STATA
TECHNICAL
BULLETIN

March 1998
STB-42

A publication to promote communication among Stata users

Editor

H. Joseph Newton
Department of Statistics
Texas A & M University
College Station, Texas 77843
409-845-3142
409-845-3144 FAX
stb@stata.com EMAIL

Associate Editors

Francis X. Diebold, University of Pennsylvania
Joanne M. Garrett, University of North Carolina
Marcello Pagano, Harvard School of Public Health
James L. Powell, UC Berkeley and Princeton University
J. Patrick Royston, Imperial College School of Medicine

Subscriptions are available from Stata Corporation, email stata@stata.com, telephone 979-696-4600 or 800-STATAPC, fax 979-696-4601. Current subscription prices are posted at www.stata.com/bookstore/stb.html.

Previous Issues are available individually from StataCorp. See www.stata.com/bookstore/stbj.html for details.

Submissions to the STB, including submissions to the supporting files (programs, datasets, and help files), are on a nonexclusive, free-use basis. In particular, the author grants to StataCorp the nonexclusive right to copyright and distribute the material in accordance with the Copyright Statement below. The author also grants to StataCorp the right to freely use the ideas, including communication of the ideas to other parties, even if the material is never published in the STB. Submissions should be addressed to the Editor. Submission guidelines can be obtained from either the editor or StataCorp.

Copyright Statement. The Stata Technical Bulletin (STB) and the contents of the supporting files (programs, datasets, and help files) are copyright © by StataCorp. The contents of the supporting files (programs, datasets, and help files), may be copied or reproduced by any means whatsoever, in whole or in part, as long as any copy or reproduction includes attribution to both (1) the author and (2) the STB.

The insertions appearing in the STB may be copied or reproduced as printed copies, in whole or in part, as long as any copy or reproduction includes attribution to both (1) the author and (2) the STB. Written permission must be obtained from Stata Corporation if you wish to make electronic copies of the insertions.

Users of any of the software, ideas, data, or other materials published in the STB or the supporting files understand that such use is made without warranty of any kind, either by the STB, the author, or Stata Corporation. In particular, there is no warranty of fitness of purpose or merchantability, nor for special, incidental, or consequential damages such as loss of profits. The purpose of the STB is to promote free communication among Stata users.

The *Stata Technical Bulletin* (ISSN 1097-8879) is published six times per year by Stata Corporation. Stata is a registered trademark of Stata Corporation.

Contents of this issue — page

## stata52 Origin/noOrigin option added to sts graph command

The `sts graph` command now plots the Kaplan–Meier survival curve starting at $t = 0$; see [R] **st sts graph**. This is now the default. Specifying the new option `noorigin` returns to the previous behavior of having the graph start at the first observed exit time (failure or censoring). All other options for this command are unchanged.

### Syntax

```
sts graph [if exp] [in range] [, by(varlist) strata(varlist) adjustfor(varlist) nolabel failure

    gwood level(#) lost enter separate tmin(#) tmax(#) xasis yasis noborder noshow ]

    noorigin graph_options ]
```

The `noorigin` option is new. See [R] **st sts graph** for a description of the other options.

`noorigin` sets the earliest time plotted to correspond to the first failure or censoring time. If this option is not specified, zero is the earliest time plotted.

### Output

Using the cancer data distributed with Stata, we will plot the Kaplan–Meier product-limit estimate for patients on drug 2, and we will repeat the plot specifying the `noorigin` option which will cause the plot to begin at `studytim` = 6, the first exit time for patients on drug 2.

```
. use cancer.dta
(Patient Survival in Drug Trial)
. stset studytim died
  (output omitted)
. sort studytim
. list if drug==2, noobs
studytim      died      drug       age
       6         1         2        67
       6         0         2        65
       7         1         2        58
       9         0         2        56
      10         0         2        49
  (output omitted)
. sts graph if drug==2
( See Figure 1 below)
. sts graph if drug==2, noorigin
( See Figure 2 below)
```

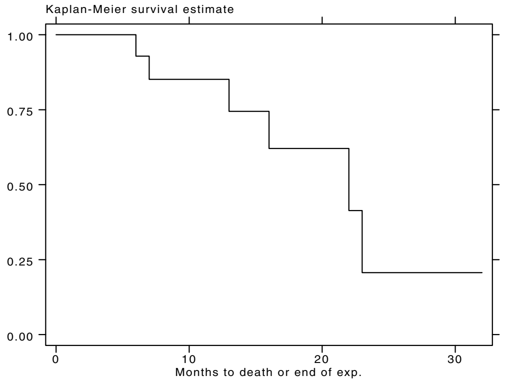

Figure 1. KM curve for patients on drug 2

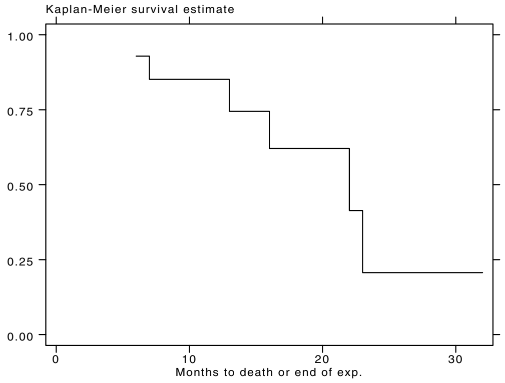

Figure 2. KM curve for patients on drug 2; noorigin specified

## dm54 Capturing comments from data dictionaries

John R. Gleason, Syracuse University, loesljrg@ican.net

Data dictionaries are an effective way to import and export data: The data are transported in an ASCII text file (a reliable mechanism on any reasonable computing platform), and the dictionary itself provides a clear description of the contents of the data file. (See [R] **infile (fixed format)** for details.) More importantly, dictionaries can be freely commented: any line with the character '`*`' in the first column will be ignored by the `infile` command when it uses the dictionary to read the data. Comments can thus be used to document the source of the data, or to explain details regarding the collection of the data—information that ought to remain attached to the data. This insert presents a command (`infilen`) that creates a Stata dataset by infiling from a fixed format data file, captures commentary from the data dictionary, and stores those comments as notes; see [R] **notes**.

To illustrate, consider the file `mammals.dat` (included with this insert) which begins thus:

```
dictionary {
*! Data originally published by Allison, T. & Cicchetti, D. V. (1976)
*! <Science, 194, 732-734>, as part of a study of correlates of sleep in
*! mammals; a subset (the first 3 variables below) appears in Table 6.6 of
*! Weisberg (1985) <Applied Regression, 2nd ed.; NY: Wiley; pp. 144-145>.
*!
*! The data show average values for 62 species of mammals; note that some
*! values are missing.  The last 3 variables below (is_prey, expos_sl, and
*! indanger) are ordered categories, where  1 = least  and  5 = most  with
*! respect to the attribute in question.  Note also that slow wave sleep
*! (swav_sl) is known as `non-dreaming' sleep, whereas paradoxical sleep
*! (pdox_sl) is known as `dreaming' sleep.
*
   str25  species    "Mammal species"
          body_wt    "Body Weight (kg)"
          brain_wt   "Brain Weight (g)"
          swav_sl    "Slow wave sleep (hrs/day)"
          pdox_sl    "Paradoxical sleep (hrs/day)"
          totsleep   "Total sleep (hrs/day)"
          lifespan   "Maximum life span (yrs)"
          gestate    "Gestation time (days)"
   byte   is_prey    "Likely to be preyed upon?"
   byte   expos_sl   "Exposed during sleep?"
   byte   indanger   "In danger from other animals?"
}
         "African elephant"   6654   5712     .    .   3.3  38.6  645  3  5  3
"African giant pouched rat"      1    6.6   6.3    2   8.3   4.5   42  3  1  3
  (output omitted)
```

The command

```
. infile using mammals.dat
```

will create a Stata dataset with 11 usefully labeled variables, but the information contained in the lines beginning with the characters '`*!`' will be left behind. By contrast, the command

```
. infilen using mammals.dat
```

creates the same Stata dataset but also imbeds the lines marked with '`*!`' as 11 notes.

The syntax of `infilen` is

`infilen using` *dfilename* [`if` *exp*] [`in` *range*] [`,` `automatic using(`*filename*$_2$`) clear nlines(`*#*`)`]

The option `nlines` controls the number of data dictionary lines that will be scanned for comments; the default value of # is 50. Otherwise, the syntax of `infilen` is exactly that of `infile` for reading data with a dictionary; see [R] **infile (fixed format)** for details.

In fact, the command

```
. infilen using mammals.dat
```

first issues the command

```
. infile using mammals.dat
```

Then, `infilen` re-reads the data dictionary in `mammals.dat`, searching for lines where the string '`*!`' appears in the first two columns. This marks a special type of comment, the kind that if found in an ado-file, will be echoed to the screen by `which`; see [R] **which**. `infilen` captures such comments and then does the equivalent of

`. note:` *text of comment*

for each ‘*!’ comment. Thus, continuing our example, after the `infilen` command there will be 11 notes attached to the current dataset:

```
. notes

_dta:
  1.  Data originally published by Allison, T. & Cicchetti, D. V. (1976)
  2.  <Science, 194, 732-734>, as part of a study of correlates of sleep in
  3.  mammals; a subset (the first 3 variables below) appears in Table 6.6 of
  4.  Weisberg (1985) <Applied Regression, 2nd ed.; NY: Wiley; pp. 144-145>.
  5.  .
  6.  The data show average values for 62 species of mammals; note that some
  7.  values are missing.  The last 3 variables below (is_prey, expos_sl, and
  8.  indanger) are ordered categories, where  1 = least  and  5 = most  with
  9.  respect to the attribute in question.  Note also that slow wave sleep
 10.  (swav_sl) is known as ´non-dreaming´ sleep, whereas paradoxical sleep
 11.  (pdox_sl) is known as ´dreaming´ sleep.
```

## Remarks

Ordinary comments (those that begin with ‘*’ only) are simply ignored, while every comment with ‘*!’ in the first two columns is stored in a note. Blank comments are often used to make commentary more readable; for compatibility with the `notes` command, such comments are rendered non-blank by saving the single character ‘.’. Note 5 (above) provides an example.

Instances of the left-quote (‘`’) character will be translated into right-quotes (‘´’); otherwise, some part of the comment will be interpreted as a macro expansion. Unfortunately, there is no way to perform a similar translation of the double-quote character (‘"’): Comments to be saved as notes may not contain ‘"’, and its presence will produce an unhelpful error message. (On the other hand, the characters ‘`’ and ‘"’ have no special status in ordinary comments.)

A maximum of 80 characters following the string ‘*!’ will be captured from comments.

## Acknowledgment

This project was supported by a grant R01-MH54929 from the National Institute on Mental Health to Michael P. Carey.

## gr28 A graphical procedure to test equality of variances

Aurelio Tobias, Institut Municipal d’Investigacio Medica (IMIM), Barcelona, atobias@imim.es

The problem of testing the equality of several variances arises in many areas. There are a wide variety of tests available for testing equality of variances of normal populations. In Stata, only Bartlett’s (1937) test for equality of variances is available in the `oneway` command.

The `grvar` command performs a graphical display for testing the equality of variances of normal populations. Rao and Krishna (1997) following Ott’s (1967) graphical procedure for testing the equality of means (ANOM) have developed this method. This graph tests equality of two or more variances from normal populations, simultaneously demonstrating statistical (the real presence of an effect) and engineering significance (the magnitude of seriousness of the effect). This graphical method could be more useful for non-statisticians than classical tests for equal variances, becoming a good supplement to the `oneway` command.

## Graphical method procedure

Consider $k$ independent samples, the $i$th of which is of size $n_i$, obtained from $k$ normal populations with means $\mu_i$ and variances $\theta_i^2$. We would like to test the null hypothesis $\theta_i^2 = \theta^2$, for all $i$, against the alternative that at least one equality does not hold. Then, we should follow four steps:

1. Calculate the variance of the $i$th sample ($S_i^2$), and then $S_i^{2/3}$, for $i = 1, \ldots, k$.
2. Calculate: $A_i = \dfrac{(9n_i - 11)}{9(n_i - 1)}$, $\quad C_i = \dfrac{18(n_i - 1)}{(9n_i - 11)^2}$, $\quad N = \sum_{i=1}^{k} n_i$, $\quad t_i = \dfrac{S_i^{2/3}}{A_i}$,

$$\bar{t}_w = \frac{1}{N}\sum_{i=1}^{k} n_i t_i, \qquad SE(t_i - \bar{t}_w) = \frac{\bar{t}_w}{N}\sqrt{N(N-2n_i)C_i + \sum_{i=1}^{k} n_i^2 C_i}$$

3. Calculate the lower and upper decision lines (LDL and UDL) for the comparison of each of the $t_i$ terms: LDL $= \bar{t}_w - z_{a/2k} SE(t_i - \bar{t}_w)$, $\qquad$ UDL $= \bar{t}_w + z_{a/2k} SE(t_i - \bar{t}_w)$, where $z_{a/2k}$ is the upper $(a/2k)$ percentage point of the standard normal distribution.
4. Finally, plot $t_i$ against LDL and UDL.

Derivations of the formulas are given in the appendix of Rao and Krishna (1997).

If any of the points plotted are outside the decision lines, we can reject the null hypothesis and conclude that there are statistically significant differences among the population variances.

## Syntax

The **grvar** command works on a dataset containing the estimated standard deviation, *sd*, and the sample size, *n*, for each sample or study. The syntax is as follows:

**grvar** *sd n* [**if** *exp*] [**in** *range*] [, **id(***strvar***) level(***#***)** *graph_options*]

## Options

**id(***strvar***)** is used to label the studies. If the data contain a labeled numeric variable, it can also be used.

**level(***#***)** specifies the level of significance of the test, in percent. The default is **level(5)** (*sic*).

*graph_options* are any of the options allowed with **graph**, **twoway** except **xlabel()**, **t2()**, **b2()**, **yline()**, **symbol()**, **pen()**, and **connect()**.

## Example

The **grvar** command has been tested using the example supplied by Rao and Krishna (1997). We have eight samples we call A, B, C, D, E, F, G, and H of lengths of the third molar of eight species of the Condylarth Hyposodus (Sokal and Rohlf 1969, 370–371).

```
. describe
Contains data from sokal.dta
  obs:             8
 vars:             4
 size:            88
-------------------------------------------------------------------------
   1. sample      byte   %8.0g                  sample number
   2. sampleid    str1   %9s                    sample identification
   3. n           byte   %8.0g                  sample size
   4. sd          float  %9.0g                  standard deviation
-------------------------------------------------------------------------
. list, noobs
       sample   sampleid         n         sd
            1          A        18   .2658947
            2          B        13   .3803945
            3          C        17    .153948
            4          D        16   .2891366
            5          E         8   .4678675
            6          F        11   .4207137
            7          G        10   .2812472
            8          H        10   .4828043

. grvar sd n, id(sampleid)
```

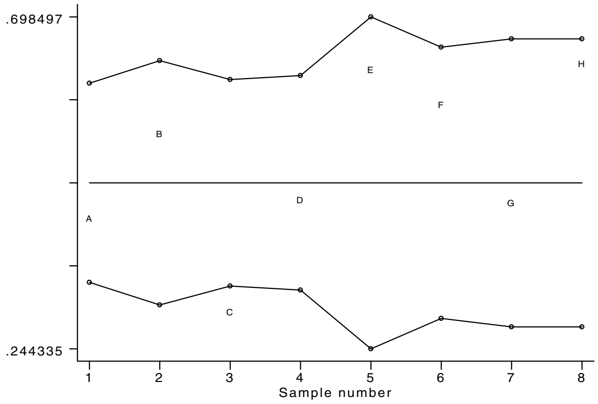

Figure 1. Graphical display to test equality of variances

We observe from Figure 1, that sample C is allocated outside its lower decision line. Hence we conclude that the population variances are not equal. Sokal and Rohlf (1969) using Bartlett's test at the 5% level obtained the same conclusion. Also Rao and Krishna (1997) got the same result using the Neyman–Pearson test, the Lehmann test, the Q test, the Bartlett 3 test, the Lehmann 2 test, and the Samiuddin–Atiqullah test with their respective critical values at the 5% level.

### Individual or frequency records

As with other Stata commands, `grvar` works on data contained in frequency records, one for each sample or study. If we have primary data, that is individual records, we must convert to frequency records using the `collapse` and `byvar` Stata commands.

### References

Bartlett, M. S. 1937. Properties of sufficiency and statistical tests. *Proceedings of the Royal Statistical Society, Series A* 160: 268–282.

Ott, E. R. 1967. Analysis of means—a graphical procedure. *Industrial and Quality Control* 24: 101–109.

Rao, C. V. and S. H. Krishna. 1997. A graphical method for testing the equality of several variances. *Journal of Applied Statistics* 24: 279–287.

Sokal, R. R. and J. F. Rohlf. 1969. *Biometry*. San Francisco, CA: W. H. Freeman.

## sbe16.1 New syntax and output for the meta-analysis command

Stephen Sharp, London School of Hygiene and Tropical Medicine, stephen.sharp@lshtm.ac.uk
Jonathan Sterne, United Medical and Dental Schools, UK, j.sterne@umds.ac.uk

The command `meta`, which performs the statistical methods involved in a systematic review of a set of individual studies, reporting the results in text and also optionally in a graph, was released in STB-38 (Sharp and Sterne 1997). This re-release contains a more flexible syntax than the earlier version and new-style output.

### Syntax

The command `meta` works on a dataset containing an estimate of the effect and its variability for each study. The syntax is

`meta` {*theta* | *exp(theta)*} {*se_theta* | *var_theta* | *ll ul* [*cl*]} [`if` *exp*] [`in` *range*]
[, `var ci eform print ebayes level(`*#*`) graph(`*string*`) id(`*varname*`)`
`fmult(`*#*`) boxysca(`*#*`) boxshad(`*#*`) cline ltrunc(`*#*`) rtrunc(`*#*`) *graph options*]

With this new syntax, the user provides the effect estimate as *theta* (i.e., a log risk ratio, a log odds ratio, or other measure of effect). Likewise, the user supplies a measure of *theta*'s variability (i.e., its standard error, *se_theta*, or its variance, *var_theta*). Alternatively, the user provides *exp(theta)* (e.g. a risk ratio or odds ratio) and its confidence interval, (*ll*, *ul*). The options remain the same as in the original version of `meta`. Type `help meta` for details.

### Required input variables

| | |
|---|---|
| *theta* | the effect estimate |
| *se_theta* | the corresponding standard error |

or

| | |
|---|---|
| *theta* | the effect estimate |
| *var_theta* | the corresponding variance |

NB: (`var` option must be included)

or

| | |
|---|---|
| *exp(theta)* | the risk (or odds) ratio |
| *ll* | the lower limit of the risk ratio's confidence interval |
| *ul* | the upper limit of the risk ratio's confidence interval |
| *cl* | optional (see below) |

NB: (`ci` option must be included)

## Optional input variable

*cl* contains the confidence level of the confidence interval defined by *ll* and *ul*. If *cl* is not provided, the procedure assumes that each confidence interval is at the 95% level. *cl* allows the user to provide the confidence level, by study, when the confidence interval is not at the default level. *cl* can be specified with or without a decimal point. For example, .90 and 90 are equivalent and may be mixed (i.e., 90, .95, 80, .90 etc.).

## Example of new syntax and output

Below we illustrate the new-style command syntax and output from `meta` applied to the example in Sharp and Sterne (1997), based on nine randomized clinical trials of the use of diuretics for pre-eclampsia in pregnancy (Collins et al. 1985).

```
. use diuretic, clear
(Diuretics and pre-eclampsia)

. describe

Contains data from diuretic.dta
  obs:             9                         Diuretics and pre-eclampsia
 vars:             6
 size:           189
-------------------------------------------------------------------------------
   1. trial       byte   %9.0g       trlab      trial identity number
   2. trialid     str8   %9s                    trial first author
   3. nt          int    %9.0g                  total treated patients
   4. nc          int    %9.0g                  total control patients
   5. rt          int    %9.0g                  pre-eclampsia treated
   6. rc          int    %9.0g                  pre-eclampsia control
-------------------------------------------------------------------------------
```

With the earlier version of `meta` it was necessary to calculate the estimated effect and standard error on the *log* scale. With the re-release, the user can specify either the odds ratio (on either the original or log scale), the standard error, the variance, or the confidence limits. So for example, suppose the odds ratio and 90% confidence limits are calculated.

```
. gen or=(rt/(nt-rt))/(rc/(nc-rc))

. gen l90=or/exp(1.645*sqrt((1/rc)+(1/(nc-rc))+(1/rt)+(1/(nt-rt))))

. gen u90=or*exp(1.645*sqrt((1/rc)+(1/(nc-rc))+(1/rt)+(1/(nt-rt))))

. l l90 or u90, noobs

      l90         or        u90
 .5404617   1.042735   2.011792
 .2257975   .3970588    .698217
  .162642   .3255814    .651758
  .093229   .2291667   .5633156
 .1426823   .2488202   .4339115
 .6090319   .7431262   .9067448
 .4347324   .7698574   1.363322
 .7602696   2.970588   11.60693
 .7458201   1.144886   1.757481

. gen cl=90

. meta or l90 u90 cl, eform ci

Meta-analysis (exponential form)

       |  Pooled      95% CI         Asymptotic      No. of
Method |     Est   Lower   Upper  z_value  p_value   studies
-------+---------------------------------------------------
Fixed  |   0.672   0.564   0.800   -4.454    0.000      9
Random |   0.596   0.400   0.889   -2.537    0.011

Test for heterogeneity: Q= 27.260 on 8 degrees of freedom (p= 0.001)
Moment-based estimate of between studies variance =  0.230
```

Note that although in this example the odds ratios and confidence intervals had to be calculated before the `meta` command could be used, in practice such variables may already have been created in the dataset for other purposes, and the new syntax saves the user from having to generate the log odds ratio and standard error. If the effect estimate is a difference in means, the same alternatives for specifying the variability are available.

## Graphics

The graphics options remain as in the earlier version of the command (Sharp and Sterne 1997).

## Acknowledgment

We are grateful to Thomas Steichen for his helpful suggestions regarding the command syntax and output.

## References

Collins, R., S. Yusuf, and R. Peto. 1985. Overview of randomised trials of diuretics in pregnancy. *British Medical Journal* 290: 17–23.

Sharp, S. and J. Sterne. 1997. sbe16: Meta-analysis. *Stata Technical Bulletin* 38: 9–14.

## sbe21 Adjusted population attributable fractions from logistic regression

Anthony R. Brady, Public Health Laboratory Service Statistics Unit, UK, tbrady@phls.co.uk

The population attributable fraction (or risk) is an epidemiological measure which attempts to quantify the proportion of disease incidence which is due to a particular factor (or "exposure"). Implicit in this measure are very strong assumptions about causality, namely that removing the exposure from the population would reduce disease incidence. However, where this assumption is met the attributable fraction can provide useful insights into the important causes of disease and thereby inform intervention strategies.

The usual measure of attributable fraction (such as that reported by Stata's epitab `cc` command) is an unadjusted measure which can be misleading in the presence of confounding. Estimating the attributable fraction (AF) from within a logistic regression framework enables confounders to be taken into account and allows estimation of the summary attributable fraction for a set of exposures. This might be of interest to those planning an intervention "package".

`aflogit` recovers estimates of the population attributable fraction for those terms specified by the user from the most recent unconditional logistic regression model. It also reports a summary attributable fraction for all the terms combined. Asymptotic standard errors and 95% confidence intervals are calculated and reported. Note that `aflogit` may also be used following Poisson regression (where the attributable fraction is interpreted as the proportion of the disease rate which is due to the exposure), but cannot be used after conditional logistic regression (`clogit`).

Negative estimates of the attributable fraction correspond to protective effects (relative risk $< 1$) where the interpretation is the proportion increase expected in the number with the disease if the protective exposure were absent from the population. Perhaps confusingly this is not the usual definition of the preventable fraction or the prevented fraction (see Last 1983).

## Background

The unadjusted population attributable fraction is given by

$$\mathrm{AF} = \frac{\Pr(\text{disease}) - \Pr(\text{disease}|\text{not exposed})}{\Pr(\text{disease})}$$

which is equivalent to

$$\frac{\text{Observed no. ill} - \text{expected no. ill on removal of the exposure}}{\text{Observed no. ill}}$$

This can be rearranged to give

$$\mathrm{AF} = \Pr(\text{exposed}|\text{disease})\left(1 - \frac{1}{\mathrm{RR}}\right)$$

where RR is the relative risk. So all that is needed to estimate the AF is the distribution of exposure among the ill, Pr(exposed | disease), and the relative risk. The distribution of exposure among the ill is assumed to apply to the population from which the sample was taken. Therefore, any sampling procedure which does not randomly sample those with the disease will invalidate estimation of the AF. Under the rare disease assumption, the relative risk is approximated by the odds ratio and so case–control data can be used to estimate the AF. This is the formula the epitab `cc` command uses.

Notice that the attributable fraction increases both as the exposure becomes more common and as the relative risk becomes larger. In addition, uncertainty in the estimated AF will come not only from the uncertainty in the relative risk but also from the distribution of exposure. This means that a risk factor which is significantly different from one may not necessarily have an attributable fraction which is significantly bigger than zero.

The principle behind using logistic regression to estimate attributable fractions is to fit the required model and then calculate the number of cases predicted by the model. This is done by generating the predicted probability of disease for each individual in the model (using the `predict` command) and then summing over all individuals. If the model has been fitted to the entire dataset then this should equal the observed number of cases of disease in the study. Next the exposure effect is "removed" from the dataset by resetting exposure covariates to zero (or some reference level for continuous variables). The `predict` command is used again to generate the predicted probability of disease for each individual with the new covariate values but under the same logistic model. Summing these probabilities gives the number of cases of disease one would expect if the exposure were absent from the population. Calculating the AF is then a simple matter of applying the second equation above.

More formally, the adjusted attributable fraction is given by

$$\widehat{\lambda}_A = 1 - \frac{n' r_z}{n' r_x}$$

where $n$ is the column vector of sample sizes, $n_i$ ($\equiv$1 where records are on individuals), and $r_x$ and $r_z$ are vectors of predicted probabilities from the model under the original covariates ($x$) and the modified covariates ($z$) respectively.

An asymptotic variance formula is given in Greenland and Drescher (1993), which they recommend be applied on the $\log(1 - \text{AF})$ scale. This is most helpfully considered as

$$\text{var}(\widehat{\lambda}) = \left(\frac{E}{O}\right)^2 \left(\frac{\text{var}_E}{E^2} + \frac{\text{var}_O}{O^2} - \frac{2\text{cov}(O,E)}{OE}\right)$$

where $O$ and $E$ refer to the number of observed and expected cases of disease respectively.

A different approach must be taken for case–control data since the Pr(disease) is fixed by the study design (0.5 if one control per case is used). Bruzzi et al. (1985) explain this approach, although it is also presented by Greenland and Drescher (1993) with some minor modifications. Essentially the method is to generalize the third equation to $J$ strata (or levels of exposure) which gives

$$\text{AF} = 1 - \sum_{j=0}^{J} \frac{\rho_j}{RR_j}$$

where $\rho_j$ = Pr(exposure group $j$ | disease) and $RR_j$ is the relative risk comparing stratum $j$ with stratum 0, the baseline group. The $RR_j$ may of course be approximated by $OR_j$ if the rare disease assumption is justified. This simplifies to the third equation when there are only two strata (exposed and not exposed)

$$\begin{aligned} AF &= 1 - \left(\frac{\text{Pr(not exposed|disease)}}{1} + \frac{\text{Pr(exposed|disease)}}{\text{RR}}\right) \\ &= 1 - 1 + \text{Pr(exposed|disease)} - \frac{\text{Pr(exposed|disease)}}{\text{RR}} \\ &= \text{Pr(exposed|disease)}\left(1 - \frac{1}{\text{RR}}\right) \end{aligned}$$

The above generalized equation gives the summary attributable fraction for all the factors combined, that is, the proportion of the disease which is due to all the exposures under consideration. In order to estimate the effect of one risk factor while adjusting for other covariates, we modify the $RR_j$ to reflect the additional risk which the risk factor of interest contributes to stratum $j$. Strata in which the risk factor does not appear therefore have $RR = 1$. The $RR_j$ for strata including the risk factor is estimated by the adjusted odds ratio from the logistic regression model (making the rare disease assumption). Providing the risk factor of interest is dichotomous and is not involved in any interactions, the estimate of the adjusted AF reduces to substituting the adjusted odds ratio into the third equation. In the notation used by Greenland and Drescher (1993), the generalized equation is

$$\widehat{\lambda}_A = 1 - \hat{\rho}'\hat{\text{s}}$$

where $\widehat{\rho}$ is a vector containing the $\widehat{\rho}_j$ and $\widehat{\text{s}}$ is a vector of $\dfrac{1}{\widehat{RR}_j}$.

The asymptotic variance of this measure has no simple form and `aflogit` uses just the first summand in Greenland and Drescher's formula, $\text{var}(\widehat{\lambda}) \approx \text{D}'_\theta \text{C} \text{D}_\theta$ , which they report gives results nearly identical to those based on the entire formula.

## Syntax

The command `aflogit` may only be used after a logistic or poisson regression model has been fitted. The syntax is

**aflogit** [term [term ...]] [*weight*] [**if** *exp*] [**in** *range*]

[**, cc** **ref**<u>erence</u>(term = # [term = # ...]) **l**evel(#)]

`fweight`s and `aweight`s are allowed.

Without arguments, `aflogit` reports attributable fractions for all terms in the model. `aflogit` tries to pick up the same weights, `if` and `in` clauses as used by the regression model, but users should manually specify these after the `logit`, `blogit`, and `poisson` commands.

## Options

`cc` indicates that the data come from an unmatched case–control study and the estimate will be based on the adjusted odds ratios. The default is to assume the data come from a cross-sectional or cohort study and base the estimate of the attributable fraction on the predicted probabilities. It is imperative to select this option if you have case–control data since the results might be seriously misleading if you use the predicted probabilities.

`reference` determines the reference level (the non-exposure category) for each term in the model. By default this is zero for every term. For continuous variables it may be desirable to set the reference level to a nonzero value (e.g. the mean for a variable such as blood pressure).

`level(`*#*`)` specifies the significance level for confidence intervals of the coefficients.

## Example: unmatched case–control data

Jacques Benichou (1991) calculated population attributable fraction estimates for a case–control study of oesophageal cancer carried out by Tuyns et al. (1977). The main interest was in the proportion of cases attributable to excess alcohol consumption. Various logistic regression models were fitted to allow for confounders (smoking and age) and for interactions between alcohol consumption and the confounders. Alcohol consumption (the “exposure”) was collected as a four-level factor but was initially used as a dichotomous factor (0 to 79 grams per day and 80 or more grams per day).

```
. use tuyns, clear
(Oesophageal cancer and alcohol)

. describe

Contains data from tuyns.dta
  obs:            59                           Oesophageal cancer and alcohol
 vars:             7                           17 Sep 1997 17:07
 size:           649 (100.0% of memory free)
-------------------------------------------------------------------------------
   1. alcohol     byte   %9.0g       alcl       g/day
   2. agegp       byte   %9.0g       agel       Age group
   3. smoke       byte   %9.0g       smokel     g/day
   4. case        byte   %9.0g                  Cancer case?
   5. n           byte   %8.0g                  Frequency
   6. alc80       byte   %9.0g                  Alcohol 0-79 vs 80+
   7. alc40       byte   %9.0g                  Alcohol 0-39 vs 40+
-------------------------------------------------------------------------------
Sorted by:  case

. cc case alc80 [fw=n]

                   Alcohol 0-79 vs 80+
                                                        Proportion
                 |   Exposed   Unexposed  |     Total     Exposed
-----------------+------------------------+----------------------
           Cases |        96         104  |       200      0.4800
        Controls |       109         666  |       775      0.1406
-----------------+------------------------+----------------------
           Total |       205         770  |       975      0.2103
                 |                        |
                 |         Pt. Est.       |  [95% Conf. Interval]
                 |------------------------+----------------------
      Odds ratio |         5.640085       |  4.003217     7.94673  (Cornfield)
 Attr. frac. ex. |         .8226977       |  .7502009    .8741621  (Cornfield)
 Attr. frac. pop |         .3948949       |
                 +-----------------------------------------------
                             chi2(1) =   110.26  Pr>chi2 = 0.0000
```

Thus using the `cc` command we estimate that 39% of the oesophageal cancers were attributable to alcohol consumptions above 80 g/day. Note there is no confidence interval around this estimate. We can obtain the same result using logistic regression and `aflogit` but with the added benefit of a confidence interval.

```
. logistic case alc80 [fw=n]

Logit Estimates                                        Number of obs =    975
                                                       chi2(1)       =  96.43
                                                       Prob > chi2   = 0.0000
Log Likelihood = -446.52782                            Pseudo R2     = 0.0975

------------------------------------------------------------------------------
    case | Odds Ratio   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
    alc80|   5.640085   .9883491      9.872   0.000        4.00059    7.951467
------------------------------------------------------------------------------

. aflogit, cc

Population attributable fraction from logistic regression
Case-control data (n=975)

Using weights [fweight= n]

Term      Ref.     A.F.        s.e.      [95% Conf. Int.]*
----------------------------------------------------------
alc80      0      0.3949      0.0409      0.3093   0.4699
----------------------------------------------------------
TOTAL             0.3949      0.0409      0.3093   0.4699

* CI calculated on log(1-AF) scale
```

The logistic regression framework also allows us to extend the model to adjust for potential confounders (age-group and smoking in this example) and even to allow for interactions between the exposure and confounders; for instance, excess alcohol consumption may have a worse effect depending on age-group. First, allowing for potential confounders we have

```
. xi: logistic case alc80 i.agegp*i.smoke [fw=n]
i.agegp              Iagegp_0-2   (naturally coded; Iagegp_0 omitted)
i.smoke              Ismoke_0-2   (naturally coded; Ismoke_0 omitted)
i.agegp*i.smoke      IaXs_#-#     (coded as above)

Logit Estimates                                        Number of obs =    975
                                                       chi2(9)       = 232.84
                                                       Prob > chi2   = 0.0000
Log Likelihood = -378.32187                            Pseudo R2     = 0.2353

------------------------------------------------------------------------------
    case | Odds Ratio   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
   alc80 |    4.88238   .9426293      8.213   0.000          3.3442   7.128055
Iagegp_1 |   11.81257    9.13967      3.191   0.001        2.592679    53.8195
Iagegp_2 |   27.87187   20.35787      4.556   0.000        6.659488   116.6518
Ismoke_1 |   6.390786   5.161163      2.297   0.022        1.312589   31.11572
Ismoke_2 |   2.33e-07   1.02e-07    -34.902   0.000        9.88e-08   5.49e-07
IaXs_1_1 |   .2689882   .2421917     -1.458   0.145        .0460605   1.570863
IaXs_1_2 |   3.28e+07   2.38e+07     23.816   0.000         7889856   1.36e+08
IaXs_2_1 |    .221224   .1856753     -1.797   0.072        .0426983   1.146182
IaXs_2_2 |   2.20e+07          .          .       .               .          .
------------------------------------------------------------------------------

Note: 33 failures and 0 successes completely determined.

. aflogit alc80, cc

Population attributable fraction from logistic regression
Case-control data (n=975)

Using weights [fweight= n]

Term      Ref.     A.F.        s.e.      [95% Conf. Int.]*
----------------------------------------------------------
alc80      0      0.3817      0.0442      0.2887   0.4625
----------------------------------------------------------
TOTAL             0.3817      0.0442      0.2887   0.4625

* CI calculated on log(1-AF) scale
```

The odds ratio on alcohol changed little (from 5.6 to 4.9) when allowing for smoking and age and so the adjusted attributable fraction is very similar (38%). Now allowing alcohol consumption to interact with age-group, we have

```
. xi: logistic case i.agegp*alc80 i.agegp*i.smoke [fw=n]
i.agegp              Iagegp_0-2   (naturally coded; Iagegp_0 omitted)
i.agegp*alc80        IaXal1_#     (coded as above)
i.smoke              Ismoke_0-2   (naturally coded; Ismoke_0 omitted)
i.agegp*i.smoke      IaXs_#-#     (coded as above)

Note: Iagegp_1 dropped due to collinearity.
Note: Iagegp_2 dropped due to collinearity.
```

```
Logit Estimates                                       Number of obs =    975
                                                      chi2(11)      = 233.13
                                                      Prob > chi2   = 0.0000
Log Likelihood =  -378.1796                           Pseudo R2     = 0.2356

------------------------------------------------------------------------------
    case | Odds Ratio   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
Iagegp_1 |   12.75294   10.61766      3.058   0.002       2.494186    65.20659
Iagegp_2 |   31.46491   24.56128      4.418   0.000       6.813728    145.3008
   alc80 |   6.527618   4.431543      2.763   0.006       1.725367    24.69608
IaXalc_1 |   .8002673   .6210458     -0.287   0.774       .1748485    3.662759
IaXalc_2 |   .7024653   .5052614     -0.491   0.623       .1715466    2.876521
Ismoke_1 |   6.226884   5.062682      2.249   0.024       1.265357     30.6428
Ismoke_2 |   2.14e-07   1.25e-07    -26.312   0.000       6.83e-08    6.72e-07
IaXs_1_1 |    .275004   .2492789     -1.424   0.154       .0465332    1.625232
IaXs_1_2 |   3.58e+07          .          .       .              .           .
IaXs_2_1 |     .22786   .1924345     -1.751   0.080       .0435315    1.192705
IaXs_2_2 |   2.38e+07   1.74e+07     23.318   0.000        5716679    9.94e+07
------------------------------------------------------------------------------
Note: 33 failures and 0 successes completely determined.

. aflogit alc80 IaXalc*, cc

Population attributable fraction from logistic regression
Case-control data (n=975)

Using weights [fweight= n]

Term      Ref.     A.F.        s.e.      [95% Conf. Int.]*
----------------------------------------------------------
alc80      0      0.4065      0.0630      0.2693   0.5179
IaXalc_1   0     -0.0312           .           .        .
IaXalc_2   0     -0.1398           .           .        .
----------------------------------------------------------
TOTAL             0.3803      0.0445      0.2866   0.4616

* CI calculated on log(1-AF) scale
```

There appears to be little evidence of an interaction between age and alcohol consumption and so the attributable fraction is unchanged. The negative values of the attributable fraction indicate that removing an exposure would make the expected number of cases larger than that observed. Hence, removing the slight protection afforded by being in the oldest age-group would result in a 14% increase in the observed number of cases of oesophageal cancer in the population. Note that algebraically the population attributable fraction has bounds $[-\infty, 1]$.

The estimates of attributable fraction produced by `aflogit` are identical to those calculated by Benichou and the standard errors are very similar (within 3%).

## Saved results

`aflogit` saves the following results in the `S_` macros:

| | |
|---|---|
| S_1 | The summary attributable fraction estimate |
| S_2 | SE of the summary attributable fraction estimate |
| S_3 | Lower confidence limit of the summary attributable fraction |
| S_4 | Upper confidence limit of the summary attributable fraction |

## Acknowledgments

I am grateful to Tony Swan for pointing the way to estimating attributable fractions from logistic regression and to Heather Dickinson for useful discussions which led to the inclusion of the Poisson regression case.

## References

Benichou, J. 1991. Methods of adjustment for estimating the attributable risk in case–control studies: A review. *Statistics in Medicine* 10: 1753–1773.

Bruzzi, P., S. B. Green, D. P. Byar, L. A. Brinton, and C. Schairer. 1985. Estimating the population attributable risk for multiple risk factors using case–control data. *American Journal of Epidemiology* 122: 904–14.

Last, J. M. ed. 1983. *A Dictionary of Epidemiology*. New York: Oxford University Press.

Greenland, S. and K. Drescher. 1993. Maximum likelihood estimation of the attributable fraction from logistic models. *Biometrics* 49: 865–872.

Tuyns, A. J., G. Pequignot, and O. M. Jensen. 1977. Le cancer de l'oesophage en Ille-et Vilaine en fonction des niveaux de consommation d'alcool et de tabac. *Bulletin of Cancer* 64: 45–60.

## sbe22 Cumulative meta-analysis

Jonathan Sterne, United Medical and Dental Schools, UK, j.sterne@umds.ac.uk

Meta-analysis is used to combine the results of several studies, and the Stata command `meta` (Sharp and Sterne 1997 and *sbe16.1* in this issue) can be used to perform meta-analyses and graph the results. In cumulative meta-analysis (Lau et al. 1992), the pooled estimate of the treatment effect is updated each time the results of a new study are published. This makes it possible to track the accumulation of evidence on the effect of a particular treatment.

The command `metacum` performs cumulative meta-analysis (using fixed- or random-effects models) and, optionally, graphs the results.

### Syntax

```
metacum {theta | exp(theta)} {se_theta | var_theta | ll ul [cl]} [if exp] [in range]
        [, var ci effect(f|r) eform level(#) id(strvar) graph
        cline fmult(#) csize(#) ltrunc(#) rtrunc(#) graph_options]
```

In common with the commands `meta` (Sharp and Sterne 1998) and `metabias` (Steichen 1988), the user provides the effect estimate as *theta* (i.e., a log risk ratio, log odds ratio, or other measure of effect). Likewise, the user supplies a measure of *theta*'s variability (i.e., its standard error, *se_theta*, or its variance, *var_theta*). Alternatively, the user provides *exp(theta)* (e.g. a risk ratio or odds ratio) and its confidence interval, (*ll*, *ul*).

### Required input variables

These are the same as for the new version of `meta` described in *sbe16.1* in this issue.

### Options for displaying results

`var` means the user has specified a variable containing the variance of the effect estimate. If this option is not included, the command assumes the standard error has been specified.

`ci` means the user has specified the lower and upper confidence limits of the effect estimate, which is assumed to be on the ratio scale (e.g. odds ratio or risk ratio).

`effect(`*f*|*r*`)` must be included. This specifies whether fixed (*f*) or random (*r*) effects estimates are to be used in the output and graph.

`eform` requests that the output be exponentiated. This is useful for effect measures such as log odds ratios which are derived from generalized linear models. If the `eform` and `graph` options are used, then the graph output is exponentiated, with a log scale for the $x$-axis.

`level(`*#*`)` specifies the confidence level, in percent, for confidence intervals. The default is `level(95)` or as set by `set level`.

`id(`*strvar*`)` is a character variable which is used to label the studies. If the data contains a labeled numeric variable, then the `decode` command can be used to create a character variable.

### Options for graphing results

`graph` requests a graph.

`cline` asks that a vertical dotted line be drawn through the combined estimate.

`fmult(`*#*`)` is a number greater than zero which can be used to scale the font size for the study labels. The font size is automatically reduced if the maximum label length is greater than 8, or the number of studies is greater than 20. However it may be possible to increase it somewhat over the default size.

`csize(`*#*`)` gives the size of the circles used in the graph (default 180).

`ltrunc(`*#*`)` truncates the left side of the graph at the number #. This is used to truncate very wide confidence intervals. However # must be less than each of the individual study estimates.

`rtrunc(`*#*`)` truncates the right side of the graph at #, and must be greater than each of the individual study estimates.

*graph_options* are any options allowed with `graph, twoway` other than `ylabel()`, `symbol()`, `xlog`, `ytick`, and `gap`.

### Background

The command `metacum` provides an alternative means of presenting the results of a meta-analysis, where instead of the individual study effects and combined estimate, the cumulative evidence up to and including each trial can be printed and/or graphed. The technique was suggested by Lau et al. (1992).

### Example

The first trial of streptokinase treatment following myocardial infarction was reported in 1959. A further 21 trials were conducted between that time and 1986, when the ISIS-2 multicenter trial (on over 17,000 patients in whom over 1800 deaths were reported) demonstrated conclusively that the treatment reduced the chances of subsequent death.

Lau et al. (1992) pointed out that a meta-analysis of trials performed up to 1977 provided strong evidence that the treatment worked. Despite this, it was another 15 years until the treatment became routinely used.

Dataset `strepto.dta` contains the results of 22 trials of streptokinase conducted between 1959 and 1986.

```
. use strepto, clear
(Streptokinase after MI)

. describe

Contains data from strepto.dta
  obs:            22                            Streptokinase after MI
 vars:             7
 size:           638
-------------------------------------------------------------------------------
   1. trial       byte   %8.0g                  Trial number
   2. trialnam    str14  %14s                   Trial name
   3. year        int    %8.0g                  Year of publication
   4. pop1        int    %12.0g                 Treated population
   5. deaths1     int    %12.0g                 Treated deaths
   6. pop0        int    %12.0g                 Control population
   7. deaths0     int    %12.0g                 Control deaths
-------------------------------------------------------------------------------
Sorted by:  trial

. list trialnam year pop1 deaths1 pop0 deaths0, noobs

      trialnam     year       pop1    deaths1       pop0    deaths0
      Fletcher     1959         12          1         11          4
         Dewar     1963         21          4         21          7
  1st European     1969         83         20         84         15
   Heikinheimo     1971        219         22        207         17
       Italian     1971        164         19        157         18
  2nd European     1971        373         69        357         94
 2nd Frankfurt     1973        102         13        104         29
1st Australian     1973        264         26        253         32
    NHLBI SMIT     1974         53          7         54          3
        Valere     1975         49         11         42          9
         Frank     1975         55          6         53          6
     UK Collab     1976        302         48        293         52
         Klein     1976         14          4          9          1
      Austrian     1977        352         37        376         65
      Lasierra     1977         13          1         11          3
      N German     1977        249         63        234         51
      Witchitz     1977         32          5         26          5
2nd Australian     1977        112         25        118         31
  3rd European     1977        156         25        159         50
          ISAM     1986        859         54        882         63
       GISSI-1     1986       5860        628       5852        758
        ISIS-2     1988       8592        791       8595       1029
```

Before doing our meta-analysis, we calculate the log odds ratio for each study, and its corresponding variance. We also create a string variable containing the trial name and year of publication:

```
. gen logor=log((deaths1/(pop1-deaths1))/((deaths0/(pop0-deaths0))))

. gen varlogor=(1/deaths1)+(1/(pop1-deaths1))+(1/deaths0)+(1/(pop0-deaths0))

. gen str4 yc=string(year)

. gen str21 trnamy=trialnam+" ("+yc+")"

. meta logor varlogor, var eform graph(f) id(trnamy) xlab(.1,.5,1,2,10) ltr(0.1)
> rtr(10) cline xline(1) print b2("Odds ratio") fmult(1.5)
```

```
Meta-analysis (exponential form)
       |  Pooled      95% CI         Asymptotic      No. of
Method |     Est   Lower   Upper  z_value  p_value   studies
-------+----------------------------------------------------
Fixed  |   0.774   0.725   0.826   -7.711    0.000     22
Random |   0.782   0.693   0.884   -3.942    0.000

Test for heterogeneity: Q= 31.498 on 21 degrees of freedom (p= 0.066)
Moment-based estimate of between studies variance =  0.017
                      |     Weights      Study       95% CI
                Study |  Fixed  Random     Est   Lower   Upper
----------------------+----------------------------------------
      Fletcher (1959) |   0.67    0.67    0.16    0.01    1.73
         Dewar (1963) |   1.91    1.85    0.47    0.11    1.94
   1st European (1969) |   6.80    6.10    1.46    0.69    3.10
    Heikinheimo (1971) |   8.72    7.61    1.25    0.64    2.42
        Italian (1971) |   8.18    7.19    1.01    0.51    2.01
   2nd European (1971) |  31.03   20.39    0.64    0.45    0.90
  2nd Frankfurt (1973) |   7.35    6.54    0.38    0.18    0.78
1st Australian (1973) |  12.75   10.50    0.75    0.44    1.31
     NHLBI SMIT (1974) |   1.93    1.87    2.59    0.63   10.60
        Valere (1975) |   3.87    3.63    1.06    0.39    2.88
         Frank (1975) |   2.67    2.55    0.96    0.29    3.19
     UK Collab (1976) |  20.77   15.39    0.88    0.57    1.35
         Klein (1976) |   0.68    0.67    3.20    0.30   34.59
      Austrian (1977) |  20.49   15.24    0.56    0.36    0.87
      Lasierra (1977) |   0.65    0.64    0.22    0.02    2.53
      N German (1977) |  21.59   15.84    1.22    0.80    1.85
      Witchitz (1977) |   2.06    1.99    0.78    0.20    3.04
2nd Australian (1977) |  10.50    8.92    0.81    0.44    1.48
  3rd European (1977) |  13.02   10.68    0.42    0.24    0.72
          ISAM (1986) |  27.13   18.63    0.87    0.60    1.27
       GISSI-1 (1986) | 303.12   49.69    0.81    0.72    0.90
        ISIS-2 (1988) | 400.58   51.76    0.75    0.68    0.82
```

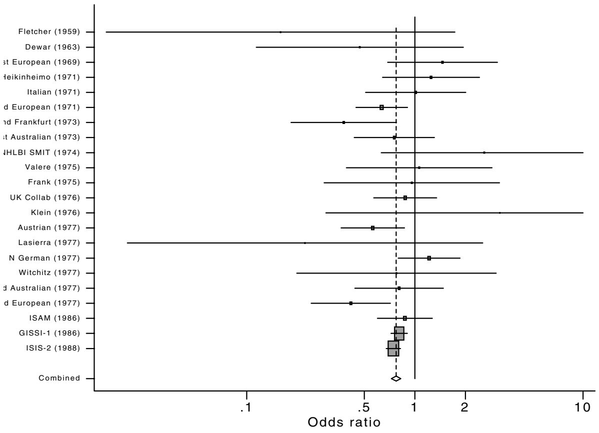

Figure 1: Streptokinase meta-analysis

It can be seen from the fixed-effects weights, and the graphical display, that the results are dominated by the two large trials reported in 1986. We now do a cumulative meta-analysis:

```
. metacum logor varlogor, var effect(f) graph eform id(trnamy) xlab(.1,.5,1,2)
> ltr(0.1) cline xline(1) b2("Odds ratio") fmult(1.5)
```

```
Cumulative fixed-effects meta-analysis of 22 studies (exponential form)
-----------------------------------------------------------------------

                      Cumulative      95% CI
Trial                   estimate    Lower  Upper         z  P value
Fletcher (1959)            0.159    0.015  1.732    -1.509    0.131
Dewar (1963)               0.355    0.105  1.200    -1.667    0.096
1st European (1969)        0.989    0.522  1.875    -0.034    0.973
Heikinheimo (1971)         1.106    0.698  1.753     0.430    0.667
Italian (1971)             1.076    0.734  1.577     0.376    0.707
2nd European (1971)        0.809    0.624  1.048    -1.607    0.108
2nd Frankfurt (1973)       0.742    0.581  0.946    -2.403    0.016
1st Australian (1973)      0.744    0.595  0.929    -2.604    0.009
NHLBI SMIT (1974)          0.767    0.615  0.955    -2.366    0.018
```

```
Valere (1975)              0.778    0.628  0.965    -2.285    0.022
Frank (1975)               0.783    0.634  0.968    -2.262    0.024
UK Collab (1976)           0.801    0.662  0.968    -2.296    0.022
Klein (1976)               0.808    0.668  0.976    -2.213    0.027
Austrian (1977)            0.762    0.641  0.906    -3.072    0.002
Lasierra (1977)            0.757    0.637  0.900    -3.150    0.002
N German (1977)            0.811    0.691  0.951    -2.571    0.010
Witchitz (1977)            0.810    0.691  0.950    -2.596    0.009
2nd Australian (1977)      0.810    0.695  0.945    -2.688    0.007
3rd European (1977)        0.771    0.665  0.894    -3.448    0.001
ISAM (1986)                0.784    0.683  0.899    -3.470    0.001
GISSI-1 (1986)             0.797    0.731  0.870    -5.092    0.000
ISIS-2 (1988)              0.774    0.725  0.826    -7.711    0.000
```

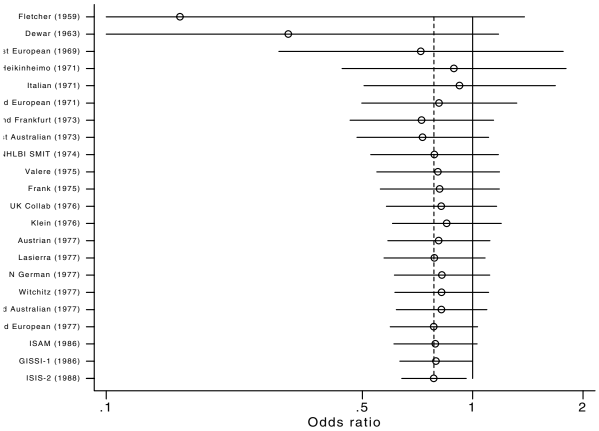

Figure 2: Streptokinase cumulative meta-analysis

By the end of 1977 there was clear evidence that streptokinase treatment prevented death following myocardial infarction. The point estimate of the pooled treatment effect was virtually identical in 1977 (odds ratio=0.771) and after the results of the large trials in 1986 (odds ratio=0.774).

## Note

The command `meta` (Sharp and Sterne 1998) should be installed before running `metacum`.

## Acknowledgment

I thank Stephen Sharp for reviewing the command, Matthias Egger for providing the streptokinase data, and Thomas Steichen for providing the alternative forms of command syntax.

## References

Lau J., E. M. Antman, J. Jimenez-Silva, et al. 1992. Cumulative meta-analysis of therapeutic trials for myocardial infarction. *New England Journal of Medicine* 327: 248–54.

Sharp S. and J. Sterne. 1997. sbe16: Meta-analysis. *Stata Technical Bulletin* 38: 9–14.

——. 1998. sbe16.1: New syntax and output for the meta-analysis command. *Stata Technical Bulletin* 42: 6–8.

Steichen, T. 1998. sbe19: Tests for publication bias in meta-analysis. *Stata Technical Bulletin* 41: 9–15.

## sbe23 Meta-analysis regression

Stephen Sharp, London School of Hygiene and Tropical Medicine, stephen.sharp@lshtm.ac.uk

The command `metareg` extends a random effects meta-analysis to estimate the extent to which one or more covariates, with values defined for each study in the analysis, explain heterogeneity in the treatment effects. Such analysis is sometimes termed “meta-regression” (Lau et al. 1998). Examples of such study-level covariates might be average duration of follow-up, some measure of study quality, or, as described in this article, a measure of the geographical location of each study. `metareg` fits models with two additive components of variance, one representing the variance within units, the other the variance between units, and therefore is applicable both to the meta-analysis situation, where each unit is one study, and to other situations such as multi-center trials, where each unit is one center. Here `metareg` is explained in the meta-analysis context.

## Background

Suppose $y_i$ represents the treatment effect measured in study $i$ ($k$ independent studies, $i = 1, \dots, k$), such as a log odds ratio or a difference in means, $v_i$ is the (within-study) variance of $y_i$, and $x_{i1}, \dots, x_{ip}$ are measured study-level covariates. A weighted normal errors regression model is

$$Y \sim N(X\beta, V)$$

where $Y = (y_1, \dots, y_k)^T$ is the $k \times 1$ vector of treatment effects, with $i$th element $y_i$, $X$ is a $k \times (p+1)$ design matrix with $i$th row $(1, x_{i1}, \dots, x_{ip})$, $\beta = (\beta_0, \dots, \beta_p)^T$ is a $(p+1) \times 1$ vector of parameters, and $V$ is a $k \times k$ diagonal variance matrix, with $i$th diagonal element $v_i$.

The parameters of this model can be estimated in Stata using `regress` with analytic weights $w_i = 1/v_i$. However, $v_i$ represents the variance of the treatment effect within study $i$, so this model does not take into account any possible residual heterogeneity in the treatment effects *between* studies. One approach to incorporating residual heterogeneity is to include an additive between-study variance component $\tau^2$, so the $i$th diagonal element of the variance matrix $V$ becomes $v_i + \tau^2$.

The parameters of the model can then be estimated using a weighted regression with weights equal to $1/v_i + \tau^2$, but $\tau^2$ must be explicitly estimated in order to carry out the regression. `metareg` allows four alternative methods for estimation of $\tau^2$, three of them are iterative, while one is noniterative and an extension of the moment estimator proposed for random effects meta-analysis without covariates (DerSimonian and Laird 1986).

## Method-of-moments estimator

Maximum-likelihood estimates of the $\beta$ parameters are first obtained by weighted regression assuming $\widehat{\tau}^2 = 0$, and then a moment estimator of $\tau^2$ is calculated using the residual sum of squares from the model,

$$RSS = \sum_{i=1}^{k} w_i (y_i - \widehat{y}_i)^2$$

as follows:

$$\widehat{\tau}^2_{mm} = \frac{RSS - (k - (p+1))}{\sum_{i=1}^k w_i - \text{tr}(V^{-1}X(X'V^{-1}X)^{-1}X'V^{-1})}$$

where $\widehat{\tau}^2_{mm} = 0$ if $RSS < k - (p+1)$ (DuMouchel and Harris 1983).

A weighted regression is then carried out with new weights $w_i^* = 1/\widehat{\tau}^2 + v_i$ to provide a new estimate of $\beta$. The formula for $\widehat{\tau}^2_{mm}$ in the case of no covariate reduces to the standard moment estimator (DerSimonian and Laird 1986).

## Iterative procedures

Three other methods for estimating $\tau^2$ have been proposed, and require an iterative procedure.

Starting with $\widehat{\tau}^2 = 0$, a regression using weights $w_i^* = 1/v_i$ gives initial estimates of $\beta$. The fitted values $\widehat{y}_i$ from this model can then be used in one of three formulas for estimation of $\tau^2$, given below:

$$\widehat{\tau}^2_{ml} = \frac{\sum_{i=1}^k w_i^{*2}[(y_i - \widehat{y}_i)^2 - v_i]}{\sum_{i=1}^k w_i^{*2}} \text{ maximum likelihood (Pocock et al. 1981)}$$

$$\widehat{\tau}^2_{reml} = \frac{\sum_{i=1}^k w_i^{*2}\left[\frac{k}{k-(p+1)}(y_i - \widehat{y}_i)^2 - v_i\right]}{\sum_{i=1}^k w_i^{*2}} \text{ restricted maximum likelihood (Berkey et al. 1995)}$$

$$\widehat{\tau}^2_{eb} = \frac{\sum_{i=1}^k w_i^{*}\left[\frac{k}{k-(p+1)}(y_i - \widehat{y}_i)^2 - v_i\right]}{\sum_{i=1}^k w_i^{*}} \text{ empirical Bayes (Berkey et al. 1995)}$$

In each case, if the estimated value $\widehat{\tau}^2$ is negative, it is set to zero.

Using the estimate $\widehat{\tau}^2$, new weights $w_i^* = 1/\widehat{\tau}^2 + v_i$ (or $1/v_i$ if $\widehat{\tau}$ is zero) are then calculated, and hence new estimates of $\beta$, fitted values $\widehat{y}_i$, and thence $\widehat{\tau}^2$. The procedure continues until the difference between successive estimates of $\tau^2$ is less than a prespecified number (such as $10^{-4}$). The standard errors of the final estimates of $\beta$ are calculated forcing the scale parameter to be 1, since the weights are equal to the reciprocal variances.

## Syntax

`metareg` has the usual syntax for a regression command, with the additional requirement that the user specify a variable containing either the within-study standard error or variance.

`metareg` *y varlist* [`if` *exp*] [`in` *range*] `,` {`wsse(`*varname*`)` | `wsvar(`*varname*`)` | `wsse(`*varname*`)` `wsvar(`*varname*`)`}
[ `bsest({reml | ml | eb | mm})` `toleran(`*#*`)` `level(`*#*`)` `noiter`]

The command supplies estimated parameters, standard errors, $Z$ statistics, $p$ values and confidence intervals, in the usual regression output format. The estimated value of $\tau^2$ is also given.

## Options

`wsse(`*varname*`)` is a variable in the dataset which contains the within-studies standard error $\sqrt{v_i}$. Either this or the `wsvar` option below (or both) must be specified.

`wsvar(`*varname*`)` is a variable in the dataset which contains the within-studies variance $v_i$. Either this or the `wsse` option above (or both) must be specified.

Note: if both the above options are specified, the program will check that the variance is the square of the standard error for each study.

`bsest({reml | ml | eb | mm})` specifies the method for estimating $\tau^2$. The default is `reml` (restricted maximum likelihood), with the alternatives being `ml` (maximum likelihood), `eb` (empirical Bayes), and `mm` (method of moments).

`toleran(`*#*`)` specifies the difference between values of $\hat{\tau}^2$ at successive iterations required for convergence. If *#* is $n$, the process will not converge until successive values of $\hat{\tau}^2$ differ by less than $10^{-n}$. The default is 4.

`level(`*#*`)` specifies the confidence level, in percent, for confidence intervals. The default is `level(95)` or as set by `set level`.

`noiter` requests that the log of the iterations in the `reml`, `ml`, or `eb` procedures be suppressed from the output.

## Example

BCG is a vaccine widely used to give protection against tuberculosis. Colditz et al. (1994) performed a meta-analysis of all published trials which randomized subjects to either BCG vaccine or placebo, and then had similar surveillance procedures to monitor the outcome, diagnosis of tuberculosis.

The data in `bcg.dta` are as reported by Berkey et al. (1995). Having read the file into Stata, the log odds ratio of tuberculosis comparing BCG with placebo, and its standard error can be calculated for each study.

```
. use bcg, clear
(BCG and tuberculosis)

. describe

Contains data from bcg.dta
  obs:           13                           BCG and tuberculosis
 vars:            8
 size:          351
-------------------------------------------------------------------------------
   1. trial       str2   %9s                  trial identity number
   2. lat         byte   %9.0g                absolute latitude from Equator
   3. nt          float  %9.0g                total vaccinated patients
   4. nc          float  %9.0g                total unvaccinated patients
   5. rt          int    %9.0g                tuberculosis in vaccinated
   6. rc          int    %9.0g                tuberculosis in unvaccinated
-------------------------------------------------------------------------------
Sorted by:

. list, noobs

    trial       lat        nt        nc        rt        rc
        1        44       123       139         4        11
        2        55       306       303         6        29
        3        42       231       220         3        11
        4        52     13598     12867        62       248
        5        13      5069      5808        33        47
```

```
      6        44       1541       1451        180        372
      7        19       2545        629          8         10
      8        13      88391      88391        505        499
      9        27       7499       7277         29         45
     10        42       1716       1665         17         65
     11        18      50634      27338        186        141
     12        33       2498       2341          5          3
     13        33      16913      17854         27         29

. gen logor=log((rt/(nt-rt))/(rc/(nc-rc)))

. gen selogor=sqrt((1/rc)+(1/(nc-rc))+(1/rt)+(1/(nt-rt)))
```

Note: if either `rt` or `rc` were 0, a standard approach would be to add 0.5 to each of `rt`, `rc`, `nt-rt`, and `nc-rc` for that study (Cox and Snell 1989).

A meta-analysis of the data can now be performed using the `meta` command described by Sharpe and Sterne (1997 and updated in *sbe16.1*).

```
. meta logor selogor, eform graph(r) id(trial) cline xlab(0.5,1,1.5) xline(1)
> boxsh(4) b2("Odds ratio - log scale")

Meta-analysis (exponential form)

       |  Pooled      95% CI         Asymptotic      No. of
Method |     Est   Lower   Upper  z_value  p_value   studies
-------+----------------------------------------------------
Fixed  |   0.647   0.595   0.702  -10.319    0.000     13
Random |   0.474   0.325   0.690   -3.887    0.000

Test for heterogeneity: Q= 163.165 on 12 degrees of freedom (p= 0.000)
Moment estimate of between-studies variance =  0.366
```

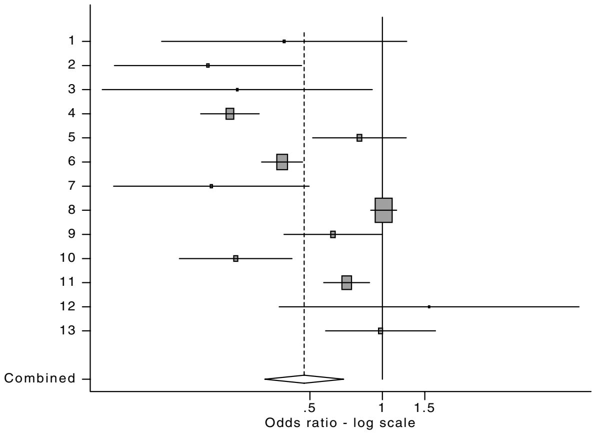

Figure 1: A meta-analysis of the BCG and Tuberculosis data

Both the graph and the statistical test indicate substantial heterogeneity between the trials, with an estimated between-studies variance of 0.366. The random effects combined estimate of 0.474, indicating a strong protective effect of BCG against tuberculosis, should not be reported without some discussion of the possible reasons for the differences between the studies (Thompson 1994).

One possible explanation for the differences in treatment effects could be that the studies were conducted at different latitudes from the equator. Berkey et al. (1995) speculated that absolute latitude, or distance of each study from the equator, may serve as a surrogate for the presence of environmental mycobacteria which provide a certain level of natural immunity against tuberculosis. By sorting on absolute latitude, the graph obtained using `meta` shows the studies in order of increasing latitude going down the page.

```
. sort lat

. meta logor selogor, eform graph(r) id(trial) cline xlab(0.5,1,1.5) xline(1)
> boxsh(4) b2("Odds ratio - log scale")
```

*(output omitted)*

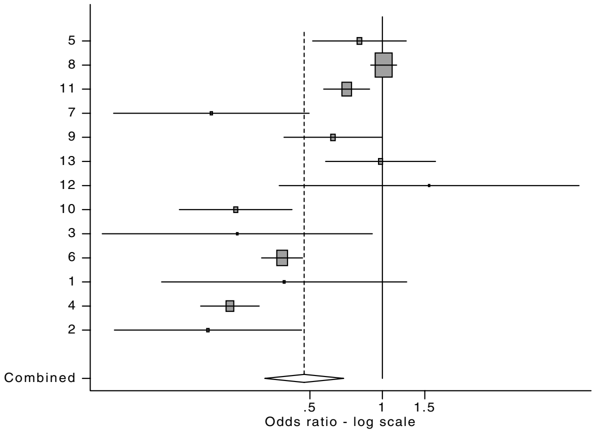

Figure 2: Same as Figure 1 but sorted by latitude

The graph now suggests that BCG vaccination is more effective at higher absolute latitudes. This can be investigated further using the `metareg` command, with a REML estimate of the between-studies variance $\tau^2$.

```
. metareg logor lat, wsse(selogor) bs(reml) noiter

Meta-analysis regression                              No of studies =   13
                                                      tau^2 method      reml
                                                      tau^2 estimate =  .0235

Successive values of tau^2 differ by less than 10^-4 - convergence achieved
------------------------------------------------------------------------------
         |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
     lat |  -.0320363   .0049432     -6.481   0.000      -.0417247   -.0223479
   _cons |   .3282194   .1659807      1.977   0.048       .0029033    .6535356
------------------------------------------------------------------------------
```

This analysis shows that after allowing for additive residual heterogeneity, there is a significant negative association between the log odds ratio and absolute latitude, i.e., the higher the absolute latitude, the lower the odds ratio, and hence the greater the benefit of BCG vaccination. The following plot of log odds ratio against absolute latitude includes the fitted regression line from the model above. The size of the circles in the plot is inversely proportional to the variance of the log odds ratio, so larger circles correspond to larger studies.

```
. gen invvlor=selogor^-2

. gen fit=0.328-0.032*lat

. gr logor fit lat [fw=invvlor], s(oi) c(.l)  xlab(0,10,20,30,40,50,60)
> ylab(-1.6094,-0.6931,0,0.6931) l1("Odds ratio (log scale)")
> b2("Distance from Equator (degrees of latitude)")
```

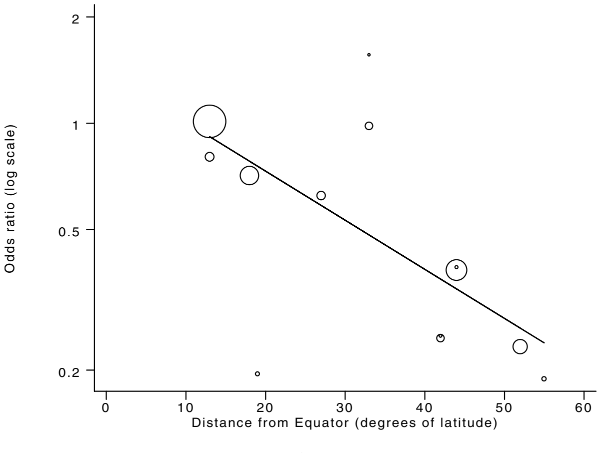

Figure 3

(Note: the axes on this graph have been modified using the STAGE software)

Here a restricted maximum-likelihood method was used to estimate $\tau^2$; the other three methods are used in turn below:

```
. metareg logor lat, wsse(selogor) bs(ml) noiter
```

```
Meta-analysis regression                              No of studies =   13
                                                      tau^2 method      ml
                                                      tau^2 estimate =  .0037
Successive values of tau^2 differ by less than 10^-4 - convergence achieved
------------------------------------------------------------------------------
         |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
     lat |  -.0327447   .0033327     -9.825   0.000      -.0392767   -.0262128
   _cons |   .3725098   .1043895      3.568   0.000       .1679102    .5771093
------------------------------------------------------------------------------
```

```
. metareg logor lat, wsse(selogor) bs(eb) noiter
Meta-analysis regression                              No of studies =   13
                                                      tau^2 method      eb
                                                      tau^2 estimate =  .1373
Successive values of tau^2 differ by less than 10^-4 - convergence achieved
------------------------------------------------------------------------------
         |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
     lat |  -.0305794   .0090005     -3.398   0.001      -.0482201   -.0129388
   _cons |   .2548214   .3138935      0.812   0.417      -.3603984    .8700413
------------------------------------------------------------------------------
```

```
. metareg logor lat, bs(mm) wsse(selogor) noiter
Warning: mm is a non-iterative method, noiter option ignored
Meta-analysis regression                              No of studies =   13
                                                      tau^2 method      mm
                                                      tau^2 estimate =  .0480

------------------------------------------------------------------------------
         |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
     lat |  -.0315724   .0061726     -5.115   0.000      -.0436704   -.0194744
   _cons |    .303035   .2108751      1.437   0.151      -.1102727    .7163427
------------------------------------------------------------------------------
```

The estimated value of $\tau^2$ using a method-of-moments estimator is 0.048, compared with 0.366 before adjusting for latitude, so absolute latitude has explained almost all of the variation between the studies.

The analyses above show that the estimate of the effect of latitude is similar using all four methods. However, the estimated values of $\tau^2$ differ considerably, with the estimate from the empirical Bayes method being largest. The restricted maximum-likelihood method corrects the bias in the maximum-likelihood estimate of $\tau^2$. The basis for using the empirical Bayes method is less clear (Morris 1983), so this method should be used with caution. The moment-based method extends the usual random-effects meta-analysis; below `metareg` is used to fit a model with no covariate:

```
. metareg logor, bs(mm) wsse(selogor)
Meta-analysis regression                              No of studies =   13
                                                      tau^2 method      mm
                                                      tau^2 estimate =  .3663

------------------------------------------------------------------------------
         |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
   _cons |  -.7473923   .1922628     -3.887   0.000      -1.124221   -.3705641
------------------------------------------------------------------------------
```

Now the estimate of $\tau^2$ is identical to that obtained earlier from `meta`, and the constant parameter is the log of the random effects pooled estimate given by `meta`.

The paper by Thompson and Sharp (1998) contains a fuller discussion both of the differences between the four methods of estimation, and other methods for explaining heterogeneity. Copies are available on request from the author.

## Saved results

`metareg` saves the following results in the `S_` macros:

| | |
|---|---|
| S_1 | $k$, number of studies |
| S_2 | $\hat{\tau}^2$, estimate of between-studies variance |

## Acknowledgment

I am grateful to Simon Thompson, Ian White, and Jonathan Sterne for their helpful comments on earlier versions of this command.

## References

Berkey, C. S., D. C. Hoaglin, F. Mosteller, and G. A. Colditz. 1995. A random-effects regression model for meta-analysis. *Statistics in Medicine* 14: 395–411.

Colditz, G. A., T. F. Brewer, C. S. Berkey, M. E. Wilson, E. Burdick, H. V. Fineberg, et al. 1994. Efficacy of BCG vaccine in the prevention of tuberculosis. Meta-analysis of the published literature. *Journal of the American Medical Association* 271: 698–702.

Cox, D. R. and E. J. Snell. 1989. *Analysis of Binary Data*. 2d ed. London: Chapman and Hall.

DerSimonian, R. and N. M. Laird. 1986. Meta-analysis in clinical trials. *Controlled Clinical Trials* 7: 177–188.

DuMouchel, W. and J. Harris. 1983. Bayes methods for combining the results of cancer studies in humans and other species. *Journal of the American Statistical Association* 78: 291–308.

Lau, J., J. P. A. Ioannidis, and C. H. Schmid. 1998. Summing up evidence: one answer is not always enough. *Lancet* 351: 123–127.

Morris, C. N. 1983. Parametric empirical Bayes inference: theory and applications. *Journal of the American Statistical Association* 78: 47–55.

Pocock, S. J., D. G. Cook, and S. A. A. Beresford. 1981. Regression of area mortality rates on explanatory variables: what weighting is appropriate? *Applied Statistics* 30: 286–295.

Sharp, S. and J. Sterne. 1997. sbe16: Meta-analysis. *Stata Technical Bulletin* 38: 9–14.

——. 1998. sbe16.1: New syntax and output for the meta-analysis command. *Stata Technical Bulletin* 42: 6–8.

Thompson, S. G. 1994. Why sources of heterogeneity in meta-analysis should be investigated. *British Medical Journal* 309: 1351–1355.

Thompson, S. G. and S. J. Sharp. 1998. Explaining heterogeneity in meta-analysis: a comparison of methods. *submitted*

# sg42.2 Displaying predicted probabilities from probit or logit regression

Mead Over, World Bank, FAX: 202-522-3230

## Syntax

`probpred` *yvar xvar* [`if` *exp*] `,` `from(`*#*`)` `to(`*#*`)` [`inc(`*#*`)`
`adjust(`*covlist*`)` `one(`*varlist*`)` `zero(`*varlist*`)` `logit`
`level(`*#*`)` `poly(`*#*`)` `nomodel` `nolist` `noplot` *graph_options* ]

## Description

`probpred` is an extension of the `logpred` program of Garrett (1995). It first estimates a probit or logit regression of a dichotomous (or binary) dependent variable on a set of independent variables. The purpose of `probpred` is to compute and graph the estimated relationship between the predicted probability from this regression and one of the independent variables, holding the others constant. By default, both `probpred` and `logpred` display the regression estimates and a graph and listing of the requested predictions and the forecast interval. `probpred` contains four additional options not included in the original `logpred` program: `logit`, `one`, `zero`, and `level`. The default is to estimate a probit regression using the Stata command `dprobit`, but the `logit` option instead estimates a logit model using the `logistic` command. The `one` and `zero` options allow the user to specify that some of the covariates listed in option `adjust` are to be set equal to one or zero instead of to their means. The `level` option allows forecast intervals to be set to confidence levels determined by the user rather than only to 95% confidence levels.

## Options

`from(`*#*`)` specifies the lowest value of *xvar* for which a prediction is to be calculated. This option is required.

`to(`*#*`)` specifies the highest value of *xvar* for which a prediction is to be calculated. This option is required.

`inc(`*#*`)` specifies the increment between adjacent values of *xvar*. The default increment is 1.

`adjust(`*covlist*`)` specifies the other covariates in the model all of which are set to their sample means in computing the predicted values unless the `one` or `zero` options are specified as described below.

`one(`*varlist*`)` specifies a subset of *covlist* to be set to one instead of to their mean values in the data.

`zero(`*varlist*`)` specifies a subset of *covlist* to be set to zero instead of to their mean values in the data.

`logit` specifies that a logit model will be used. The default is probit.

`level(`*#*`)` specifies the confidence level to be used in computing and displaying the forecast interval.

`poly(`*#*`)` indicates that *xvar* enters the model as a polynomial. Quadratic and cubic models are allowed. They are indicated by `poly(2)` and `poly(3)`, respectively. The polynomial terms are created and entered in the regression automatically.

`nomodel` suppresses the display of the estimated regression.

`nolist` suppresses the list of predicted values.

`noplot` suppresses the graph of predicted values.

*graph_options* allowed include `xlabel`, `ylabel`, `saving(`*filename*`)`, and `titles`.

## Example

Using the coronary heart disease data that Garrett used (Garrett 1995), we have that

```
. probpred chd age, from(20) to(80) inc(5) adj(smk chl)
```

calculates the predicted probability of coronary heart disease (`chd`) for five-year increments of age from 20 years to 80 years adjusted for smoking status (`smk`) and serum cholesterol (`chl`), and results in the following output:

```
Iteration 0:  Log Likelihood = -438.5583
Iteration 1:  Log Likelihood =-413.31477
Iteration 2:  Log Likelihood =-412.92502
Iteration 3:  Log Likelihood =-412.92466

Probit Estimates                                        Number of obs =   1218
                                                        chi2(3)       =  51.27
                                                        Prob > chi2   = 0.0000
Log Likelihood = -412.92466                             Pseudo R2     = 0.0584
------------------------------------------------------------------------------
     chd |      dF/dx   Std. Err.      z    P>|z|     x-bar  [    95% C.I.   ]
---------+--------------------------------------------------------------------
     age |   .0049078   .0009072     5.32   0.000   53.7061    .00313  .006686
    smk*|   .0768262   .0167212     4.19   0.000   .635468   .044053  .109599
     chl |    .000716   .0002149     3.30   0.001   211.739   .000295  .001137
---------+--------------------------------------------------------------------
  obs. P |   .1165846
 pred. P |   .1033125  (at x-bar)
------------------------------------------------------------------------------
(*) dF/dx is for discrete change of dummy variable from 0 to 1
    z and P>|z| are the test of the underlying coefficient being 0

Probabilities and 95% Confidence Intervals

  Outcome Variable:     Coronary heart disease -- chd
  Independent Variable: Age in years -- age
  Covariates:           smk chl
  Variables set to Zero:
  Variables set to One:
  Total Observations:   1218

            age       pred      lower      upper
  1.         20   .0145038   .0052657   .0352342
  2.         25    .020337   .0088225   .0426535
  3.         30   .0280478   .0143056   .0513626
  4.         35   .0380526   .0224407   .0615699
  5.         40   .0507943   .0340208   .0735753
  6.         45   .0667224   .0497244   .0879035
  7.         50   .0862665   .0696591   .1056709
  8.         55   .1098053   .0925022    .129317
  9.         60   .1376323   .1156769   .1622731
 10.         65   .1699221   .1382471   .2059036
 11.         70   .2066993   .1610617   .2592201
 12.         75   .2478149   .1848735   .3206733
 13.         80    .292933   .2100355   .3884447
```

The graphical output of the command is given in Figure 1.

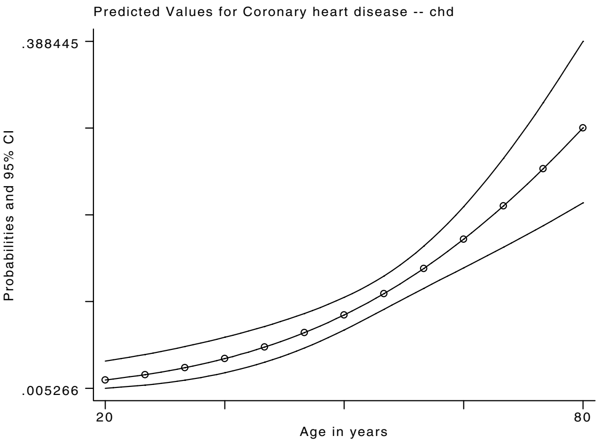

Figure 1. Probability of coronary heart disease as function of age

## References

Garrett, J. 1995. sg42: Plotting predicted values from linear and logistic regression models. *Stata Technical Bulletin* 26: 18–23. Reprinted in *Stata Technical Bulletin Reprints*, vol. 5, pp. 111–116.

## sg76 An approximate likelihood-ratio test for ordinal response models

Rory Wolfe, Royal Children's Hospital, Australia, wolfer@cryptic.rch.unimelb.edu.au
William Gould, Stata Corporation, wgould@stata.com

## Introduction

An ordinal response is made on a $J$-category ordered scale. A popular model for $i = 1, \ldots, n$ ordinal responses $y_i$ is

$$\mathrm{link}(\gamma_{ij}) = \kappa_j - \mathbf{x}_i\beta, \qquad j = 1, \ldots, J-1$$

where $\gamma_{ij} = \Pr(y_i \leq j|\mathbf{x}_i)$ is the cumulative probability for category $j$ which has associated with it a "cut-point" parameter $\kappa_j$ related to the cumulative probability of $y_i$ when $\mathbf{x}_i = 0$. The vector $\mathbf{x}_i$ contains values of explanatory variables and the effects of these variables are quantified by $\beta$ (a parameter vector of length $s$). Stata 5.0 has the facility to fit three such models; a model with a logit link can be fitted using the `ologit` command, a probit link model can be fitted using `oprobit`, and the complementary-log-log link model can also be fitted because of its equivalence to a continuation-ratio model (Läärä and Matthews 1985) which in turn can be fitted (after appropriate modifications to the data) by specifying a complementary-log-log link for binary data using the `glm` command.

The model given above defines effects of explanatory variables on cumulative response probabilities that are constant across all categories of the ordinal response. In the case of a logit link this constancy of effect across categories is usually called the proportional-odds assumption, for the complementary-log-log link it is called the proportional-hazards assumption.

Suppose, for example, that $y_i$ is a measure of smoking frequency with categories "None", "Occasional", and "Frequent", and the model only contains the explanatory variable "Sex", coded as females = 0 and males = 1. Then the difference in effect size between males and females is

$$\mathrm{link}(\gamma_{0j}) - \mathrm{link}(\gamma_{1j}) = \beta_{sex}, \qquad j = 1, 2$$

i.e. a constant effect across the three categories. In terms of underlying prevalence rates, this model specifies that the difference between males and females is that category probabilities are shifted up (if $\beta_{sex}$ is positive) the response scale for males, relative to females.

In this example we might suspect that the prevalence of "Frequent" smoking is greater for males than females but that the combined prevalence of "Occasional" and "Frequent" smoking is greater for females than males. This implies a sex effect that is not constant across the categories of the response. A generalization of the first model that allows for such a nonconstant effect is

$$\mathrm{link}(\gamma_{ij}) = \kappa_j - \mathbf{x}_i\beta_j, \qquad j = 1, \ldots, J-1$$

where now the effect parameters $\beta_j$ are allowed to vary over the categories.

The purpose of the command `omodel` is to fit the first model and to test whether the fit to the data provided by the built-in assumption of effect constancy across categories can be improved upon by fitting the one above (in which the effects of all explanatory variables are different at each category).

## Method

The fitting of the first model is described in the Stata manuals under `ologit` or `oprobit` as appropriate. The subsequent test performed by `omodel` is an approximate likelihood-ratio test. Likelihood-ratio tests are described in general by McCullagh and Nelder (1989, Appendix A).

For comparing the first and second models, the likelihood-ratio test is defined as

$$-2\left[l_m(\mathbf{y};\widehat{\beta}^{mle}_{(1)}) - l_m(\mathbf{y};\widehat{\beta}^{mle}_{(2)})\right] \sim \chi^2_{s(J-2)}$$

where $l_m(.)$ represents the multinomial log-likelihood function (which implicitly takes the specified link function into account) and $\widehat{\beta}^{mle}_{(k)}$ denotes the maximum-likelihood estimate of the parameters in model $(k)$ (here we consider the $\tau_j$ parameters and the vector $\beta$ to be included in $\beta_{(k)}$). The problem with using this test is that the second model cannot be fitted in Stata by maximum likelihood. Hence we employ an approximation to the above test using

$$-2\left[l_m(\mathbf{y};\widehat{\beta}^{ib}_{(1)}) - l_m(\mathbf{y};\widehat{\beta}^{ib}_{(2)})\right] \sim \chi^2_{s(J-2)}$$

where $\widehat{\beta}^{ib}_{(k)}$ denote estimates that are obtained from an alternative estimation method which will be referred to as “independent binaries”.

In brief, independent binaries estimation of a model for an ordinal response employs the $J-1$ binary responses defined as

$$y^{\dagger}_{ij} = \begin{cases} 1 & \text{if } y_i \le j \\ 0 & \text{otherwise} \end{cases}$$

for $j = 1,\ldots,J-1$ where $E(y^{\dagger}_{ij}|\mathbf{x}_i) = \gamma_{ij}$. The correlation between these binary responses is given by

$$\mathrm{corr}(y^{\dagger}_{ij}, y^{\dagger}_{ik}) = \sqrt{\frac{\gamma_{ij}(1-\gamma_{ik})}{\gamma_{ik}(1-\gamma_{ij})}}, \qquad j < k$$

but in estimation this fact is ignored. Hence for estimation purposes, the model is specified in terms of independent binary responses $y^{\dagger}_{ij}$ and a command for fitting binary responses (e.g. `logit` or `probit`) can be used to obtain the estimates.

As with maximum likelihood, the independent binaries method of estimation provides consistent estimates of the model parameters, but the estimation is inefficient compared to maximum likelihood. The approximation $l_m(\mathbf{y};\widehat{\beta}^{ib}_{(k)})$ will always be smaller than the true likelihood $l_m(\mathbf{y};\widehat{\beta}^{mle}_{(k)})$ since this quantity is by definition the maximum value of $l_m(\mathbf{y};\beta_{(k)})$. In our experience the approximation is remarkably good, and it certainly should be for large $n$ or any likelihood that is fairly flat around the maximum. Also note that in performing the approximate test, we subtract two quantities that are both underestimates so any error in the approximation will be reduced rather than magnified at this stage.

The null hypothesis of our test is that the $(J-1)$ $\beta_j$ parameter vectors in the second model are all equal to $\beta$ as in the first model. The test statistic is the second $\chi^2$ statistic above and a significant $p$ value is evidence to reject this null hypothesis and accept that the effects of the explanatory variables are not constant across the categories of the ordinal response.

## Syntax

`omodel logit` *varlist* [*weight*] [`if` *exp*] [`in` *range*]

`omodel probit` *varlist* [*weight*] [`if` *exp*] [`in` *range*]

`fweights` are allowed and provide a very useful way of dealing with data that are stored in a contingency table format; see the second example below.

## Example of ordinal by ordinal 2-way tables

The data presented by Agresti (1984, 12) on an undesirable side-effect outcome from different operations for treating duodenal ulcer patients are reproduced in Table 1.

Table 1: Cross-classification of dumping severity and operation

| | Dumping severity | | | |
|---|---|---|---|---|
| Operation | None | Slight | Moderate | Total |
| Drainage and vagotomy | 61 | 28 | 7 | 96 |
| 25% resection and vagotomy | 68 | 23 | 13 | 104 |
| 50% resection and vagotomy | 58 | 40 | 12 | 110 |
| 75% resection | 53 | 38 | 16 | 107 |
| Total | 240 | 129 | 48 | 417 |

Suppose we have these data in Stata as a record for each patient with variables “operate” (numbered 1–4) and “outcome” (response category numbered 1–3). An ordered logit model is fitted to the data using the command

```
xi: omodel logit outcome i.operate
```

and the output from the test of the proportional-odds assumption is

```
      chi2(3) =     3.72
  Prob > chi2 =   0.2932
```

If we had data on the patients age and sex we could fit a more elaborate model

```
xi: omodel logit outcome i.operate*age i.sex
```

and of course we could fit an ordered probit model by simply replacing `logit` with `probit` in the above commands.

## Cheese tasting experiment

Suppose that the following data (from McCullagh and Nelder 1989, 175) are to be analyzed with a cumulative logit model to detect for differences in preference of four cheese additives.

Table 2: Multiple-response cheese-tasting experiment

| | Response category | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Cheese | I* | II | III | IV | V | VI | VII | VIII | IX† | Total |
| A | 0 | 0 | 1 | 7 | 8 | 8 | 19 | 8 | 1 | 52 |
| B | 6 | 9 | 12 | 11 | 7 | 6 | 1 | 0 | 0 | 52 |
| C | 1 | 1 | 6 | 8 | 23 | 7 | 5 | 1 | 0 | 52 |
| D | 0 | 0 | 0 | 1 | 3 | 7 | 14 | 16 | 11 | 52 |
| Total | 7 | 10 | 19 | 27 | 41 | 28 | 39 | 25 | 12 | 208 |

*I=strong dislike; †IX=excellent taste

These data can be entered into Stata most simply as the contingency table of counts that is presented here, i.e., three variables: “response category” (with values 1–9, labeled I–IX), “additive” (values 1–4, labeled A–D) and “count” (which contains the cell counts of the table). When data are stored in this form, `omodel` can be used as follows

```
xi: omodel logit response i.additive [fweight=count]
```

with the result from the approximate likelihood-ratio test statistic in this case being 44.64 on 16 degrees of freedom. Note that the degrees of freedom are less than $s(J-2)=21$ because of the zeroes in the table.

The $p$ value for the test is 0.0002 so the null hypothesis is rejected and we conclude that the preferences between the cheeses cannot be described by effects that are constant across all categories.

## Technical notes

In this second example, the first attempt to use `omodel` resulted in the error “matsize too small”, which is remedied by, e.g. `set matsize 150`. When `omodel` performs the approximate test it is necessary to expand the data to $(J-1)$ times $n$ and fit two logistic models to the expanded data. If $J$ and $n$ are large, this can take an appreciable amount of time. Also, if as in this example, $J$ is large, then the `matsize` problem is likely to be encountered.

In the second example, the log-likelihood for the model is $-355.67$ and the approximation to this from independent binaries estimation is $-355.98$.

## Saved results

| | |
|---|---|
| `S_1` | Log-likelihood ratio |
| `S_2` | Test degrees of freedom |

## Discussion

The new command `omodel` provides an approximate likelihood-ratio test for the cumulative logit and probit models. Note that in the case of the complementary-log-log model an exact test can be calculated by fitting the two models with judicious use of `glm`.

It is to be emphasized that the likelihood-ratio test does not provide a test of whether the first model fits the data adequately. This goodness-of-fit-type interpretation is not the correct conclusion to be reached from a nonsignificant $p$ value. A significant $p$ value provides strong evidence that a more general model is required. A nonsignificant $p$ value shows a lack of evidence that the general model provides an improvement in fit to the data.

The likelihood-ratio tests the first model against a completely general alternative, the second model. For $J > 3$ there is an intermediate, nested model in which the effect of all explanatory variables is described by two parameters (as opposed to one parameter in the first and $J-1$ parameters in the second). This model is defined as

$$\mathrm{link}(\gamma_{ij}) = \kappa_j \exp(\mathbf{x}_i\tau) - \mathbf{x}_i\beta, \qquad j = 1,\ldots,J-1$$

where as before, the parameters $\beta$ describe shifts of the underlying category probabilities up the response scale, and now the parameters $\tau$ describe concentration of underlying category probabilities towards the middle of the response scale. The above model cannot at present be estimated in Stata but is worthy of consideration since it describes a phenomenon that often occurs with ordinal response scales.

If a significant $p$ value is obtained from the likelihood-ratio test, it is possible that the above model is sufficient to describe all of the non-constancy of effects across categories. A further likelihood-ratio test comparing the second model and the one above would determine whether this was the case. For the data in Example 2, McCullagh and Nelder (1989, 177) examine the deviance reduction due to the above model and report it to be 3.3 on 3 degrees of freedom indicating that the non-proportionality of odds cannot be described in the simple fashion defined by the model above.

In the case of a nonsignificant $p$ value from the likelihood-ratio test, the deviance value can be compared to a $\chi^2$ distribution on $s$ degrees of freedom. This informal comparison tests whether the above model could provide an improved fit to the data (over the first model) if it accounted for the entire reduction in deviance between the first and second models.

An alternative to the likelihood-ratio test is to test a score-statistic (see McCullagh & Nelder 1989, Appendix A) with a null hypothesis that the general model parameter structure of the second model is equal to the structure defined by the first model. Such a score statistic test is provided by `PROC LOGISTIC` in SAS (SAS Institute 1989). For the dumping severity example the score statistic is 4.021 on 3 degrees of freedom ($p = 0.2592$), a similar result to the approximate likelihood-ratio test.

For the cheese tasting example, the test statistic from SAS is 17.29 on 21 degrees of freedom. In contrast to the likelihood-ratio test this is nonsignificant, hence the conclusion from the score test is that there is no evidence to reject the null hypothesis of proportional odds in favor of the general model. Note that the degrees of freedom presented by SAS makes no allowance for the sparse nature of the data.

Note that in the case of a single parameter the approximate theoretical error associated with a score statistic is of order $n^{-1/2}$ whereas for the likelihood-ratio test the approximate theoretical error is of order $n^{-1}$, hence the latter is theoretically preferable. The advantage of the score statistic in practice is in terms of computing effort, with only the first model needing to be fitted in comparison to the likelihood-ratio test which requires the fitting of both the first and second models.

## References

Agresti, A. 1984. *Analysis of Ordinal Categorical Data*. New York: John Wiley & Sons.

Läärä, E. and J. Matthews. 1985. The equivalence of two models for ordinal data. *Biometrika* 72: 206–207.

McCullagh, P. and J. A. Nelder. 1989. *Generalized Linear Models*. 2d ed. London: Chapman & Hall.

SAS Institute, Inc. 1989. *SAS/STAT Users Guide Vol.2, Version 6*. 4th ed. Cary, NC: SAS Institute.

## sg77 Regression analysis with multiplicative heteroscedasticity

Jeroen Weesie, Utrecht University, Netherlands, weesie@weesie.fsw.ruu.nl

One of the main assumptions in the standard linear regression model—as estimated in Stata by the commands `regress` and `fit`—is that the residuals are homoscedastic, i.e., all residuals have the same variance $\sigma^2$. What are the consequences when we use OLS-estimates while the homoscedasticity assumption is violated, i.e., the residuals are heteroscedastic? It is well-known that the OLS-estimates $\widehat{\beta} = (X'X)^{-1}Xy$ are still *unbiased* and consistent under standard regularity conditions. However, the standard OLS-based standard errors estimated as $\widehat{\sigma}^2(X'X)^{-1}$ are *wrong*, and hence most inferences about the regression coefficients are in jeopardy. It is possible to use a different estimator of the standard errors of the regression coefficient, known as the Huber/White/robust/sandwich estimator, and available in Stata with the option `robust`. This estimator of the standard error is consistent even if the residuals are heteroscedastic. Consequently, tests of hypotheses about regression coefficients that are based on these robust standard errors are consistent as well.

Robust estimation of standard errors, however, is *not* a panacea for possible heteroscedasticity. The reason is that under heteroscedasticity the OLS-estimators are *inefficient*: there exist alternative "more accurate" estimators, i.e. estimators with a uniformly smaller variance. This is known as the Gauss–Markov theorem in the theory of the linear model (Eaton 1983, Ch. 4). Consequently, the "robust" test of hypotheses on parameters is suboptimal.

How can we diagnose whether the homoscedasticity assumption is violated in some regression problem? A variety of tests have been proposed. Currently, Stata contains a score test (`hettest`) of homoscedasticity against a parametric model that specifies how the residual variance depends on a vector of covariates. In a variant of this score test, the residual variance is specified as a function of the expected value of the dependent variable. Stata's reference manual is silent on how to proceed if any of these tests are significant, i.e., if the homoscedasticity hypothesis has to be rejected. Heteroscedasticity of the latter form, i.e., $\sigma^2 = f(\mathrm{E}y)$, can usually be dealt with by a transformation of the dependent variable. The Box–Cox regression model, in which the transformation is estimated within the Box–Cox family, is quite suitable here. However, heteroscedasticity of the first type is more awkward. Here, we usually cannot avoid actually modeling the variance of the residual (or the dependent variable) in terms of covariates, just like we model the expected value of the dependent variable in terms of some covariates.

The command `regh` estimates a linear regression model with normally distributed residuals with the specific functional form for heteroscedasticity that is also assumed by `hettest`,

$$\begin{aligned} y_i =& \mu_i + \sigma_i e_i \\ \mu_i =& \mathrm{E}y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} \\ \sigma_i^2 =& \mathrm{Var}y(i) = \exp(\gamma_0 + \gamma_1 z_{i1} + \ldots + \gamma_m z_{im}) \end{aligned}$$

where $y_i$ is a random variable (the "dependent variable") with mean $\mu_i$ and variance $\sigma_i^2$, and $x_i$ and $z_i$ are (vectors of) covariates predicting the mean and log-variance of $y$ respectively. Thus, we have a linear model for the expected value (mean), and a log-linear model for the variance of a response variable, conditional on a set of covariates that predict the mean and variance. In addition, the residuals $e_i$ are assumed to be standard normal distributed and independent. (See the option `cluster` below for cluster-correlated observations, and for estimating a model with nonnormal residuals.) The (vector-) coefficients $\beta$ and $\gamma$ are to be estimated.

### Syntax diagrams

```
regh    eqmean eqlnvar [if exp] [in range] [, robust cluster(varname)
        twostage level(#) from(matname) maximize_options]

reghv   depvar [varlist] [if exp] [in range] , var(varlist) [ robust cluster(varname)
        twostage level(#) maximize_options]
```

*eqmean* is an equation that contains the dependent variable, followed by the $x$-variables. *eqlnvar* is an equation that contains the $z$-variables. A constant is automatically appended to both *eqmean* and *eqlnvar*. (Thus, if *eqlnvar* is empty, `regh` fits the same model as `regress`.) In the current implementation, the constants cannot be dropped. `regh` typed without arguments replays results. `regh` interfaces to post-estimation commands such as `test`, `predict` etc. in the standard way.

`reghv` simply sets up appropriate equations (named `lp_mean` and `lp_lnvar`), and interfaces to `regh`. Thus, to replay results type `regh`.

Note that the current implementations of `regh` and `reghv` do not support weights.

## Options

`var(`*varlist*`)` specifies the variables used to model the log(variance) of the residuals. (Only with `reghv`).

`robust` specifies that the robust method of calculating the (co)variance matrix is to be used instead of the traditional calculation (Harvey 1976). The robust variant of the covariance matrix is also computed for the 2SLS estimator.

`cluster(`*varname*`)` implies `robust` and specifies a variable on which clustering is to be based. The cluster-variable may be of type string.

`twostage` specifies that Harvey's 2SLS estimator (and the associated consistent (co)variance matrix estimate) should be computed, otherwise the maximum-likelihood estimator is used. If the residuals are normally distributed, the variance of maximum-likelihood estimators is approximately half the variance of the 2SLS-estimator.

`level(`*#*`)` specifies the confidence level, in percent, for confidence intervals. The default is `level(95)` or as set by `set level`.

`from(`*matname*`)` specifies a matrix (row-vector) with initial values. `from` should be properly named (see online help for `ml` for details). `from` enables efficient bootstrapping where one may use "full sample" estimates as starting values for the resamples.

*maximize_options* control the maximization process; see [R] **maximize**. You should never have to specify the more technical of these options, although we do recommend specifying the `trace` option.

## Additional programs

We have a number of additional programs (commands) to test for and model heteroscedasticity in a regression-like context. Contact the author if you are interested in any of these commands.

`htest` is a modification of Stata's `hettest` that easily provides univariate tests as well.

`white` is White's test for homoscedasticity ($H_0$) against general forms of heteroscedasticity ($H_a$) (see White 1980).

`szroeter` is Szroeter's semi-parametric test whether the residual variance is monotone in an exogenous discrete or continuous variable.

`regh2` is similar to regh but uses `ml/deriv2` (but Fisher-scoring rather than Newton–Raphson) rather than the hand-coded alternating scoring algorithm. `regh2` is considerably slower, but, may be more stable for very ill-conditioned problems.

`reghf` computes maximum-likelihood estimators for (location, scale) models with nonnormal residuals (e.g., Cauchy, logistic). There is some evidence that inference on variance-models in the context of regression models is less robust than inference on the model of the mean. Using `reghf` it is possible to verify whether conclusions depend on the assumption of normality. In case of doubt, use the 2SLS-estimator that is less efficient than the maximum-likelihood estimation under normality, but consistent for nonnormal residuals as well.

`reghv` computes maximum-likelihood estimators (for normally distributed residuals) for a variant of the multiplicative heteroscedasticity model that assumes

$$\log \sigma_i^2 = \alpha\mu_i + \gamma' z_{i*}$$

where $\alpha$ is a scalar parameter.

`probith` estimates probit models with multiplicative heteroscedasticity. Note that the probit model can be interpreted as a regression model with normally distributed residuals and severe missing values in the dependent variable $y$: it is only observed when $y \geq 0$. This generalizes immediately to the heteroscedastic probit regression model.

## Example

In this stylized example, I discuss an analysis of data on the management of R&D cooperative alliances, modeled after Blumberg (1997). We wanted to analyze the extent to which firms that engage in an R&D-oriented transaction with a partner, invest in ex-ante planning such as contracting (`contract`). From Williamson's transactions–costs economics and Granovetter-like sociological arguments about the effects of social embeddedness of the functioning of markets, it is expected that these investments increase with the risks involved (here indicated by the variables `volume` (log-volume-in-money) and `depend` (mutual dependency)), decrease to the extent that the partners involved have a common past (`past`) and expect to have a common future (`future`), and decrease to the extent that the partners have common ties to other firms (`network`).

```
. fit contract volume depend past future network
```

```
  Source |       SS       df       MS              Number of obs =      94
---------+------------------------------           F(  5,    88) =   49.86
   Model |  458.319345     5  91.663869            Prob > F      =  0.0000
Residual |  161.770291    88  1.83829876           R-squared     =  0.7391
---------+------------------------------           Adj R-squared =  0.7243
   Total |  620.089636    93  6.66763049           Root MSE      =  1.3558

------------------------------------------------------------------------------
contract |      Coef.   Std. Err.       t     P>|t|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
  volume |   .9070136   .1587244      5.714   0.000       .5915822    1.222445
  depend |   .4112368   .1490966      2.758   0.007       .1149387    .7075349
    past |   -.224738   .0171927    -13.072   0.000      -.2589049    -.190571
  future |   -.218904   .1034922     -2.115   0.037      -.4245731    -.013235
 network |  -.3401932   .1011277     -3.364   0.001      -.5411632   -.1392232
   _cons |  -1.679443   1.160729     -1.447   0.151      -3.986149    .6272628
------------------------------------------------------------------------------
```

Let us compare these results with OLS-estimates with “robust” standard errors.

```
. fit contract volume depend past future network, robust
Regression with robust standard errors                 Number of obs =      94
                                                       F(  5,    88) =   22.95
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.7391
                                                       Root MSE      =  1.3558

------------------------------------------------------------------------------
         |               Robust
contract |      Coef.   Std. Err.       t     P>|t|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
  volume |   .9070136   .1904846      4.762   0.000       .5284655    1.285562
  depend |   .4112368   .1038016      3.962   0.000       .2049529    .6175207
    past |   -.224738   .0259251     -8.669   0.000      -.2762586   -.1732174
  future |   -.218904   .0827287     -2.646   0.010      -.3833099   -.0544982
 network |  -.3401932   .0987157     -3.446   0.001        -.53637   -.1440164
   _cons |  -1.679443    .898974     -1.868   0.065      -3.465965    .1070788
------------------------------------------------------------------------------
```

The changes in estimates for standard errors are generally conservative. Next we test for homoscedasticity with a Weisberg–Cook’s score test using **htest**.

```
. htest, rhs u
Score tests for heteroscedasticity (Ho: Constant variance)
         |     var = f(a + b*z)     |
Variable |     score   df      p    |
---------+--------------------------+
volume   |     40.22    1   0.0000  |
depend   |      0.09    1   0.7604  |
past     |     94.68    1   0.0000  |
future   |      2.42    1   0.1201  |
network  |      0.27    1   0.6021  |
---------+--------------------------+
combined |    128.14    5   0.0000  |
```

According to the univariate tests of **htest**, it appears that the variance of the residuals can be modeled using only **volume** and **past**. White’s omnibus test for heteroscedasticity also indicates that the regression analysis is misspecified, possibly due to heteroscedasticity.

```
. white
White´s test for Ho: homoscedasticity
         against Ha: unrestricted heteroscedasticity
    test statistic W =      34.44544
    Pr(chi2(19) > W) =        0.0163
```

The multiple-equations command **regh** expects input via equations.

```
. eq mean : contract volume depend past future network
. eq var  : volume past
. regh mean var
iteration 1:  Log-likelihood = -80.76063  |delta-b|  2.126929
iteration 2:  Log-likelihood = -78.22403  |delta-b|  .2676002
iteration 3:  Log-likelihood = -78.05332  |delta-b|  .0455286
iteration 4:  Log-likelihood = -78.04507  |delta-b|  .0167279
iteration 5:  Log-likelihood = -78.04482  |delta-b|  .0032507
iteration 6:  Log-likelihood = -78.04482  |delta-b|  .0005627
Multiplicative heteroscedastic regression           Number of obs  =      94
Estimator: mle                                      Model chi2(7)  = 288.008
                                                    Prob > chi2    =   0.000
Log Likelihood              =   -78.045             Pseudo R2      =  0.6485
                                                    VWLS R2        =  0.9659
------------------------------------------------------------------------------
contract |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
mean     |
  volume |    .631765   .0351336     17.982   0.000       .5629045    .7006255
  depend |   .4844798   .0297727     16.273   0.000       .4261265    .5428332
    past |  -.2116046    .005701    -37.117   0.000      -.2227784   -.2004308
  future |  -.4432523   .0231963    -19.109   0.000      -.4887163   -.3977883
 network |  -.2814385   .0231456    -12.159   0.000       -.326803    -.236074
   _cons |   .2489145   .2559325      0.973   0.331      -.2527041    .7505331
---------+--------------------------------------------------------------------
var      |
  volume |   .5516058   .1561316      3.533   0.000       .2455935     .857618
    past |  -.2210131   .0179103    -12.340   0.000      -.2561166   -.1859097
   _cons |  -3.527318   .9874938     -3.572   0.000      -5.462771   -1.591866
------------------------------------------------------------------------------
```

The multi-panel output will look familiar to experienced Stata users. According to the estimated model, (the expected amount of) contracting increases with the volume and dependency associated with the transaction, and decreases with common past, expected common future, and common network ties. The variance of the residuals increases with the volume of the transaction, and decreases with a common past. Comparing the output of `fit`, `robust` to `regh`, it is obvious that the robust standard errors of OLS-estimates are much larger than those of the `regh`-model, and so the 95% confidence intervals have shrunk drastically.

Having improved our estimates of regression-coefficients, we cannot, however, be very confident that we modeled the heteroscedasticity correctly. Thus, we may be estimating the standard errors of the `regh`-coefficients inconsistently. This problem can be remedied via robust estimation of standard errors in the `regh` model.

```
.   regh mean var, robust nolog
Multiplicative heteroscedastic regression           Number of obs  =      94
Estimator: mle                                      Model chi2(7)  = 288.008
                                                    Prob > chi2    =   0.000
Log Likelihood              =   -78.045             Pseudo R2      =  0.6485
                                                    VWLS R2        =  0.9659
------------------------------------------------------------------------------
         |               Robust
contract |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
mean     |
  volume |    .631765   .0373337     16.922   0.000       .5585923    .7049378
  depend |   .4844798   .0301037     16.094   0.000       .4254776    .5434821
    past |  -.2116046   .0057706    -36.669   0.000      -.2229148   -.2002944
  future |  -.4432523    .024733    -17.922   0.000      -.4917281   -.3947766
 network |  -.2814385   .0164501    -17.109   0.000      -.3136801   -.2491969
   _cons |   .2489145   .2852164      0.873   0.383      -.3100994    .8079285
---------+--------------------------------------------------------------------
var      |
  volume |   .5516058   .1722091      3.203   0.001       .2140822    .8891293
    past |  -.2210131   .0208703    -10.590   0.000      -.2619182   -.1801081
   _cons |  -3.527318   1.090857     -3.234   0.001      -5.665359   -1.389277
------------------------------------------------------------------------------
```

Comparing the standard errors of `regh` with the robust standard errors of `regh, robust`, we conclude the effects of the possible misspecification of heteroscedasticity are apparently rather mild.

Note that `regh` produces a trace of the iterative process of computing maximum-likelihood estimates that is similar to the standard Stata's `trace`—I added the `L2-norm` of the change in the parameter vectors to better trace progress for ill-conditioned problems for which convergence was problematic before I introduced step-halving in the algorithm.

So far we have shown output of **regh** for maximum-likelihood estimators based on the assumption that the residuals are normally distributed. If you distrust this assumption strongly, you may prefer Harvey's two-stage least squares estimates (2SLS). Note that under normality, the approximate standard errors of maximum-likelihood estimators are roughly 1.4 times smaller than those of 2SLS estimators (Harvey 1976).

```
. regh mean var, twostage
Multiplicative heteroscedastic regression            Number of obs  =      94
Estimator: 2sls                                      Model chi2(7)  = 222.900
                                                     Prob > chi2    =   0.000
Log Likelihood             =  -110.599               Pseudo R2      =  0.5019
                                                     VWLS R2        =  0.9027
------------------------------------------------------------------------------
contract |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
mean     |
  volume |   .9070136   .0651181     13.929   0.000       .7793844    1.034643
  depend |   .4112368   .0543363      7.568   0.000       .3047396     .517734
    past |  -.224738    .0088016    -25.534   0.000      -.2419888   -.2074872
  future |  -.218904    .0426059     -5.138   0.000       -.30241    -.1353981
 network |  -.3401932   .0420307     -8.094   0.000      -.4225718   -.2578146
   _cons |  -1.679443   .4606367     -3.646   0.000      -2.582274   -.7766116
---------+--------------------------------------------------------------------
var      |
  volume |   .3422644   .2452509      1.396   0.163      -.1384185    .8229472
    past |  -.1401075   .0281334     -4.980   0.000      -.1952479   -.0849671
   _cons |  -2.081192   1.551151     -1.342   0.180      -5.121393    .9590086
------------------------------------------------------------------------------
```

## Notes and references

Harvey (1976) suggests a 2SLS-estimator in which $y_i$ is regressed on $x_i$ to estimate $\beta$, and to regress the log-squared-residuals of this regression on $z_i$ to estimate $\gamma$. After adding 1.2704 to $\widehat{\gamma}_0$, this estimator of $\beta, \gamma)$ is consistent (Harvey 1976). This estimator is available with the option `twostage`.

To compute maximum-likelihood estimators for normally distributed residuals, **regh** uses Harvey's alternating scoring algorithm, starting from the 2SLS estimator (See also Greene 1993). I added step-halving to improve stability of the algorithm. In my experience, this algorithm is fast and converges well.

Robust standard errors and the adaptation of the standard errors for clustered observations follows the method implemented by the `_robust` command in Stata.

## Acknowledgment

I am grateful to Chris Snijders and Vincent Buskens for testing these commands. This research was supported by PIONIER-grant PGS 50-370 of the Netherlands Organization for Scientific Research.

## References

Blumberg, B. 1997. Das Management von Technologiekooperationen. Partnersuche und Verhandlungen mit der Partner aus empirisch-theoretischer Perspektive. Amsterdam: Thesis.

Eaton, M. L. 1983. *Multivariate Statistics. A Vector Space Approach*. New York: John Wiley & Sons.

Greene, W. H. 1993. *Econometric Analysis*. 2d ed. New York: Macmillan.

Harvey, A. C. 1976. Estimating regression models with multiplicative heteroscedasticity. *Econometrica* 44: 461–465.

White, H. 1980. A heteroskedastic-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica* 48: 817–830.

## sg78 Simple and multiple correspondence analysis in Stata

Philippe Van Kerm, University of Namur, Belgium, philippe.vankerm@fundp.ac.be

Correspondence analysis is a technique that often turns out to be helpful when one wants to analyze large cross-tabular data (see Benzecri and Benzecri 1980 or Greenacre and Blasius 1994). It results in a simple graphical display which permits rapid interpretation and understanding of the data. It is currently available in a number of statistical packages (e.g. SPSS, SAS, or BMDP), but not in Stata 5.0. This insert aims at filling this gap (at least partly). The commands `coranal` and `mca` provide the general features of simple and multiple correspondence analysis and are designed for standard applications. Admittedly, they might be inefficient for more specific uses and suffer from some deficiencies outlined below. As I am not myself an experienced user of these techniques, I will neither enter into technicalities nor provide precise explanations on how to use and interpret the results. Furthermore, the very simple example provided here is just intended to give an idea about how the output can look; it does not substitute for an introductory text. The references given at the end of the insert should accomplish these tasks much better than I can.

The two major drawbacks of the current versions of the procedures are 1) the aspect ratio distortion of the maps (the scales on the vertical axes are usually different from that on the horizontal axes—this is innocuous in most graphs but it may sometimes be misleading if one is not aware of this); it may be necessary to edit the graphs to solve this problem, and 2) the absence of an option for including supplementary points in the analysis; this ought to be made available in the future.

Future developments of these commands should also include the creation of immediate command versions or quasi-immediate commands; that is, using cross-tabulations or matrices as inputs and not just raw data. However, a trick for doing so with the current version is suggested in the help file for `coranal`.

## Simple correspondence analysis

The command `coranal` produces numerical results as well as graphical outputs for simple correspondence analyses (that is, analyses of two-way cross-tabulations of variables, or said differently, analyses of contingency tables). The computation algorithm draws, on an almost step-by-step basis, on Blasius and Greenacre (1994).

## Syntax

`coranal` *var1 var2* [*weight*] [`if` *exp*] [`in` *range*] [`,` `d(`*#*`)` `q(`*#*`)` `asymmetric`]

`aweight`s, `fweight`s, and `iweight`s are allowed.

## Options

`d(`*#*`)` specifies the number of dimensions to be considered (for both numerical and graphical displays). If `d(0)` is specified, then `coranal` provides no graphical display and returns the numerical output for all nontrivial dimensions. For maps to be readable, # must be set larger than 1. Furthermore, consistent maps can only be obtained by specifying # lower than or equal to the number of underlying non-trivial dimensions. The default # is 0.

`q(`*#*`)` specifies a quality of representation threshold ($0 < \# \leq 1$). It restricts the mappings to points satisfying the condition that their quality of representation (sum of contributions of principal axes) in the `d(`*#*`)` first dimensions is higher than or equal to *#*. Rejected points are still mapped but symbolized by a dot.

`asymmetric` specifies that the joint displays of *var1* and *var2* are to be presented in the form of asymmetric maps (both variables are taken as vertices consecutively). By default, symmetric maps are displayed.

## Example

To briefly illustrate the command, consider the following example. The dataset `exca.dta` (included with this insert) contains a fictional survey. Available individual characteristics are the main leisure-time activity (`VAR1`), the age group (`VAR2`), the place of birth (`VAR3`) and the gender (`VAR4`). Say we want to explore the links between age groups and leisure-time activity.

```
. tab VAR1 VAR2 ,chi row

           | VAR2
      VAR1 |     Child    Teenage      Adult    Elderly |     Total
-----------+--------------------------------------------+----------
        TV |        12          5          3          4 |        24
           |     50.00      20.83      12.50      16.67 |    100.00
-----------+--------------------------------------------+----------
      Walk |         1          1          4          1 |         7
           |     14.29      14.29      57.14      14.29 |    100.00
-----------+--------------------------------------------+----------
      Cook |         1          0          2          6 |         9
           |     11.11       0.00      22.22      66.67 |    100.00
-----------+--------------------------------------------+----------
     Sport |         5          8          4          1 |        18
           |     27.78      44.44      22.22       5.56 |    100.00
-----------+--------------------------------------------+----------
     Games |        11          2          1          3 |        17
           |     64.71      11.76       5.88      17.65 |    100.00
-----------+--------------------------------------------+----------
     Total |        30         16         14         15 |        75
           |     40.00      21.33      18.67      20.00 |    100.00

          Pearson chi2(12) =  33.9076   Pr = 0.001
```

This table and the $\chi^2$ statistic reveals that there indeed exists a strong relation between the two variables. Let us apply now a simple correspondence analysis on this table for an easy representation of the apparent associations between the various categories (see the references given below for complete explanations about the interpretations of the following numerical and graphical output).

```
. coranal VAR1 VAR2
------------------------------------------------------------------------------
                   SIMPLE CORRESPONDENCE ANALYSIS
------------------------------------------------------------------------------

 Total Inertia :      0.452

 Principal Inertias and Percentages :

      Inertia    Share    Cumul
Dim1    0.235    0.519    0.519
Dim2    0.173    0.382    0.902
Dim3    0.045    0.098    1.000

 VAR1 coordinates :

           Mass  Inertia     Dim1     Dim2     Dim3
VAR1:1    0.320    0.016   -0.121    0.191    0.014
VAR1:2    0.093    0.093    0.221   -0.842    0.490
VAR1:3    0.120    0.182    1.213   -0.039   -0.213
VAR1:4    0.240    0.096   -0.443   -0.388   -0.229
VAR1:5    0.227    0.065   -0.094    0.509    0.133

 VAR2 coordinates :

           Mass  Inertia     Dim1     Dim2     Dim3
VAR2:1    0.400    0.092   -0.224    0.406    0.121
VAR2:2    0.213    0.098   -0.531   -0.297   -0.297
VAR2:3    0.187    0.103    0.159   -0.673    0.269
VAR2:4    0.200    0.160    0.866    0.132   -0.177

 Explained inertia of axis by VAR1 :

          Dim1    Dim2    Dim3
VAR1:1  0.0198  0.0672  0.0014
VAR1:2  0.0193  0.3829  0.5038
VAR1:3  0.7518  0.0011  0.1225
VAR1:4  0.2006  0.2088  0.2817
VAR1:5  0.0085  0.3400  0.0906

 Explained inertia of axis by VAR2 :

          Dim1    Dim2    Dim3
VAR2:1  0.0856  0.3820  0.1324
VAR2:2  0.2558  0.1086  0.4222
VAR2:3  0.0202  0.4892  0.3040
VAR2:4  0.6384  0.0202  0.1415

 Contributions of principal axes to VAR1 :

          Dim1    Dim2    Dim3
VAR1:1  0.2850  0.7112  0.0038
VAR1:2  0.0488  0.7104  0.2408
VAR1:3  0.9690  0.0010  0.0299
VAR1:4  0.4919  0.3771  0.1310
VAR1:5  0.0307  0.9071  0.0623

 Contributions of principal axes to VAR2 :

          Dim1    Dim2    Dim3
VAR2:1  0.2185  0.7175  0.0641
VAR2:2  0.6151  0.1923  0.1925
VAR2:3  0.0461  0.8223  0.1316
VAR2:4  0.9387  0.0219  0.0395
var1lb:
             1 TV
             2 Walk
             3 Cook
             4 Sport
             5 Games

var2lb:
             1 Child
             2 Teenage
             3 Adult
             4 Elderly
```

The first table reveals that it is possible to obtain a two-dimensional graphical representation of the observed relation between VAR1 and VAR2 by, very roughly speaking, retaining 90 percent of information as Dim1 and Dim2 account for 90 percent of total inertia (note that the total inertia is equal to the $\chi^2$ statistic divided by the number of observations).

```
. coranal VAR1 VAR2 ,d(2)
```

(*output omitted*)

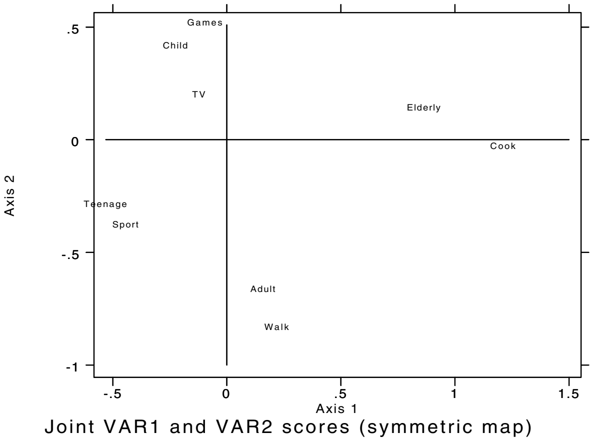

Figure 1: An example of simple correspondence analysis

On this graph, we can now observe how (dis-)similar are the profiles of the row points (or column points): categories that are similar to each other regarding their distribution over the categories of the other variable are plotted close to each other in the figure. Basically, the distances between the row points (or column points) in this display are to be seen as the best approximations in two dimensions of the $\chi^2$ distances between the row profiles (or column profiles) in Euclidian space. Note that the exact location of a point is situated right in the middle of the label name.

By default, coranal produces such "symmetric" maps which are usually more readable. Yet the distances between VAR1 points and VAR2 points are not directly comparable. To interpret such distances one has to display "asymmetric" maps (set the asymmetric option). We might also seek to avoid interpreting points for which the quality of representation is judged to be low (below 85 percent, say), where quality is measured by the contributions of the first two principal axes. Note that this is a very high threshold level.

```
. coranal VAR1 VAR2 ,d(2) as q(0.85)
```

(*output omitted*)

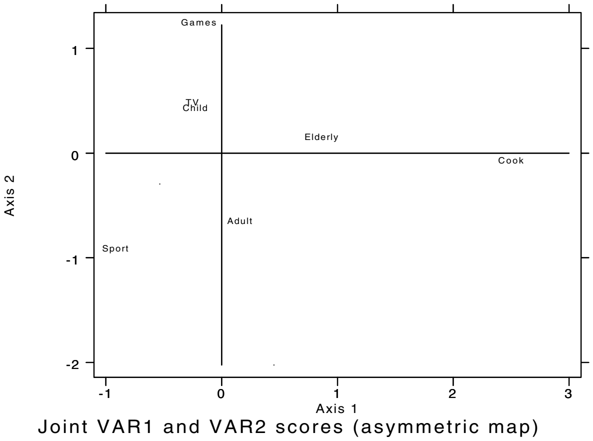

Figure 2: Using the asymmetric and q options

The points representing the profiles of the "Teenage" category and the "Walk" activity have been removed from the display since their quality of representation is below 85 percent. In this display, the activities have been taken as vertices to which the position of the age groups are to be compared. We could now observe, for instance how "Cook" is related to the remaining age groups: this occupation is apparently more a delight for the "Elderly" then for the "Adult" and eventually for the "Child" since the distance between the vertex "Cook" and the "Elderly" group is the smallest, the next smallest being with the "Adult" group and so on. This actually represents the row profile of "Cook" in the cross-tabulation of VAR1 and VAR2 above in the right order.

## Multiple correspondence analysis

The command `mca` produces numerical results and graphical representations for multiple correspondence analyses (MCA).

It must be emphasized that there is some controversy about the use and implementation of multiple correspondence analyses (see Greenacre 1984, 1993, 1994). `mca` actually conducts an adjusted simple correspondence analysis on the Burt matrix constructed with *varlist* (i.e. matrix of frequency counts resulting from all two-way cross-tabulations of the variables in *varlist* including the cross-tabulations of each variable with itself). It can be shown that the total inertia of the Burt matrix is high due to the fitting of the diagonal sub-matrices. Consequently, a simple correspondence analysis applied to this matrix usually results in maps of apparently poor quality. As a remedy, if not otherwise specified, `mca` adjusts the obtained principal inertias (eigenvalues) following a method suggested by Benzecri and presented in Greenacre (1984).

The computation algorithm draws most largely on Blasius and Greenacre (1994) and Greenacre (1994).

## Syntax

`mca` *varlist* [*weight*] [`if` *exp*] [`in` *range*] [`, d(`*#*`) q(`*#*`) notrans`]

`aweight`s, `fweight`s, and `iweight`s are allowed.

## Options

`d(`*#*`)` specifies the number of dimensions to be considered (for both numerical and graphical displays). If `d(0)` is specified, then `mca` provides no graphical display and returns the numerical output for all nontrivial dimensions. For maps to be readable, # must be set larger than 1. Furthermore, consistent maps can only be obtained by specifying # lower than or equal to the number of underlying nontrivial dimensions. Default # is 0.

`q(`*#*`)` specifies a quality of representation threshold (0 <#≤ 1). It restricts the mappings to points satisfying the condition that their quality of representation (sum of contributions of principal axes) in the `d(`*#*`)` first dimensions is higher than or equal to #. Rejected points are still mapped but symbolized by a dot. Note that the appropriateness of this option is somewhat questionable in the case of multiple correspondence analyses due to the above outlined inertia inflation of Burt matrices.

`notrans` requests that no adjustments to eigenvalues be made.

## Example

To illustrate `mca`, let us return to our fictional survey dataset. We want now to have an idea about how all the variables relate to each other and not only `VAR1` and `VAR2`. Say we conjecture that keeping two dimensions will retain enough information. We can use

```
. mca VAR1 VAR2 VAR3 VAR4, d(2)
------------------------------------------------------------------------------
                   MULTIPLE CORRESPONDENCE ANALYSIS
------------------------------------------------------------------------------

 Total Inertia :      0.108
 Principal Inertia Components :
      Inertia    Share    Cumul
Dim1    0.060    0.557    0.557
Dim2    0.027    0.248    0.805
 Coordinates :
           Mass  Inertia     Dim1     Dim2
VAR1_1    0.080    0.004   -0.105    0.155
VAR1_2    0.023    0.008    0.402   -0.262
VAR1_3    0.030    0.016    0.652   -0.037
VAR1_4    0.060    0.011   -0.309   -0.300
VAR1_5    0.057    0.004   -0.034    0.226
VAR2_1    0.100    0.006   -0.113    0.195
VAR2_2    0.053    0.011   -0.371   -0.251
VAR2_3    0.047    0.011    0.338   -0.243
VAR2_4    0.050    0.009    0.306    0.105
VAR3_1    0.103    0.005    0.192   -0.087
VAR3_2    0.027    0.007    0.002    0.369
VAR3_3    0.120    0.005   -0.165   -0.007
VAR4_1    0.110    0.007   -0.239   -0.008
VAR4_2    0.140    0.005    0.188    0.006
 Explained inertia of axes  :
           Dim1    Dim2
VAR1_1  0.0148  0.0717
VAR1_2  0.0629  0.0599
```

```
VAR1_3  0.2120  0.0015
VAR1_4  0.0954  0.2016
VAR1_5  0.0011  0.1084
VAR2_1  0.0211  0.1421
VAR2_2  0.1221  0.1261
VAR2_3  0.0887  0.1027
VAR2_4  0.0778  0.0205
VAR3_1  0.0631  0.0293
VAR3_2  0.0000  0.1355
VAR3_3  0.0546  0.0002
VAR4_1  0.1044  0.0003
VAR4_2  0.0820  0.0002
 Contributions of principal axes :
          Dim1    Dim2
VAR1_1  0.2389  0.5144
VAR1_2  0.4695  0.1991
VAR1_3  0.7990  0.0025
VAR1_4  0.5110  0.4807
VAR1_5  0.0167  0.7186
VAR2_1  0.2229  0.6668
VAR2_2  0.6621  0.3043
VAR2_3  0.4874  0.2513
VAR2_4  0.5265  0.0618
VAR3_1  0.7664  0.1585
VAR3_2  0.0000  0.5529
VAR3_3  0.6792  0.0012
VAR4_1  0.9385  0.0011
VAR4_2  0.9385  0.0011
```

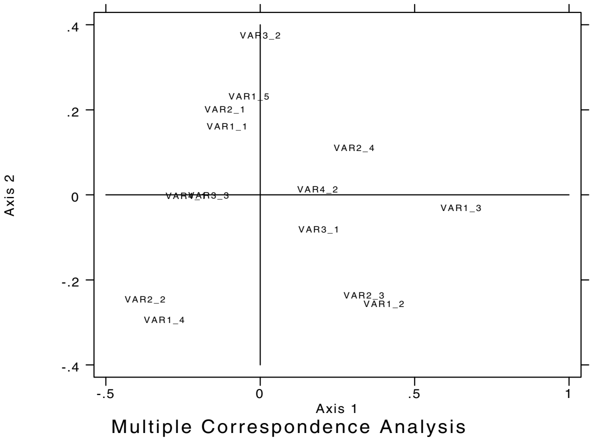

Figure 3

Unfortunately, maps resulting from multiple correspondence analyses may result in rather inextricable displays (as in this example) which may be difficult to interpret. By looking carefully at the numerical output and the figure, we can observe that the preceding MCA reveals, for instance, a certain tendency for the activities of the type “Walk” (VAR1_2) and “Cook” (VAR1_3) to be more associated with women (VAR4_2) while “TV” (VAR1_1) and “Sport” (VAR1_4) seem more associated with men (VAR4_1). It appears also that there is apparently relatively more women in the oldest age groups (VAR2_3 and VAR2_4) than in the youngest (VAR2_1 and VAR2_2).

## References

Benzecri J. P. and F. Benzecri. 1980. Analyse des correspondances: expose elementaire, Dunod, Paris.

Blasius J. and M. Greenacre. 1994. Computation of correspondence analysis. In *Correspondence Analysis in the Social Sciences—Recent Developments and Applications*, ed. M. Greenacre and J. Blasius. London: Academic Press.

Greenacre, Michael J. 1984. *Theory and Applications of Correspondence Analysis*. London: Academic Press.

——. 1993. *Correspondence Analysis in Practice*. London: Academic Press.

——. 1994. Multiple and joint correspondence analysis. In *Correspondence Analysis in the Social Sciences—Recent Developments and Applications*, ed. M. Greenacre and J. Blasius. London: Academic Press.

Greenacre M. and J. Blasius, eds. 1994. *Correspondence Analysis in the Social Sciences—Recent Developments and Applications*. London: Academic Press.

## sg79 Generalized additive models

Patrick Royston, Imperial College School of Medicine, UK, proyston@rpms.ac.uk
Gareth Ambler, Imperial College School of Medicine, UK, gambler@rpms.ac.uk

[*Editor's note: the* gam *program is only available for Stata for Windows.*]

### Introduction

Generalized Additive Models (GAMs) are flexible extensions of Generalized Linear Models (GLMs). Whereas the link function in a GLM is a linear combination of the predictors, in a GAM the link function may include flexible functions of arbitrary complexity based (for example) on smoothing splines. The amount of smoothing applied to each predictor is controlled by the user according to a quantity known as the equivalent degrees of freedom (df). The GLM is the special case with 1 df for each predictor.

GAMs may be used for exploratory analysis when one knows very little about the functional forms in a dataset. Plots of the smooth functions and a related significance test may assist the user to detect nonlinear relationships that were previously hidden. GAMs may also be useful to suggest functional forms for parametric modeling or for checking an existing parametric model for bias. A plot of the smooth function together with the corresponding parametric function for a given predictor may reveal "hot spots" of poor fit.

Generalized additive modeling is implemented in a FORTRAN program named GAMFIT. It was written by T. J. Hastie and R. J. Tibshirani, who introduced and developed the GAM, and is available from Statlib `http://lib.stat.cmu.edu/general/`. Here we provide an ado-file, `gam`, which is an interface between Stata and a slightly modified version of the FORTRAN program. We have only implemented the procedure to work on Windows platforms. The user is completely shielded from the original program. `gam` has the following basic syntax:

`gam` *yvar xvars* [`if` *exp*] [`in` *range*] [*major_options minor_options*]

Full details are given in the section *Syntax of the* `gam` *command.*

In this implementation cubic smoothing splines are used to estimate the flexible functions, although many other types of estimator could have been used such as kernel smoothers or locally weighted running line smoothers (`lowess`). In addition, `gam` can fit additive proportional hazards models to survival data. Details of all these methods can be found in Hastie and Tibshirani (1990a).

### The generalized additive model

We now discuss the generalized additive model, dealing with the additive proportional hazards model in the next section. The link function $g(\mu)$ for a GAM with $q$ predictors $x_1, \ldots, x_q$ has the form

$$g(\mu) = \alpha + \sum_{j=1}^{q} f_j(x_j)$$

where the $f_j$'s are smooth functions of the predictors. The "additive" in GAM refers to the fact that the predictor effects are assumed to be additive on the scale of the link function. The "generalized" alludes to the fact that we can use the error structures and link functions that are available with generalized linear models.

As a running example, we use data in the file `kyph.dat` derived from 81 patients who underwent corrective spinal surgery. The aim of the study was to assess the incidence of spinal deformities following surgery, and discover how various predictors affect this. The response $y$ is binary with $y = 1$ denoting the presence of kyphosis, which is defined to be a forward flexion of the spine of at least 40 degrees, and $y = 0$ denoting an absence. There are 3 predictors, age at time of surgery (`age`), the starting range of vertebrae levels involved in the operation (`start`), and the number of vertebral levels involved (`number`). These data were collected by Bell et al. (1994) and have previously been analyzed with GAMs by Hastie and Tibshirani (1990a).

We use a logistic additive model for these data and allow each function to have 4 degrees of freedom (*df*) via the `df` option. This option is very important since it controls the flexibility of the function estimators. The actual value of *df* for a predictor may be viewed as the number of parameters it uses. We now illustrate how the value of *df* can affect function estimates.

We fit the model logit $(p) = \alpha + f(\texttt{age})$ using different values of *df*. We chose the values 1, 4 and 15. Figure 1 shows the fitted values from these models on the probability scale. The 1 *df* model is the linear model logit $(p) = \alpha + \beta \cdot \texttt{age}$ and the estimates are the same as those produced by `logit` or `logistic`. The estimated smooth functions for the 4 *df* and 15 *df* exhibit a large amount of curvature, with the 15 *df* model displaying implausible wiggles everywhere, showing that it is overfitted. This demonstrates that care needs to be taken when using the `df` option.

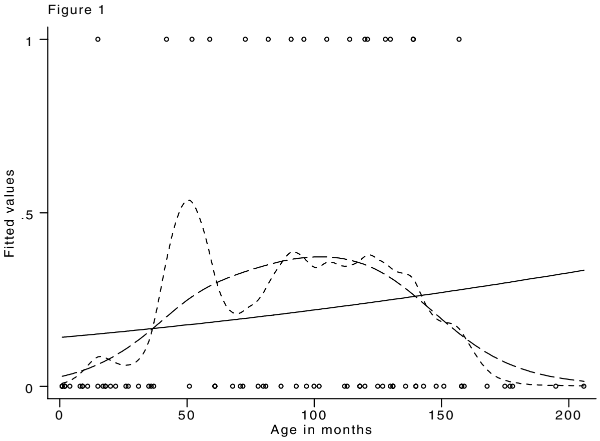

Figure 1: Fitted values for three values of degrees of freedom

The Stata command to fit the multivariable model with `age`, `start`, and `number` is

```
. gam y age start number,df(4) family(binomial)
```

which results in the following output:

```
Generalized Additive Model with family binom, link logit.
Model df     =    13.299                       No. of obs =        81
Deviance     =   44.1354                       Dispersion =         1
----------+-------------------------------------------------------------
        y |   df    Lin. Coef.  Std. Err.       z        Gain    P>Gain
----------+-------------------------------------------------------------
      age |  3.874   .0191194   .0055689     3.433     7.929    0.0430
    start |  4.220  -.3131638   .0633463    -4.944     7.922    0.0563
   number |  4.205   .3795253   .1825391     2.079     4.735    0.2159
    _cons |      1   -2.72886    .311933    -8.748         .         .
------------------------------------------------------------------------
Total gain (nonlinearity chisquare) =    20.586 (9.299 df), P = 0.0171
```

This table is similar in many ways to the table produced using the `logit` command with columns relating to the linear coefficient, its standard error, and $z$ score. However, there are additional columns relating to the degrees of freedom (`df`) and gain. The table reveals that the *df* used with each predictor is not equal to the exact value requested. This is because the program makes its best guess at the requested *df* initially and calculates the exact values at the convergence of the procedure. These values rarely agree exactly. We have noticed that transforming the predictors to approximate normality (see [R] **lnskew0**) can improve the accuracy of the final values, although the fitted functions will be different.

The gain allows us to assess apparent nonlinearities in the fitted functions. The gain for a given predictor is the deviance increase (i.e. worsening of fit) that occurs if a linear term is used to model the predictor instead of the smooth term. Predictors which are modeled linearly (*df* $= 1$) have no entries in the `Gain` column. The final column refers to an approximate significance test on the smooth function. Under the hypothesis that any nonlinearities in the smooth function are artifactual, the gain is assumed to have a chi-squared distribution with degrees of freedom equal to *df* $- 1$, which can be interpreted as the number of degrees of freedom associated with the nonlinear part of the relationship with the outcome variable. We subtract one to allow for the linear component of the smoothing spline fit whose coefficient is given in the `Lin. Coef` column, followed by its standard error. The nonlinear *df* is usually noninteger valued and the test statistic has approximately a gamma distribution with the appropriate parameters.

An examination of the table indicates that only the functions for `age` and `start` show any signs of nonlinearity. The model may be simplified and we do this using a stepwise procedure suggested by Hastie and Tibshirani (1990a). Each term may be specified as smooth with 4 *df*, linear, or excluded from the model. Starting with the model with every term specified as smooth we arrive at the model without `number`, whose command line is

```
. gam y age start,df(4) family(binomial)
```

This produces the following table of results:

```
Generalized Additive Model with family binom, link logit.

Model df     =     9.071                      No. of obs =        81
Deviance     =   49.9137                      Dispersion =         1
----------+----------------------------------------------------------
        y |   df    Lin. Coef.  Std. Err.       z        Gain    P>Gain
----------+----------------------------------------------------------
      age |  3.848   .0131614   .0062991     2.089      7.826    0.0441
    start |  4.223  -.2850923   .0641426    -4.445      9.184    0.0323
    _cons |      1    -2.0261    .349529    -5.797          .         .
---------------------------------------------------------------------
Total gain (nonlinearity chisquare) =    17.011 (6.071 df), P = 0.0097
```

In addition to the table, we obtain information about the fit from several new variables created automatically by gam. GAM_mu holds the fitted values on the original scale of the response, and information on the individual predictors is found in the variables prefixed by s_, e_ and r_. For example, including age in the model leads to creation of the variables s_age, r_age, and e_age. The variables prefixed by s_ are the smooth function estimates, those prefixed by e_ are the pointwise standard errors of the estimates, and those prefixed by r_ are the partial residuals which we now describe in detail.

The algorithm for fitting GAMs involves two iterative loops which are executed until convergence of the estimates occurs. The outer loop, the so-called local scoring algorithm, is similar to the Fisher scoring algorithm used to fit GLMs. This sets up an "adjusted dependent variable" used by the inner loop and recalculates weightings. The inner loop, the backfitting algorithm, fits a weighted additive model with the adjusted dependent variable as the response. The latter is assumed to be normally distributed although the variances are heterogeneous. Within the backfitting algorithm the weighted partial residuals for each predictor are smoothed as functions of the predictor values. The partial residuals for a given predictor are the raw residuals with the effect of the predictor removed. For example, if we denote the adjusted dependent variable by $z_i$ and the current estimated functions by $\widehat{f}_j(x_{ij})$ then the partial residuals for the $k$th predictor are given by

$$r_{ik} = z_i - \widehat{\alpha} - \sum_{j \neq k} \widehat{f}_j(x_{ij})$$

We give plots of the smooth functions and their pointwise standard errors for age and start in Figure 2. The standard errors give an idea of the accuracy of the estimated function at each observed predictor value but they do not provide global estimates of the uncertainty in the function.

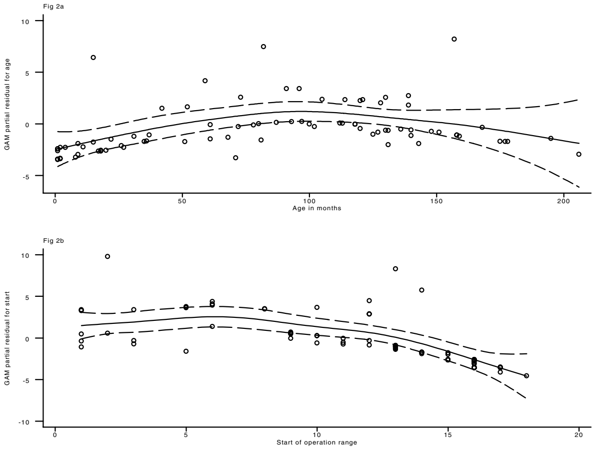

Figure 2: Smooth functions and standard errors for age and start

These plots reveal that each function appears to model the relationships in the data well.

## The additive proportional hazards model

One can use the gam command to fit the additive proportional hazards model. This model, introduced by Hastie and Tibshirani (1990b), extends the familiar semiparametric model of Cox (1972) by allowing flexible predictor effects. The additive proportional hazards model has the form

$$\lambda(t|x_1,\ldots,x_q) = \lambda_0(t)\exp\left\{\sum_{j=1}^{q} f_j(x_j)\right\}$$

where $\lambda_0(t)$ is the baseline hazard and the $f_j(x_j)$ are smooth functions which are estimated using smoothing splines.

We now present an example of the use of `gam` with survival data (given in the file `pbc.dat`) on 312 patients with primary biliary cirrhosis (PBC) who were referred to the Mayo Clinic between 1974 and 1984. They participated in a randomized, placebo-controlled trial of the drug D-penicillamine. Patients were followed up until July 1986 for survival status, and 125 died. The response `t` is survival time and the predictors we use are age (`age`), presence of edema (`edm`), serum bilirubin (`bil`), albumin (`alb`) and prothombin time (`pro`). Of these predictors, `edm` has 3 categories and is split into two indicator variables `edm2` and `edm3`. The censoring variable is denoted by `cens`. No difference in the survival times of the treated and untreated patients were detected, so we exclude the treatment variable from the model. More details of the study can be found in Fleming and Harrington (1991).

Initially we fit an additive proportional hazards model using

```
. gam t age bil alb pro edm2 edm3,dead(cens) df(4, edm2 edm3:1) family(cox)
```

which allows every predictor 4 *df* except `edm2` and `edm3` which are binary and so are given 1 *df*. This results in the following table:

```
Generalized Additive Model with family cox, link cox.
Model df     =    19.606                          No. of obs =       312
Deviance     =   1065.37                          Dispersion =         1
----------+-------------------------------------------------------------
        t |   df    Lin. Coef.  Std. Err.      z        Gain    P>Gain
----------+-------------------------------------------------------------
      age |  4.108   .0341721   .0090497     3.776     7.352    0.0666
      bil |  4.687   .1273225   .0156951     8.112    35.737    0.0000
      alb |  4.451  -1.054681   .2099035    -5.025     2.546    0.5472
      pro |  4.360   .2878167   .0935668     3.076     5.002    0.2108
     edm2 |      1    .186678      .2781     0.671         .         .
     edm3 |      1    .949492    .298307     3.183         .         .
------------------------------------------------------------------------
Total gain (nonlinearity chisquare) =    50.637 (13.606 df), P = 0.0000
```

After dropping `edm2` and any smooth terms which do not significantly contribute to the fit we arrive at the model presented below:

```
Generalized Additive Model with family cox, link cox.
Model df     =     9.447                          No. of obs =       312
Deviance     =   1076.57                          Dispersion =         1
----------+-------------------------------------------------------------
        t |   df    Lin. Coef.  Std. Err.      z        Gain    P>Gain
----------+-------------------------------------------------------------
      age |      1   .0355328   .0086492     4.108         .         .
      bil |  5.447   .1286926   .0158869     8.101    37.744    0.0000
      alb |      1  -1.031641   .2200582    -4.688         .         .
      pro |      1   .2603843   .0838921     3.104         .         .
     edm3 |      1    .947819    .285888     3.315         .         .
------------------------------------------------------------------------
Total gain (nonlinearity chisquare) =    37.744 (4.447 df), P = 0.0000
```

This table reveals that there is a strong nonlinear relationship involving serum bilirubin (`bil`). Figure 3 shows the smooth function for this variable and suggests that a logarithmic transformation of `bil` would be suitable as a simple parametric model. No partial residuals are shown as these are not created for the Cox model.

(*Figure 3 on the next page*)

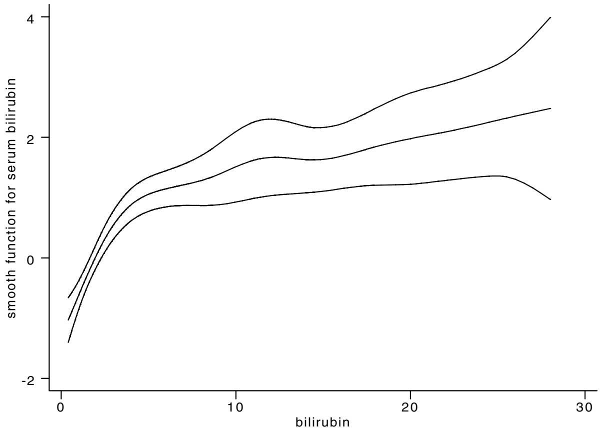

Figure 3: The smooth function for serum bilirubin for PBC data

We note that Figures 2 and 3 were produced using the `gamplot` program which we also provide. `gamplot` allows the user to easily plot the smooth functions produced by `gam` with options to include partial residuals and standard error bands. Full details are given in the section entitled *Syntax of the* `gamplot` *command*.

### Syntax of the gam command

`gam` *yvar xvars* [*weight*] [`if` *exp*] [`in` *range*] [`,` *major_options minor_options*]

`aweight`s and `fweight`s are allowed.

We note that `gam` expects to find the FORTRAN executables in the directory *c:\ado*. If they are not stored here, the global macro `GAMDIR` needs to be set. For example if the FORTRAN executable is stored in *d:\myado\*, you would enter

```
global GAMDIR d:\myado\
```

before calling `gam`.

### Options

The *major_options* (most used options) are

`family(`*family_name*`)` `link(`*link_name*`)` `df(`*dflist*`)` `dead(`*deadvar*`)`

where *family_name* is one of the following

`gaussian` | `binomial` | `poisson` | `gamma` | `cox`

and *link_name* is one of the following

`identity` | `logit` | `log` | `inverse` | `cox`

The *minor_options* (less used options) are

`big` `missing(`*#*`)` `noconstant`

### Major options

`family(`*family_name*`)` specifies the distribution of *yvar*. The default is `family(gaussian)`.

`link(`*link_name*`)` specifies the link function. The default for each family are the canonical links: `identity` for *family* `gauss`, `logit` for `binomial`, `log` for `poisson`, `inverse` for `gamma` and (by convention) `cox` for `cox`.

`df(`*dflist*`)` sets up the degrees of freedom (*df*) for each predictor. These values need not be integers. An item in *dflist* may be either # or *xvarlist*: #, where *xvarlist* is a subset of *xvars*. Items are separated by commas. With the first type of item, the *df* for all predictors are taken to be #. With the second type of item, all members of *xvarlist* have # df. If an item of the second type follows one of the first type, the later # overrides the earlier # for each variable in *xvars*. The default value is 1 for each predictor.

`dead(`*deadvar*`)` only applies to Cox regression where *deadvar* is the censoring status variable (0 for censored, 1 for “dead”).

## Minor options

`big` asks for the large-problem version of the GAMFIT FORTRAN program, `gambig.exe`. The largest problem that can be fit is 70,000 reals in the standard version of GAMFIT (`gamfit.exe`) and 1,000,000 reals in the big version (`gambig.exe`). These quantities are the amount of storage needed by the FORTRAN programs, not the amount of data stored in Stata. The problem size is approximated by the following formula: $\textit{reals} = 1000 \times N \times (\textit{vsize}^{0.2})/25$ where $N$ is the number of observations in the problem and *vsize* is the total number of variables, including the constant and *cens_var* if a Cox model is fit. For example, for a model with a constant and a single predictor ($\textit{vsize} = 2$) the biggest problems that can be fit are $N = 1523$ and $N = 21764$ for the standard and big versions respectively.

`noconstant` specifies that the model shall not have an intercept.

`missing(`*#*`)` defines the missing value code seen by `gam` to be *#*, which must be a number. The default is 9999. The FORTRAN program is able to impute missing values. See Hastie and Tibshirani (1990a) for more details.

## Syntax of the gamplot command

`gamplot` *xvar* [*xvar2*] [`if` *exp*] [`in` *range*] [`, noconf nopres se(`*#*`) abs(`*#*`)` *graph_cmd_options*]

where *xvar* is a member of *xvarlist*. If *xvar2* is specified the functions are plotted against this variable.

## Options

`noconf` suppresses the plotting of the pointwise standard errors.

`nopres` suppresses the addition of partial residuals on the plot.

`se(`*#*`)` determines the width of the standard error bands. The default is 0.95 which corresponds to 95% coverage (using assumptions of normality).

`abs(`*#*`)` is the maximum permitted distance between the smooth function and the partial residuals. Any partial residuals which exceed this value will not be plotted. The user is made aware of this. The default is $10^{-15}$.

*graph_cmd_options* are options appropriate to `graph`. The only options that cannot be used are `sort`, `symbol`, `connect`, and `pen` as these are already being used.

## Warning

We cannot vouch for the results from the FORTRAN software `gamfit.exe` and have occasionally noticed anomalies. However we believe it to be reliable in the vast majority of instances. `gam` can fail to converge with Cox regression and can occasionally cause Stata to shut down without warning. We find that this problem can usually be cured by changing the values of *df* slightly.

We also note that (non-binary) predictors are standardized before analysis. As a result the estimate and standard error of the intercept will differ from those produced using Stata commands such as `logit`, `cox`, and `regress`.

## Acknowledgment

This research received financial support from project grant number 045512/Z/95/Z from the Wellcome Trust.

## References

Bell, D. F., J. L. Walker, G. O' Conner, and R. Tibshirani. 1994. Spinal deformation after multiple-level cervical laminectomy in children. *Spine* 19: 406–411.

Cox, D. R. 1972. Regression models and life tables (with discussion). *Journal of the Royal Statistical Society, Series B* 74: 187–220.

Fleming, T. and D. Harrington. 1991. *Counting Processes and Survival Analysis*. New York: John Wiley & Sons.

Hastie, T. J. and R. J. Tibshirani. 1990a. *Generalized Additive Models*. London: Chapman and Hall.

——. 1990b. Exploring the nature of covariate effects in the proportional hazards model. *Biometrics* 46: 1005–1016.

# sg80 Indirect standardization

Mario Cleves, Stata Corporation, mcleves@stata.com

## Syntax

`istdize` *casevar*$_s$ *popvar*$_s$ *stratavars* `using` *filename* [`if` *exp*] [`in` *range*] `,`
{`popvars(`*casevar*$_p$ *popvar*$_p$`)` | `rate(`*ratevar*$_p$ *#*| *crudevar*$_p$`)`} [`by(`*groupvars*`) print format(%`*fmt*`) level(`*#*`)`]

## Description

`istdize` produces indirectly standardized rates for a study population based on a standard population. This standardization method is indicated when stratum-specific rates for the population being studied are either unavailable or are based on small samples and thus unreliable. The standardization uses the stratum-specific rates of a standard population to calculate the expected number of cases in the study population(s), sums them, and then compares them to the actual number of cases observed. The standard population's information is contained in another Stata data file and specified via `using` on the command line.

In addition to the indirectly standardized rate, the command produces point estimates and exact confidence intervals of the study population's standardized mortality ratio (SMR) if death is the event of interest or the standardized incidence ratio (SIR) for studies of disease incidence. Here we refer to both ratios as SMR.

*casevar*$_s$ is the variable name for the study population's number of cases (deaths). It must contain integers and each subpopulation identified by *groupvar* must have the same values or missing.

*popvar*$_s$ identifies the number of subjects represented by each observation in the study population.

*stratavars* define the strata.

## Options

`popvars(`*casevar*$_p$ *popvar*$_p$`)` or `rate(`*ratevar*$_p$ *#*|*crudevar*$_p$`)` must be specified. Only one of these two options is allowed. These options are used to describe the standard population's data where:

`popvars(`*casevar*$_p$ *popvar*$_p$`)`: *casevar*$_p$ records the number of cases (deaths) for each stratum in the standard population. *popvar*$_p$ records the total number of individuals in each stratum (individuals at risk).

`rate(`*ratevar*$_p$ *#*|*crudevar*$_p$`)`: *ratevar*$_p$ contains the strata specific rates. *#*|*crudevar*$_p$ is used to specify the crude case rate either via a variable name or optionally by the crude case rate value. If a crude rate variable is used it must be the same for all observations, although, it could be missing for some.

`by(`*groupvars*`)` specifies variables identifying study populations when more than one exist in the data. If this option is not specified the entire study population is treated as one group.

`print` outputs a table summary of the standard population before displaying the study population results.

`format(%`*fmt*`)` specifies the format to use for displaying the summary table. The default is `%10.0g`.

`level(`*#*`)` specifies the confidence level, in percent, for the confidence interval of the adjusted rate and the SMR; see [R] **level**.

## Remarks

Standardization of rates can be performed via the indirect method whenever the stratum-specific rates are either unknown or unreliable. In situations were the stratum specific rates are known, the direct standardization method is preferred. See [R] **dstdize**.

In order to apply the indirect method the following must be available:

1. The observed number, $O$, of cases in each population to be standardized. For example, if death rates in two states are being standardized using the US death rate for the same time period, then you must know the total number of deaths in each state.
2. The distribution, $n_1, \ldots, n_k$, across the various strata for the population being studied. If you are standardizing the death rate in the two states adjusting for age, then you must know the number of individuals in each of the *k* age groups.
3. The strata specific rates, $p_1, \ldots, p_k$, for the standard population. You must have the US death rate for each stratum (age group).
4. The crude rate, $C$, of the standard population. For example, the mortality rate for all of the US for the year.

Then the expected number of cases (deaths), $E$, in each population is obtained by applying the standard population strata-specific rates, $p_1, \ldots, p_k$, to the study population's as

$$E = \sum_{i=1}^{k} n_i p_i$$

The indirect adjusted rate is then

$$R_{\text{indirect}} = C\frac{O}{E}$$

and $O/E$ is the study population's standardized mortality ratio (SMR) if death is the event of interest or the standardized incidence ratio (SIR) for studies of disease incidence.

The exact confidence interval is calculated for each estimated SMR by assuming a Poisson process as described in Breslow and Day (1980, 69–71). These are obtained by first calculating the upper and lower bounds for the confidence interval of the Poisson distributed observed events, $O$, say $L$ and $U$ respectively, and then computing: $\text{SMR}_L = L/E$ and $\text{SMR}_U = U/E$.

## Example

This example is borrowed from Kahn and Sempos (1989, 95–105). We want to compare 1970 mortality rates in California and Maine adjusting for age. Although we have age-specific population counts for the two states, we lack age-specific death rates. In this situation, direct standardization is not feasible. Since we have the US population census data for the same year, we can produce indirectly standardized rates for the two states.

The United States census, the standard population in this example, was entered into Stata and saved in `popkahn.dta`.

```
. use popkahn

. list age pop deaths rate
           age          pop     deaths       rate
  1.       <15     57900000     103062     .00178
  2.     15-24     35441000      45261     .00128
  3.     25-34     24907000      39193     .00157
  4.     35-44     23088000      72617     .00315
  5.     45-54     23220000     169517      .0073
  6.     55-64     18590000     308373     .01659
  7.     65-74     12436000     445531     .03583
  8.       75+      7630000     736758     .09656
```

Note that the standard population contains for each age stratum the total number of individuals (`pop`) and both the age-specific mortality rate (`rate`) and the number of deaths. It is not necessary that the standard population contain all three. If you only have the age-specific mortality rate you can use the `rate(`*ratevar*$_p$ *crudevar*$_p$`)` or `rate(`*ratevar*$_p$ *#*`)` options, where *crudevar*$_p$ refers to the variable containing the total population's crude death rate or # is the total population's crude death rate.

Now let's look at the states data (study population).

```
. use Kahn

. list
          state        age        pop      death
  1. California        <15    5524000     166285
  2. California      15-24    3558000     166285
  3. California      25-34    2677000     166285
  4. California      35-44    2359000     166285
  5. California      45-54    2330000     166285
  6. California      55-64    1704000     166285
  7. California      65-74    1105000     166285
  8. California        75+     696000     166285
  9.      Maine        <15     286000      11051
 10.      Maine      15-24     168000          .
 11.      Maine      25-34     110000          .
 12.      Maine      35-44     109000          .
 13.      Maine      45-54     110000          .
 14.      Maine      55-64      94000          .
 15.      Maine      65-74      69000          .
 16.      Maine        75+      46000          .
```

Note that for each state, the number of individuals in each stratum (age group) is contained in the variable `pop`. The death variable is the total number of deaths observed in the state during the year. It must have the same value for all observations in the group, such as for California, or it could be missing in all but one observation per group, such as in the case of Maine.

In terms of matching these two datasets, it is important that the strata variables have the same name in both datasets and ideally the same levels. If a level is missing from either dataset, the level will not be included in the standardization.

With these data loaded in memory (`use Kahn`), we now execute the command. We will use the `print` option to obtain the standard population's summary table and since we have both the standard population's age specific count and deaths, we will specify the `popvars(`*casevar*$_p$ *popvar*$_p$`)` option. Alternatively, we could specify the `rate(rate 0.00945)` option since we know that 0.00945 is the US crude death rate for 1970.

(*Continued on next page*)

```
. istdize death pop age using popkahn,by(state) pop(deaths pop) print

------Standard Population------
  Stratum              Rate
-------------------------------
      <15              0.00178
    15-24              0.00128
    25-34              0.00157
    35-44              0.00315
    45-54              0.00730
    55-64              0.01659
    65-74              0.03583
      75+              0.09656
-------------------------------
Standard population´s crude rate:       0.00945

----------------------------------------------------------
-> state= California
                   Indirect Standardization
                   Standard
                  Population     Observed      Cases
  Stratum            Rate       Population    Expected
----------------------------------------------------------
      <15           0.0018       5524000      9832.72
    15-24           0.0013       3558000      4543.85
    25-34           0.0016       2677000      4212.46
    35-44           0.0031       2359000      7419.59
    45-54           0.0073       2330000     17010.10
    55-64           0.0166       1704000     28266.14
    65-74           0.0358       1105000     39587.63
      75+           0.0966        696000     67206.23
----------------------------------------------------------
Totals:                         19953000    178078.73

                                                 Observed Cases:   166285
                                                  SMR (Obs/Exp):     0.93
                            SMR exact 95% Conf. Interval: [0.9293, 0.9383]
                                                     Crude Rate:   0.0083
                                                  Adjusted Rate:   0.0088
                                      95% Conf. Interval: [0.0088, 0.0089]

----------------------------------------------------------
-> state= Maine
                   Indirect Standardization
                   Standard
                  Population     Observed      Cases
  Stratum            Rate       Population    Expected
----------------------------------------------------------
      <15           0.0018        286000       509.08
    15-24           0.0013        168000       214.55
    25-34           0.0016        110000       173.09
    35-44           0.0031        109000       342.83
    45-54           0.0073        110000       803.05
    55-64           0.0166         94000      1559.28
    65-74           0.0358         69000      2471.99
      75+           0.0966         46000      4441.79
----------------------------------------------------------
Totals:                           992000     10515.67
```

(*Continued on next page*)

```
                                           Observed Cases:    11051
                                            SMR (Obs/Exp):     1.05
                      SMR exact 95% Conf. Interval: [1.0314, 1.0707]
                                               Crude Rate:   0.0111
                                            Adjusted Rate:   0.0099
                                95% Conf. Interval: [0.0097, 0.0101]

Summary of Study Populations (Rates):
                Cases
    state    Observed     Crude   Adj_Rate  Confidence Interval
 --------------------------------------------------------------------------

California     166285  0.008334   0.008824  [0.008782, 0.008866]
     Maine      11051  0.011140   0.009931  [0.009747, 0.010118]

Summary of Study Populations (SMR):
                Cases        Cases                   Exact
    state    Observed     Expected    SMR     Confidence Interval
 --------------------------------------------------------------------------

California     166285    178078.73   0.934    [0.929290, 0.938271]
     Maine      11051     10515.67   1.051    [1.031405, 1.070687]
```

## Acknowledgment

We gratefully acknowledge the collaboration of Dr. José Maria Pacheco from the Departamento de Epidemiologia, Faculdade de Saúde Pública/USP, Sao Paulo, Brazil and Dr. John L. Moran from The Queen Elizabeth Hospital, Woodville, Australia.

## References

Breslow, N. E. and N. E. Day. 1980. *Statistical Methods in Cancer Research*, Vol II. New York: Oxford University Press.

Kahn, H. A. and C. T. Sempos. 1989. *Statistical Methods in Epidemiology*. New York: Oxford University Press.

## snp14 A two-sample multivariate nonparametric test

Andrew M. Perkins, University of East Anglia, UK, a.perkins@uea.ac.uk

`mv2snp` performs the two-sample multivariate nonparametric test given by Leach (1991). The syntax of `mv2snp` is

`mv2snp` *varlist*`,` `by(`*groupvar*`)` [`verbose`]

where *varlist* consists of at least 2 variable names.

## Options

`by(`*groupvar*`)` is not optional; it specifies the variable identifying the groups.

`verbose` is optional. If included, the vector of test statistics ($U$), variance-covariance matrix ($V$) and inverse of matrix $NV$ will be reported (see below).

## Methods and formulas

Data are ranked as a combined sample (as is carried-out for the univariate Mann–Whitney test) on each variable separately. The test statistic $U^*$ is given as

$$U^* = U'(NV)^{-1}U$$

$U^*$ is asymptotically distributed as $\chi^2_p$ where $p$ is the number of variables. Significant values of $U^*$ indicate a difference between the two groups.

$U$ is the vector of test statistics consisting of

$$U_i = \frac{R_i}{(N+1)} - \frac{n}{2}$$

for each variable, where $R_i$ is the sum of ranks of the smaller group on the $i$th variable, $n$ is the size of the smaller group, and $N$ is the size of the combined sample. $U'$ is the transpose of $U$. The variance-covariance matrix $V$ is defined as having diagonal entries given by

$$v_{ii} = \frac{mn}{12N(N+1)}$$

and off-diagonal entries given by

$$v_{ij} = \frac{mn}{N^2(N-1)(N+1)^2}\left(\sum_{t=1}^{N} R_{it}R_{jt} - \frac{N(N+1)^2}{4}\right)$$

where $m$ and $n$ are the sizes of the two groups, $N$ is the size of the combined sample $(m+n)$ and $\sum R_{it}R_{jt}$ is the sum of the cross-products of each pair of rankings across the combined sample. The matrix $(NV)^{-1}$ is the inverse of $NV$ where $N$ is the combined sample size.

## Example

The following example is that given by Leach (1991). Two psychological questionnaires (GHQ and IES) were given to two groups of policemen following a disaster at a soccer stadium. The first group (H) of 12 police were involved with the disaster; the second group (S) of 11 police were not involved.

The data are entered into Stata as three variables, `ghq`, `ies` and `group`:

```
. list ghq ies group
          ghq       ies      group
  1.       32        38          h
  2.       34        17          h
  3.       49        49          h
  4.       33        45          h
  5.       42        41          h
  6.       49        49          h
  7.       17        29          h
  8.       48        43          h
  9.       37        51          h
 10.       48        48          h
 11.       37        32          h
 12.       22        51          h
 13.       13         6          s
 14.       13        14          s
 15.       11        13          s
 16.       21        36          s
 17.        8         0          s
 18.       24        23          s
 19.       28        19          s
 20.        9         0          s
 21.       24        34          s
 22.       22         7          s
 23.       11         5          s
```

Running the test with the `verbose` option gives

```
. mv2snp ghq ies, by(group) verbose
Multi-variate non-parametric test (Leach, 1991)
Vector of test statistics, U
  -2.39583
  -2.41667
Variance-Covariance matrix, V
  0.01993  0.01561
  0.01561  0.01993
Inverse of matrix NV
   5.64142  -4.41781
  -4.41781   5.64142
-------------------------------------------------------------------------------
Test on variables: ghq ies
                   ghq      - GHQ Results
                   ies      - IES Results
Grouping variable: group    - Grouping Variable
Size of smaller group = 11        Number of variables = 2
Size of larger group  = 12        U* =     14.17
Combined sample size  = 23        P =   0.00084
-------------------------------------------------------------------------------
```

Omitting the `verbose` option excludes output of vector $U$ and matrices $V$ and $(NV)^{-1}$.

## Saved results

`mv2snp` saves the following results:

| | |
|---|---|
| `S_1` | $U^*$ |
| `S_2` | Combined sample size |
| `S_3` | Size of smaller sample |
| `S_4` | Number of variables |

## Remarks

The grouping variable may be alpha or numeric and need not be named `group`. The data do not need to be sorted before applying `mv2snp`. When ranking the data, `mv2snp` does not exclude data values of zero. `mv2snp` does, however, perform casewise deletion for subjects with missing data because the nature of a multivariate test examines the correlation between variables; thus including data against which no correlation can be made will lead to erroneous results.

## Reference

Leach, C. 1991. Nonparametric methods for complex data sets. In *New Developments in Statistics for Psychology and the Social Sciences*, Vol. 2. ed. P. Lovie and A. D. Lovie. British Psychological Society and Routledge.

## STB categories and insert codes

Inserts in the STB are presently categorized as follows:

| | | | |
|---|---|---|---|
| *General Categories:* | | | |
| *an* | announcements | *ip* | instruction on programming |
| *cc* | communications & letters | *os* | operating system, hardware, & interprogram communication |
| *dm* | data management | | |
| *dt* | datasets | *qs* | questions and suggestions |
| *gr* | graphics | *tt* | teaching |
| *in* | instruction | *zz* | not elsewhere classified |
| *Statistical Categories:* | | | |
| *sbe* | biostatistics & epidemiology | *ssa* | survival analysis |
| *sed* | exploratory data analysis | *ssi* | simulation & random numbers |
| *sg* | general statistics | *sss* | social science & psychometrics |
| *smv* | multivariate analysis | *sts* | time-series, econometrics |
| *snp* | nonparametric methods | *svy* | survey sampling |
| *sqc* | quality control | *sxd* | experimental design |
| *sqv* | analysis of qualitative variables | *szz* | not elsewhere classified |
| *srd* | robust methods & statistical diagnostics | | |

In addition, we have granted one other prefix, *stata*, to the manufacturers of Stata for their exclusive use.

## Guidelines for authors

The Stata Technical Bulletin (STB) is a journal that is intended to provide a forum for Stata users of all disciplines and levels of sophistication. The STB contains articles written by StataCorp, Stata users, and others.

Articles include new Stata commands (ado-files), programming tutorials, illustrations of data analysis techniques, discussions on teaching statistics, debates on appropriate statistical techniques, reports on other programs, and interesting datasets, announcements, questions, and suggestions.

A submission to the STB consists of

1. An insert (article) describing the purpose of the submission. The STB is produced using plain TeX so submissions using TeX (or LaTeX) are the easiest for the editor to handle, but any word processor is appropriate. If you are not using TeX and your insert contains a significant amount of mathematics, please FAX (409–845–3144) a copy of the insert so we can see the intended appearance of the text.

2. Any ado-files, `.exe` files, or other software that accompanies the submission.

3. A help file for each ado-file included in the submission. See any recent STB diskette for the structure a help file. If you have questions, fill in as much of the information as possible and we will take care of the details.

4. A do-file that replicates the examples in your text. Also include the datasets used in the example. This allows us to verify that the software works as described and allows users to replicate the examples as a way of learning how to use the software.

5. Files containing the graphs to be included in the insert. If you have used STAGE to edit the graphs in your submission, be sure to include the `.gph` files. Do not add titles (e.g., "Figure 1: ...") to your graphs as we will have to strip them off.

The easiest way to submit an insert to the STB is to first create a single "archive file" (either a `.zip` file or a compressed `.tar` file) containing all of the files associated with the submission, and then email it to the editor at `stb@stata.com` either by first using `uuencode` if you are working on a Unix platform or by attaching it to an email message if your mailer allows the sending of attachments. In Unix, for example, to email the current directory and all of its subdirectories:

```
tar -cf - . | compress | uuencode xyzz.tar.Z > whatever
mail stb@stata.com < whatever
```

## International Stata Distributors

International Stata users may also order subscriptions to the *Stata Technical Bulletin* from our International Stata Distributors.

| | |
|---|---|
| Company: | Applied Statistics &<br>Systems Consultants |
| Address: | P.O. Box 1169<br>Nazerath-Ellit 17100<br>Israel |
| Phone: | +972 66-554254 |
| Fax: | +972 66-554254 |
| Email: | sasconsl@actcom.co.il |
| Countries served: | Israel |

| | |
|---|---|
| Company: | Dittrich & Partner Consulting |
| Address: | Prinzenstrasse 2<br>D-42697 Solingen<br>Germany |
| Phone: | +49 212-3390 200 |
| Fax: | +49 212-3390 295 |
| Email: | evhall@dpc.de |
| URL: | http://www.dpc.de |
| Countries served: | Germany, Austria, Italy |

| | |
|---|---|
| Company: | IEM |
| Address: | P.O. Box 2222<br>PRIMROSE 1416<br>South Africa |
| Phone: | 27 11 828-6169 |
| Fax: | 27 11 822-1377 |
| Email: | iem@hot.co.za |
| Countries served: | South Africa, Botswana, Lesotho,<br>Namibia, Mozambique, Swaziland,<br>Zimbabwe |

| | |
|---|---|
| Company: | MercoStat Consultores |
| Address: | 9 de junio 1389<br>CP 11400 MONTEVIDEO<br>Uruguay |
| Phone: | 598-2-613-7905 |
| Fax: | +Same |
| Email: | andres.gil@usa.net |
| Countries served: | Uruguay, Argentina, Brazil<br>Paraguay |

| | |
|---|---|
| Company: | Metrika Consulting |
| Address: | Mosstorpsvagen 48<br>183 30 Taby Stockholm<br>Sweden |
| Phone: | +46-708-163128 |
| Fax: | +46-8-7924747 |
| Email: | sales@metrika.se |
| Countries served: | Sweden, Baltic States, Denmark<br>Finland, Iceland, Norway |

| | |
|---|---|
| Company: | Ritme Informatique |
| Address: | 34 Boulevard Haussmann<br>75009 Paris<br>France |
| Phone: | +33 1 42 46 00 42 |
| Fax: | +33 1 42 46 00 33 |
| Email: | info@ritme.com |
| URL: | http://www.ritme.com |
| Countries served: | France, Belgium<br>Luxembourg, Switzerland |

| | |
|---|---|
| Company: | Smit Consult |
| Address: | Doormanstraat 19<br>5151 GM Drunen<br>Netherlands |
| Phone: | +31-416-378 125 |
| | +31-416-378 385 |
| Email: | j.a.c.m.smit@smitcon.nl |
| URL: | http://www.smitconsult.nl |
| Countries served: | Netherlands |

(*List continued on next page*)

## International Stata Distributors

(*Continued from previous page*)

Company: Survey Design & Analysis
Services P/L
Address: 249 Eramosa Road West
Moorooduc VIC 3933
Australia
Phone: +61 3 5978 8329
Fax: +61 3 5978 8623
Email: sales@survey-design.com.au
URL: http://survey-design.com.au
Countries served: Australia, New Zealand

Company: Timberlake Consultants
Address: 47 Hartfield Crescent
West Wickham
Kent BR4 9DW
United Kingdom
Phone: +44 181 462 0495
Fax: +44 181 462 0493
Email: info@timberlake.co.uk
URL: http://www.timberlake.co.uk
Countries served: United Kingdom, Eire

Company: Timberlake Consulting S.L.
Address: Calle Montecarmelo n°36 Bajo
41011 Seville
Spain
Phone: +34.5.428.40.94
Fax: +34.5.428.40.94
Email: timberlake@zoom.es
Countries served: Spain

Company: Timberlake Consultores
Address: Praceta do Comércio, 13 - 9° Dto.
Quinta Grande
2720 Alfragide
Portugal
Phone: +351 (01) 4719337
Telemóvel: 0931 62 7255
Email: timberlake.co@mail.telepac.pt
Countries served: Portugal

Company: Unidost A.S.
Address: Rihtim Cad. Polat Han D:38
Kadikoy
81320 ISTANBUL
Turkey
Phone: +90-(216)-4141958
Fax: +90-(216)-3368923
Email: info@unidost.com
URL: http://www.turk.net/mhendekli/unidost.htm
Countries served: Turkey