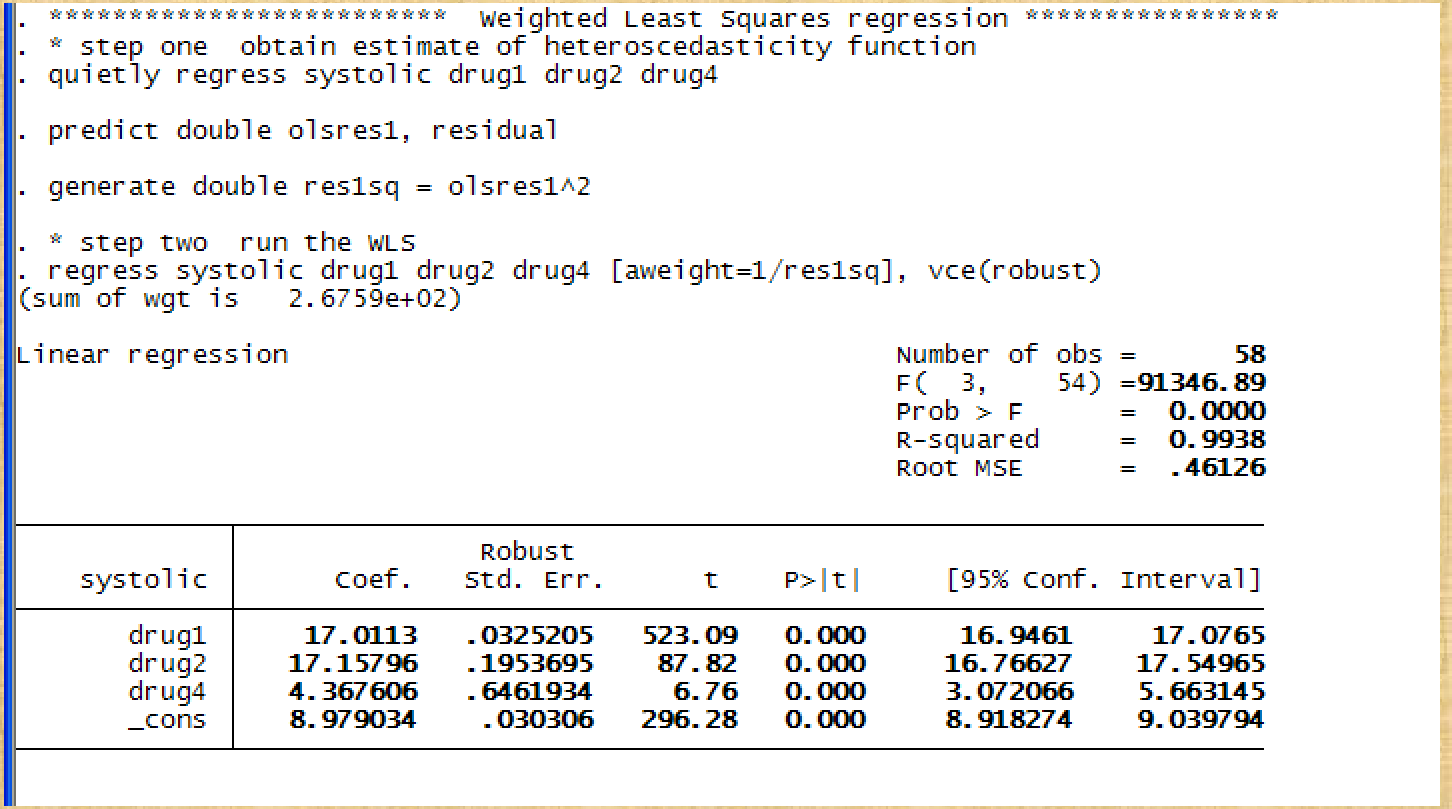
Amount of weight given to an observation given distance t from mean of bandwidth



A Gaussian weight



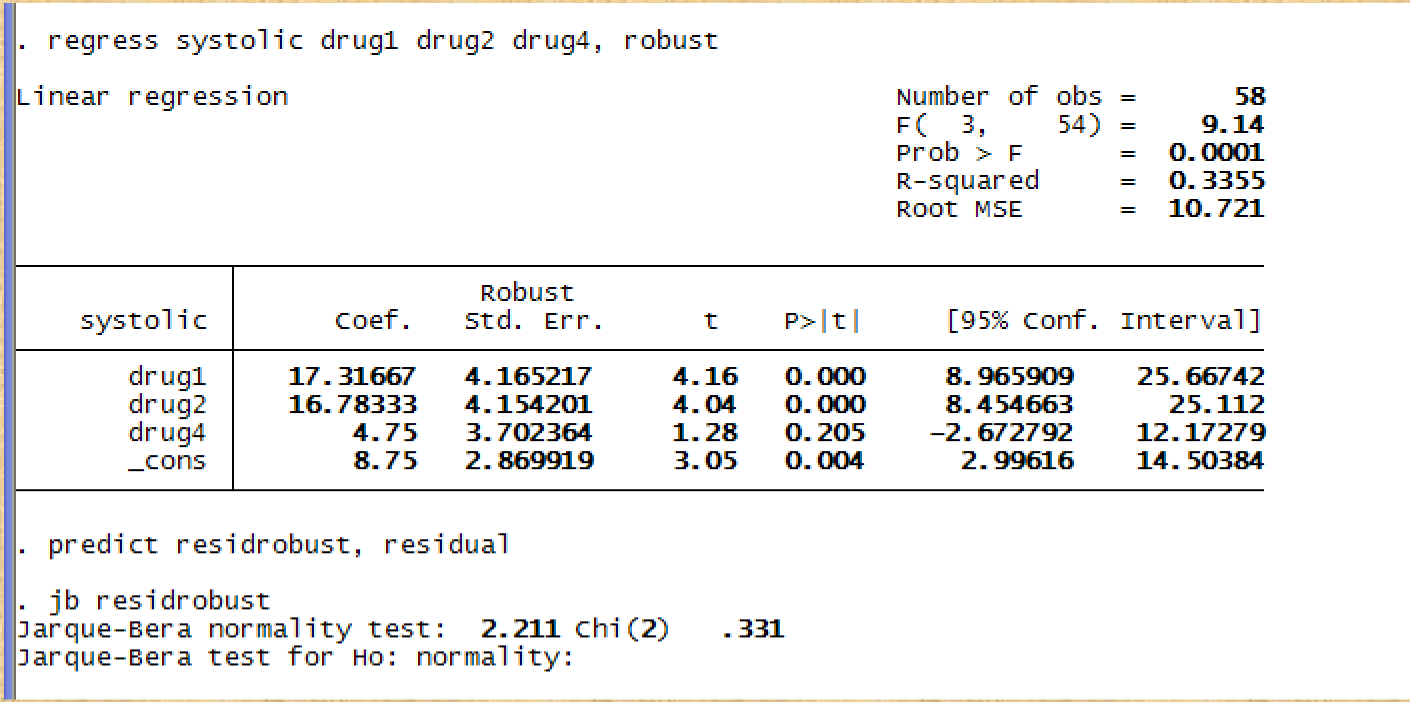
Weighted Least squares regression



```
. *************************  Weighted Least Squares regression ****************
. * step one  obtain estimate of heteroscedasticity function
. quietly regress systolic drug1 drug2 drug4

. predict double olsres1, residual

. generate double res1sq = olsres1^2

. * step two  run the WLS
. regress systolic drug1 drug2 drug4 [aweight=1/res1sq], vce(robust)
(sum of wgt is   2.6759e+02)

Linear regression                                      Number of obs =      58
                                                       F(  3,    54) =91346.89
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.9938
                                                       Root MSE      =  .46126

------------------------------------------------------------------------------
             |               Robust
    systolic |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       drug1 |    17.0113   .0325205   523.09   0.000      16.9461     17.0765
       drug2 |   17.15796   .1953695    87.82   0.000     16.76627    17.54965
       drug4 |   4.367606   .6461934     6.76   0.000     3.072066    5.663145
       _cons |   8.979034    .030306   296.28   0.000     8.918274    9.039794
------------------------------------------------------------------------------
```

Robust regression (heteroscedastically consistent)

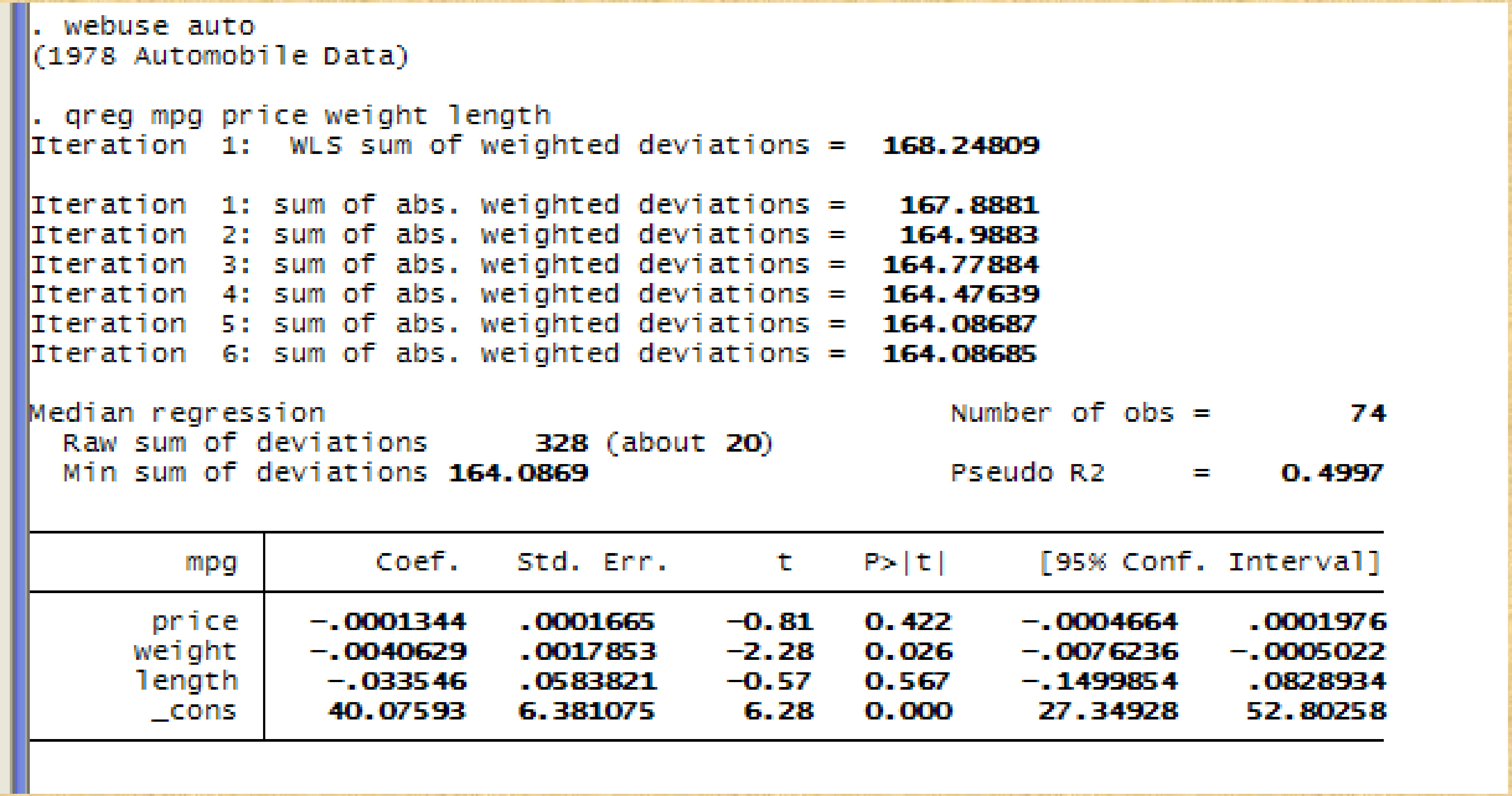
Using a sandwich estimator of the variance developed by Hal White in 1980, which is asymptotically heteroscedastically consistent



```
. regress systolic drug1 drug2 drug4, robust

Linear regression                                      Number of obs =      58
                                                       F(  3,    54) =    9.14
                                                       Prob > F      =  0.0001
                                                       R-squared     =  0.3355
                                                       Root MSE      =  10.721

------------------------------------------------------------------------------
             |               Robust
    systolic |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       drug1 |   17.31667   4.165217     4.16   0.000     8.965909    25.66742
       drug2 |   16.78333   4.154201     4.04   0.000     8.454663      25.112
       drug4 |       4.75   3.702364     1.28   0.205    -2.672792    12.17279
       _cons |       8.75   2.869919     3.05   0.004      2.99616    14.50384
------------------------------------------------------------------------------

. predict residrobust, residual

. jb residrobust
Jarque-Bera normality test:  2.211 Chi(2)  .331
Jarque-Bera test for Ho: normality:
```

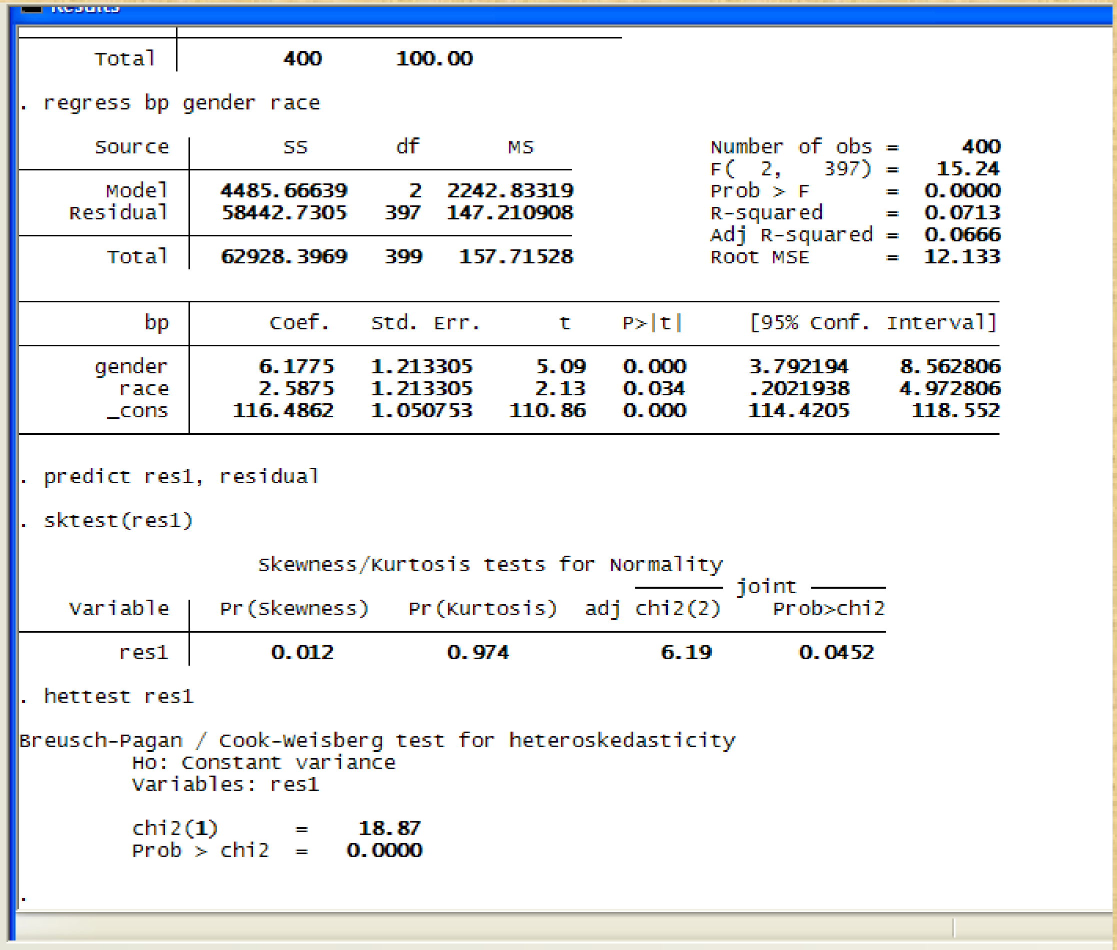
Median regression

predicts the 50th percentile of the dependent variable



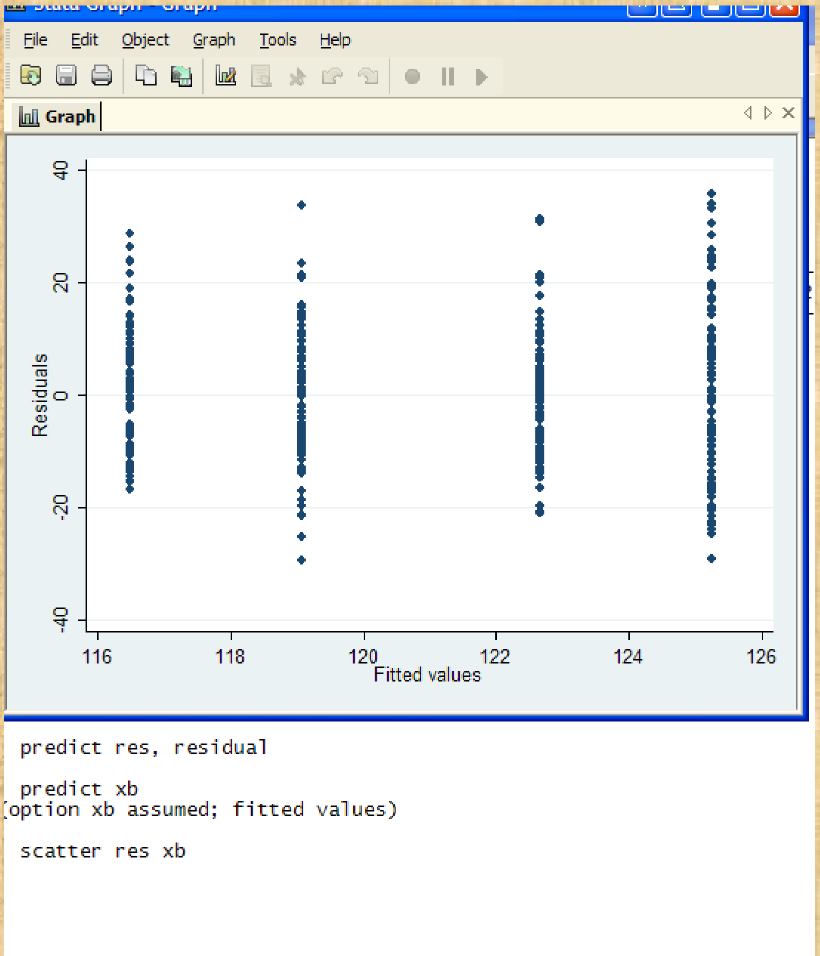
```
. webuse auto
(1978 Automobile Data)

. qreg mpg price weight length
Iteration  1:  WLS sum of weighted deviations =  168.24809

Iteration  1: sum of abs. weighted deviations =   167.8881
Iteration  2: sum of abs. weighted deviations =   164.9883
Iteration  3: sum of abs. weighted deviations =  164.77884
Iteration  4: sum of abs. weighted deviations =  164.47639
Iteration  5: sum of abs. weighted deviations =  164.08687
Iteration  6: sum of abs. weighted deviations =  164.08685

Median regression                                    Number of obs =         74
  Raw sum of deviations      328 (about 20)
  Min sum of deviations 164.0869                     Pseudo R2     =     0.4997

------------------------------------------------------------------------------
         mpg |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       price |  -.0001344   .0001665    -0.81   0.422    -.0004664    .0001976
      weight |  -.0040629   .0017853    -2.28   0.026    -.0076236   -.0005022
      length |   -.033546   .0583821    -0.57   0.567    -.1499854    .0828934
       _cons |   40.07593   6.381075     6.28   0.000     27.34928    52.80258
------------------------------------------------------------------------------
```

Variance weighted least squares regression for severe heteroscedasticity Stata Reference Guide Q-Z(2007), pp 554-559.



```
------------+-----------------------------------
      Total |        400      100.00

. regress bp gender race

      Source |       SS       df       MS              Number of obs =     400
-------------+------------------------------           F(  2,   397) =   15.24
       Model |  4485.66639     2  2242.83319           Prob > F      =  0.0000
    Residual |  58442.7305   397  147.210908           R-squared     =  0.0713
-------------+------------------------------           Adj R-squared =  0.0666
       Total |  62928.3969   399   157.71528           Root MSE      =  12.133

------------------------------------------------------------------------------
          bp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      gender |     6.1775   1.213305     5.09   0.000     3.792194    8.562806
        race |     2.5875   1.213305     2.13   0.034     .2021938    4.972806
       _cons |   116.4862   1.050753   110.86   0.000     114.4205     118.552
------------------------------------------------------------------------------

. predict res1, residual

. sktest(res1)

                   Skewness/Kurtosis tests for Normality
                                                        ------- joint ------
    Variable |  Pr(Skewness)   Pr(Kurtosis)  adj chi2(2)    Prob>chi2
-------------+-------------------------------------------------------
        res1 |      0.012          0.974           6.19         0.0452

. hettest res1

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: res1

         chi2(1)      =    18.87
         Prob > chi2  =   0.0000

.
```

Graphical diagnosis



```
 predict res, residual

 predict xb
(option xb assumed; fitted values)

 scatter res xb
```

Weighting the variables by inverting s^2

- Stata can weight each variable xi by its variance, thus normalizing the effect of the variable by its spread across the line of estimation (prediction).
- Thus, heteroscedasticity is automatically corrected for by this procedure.

$$V = diagonal(s_1^2, s_2^2, \dots, s_k^2)$$

$$where$$

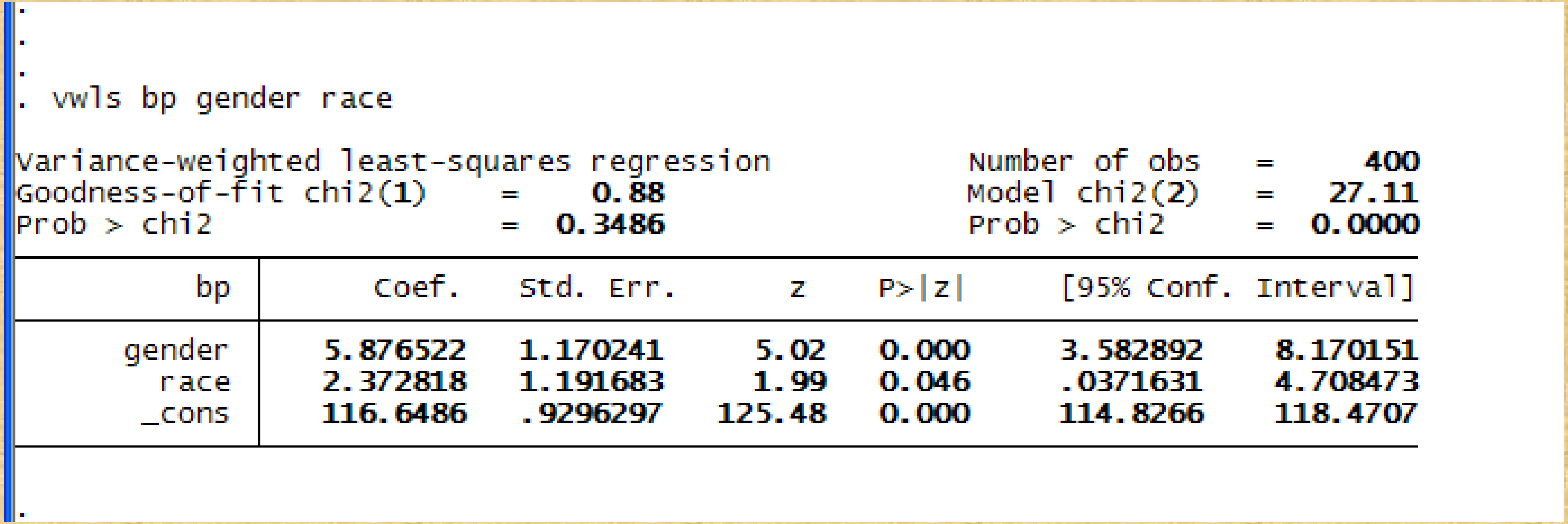
$$k = number\ of\ variables\ (not\ including\ the\ constant)$$

$$s_k = std\ deviation\ of\ variable\ k$$

$$b = (X'V^{-1}X)^{-1}(X'V^{-1}Y)$$

$$Goodness\ of\ fit\ \chi^2_{n-k} = (y - Xb)V^{-1}(y - Xb)$$

Stata command: vwls



```
.
.
.
. vwls bp gender race

Variance-weighted least-squares regression          Number of obs   =      400
Goodness-of-fit chi2(1)     =     0.88              Model chi2(2)   =    27.11
Prob > chi2                 =   0.3486              Prob > chi2     =   0.0000

------------------------------------------------------------------------------
          bp |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      gender |   5.876522   1.170241     5.02   0.000     3.582892    8.170151
        race |   2.372818   1.191683     1.99   0.046     .0371631    4.708473
       _cons |   116.6486   .9296297   125.48   0.000     114.8266    118.4707
------------------------------------------------------------------------------

.
```

Bootstrap Methods (Efron, B.) Cameron and Trivedi, 417.

- Resampling methods.
  - Saving the means for repeated samples.
  - Obtain a sampling distribution of means.

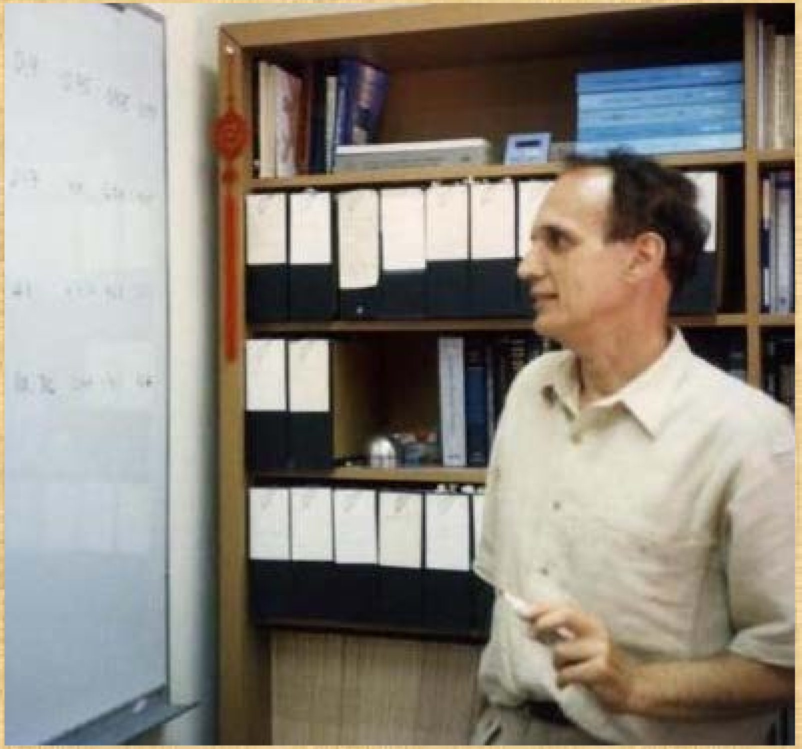
$$With\ 400\ samples,\ B = 400.$$

$$\bar{\hat{\theta}}^{*} = \frac{\sum_{b=1}^{B} \hat{\theta}_b^{*}}{B} = mean\ \ of\ \ bootstrap$$

$$Var(\theta_{boot}) = \frac{\sum_{b=1}^{B} (\hat{\theta}_b^{*} - \bar{\hat{\theta}}^{*})^2}{B - 1}$$

$$SE_{boot} = \sqrt{Var(\theta_{boot})}$$

Bradley Efron (Stanford University developed bootstrapping



Bootstrap confidence intervals

- 95% confidence intervals

$$\mathbf{95}\%\ confidence\ intervals = \hat{\theta}_b^* \pm \mathbf{1.96}\sqrt{Var_{boot}}$$

Bootstrap estimate of bias

- Suppose that the $\hat{\Theta}$ estimator of θ is biased:

$$bias = \bar{\hat{\theta}}_b^* - \hat{\theta}$$

$$where$$

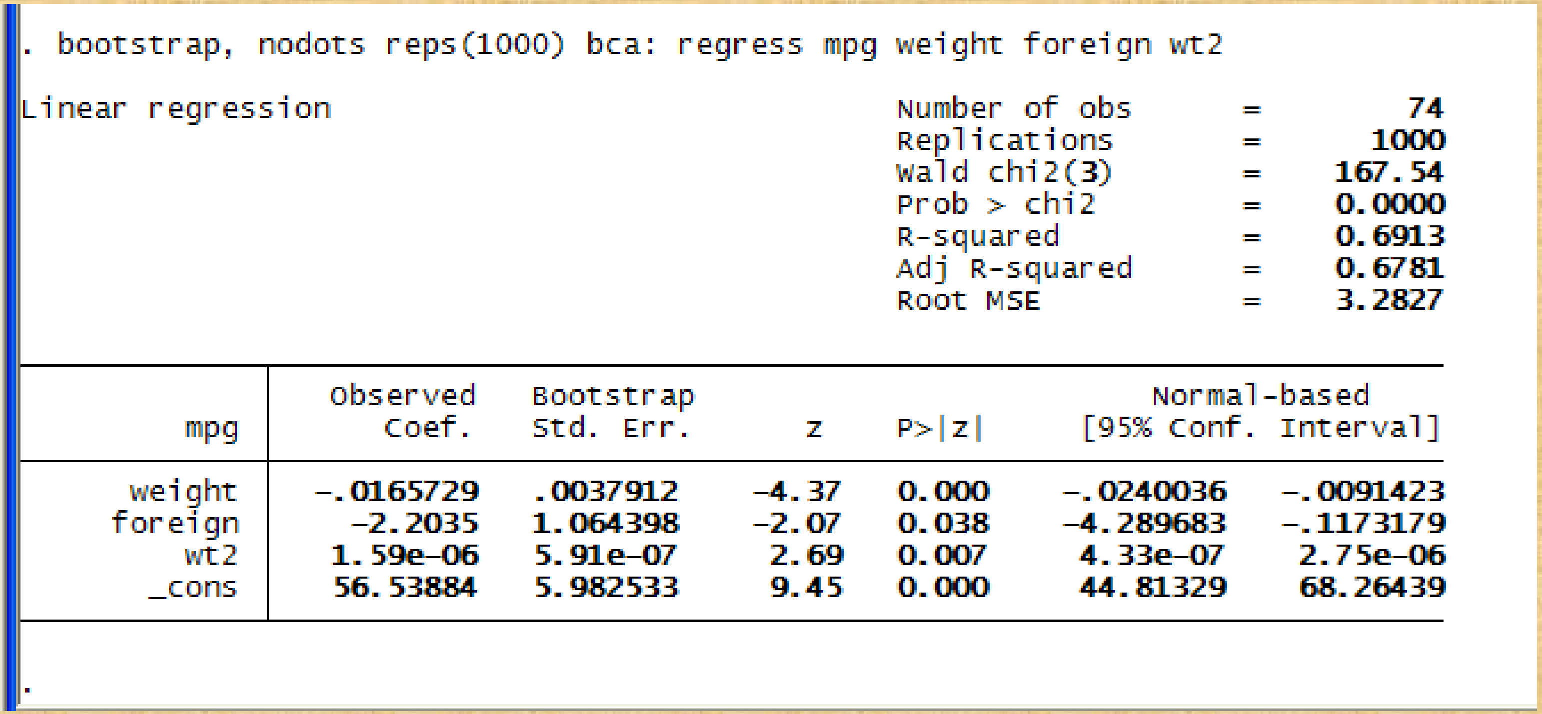
$$\hat{\theta} = DGP \ value$$

$$\bar{\hat{\theta}}_b^* = mean \ of \ the \ estimator, \ given \ the \ DGP \, value$$

How many bootstraps are needed?

- Efron and Tibshirani(1993), B=50 is good enough and very seldom are more than 200 needed.
- Cameron and Trivedi suggest 400. When I read Efron, I recall the number 10000 seems to be the number of replications needed.

Bootstrapped Regression



```
. bootstrap, nodots reps(1000) bca: regress mpg weight foreign wt2

Linear regression                               Number of obs      =        74
                                                Replications       =      1000
                                                Wald chi2(3)       =    167.54
                                                Prob > chi2        =    0.0000
                                                R-squared          =    0.6913
                                                Adj R-squared      =    0.6781
                                                Root MSE           =    3.2827

------------------------------------------------------------------------------
             |   Observed   Bootstrap                         Normal-based
         mpg |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      weight |  -.0165729    .0037912    -4.37   0.000    -.0240036   -.0091423
     foreign |    -2.2035    1.064398    -2.07   0.038    -4.289683   -.1173179
         wt2 |   1.59e-06    5.91e-07     2.69   0.007     4.33e-07    2.75e-06
       _cons |   56.53884    5.982533     9.45   0.000     44.81329    68.26439
------------------------------------------------------------------------------

.
```

BCA option

- Bias correction: Corrects for bias in the bootstrap.
- Acceleration: allows for more asymmetric distributions.

Poisson count models

- **" Much of the world is distributed lognormally, "   E. Tufte.**
  - when the dependent variable is an integer or a rare event.
  - Disadvantage with this model is that it assumes that the mean= variance.

$$\text{Poisson model}:$$

$$\mu = e^{(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p)} \quad \text{or} \quad \ln(\mu) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

$$\text{Poisson distribution}:$$

$$Prob(y \mid \mu) = \frac{e^{-\mu} \mu^{y}}{y!}, \quad \text{for } y = 0, 1, 2, \ldots$$

where

μ = *expected count* *and* μ *is called the rate parameter*

and when dealing with 1 time frame (the predicted # events)

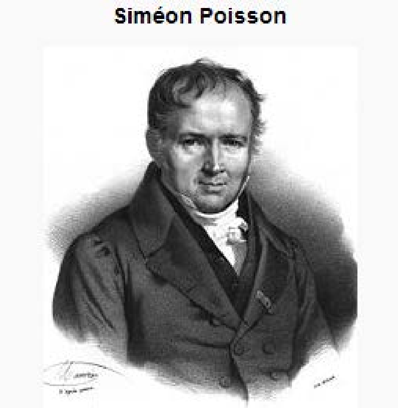
y = *observed count*

assumes $\mu > 0$,

μ = *mean* = *variance.*

Poisson regression model

- named after Simeon–Denis Poisson, who discovered the distribution on which this was based.


Siméon Poisson

$$\textit{Poisson regression}$$

$$\textit{standard model}: \quad \ln(E(Y)) = a + bx + e$$

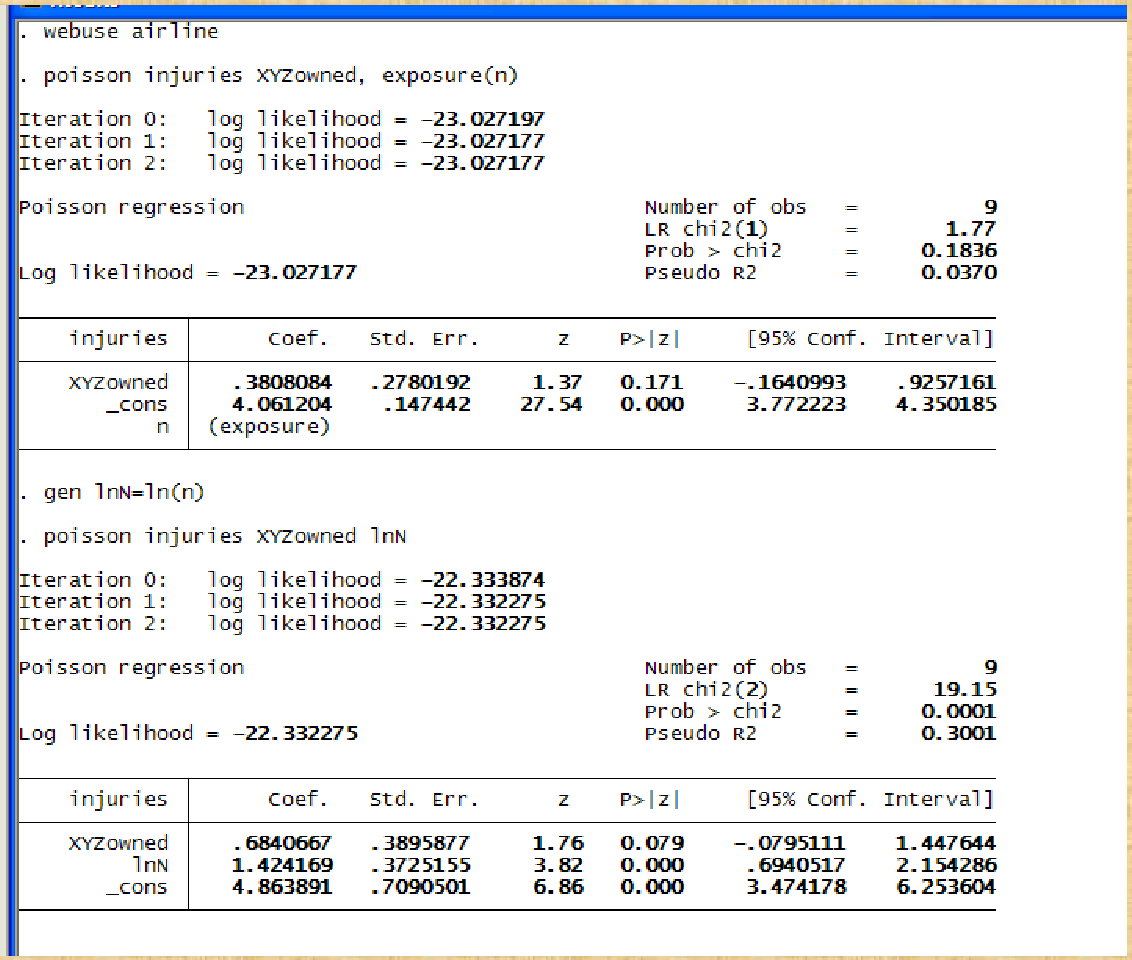
$$\textit{Poisson regression}$$

$$\textit{rate model}: \quad \ln(E(Y)) - \ln(exposure) = \ln\left(\frac{E(Y)}{exposure}\right) = a + bx$$

$$\ln(E(Y)) = \ln(exposure) + \ln\left(\frac{E(Y)}{exposure}\right) = a + bx$$

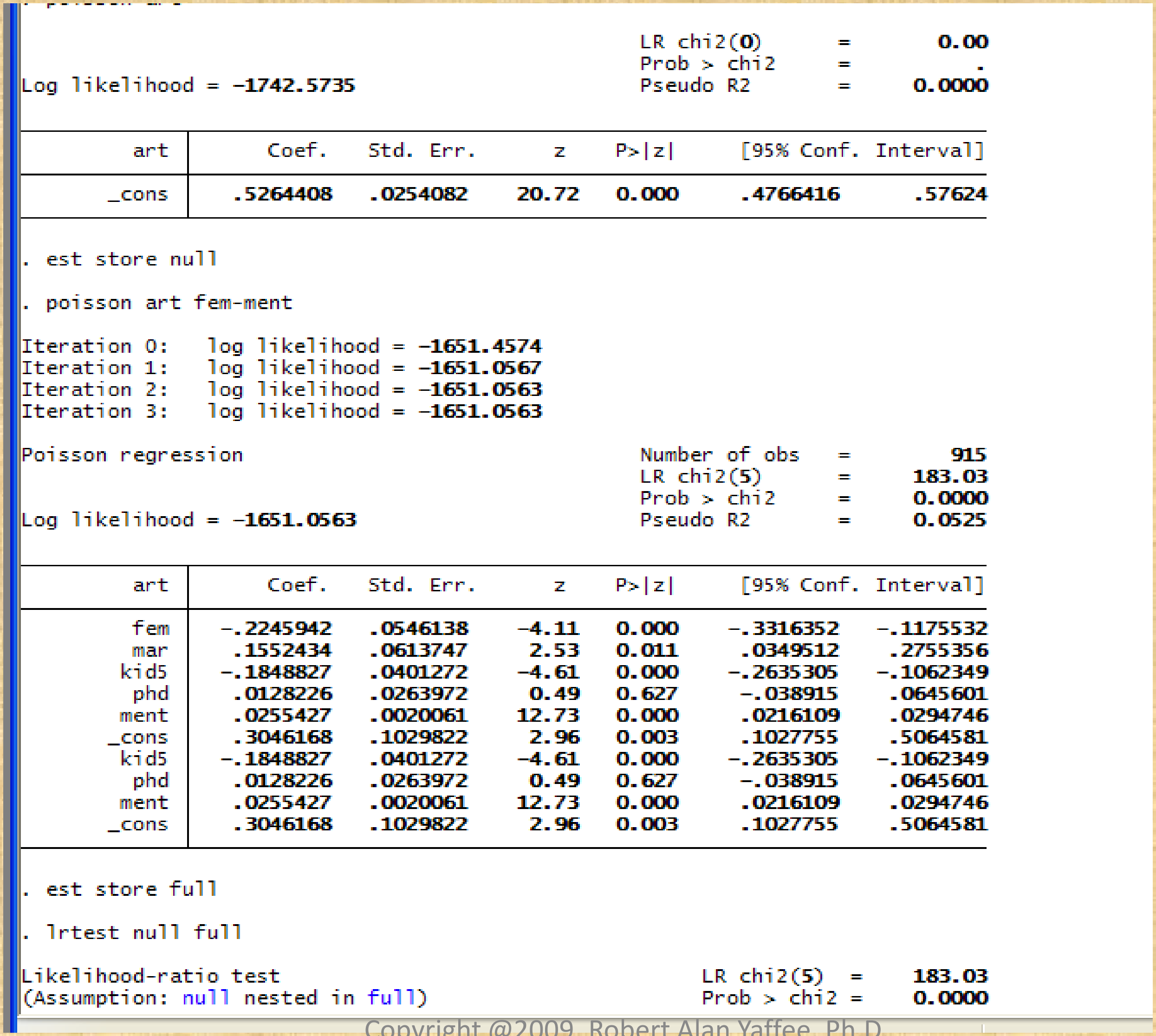
The offset

Poisson count models



```
. webuse airline

. poisson injuries XYZowned, exposure(n)

Iteration 0:   log likelihood = -23.027197
Iteration 1:   log likelihood = -23.027177
Iteration 2:   log likelihood = -23.027177

Poisson regression                                Number of obs   =          9
                                                  LR chi2(1)      =       1.77
                                                  Prob > chi2     =     0.1836
Log likelihood = -23.027177                       Pseudo R2       =     0.0370

------------------------------------------------------------------------------
    injuries |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    XYZowned |   .3808084   .2780192     1.37   0.171    -.1640993    .9257161
       _cons |   4.061204    .147442    27.54   0.000     3.772223    4.350185
           n |  (exposure)
------------------------------------------------------------------------------

. gen lnN=ln(n)

. poisson injuries XYZowned lnN

Iteration 0:   log likelihood = -22.333874
Iteration 1:   log likelihood = -22.332275
Iteration 2:   log likelihood = -22.332275

Poisson regression                                Number of obs   =          9
                                                  LR chi2(2)      =      19.15
                                                  Prob > chi2     =     0.0001
Log likelihood = -22.332275                       Pseudo R2       =     0.3001

------------------------------------------------------------------------------
    injuries |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    XYZowned |   .6840667   .3895877     1.76   0.079    -.0795111    1.447644
         lnN |   1.424169   .3725155     3.82   0.000     .6940517    2.154286
       _cons |   4.863891   .7090501     6.86   0.000     3.474178    6.253604
------------------------------------------------------------------------------
```

Comparing Poisson models



```
                                                  LR chi2(0)      =       0.00
                                                  Prob > chi2     =          .
Log likelihood = -1742.5735                       Pseudo R2       =     0.0000

------------------------------------------------------------------------------
         art |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _cons |   .5264408   .0254082    20.72   0.000     .4766416      .57624
------------------------------------------------------------------------------

. est store null

. poisson art fem-ment

Iteration 0:   log likelihood = -1651.4574
Iteration 1:   log likelihood = -1651.0567
Iteration 2:   log likelihood = -1651.0563
Iteration 3:   log likelihood = -1651.0563

Poisson regression                                Number of obs   =        915
                                                  LR chi2(5)      =     183.03
                                                  Prob > chi2     =     0.0000
Log likelihood = -1651.0563                       Pseudo R2       =     0.0525

------------------------------------------------------------------------------
         art |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         fem |  -.2245942   .0546138    -4.11   0.000    -.3316352   -.1175532
         mar |   .1552434   .0613747     2.53   0.011     .0349512    .2755356
        kid5 |  -.1848827   .0401272    -4.61   0.000    -.2635305   -.1062349
         phd |   .0128226   .0263972     0.49   0.627     -.038915    .0645601
        ment |   .0255427   .0020061    12.73   0.000     .0216109    .0294746
       _cons |   .3046168   .1029822     2.96   0.003     .1027755    .5064581
        kid5 |  -.1848827   .0401272    -4.61   0.000    -.2635305   -.1062349
         phd |   .0128226   .0263972     0.49   0.627     -.038915    .0645601
        ment |   .0255427   .0020061    12.73   0.000     .0216109    .0294746
       _cons |   .3046168   .1029822     2.96   0.003     .1027755    .5064581
------------------------------------------------------------------------------

. est store full

. lrtest null full

Likelihood-ratio test                                  LR chi2(5)  =    183.03
(Assumption: null nested in full)                      Prob > chi2 =    0.0000
```

Poisson model assumptions

- Observations are independent
- Measures integers (counts) of events
- No multicollinearity
- Residuals are skewed; in OLS they are symmetric.
- Variance increases as the mean increases whereas in traditional regression models the variance is constant.
- Overdispersion (the variance is larger than the mean) for any number of cases does not exist.
    - If it occurs, there are corrections for it that can be applied as in the next model we discuss.

Test for overdispersion

Cameron and Trivedi Microeconomics using Stata (p.561)

- Overdispersion is where the conditional variance is greater than the conditional mean.
- If you have a random variable with measurement error, v, you could have an error such as uv instead of just u. If E(v)=1, it would preserve the mean but increase the variance (for logged dependent variables).
- $E(y)=\mu$, $Var(y)= \mu(1+ \mu\sigma^2)$

Test for Overdispersion--continued

- Test = if $x = (\text{resid}^2 - \text{dv})/\text{dv}$
- We regress dv on x, and if it is significant, there is overdispersion.
- If $p(b) < 0.05$, then we use the negative binomial model.

Negative binomial Models
nbreg models

- Also used for count data
- The mean does not have to equal the variance
  - Stata has this for zero-inflated and regular
  - It has it in the complex survey module as well as in the regular options.
  - It fits both the Poisson and the Negative binomial regression.
- A Poisson likelihood with a gamma prior (for all the Bayesians)

Assumptions of the negative binomial

- no multicollinearity
- Overdispersion is permitted here
- The Poisson parameter is itself a gamma distribution.

Binary dependent variable models

- The probit regression model
  - Assumes an underlying latent variable that is normally distributed. The proportions determine the cut-point in the normal distribution. If the value is greater than the cut-point, the respondent gets a 1, otherwise his value is scored as a zero.
- The logistic regression model
  - Uses the natural log of the odds ratio (the logit) as the dependent variable.
  - Odds ratio = prob(event)/(1-prob(event))

Logistic regression

- Formula for logistic regression:

Regression models for binary dependent variables

$$odds = \frac{prob}{\mathbf{1} - prob}$$

$$odds = e^{(a + b_1 x_1 + b_2 x_2 + \ldots + b_p x_p)}$$

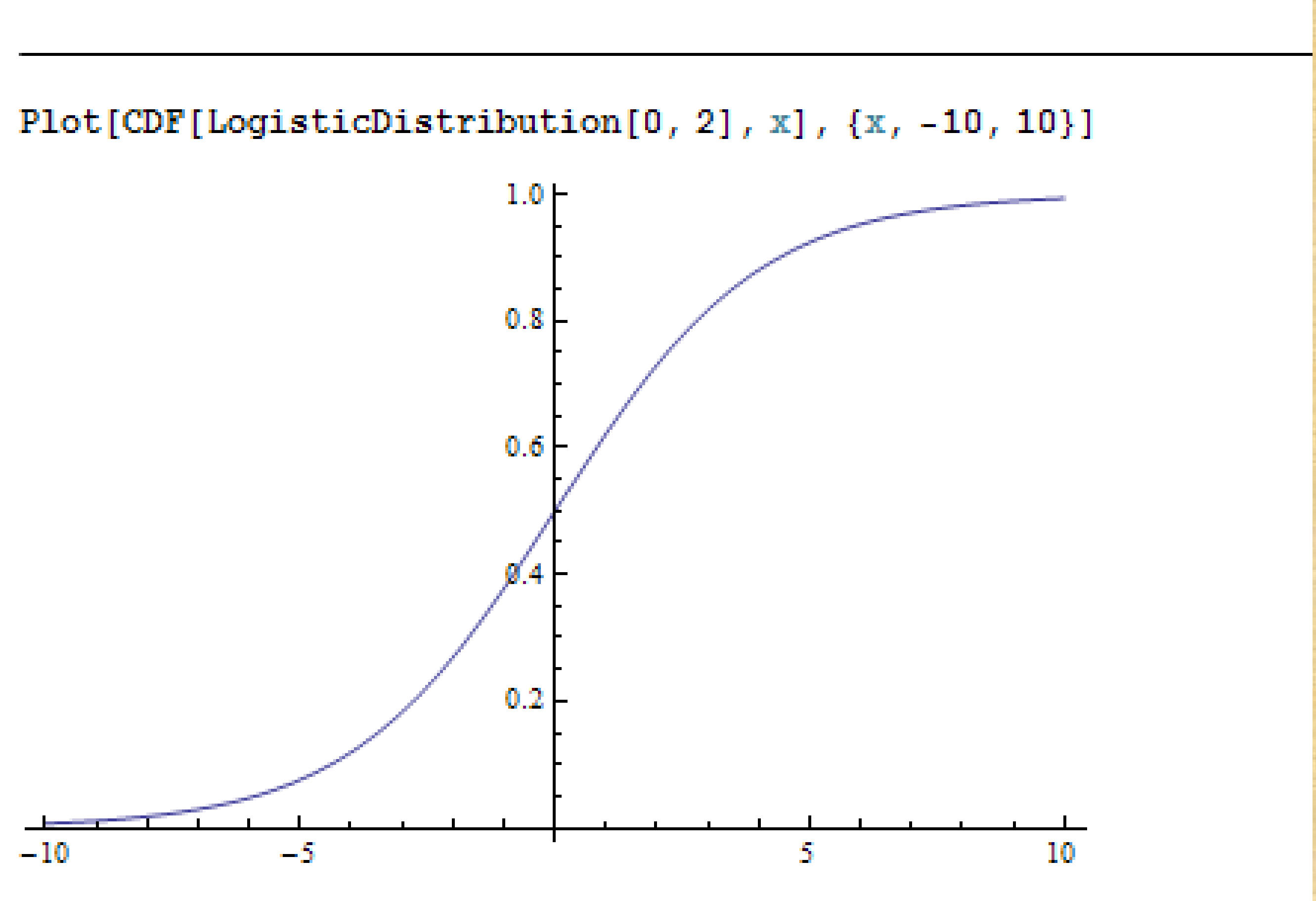
$$logit = ln(odds)$$

$$logit = a + b_1 x_1 + b_2 x_2 + \ldots + b_p x_p$$

Converting an odds to a probability

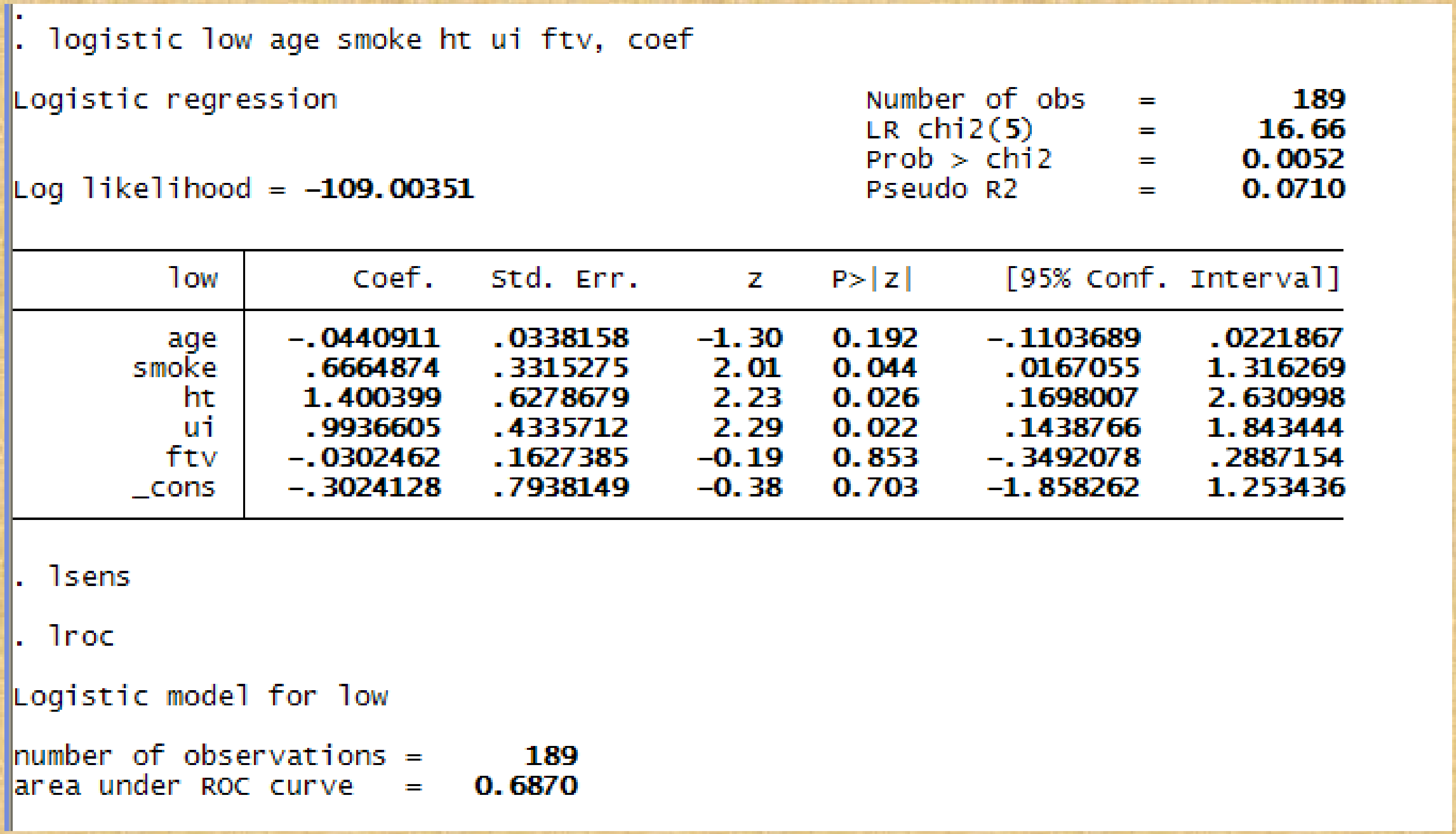
- Odds= prob/(1-prob)
- Odds(1-prob)=prob
- Odds- Odds*prob= prob
- Prob = Odds/(1+Odds)
- Prob = e^(X'B)/[1+ e^(X'B)]
- Prob = $e^{(a + b_1 x_1 + \ldots)}/[1 + e^{(a + b_1 x_1 + \ldots)}]$

Cumulative Density Function of Logistic transformation



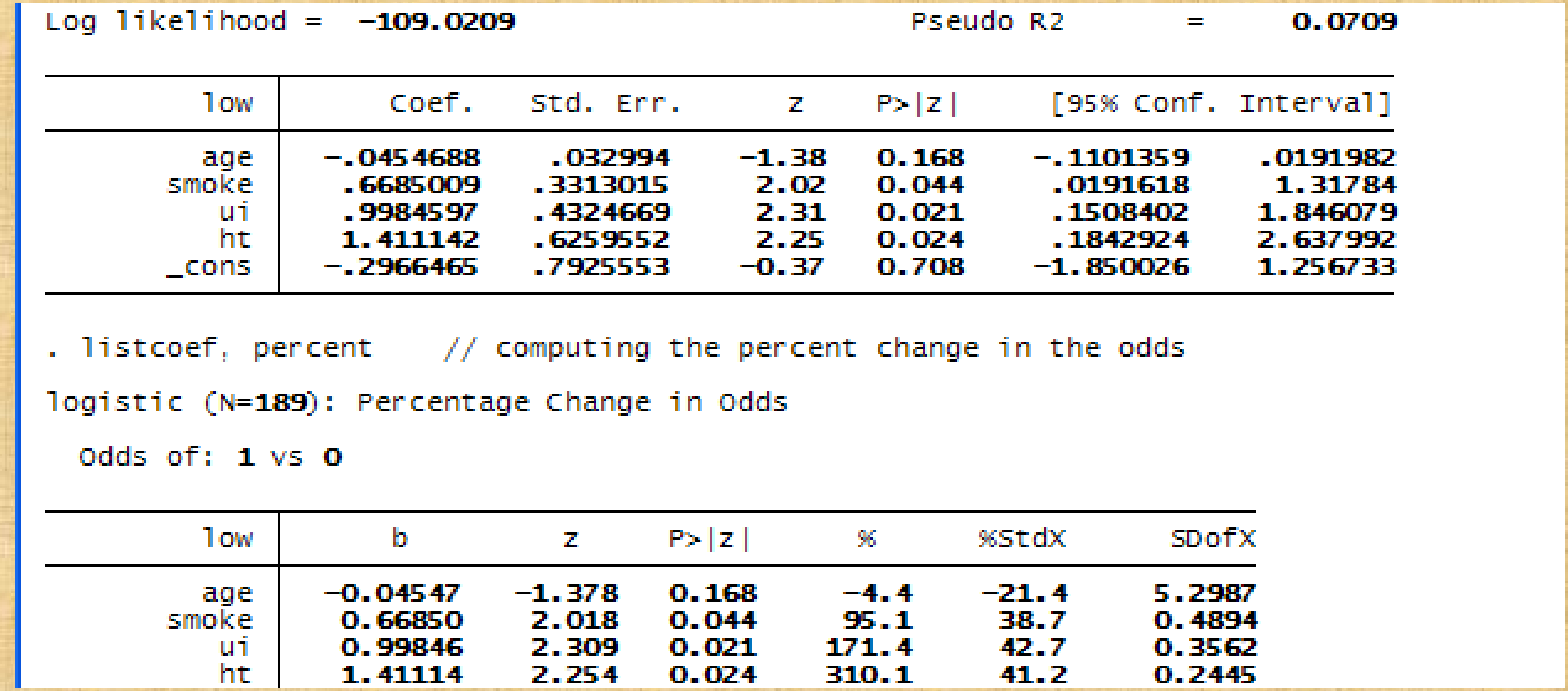
```
Plot[CDF[LogisticDistribution[0, 2], x], {x, -10, 10}]
```

Logistic regression



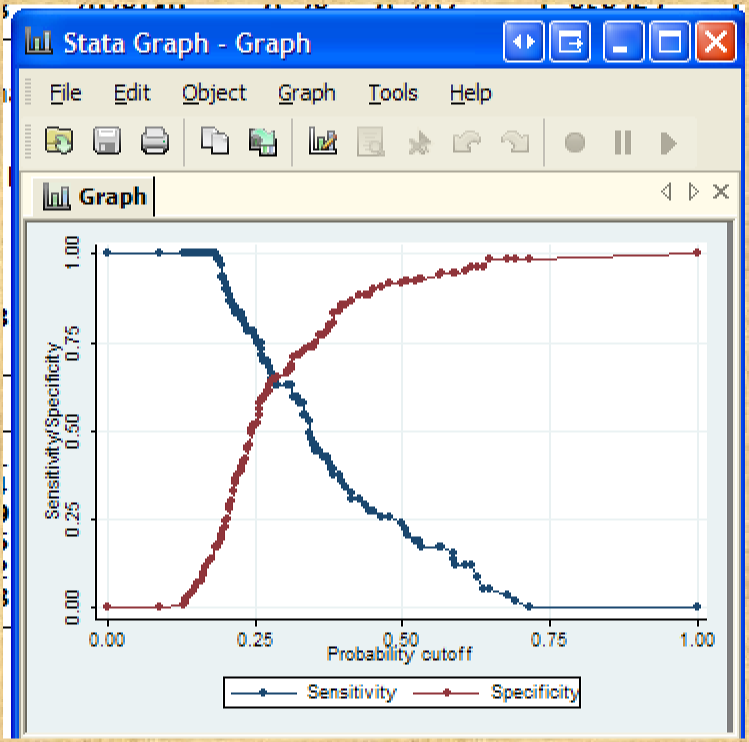
```
.
. logistic low age smoke ht ui ftv, coef

Logistic regression                               Number of obs   =        189
                                                  LR chi2(5)      =      16.66
                                                  Prob > chi2     =     0.0052
Log likelihood = -109.00351                       Pseudo R2       =     0.0710

------------------------------------------------------------------------------
         low |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |  -.0440911   .0338158    -1.30   0.192    -.1103689    .0221867
       smoke |   .6664874   .3315275     2.01   0.044     .0167055    1.316269
          ht |   1.400399   .6278679     2.23   0.026     .1698007    2.630998
          ui |   .9936605   .4335712     2.29   0.022     .1438766    1.843444
         ftv |  -.0302462   .1627385    -0.19   0.853    -.3492078    .2887154
       _cons |  -.3024128   .7938149    -0.38   0.703    -1.858262    1.253436
------------------------------------------------------------------------------

. lsens

. lroc

Logistic model for low

number of observations =      189
area under ROC curve   =   0.6870
```

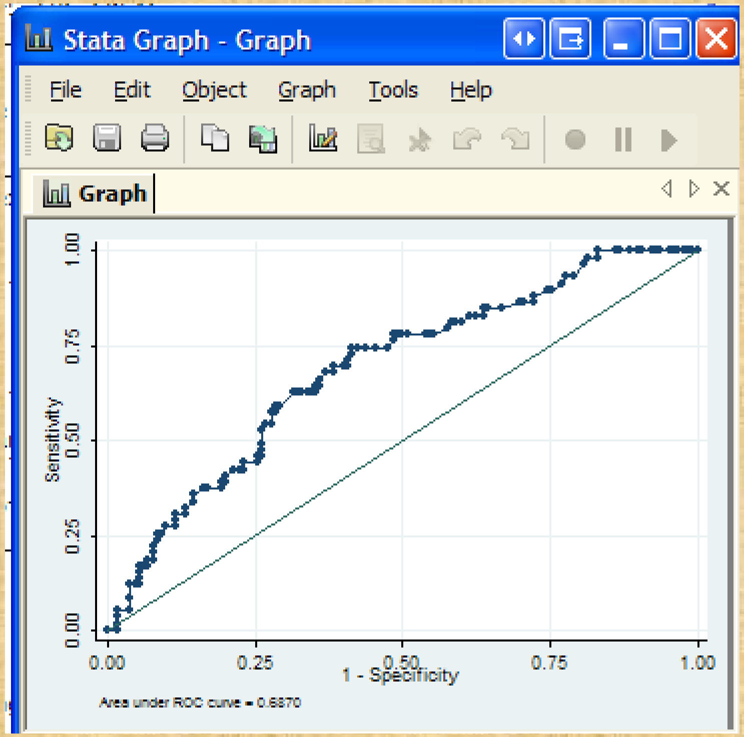
Converting coefficients to percentage change



```
Log likelihood =  -109.0209                          Pseudo R2       =     0.0709

------------------------------------------------------------------------------
         low |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |  -.0454688    .032994    -1.38   0.168    -.1101359    .0191982
       smoke |   .6685009   .3313015     2.02   0.044     .0191618     1.31784
          ui |   .9984597   .4324669     2.31   0.021     .1508402    1.846079
          ht |   1.411142   .6259552     2.25   0.024     .1842924    2.637992
       _cons |  -.2966465   .7925553    -0.37   0.708    -1.850026    1.256733
------------------------------------------------------------------------------

. listcoef, percent    // computing the percent change in the odds

logistic (N=189): Percentage Change in Odds

  Odds of: 1 vs 0

----------------------------------------------------------------------
         low |        b         z     P>|z|        %     %StdX    SDofX
-------------+--------------------------------------------------------
         age |  -0.04547   -1.378    0.168     -4.4     -21.4   5.2987
       smoke |   0.66850    2.018    0.044     95.1      38.7   0.4894
          ui |   0.99846    2.309    0.021    171.4      42.7   0.3562
          ht |   1.41114    2.254    0.024    310.1      41.2   0.2445
```

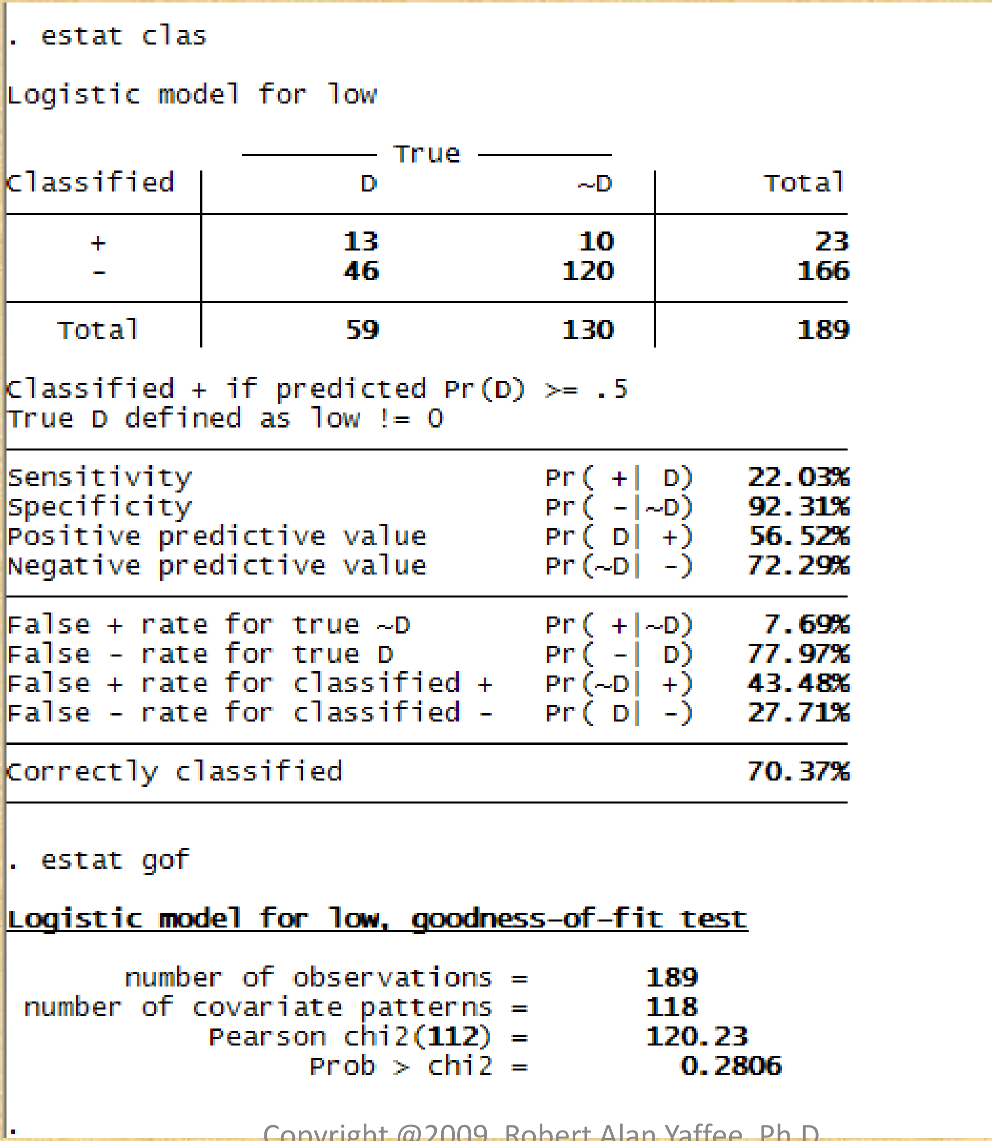
lsens and lroc





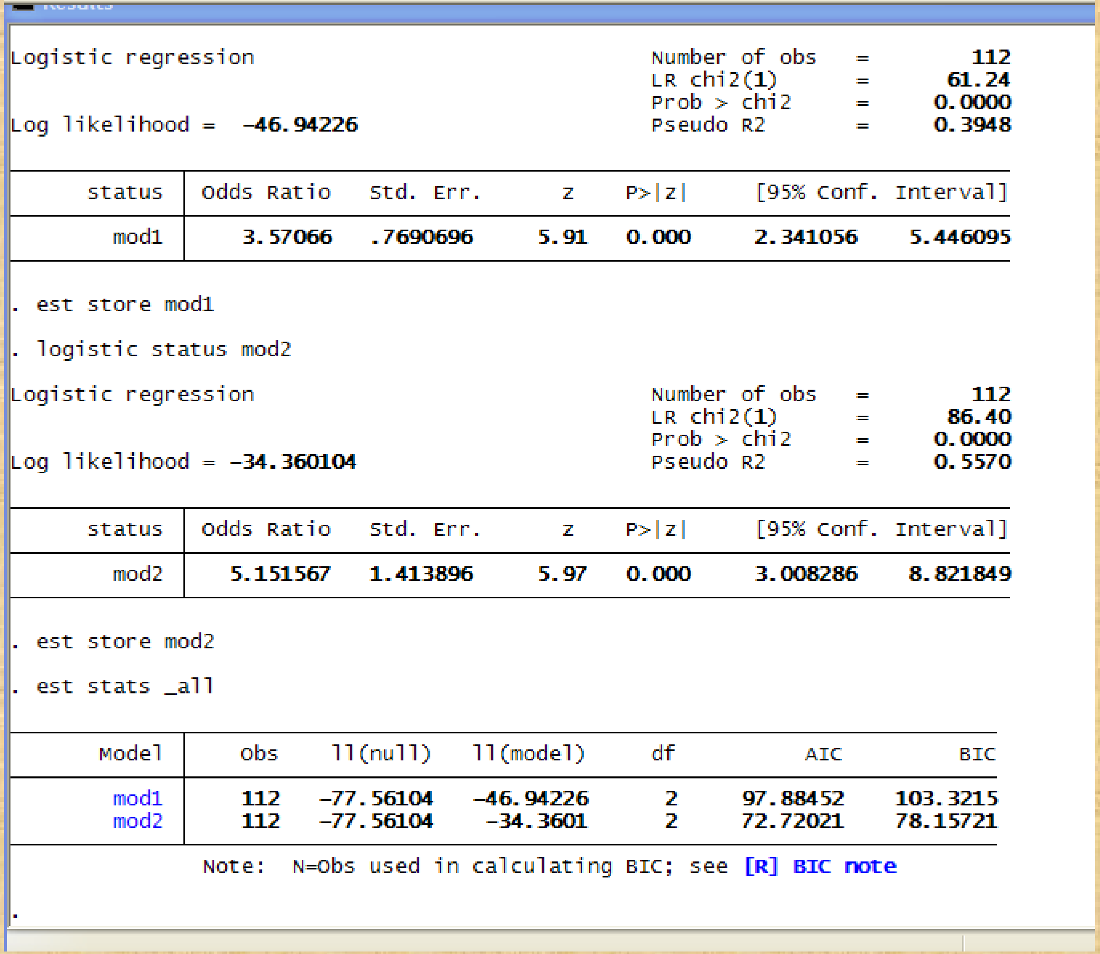
Area under ROC curve = 0.6870

estat class and estat gof



```
. estat clas

Logistic model for low

              -------- True --------
Classified |         D            ~D  |      Total
-----------+--------------------------+-----------
     +     |        13            10  |         23
     -     |        46           120  |        166
-----------+--------------------------+-----------
   Total   |        59           130  |        189

Classified + if predicted Pr(D) >= .5
True D defined as low != 0
--------------------------------------------------
Sensitivity                     Pr( +| D)   22.03%
Specificity                     Pr( -|~D)   92.31%
Positive predictive value       Pr( D| +)   56.52%
Negative predictive value       Pr(~D| -)   72.29%
--------------------------------------------------
False + rate for true ~D        Pr( +|~D)    7.69%
False - rate for true D         Pr( -| D)   77.97%
False + rate for classified +   Pr(~D| +)   43.48%
False - rate for classified -   Pr( D| -)   27.71%
--------------------------------------------------
Correctly classified                        70.37%
--------------------------------------------------

. estat gof

Logistic model for low, goodness-of-fit test

       number of observations =       189
 number of covariate patterns =       118
            Pearson chi2(112) =       120.23
                  Prob > chi2 =         0.2806

.
```

Comparing nested models with information criteria



```
Logistic regression                               Number of obs   =        112
                                                  LR chi2(1)      =      61.24
                                                  Prob > chi2     =     0.0000
Log likelihood =  -46.94226                       Pseudo R2       =     0.3948

------------------------------------------------------------------------------
      status | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        mod1 |    3.57066   .7690696     5.91   0.000     2.341056    5.446095
------------------------------------------------------------------------------

. est store mod1

. logistic status mod2

Logistic regression                               Number of obs   =        112
                                                  LR chi2(1)      =      86.40
                                                  Prob > chi2     =     0.0000
Log likelihood = -34.360104                       Pseudo R2       =     0.5570

------------------------------------------------------------------------------
      status | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        mod2 |   5.151567   1.413896     5.97   0.000     3.008286    8.821849
------------------------------------------------------------------------------

. est store mod2

. est stats _all

-----------------------------------------------------------------------------
       Model |    Obs    ll(null)   ll(model)     df          AIC         BIC
-------------+---------------------------------------------------------------
        mod1 |    112   -77.56104   -46.94226      2     97.88452    103.3215
        mod2 |    112   -77.56104    -34.3601      2     72.72021    78.15721
-----------------------------------------------------------------------------
               Note:  N=Obs used in calculating BIC; see [R] BIC note

.
```

Information Criteria

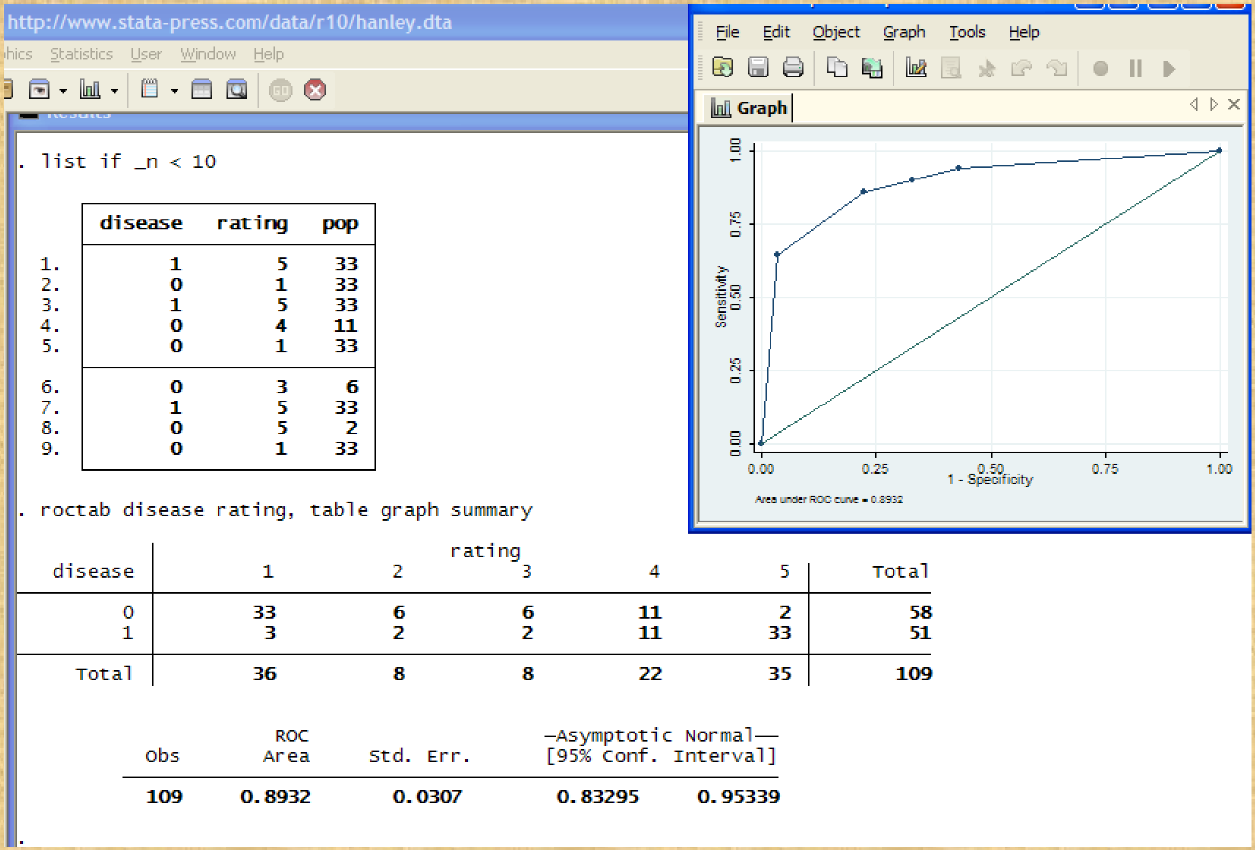
Information Criterion = deviance + penalty for number of parameters

$-2LL \sim SSE$ (deviance)

$AIC = -2LL + 2p$

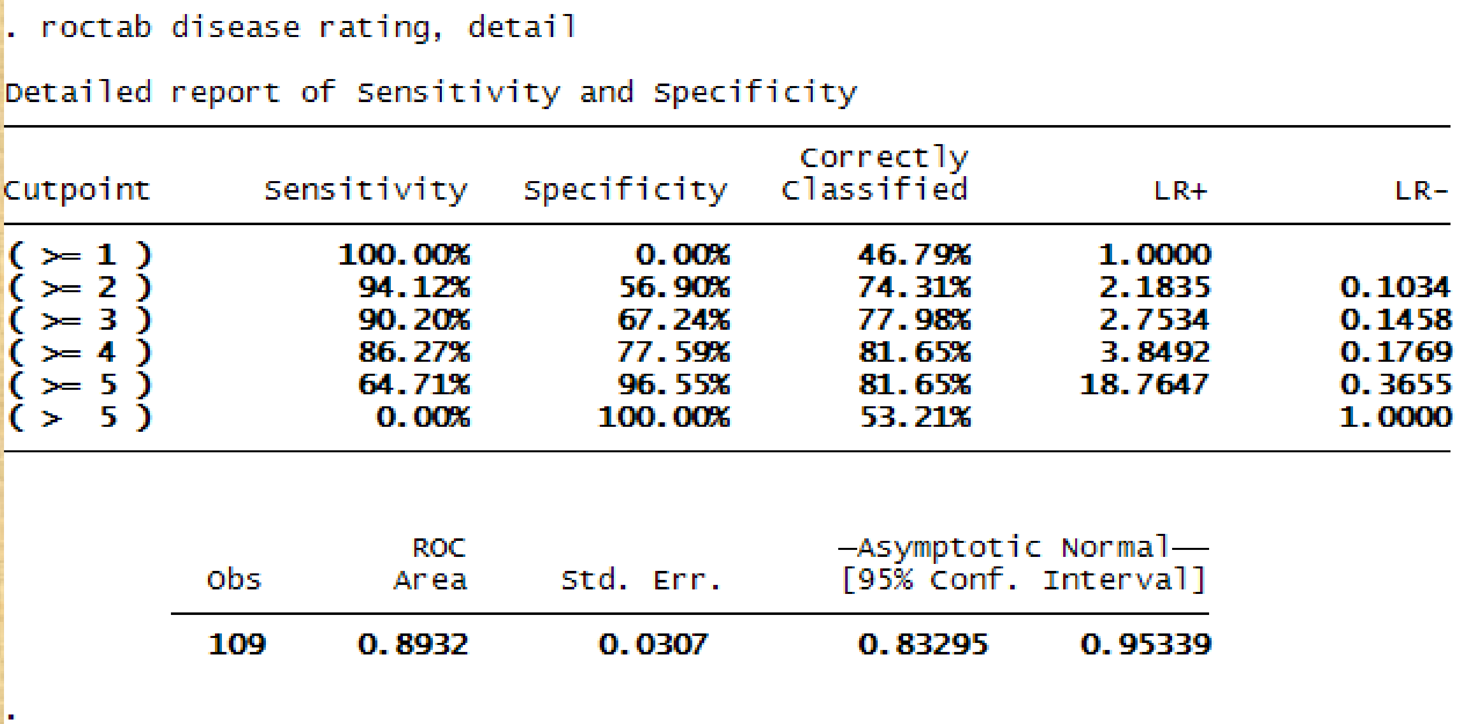
$BIC = -2LL + p\log(n)$

Receiver Operating Characteristic Analysis



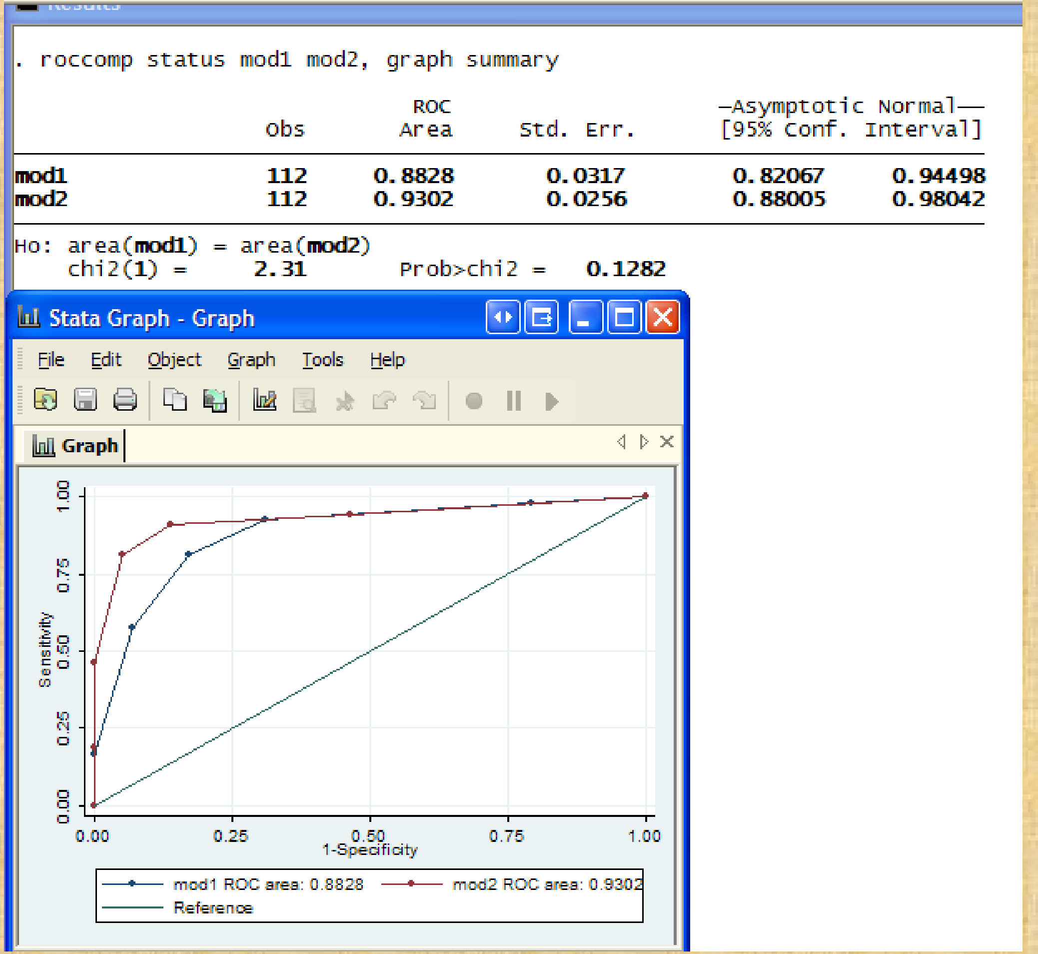
```
. list if _n < 10

     +--------------------------+
     | disease   rating     pop |
     |--------------------------|
  1. |       1        5      33 |
  2. |       0        1      33 |
  3. |       1        5      33 |
  4. |       0        4      11 |
  5. |       0        1      33 |
     |--------------------------|
  6. |       0        3       6 |
  7. |       1        5      33 |
  8. |       0        5       2 |
  9. |       0        1      33 |
     +--------------------------+

. roctab disease rating, table graph summary

           |                         rating
   disease |         1          2          3          4          5 |     Total
-----------+-------------------------------------------------------+----------
         0 |        33          6          6         11          2 |        58
         1 |         3          2          2         11         33 |        51
-----------+-------------------------------------------------------+----------
     Total |        36          8          8         22         35 |       109

                      ROC                    -Asymptotic Normal--
           Obs       Area     Std. Err.      [95% Conf. Interval]
         ------------------------------------------------------------
           109     0.8932       0.0307        0.83295     0.95339

.
```

Screening analysis



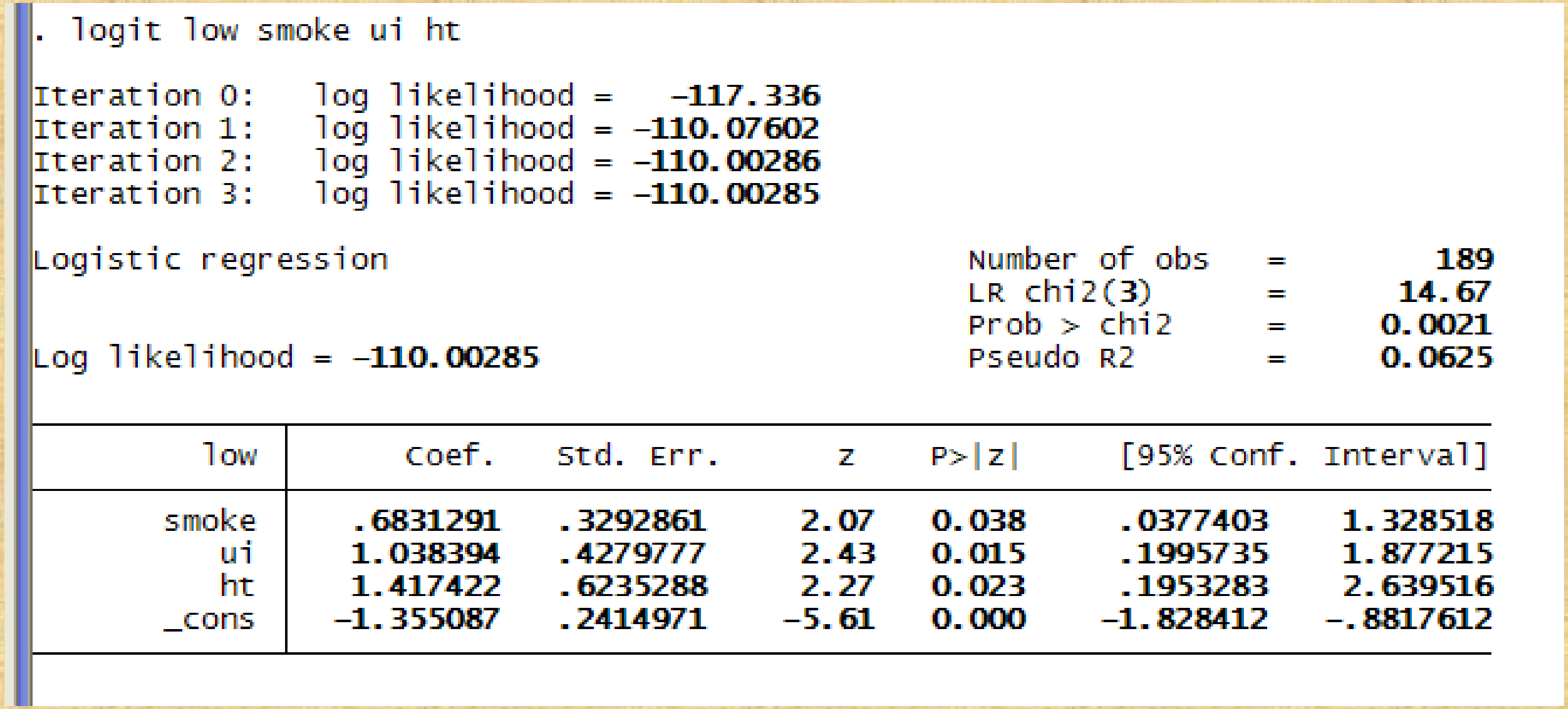
```
. roctab disease rating, detail

Detailed report of Sensitivity and Specificity
------------------------------------------------------------------------------
                                          Correctly
Cutpoint      Sensitivity   Specificity   Classified          LR+          LR-
------------------------------------------------------------------------------
( >= 1 )         100.00%         0.00%       46.79%       1.0000
( >= 2 )          94.12%        56.90%       74.31%       2.1835       0.1034
( >= 3 )          90.20%        67.24%       77.98%       2.7534       0.1458
( >= 4 )          86.27%        77.59%       81.65%       3.8492       0.1769
( >= 5 )          64.71%        96.55%       81.65%      18.7647       0.3655
( >  5 )           0.00%       100.00%       53.21%                    1.0000
------------------------------------------------------------------------------


                      ROC                    -Asymptotic Normal--
           Obs       Area     Std. Err.      [95% Conf. Interval]
         ------------------------------------------------------------
           109     0.8932       0.0307        0.83295     0.95339

.
```

Comparison of ROC curves to compare logistic models



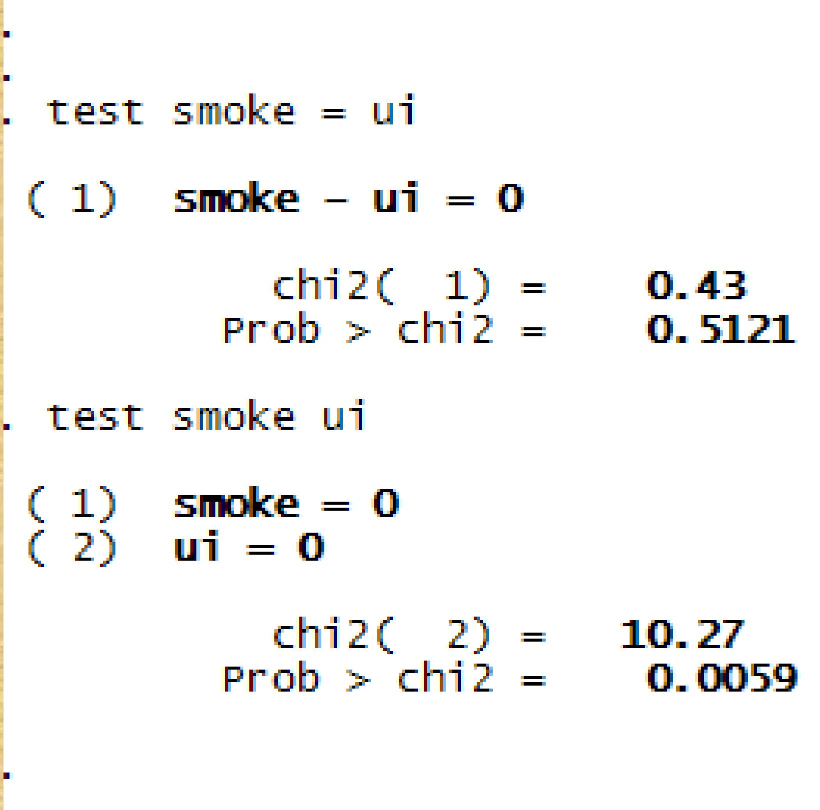
```
. roccomp status mod1 mod2, graph summary

                              ROC                    -Asymptotic Normal--
                  Obs       Area     Std. Err.      [95% Conf. Interval]
-------------------------------------------------------------------------
mod1              112     0.8828       0.0317        0.82067     0.94498
mod2              112     0.9302       0.0256        0.88005     0.98042
-------------------------------------------------------------------------
Ho: area(mod1) = area(mod2)
    chi2(1) =      2.31       Prob>chi2 =    0.1282
```

Logistic regression



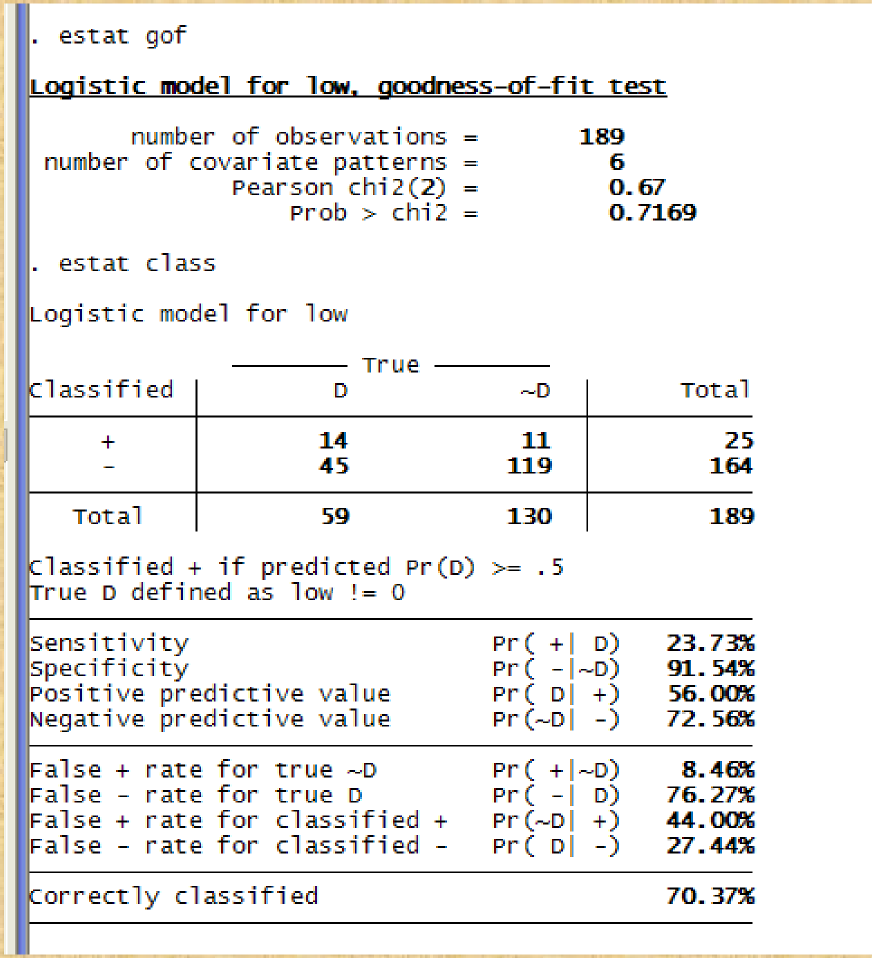
```
. logit low smoke ui ht

Iteration 0:   log likelihood =   -117.336
Iteration 1:   log likelihood = -110.07602
Iteration 2:   log likelihood = -110.00286
Iteration 3:   log likelihood = -110.00285

Logistic regression                               Number of obs   =        189
                                                  LR chi2(3)      =      14.67
                                                  Prob > chi2     =     0.0021
Log likelihood = -110.00285                       Pseudo R2       =     0.0625

------------------------------------------------------------------------------
         low |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       smoke |   .6831291   .3292861     2.07   0.038     .0377403    1.328518
          ui |   1.038394   .4279777     2.43   0.015     .1995735    1.877215
          ht |   1.417422   .6235288     2.27   0.023     .1953283    2.639516
       _cons |  -1.355087   .2414971    -5.61   0.000    -1.828412   -.8817612
------------------------------------------------------------------------------
```

Testing multiple coefficients



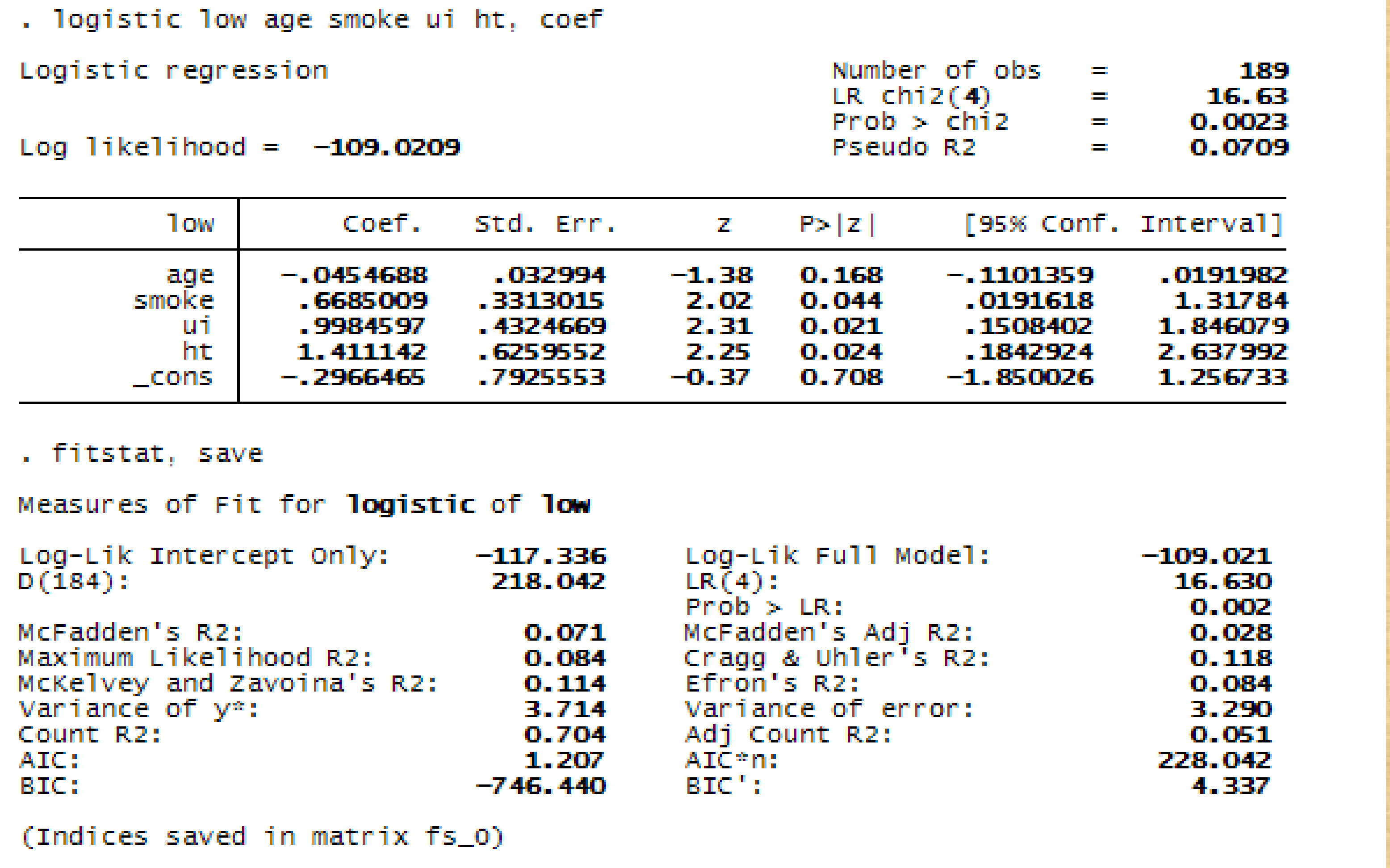
```
.
.
. test smoke = ui

 ( 1)  smoke - ui = 0

           chi2(  1) =    0.43
         Prob > chi2 =    0.5121

. test smoke ui

 ( 1)  smoke = 0
 ( 2)  ui = 0

           chi2(  2) =   10.27
         Prob > chi2 =    0.0059

.
```

Logistic regression postestimation



```
. estat gof

Logistic model for low, goodness-of-fit test

       number of observations =       189
 number of covariate patterns =         6
             Pearson chi2(2) =         0.67
                 Prob > chi2 =         0.7169

. estat class

Logistic model for low

              -------- True --------
Classified |         D            ~D  |      Total
-----------+--------------------------+-----------
     +     |        14            11  |         25
     -     |        45           119  |        164
-----------+--------------------------+-----------
   Total   |        59           130  |        189

Classified + if predicted Pr(D) >= .5
True D defined as low != 0
--------------------------------------------------
Sensitivity                     Pr( +| D)   23.73%
Specificity                     Pr( -|~D)   91.54%
Positive predictive value       Pr( D| +)   56.00%
Negative predictive value       Pr(~D| -)   72.56%
--------------------------------------------------
False + rate for true ~D        Pr( +|~D)    8.46%
False - rate for true D         Pr( -| D)   76.27%
False + rate for classified +   Pr(~D| +)   44.00%
False - rate for classified -   Pr( D| -)   27.44%
--------------------------------------------------
Correctly classified                        70.37%
--------------------------------------------------
```

Fitstat, save



```
. logistic low age smoke ui ht, coef

Logistic regression                               Number of obs   =        189
                                                  LR chi2(4)      =      16.63
                                                  Prob > chi2     =     0.0023
Log likelihood =  -109.0209                       Pseudo R2       =     0.0709

------------------------------------------------------------------------------
         low |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |  -.0454688    .032994    -1.38   0.168    -.1101359    .0191982
       smoke |   .6685009   .3313015     2.02   0.044     .0191618     1.31784
          ui |   .9984597   .4324669     2.31   0.021     .1508402    1.846079
          ht |   1.411142   .6259552     2.25   0.024     .1842924    2.637992
       _cons |  -.2966465   .7925553    -0.37   0.708    -1.850026    1.256733
------------------------------------------------------------------------------

. fitstat, save

Measures of Fit for logistic of low

Log-Lik Intercept Only:       -117.336   Log-Lik Full Model:           -109.021
D(184):                        218.042   LR(4):                          16.630
                                         Prob > LR:                       0.002
McFadden's R2:                   0.071   McFadden's Adj R2:               0.028
Maximum Likelihood R2:           0.084   Cragg & Uhler's R2:              0.118
McKelvey and Zavoina's R2:       0.114   Efron's R2:                      0.084
Variance of y*:                  3.714   Variance of error:               3.290
Count R2:                        0.704   Adj Count R2:                    0.051
AIC:                             1.207   AIC*n:                         228.042
BIC:                          -746.440   BIC':                            4.337

(Indices saved in matrix fs_0)
```

Formulae for fitstats

$$Likelihood\ \ ratio\ \ \chi^2 = 2Ln(Full\ \ model) - 2ln(null\ \ model)\quad df = diff\ \ in\ \ parms$$

$$Deviance = -2lnL(Full\ Model)\quad df\ = N - parms$$

$$R^2 = \frac{Var(\hat{y})}{Var(\hat{y}) + Var(\hat{\varepsilon})} = 1 - \left(\frac{L(\text{mod}\,null)}{L(\text{mod}\ full)}\right)^{2/N}$$

$$adj\ \ R^2 = \left(R^2 - \frac{p}{N-1}\right)\left(\frac{N-1}{N-p-1}\right)\ where\ \ p = number\ \ of\ \ parameters$$

$$N = number\ \ of\ \ observations.$$

$$McFadden's\ \ R^2 = 1 - \frac{LL(mod\ \ full)}{LL(mod\ \ null)}\ \ always\ increases\ with\ addition\ of\ variables.$$

$$Maximum\ \ likelihood\,(Cox - Snell)\ \ R^2$$

$$= 1 - \left(\frac{L(\text{mod}\,null)}{L(\text{mod}\ full)}\right)^{2/N} = 1 - \exp(-G^2 / N)$$

$$Cragg\ \ \&\ Uhler's\,(Nagelkerke)\ \ R^2 = \frac{R^2_{ML}}{\max R^2_{ML}}$$

$$= \frac{\left(1 - \left\{L(Modnull / LModfull\right\}\right)^{2/N}}{1 - L(Modnull)^{2/N}}$$

More Fitstat

Efron's pseudo R^2 *for binary outcomes*

defines $\hat{y} = \hat{\pi} = \Pr(y = 1 \mid x)$

$$= 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{\pi}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}$$

$$\text{Count } R^2 = \text{measure of proportion of correct predictions} = \frac{\sum n_{jj}}{N}.$$

$$\text{Adjusted Count } R^2 = \frac{\sum n_{jj} - \max(n_{r+})}{N - \max(n_{r+})}.$$

Information criteria

$$AIC = -2LL(\text{Model with } k \text{ parms}) - 2P_k$$

$$AIC' = \frac{-2LL - 2P}{N}$$

$$BIC = -2LL = dfk * \ln(N) \quad \text{where } dfk = df \text{ of deviance}(-2LL)$$

$$BIC' = -G^2(\text{Model with } k \text{ parms}) - df' Ln(N) \quad \text{where } df' = \#\text{ regressors in model}$$

$$BIC^s = -2L * N * Ln(\text{Model with } k \text{ parms}) + dfks\ln(N) \text{ where } dfks = \#\text{ parms in model}$$

including the constant.

More fitstat

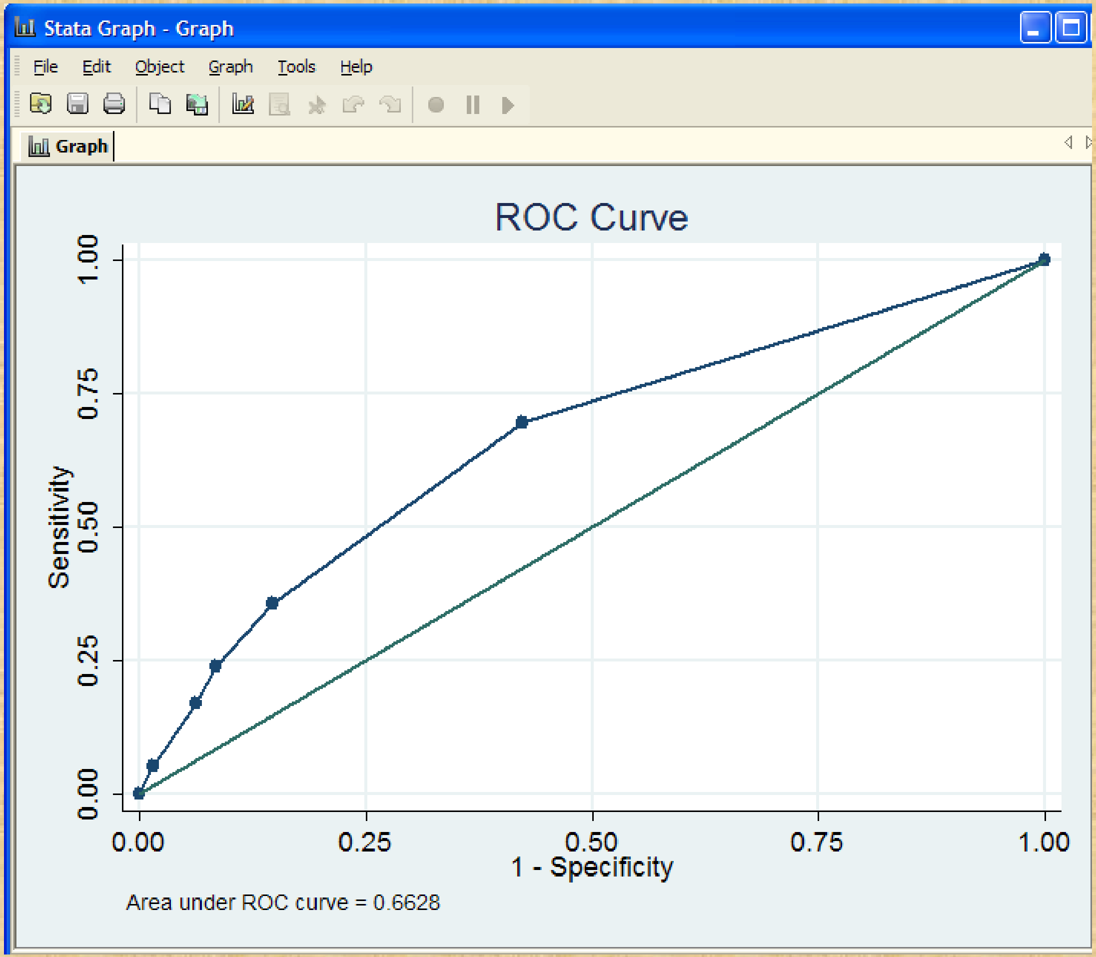
- Source of fitstat info: Long and Reese, op. cit., pp. 107-111.

$$McKelvey\ \&\ Zavoina's\ R^2 = \frac{Var(y^*)}{Var(y^*) + Var(e)}$$

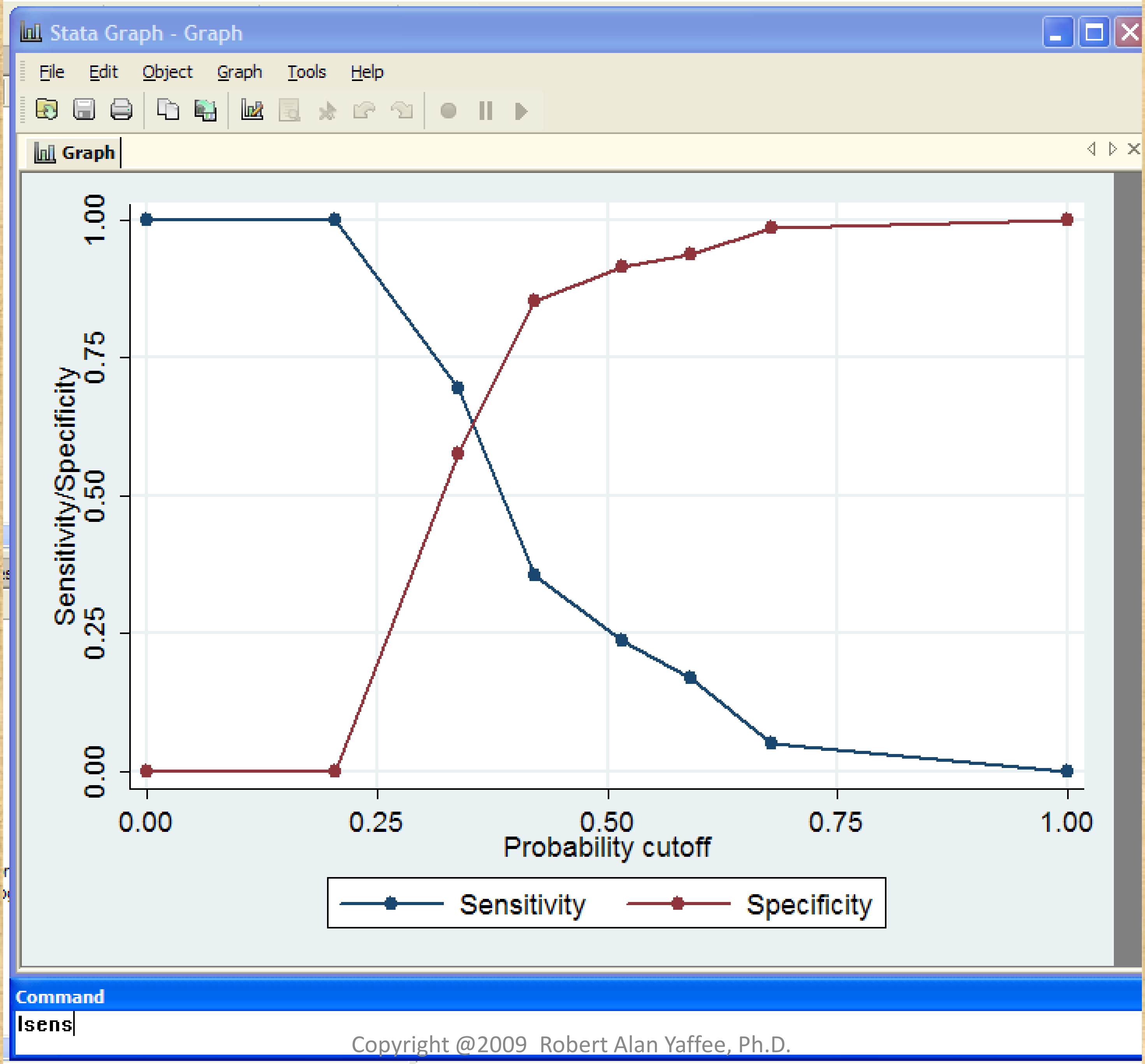
where

$y^* = latent\ variable.$

ROC curve



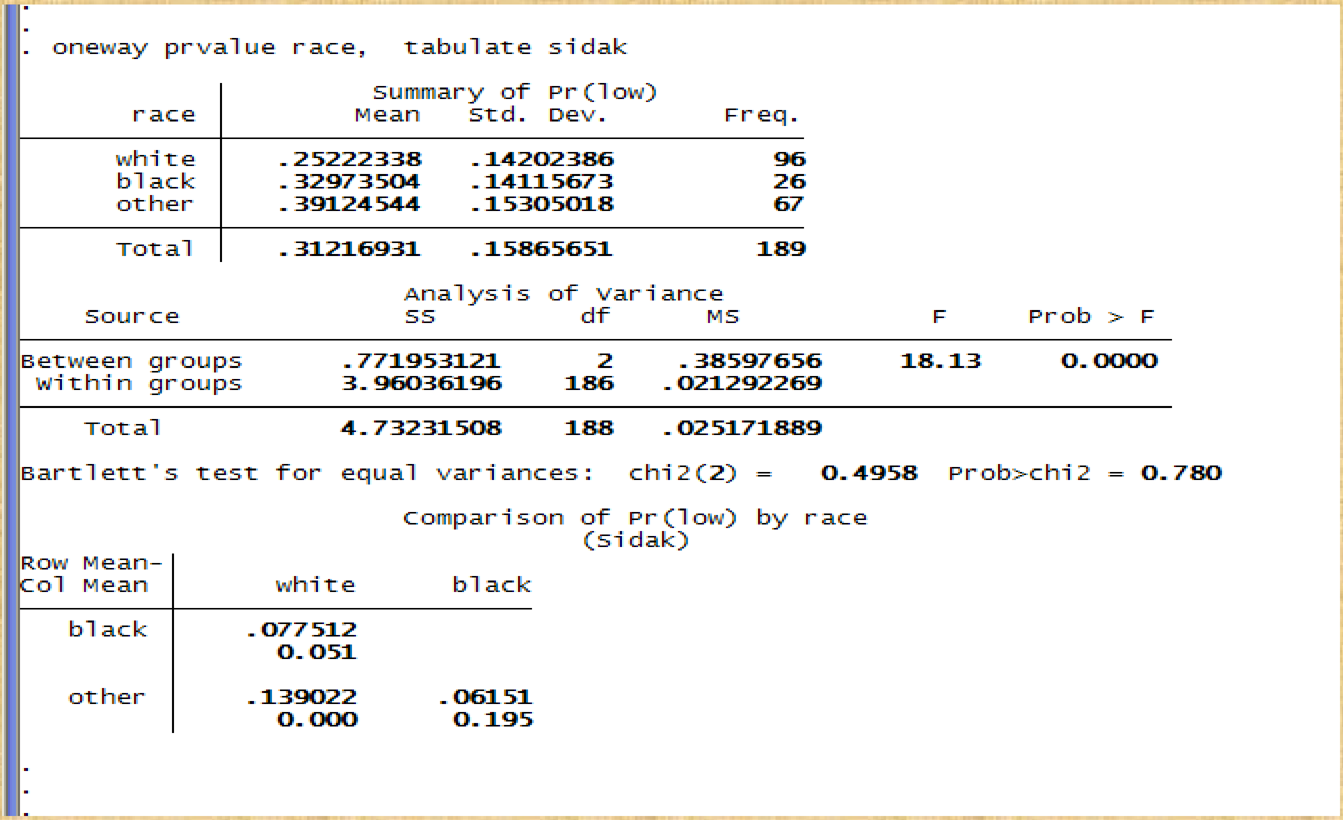
Area under ROC curve = 0.6628


Stata Graph - Graph
File
Edit
Object
Graph
Tools
Help
Graph
Sensitivity/Specificity
0.00
0.25
0.50
0.75
1.00
0.00
0.25
0.50
0.75
1.00
Probability cutoff
Sensitivity
Specificity
Command

```
lsens
```

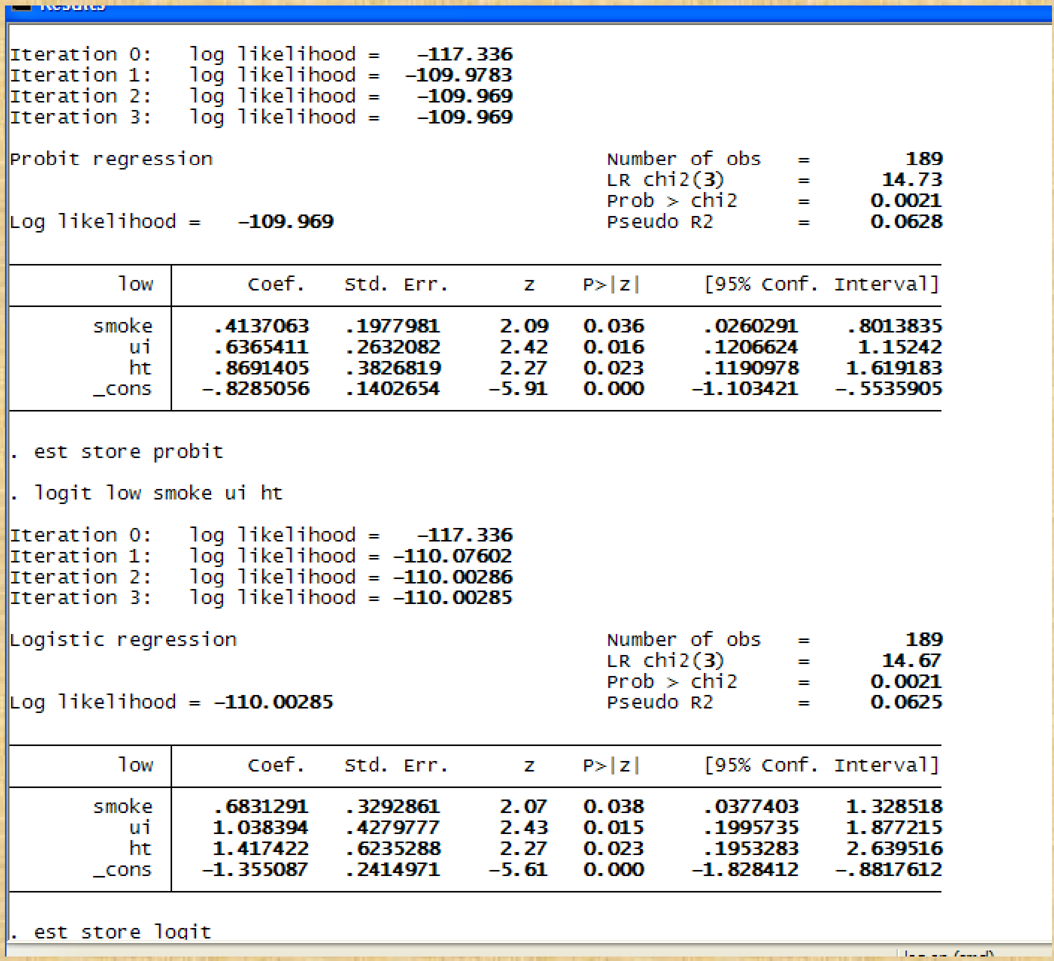
Save and analyze predicted probabilities of outcome

- Command: predict prvalue, pr



```
.
.
. oneway prvalue race,  tabulate sidak

            |          Summary of Pr(low)
       race |        Mean   Std. Dev.       Freq.
------------+------------------------------------
      white |   .25222338   .14202386          96
      black |   .32973504   .14115673          26
      other |   .39124544   .15305018          67
------------+------------------------------------
      Total |   .31216931   .15865651         189

                        Analysis of Variance
    Source              SS         df      MS            F     Prob > F
------------------------------------------------------------------------
Between groups      .771953121      2   .38597656     18.13     0.0000
 Within groups      3.96036196    186  .021292269
------------------------------------------------------------------------
    Total           4.73231508    188  .025171889

Bartlett's test for equal variances:  chi2(2) =   0.4958  Prob>chi2 = 0.780

                        Comparison of Pr(low) by race
                                   (Sidak)
Row Mean-|
Col Mean |      white      black
---------+----------------------
   black |    .077512
         |      0.051
         |
   other |    .139022     .06151
         |      0.000      0.195

.
.
.
```

Storing results for model comparison



```
Iteration 0:   log likelihood =   -117.336
Iteration 1:   log likelihood =  -109.9783
Iteration 2:   log likelihood =   -109.969
Iteration 3:   log likelihood =   -109.969

Probit regression                                 Number of obs   =        189
                                                  LR chi2(3)      =      14.73
                                                  Prob > chi2     =     0.0021
Log likelihood =   -109.969                       Pseudo R2       =     0.0628

------------------------------------------------------------------------------
         low |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       smoke |   .4137063   .1977981     2.09   0.036     .0260291    .8013835
          ui |   .6365411   .2632082     2.42   0.016     .1206624     1.15242
          ht |   .8691405   .3826819     2.27   0.023     .1190978    1.619183
       _cons |  -.8285056   .1402654    -5.91   0.000    -1.103421   -.5535905
------------------------------------------------------------------------------

. est store probit

. logit low smoke ui ht

Iteration 0:   log likelihood =   -117.336
Iteration 1:   log likelihood = -110.07602
Iteration 2:   log likelihood = -110.00286
Iteration 3:   log likelihood = -110.00285

Logistic regression                               Number of obs   =        189
                                                  LR chi2(3)      =      14.67
                                                  Prob > chi2     =     0.0021
Log likelihood = -110.00285                       Pseudo R2       =     0.0625

------------------------------------------------------------------------------
         low |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       smoke |   .6831291   .3292861     2.07   0.038     .0377403    1.328518
          ui |   1.038394   .4279777     2.43   0.015     .1995735    1.877215
          ht |   1.417422   .6235288     2.27   0.023     .1953283    2.639516
       _cons |  -1.355087   .2414971    -5.61   0.000    -1.828412   -.8817612
------------------------------------------------------------------------------

. est store logit
```

Assumptions of binary logistic regression

- Linearity: The model is linear for logits.
- Additivity: There are no significant interactions.
- The residuals are binomially distributed until the sample size gets large when the binomial assumes a normal shape. Therefore, large samples are necessary for such Maximum likelihood estimation.
- No multicollinearity.
- No overly influential observations (Daniel Pregebon)

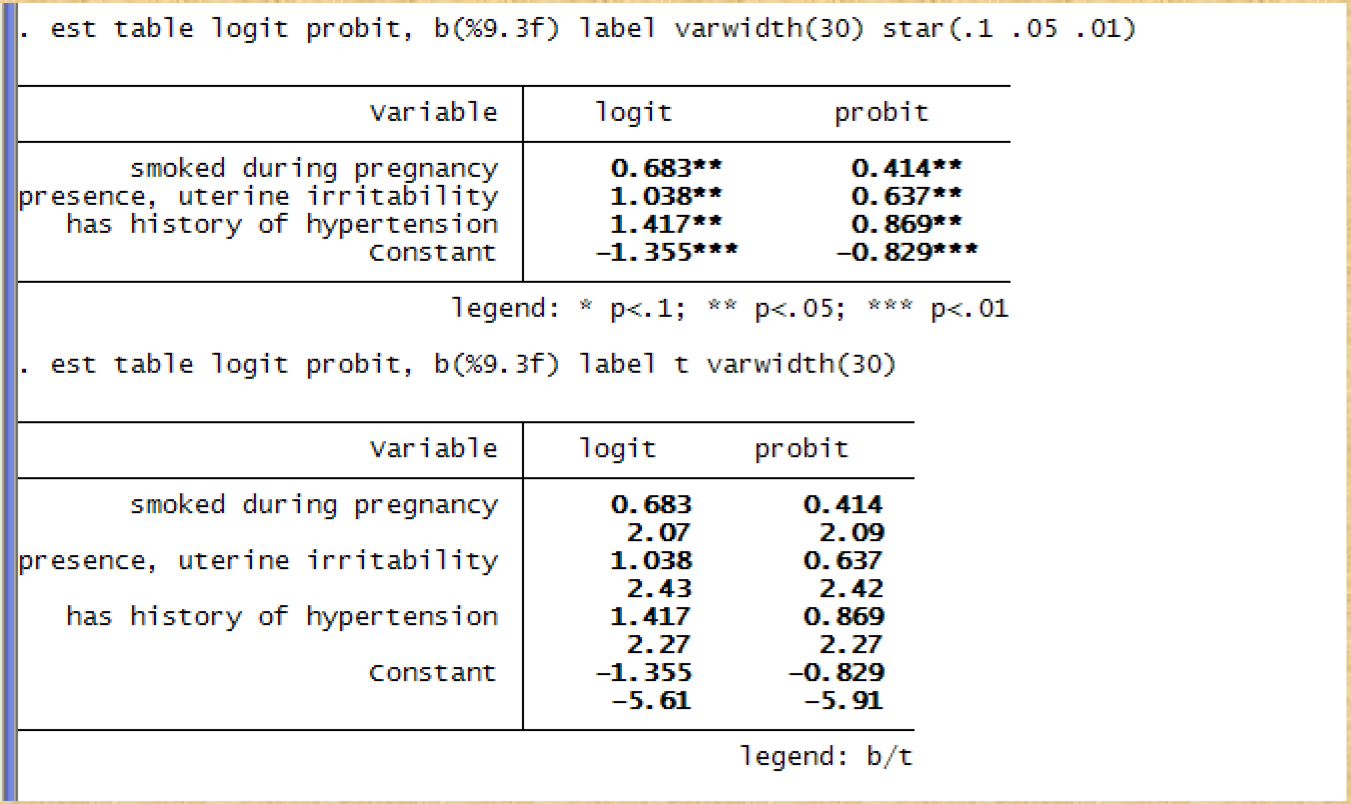
Model Validation by testing assumptions

- Test each assumption to be sure it holds for the model. Check the sample size to be sure that it is large.
- Check for multicollinearity with the iv corr matrix.
- Test for interactions between variables.
- Plot the probabilities to check for linearity.
- .

Validating tests

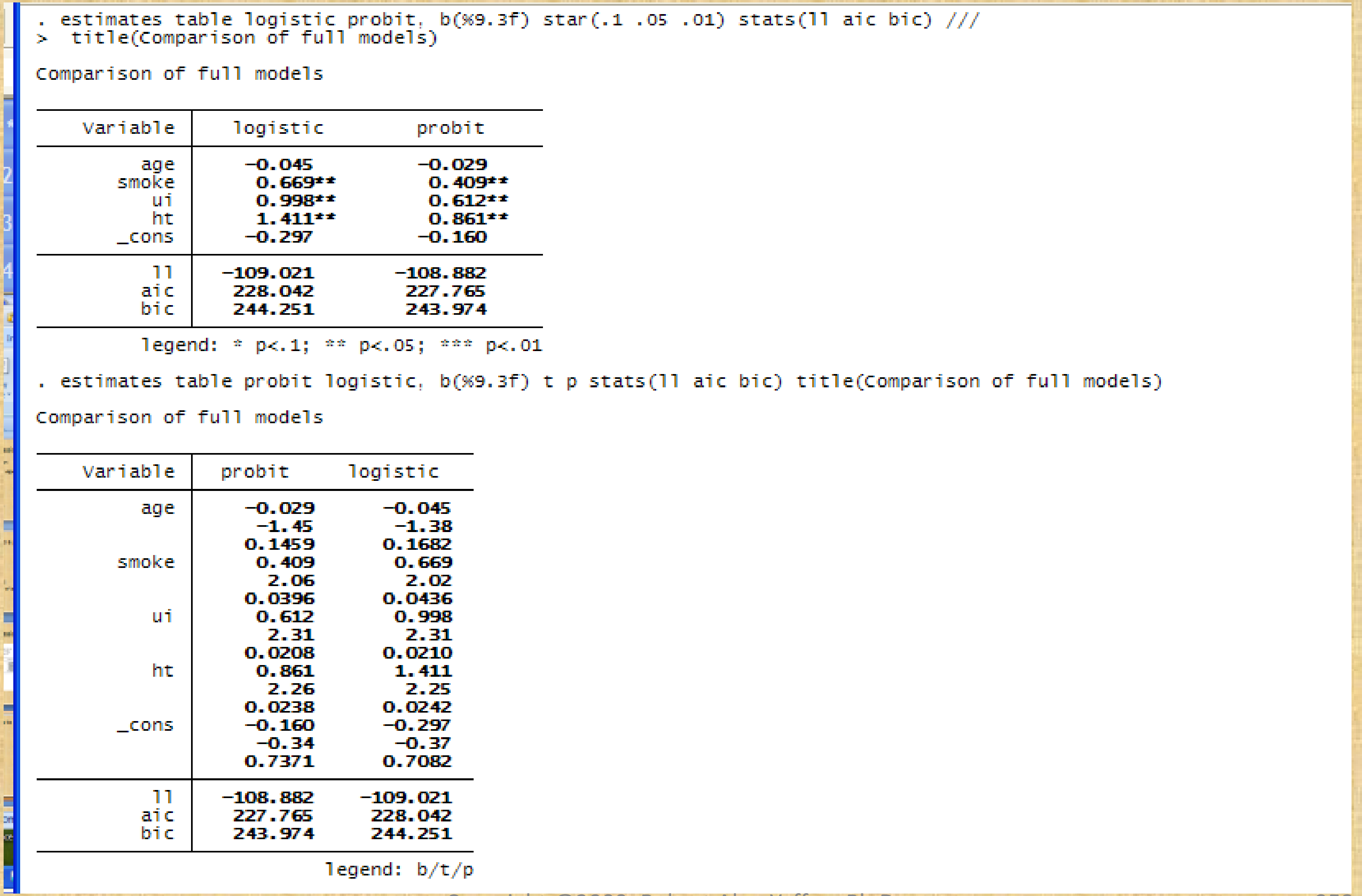
- Check the Classification chart to be sure that the percentage correctly classified is high.
- Test for predictive validity of classification on an out-of-sample analysis. You can use Brier's Score = Error variance (with n as denominator). Nonparametric correlation between observed and predicted scores (Somer's D) should be high.
- Bootstrap or jacknife to be sure that the empirical std errors do not deviate much from those estimated.
- Compute the Q (the overall quality index). Q = D-U,
  - where D=discrimination score (LR chi-square-1)/n and the U = unreliability index (- 2LL between uncalilbrated XB and the calibrated XB (with overall intercept and slope calibrated to the test sample).

Model comparison tables



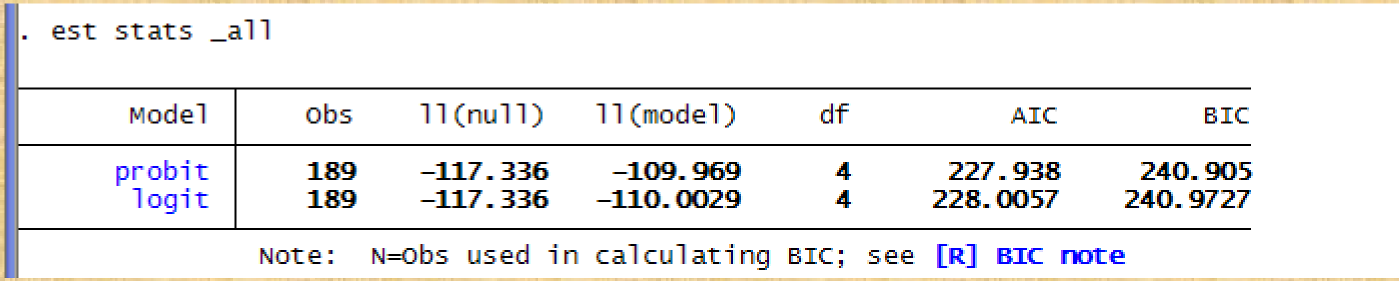
```
. est table logit probit, b(%9.3f) label varwidth(30) star(.1 .05 .01)

--------------------------------------------------------------
                      Variable |     logit          probit
-------------------------------+------------------------------
       smoked during pregnancy |    0.683**         0.414**
presence, uterine irritability |    1.038**         0.637**
   has history of hypertension |    1.417**         0.869**
                      Constant |   -1.355***       -0.829***
--------------------------------------------------------------
                            legend: * p<.1; ** p<.05; *** p<.01

. est table logit probit, b(%9.3f) label t varwidth(30)

--------------------------------------------------------
                      Variable |    logit       probit
-------------------------------+------------------------
       smoked during pregnancy |    0.683        0.414
                               |     2.07         2.09
presence, uterine irritability |    1.038        0.637
                               |     2.43         2.42
   has history of hypertension |    1.417        0.869
                               |     2.27         2.27
                      Constant |   -1.355       -0.829
                               |    -5.61        -5.91
--------------------------------------------------------
                                             legend: b/t
```

Other model comparison tables



```
. estimates table logistic probit, b(%9.3f) star(.1 .05 .01) stats(ll aic bic) ///
>  title(Comparison of full models)

Comparison of full models

------------------------------------------
    Variable |   logistic        probit
-------------+----------------------------
         age |    -0.045        -0.029
       smoke |     0.669**       0.409**
          ui |     0.998**       0.612**
          ht |     1.411**       0.861**
       _cons |    -0.297        -0.160
-------------+----------------------------
          ll |  -109.021      -108.882
         aic |   228.042       227.765
         bic |   244.251       243.974
------------------------------------------
         legend: * p<.1; ** p<.05; *** p<.01

. estimates table probit logistic, b(%9.3f) t p stats(ll aic bic) title(Comparison of full models)

Comparison of full models

------------------------------------
    Variable |   probit    logistic
-------------+----------------------
         age |    -0.029     -0.045
             |     -1.45      -1.38
             |    0.1459     0.1682
       smoke |     0.409      0.669
             |      2.06       2.02
             |    0.0396     0.0436
          ui |     0.612      0.998
             |      2.31       2.31
             |    0.0208     0.0210
          ht |     0.861      1.411
             |      2.26       2.25
             |    0.0238     0.0242
       _cons |    -0.160     -0.297
             |     -0.34      -0.37
             |    0.7371     0.7082
-------------+----------------------
          ll |  -108.882   -109.021
         aic |   227.765    228.042
         bic |   243.974    244.251
------------------------------------
                         legend: b/t/p
```

Model comparison with information criteria



```
. est stats _all

-----------------------------------------------------------------------------
       Model |    Obs    ll(null)   ll(model)     df          AIC         BIC
-------------+---------------------------------------------------------------
      probit |    189    -117.336    -109.969      4      227.938     240.905
       logit |    189    -117.336   -110.0029      4     228.0057    240.9727
-----------------------------------------------------------------------------
               Note:  N=Obs used in calculating BIC; see [R] BIC note
```

Probit regression model

introduced by Chester Ittner Bliss (1935)

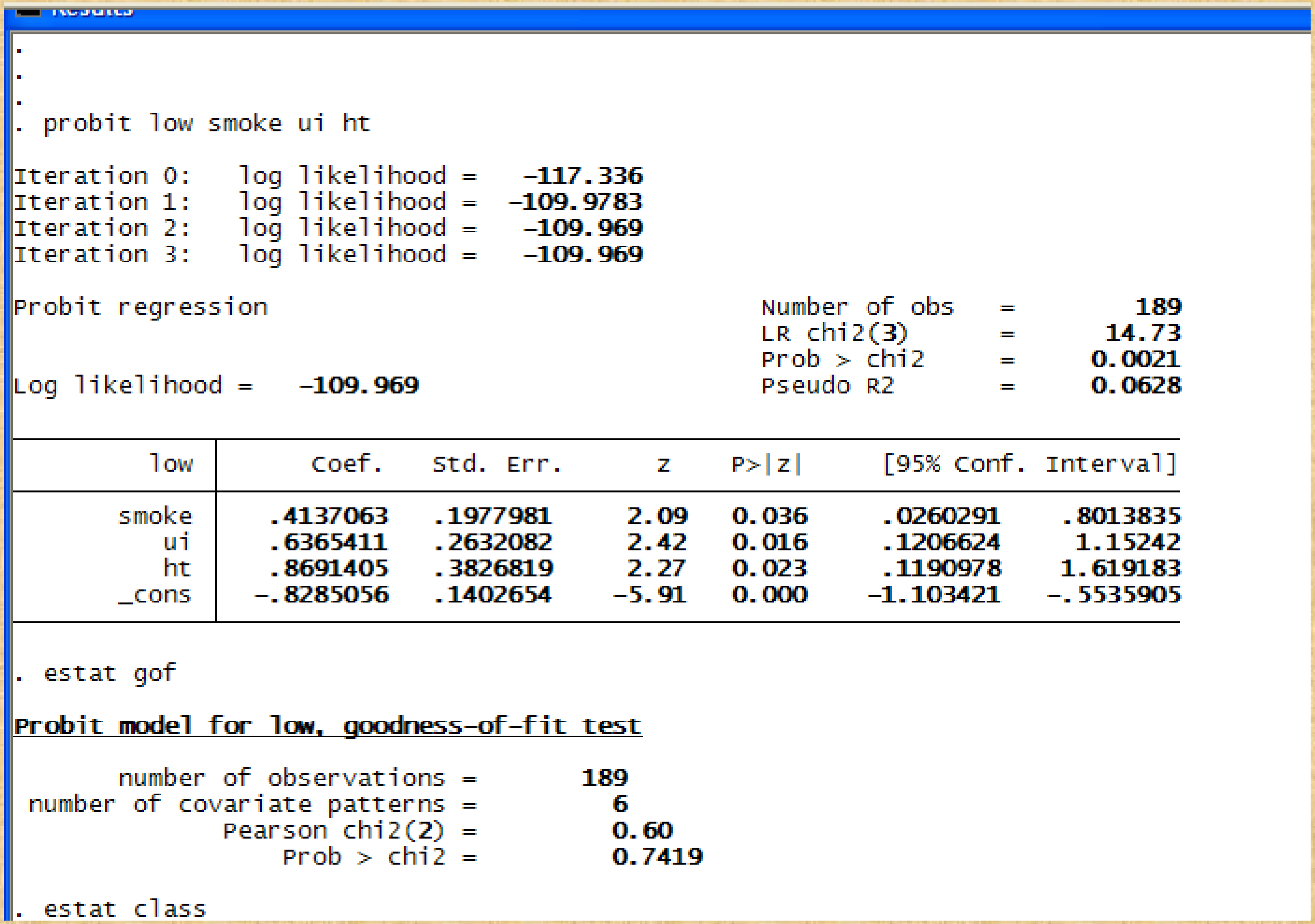
- It assumes that the underlying area of the normal curve is bifurcated so the proportions of counts in one group and those in the other represent the percentages of counts in the 0s and 1s.

$$\lim_{x'\beta -> \infty} Prob(Y = 1 | x) = 1$$

$$\lim_{x'\beta -> -\infty} Prob(Y = 1 | x) = 0$$

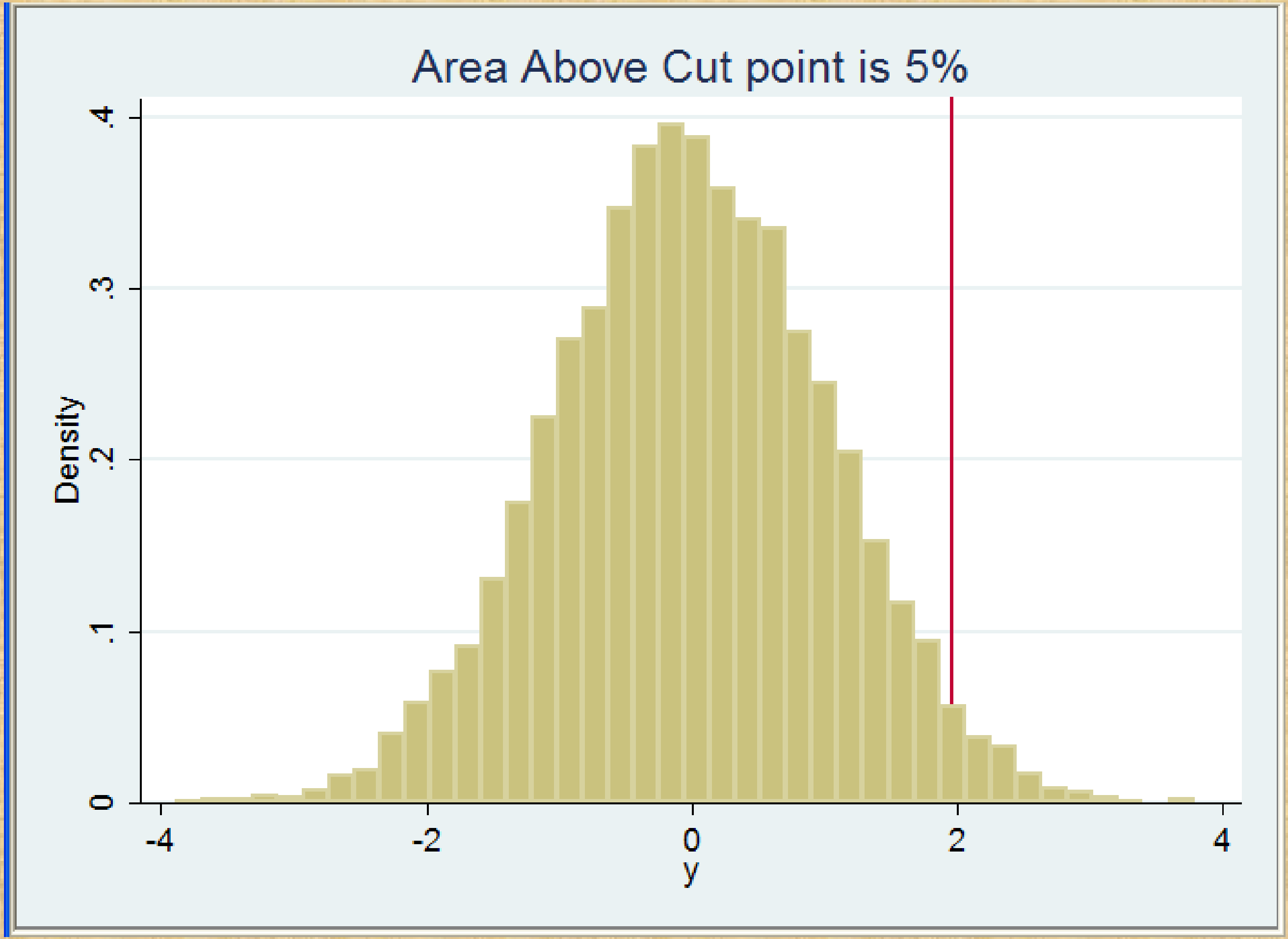
$$Prob\ (Y = 1 | x) = \int_{-\infty}^{\infty} \phi(t)dt = \Theta(x'\beta)$$

Probit model

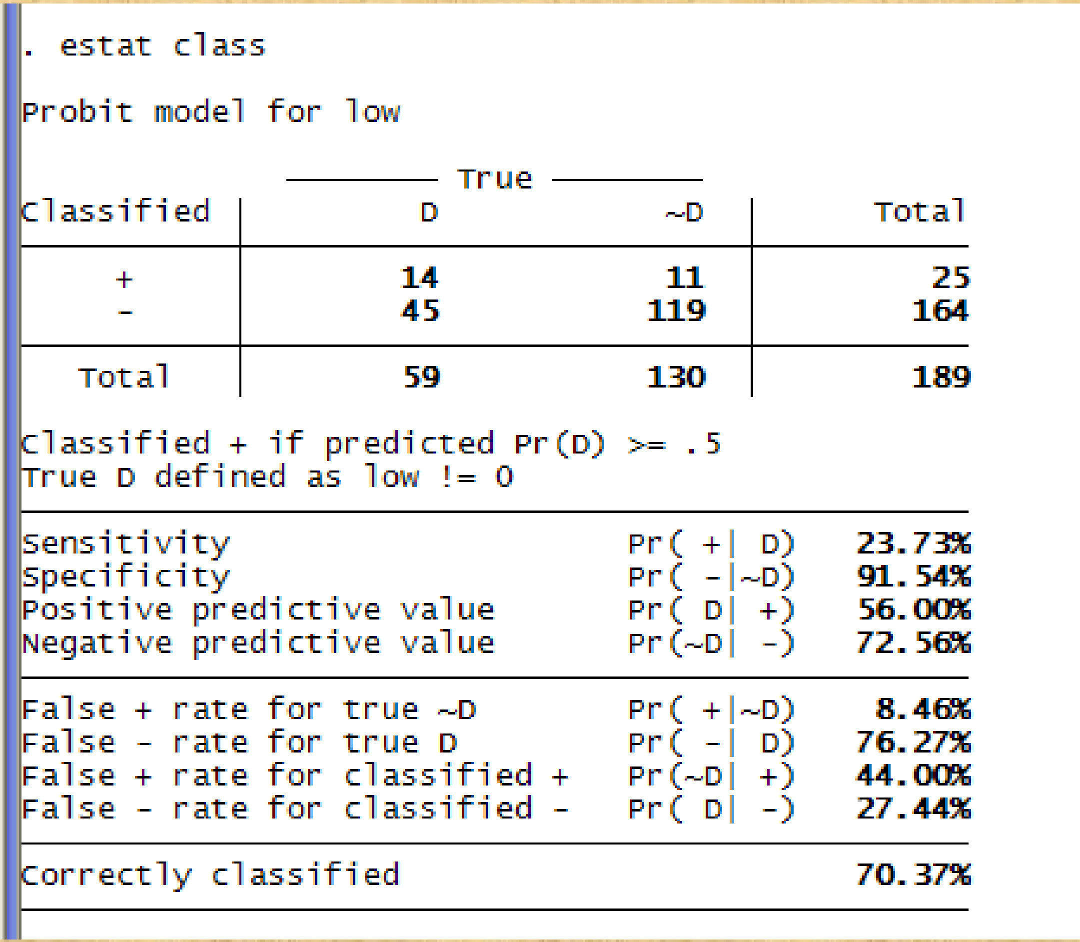


```
.
.
.
. probit low smoke ui ht

Iteration 0:   log likelihood =   -117.336
Iteration 1:   log likelihood =  -109.9783
Iteration 2:   log likelihood =   -109.969
Iteration 3:   log likelihood =   -109.969

Probit regression                                 Number of obs   =        189
                                                  LR chi2(3)      =      14.73
                                                  Prob > chi2     =     0.0021
Log likelihood =   -109.969                       Pseudo R2       =     0.0628

------------------------------------------------------------------------------
         low |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       smoke |   .4137063   .1977981     2.09   0.036     .0260291    .8013835
          ui |   .6365411   .2632082     2.42   0.016     .1206624     1.15242
          ht |   .8691405   .3826819     2.27   0.023     .1190978    1.619183
       _cons |  -.8285056   .1402654    -5.91   0.000    -1.103421   -.5535905
------------------------------------------------------------------------------

. estat gof

Probit model for low, goodness-of-fit test

       number of observations =       189
 number of covariate patterns =         6
             Pearson chi2(2) =         0.60
                 Prob > chi2 =         0.7419

. estat class
```

An assumption of an underlying normal variable

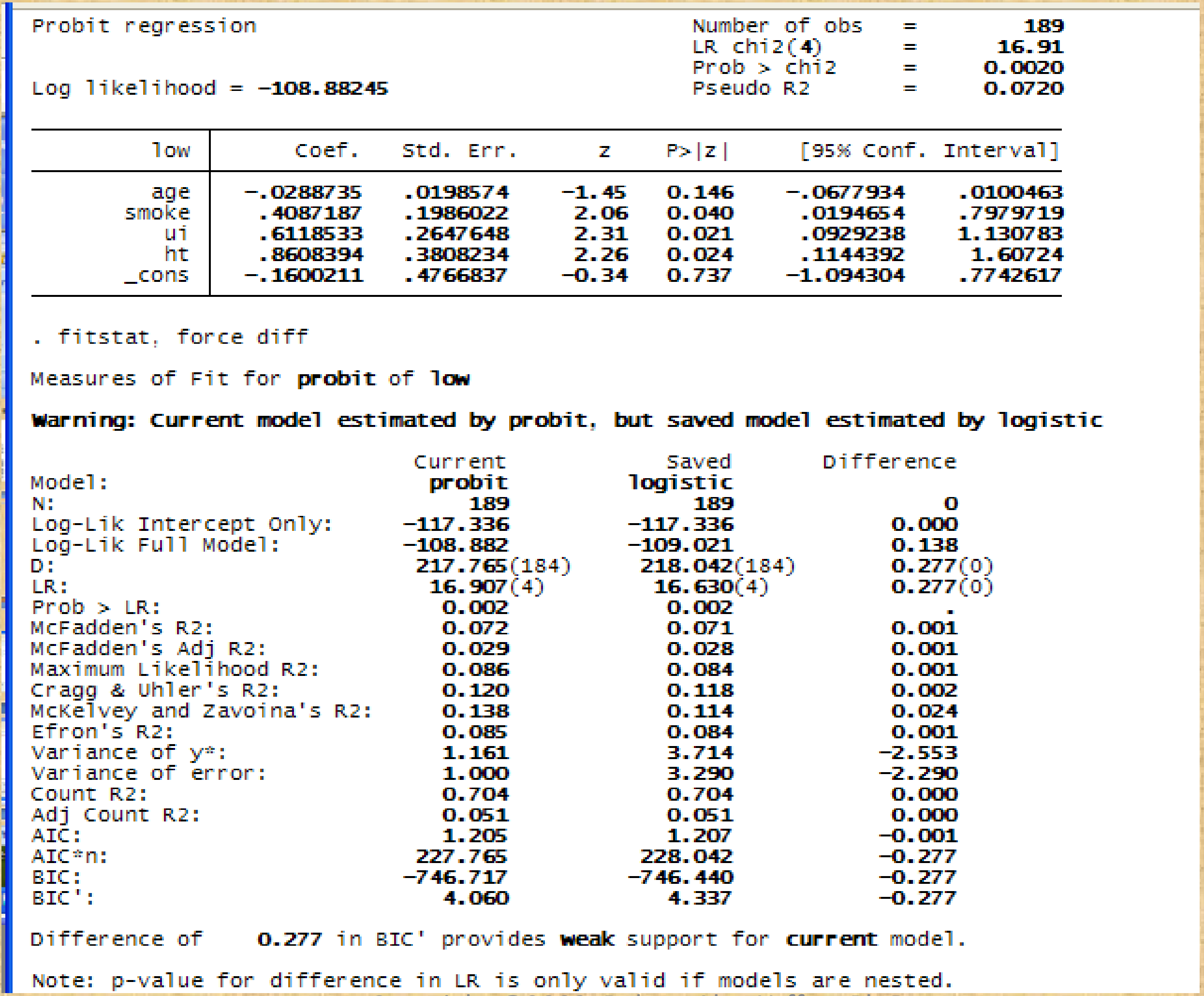


Classification table



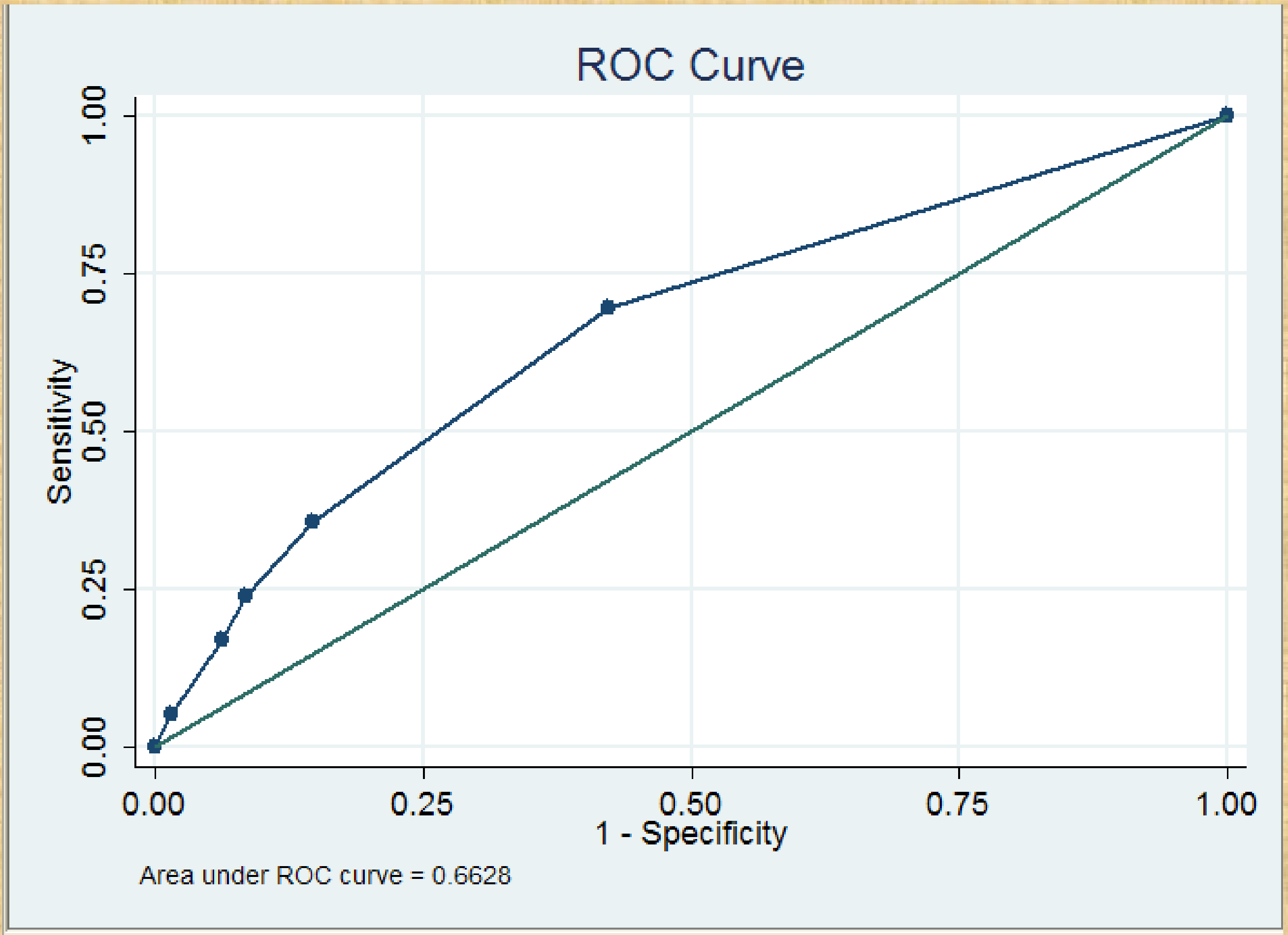
```
. estat class

Probit model for low

              -------- True --------
Classified |         D            ~D  |      Total
-----------+--------------------------+-----------
     +     |        14            11  |         25
     -     |        45           119  |        164
-----------+--------------------------+-----------
   Total   |        59           130  |        189

Classified + if predicted Pr(D) >= .5
True D defined as low != 0
--------------------------------------------------
Sensitivity                     Pr( +| D)   23.73%
Specificity                     Pr( -|~D)   91.54%
Positive predictive value       Pr( D| +)   56.00%
Negative predictive value       Pr(~D| -)   72.56%
--------------------------------------------------
False + rate for true ~D        Pr( +|~D)    8.46%
False - rate for true D         Pr( -| D)   76.27%
False + rate for classified +   Pr(~D| +)   44.00%
False - rate for classified -   Pr( D| -)   27.44%
--------------------------------------------------
Correctly classified                        70.37%
--------------------------------------------------
```

Fitstat, diff force



```
Probit regression                                 Number of obs   =        189
                                                  LR chi2(4)      =      16.91
                                                  Prob > chi2     =     0.0020
Log likelihood = -108.88245                       Pseudo R2       =     0.0720

------------------------------------------------------------------------------
         low |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |  -.0288735   .0198574    -1.45   0.146    -.0677934    .0100463
       smoke |   .4087187   .1986022     2.06   0.040     .0194654    .7979719
          ui |   .6118533   .2647648     2.31   0.021     .0929238    1.130783
          ht |   .8608394   .3808234     2.26   0.024     .1144392     1.60724
       _cons |  -.1600211   .4766837    -0.34   0.737    -1.094304    .7742617
------------------------------------------------------------------------------

. fitstat, force diff

Measures of Fit for probit of low

Warning: Current model estimated by probit, but saved model estimated by logistic

                              Current            Saved       Difference
Model:                         probit         logistic
N:                                189              189                0
Log-Lik Intercept Only:      -117.336         -117.336            0.000
Log-Lik Full Model:          -108.882         -109.021            0.138
D:                        217.765(184)     218.042(184)        0.277(0)
LR:                         16.907(4)        16.630(4)         0.277(0)
Prob > LR:                      0.002            0.002                .
McFadden's R2:                  0.072            0.071            0.001
McFadden's Adj R2:              0.029            0.028            0.001
Maximum Likelihood R2:          0.086            0.084            0.001
Cragg & Uhler's R2:             0.120            0.118            0.002
McKelvey and Zavoina's R2:      0.138            0.114            0.024
Efron's R2:                     0.085            0.084            0.001
Variance of y*:                 1.161            3.714           -2.553
Variance of error:              1.000            3.290           -2.290
Count R2:                       0.704            0.704            0.000
Adj Count R2:                   0.051            0.051            0.000
AIC:                            1.205            1.207           -0.001
AIC*n:                        227.765          228.042           -0.277
BIC:                         -746.717         -746.440           -0.277
BIC':                           4.060            4.337           -0.277

Difference of    0.277 in BIC' provides weak support for current model.

Note: p-value for difference in LR is only valid if models are nested.
```

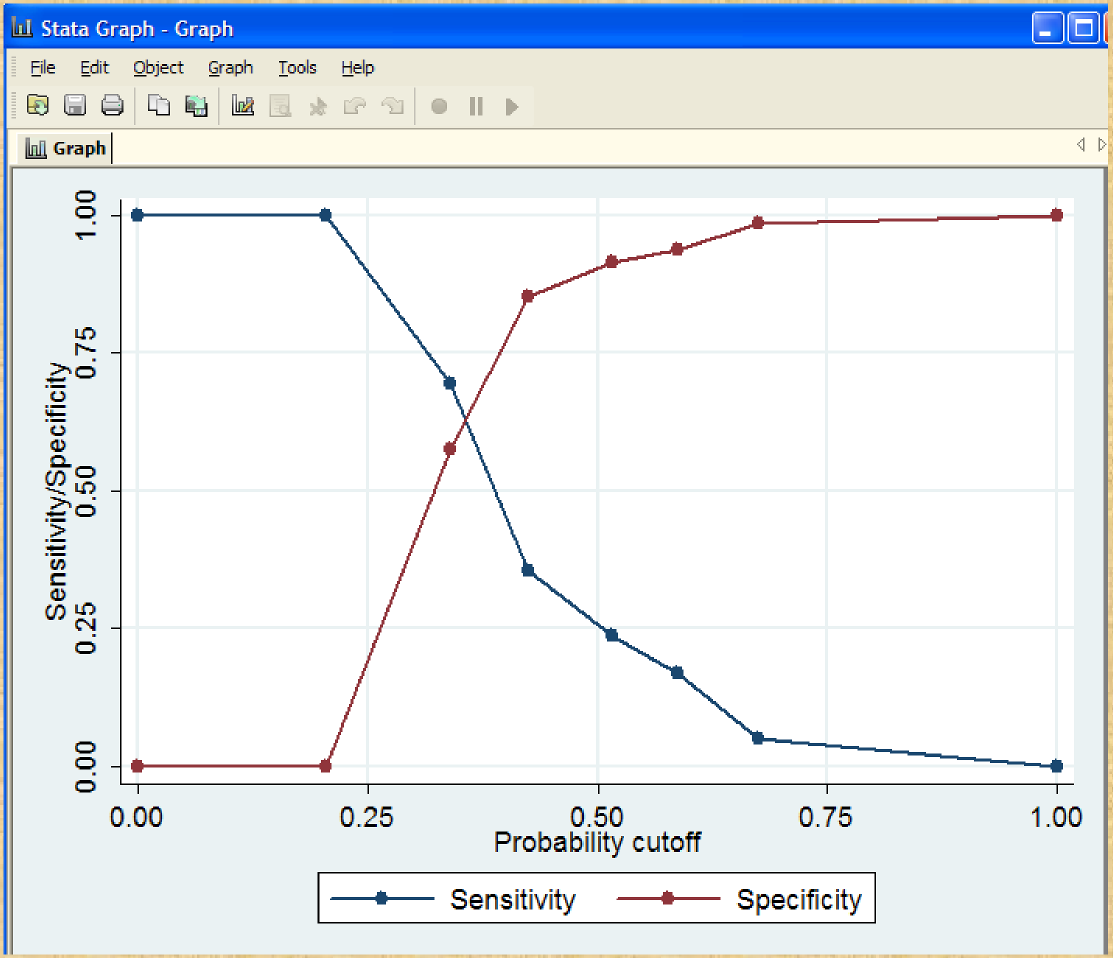
Receiver Operating Characteristic Curve



Area under ROC curve = 0.6628

Stata postestimation command: lroc

Stata Postestimation command: lsens



Regression analysis for Ordinal Dependent Variables

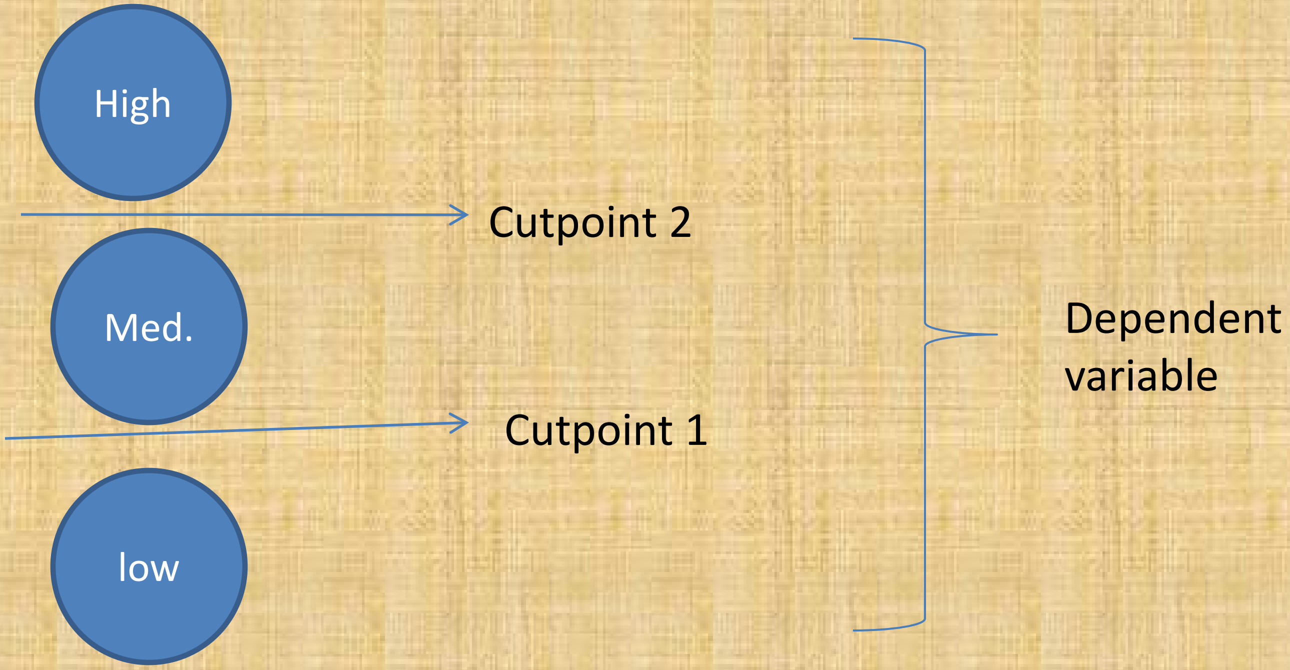
- Ordinal logistic regression using cumulative logits
- Ordinal probit regression

Ordinal logistic regression

- Is the dependent variable an ordered typology? Is it actually an ordered variable?
- Ordinal logistic regression uses cumulative logits.
- Several cutpoints split the dependent variable into a reference set and the remainder for comparison with the reference set. These cutpoint usually increase with the number of levels in the dependent variable.
- There are # levels minus 1 cutpoints for the dependent variable.

Cumulative logits

- Suppose a dependent variable has three ordered categories: low, medium, and high.



High

Cutpoint 2

Med.

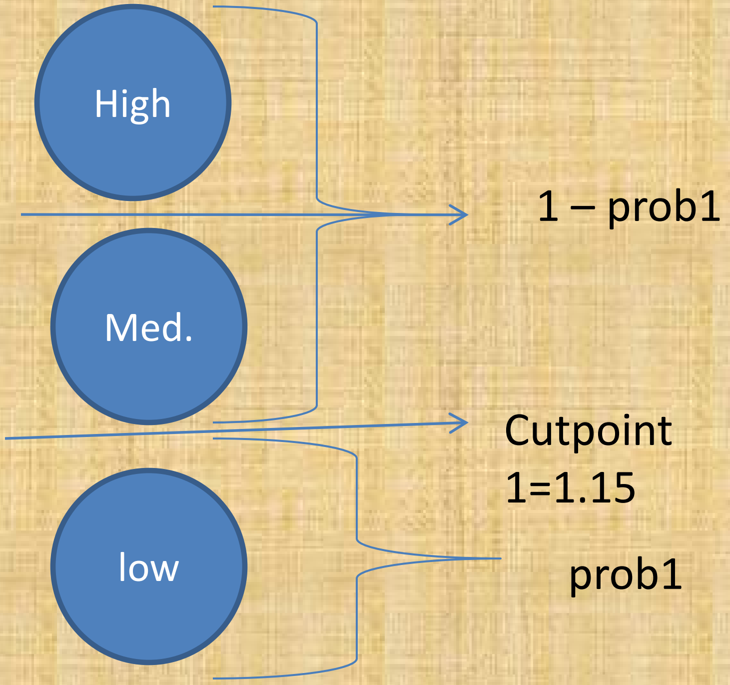
Dependent variable

Cutpoint 1

low

Cumulative logits

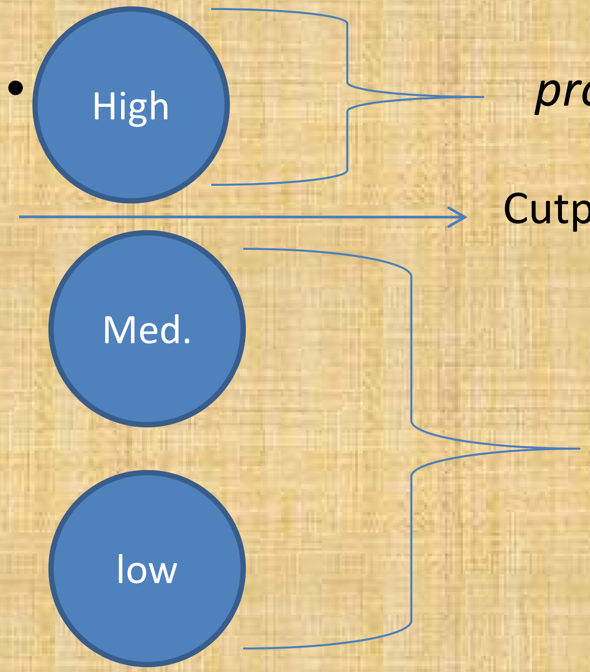
- Logit 1 uses cutpoint 1 to divide the sets into probability 1 and 1- probability 1.



Logit1 = ln(p/(1-p)

Cumulative logits

Logit 2 uses cutpoint 2 to divide the sets into probability 2 and 1- probability 2.



- High

prob2

Logit2 = ln(prob 2/(1-prob2)

Cutpoint 2=3.24

Med.

1- prob2

low

Assumptions of the ordinal logistic regression model

- Logits (ln(odds)) are linearly related to the predictors, such that an ordered structure of response (is preserved without necessarily revealing the precise extent of this ordering) of the dependent variable is maintained with respect to each predictor variable.
- Linearity and additivity: Regression coefficients are independent of the cut-point for the level of Y employed. This prevents any interaction between the X variables from being significant.
- Additivity: **Proportional odds assumption(parallel regression assumption)** holds: There are no significant interactions among independent variables. If interactions were significant, then there would be valid nonlinear or multiplicative effects.

Formulation of this assumption

- Assume that the cutpoints are represented by τ1, τ2, and τ3 . The response variable measure low, medium, and high satisfaction wrt a treatment.

$$
\begin{aligned}
\Pr(y \le \mathbf{1} \mid x) &= F(\tau_{\mathbf{1}} - \beta x) \\
\Pr(y \le \mathbf{2} \mid x) &= F(\tau_{\mathbf{2}} - \beta x) \\
\Pr(y \le \mathbf{3} \mid x) &= F(\tau_{\mathbf{3}} - \beta x) \\
\vdots \qquad & \qquad \vdots \\
\Pr(y \le J \mid x) &= F(\tau_{j} - \beta x)
\end{aligned}
$$

F=cumulative
probability
density function.

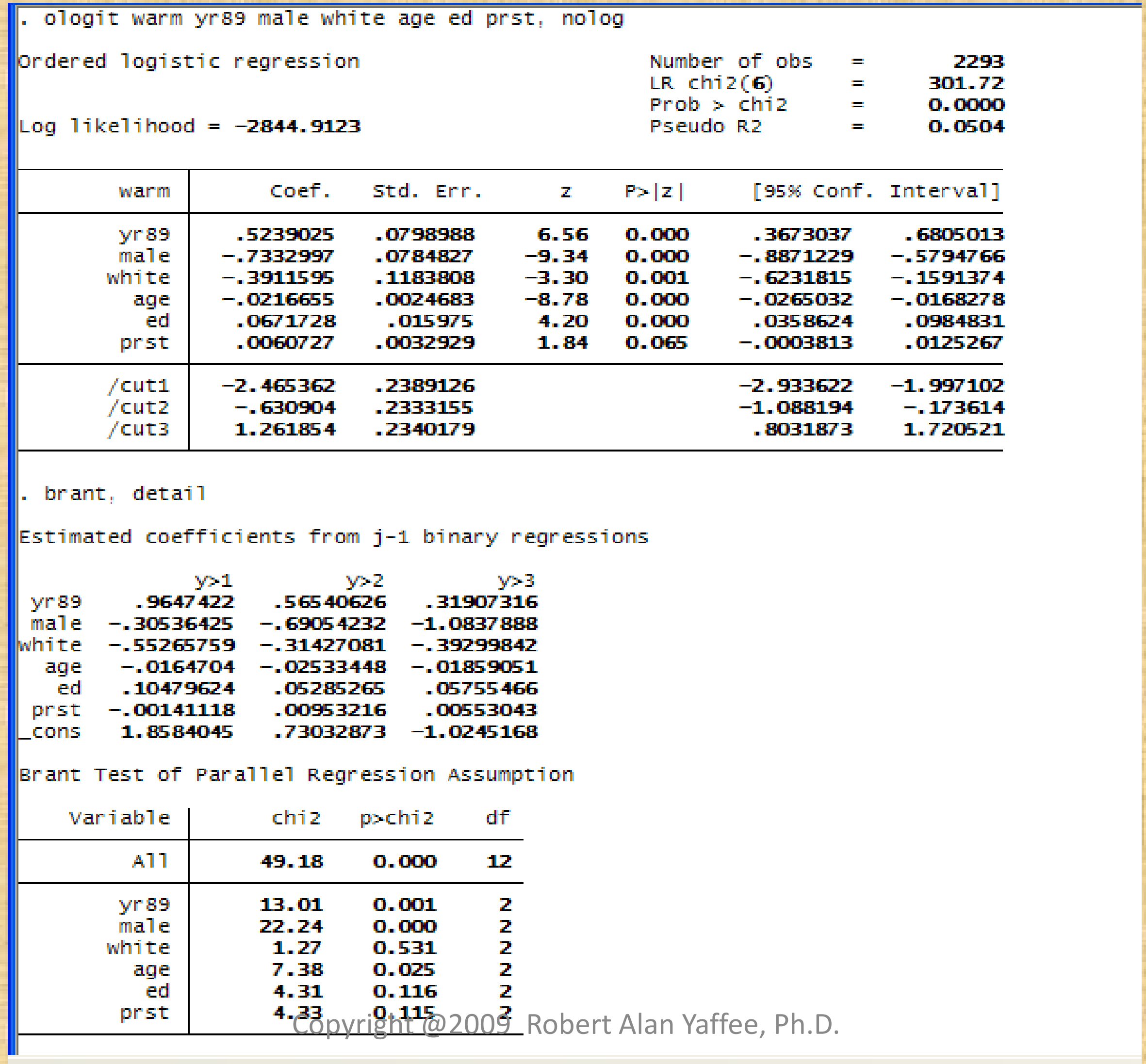
Stata’s estimate

Model parameter		Stata estimate		different Parameterization	
β_0		$B_0 - B_0 = 0$		$B_0 - \tau_1$	
τ_1		$\tau_1 - B_0$		$\tau_1 - \tau_1 = 0$	
τ_2		$\tau_2 - B_0$		$\tau_2 - \tau_1$	
τ_3		$\tau_3 - B_0$		$\tau_3 - \tau_1$	

J. Scott Long and Jeremy Freese ,2nd ed., 2006, p.196

Tau values are the cutpoints

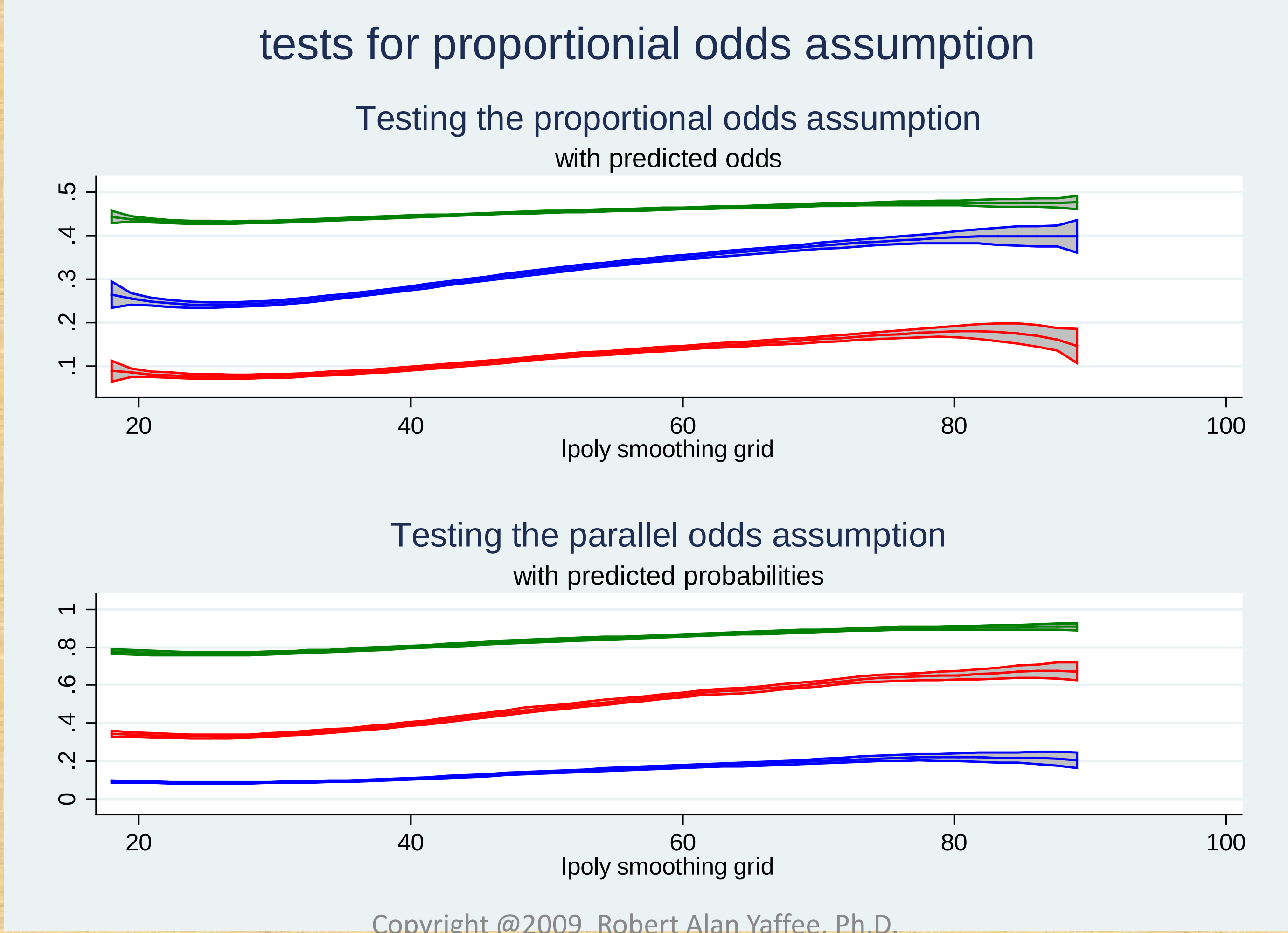
The Brant test of this assumption significant result=> violation of || odds



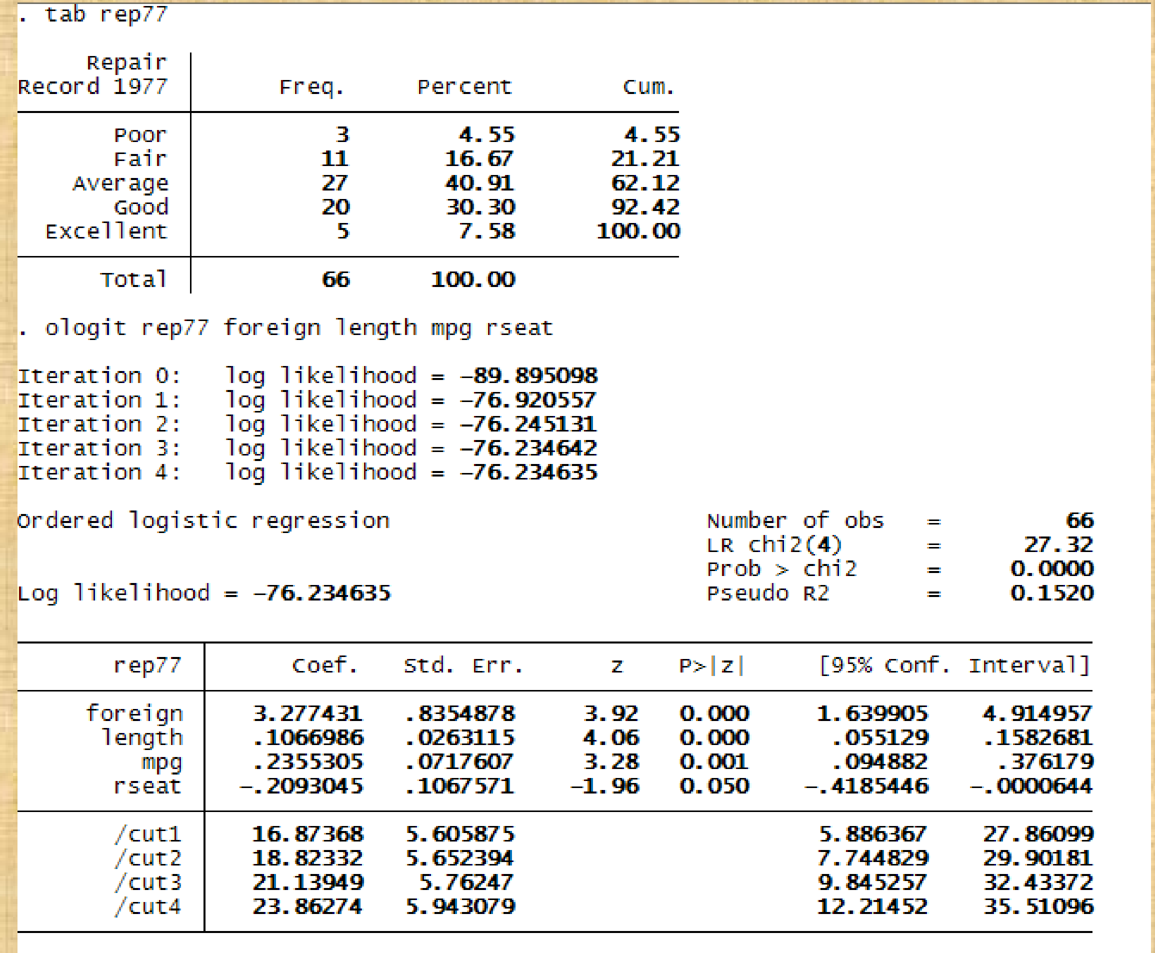
```
. ologit warm yr89 male white age ed prst, nolog

Ordered logistic regression                       Number of obs   =       2293
                                                  LR chi2(6)      =     301.72
                                                  Prob > chi2     =     0.0000
Log likelihood = -2844.9123                       Pseudo R2       =     0.0504

------------------------------------------------------------------------------
        warm |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        yr89 |   .5239025   .0798988     6.56   0.000     .3673037    .6805013
        male |  -.7332997   .0784827    -9.34   0.000    -.8871229   -.5794766
       white |  -.3911595   .1183808    -3.30   0.001    -.6231815   -.1591374
         age |  -.0216655   .0024683    -8.78   0.000    -.0265032   -.0168278
          ed |   .0671728    .015975     4.20   0.000     .0358624    .0984831
        prst |   .0060727   .0032929     1.84   0.065    -.0003813    .0125267
-------------+----------------------------------------------------------------
       /cut1 |  -2.465362   .2389126                     -2.933622   -1.997102
       /cut2 |   -.630904   .2333155                     -1.088194    -.173614
       /cut3 |   1.261854   .2340179                      .8031873    1.720521
------------------------------------------------------------------------------

. brant, detail

Estimated coefficients from j-1 binary regressions

              y>1          y>2          y>3
 yr89     .9647422    .56540626    .31907316
 male   -.30536425   -.69054232   -1.0837888
white   -.55265759   -.31427081   -.39299842
  age    -.0164704   -.02533448   -.01859051
   ed    .10479624    .05285265    .05755466
 prst   -.00141118    .00953216    .00553043
_cons    1.8584045    .73032873   -1.0245168

Brant Test of Parallel Regression Assumption

    Variable |      chi2   p>chi2    df
-------------+--------------------------
         All |     49.18    0.000    12
-------------+--------------------------
        yr89 |     13.01    0.001     2
        male |     22.24    0.000     2
       white |      1.27    0.531     2
         age |      7.38    0.025     2
          ed |      4.31    0.116     2
        prst |      4.33    0.115     2
----------------------------------------
```

Graphical test of the Proportional Odds assumptions

do graphpodds

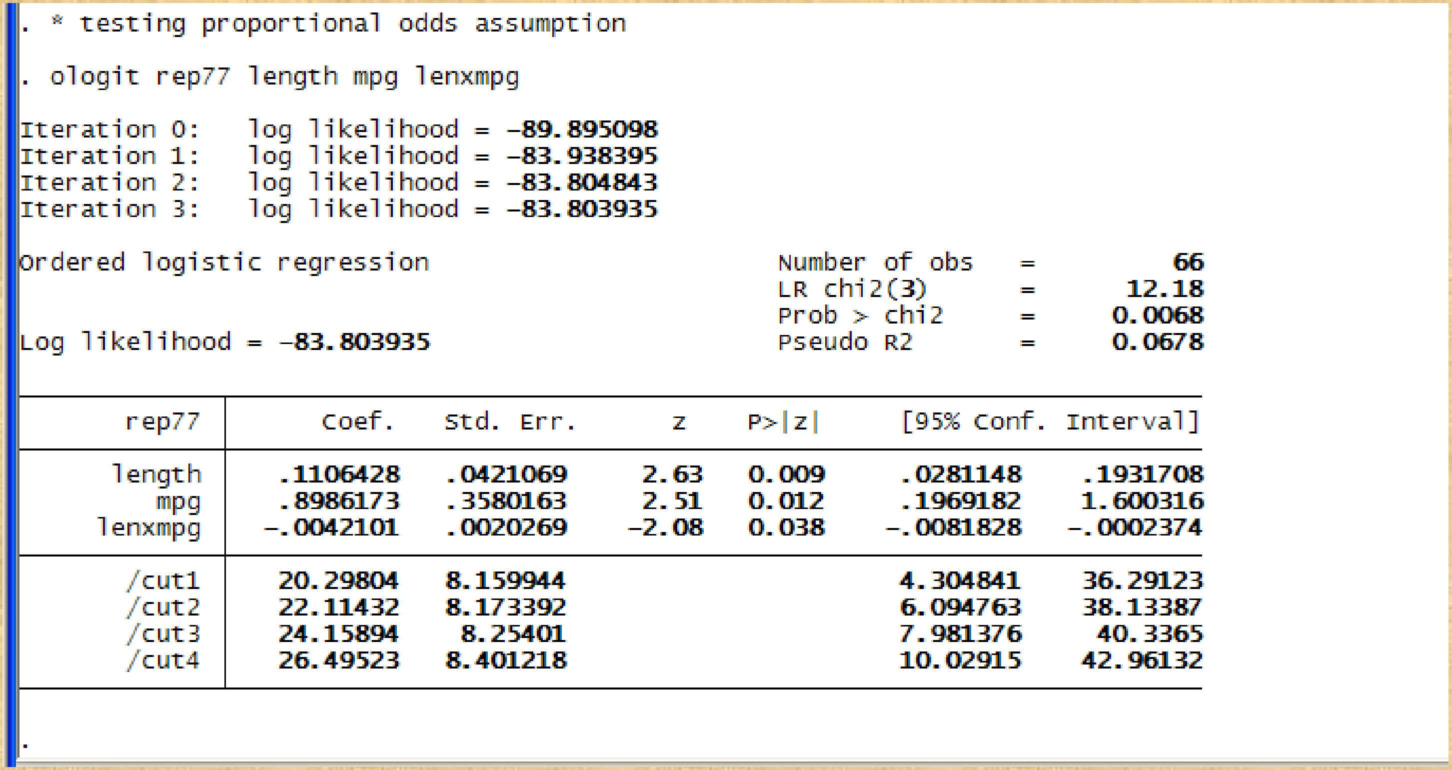


Ordinal logistic regression



```
. tab rep77

     Repair |
Record 1977 |      Freq.     Percent        Cum.
------------+-----------------------------------
       Poor |          3        4.55        4.55
       Fair |         11       16.67       21.21
    Average |         27       40.91       62.12
       Good |         20       30.30       92.42
  Excellent |          5        7.58      100.00
------------+-----------------------------------
      Total |         66      100.00

. ologit rep77 foreign length mpg rseat

Iteration 0:   log likelihood = -89.895098
Iteration 1:   log likelihood = -76.920557
Iteration 2:   log likelihood = -76.245131
Iteration 3:   log likelihood = -76.234642
Iteration 4:   log likelihood = -76.234635

Ordered logistic regression                       Number of obs   =         66
                                                  LR chi2(4)      =      27.32
                                                  Prob > chi2     =     0.0000
Log likelihood = -76.234635                       Pseudo R2       =     0.1520

------------------------------------------------------------------------------
       rep77 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     foreign |   3.277431   .8354878     3.92   0.000     1.639905    4.914957
      length |   .1066986   .0263115     4.06   0.000      .055129    .1582681
         mpg |   .2355305   .0717607     3.28   0.001      .094882     .376179
       rseat |  -.2093045   .1067571    -1.96   0.050    -.4185446   -.0000644
-------------+----------------------------------------------------------------
       /cut1 |   16.87368   5.605875                      5.886367    27.86099
       /cut2 |   18.82332   5.652394                      7.744829    29.90181
       /cut3 |   21.13949    5.76247                      9.845257    32.43372
       /cut4 |   23.86274   5.943079                      12.21452    35.51096
------------------------------------------------------------------------------
```

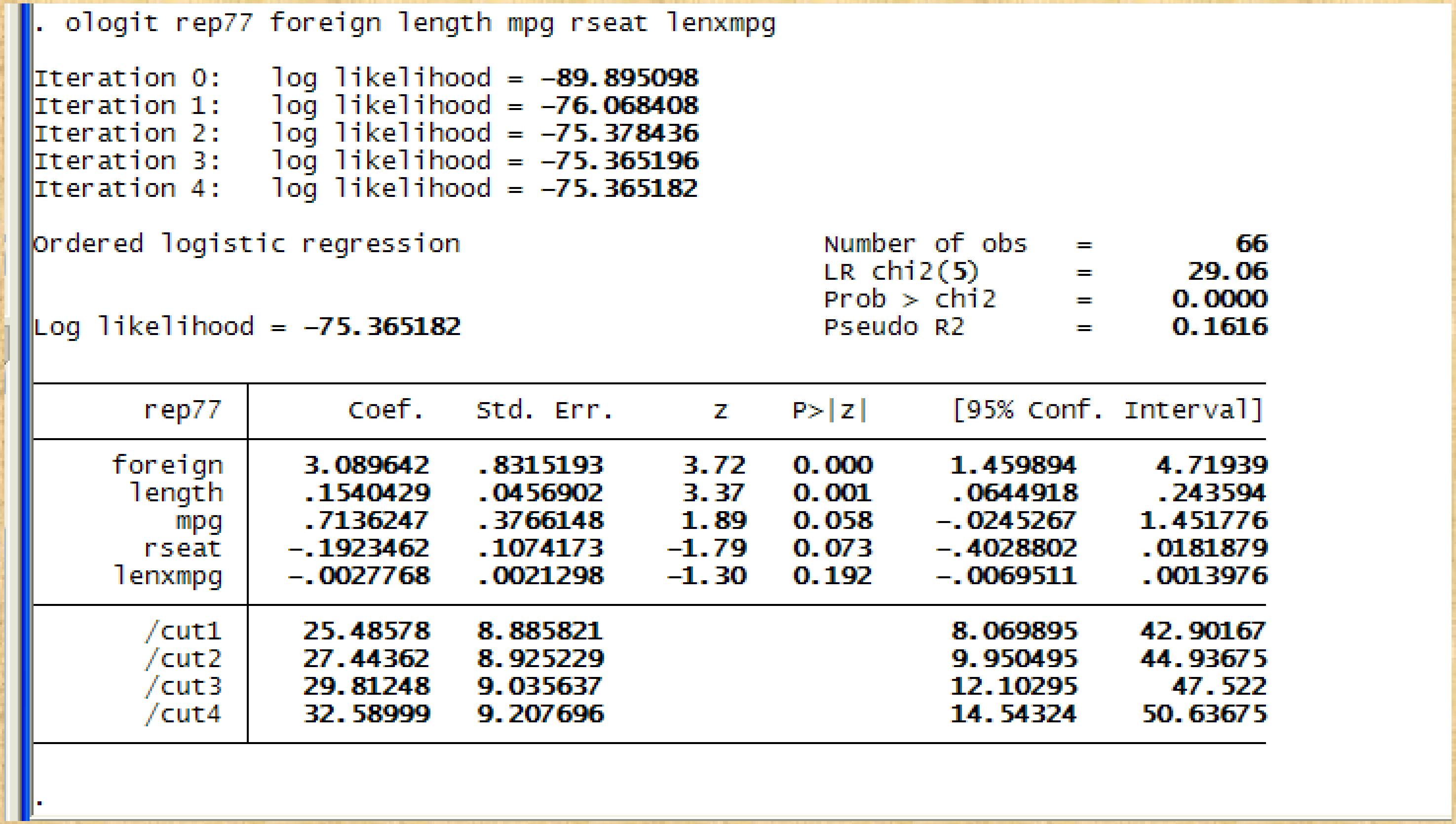
Testing proportional odds assumption



```
. * testing proportional odds assumption

. ologit rep77 length mpg lenxmpg

Iteration 0:   log likelihood = -89.895098
Iteration 1:   log likelihood = -83.938395
Iteration 2:   log likelihood = -83.804843
Iteration 3:   log likelihood = -83.803935

Ordered logistic regression                       Number of obs   =         66
                                                  LR chi2(3)      =      12.18
                                                  Prob > chi2     =     0.0068
Log likelihood = -83.803935                       Pseudo R2       =     0.0678

------------------------------------------------------------------------------
       rep77 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      length |   .1106428   .0421069     2.63   0.009     .0281148    .1931708
         mpg |   .8986173   .3580163     2.51   0.012     .1969182    1.600316
     lenxmpg |  -.0042101   .0020269    -2.08   0.038    -.0081828   -.0002374
-------------+----------------------------------------------------------------
       /cut1 |   20.29804   8.159944                      4.304841    36.29123
       /cut2 |   22.11432   8.173392                      6.094763    38.13387
       /cut3 |   24.15894    8.25401                      7.981376     40.3365
       /cut4 |   26.49523   8.401218                      10.02915    42.96132
------------------------------------------------------------------------------

.
```

Testing the interaction is a test of the proportional odds assumption.

Corrected Ordinal Logistic Regression



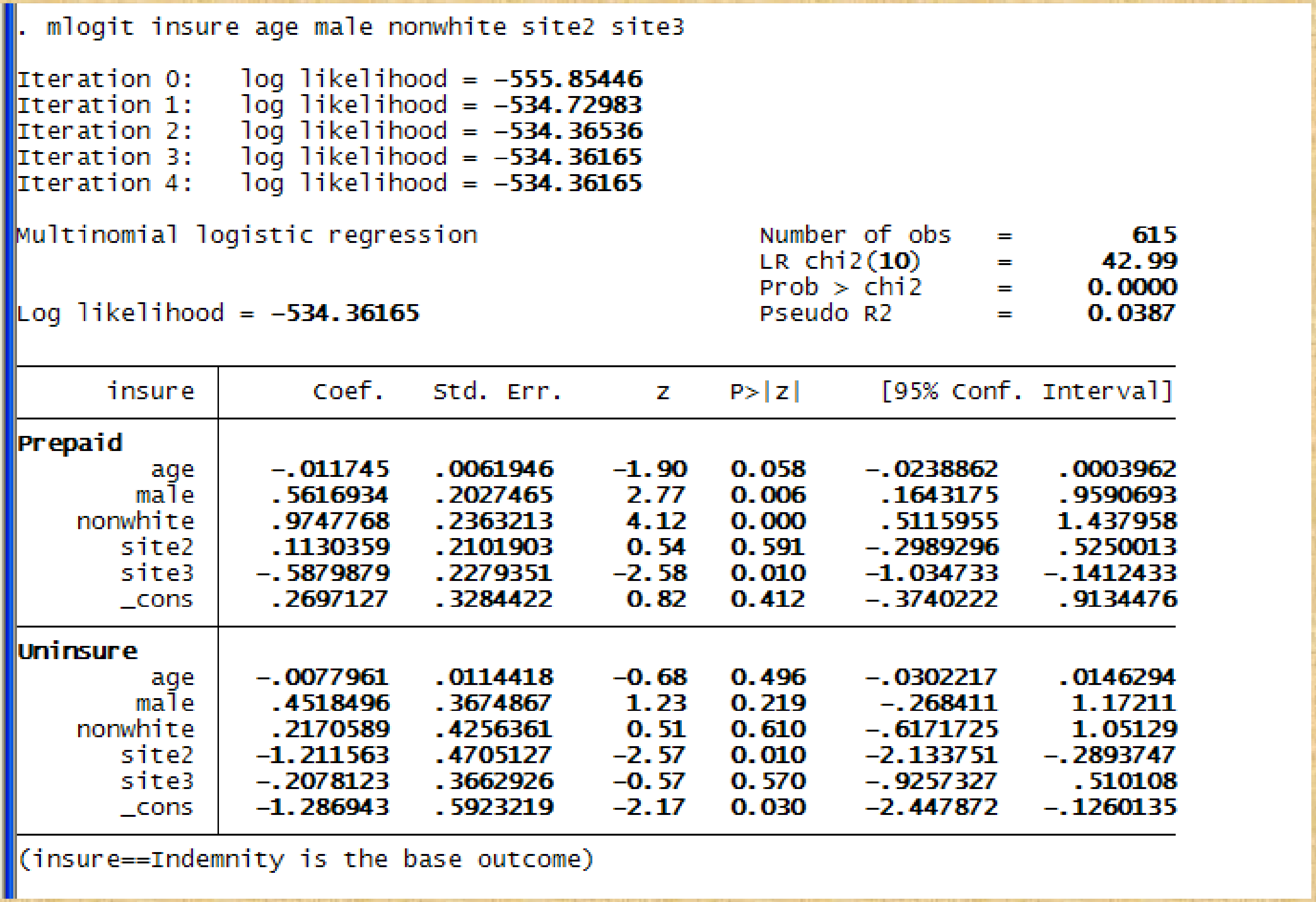
```
. ologit rep77 foreign length mpg rseat lenxmpg

Iteration 0:   log likelihood = -89.895098
Iteration 1:   log likelihood = -76.068408
Iteration 2:   log likelihood = -75.378436
Iteration 3:   log likelihood = -75.365196
Iteration 4:   log likelihood = -75.365182

Ordered logistic regression                       Number of obs   =         66
                                                  LR chi2(5)      =      29.06
                                                  Prob > chi2     =     0.0000
Log likelihood = -75.365182                       Pseudo R2       =     0.1616

------------------------------------------------------------------------------
       rep77 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     foreign |   3.089642   .8315193     3.72   0.000     1.459894     4.71939
      length |   .1540429   .0456902     3.37   0.001     .0644918     .243594
         mpg |   .7136247   .3766148     1.89   0.058    -.0245267    1.451776
       rseat |  -.1923462   .1074173    -1.79   0.073    -.4028802    .0181879
     lenxmpg |  -.0027768   .0021298    -1.30   0.192    -.0069511    .0013976
-------------+----------------------------------------------------------------
       /cut1 |   25.48578   8.885821                      8.069895    42.90167
       /cut2 |   27.44362   8.925229                      9.950495    44.93675
       /cut3 |   29.81248   9.035637                      12.10295      47.522
       /cut4 |   32.58999   9.207696                      14.54324    50.63675
------------------------------------------------------------------------------

.
```

Ordered Probit models are an alternative

- Assumptions include constant proportionality across the cutpoints.
- An underlying normal distribution is assumed.
- The scale is such that one is dealing with standardized units.

Multinomial logistic regression

for ordinal or categorical choice in the dependent variable



```
. mlogit insure age male nonwhite site2 site3

Iteration 0:   log likelihood = -555.85446
Iteration 1:   log likelihood = -534.72983
Iteration 2:   log likelihood = -534.36536
Iteration 3:   log likelihood = -534.36165
Iteration 4:   log likelihood = -534.36165

Multinomial logistic regression                   Number of obs   =        615
                                                  LR chi2(10)     =      42.99
                                                  Prob > chi2     =     0.0000
Log likelihood = -534.36165                       Pseudo R2       =     0.0387

------------------------------------------------------------------------------
      insure |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
Prepaid      |
         age |   -.011745   .0061946    -1.90   0.058    -.0238862    .0003962
        male |   .5616934   .2027465     2.77   0.006     .1643175    .9590693
    nonwhite |   .9747768   .2363213     4.12   0.000     .5115955    1.437958
       site2 |   .1130359   .2101903     0.54   0.591    -.2989296    .5250013
       site3 |  -.5879879   .2279351    -2.58   0.010    -1.034733   -.1412433
       _cons |   .2697127   .3284422     0.82   0.412    -.3740222    .9134476
-------------+----------------------------------------------------------------
Uninsure     |
         age |  -.0077961   .0114418    -0.68   0.496    -.0302217    .0146294
        male |   .4518496   .3674867     1.23   0.219     -.268411     1.17211
    nonwhite |   .2170589   .4256361     0.51   0.610    -.6171725     1.05129
       site2 |  -1.211563   .4705127    -2.57   0.010    -2.133751   -.2893747
       site3 |  -.2078123   .3662926    -0.57   0.570    -.9257327     .510108
       _cons |  -1.286943   .5923219    -2.17   0.030    -2.447872   -.1260135
------------------------------------------------------------------------------
(insure==Indemnity is the base outcome)
```

References

- Ben Jann, personal communication about mrtab, summer 2009.
- Baum, C. Stata programming, College Station, Tx: StataCorp.
- Cameron, A.c. and Trevedi, P.K. (2008) Microeconomics using Stata, College Station, Tx: StataCorp. 147-156.
- Bollen, K. and Jackman, R. W. (1990) “Regression Diagnostics: An Expository Treatment of Outliers and Influential Cases, Fox, J. and Long, J.S. (1990), eds. Modern Methods of Data Analysis, Newbury Park: Sage, 268.

References-continued

- J. Scott Long and Jeremy Freese (2006). Regression Models for Categorical and Dependent Variables using Stata. Stata Press: College Station, Tx. Chapters 8 and 9.
- Greene, W.H. (2008). Econometric Analysis, chapters on limited dependent variables are worth reading.
- Harrell, F.E., Jr. (2001). Regression Modeling Strategies. Springer: New York, Chapter 2.
- Mitchell, Michael N. A Visual Guide to Stata Graphics, 2nd edition, 2008, sections on legends and labels for graphs is very helpful.
- Pindyck and Rubenfeld, Economic Models and Forecasts, MIT Press.

  UCLA Academic Technology Services Stata Portal on the www is recommended to all students.
- Stata User's Reference Guides. Release 10. Stata-Press. College Station, Tx.
- Winer, B.J., Brown, D.R., and Michaels, K.M. (1991). Statistical Principles of Experimental Design, 3rd ed., McGraw Hill: New York, NY, 123-129.