A publication to promote communication among Stata users

Editor
H. Joseph Newton
Department of Statistics
Texas A & M University
College Station, Texas 77843
979-845-3142
979-845-3144 FAX
stb@stata.com EMAIL

Associate Editors
Nicholas J. Cox, University of Durham
Joanne M. Garrett, University of North Carolina
Marcello Pagano, Harvard School of Public Health
J. Patrick Royston, UK Medical Research Council
Jeroen Weesie, Utrecht University

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

| an72 | STB-49–STB-54 available in bound format | 2 |
| ip91 | Update of the byvar command | 2 |
| bse32 | Errata for bse32 | 2 |
| bse33 | Comparing several methods of measuring the same quantity | 2 |
| bse34 | Loglinear modeling using iterative proportional fitting | 10 |
| sg135 | Test for autoregressive conditional heteroskedasticity in regression error distribution | 13 |
| sg136 | Tests for serial correlation in regression error distribution | 14 |
| sg137 | Tests for heteroskedasticity in regression error distribution | 15 |
| sg138 | Bootstrap inferences about measures of correlation | 17 |
| sg139 | Logistic regression when binary outcome is measured with uncertainty | 20 |
| sg140 | The Gumbel quantile plot and a test for choice of extreme models | 23 |
| sg141 | Treatment effects model | 25 |
| sg142 | Uniform layer effect models for the analysis of differences in two-way associations | 33 |
| snps | somersd—Confidence intervals for nonparametric statistics and their differences | 47 |
| zz10 | Cumulative index for STB-49–STB-54 | 55 |
The ninth year of the *Stata Technical Bulletin* (issues 49–54) has been reprinted in a bound book called *The Stata Technical Bulletin Reprints, Volume 9*. The volume of reprints is available from StataCorp for $25, plus shipping. Authors of inserts in STB-49–STB-54 will automatically receive the book at no charge and need not order.

This book of reprints includes everything that appeared in issues 49–54 of the STB. As a consequence, you do not need to purchase the reprints if you saved your STBs. However, many subscribers find the reprints useful since they are bound in a convenient volume. Our primary reason for reprinting the STB, though, is to make it easier and cheaper for new users to obtain back issues. For those not purchasing the Reprints, note that ziz10 in this issue provides a cumulative index for the ninth year of the original STBs.

You may order the Reprints Volumes online at www.stata.com/bookstore/stbr.html or use the enclosed order form.

**Update of the byvar command**

Patrick Royston, Imperial College School of Medicine, London, UK, p.royston@ic.ac.uk

**Abstract:** The *byvar* command has been updated for Stata 6 and a few new features added.

**Keywords:** Stata commands.

The *byvar* command introduced in Royston (1995) has been updated for Stata 6 and a few new features added.

**References**


**Errata for sbe32**

López Vizcaíno, M. E., Santiago Pérez M. I., Abraira García L., Dirección Xeral de Saúde Publica, Spain, dxsp3@jet.es

**Abstract:** Errors in the Methodology section of López Vizcaíno et al. (2000) are corrected.

**Keywords:** Outbreak, regression, threshold, public health surveillance.

In the process of editing López Vizcaíno et al. (2000), errors were introduced into the Methodology section. In the first two equations in that section the $g_i$’s should have been $\gamma_i$’s, while the $m_i$ should have been $\mu_i$. Finally, the sentence that begins after the third displayed equation should say that the Poisson model corresponds to $\phi = 1$ rather that $\phi = 0$.

**References**


**Comparing several methods of measuring the same quantity**

Paul Seed, GKT School of Medicine, King’s London, UK, paul.seed@kcl.ac.uk

**Abstract:** New commands are given, based on the Bland–Altman approach to the analysis of studies comparing two or more methods for measuring the same quantity. An extension to more than two methods is explained, with an associated command. A new command, based on Pitman’s method, gives confidence intervals for the variance ratio of paired data. It is more powerful than Stata’s *sdtest*, particularly for large correlations. For more than two methods, with no reference standard, a new generalization of Bland–Altman methods is shown and compared with an approach based on factor analysis.

**Keywords:** Method comparison, Bland–Altman, variance ratio.

**The problem**

New techniques for taking clinical measurements are always being developed. How can we decide which is best? Sometimes a new measurement technique is compared with an established “Gold Standard,” which may or may not be regarded as exact. How good is the new technique? Alternatively, there may be several methods, all seen as imperfect. Which is best?
Some typical datasets

Consider blood iron where one might want to compare an established method (colorimetry) with two new clinical techniques: inductively coupled plasma optical emission spectrometry (ICPOES) with 18 pairs of measurements:

```
. use col_icp
. summarize
```

<table>
<thead>
<tr>
<th>Variable</th>
<th>Obs</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>colorime</td>
<td>20</td>
<td>19.3</td>
<td>4.878524</td>
<td>11</td>
<td>26</td>
</tr>
<tr>
<td>icpoes</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>icp</td>
<td>18</td>
<td>31.6667</td>
<td>10.07034</td>
<td>16</td>
<td>58</td>
</tr>
<tr>
<td>mean</td>
<td>18</td>
<td>25.4444</td>
<td>6.187249</td>
<td>13.5</td>
<td>40</td>
</tr>
<tr>
<td>diff</td>
<td>18</td>
<td>-12.4444</td>
<td>10.26257</td>
<td>-36</td>
<td>-5</td>
</tr>
</tbody>
</table>

and ICPOES following protein precipitation with trichloroacetic acid (TCA) with 52 pairs:

```
. use tca_col, clear
. summarize
```

<table>
<thead>
<tr>
<th>Variable</th>
<th>Obs</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>colorime</td>
<td>52</td>
<td>15.09615</td>
<td>5.665141</td>
<td>8</td>
<td>26</td>
</tr>
<tr>
<td>tca_ppt_</td>
<td>52</td>
<td>13.2154</td>
<td>5.413623</td>
<td>5</td>
<td>23</td>
</tr>
<tr>
<td>mean</td>
<td>52</td>
<td>14.15385</td>
<td>5.406251</td>
<td>6.5</td>
<td>24.5</td>
</tr>
<tr>
<td>diff</td>
<td>52</td>
<td>1.884615</td>
<td>1.395425</td>
<td>-1</td>
<td>5</td>
</tr>
</tbody>
</table>

The question is how does the new test compare with the old?

The second example compares five eyesight tests carried out on 15 patients before and after operations for astigmatism. We are interested in the percentage improvement in eyesight as measured by each of the five tests:

```
. use tan_part, clear
. summarize pct_*
```

<table>
<thead>
<tr>
<th>Variable</th>
<th>Obs</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>pct_1</td>
<td>15</td>
<td>20.19054</td>
<td>18.39144</td>
<td>0</td>
<td>60</td>
</tr>
<tr>
<td>pct_2</td>
<td>15</td>
<td>33.52114</td>
<td>35.62064</td>
<td>.4051345</td>
<td>114.679</td>
</tr>
<tr>
<td>pct_3</td>
<td>15</td>
<td>36.47013</td>
<td>57.3427</td>
<td>-13.49716</td>
<td>220.439</td>
</tr>
<tr>
<td>pct_4</td>
<td>15</td>
<td>15.95635</td>
<td>11.4139</td>
<td>.7299263</td>
<td>41.23711</td>
</tr>
<tr>
<td>pct_5</td>
<td>15</td>
<td>24.67196</td>
<td>11.49144</td>
<td>9.966988</td>
<td>42.85716</td>
</tr>
</tbody>
</table>

One last example is tumor activity adjusted for partial volume and glucose uptake (the variable \( \log_{gp} \)) and that adjusted for partial volume alone (the variable \( \log_p \)):

```
. use suv2, clear
. summarize log_*
```

<table>
<thead>
<tr>
<th>Variable</th>
<th>Obs</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>log_p</td>
<td>86</td>
<td>2.260328</td>
<td>.636188</td>
<td>.2615947</td>
<td>4.099669</td>
</tr>
<tr>
<td>log_gp</td>
<td>86</td>
<td>2.183237</td>
<td>.7080229</td>
<td>-.8204135</td>
<td>3.929667</td>
</tr>
</tbody>
</table>

Simple methods that don’t work

I have found two methods in particular that don’t perform very well in such problems. First, very high correlations are almost always found. The null hypothesis that there is no association is just not credible. The significance test tells us nothing we don’t already know. Secondly, the \( F \) test, provided by Stata’s `sdtest` command, is not appropriate for paired data. Pitman’s test (below) is more powerful, particularly with a large correlation. It is not safe to assume that the measure with the smaller variance has the smaller component of error. While this is often the case, it might just be less sensitive to genuine variation.

Other methods to use with caution

Linear regression looks for any linear relationship: \( m_2 = a + bm_1 \), whereas we are often interested in \( m_2 = m_1 \), that is, \( a = 0 \) and \( b = 1 \). The method also assumes that \( m_1 \) is measured without error. This is scarcely likely. If \( m_1 \) is measured with error, the estimate of \( b \) is biased towards zero.

Paired \( t \) tests and confidence intervals for differences in means are useful as evidence of systematic bias, but measures with large random error can have a nonsignificant \( t \) test, even when bias exists. We are mainly interested in the error in each individual measurement. Bias is not as important as the absolute size of the likely difference.
Simple approaches that can be useful

The reference range for differences between individual measurements is defined as the mean plus or minus two standard deviations. Approximately 95% of values will be between these limits. If two measures agree well, the reference range will be very narrow. Note that the reference range is in the same units as the actual measurement. In Stata we can use

```
. gen diff = m2 - m1
. summarize diff
. local mdiff = r(mean)
. local lrr = "m2" - 2*r(Var)^.5
. local urr = "m2" + 2*r(Var)^.5
. display "Mean difference = " mdiff
. display "Reference Range = " lrr " to " urr
```

Bland–Altman plots

Bland and Altman (1983) introduced the idea of plotting the difference of paired variables versus their average, with horizontal lines for the reference range for the difference. Any plots of the actual data are useful to show oddities. The plots will show no trend if the variance of \( m_1 \) and \( m_2 \) are the same. A positive trend shows the variance of \( e_2 \) is larger than that of \( e_1 \). We can use

```
. gen av = (m1 + m2)/2
. graph diff av, xlab ylab yline(mdiff, lrr, urr)
```

or use the command `baplot` included with this insert

```
. baplot m2 m1
```

Syntax for baplot command

```
baplot varname1 varname2 [if exp] [in range] [ , format(str) avlab(str) difflab(str) graph_options ]
```

Options for baplot

- `format(str)` sets the format for the results given.
- `avlab(str)` gives a variable label to the average before plotting the graph.
- `difflab(str)` gives a variable label to the difference before plotting the graph.
- `graph_options` are any of the options allowed with `graph, twoway`; see [G] graph options.

Examples

Consider comparing a new measure with a gold standard. For the blood-iron data, we can compare ICPOES with Colorimetry giving the output below and the plot in Figure 1.

```
. use col_icp, clear
. summarize icp colorime
```

```
Variable |
---------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|
         | Obs | Mean | Std. Dev. | Min | Max | 16 | 56|
---------|-----|------|-----------|-----|-----|----|----|
icp | 18  | 31.66667 | 10.07034 | 16 | 56 |    |    |
colorime | 18 | 19.22222 | 5.105425 | 11 | 26 |    |    |
---------|-----|------|-----------|-----|-----|----|----|
. baplot icp colorime, xlab(0,.10,.20,.30,.40) ylab(-.40,-.20,0,.20,.40) avlab("ICPOES vs Colorimetry")
```

Bland–Altman comparison of icp and colorime

Limits of agreement (Reference Range for difference): -8.081 to 32.970

Mean difference: 12.444 (CI 7.341 to 17.548)

Range: 13.500 to 40.000

Pitman's Test of difference in variance: r = 0.600, n = 18, p = 0.008

(Figure 1 on next page)
**Pitman's test of difference in variance**

As mentioned earlier, if $m_1$ and $m_2$ have equal variance, the covariance (and hence the correlation) between their average and their difference will be zero. Pitman’s test looks for a significant correlation between the difference and the average of $m_1$ and $m_2$. If one exists, this is evidence that the variances are not the same. Because this uses the fact that the data are paired, it can be much more powerful than the usual $F$ test (consider paired and unpaired $t$ tests).

Pitman, quoted in Snedecor and Cochran (1967), extended this test to give a confidence interval for the variance ratio. This can be obtained by using the new command `sdpair` included with this insert.
Syntax for the sdpair command

```
sdpair varname1 varname2 [weight] [if exp] [in range] [, format(str) level(#) ]
```

Weights and aweights are allowed.

Options for sdpair

- **format(str)** sets the format for the display of results.
- **level(#)** specifies the confidence level, in percent, for confidence intervals. The default is `level(95)` or as set by `set level`.

Example of Pitman's test:

Compare tumor activity adjusted for partial volume and glucose uptake with that for partial volume alone:

```
. sdtest log_p = log_gp
(output omitted)
P < F_obs = 0.1627  P < F_L + P > F_U = 0.3254  P > F_obs = 0.8373
. sdpair log_p log_gp
```

Pitman's variance ratio test between log_p and log_gp:

- Ratio of Standard deviations = 0.8964
- 95% Confidence Interval 0.8365 to 0.9669
- t = -2.966, df = 84, p = 0.004

Multiple Bland–Altman plots for comparing more than two methods

The command `bamat` produces a matrix of Bland–Altman plots for all possible pairs of methods. This is very useful for a first comparison of methods, and may identify a method that is clearly inferior to the others. It is illustrated with the eyesight data.

Syntax for bamat

```
bamat varlist [if exp] [in range], format(str) notable data avlab(str) difflab(str) obs(#)
```

Options for bamat

- **format(str)** sets the format for display of results.
- **notable** suppresses display of results.
- **data** lists data used in plotting each graph.
- **avlab(str)** gives a variable label to the average before plotting the graph.
- **difflab(str)** gives a variable label to the difference before plotting the graph.
- **obs(#)** specifies the minimum number of nonmissing values per observation needed for a point to be plotted. The default value is 2 (pairwise deletion).
- **listwise** specifies listwise deletion of missing data. Default is pairwise. Only observations with no missing values are used.
- **title(str)** adds a single title to the block of graphs.
- **graph_options** are any of the options allowed with `graph, twoway`; see [G] graph options.

Example of bamat

Once again we consider the eyesight data.

```
. use tan_part, clear
. bamat pct_/*
```

Reference ranges for differences between two methods

<table>
<thead>
<tr>
<th>Method 1</th>
<th>Method 2</th>
<th>Mean</th>
<th>[95% Reference Range]</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>pct_2</td>
<td>pct_1</td>
<td>13.331</td>
<td>-50.777</td>
<td>-31.678</td>
<td>88.525</td>
</tr>
<tr>
<td>pct_3</td>
<td>pct_1</td>
<td>16.290</td>
<td>-100.401</td>
<td>-42.994</td>
<td>195.588</td>
</tr>
<tr>
<td>pct_3</td>
<td>pct_2</td>
<td>2.949</td>
<td>-100.526</td>
<td>-44.137</td>
<td>162.944</td>
</tr>
<tr>
<td>pct_4</td>
<td>pct_1</td>
<td>-4.234</td>
<td>-47.425</td>
<td>-46.533</td>
<td>41.237</td>
</tr>
</tbody>
</table>
Modified Bland–Altman plots

We would like to modify Bland–Altman plots for use with more than two measures when there is no gold standard measure. For example, if we have eight measures, there would be 28 Bland–Altman plots. We consider a modification that gives one comparison per measure. The average is just the average of all the measures. We hope this is close to the truth. The difference we use for the ith measure is the average of the ith measure minus the average of the other measures. We work out a reference range as before.

If each measure is of the form \( m_i = t + e_i \), with the errors independent and of equal variance, then the correlation between the average and the difference will be zero. If for some particularly useful method, \( m_i \) has smaller than average variance, there will be a negative trend.

This method has difficulties if the errors are correlated or the model breaks down in other ways; for example, if \( m_i \) is a linear function of the truth, that is, \( m_i = a_i + b_i t + e_i \).

We can do this by brute force in Stata by

\[
\begin{align*}
\text{. egen av = rmean(m1-m5)} \\
\text{. egen mean1 = rmean(m2-m5)} \\
\text{. gen diff = m1 - mean1} \\
\text{. summ diff} \\
\text{. local mdiff = \_r(mean)} \\
\text{. local lrr = \_diff - 2*\_r(Var)}^{.5} \\
\text{. local urr = \_diff + 2*\_r(Var)}^{.5} \\
\text{. graph diff av, xlab ylab yline(lrr, mdiff, urr)}
\end{align*}
\]

or use the new command `bagroup` included with this insert.

Syntax for bagroup

```
bagroup varlist [if exp] [in range] [ , format(str) rows(#) avlab(str) difflab(str) 
   title(str) obs(#) listwise graph_options ]
```

Options for bagroup

- `format(str)` sets the format for display of results.
- `rows(#)` specifies the number of rows of graphs to be shown.
- `avlab(str)` gives a variable label to the average before plotting the graph.
- `difflab(str)` gives a variable label to the difference before plotting the graph.
**title(str)** adds a single title to the block of graphs.

**obs(#)** specifies the minimum number of nonmissing values per observation needed for a point to be plotted. The default value is 2 (pairwise deletion).

**listwise** specifies listwise deletion of missing data. Default is pairwise. Only observations with no missing values are used.

**graph_options** are any of the options allowed with **graph, twoway**; see [G] **graph options**.

### Example of bagroup

For the eyesight data we obtain the results below and the plot in Figure 4.

```stata
use tan_part, clear
summarize pct_*, obs

bgroup pct_*, by
describe pct_*, obs
describe pct_*, by
```

### Factor analysis

Principal component factor analysis finds linear combinations of the variables. The first accounts for the largest possible proportion of the total variation. Later factors account for as much as possible of what is left. Correlations, not covariances are used. Effectively, each variable is standardized to have mean zero and variance one. This gives each the same importance in determining the factors.

In a factor analysis, the first factor should be a good measure of the truth. If some methods are measuring the wrong thing, their errors will be correlated. This confounder will tend to appear in secondary, orthogonal factors not in the main measure. Correlations of each measure with the principal factor are a useful measure of which is most predictive. Significance tests are not available.

Because the variables are first standardized, factor analysis is not affected by calibration problems of the form $m_i = a_i + b_i t + e_i$. If there is a standard scale (as with the blood iron), this may be a problem. If not (as with the eyesight data), it may be a bonus.

As an example, consider the eyesight data.
. factor pct_/*
  (principal factors; 2 factors retained)

<table>
<thead>
<tr>
<th>Factor</th>
<th>Eigenvalue</th>
<th>Difference</th>
<th>Proportion</th>
<th>Cumulative</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2.34432</td>
<td></td>
<td>0.9572</td>
<td>1.0000</td>
</tr>
<tr>
<td>2</td>
<td>0.42554</td>
<td>1.81878</td>
<td>0.9572</td>
<td>1.0000</td>
</tr>
<tr>
<td>3</td>
<td>-0.06390</td>
<td>0.00903</td>
<td>-0.0677</td>
<td>1.1466</td>
</tr>
<tr>
<td>4</td>
<td>-0.15012</td>
<td>0.03304</td>
<td>-0.0660</td>
<td>1.0806</td>
</tr>
<tr>
<td>5</td>
<td>-0.18316</td>
<td></td>
<td>-0.0806</td>
<td>1.0000</td>
</tr>
</tbody>
</table>

Factor Loadings

<table>
<thead>
<tr>
<th>Variable</th>
<th>1</th>
<th>2</th>
<th>Uniqueness</th>
</tr>
</thead>
<tbody>
<tr>
<td>pct_1</td>
<td>0.33081</td>
<td>0.38083</td>
<td>0.72166</td>
</tr>
<tr>
<td>pct_2</td>
<td>0.81906</td>
<td>0.25653</td>
<td>0.2832</td>
</tr>
<tr>
<td>pct_3</td>
<td>0.65937</td>
<td>-0.26085</td>
<td>0.49268</td>
</tr>
<tr>
<td>pct_4</td>
<td>0.52325</td>
<td>-0.36159</td>
<td>0.59546</td>
</tr>
<tr>
<td>pct_5</td>
<td>0.86169</td>
<td>0.02151</td>
<td>0.25702</td>
</tr>
</tbody>
</table>

. score pct_fac
  (based on unrotated factors)
  (1 scoring not used)

Scoring Coefficients

<table>
<thead>
<tr>
<th>Variable</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>pct_1</td>
<td>0.93599</td>
</tr>
<tr>
<td>pct_2</td>
<td>0.30460</td>
</tr>
<tr>
<td>pct_3</td>
<td>0.18434</td>
</tr>
<tr>
<td>pct_4</td>
<td>0.12185</td>
</tr>
<tr>
<td>pct_5</td>
<td>0.41533</td>
</tr>
</tbody>
</table>

. corr pct_/*
  (obs=15)

<table>
<thead>
<tr>
<th></th>
<th>pct_1</th>
<th>pct_2</th>
<th>pct_3</th>
<th>pct_4</th>
<th>pct_5</th>
<th>pct_fac</th>
</tr>
</thead>
<tbody>
<tr>
<td>pct_1</td>
<td>1.0000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>pct_2</td>
<td>0.4424</td>
<td>1.0000</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>pct_3</td>
<td>0.1321</td>
<td>0.4704</td>
<td>1.0000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>pct_4</td>
<td>0.0077</td>
<td>0.3370</td>
<td>0.5163</td>
<td>1.0000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>pct_5</td>
<td>0.2959</td>
<td>0.7727</td>
<td>0.5814</td>
<td>0.4527</td>
<td>1.0000</td>
<td></td>
</tr>
<tr>
<td>pct_fac</td>
<td>0.3803</td>
<td>0.6905</td>
<td>0.7169</td>
<td>0.9689</td>
<td>0.9369</td>
<td>1.0000</td>
</tr>
</tbody>
</table>

Modeling approaches

If we use the model $m_i = f_i + b_i t + e_i$, there are several possibilities, depending on the data. With repeated measures, we could use errors-in-variables regression (Strike 1991, 1996). With data from more than two methods of measurement, either restricted factor analysis (Dunn 1989) or multilevel modeling (Goldstein 1995) are possible. None of these are yet available in Stata.

Conclusions

Bland–Altman plots are a simple, effective way of comparing two methods of measuring the same quantity. More obvious methods, such as $t$ tests, correlation, and regression can be seriously misleading.

The Stata command `sdtest` is not appropriate for comparisons of variances with paired data, while the new command `sdpair`, based on Pitman’s method, is more powerful, and gives confidence intervals for the variance ratio.

Bland–Altman plots can be generalized to handle more than two methods, while factor analysis allows comparison of each measure with a good estimate of the truth and is not affected by calibration problems.

References

sbe34  Loglinear modeling using iterative proportional fitting

Adrian Mander, MRC Biostatistics Unit, Cambridge, UK, adrian.mander@mrc-bsu.cam.ac.uk

Abstract: Iterative proportional fitting is a procedure that calculates the expected frequencies within a contingency table. The algorithm converges to maximum likelihood estimates even when the likelihood is badly behaved and is extremely fast when the contingency table has a large number of cells.

Keywords: Loglinear modeling, contingency tables, constrained estimation.

Syntax

```
ipf [varlist] [weight], fit(string) confile(filename) convars(varlist) save(filename)

expect constr(string) nolog
```

Description

The iterative proportional fitting (IPF) algorithm is a simple method to calculate the expected counts of a hierarchical loglinear model. The algorithm’s rate of convergence is first order. The more commonly used Newton–Raphson algorithm is second order. However, each iteration of the IPF algorithm is quicker because Newton–Raphson inverts matrices. This makes the IPF algorithm much quicker for contingency tables with numerous cells.

The IPF algorithm has the following steps:

1. Initial estimates of the expected frequencies are given. The initial estimates should have associations and interactions that are less complex than the model being fitted. By default the initial frequencies are 1.
2. The estimates of the expected frequencies are successively adjusted by scaling factors so they match each marginal table.
3. The scaling continues until the log likelihood converges.

The algorithm always converges to the correct expected frequencies even when the likelihood is poorly behaved, for example, when there are zero fitted counts.

The `varlist` defines the dimension of the contingency table that the Poisson likelihood is calculated over. If the `varlist` is not specified, the variables in the `fit` option define the dimensions of the contingency table.

Options

- `fit(string)` specifies the loglinear model. It requires special syntax of the form `var1*var2+var3+var4`. The term `var1*var2` includes all the interactions between the two variables and also the main effects of `var1` and `var2`. The main effects `var3` and `var4` are also included in the model but no interactions. This syntax is used in most books on loglinear modeling.
- `confile(filename)` specifies a `.dta` file that contains initial values for the expected counts, the variable containing the frequencies must be called `Efreqold`. Any missing values in this file will be replaced by 1. This option requires the use of the option `convars`.
- `convars(varlist)` specifies the variables in the file specified by `confile`, excluding `Efreqold`. This `varlist` may be a subset of the variables in the model. All cells not specified with an initial expected frequency will have initial value of 1.
- `save(filename)` specifies the expected frequencies, observed frequencies and estimated probabilities for every cell to be saved in a `.dta` file.
- `expect` specifies that the expected frequencies are displayed.
- `constr(string)` specifies initial values for the expected frequencies. The syntax requires a condition in square brackets followed by a value for the expected frequency. Hence `[sex="male"]2` replaces all initial values for males to be 2.
- `nolog` specifies whether the log likelihood is displayed at each iteration.

Examples

To illustrate the command, data has been taken from Agresti (1990, 308).

```
call use fish
.describe
```
Contains data from fish. dta

Contains data from fish. dta
obs: 56
vars: 5
size: 1,344 (99.8% of memory free)

1. lake float %9.0g 1
2. gender float %9.0g g
3. size float %9.0g
4. food float %9.0g f
5. freq float %9.0g

Sorted by:

and we reconstruct the table on page 309 of Agresti (1990) via the IPF algorithm:

Table 1. Goodness of fit of models

<table>
<thead>
<tr>
<th>Model</th>
<th>$G^2$</th>
<th>$X^2$</th>
<th>df</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) food + lake * size * gender</td>
<td>116.76114</td>
<td>106.49216</td>
<td>60</td>
</tr>
<tr>
<td>(2) food * gender + lake * size * gender</td>
<td>114.65707</td>
<td>101.24765</td>
<td>56</td>
</tr>
<tr>
<td>(3) food * size + lake * size * gender</td>
<td>101.61156</td>
<td>86.887138</td>
<td>56</td>
</tr>
<tr>
<td>(4) food * lake + lake * size * gender</td>
<td>73.565895</td>
<td>79.579025</td>
<td>56</td>
</tr>
<tr>
<td>(5) food * lake + food * size + lake * size * gender</td>
<td>52.478477</td>
<td>58.016632</td>
<td>56</td>
</tr>
<tr>
<td>(6) food * lake + food * size + food * gender + lake * size * gender</td>
<td>50.263695</td>
<td>52.566868</td>
<td>56</td>
</tr>
</tbody>
</table>

In Table 2, we collapse the information in Table 1 over gender:

Table 2. Goodness of fit of models for a table collapsed over gender

<table>
<thead>
<tr>
<th>Model</th>
<th>$G^2$</th>
<th>$X^2$</th>
<th>df</th>
</tr>
</thead>
<tbody>
<tr>
<td>(7) food + lake * size</td>
<td>81.36248</td>
<td>73.059517</td>
<td>28</td>
</tr>
<tr>
<td>(8) food * size + lake * size</td>
<td>66.212906</td>
<td>54.29039</td>
<td>24</td>
</tr>
<tr>
<td>(9) food * lake + lake * size</td>
<td>38.167236</td>
<td>32.742958</td>
<td>16</td>
</tr>
<tr>
<td>(10) food * lake + food * size + lake * size</td>
<td>17.079826</td>
<td>15.043343</td>
<td>12</td>
</tr>
</tbody>
</table>

The study is about the factors that influence the primary food choice of alligators. The response variable is the food and the choices are subclassified by size of alligator, gender of alligator, and one of four lakes the alligators are sampled from. There were 219 alligators distributed over 80 possible cells. As the data are sparse, the likelihood-ratio test ($G^2$) and the Pearson $X^2$ test are not reliable, but comparison of the models can be made using $G^2$. Let $F = food$, $L = lake$, $G = gender$, and $S = size$, and the following shorthand $G^2[(F, LGS) | (FG, LGS)] = 2.1$ and $G^2[(FS, FL, LGS) | (FG, FL, FS, LGS)] = 2.2$ is used to compare models (1) and (2) and models (5) and (6), respectively. Both tests are based on 4 degrees of freedom, suggesting that the table should be collapsed over gender. From the collapsed table, it is clear that both lake and size have effects on the food choice of the alligator.

Constrained estimation

Constrained estimation can be implemented by selecting appropriate models and initial expected frequencies. This will be illustrated using a case–control study. Let the variables $E$ and $D$ be exposure and disease (both variables are binary, exposed cases are defined by $D = 1$ and $E = 1$, respectively). The command that fits a model of independence of disease and exposure is

```
.ipf [fw=freq], fit(E+D) exp
```

This model constrains the odds ratio to be 1. To constrain the odds ratio to equal 2 requires the initial expected frequency in either the cell (0,0) or the cell (1,1) for (D,E) to equal 2. The simplest way to alter one cell’s initial expected frequency is by using the constr option.

```
.ipf [fw=freq], fit(D + E) constr([D==0 & E==0] 2) exp
```
An alternative method uses `convars` and `confile`. First, create a file of initial values for table and save this file as `constr/.dta` making sure that it is sorted on `D` and `E`. The `ipf` command will merge this file with the main dataset. Any cells that have no initial frequency after the merge will not be constrained.

```
. list
     D   E   Efreqold
  1.   0   0        2
  2.   0   1        1
  3.   1   0        1
  4.   1   1        1
```

The model fit using the `constrain` file is shown below. Note that all the variables of the `constrain` file must be specified in the `convars` option.

```
. ipf [fw=]D, fit/(D+E) convars(D E) confile(constr) exp
```

### Partial constraints in a marginal table

For illustration purposes, the variables `D` and `E` are extended to include one extra category each, call this 2. The basic fit is now given below.

```
. ipf [fw=]D, fit/(D+E) exp
```

The same `constrain.dta` file as used previously gives the following output.

```
. ipf [fw=]D, fit/(D+E) convars(D E) confile(constr) exp
```

Observe that the initial values are missing for all cells except the top left $2 \times 2$ table. Hence this table is partially constrained to have an odds ratio of 2 in the top left part of the table, but the rest of the table is unconstrained. Note that the partial constraints are a subset of the marginal table defined by the `varlist` in the `convars` option; thus, in this example, the model being fit is actually `D * E` with the partially constrained odds ratio 2. If the `constr.dta` file contained only missing values, then this would be equivalent to fitting the model `D * E`.

### References

Test for autoregressive conditional heteroskedasticity in regression error distribution

Christopher F. Baum, Boston College, baum@bc.edu
Vince Wiggins, Stata Corporation, vwiggins@stata.com


Keywords: Conditional heteroskedasticity, ARCH, Engle.

Syntax

```
archlm [if exp] [in range] [, lags(numlist) nosample ]
```

Description

Consider a regression of a time series of \( T \) values of a response \( y_t \) on a regressor matrix \( X \). The errors in this regression model may be unconditionally heteroskedastic and independently distributed, satisfying the assumptions for the application of ordinary least squares estimation, but their distribution may exhibit autoregressive conditional heteroskedasticity (ARCH), as defined by Engle (1982).

`archlm` computes Engle’s Lagrange multiplier test for ARCH\((p)\), that is, for the absence of ARCH effects up to and including \( p \)th-order, in a time-series model. See Davidson and MacKinnon (1993, 557).

This command is to be used after `regress`. The test is for use with time-series data; you must `tsset` your data before using these tests. The command displays the test statistic, degrees of freedom and \( p \)-value, and places values in the `return` array. Type `return list` to see such values.

Options

- `lags(numlist)` specifies the lag order(s) to be tested by `archlm`. Test results will then be produced for each specified lag order in `numlist`. By default, `archlm` will use \( p = 1 \), that is, a single lag.
- `nosample` indicates that the test be performed on either all observations or all observations included in `archlm`’s `if` and `in` conditions if specified. By default, `archlm` includes only observations from the estimation sample.

Remarks

The ARCH Lagrange multiplier test is a general test of the null hypothesis that the regression errors \( \epsilon_t \) are not conditionally heteroskedastic, versus the alternative that their distribution involves a \( p \)th-order ARCH process:

\[
H_1 : \quad \epsilon_t^2 = \gamma_0 + \gamma_1 \epsilon_{t-1}^2 + \gamma_2 \epsilon_{t-2}^2 + \ldots + \gamma_p \epsilon_{t-p}^2
\]

Under the null hypothesis, all of the slope coefficients, \( \gamma_1 \) through \( \gamma_p \), are zero. As Engle (1982) first showed, this hypothesis may be tested by regressing the squares of the regression residuals on a constant and \( p \) lagged values of the squared residuals. Under the null hypothesis, \( T \) times the centered \( R^2 \) from this regression will be distributed \( \chi^2(p) \), where \( T \) is the sample size and \( p \) is the number of lagged residual vectors included in the regression. If rejections are encountered, Stata’s `arch` command may be used to estimate variations of the ARCH model.

Examples

We access the Klein (1950) Model I data used as an example in the discussion of Stata’s `reg3` discussion via net-aware Stata,

```
. do http://fmwww.bc.edu/RePEc/bocode/k/klein.do
. tsset year, yearly
. regress consump wagegovt
(output omitted)
. archlm,lags(1 2 3 4)
ARCH LM test statistic, order( 1):  5.542637  Chi-sq( 1)  P-value =  .0186
ARCH LM test statistic, order( 2):  9.431075  Chi-sq( 2)  P-value =  .009
ARCH LM test statistic, order( 3):  9.039037  Chi-sq( 3)  P-value =  .0288
ARCH LM test statistic, order( 4):  8.787176  Chi-sq( 4)  P-value =  .0666
```
Consumption is regressed on the government wage bill. The tests for ARCH\((p)\) effects for orders 1, 2, 3 and 4 each reject the null hypothesis of no ARCH effects at stronger than the 10% level of significance. As Davidson and MacKinnon stress (1993, 557), such a finding may or may not indicate the presence of conditional heteroskedasticity; it may also point to other forms of misspecification.

References

### sg136 Tests for serial correlation in regression error distribution

Christopher F. Baum, Boston College, baum@bc.edu
Vince Wiggins, Stata Corporation, vwiggins@stata.com

**Abstract:** Implements Durbin’s (1970) \(h\) test and Breusch (1978) and Godfrey’s (1978) tests for autocorrelation in the disturbances. Both tests are valid in the presence of stochastic regressors, including lagged dependent variables. The \(h\) test is strictly for first-order autocorrelation whereas the Breusch–Godfrey test is applicable to autocorrelation or moving average of arbitrary degree.

**Keywords:** Autocorrelation, moving average, Durbin, Breusch, Godfrey, stochastic regressor, lagged dependent variable.

**Syntax**

```
  durbinh
  bgtest [, lags(p) ]
```

Both commands are to be used after `regress`; see [R] `regress`. Both tests are for use with time-series data. You must `tsset` your data before using these tests; see [R] `tsset`.

**Description**

Consider a regression of a time series of \(T\) values of a response \(y_t\) on a regressor matrix \(X\), possibly including one or more lagged values of the response variable. For ordinary least squares (OLS) to be the appropriate estimator, the error process \(e_t\) should be independently and identically distributed. In the context of time-series data, serial correlation is often encountered, violating the distributional assumptions on the error process. If lagged dependent variables are included in the regressor matrix, alternative tests of those distributional assumptions are required.

`durbinh` computes a form of the Durbin \(h\) test (1970) for first-order serial correlation in a model containing a lagged dependent variable among the regressors. In that context, the commonly applied Durbin–Watson test (see `dwstat`) is biased toward acceptance of the null hypothesis of zero autocorrelation. The Durbin \(h\) test provides a consistent estimate of the first-order autocorrelation coefficient \(\rho\) in the AR(1) process \(e_t = \rho e_{t-1} + v_t\) when the regressors include \(y_{t-1}\). See Davidson and MacKinnon (1993, 357–364) for details.

`bgtest` computes the Breusch–Godfrey Lagrange multiplier test (Breusch 1978, Godfrey 1978) for nonindependence in the error distribution, conditional on the lag order \(p\). The test’s null hypothesis of independence in the error distribution has “locally equivalent” alternatives (Godfrey and Wickens 1982) of either AR\((p)\) or MA\((p)\): that is, a \(p\)th-order autoregressive or moving average process. The test statistic, a \(TTR^2\) Lagrange multiplier measure, is distributed \(\chi^2(p)\) under the null hypothesis. The test is asymptotically equivalent to the Box–Pierce or Ljung–Box portmanteau tests (the \(Q\) statistic implemented in the `wntestq` command) for \(p\) lags. Unlike either form of the \(Q\) statistic, the Breusch–Godfrey test is valid in the presence of stochastic regressors such as lagged values of the dependent variable.

Both commands display the test statistic, degrees of freedom and \(p\)-value, and save results in `r()`; see [R] `saved results`. Type `return list` to see such values.

The Breusch–Godfrey test for \(p = 1\) is asymptotically equivalent to the Durbin \(h\) test. The Durbin \(h\) test statistic is presented as a Student-\(t\) test with one degree of freedom.

**Options**

`lags(p)` specifies that an autoregressive or moving average process of order \(p\) for the regression errors is to be tested. This option only applies to `bgtest`.

`bgtest` by default will use only a single lag. A greater number of lagged values may be included in the test via the `lags` option.
Remarks

The Breusch–Godfrey test is a general test of the null hypothesis that the regression errors $\epsilon_t$ are independently distributed, versus the alternative that their distribution involves a $p$th-order process:

$$H_1 : \epsilon_t = \text{AR}(p) \text{ or } \epsilon_t = \text{MA}(p)$$

where $\text{AR}(p)$ denotes the $p$th-order autoregressive process, and $\text{MA}(p)$ denotes the $p$th-order moving average process. The test statistic is computed from the regression of the least squares residuals $e_t$ on the full matrix of regressors, $X$, and $p$ lags of the residuals. Under the null hypothesis, $T$ times the uncentered $R^2$ from this regression will be distributed $\chi^2(p)$, where $T$ is the sample size and $p$ is the number of lagged residual vectors included in the regression. A rejection of the null hypothesis implies that the errors are distributed as $\text{AR}(p)$ or $\text{MA}(p)$. The indeterminacy arises from the equivalence of the derivatives of these two models when evaluated under the null hypothesis; in Godfrey and Wickens (1982) terms, they are locally equivalent alternatives under the null hypothesis.

The Durbin $h$ test is a special case of the Breusch–Godfrey test where $p = 1$. Textbook discussions of this test often provide an alternative formula which can be problematic due to the square root of a potentially negative quantity. The Breusch–Godfrey form of the test may always be computed, and is asymptotically equivalent.

Examples

We access the Klein (1950) Model I data used as an example in Stata’s discussion of the `reg3` command via net-aware Stata.

```
. do http://fmwww.bc.edu/RePEc/bocode/k/klein.do
. tsset year, yearly
. regress consump wagegovt L.consump
   (output omitted)
. durbin
   Durbin–Watson $h$-statistic:  0.7848839  t =  3.401193  P-value = 0.0037
. bgtest
   Breusch–Godfrey LM statistic:  8.393221  Chi-sq(1) P-value = 0.0038
   Breusch–Godfrey LM statistic:  7.066155  Chi-sq(2) P-value = 0.0196
```

Consumption is regressed on the government wage bill and lagged consumption.

The presence of the lagged dependent variable necessitates the use of the Durbin $h$ or Breusch–Godfrey tests. Both tests overwhelmingly reject the null hypothesis of independent errors, as does the Breusch–Godfrey test with two lags (an alternative hypothesis of $\text{AR}(2)$ or $\text{MA}(2)$ in the error distribution).

References


Syntax

\texttt{whitetst} [ \texttt{if} \ exp ] [ \texttt{in} \ range ] \ [ , \texttt{nosample} ]

\texttt{bpagan} \ varlist \ [ \texttt{if} \ exp ] \ [ \texttt{in} \ range ]

Description

Consider a regression of \( n \) values of a response on a regressor matrix \( X \) including \( p \) nonconstant regressors.

\texttt{whitetst} computes the White (1980) general test for heteroskedasticity in the error distribution by regressing the squared residuals on all distinct regressors, and their squares and cross-products. The test statistic, a Lagrange multiplier measure, is distributed as \( \chi^2(p) \) under the null hypothesis of homoskedasticity. See Greene (2000, 507–511).

\texttt{bpagan} computes the Breusch–Pagan (1979) Lagrange multiplier test for heteroskedasticity in the error distribution, conditional on a set of variables which are presumed to influence the error variance. The test statistic, a Lagrange multiplier measure, is distributed as \( \chi^2(p) \) under the null hypothesis of homoskedasticity.

Both commands are to be used after \texttt{regress}. Both commands display the test statistic, degrees of freedom and \( p \)-value, and return results in \texttt{r(\). Type \texttt{return list} to see such values.

The Breusch–Pagan test is asymptotically equivalent to White’s (1980) general test for heteroskedasticity performed by \texttt{whitetst} if the same auxiliary variables are specified (for White’s test, all distinct regressors, and their squares and cross-products). This test should not be confused with another Breusch–Pagan test implemented in Stata’s \texttt{mvreg} for the independence of error vectors in a multivariate setting.

Options

\texttt{nosample} when specified with \texttt{whitetst} indicates that the test be performed on either all observations or all observations included in \texttt{whitetst}’s \texttt{if} and \texttt{in} conditions if specified. By default, \texttt{whitetst} includes only observations from the estimation sample.

Remarks

Both these tests are general tests of heteroskedasticity which allow the researcher to take advantage of the consistency of the least squares point estimates of the coefficient vector, even in the presence of heteroskedasticity. This implies that the least squares residuals may be used to construct a test to detect heteroskedastic behavior in the true disturbances.

The White test may be described as a general test of the null hypothesis

\[ H_0 : \sigma_i^2 = \sigma^2 \text{ for all } i \]

If the null hypothesis is satisfied, the appropriate covariance matrix for the least squares coefficients will be the conventional estimator, which is based on the correct estimated covariance matrix of the least squares estimator

\[ V = s^2(X'X)^{-1} \]

If the null hypothesis is not appropriate, the correct covariance matrix will be

\[ V = s^2(X'X)^{-1} [X'\Omega X] (X'X)^{-1} \]

where \( \Omega \) is a diagonal matrix containing \( \sigma_i^2 \) on the diagonal. \( V \) may be consistently estimated by

\[ \tilde{V} = s^2(X'X)^{-1} \left[ \sum_{i=1}^{n} e_i^2 x_i x_i' \right] (X'X)^{-1} \]

where \( e_i \) are the least squares residuals and \( x_i \) is the \( i \)th row of the regressor matrix. This is the variance estimated by \texttt{regress} when the \texttt{robust} option is specified. The two estimates of the covariance matrix will differ if the null hypothesis is not supported by the data. White’s test takes advantage of this difference. It is computed as \( nR^2 \) in the regression of \( e_i^2 \), the squared residuals, on a constant and all unique variables in \( X \otimes X \). The statistic is asymptotically distributed as \( \chi^2(p) \) where \( p \) is the number of nonconstant regressors in the equation.

Although the White test is extremely general, this is also its weakness. A rejection may reveal heteroskedasticity, but it may also identify some form of misspecification, such as the exclusion of relevant variables from the equation. It is a nonconstructive test, in that a rejection does not provide a suggested remedy.
The Breusch–Pagan test is a more specific test in which the null hypothesis may be specified as

\[ H_0: \quad \sigma_i^2 = \sigma^2 f(\alpha_0 + \alpha' z_i) \]

where \( z_i \) is a set of independent variables. The model is homoskedastic if \( \alpha = 0 \). Like the White test, the test produces a Lagrange multiplier statistic, one-half the explained sum of squares in the regression of \( e_i^2 / (\hat{e}_i e/n) \) on \( z_i \). Under the null hypothesis, this statistic is asymptotically distributed as \( \chi^2(p) \) where \( p \) is the number of variables in \( z \).

**Examples**

With Stata’s auto data read in,

```
. regress price mpg weight length
. whitetst
White’s general test statistic : 39.59324    Chi-sq( 9)  P-value = 9.0e-06
```

The nine degrees of freedom for this test statistic correspond to the three regressors, \( mpg, weight, length \), their squares, and their three unique crossproducts. The small \( p \)-value indicates that the null hypothesis of homoskedasticity is overwhelmingly rejected.

```
. gen gpm=1/mpg
. regress price mpg weight length
. bpagan mpg gpm
Breusch-Pagan LM statistic:  6.75232    Chi-sq( 2)  P-value = 0.0342
```

The two degrees of freedom for the test statistic correspond to the two variables, \( mpg \) and \( gpm \), given on the \texttt{bpagan} command. The \( p \)-value indicates that the null hypothesis of homoskedasticity of the errors may be rejected at stronger than the 5% level of significance.

**Note on authorship**

\texttt{whitetst} was authored by Baum and Cox; the code was much improved by the availability of \texttt{rmcoll} (documented online in Stata updated after 28 September 1999). \texttt{bpagan} was authored by Baum and Wiggins.

**References**


There are drawbacks to the Fisher $r$-to-$z$ technique. One drawback is the assumption that the original data are distributed bivariate normal. In applied research, this is rarely the case, and when the assumption of bivariate normality breaks down, confidence intervals and inferences about correlations can be inaccurate. A second drawback, and one that is much more problematic, is that the researcher often wants to compare correlations calculated from the same sample of observations, that is, elements of a correlation matrix. Such coefficients are not independent of each other, and therefore formulas for the standard error of the difference in $z$-transforms may not be readily available.

This insert presents `bootcor`, a program that uses bootstrapping (Efron and Tibshirani 1993) to make more accurate inferences about the difference of correlation coefficients. `bootcor` creates a user-specified number of bootstrap resamples of the dataset, and computes the two correlation coefficients being compared for each resample. These two correlation coefficients are then $r$-to-$z$ transformed (to improve the symmetry of the distributions) and a difference score is calculated. A $z$-test is used on the distribution of difference scores.

`bootcor` can make inferences about Pearson product-moment correlation coefficients, intraclass correlation coefficients, and concordance coefficients. The user can specify three or four variables. If three variables are selected, a comparison is made between $r(\text{var1}, \text{var2})$ and $r(\text{var1}, \text{var3})$. If four variables are selected, a comparison is made between $r(\text{var1}, \text{var2})$ and $r(\text{var3}, \text{var4})$.

**Options**

`reps(#)` allows the user to specify how many bootstrap replications $B$ to compute. The default value of $B$ is 50. It is recommended that $B$ be at least 1000 for adequate accuracy when estimating percentiles of sampling distributions.

`stat(pearson | icc | concord)` specifies which measure of correlation should be used in the comparison. The Pearson product-moment correlation coefficient (`pearson`) is the default setting. The user can also choose from two other measures of agreement: the intraclass correlation coefficient (`icc`) or the concordance coefficient (`concord`). The user does not need to have additional commands installed for computing intraclass or concordance coefficients.

`level(#)` allows the user to specify the level of confidence for the individual correlation coefficients. Level can range from 1 to 99.9. The default is 95.

`saving(newfile)` will export the bootstrap replications to a `.dct` file that the user can later analyze in more detail. Five variables are saved, with each resample listed casewise. $r_{\text{boot}1}$ is $r_1$, $r_{\text{boot}2}$ is $r_2$, $z_{\text{boot}1}$ is the Fisher-transformed value of $r_{\text{boot}1}$, $z_{\text{boot}2}$ is the Fisher-transformed value of $r_{\text{boot}2}$, and $z_{\text{bootd}}$ is the difference of $z_{\text{boot}1}$ and $z_{\text{boot}2}$.

The user can load this file into Stata using the command

```
    . infile using newfile.dct
```

**Examples**

The following examples are demonstrated on a subset of data from a dataset of alcohol-related measures in college students. Data were collected at two times, within one week of each other. The dataset is called `bootcor.dta` and is provided on the STB diskette.

```
    . use bootcor.dta, clear
    . describe
    Contains data from bootcor.dta
    obs: 82
    vars: 8
    size: 1,476 (99.1% of memory free)
    --------------------------------------------
    1. ads1 byte %8.0g Alcohol Dependence Time 1
    2. ads2 byte %8.0g Alcohol Dependence Time 2
    3. rapy1 byte %8.0g Alcohol Related Problems in the Last Year Time 1
    4. rapym1 byte %8.0g Alcohol Related Problems in the Last Month Time 1
    5. rapy2 byte %8.0g Alcohol Related Problems in the Last Year Time 2
    6. rapym2 byte %8.0g Alcohol Related Problems in the Last Month Time 2
    7. bac1 float %9.0g Peak BAC Time 1
    8. bac2 float %9.0g Peak BAC Time 2
    --------------------------------------------
    Sorted by:
```

The user can load this file into Stata using the command

```
    . infile using newfile.dct
```
Comparing the test-retest reliabilities of measures

In this first analysis, it is of interest whether the intraclass test-retest correlation coefficients of the two measures of alcohol-related problems are equal. In other words, is there any difference in the reliability of estimates of the number of alcohol problems in the past month versus problems in the last year?

```
. bootcor rapiy1 rapiy2 rapim1 rapim2, reps(1000) stat(icc) level(90)
Results of Bootstrap Comparison of Intraclass Correlation
Bootstrap Replications: 1000 Observations: 82

Variables Observed Bootstrap Mean(ρ) [ 90% CI ]
------------- ------------- ------------- ------------- -------------
rapiy1 & rapiy2 0.901 0.900 0.865 0.926
rapim1 & rapim2 0.676 0.667 0.657 0.707

Z-score of Fisher R-to-Z Difference: 4.671 P-Value: 0.000
```

The results of this bootstrap comparison yield a highly significant result, with a $z = 4.671$. We would reject the null hypothesis that these two assessments have the same test-retest reliability; it appears that people are reliably better at reporting alcohol-related problems over the past year than in the past month. Also of interest are the confidence intervals for the two parameter estimates. The 90% confidence intervals for the parameters $\rho(rapiy1, rapiy2)$ and $\rho(rapim1, rapim2)$ are listed above.

In the second analysis, the question of interest is whether the strength of the relationship between peak blood alcohol content and alcohol-related problems is the same as peak blood alcohol content and alcohol dependence symptoms.

```
. bootcor bac1 rapim1 ads1 reps(1000) level(90)
(0 observations deleted)
Results of Bootstrap Comparison of Pearson’s R
Bootstrap Replications: 1000 Observations: 82

Variables Observed Bootstrap Mean(ρ) [ 90% CI ]
bac1 & rapim1 0.652 0.655 0.504 0.786
bac1 & ads1 0.502 0.502 0.366 0.624

Z-score of Fisher R-to-Z Difference: 1.577 P-Value: 0.115
```

The results of this bootstrap comparison yield a nonsignificant result, with a $z = 1.577$ and $p = .115$. The 90% confidence intervals for the parameters $\rho(bac1, rapim1)$ and $\rho(bac1, ads1)$ are listed above as well.
Saved Results

\texttt{bootcor} saves in \texttt{r():}

\texttt{Scalars}

\texttt{r(z)} \hspace{1cm} \text{observed } z\text{-value of the mean of the difference scores}

\texttt{r(p)} \hspace{1cm} \text{probability of observing a } z\text{ equal to or more extreme than observed}

\texttt{r(corr1)} \hspace{1cm} \text{value of } r_1\text{ as calculated from the dataset}

\texttt{r(bcorr1)} \hspace{1cm} \text{observed mean of the bootstrap distribution of } r_1

\texttt{r(bcorr1l)} \hspace{1cm} \text{lower limit of the confidence interval of } r_1

\texttt{r(bcorr1u)} \hspace{1cm} \text{upper limit of the confidence interval of } r_1

\texttt{r(corr2)} \hspace{1cm} \text{value of } r_2\text{ as calculated from the dataset}

\texttt{r(bcorr2)} \hspace{1cm} \text{observed mean of the bootstrap distribution of } r_2

\texttt{r(bcorr2l)} \hspace{1cm} \text{lower limit of the confidence interval of } r_2

\texttt{r(bcorr2u)} \hspace{1cm} \text{upper limit of the confidence interval of } r_2

\texttt{r(bse1)} \hspace{1cm} \text{standard error of the bootstrap distribution of } r_1

\texttt{r(bse2)} \hspace{1cm} \text{standard error of the bootstrap distribution of } r_2

\texttt{r(bsed)} \hspace{1cm} \text{standard error of the bootstrap distribution of difference scores}

Acknowledgment

The author thanks Elizabeth T. Miller for providing the dataset used in the examples.

References


---

\textbf{sg139} \hspace{1cm} \textbf{Logistic regression when binary outcome is measured with uncertainty}

\begin{tabular}{l}
Mario Cleves, Stata Corporation, mcleves@stata.com \\
Alberto Tosetto, S. Bortolo Hospital, Vicenza, Italy, tosetto@hemato.ven.it
\end{tabular}

\textbf{Abstract:} \textit{Traditional logit or logistic regression assumes that the outcome variable is measured without error. In some studies, however, the outcome variable is measured with imperfect sensitivity and specificity. It is known that the resulting misclassification will lead to biased parameter point estimates and variances. In this insert we implement an EM algorithm suggested by Magder and Hughes (1997) that produces unbiased estimates of parameters and their variances.}

\textbf{Keywords:} Logit, logistic models, sensitivity, specificity, EM algorithm, measurement error.

\textbf{Syntax}

\begin{verbatim}
logitem depvar [indepvars] [if exp] [in range], sens(sensvar | #) spec(specvar | #)
[ level(#) robust nolog noorite rate(#) tolerance(#) ltolerance(#) ]
\end{verbatim}

\textbf{Syntax for predict}

\begin{verbatim}
predict [type] newvarname [if exp] [in range] [, statistic ]
\end{verbatim}

where \textit{statistic} is

\begin{tabular}{ll}
\texttt{p} & \text{probability of a positive outcome (the default)} \\
\texttt{xb} & \text{x}_i\beta, \text{fitted values} \\
\texttt{stdp} & \text{standard error of the prediction} \\
\texttt{number} & \text{sequential number of the covariate pattern}
\end{tabular}

Unstarred statistics are available both in and out of sample; type \texttt{predict ... if e(sample)} ... if wanted only for the estimation sample. Starred statistics are calculated only for the estimation sample even when \texttt{if e(sample)} is not specified.

\textbf{Description}

\texttt{logitem} uses an expectation-maximization (EM) algorithm to estimate a maximum-likelihood logit regression model when the outcome variable is measured with an imperfect test of known sensitivity and specificity.

The method allows the sensitivity and specificity to vary across observations.
Options

sens(sensvar | #) specifies the value or the name of the sensitivity variable. Sensitivity should be between 0 and 1.
spec(specvar | #) specifies the value or the name of the specificity variable. Specificity should be between 0 and 1.
level(#) specifies the confidence level, in percent, for confidence intervals. The default is level(95) or as set by set level.
robust specifies that the Huber/White/sandwich estimator of variance is to be used in place of the traditional calculation.
nolog prevents logitem from showing the iteration log.
noor reports the estimated coefficients instead of odds ratios. This option affects how results are displayed, not how they are estimated. noor may be specified at estimation or when redisplaying previously estimated results.
iterate(#), tolerance(#), and ltolerance(#) specify the definition of convergence.
iterate(16000) tolerance(1e-6) ltolerance(0) is the default.

Convergence is declared when

\[ \text{mreldif}(b_{i+1}, b_i) \leq \text{tolerance()} \]
\[ \text{or reldif}(\ln L(b_{i+1}), \ln L(b_i)) \leq \text{ltolerance}() \]

for two consecutive EM steps. In addition, iteration stops when \( i = \text{iterate}() \); in that case, results along with the message “convergence not achieved” are presented. The return code is still set to 0.

Options for predict

p, the default, calculates the probability of a positive outcome.
xb calculates the linear prediction.
stdp calculates the standard error of the linear prediction.
number numbers the covariate patterns—observations with the same covariate pattern have the same number. Observations not used in estimation have the prediction set to missing. The “first” covariate pattern is numbered 1, the second 2, and so on.

Remarks

Traditional logit or logistic regression assumes that the outcome variable is measured without error. In some studies, however, the outcome variable is not measured perfectly. This can occur, for example, when using a diagnostic test having sensitivity and/or specificity lower than 100%. The resulting misclassification can lead to bias in the coefficients estimated and related statistics (Copeland, et al. 1977).

Magder and Hughes (1997) proposed an EM algorithm that incorporates the values of the sensitivity and specificity of the classification test into the estimation of the logistic parameters. They showed that in the presence of misclassification, their procedure produced unbiased estimates of both the coefficients and their variances. It is this EM algorithm that we have implemented in logitem. Note that when sensitivity and specificity are both set to one, logitem and logistic produce the same estimates.

Examples

Tosetto, et al. (1999) conducted a case–control study to determine the importance of the prothrombin gene allele G20210A as a risk factor in venous thromboembolism (VTE). The study consisted of 116 VTE patients and 232 healthy individuals ascertained randomly from a well defined population. For each subject in the study, they obtained information regarding previous diagnosis of VTE using a survey tool with an estimated sensitivity of 71.3% and specificity of 98.9%.

Each subject in the study was also typed at the prothrombin locus. No homozygous carriers of the mutated allele (G20210A) were found. Thirteen (3.7%) subjects were heterozygous for the mutation and the remaining 335 subjects did not have the mutation.

In our data, case indicates whether the patient has been diagnosed with VTE, and pro whether the individual has the mutation. Here are the results from logistic:

```
. logistic case pro
Logit estimates
Number of obs = 348
LR chi2(1) = 0.16
Prob > chi2 = 0.6926
Pseudo R2 = 0.0004
Log likelihood = -221.42878
```
and those from **logitem** incorporating the sensitivity and specificity:

```
. logitem case pro, sens(.713) spec(.969) nolog
```

**logistic** regression when outcome is uncertain

| Odds Ratio | Std. Err. | z | P>|z| | [95% Conf. Interval] |
|------------|-----------|---|-------|----------------------|
| pro        | 1.35498   | 1.065479 | 0.387 | 0.699 | .5904148 | 6.320728 |

Neither model provides evidence supporting the hypothesis of an association between the mutated allele and VTE. Note that although the odds ratio reported by **logitem** is larger—further from the null—than that reported by standard logistic regression, its standard error is larger, reflecting the added uncertainty about the outcome variable. This is a known property of this method; namely, the EM algorithm typically produces larger odds ratios and larger variances.

**Saved Results**

**logitem** saves in `e()`:

Scalars

<table>
<thead>
<tr>
<th>e(N)</th>
<th>number of observations</th>
</tr>
</thead>
<tbody>
<tr>
<td>e(ll)</td>
<td>log likelihood</td>
</tr>
<tr>
<td>e(ll_0)</td>
<td>log likelihood, constant-only model</td>
</tr>
<tr>
<td>e(df_m)</td>
<td>model degrees of freedom</td>
</tr>
<tr>
<td>e(chi2)</td>
<td>(\chi^2)</td>
</tr>
<tr>
<td>e(r2_p)</td>
<td>pseudo (R^2)</td>
</tr>
</tbody>
</table>

Macros

| e(cmd) | logitem |
| e(depvar) | name of dependent variable |
| e(chi2type) | LR: type of model \(\chi^2\) test |

Matrices

| e(b) | coefficient vector |
| e(V) | variance–covariance matrix of the estimators |

Functions

| e(sample) | marks estimation sample |

**Methods and Formulas**

Let \(Y_i = 1\) if individual \(i\) truly has the outcome of interest (diseased) and 0 otherwise (nondiseased). Let \(T_i = 1\) if individual \(i\) is classified as having the outcome and 0 otherwise. Assume that \(\hat{Y}_i\) is the probability that the \(i\)th individual truly has the condition being studied given the values of \(T_i\) and \(k \times 1\) covariate vector \(X_i\). Then if individual \(i\) is classified as having the outcome (\(T_i = 1\)),

\[
\hat{Y}_i = \frac{\text{Prob}(Y_i = 1|X_i, \beta) \times \text{sensitivity}}{\text{Prob}(Y_i = 1|X_i, \beta) \times \text{sensitivity} + \text{Prob}(Y_i = 0|X_i, \beta) \times (1 - \text{specificity})}
\]

and if \(T_i = 0\),

\[
\hat{Y}_i = \frac{\text{Prob}(Y_i = 1|X_i, \beta) \times (1 - \text{sensitivity})}{\text{Prob}(Y_i = 1|X_i, \beta) \times (1 - \text{sensitivity}) + \text{Prob}(Y_i = 0|X_i, \beta) \times \text{specificity}}
\]

where \(\beta\) is a \(k \times 1\) coefficient vector to be estimated, and

\[
\text{Prob}(Y_i = 1|X_i, \beta) = \frac{\exp(\sum_{j=0}^{k} \beta_j X_{ij})}{1 + \exp(\sum_{j=0}^{k} \beta_j X_{ij})}
\]
The EM algorithm begins by first setting $\beta$ to an arbitrary value and computing $\hat{Y}_i$ for each observation. This is the expectation step.

The data are then duplicated and each observation included twice, once with the outcome variable set to 1 and another with the outcome set to zero. A weighted logistic regression model is fitted with weights equal to $b_i$ if the outcome variable is 1 and $(1 - \hat{Y}_i)$ if it is zero. This constitutes the maximization step.

The new $\beta$’s obtained from the fitted logistic model are used to calculate new $\hat{Y}_i$’s and the process repeated until convergence is declared.

References

sg140 The Gumbel quantile plot and a test for choice of extreme models

Manuel G. Scotto, University of Lisbon, arima@mail.telepac.pt

Abstract: Some statistical tools for exploratory data analysis are presented. The Gumbel quantile plot is described as an informal way to test if the Gumbel distribution provides a good fit for data. Furthermore, we include a method of statistical choice among the three extreme value distributions.

Keywords: Generalized extreme value distribution, hypothesis testing, Gumbel quantile plot.

Syntax

gqpt varname [if exp] [in range]

Introduction

The main goal of this work is in dealing with the statistical choice of extreme models. This is essential in applications where the attention is focused at rarely occurring events, such as an annual maximal flood exceeding dykes, or a seasonal minimal temperature below the critical value for crop production. We restrict ourselves to the one-dimensional case and start with a discussion of the problem.

Let $X_1, \ldots, X_n$ be independent and identically distributed random variables with underlying marginal distribution given by

$$G_\xi(x; \lambda, \delta) = \exp\left(-\left(1 + \frac{x - \lambda}{\delta}\right)^{-\frac{1}{\xi}}\right), \quad 1 + \frac{x - \lambda}{\delta} > 0, \quad -\infty < \xi < \infty$$

which is the well-known generalized extreme value distribution (GEV). The parameters $\lambda$ and $\delta$ are the location and scale parameters respectively and $\xi$ is the shape parameter and may be used to model a wide range of tail behaviors. There are three particular forms of $G$ corresponding to $\xi > 0$ (Fréchet distribution), $\xi < 0$ (Weibull distribution), and $\xi = 0$ being interpreted as the limit as $\xi \to 0$, widely called the Gumbel distribution. We use the Gumbel quantile plot (GQP) and the statistic first introduced by Gumbel and developed by Tiago de Oliveira and Gomes (1984), hereafter referred to as OG, for a quick statistical choice between the extreme models.

The quantile plot for the Gumbel distribution

Probability plotting papers are commonly used to assess, in an informal way, whether a sample comes from a particular distribution. For the Gumbel distribution, the quantile function is given by

$$\Lambda(x; \lambda, \delta) = \exp\left(-\exp\left(-\frac{x - \lambda}{\delta}\right)\right), \quad -\infty < x < \infty$$

which leads to so called double logarithmic plotting. To this end, we first take the ordered sample $X_{1:n} \leq \ldots \leq X_{n:n}$ and plot $X_{k:n}$ versus $-\log(-\log(p_k))$, where $p_k = i/(n + 1)$ is the classical plotting position. If the Gumbel distribution provides a good fit to our data, then the GQP should look roughly linear. Furthermore, both Fréchet and Weibull models can also be validated by
means of the GQP. If the plot has a downside concavity we can assume a Fréchet model whereas an upside concavity indicates a Weibull model. Finally, note that

\[ -\log(-\log(p_i)) = -\frac{\lambda}{\delta} + \frac{X_{i:n}}{\delta} \]

Using linear regression, quick estimates for \( \lambda \) and \( \delta \) can be deduced from the slope and the intercept. Maximum likelihood estimators can be obtained by means of the \texttt{gumbel} command introduced in Scotto and Tobias (1998).

**Statistical choice between the extreme models**

Statistical choice among the extreme models gives a central and preeminent position to the Gumbel distribution due to the simplicity of inferences associated with this distribution. We present a test for \( H_0: \xi = 0 \) in the GEV(\( \xi \)) model. We consider the statistic,

\[ Q_n = \frac{X_{n:n} - X_{\lfloor n/2 \rfloor + 1:n}}{X_{\lfloor n/2 \rfloor + 1:n} - X_{1:n}} \]

which is location and dispersion-parameter free. Under the validity of \( H_0 \), it was shown by OG that there exist \( a_n > 0 \) and \( b_n \) such that \( W_n = a_n (Q_n - b_n) \rightarrow \Lambda(\cdot) \). One choice is \( a_n = \log \log 2 \) and \( b_n = (\log n + \log \log 2)/(\log \log n - \log \log 2) \). OG proposed a simple deciding rule in order to decide among the extreme models; choose \( 0 < b < a < \infty \) and decide for the Gumbel distribution when \( b \leq W_n \leq a \), for the Fréchet distribution when \( W_n > a \), and for the Weibull distribution when \( W_n < b \). The values of \( a \) and \( b \) corresponding to the usual significance are given in the table below.

<table>
<thead>
<tr>
<th>( \alpha )</th>
<th>( a )</th>
<th>( b )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.050</td>
<td>-1.561334</td>
<td>3.161461</td>
</tr>
<tr>
<td>0.025</td>
<td>-1.719620</td>
<td>3.843121</td>
</tr>
<tr>
<td>0.010</td>
<td>-1.893530</td>
<td>4.740459</td>
</tr>
<tr>
<td>0.001</td>
<td>-2.222295</td>
<td>7.010001</td>
</tr>
</tbody>
</table>

**Example**

We applied both the GQP and the statistical test described above, to the annual maximum sea levels in Venice dataset during the period 1981–82 (Smith 1986).

```plaintext
  . gqpt seal
  Variable | Delta Lambda Q W
  15.935767 96.122623 2.3609653 .41984981

Values corresponding to the usual significance levels

<table>
<thead>
<tr>
<th>alpha</th>
<th>a</th>
<th>b</th>
</tr>
</thead>
<tbody>
<tr>
<td>.050</td>
<td>-1.561334</td>
<td>3.161461</td>
</tr>
<tr>
<td>.025</td>
<td>-1.719620</td>
<td>3.843121</td>
</tr>
<tr>
<td>.010</td>
<td>-1.893530</td>
<td>4.740459</td>
</tr>
<tr>
<td>.001</td>
<td>-2.222295</td>
<td>7.010001</td>
</tr>
</tbody>
</table>
```

This gives rise to the graph in Figure 1.
Figure 1. Gumbel quantile plot of annual maximum sea level in Venice, for the period 1981–82.

References


Syntax
Basic syntax
```
treatreg depvar [varlist], treat(depvar = varlist) [ twostep ]
```

Full syntax for maximum likelihood estimates only
```
treatreg depvar [varlist] [weight] [if exp] [in range], treat(depvar = varlist, [ noconstant ])
[ robust cluster(varname) hazard(newvarname) noconstant first noskip level(#)
iterate(#) maximize_options ]
```

Full syntax for two-step consistent estimates only
```
treatreg depvar [varlist] [if exp] [in range], twostep treat(depvar = varlist, [ noconstant ])
[ hazard(newvarname) noconstant first level(#) ]
```

pweights, aweights, fweights, and iweights are allowed with maximum likelihood estimation; see [U] 14.1.6 weight. No weights are allowed if twostep is specified.

treatreg shares the features of all estimation commands; see [U] 23 Estimation and post-estimation commands.
Syntax for predict

```
predict [type] newvarname [if exp] [in range] [., statistic]
```

where statistic is

- `xb`: $x_j\hat{b}$, fitted values (the default)
- `yctrt`: $E(y_j \mid \text{treatment} = 1)$
- `ycntrt`: $E(y_j \mid \text{treatment} = 0)$
- `ptrt`: $P(\text{treatment} = 1)$
- `xbtrt`: linear prediction for treatment equation
- `stdptrt`: standard error of the linear prediction for treatment equation
- `stdp`: standard error of the prediction
- `stdf`: standard error of the forecast

Description

`treatreg` estimates a treatment effects model using either a two-step consistent estimator or full maximum likelihood. The treatment effects model considers the effect of an endogenously chosen binary treatment on another endogenous continuous variable, conditional on two sets of independent variables.

Options

`treat(...)`: specifies the variables and options for the treatment equation. It is an integral part of specifying a treatment effects model and is not optional.

`twostep`: specifies that two-step efficient estimates of the parameters, standard errors, and covariance matrix are to be produced.

`robust`: specifies that the Huber/White/sandwich estimator of the variance is to be used in place of the conventional MLE variance estimator. `robust` combined with `cluster(/)` further allows observations which are not independent within cluster (although they must be independent between clusters).

If you specify `pweights`, `robust` is implied; See [U] 23.11 Obtaining robust variance estimates.

`cluster(varname)`: specifies that the observations are independent across groups (clusters) but not necessarily independent within groups. `varname` specifies to which group each observation belongs. `cluster()` affects the estimation of the variance–covariance matrix and, thus, of the standard errors (VCE), but not the estimated coefficients. `cluster()` can be used with `pweights` to produce estimates for unstratified cluster-sampled data.

`cluster()` implies `robust`; that is, specifying `robust cluster()` is equivalent to typing `cluster()` by itself.

`hazard(newvarname)`: will create a new variable containing the hazard from the treatment equation. The hazard is computed from the estimated parameters of the treatment equation.

`noconstant`: suppresses the constant term (intercept) in the model. This option may be specified on the regression equation, the treatment equation, or both.

`first`: specifies that the first-step probit estimates of the treatment equation be displayed prior to estimation.

`noskip`: specifies that a full maximum likelihood model with only a constant for the regression equation be estimated. This model is not displayed but is used as the base model to compute a likelihood-ratio test for the model test statistic displayed in the estimation header. By default, the overall model test statistic is an asymptotically equivalent Wald test of all the parameters in the regression equation being zero (except the constant). For many models, this option can significantly increase estimation time.

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

`iterate(#)`: restricts the maximum number of iterations during optimization to the specified number; see [R] maximize.

`iterate(0)`: produces two-step parameter estimates with standard errors computed from the inverse Hessian of the full information matrix at the two-step solution for the parameters. As an alternative, the `twostep` option computes two-step consistent estimates of the standard errors.

`maximize_options` control the maximization process; see [R] maximize. You will seldom need to specify any of the maximize options except for `iterate(0)` and possibly `difficult`. If the iteration log shows many “not concave” messages and it is taking many iterations to converge, try the `difficult` option to see if that helps it to converge in fewer steps.
Options for predict

**xb** the default, calculates the linear prediction \( x_j \beta \).

**yctr** calculates the expected value of the dependent variable conditional on the presence of the treatment; \( E(y_j \mid \text{treatment} = 1) \).

**ycntr** calculates the expected value of the dependent variable conditional on the absence of the treatment; \( E(y_j \mid \text{treatment} = 0) \).

**ptr** calculates the probability of the presence of the treatment: \( P(\text{treatment} = 1) = \Pr(w_j \gamma + u_j > 0) \).

**xbtr** calculates the linear prediction for the treatment equation.

**stptr** calculates the standard error of the linear prediction for the treatment equation.

**stdp** calculates the standard error of the prediction. It can be thought of as the standard error of the predicted expected value or mean for the observation’s covariate pattern. This is also referred to as the standard error of the fitted value.

**stdf** calculates the standard error of the forecast. This is the standard error of the point prediction for a single observation. It is commonly referred to as the standard error of the future or forecast value. By construction, the standard errors produced by **stdf** are always larger than those produced by **stdp**; see [R] regress Methods and Formulas.

**Remarks**

The treatment effects model estimates the effect of an endogenous binary treatment, \( Z_j \), on a continuous, fully-observed variable \( y_j \), conditional on the independent variables \( x_j \) and \( w_j \). The primary interest is in the regression function

\[
y_j = x_j \beta + \delta z_j + \epsilon_j
\]

where \( z_j \) is an endogenous dummy variable indicating whether the treatment is assigned or not. The binary decision to obtain the treatment \( z_j \) is modeled as the outcome of an unobserved latent variable, \( z_j^* \). It is assumed that \( z_j^* \) is a linear function of the exogenous covariates \( w_j \) and a random component \( u_j \). Specifically,

\[
z_j^* = w_j \gamma + u_j
\]

and the observed decision is

\[
z_j = \begin{cases} 
1, & \text{if } z_j^* > 0 \\
0, & \text{otherwise}
\end{cases}
\]

where \( \epsilon \) and \( u \) are bivariate normal with mean zero and covariance matrix

\[
\begin{bmatrix}
\sigma & \rho \\
\rho & 1
\end{bmatrix}
\]

There are many variations of this model in the literature. Maddala (1983) derives the maximum likelihood and two-step estimators of the version implemented here. Maddala (1983) also gives a brief review of several empirical applications of this model. Barnow, et al. (1981) provide another useful derivation of this model. Barnow et al. (1981) concentrate on deriving the conditions in which the self-selection bias of the simple OLS estimator of the treatment effect, \( \delta \), is nonzero and of a specific sign.

**Example**

We will illustrate **treatreg** using a subset of the Mroz data distributed with Berndt (1991). This dataset contains 753 observations on women’s labor supply. Our subsample is of 250 observations, with 150 market laborers and 100 nonmarket laborers. Since 40% of the women in our sample chose not to enter the labor market, the simple treatment regression model is not the correct model for these data. Ideally, we would like a model that accounts for the sample selection on entering the labor force and the endogeneity of the college degree. Despite this misspecification, this dataset can be used to illustrate how the **treatreg** command works.

```
. use labor, clear
. describe
Contains data from labor.dta
    obs:     250
    vars:     15
    size:   16,000 (98.4% of memory free)

1. lfp   float %9.0g  1 if woman worked in 1975
2. whrs   float %9.0g  wife's hours of work
```

...
We will assume that the wife went to college if her educational attainment was more than 12 years. Let \( wc \) be the dummy variable indicating whether the individual went to college. With this definition, our sample contains the following distribution of college education.

\[
\text{. gen wc = 0}
\]
\[
\text{. replace wc = 1 if we > 12}
\]
(69 real changes made)
\[
\text{. tab wc}
\]

<table>
<thead>
<tr>
<th>wc</th>
<th>Freq.</th>
<th>Percent</th>
<th>Cum.</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>181</td>
<td>72.40</td>
<td>72.40</td>
</tr>
<tr>
<td>1</td>
<td>69</td>
<td>27.60</td>
<td>100.00</td>
</tr>
</tbody>
</table>

Total | 250   | 100.00  |

We will model the wife's wage as a function of her age, whether the family was living in a big city, and whether she went to college. An ordinary least squares estimation produces the following results:

\[
\text{. regress ww wa cit wc}
\]

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>Number of obs = 250</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>93.2396568</td>
<td>3</td>
<td>31.0799523</td>
<td>F(3, 246) = 4.82</td>
</tr>
<tr>
<td>Residual</td>
<td>1880.32762</td>
<td>246</td>
<td>7.45157627</td>
<td>R-squared = 0.0555</td>
</tr>
<tr>
<td>Total</td>
<td>1880.32762</td>
<td>249</td>
<td>7.45157627</td>
<td>Adj R-squared = 0.0440</td>
</tr>
</tbody>
</table>

| ww | Conf. Std. Err. | t | P>|t| | [95% Conf. Interval] |
|----|-----------------|---|------|-------------------------|
| wa | -0.0104985 | 0.193667 | -0.545 | 0.586 | -0.484472 | 0.274502 |
| cit | 0.1278922 | 0.338908 | 0.377 | 0.706 | -0.5396351 | 0.7954194 |
| wc | 1.332192 | 0.364843 | 3.656 | 0.000 | 0.6143619 | 2.050001 |
| _cons | 2.278337 | 0.432385 | 2.702 | 0.007 | 0.6174459 | 3.939225 |

Is 1.332 a consistent estimate of the marginal effect of a college education on wages? If individuals choose whether or not to attend college and the error term of the model that gives rise to this choice is correlated with the error term in the wage
equation, then the answer is no. (See Barnow et al. 1981 for a good discussion of the existence and sign of selectivity bias.) One might suspect that individuals with higher abilities, either innate or due to the circumstances of their birth, would be more likely to go to college and to earn higher wages. Such ability is, of course, unobserved. Furthermore, if the error term in our model for going to college is correlated with ability, and the error term in our wage equation is correlated with ability, then the two terms should be positively correlated. These conditions make the problem of signing the selectivity bias equivalent to an omitted-variable problem. In the case at hand, since we would anticipate the correlation between the omitted variable and a college education to be positive, we suspect that OLS is biased upwards.

To account for the bias, we fit the treatment effects model. We model the wife’s college decision as a function of her mother’s and her father’s educational attainment. Thus, we are interested in estimating the model

\[ \begin{align*}
    \text{ww} &= \beta_0 + \beta_1 \text{wa} + \beta_2 \text{cit} + \delta \text{wc} + \epsilon \\
    \text{wc}^* &= \gamma_0 + \gamma_1 \text{wmed} + \gamma_2 \text{wfed} + \mu
\end{align*} \]

where

\[ \text{wc} = \begin{cases} 
    1, & \text{wc}^* > 0, \text{ i.e., wife went to college} \\
    0, & \text{otherwise}
\end{cases} \]

and where \(\epsilon\) and \(u\) have a bivariate normal distribution with covariance matrix

\[
\begin{bmatrix}
\sigma & \rho \\
\rho & 1
\end{bmatrix}
\]

The following output gives the maximum likelihood estimates of the parameters of this model.

```
   . treatreg ww wa cit, treat(wc=wmed wfed)
   Iteration 0:  log likelihood = -707.07237
   Iteration 1:  log likelihood = -707.07215
   Iteration 2:  log likelihood = -707.07215
Treatment effects model -- MLE
Number of obs = 250
Wald chi2(3)  = 4.11
Prob > chi2 = 0.2501
Log likelihood = -707.07215
                    | Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------------+-----------------------------------------------
                   ww |  .0110424   .0199652  -0.553    0.580    -.0501736    .0290877
                    wa |  .127636   .3361988   0.380    0.704    -.5312917    .7866538
                   cit |  1.271377   .7412951   1.715    0.086    -.1815842    2.724239
                   wc |  2.316638   .9397573   2.457    0.014     .4767477    4.160529
                   _cons |  -.2631876   .3309128  -0.795    0.426    -.9380458    .3966916
                   wc |  .1198055   .0320056   3.743    0.000     .0570757    .1825352
                   wmed |  .0961896   .0290668   3.307    0.001     .0391795    .1531977
                  wfed |  -.2631876   .3309128  -0.795    0.426    -.9380458    .3966916
                    _cons |  -.2631876   .3309128  -0.795    0.426    -.9380458    .3966916
                   rshrho |  .1789688   .1599998   0.994    0.325    -.3548063    .3902399
                   lnsigma |  .9241584   .0447455  20.654    0.000     .8364888    1.018658
                   lambda |  .0786359    .175337   4.502    0.000     .0361452    .1211266
                   rho |  .0179469    .199959   0.094    0.925    -.3450603    .3902399
                   sigma |  2.519747   1.127473   2.260    0.024     1.275601    3.763904
                   lambda |  .0450149    .496442  -0.090    0.927    -.8893105    .9893104
-------------------+-----------------------------------------------
LR test of indep. eqns. (rho = 0):  chi2(1) =  0.01  Prob > chi2 = 0.9251
```

In the input, we specified that the continuous dependent variable, \(\text{ww}\) (wife’s wage), is a linear function of \(\text{cit}\) and \(\text{wa}\). Note the syntax for the treatment variable. The treatment \(\text{wc}\) is not included in the first variable list; it is specified in the \(\text{treat()}\) option. In this example, \(\text{wmed}\) and \(\text{wfed}\) are specified as the exogenous variables in the treatment equation.

The output has the form of many two-equation estimators in Stata. We note that our conjecture that the OLS estimate was biased upwards is verified. But perhaps more interesting, the size of the bias is negligible and the likelihood-ratio test at the bottom of the output indicates that we cannot reject the null hypothesis that the two error terms are uncorrelated. This result might be due to several specification errors. We ignored the selectivity bias due to the endogeneity of entering the labor market. We have also written both the wage equation and the college education equation in crude linear form, ignoring any higher power terms or interactions.
The results for the two ancillary parameters require explanation. For numerical stability during optimization, \texttt{treatreg} does not directly estimate $\rho$ or $\sigma$. Instead, \texttt{treatreg} estimates the inverse hyperbolic tangent of $\rho$,

$$\text{atanh } \rho = \frac{1}{2} \ln \left( \frac{1 + \rho}{1 - \rho} \right)$$

and $\ln \sigma$. Also, \texttt{treatreg} reports $\lambda = \rho \sigma$, along with an estimate of the standard error of the estimate and a confidence interval for it.

**Technical Note**

If each of the equations in the model had contained many regressors, the \texttt{treatreg} command could become quite long. An alternate way of specifying our wage model would be to make use of Stata’s local macros. The following lines are an equivalent way of estimating our model.

```
* local wageeq "ww wa cit"
* local trteq "wc/=wmed wfed"
* treatreg `wageeq', treat(`trteq')
```

**Example (continued)**

Stata will also produce a two-step estimator of the model with the \texttt{twostep} option. Maximum likelihood estimation of the parameters can be time-consuming with large datasets, and the two-step estimates may provide a good alternative in such cases. Continuing with the women’s wage model, we can obtain the two-step estimates with consistent covariance estimates by typing

```
* treatreg ww wa cit, treat(wc/=wmed wfed) twostep
```

The reported $\lambda$ is the parameter estimate on the hazard from the augmented regression. The augmented regression is derived in Maddala (1983) and presented in the Methods and Formulas section below.

The default statistic produced by \texttt{predict} after \texttt{treatreg} is the expected value of the dependent variable from the underlying distribution of the regression model. For the case at hand this statistic is

$$\texttt{ww} = \beta_0 + \beta_1 \text{wa} + \beta_2 \text{cit} + \delta \text{wc} + \epsilon$$

Several other interesting aspects of the treatment effects model can be explored with \texttt{predict}. Continuing with our wage model, the wife’s expected wage, conditional on attending college, can be obtained with the \texttt{yctrt} option. The wife’s expected wages, conditional on not attending college, can be obtained with the \texttt{ycntrt} option. Thus, the difference in expected wages between participants and nonparticipants is the difference between \texttt{yctrt} and \texttt{ycntrt}. For the case at hand, we have the following calculation:

```
* predict wctrt, yctrt
```
Technical Note

The difference in expected earnings between participants and nonparticipants is

\[ E[y_i \mid z_i = 1] - E[y_i \mid z_i = 0] = \delta + \rho \sigma \left( \frac{\Phi_k}{\Phi_k (1 - \Phi_i)} \right) \]

If the correlation between the error terms, \( \rho \), is zero, then the problem reduces to one estimable by OLS and the difference is simply \( \delta \). Since \( \rho \) is positive in our example, we see that least squares overestimates the treatment effect.

Saved Results

treatreg saves in e():

Scalars
- \( e(N) \) number of observations
- \( e(k) \) number of variables
- \( e(kp) \) number of equations
- \( e(kd) \) number of dependent variables
- \( e(df) \) model degrees of freedom
- \( e(ll) \) log likelihood
- \( e(rc) \) return code
- \( e(sigma) \) \( \sigma^2 \)
- \( e(chi2) \) \( \chi^2 \) for comparison test
- \( e(chi2c) \) \( \chi^2 \) for comparison test
- \( e(p) \) p-value for \( \chi^2 \) test
- \( e(ic) \) number of iterations
- \( e(lambda) \) \( \lambda \)

Macros
- \( e(cmd) \) \texttt{treatreg} name of likelihood-evaluator program
- \( e(depvar) \) name(s) of dependent variable(s)
- \( e(title) \) title in estimation output
- \( e(clustervar) \) name of cluster variable
- \( e(method) \) requested estimation method
- \( e(wtype) \) weight type corresponding to \( e(chi2ct) \)
- \( e(wexp) \) weight expression
- \( e(predict) \) program used to implement \texttt{predict}
- \( e(clusvar) \) name of cluster variable
- \( e(chi2type) \) Wald or LR; type of model \( \chi^2 \) test
- \( e(chi2ct) \) Wald or LR; type of model \( \chi^2 \) test corresponding to \( e(chi2c) \)
- \( e(hazard) \) variable containing hazard
- \( e(opt) \) type of optimization

Matrices
- \( e(b) \) coefficient vector
- \( e(V) \) variance–covariance matrix of the estimators

Functions
- \( e(sample) \) marks estimation sample

Methods and Formulas


The primary regression equation of interest is

\[ y_j = x_j \beta + \delta z_j + \epsilon_j \]

where \( z_j \) is a binary decision variable. The binary variable is assumed to stem from an unobservable latent variable

\[ z^*_j = w_j \gamma + u_j \]

The decision to obtain the treatment is made according to the rule

\[ z_j = \begin{cases} 
1, & \text{if } z^*_j > 0 \\
0, & \text{otherwise}
\end{cases} \]
where \( \epsilon \) and \( u \) are bivariate normal with mean zero and covariance matrix

\[
\begin{bmatrix}
\sigma & \rho \\
\rho & 1
\end{bmatrix}
\]

The likelihood function for this model is given in Maddala (1983, 122). Greene (2000, 180) discusses the standard method of reducing a bivariate normal to a function of a univariate normal and the correlation \( \rho \). Combining the two yields the following log likelihood for observation \( j \):

\[
l_j = \begin{cases} 
\ln \Phi \left( \frac{w_j \gamma + \left(y_j - x_j \beta - \delta\rho / \sigma \right)}{\sqrt{1 - \rho^2}} \right) - \frac{1}{2} \left( \frac{y_j - x_j \beta - \delta}{\sigma} \right)^2 - \ln(\sqrt{2\pi}\sigma) & z_j = 1 \\
\ln \Phi \left( \frac{-w_j \gamma - \left(y_j - x_j \beta\rho / \sigma \right)}{\sqrt{1 - \rho^2}} \right) - \frac{1}{2} \left( \frac{y_j - x_j \beta}{\sigma} \right)^2 - \ln(\sqrt{2\pi}\sigma) & z_j = 0
\end{cases}
\]

where \( \Phi() \) is the distribution function of the standard normal distribution.

In the maximum likelihood estimation, \( \sigma \) and \( \rho \) are not directly estimated. Directly estimated are \( \ln\sigma \) and \( \text{atanh} \rho \), where

\[
\text{atanh} \rho = \frac{1}{2} \ln \left( \frac{1 + \rho}{1 - \rho} \right)
\]

The standard error of \( \lambda = \rho\sigma \) is approximated through the propagation of error (delta) method, which is given by

\[
\text{Var}(\lambda) \approx D \text{Var}(\text{atanh} \rho \ \ln\sigma) D'
\]

where \( D \) is the Jacobian of \( \lambda \) with respect to \( \text{atanh} \rho \) and \( \ln\sigma \).

Maddala (1983, 120–122), also derives the two-step estimator. In the first stage, one obtains probit estimates of the treatment equation

\[
\Pr(z_j = 1 \mid w_j) = \Phi(w_j \gamma)
\]

From these estimates the hazard, \( h_j \), for each observation \( j \) is computed as

\[
h_j = \begin{cases} 
\frac{\phi(w_j \gamma)}{\Phi(w_j \gamma)} & z_j = 1 \\
\frac{-\phi(w_j \gamma)}{1 - \Phi(w_j \gamma)} & z_j = 0
\end{cases}
\]

where \( \phi \) is the standard normal density function. We also define

\[
d_j = h_j (h_j + \hat{\gamma} w_j)
\]

Then,

\[
\begin{align*}
E[y_i \mid z_i] &= X_i \beta + \delta z_i + \rho \sigma h_j \\
\text{Var}[y_i \mid z_i] &= \sigma^2 (1 - \rho^2 d_j)
\end{align*}
\]

The two-step parameter estimates of \( \beta \) and \( \delta \) are obtained by augmenting the regression equation with the hazard \( h \). Thus, the regressors become \([X \ z \ h] \) and we obtain the additional parameter estimate \( \hat{\beta}_h \) on the variable containing the hazard. A consistent estimate of the regression disturbance variance is obtained using the residuals from the augmented regression and the parameter estimate on the hazard

\[
\hat{\sigma}^2 = \frac{e' e + \hat{\beta}_h^2 \sum_{j=1}^N d_j}{N}
\]

The two-step estimate of \( \rho \) is then

\[
\hat{\rho} = \frac{\hat{\beta}_h}{\hat{\sigma}}
\]
We will now describe how the consistent estimates of the coefficient covariance matrix based on the augmented regression are derived. Let \( A = [X \, Z \, h] \) and \( D \) be a square diagonal matrix of rank \( N \) with \( (1 - \hat{\phi}^2 d_j) \) on the diagonal elements.

\[
V_{\text{wrestep}} = \hat{\sigma}^2(A'A)^{-1}(A'DA + Q)(A'A)^{-1}
\]

where

\[
Q = \hat{\rho}^2(A'DA)V_p(A'DA)
\]

and \( V_p \) is the variance–covariance estimate from the probit estimation of the treatment equation.

**Reference**


This insert focuses on two different specifications of the uniform layer effect model: the additive model and the multiplicative model. The additive model has been proposed by Yamaguchi (1987) and can be formulated as

$$\log(F_{ijk}) = \lambda + \lambda_i^R + \lambda_j^C + \lambda_k^L + \lambda_{ij}^{RL} + \lambda_{ik}^{CL} + \lambda_{ij}^{RC} + i\beta_k$$

where the $\lambda$ and $\beta$ parameters are subject to appropriate constraints that make them identifiable. As we can see, the additive model retains the two-way interaction term but replaces the three-way interaction term with the product $i\beta_k$. This means that the conditional log-odds ratios pertaining to each layer $k$ take the parametric form

$$\log(\theta_{ijk}) = \lambda_{ij}^{RC} + \lambda_{(i+1)(j+1)}^{RC} - \lambda_{(i+1)j}^{RC} - \lambda_{ij}^{RC} + \beta_k$$

Hence, in the additive model the $\beta$ parameters express the extent to which the strength of the association between variables $R$ and $C$ varies across layers. More precisely, the difference between any pair of $\beta$ parameters (say $\beta_k$ and $\beta_{k^*}$) expresses in absolute terms how much the $R$ by $C$ association is uniformly stronger or weaker in layer $k$ than in layer $k^*$ (Goodman and Hout 1998, 184). Formally,

$$\delta_{k-k^*} = \log(\theta_{ijk}) - \log(\theta_{ijk^*}) = \beta_k - \beta_{k^*}, \quad i = 1, \ldots, I - 1, \ j = 1, \ldots, J - 1$$

It should be noted that because of the presence of the $ij$ product in the equation, the results produced by the additive model depend on the ordering of both the row and column categories (Goodman and Hout 1998, 184).

Sometimes the layers can be assigned exogenous scores that have a theoretical meaning (Yamaguchi 1987, 486). In such cases, the additive model can be reformulated as

$$\log(F_{ijk}) = \lambda + \lambda_i^R + \lambda_j^C + \lambda_k^L + \lambda_{ij}^{RL} + \lambda_{ik}^{CL} + \lambda_{ij}^{RC} + \sum_{v=1}^{V} i\beta_v$$

where $v$ indexes the $V$ exogenous scores assigned to the layers, $S_{vk}$ denotes the value taken on by score $v$ in layer $k$, and $\beta_v$ denotes the linear effect exerted by score $v$ on the log-odds ratios. Thus, according to this version of the additive model, which I will refer to as the linear additive model, the difference between any pair of conditional log-odds ratios pertaining to layers $k$ and $k^*$ is equal to

$$\delta_{k-k^*} = \sum_{v=1}^{V} (S_{vk} - S_{v{k^*}}) \beta_v, \quad i = 1, \ldots, I, \ j = 1, \ldots, J$$

It should be noted that in most cases, to ensure both identification and meaningfulness of the $\beta$ parameters, it is required that $V \leq K - 2$.

The multiplicative model has been proposed by Xie (1992), see also Erikson and Goldthorpe (1992, 91–93) and can be formulated as

$$\log(F_{ijk}) = \lambda + \lambda_i^R + \lambda_j^C + \lambda_k^L + \lambda_{ij}^{RL} + \lambda_{ik}^{CL} + \lambda_{ij}^{RC} + \psi_{ijk}\phi_k$$

where the $\lambda$, $\psi$, and $\phi$ parameters are subject to appropriate constraints that make them identifiable. As we can see, the multiplicative model replaces both the two-way interaction term $\lambda_{ij}^{RC}$ and the three-way interaction term $\lambda_{ij}^{RC}$ with the product $\psi_{ijk}\phi_k$, where $\psi_{ijk}$ denotes cell-specific scores that express the baseline pattern of association between variables $R$ and $C$, and $\phi_k$ denotes layer-specific scores that express the strength of the $R$ by $C$ association in each layer. The $\psi$ and $\phi$ parameters can be seen as latent scores estimated from the data using iterative procedures (Xie 1992, 382; Goodman and Hout 1998, 181–182).

The formula for the multiplicative model implies that the conditional log-odds ratios pertaining to each layer $k$ take the parametric form

$$\log(\theta_{ijk}) = (\psi_{ij} + \psi_{(i+1)(j+1)} - \psi_{(i+1)j} - \psi_{ij}) \phi_k$$

Consequently, in the multiplicative model the ratio between any pair of $\phi$ parameters (say $\phi_k$ and $\phi_{k^*}$) expresses in relative terms how much the association between variables $R$ and $C$ is uniformly stronger or weaker in layer $k$ than in layer $k^*$ (Goodman and Hout 1998, 185). Formally,

$$\delta_{k/k^*} = \log(\theta_{ijk}) / \log(\theta_{ijk^*}) = \phi_k / \phi_{k^*}, \quad i = 1, \ldots, I - 1, \ j = 1, \ldots, J - 1$$

Both the additive and the multiplicative uniform layer effect models have been originally devised to compare social mobility tables across countries or over time. However, both models can be applied to any research question where the $R$ by $C$ association is assumed to have the same pattern but possibly different strengths across layers (see Xie 1991, Goodman and Hout 1998).
Syntax

```
unidiff cellvar, row(rowvar) column(colvar) layer(layvar) effect(null | add | addlin | mult)
    pattern(f| | qpm | qs | cp | ua | re | ce | rce | hrce | own1 | own2) quasi
    design(varlist) scores(varlist) extra(varlist) refcat(#) constraints(numlist)
    lambda(rawlog | rawexp | stdlog | stdexp) shd(log | exp) saveexp(newvar)
    savelambda(newvar) nodetail nodisproc nodispectra
```

Description

`unidiff` estimates the null, additive, linear additive, and multiplicative uniform layer effect models, displays relevant goodness-of-fit statistics and parameter estimates, and optionally computes several ancillary quantities of interest. The dataset to be analyzed must include at least four variables:

- `cellvar` contains the observed cell frequencies that make up the three-way contingency table object of analysis.
- `rowvar` indexes (and optionally labels) the \(I\) categories of the row variable.
- `colvar` indexes (and optionally labels) the \(J\) categories of the column variable.
- `layvar` indexes (and optionally labels) the \(K\) categories of the layer variable.

Note that `unidiff` drops without warning all variables starting with `rc_`.

Options

`row(rowvar)` is required. It specifies the name of the row variable.

`column(colvar)` is required. It specifies the name of the column variable.

`layer(layvar)` is required. It specifies the name of the layer variable.

`effect(null | add | addlin | mult)` is required. It specifies the type of uniform layer effect model to be estimated.

- `effect(null)` estimates the null effect model, that is, a model that postulates constant pattern and strength of the \(R\) by \(C\) association across layers.
- `effect(add)` estimates the additive model.
- `effect(addlin)` estimates the linear additive model.
- `effect(mult)` estimates the multiplicative model.

`pattern(f| | qpm | qs | cp | ua | re | ce | rce | hrce | own1 | own2)` is required. It specifies the baseline pattern of association between variables \(R\) and \(C\), that is, the form taken by the two-way interaction term \(\lambda_{ij}^{RC}\) or, in the case of the multiplicative model, by the \(\psi_{ij}\) parameters. Some patterns are allowed only when \(I = J\). For details on all these patterns of association, see Hout (1983).

- `pattern(f)` specifies the “full interaction” (saturated) pattern of association.
- `pattern(qpm)` specifies the “quasi-perfect mobility” pattern of association. It is allowed only when \(I = J\).
- `pattern(qs)` specifies the “quasi-symmetry” pattern of association. It is allowed only when \(I = J\).
- `pattern(cp)` specifies the “crossing parameters” pattern of association. It is allowed only when \(I = J\).
- `pattern(ua)` specifies the “uniform association” pattern of association.
- `pattern(re)` specifies the “row effects” pattern of association.
- `pattern(ce)` specifies the “column effects” pattern of association.
- `pattern(rce)` specifies the “row and column effects \(I\)” pattern of association.
- `pattern(hrce)` specifies the “homogeneous row and column effects \(I\)” pattern of association. It is allowed only when \(I = J\).
- `pattern(own1)` specifies a user-defined pattern of association expressed by a single “topological,” that is, categorical variable.
- `pattern(own2)` specifies a user-defined pattern of association expressed by one or more quantitative variables.
analysis is to detect any cross-national variation in the
The pattern of association between father’s class and son’s class is assumed to be constant across countries. The purpose of the
father’s occupational class (the row variable), son’s occupational class (the column variable), and country (the layer variable).
respectively, the additive and the multiplicative-uniform layer effect models. It is a 5
Example 1

In this first example, I reanalyze the social mobility data used by Yamaguchi (1987) and Xie (1992) in their illustration of,
association pattern.

. use example1.dta, clear
. describe
Let us start with the additive model. To reproduce Yamaguchi's (1987) results, two assumptions must be taken into account. First, occupational classes are ordered hierarchically along a vertical status dimension ranging from upper nonmanual (highest) to farm (lowest). Second, models are applied to off-diagonal cells only, due to the particular meaning that diagonal cells have in mobility table analysis. This means that diagonal cells must be "blocked," that is, their frequencies must be exactly reproduced by the models. To this aim, we must create and include in the models as "extra" variables, $15 (= I \times K)$ indicator variables, one for each diagonal cell of each layer.

```
. local COUNTRY = "US GB JA"
. local i=1
  while `i'<6 {  
    2. local ITEM = word `i' of `COUNTRY'
    3. local j=1
    4. while `j'<5 {  
      5. generate diag`i'`j'=country==`i' & father==`j' & son==`j'
      6. lab var diag`i'`j' "ITEM: Immobility in class `j'"
      7. local j=`j'+1
      8. }
    9. local i=`i'+1
  10. }
. describe
```

In his analysis, Yamaguchi (1987) tests several specifications of the pattern of association between father’s class and son’s
For illustration purposes, I will focus on two of them; the “full interaction” pattern and the “homogeneous row and column effects” pattern. The additive model with full interaction pattern of the $R$ by $C$ association and “blocked” diagonal cells can be estimated by

```
. unidiff obs, row(father) col(son) effect(add) pattern(fi)
> extra(diag1 - diag35)
```

**Analysis of differences in two-way associations**

**Table structure**

<table>
<thead>
<tr>
<th>Name</th>
<th>Label</th>
<th>N. of categories</th>
</tr>
</thead>
<tbody>
<tr>
<td>Row</td>
<td>father</td>
<td>5</td>
</tr>
<tr>
<td>Column</td>
<td>son</td>
<td>5</td>
</tr>
<tr>
<td>Layer</td>
<td>country</td>
<td>3</td>
</tr>
</tbody>
</table>

**Model specification**

- Layer effect: additive
- $R$-$C$ association pattern: full interaction
- Additional variables: diag1 diag2 diag3 diag4 diag5 diag6 diag7 diag8 diag9 diag10 diag11 diag12 diag13 diag14 diag15 diag16 diag17 diag18 diag19 diag20 diag21 diag22 diag23 diag24 diag25 diag26 diag27 diag28 diag29 diag30 diag31 diag32 diag33 diag34 diag35

**Goodness-of-fit statistics**

<table>
<thead>
<tr>
<th>Model</th>
<th>N</th>
<th>df</th>
<th>X2</th>
<th>p</th>
<th>G2</th>
<th>p</th>
<th>rG2</th>
<th>BIC</th>
<th>DI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cond. ind.</td>
<td>2887</td>
<td>48</td>
<td>6659.6</td>
<td>0.00</td>
<td>5591.5</td>
<td>0.00</td>
<td>0.0</td>
<td>5098.5</td>
<td>16.0</td>
</tr>
<tr>
<td>Null effect</td>
<td>2887</td>
<td>22</td>
<td>36.2</td>
<td>0.03</td>
<td>36.2</td>
<td>0.03</td>
<td>99.4</td>
<td>-189.7</td>
<td>0.9</td>
</tr>
<tr>
<td>Additive effect</td>
<td>2887</td>
<td>20</td>
<td>30.7</td>
<td>0.06</td>
<td>30.7</td>
<td>0.06</td>
<td>99.5</td>
<td>-174.7</td>
<td>0.7</td>
</tr>
</tbody>
</table>

**Beta parameters**

<table>
<thead>
<tr>
<th>Country</th>
<th>estimate</th>
<th>s.e.</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>United States</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>Britain</td>
<td>0.0035</td>
<td>0.0147</td>
<td>0.8133</td>
</tr>
<tr>
<td>Japan</td>
<td>-0.0411</td>
<td>0.0180</td>
<td>0.0227</td>
</tr>
</tbody>
</table>

**$R$-$C$ association parameters**

<table>
<thead>
<tr>
<th>Variable</th>
<th>Label</th>
<th>estimate</th>
<th>s.e.</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>rc_f12</td>
<td>Full interaction: level 2</td>
<td>-3.1828</td>
<td>0.2629</td>
<td>0.0000</td>
</tr>
<tr>
<td>rc_f13</td>
<td>Full interaction: level 3</td>
<td>-3.1345</td>
<td>0.2618</td>
<td>0.0000</td>
</tr>
<tr>
<td>rc_f14</td>
<td>Full interaction: level 4</td>
<td>-3.0926</td>
<td>0.2598</td>
<td>0.0000</td>
</tr>
<tr>
<td>rc_f15</td>
<td>Full interaction: level 5</td>
<td>-3.3612</td>
<td>0.2348</td>
<td>0.0000</td>
</tr>
<tr>
<td>rc_f16</td>
<td>Full interaction: level 6</td>
<td>-3.2003</td>
<td>0.2598</td>
<td>0.0000</td>
</tr>
<tr>
<td>rc_f17</td>
<td>Full interaction: level 7</td>
<td>-1.6144</td>
<td>0.2494</td>
<td>0.0000</td>
</tr>
<tr>
<td>rc_f18</td>
<td>Full interaction: level 8</td>
<td>-2.3159</td>
<td>0.2541</td>
<td>0.0000</td>
</tr>
<tr>
<td>rc_f19</td>
<td>Full interaction: level 9</td>
<td>-2.8686</td>
<td>0.2285</td>
<td>0.0000</td>
</tr>
<tr>
<td>rc_f20</td>
<td>Full interaction: level 10</td>
<td>-3.0470</td>
<td>0.2584</td>
<td>0.0000</td>
</tr>
<tr>
<td>rc_f21</td>
<td>Full interaction: level 11</td>
<td>-2.0611</td>
<td>0.2552</td>
<td>0.0000</td>
</tr>
<tr>
<td>rc_f22</td>
<td>Full interaction: level 12</td>
<td>-1.6946</td>
<td>0.2556</td>
<td>0.0000</td>
</tr>
<tr>
<td>rc_f23</td>
<td>Full interaction: level 13</td>
<td>-2.7080</td>
<td>0.2631</td>
<td>0.0000</td>
</tr>
<tr>
<td>rc_f24</td>
<td>Full interaction: level 14</td>
<td>-2.9281</td>
<td>0.2545</td>
<td>0.0000</td>
</tr>
<tr>
<td>rc_f25</td>
<td>Full interaction: level 15</td>
<td>-1.8064</td>
<td>0.2493</td>
<td>0.0000</td>
</tr>
<tr>
<td>rc_f26</td>
<td>Full interaction: level 16</td>
<td>-1.3609</td>
<td>0.2445</td>
<td>0.0000</td>
</tr>
<tr>
<td>rc_f27</td>
<td>Full interaction: level 17</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
</tbody>
</table>

**Extra variable parameters**

<table>
<thead>
<tr>
<th>Variable</th>
<th>Label</th>
<th>estimate</th>
<th>s.e.</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>diag1</td>
<td>US: Mobility in class 1</td>
<td>3.8151</td>
<td>0.2524</td>
<td>0.0000</td>
</tr>
<tr>
<td>diag2</td>
<td>US: Mobility in class 2</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>diag3</td>
<td>US: Mobility in class 3</td>
<td>-0.5863</td>
<td>0.1402</td>
<td>0.0000</td>
</tr>
<tr>
<td>diag4</td>
<td>US: Mobility in class 4</td>
<td>0.0235</td>
<td>0.0690</td>
<td>0.7334</td>
</tr>
<tr>
<td>diag5</td>
<td>US: Mobility in class 5</td>
<td>0.2605</td>
<td>0.2189</td>
<td>0.2540</td>
</tr>
<tr>
<td>diag6</td>
<td>GB: Mobility in class 1</td>
<td>4.0496</td>
<td>0.2775</td>
<td>0.0000</td>
</tr>
<tr>
<td>diag7</td>
<td>GB: Mobility in class 2</td>
<td>0.1899</td>
<td>0.1024</td>
<td>0.0696</td>
</tr>
</tbody>
</table>
As we can see, the output consists of seven items:

- A description of the structure of the contingency table object of analysis.
- A description of the specification of the fitted model. The output of these first two items can be suppressed by specifying the option `nodetail`.
- A table reporting goodness-of-fit statistics for both the main model (in this case the additive model) and two benchmark models: the conditional independence model and the null effect model (see above). In this table, \(N\) denotes the total number of observations, \(df\) the residual degrees of freedom, \(X^2\) the Pearson chi-squared statistic (with corresponding \(p\)-value), \(G^2\) the likelihood-ratio chi-squared statistic (with corresponding \(p\)-value), \(rG^2\) the percent reduction in \(G^2\) compared to the conditional independence model, \(BIC\) the Bayesian information criterion, and \(DI\) the dissimilarity index. For more details on these measures, see the Methods and Formulas section below.
- A table reporting the maximum likelihood estimates (with corresponding standard errors and \(p\)-values) of the parameters.
- A table reporting the maximum likelihood estimates (with corresponding standard errors and \(p\)-values) of the parameters associated with the variables that express the \(R\) by \(C\) association pattern. The output of this table can be suppressed by specifying the option `nodisprc`.
- A table reporting the maximum likelihood estimates (with corresponding standard errors and \(p\)-values) of the parameters associated with the extra variables. The output of this table can be suppressed by specifying the option `nodispextra`.
- A table reporting kappa indices, which express in standardized form the strength of the \(R\) by \(C\) association within each layer (Hout, et al. 1995, 813; Goodman 1991, 1089). For more details on the kappa index, see the Methods and Formulas section below.

Yamaguchi (1987) estimates a second version of this model which constrains the beta parameters for the United States and Great Britain to be equal. To this aim, we use the option `constraints` as follows:

```
. unidiff obs row(father) col(son) lay(country) effect(add) pattern(fi)
> extra(diag1-diag35) constraints(1 1 2) nodetail nodisprc nodispextra
```

```
Analysis of differences in two-way associations
Goodness-of-fit statistics

<table>
<thead>
<tr>
<th>Model</th>
<th>(N)</th>
<th>(df)</th>
<th>(X^2)</th>
<th>(p)</th>
<th>(G^2)</th>
<th>(p)</th>
<th>(rG^2)</th>
<th>(BIC)</th>
<th>(DI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cond. indep.</td>
<td>28887</td>
<td>48</td>
<td>6659.6</td>
<td>0.00</td>
<td>5951.5</td>
<td>0.00</td>
<td>0</td>
<td>5098.5</td>
<td>16.0</td>
</tr>
<tr>
<td>Null effect</td>
<td>28887</td>
<td>22</td>
<td>36.2</td>
<td>0.03</td>
<td>36.2</td>
<td>0.03</td>
<td>99.4</td>
<td>-189.7</td>
<td>0.9</td>
</tr>
<tr>
<td>Additive effect</td>
<td>28887</td>
<td>21</td>
<td>30.8</td>
<td>0.06</td>
<td>30.8</td>
<td>0.06</td>
<td>99.4</td>
<td>-184.9</td>
<td>0.7</td>
</tr>
</tbody>
</table>

Beta parameters

<table>
<thead>
<tr>
<th>Country</th>
<th>estimate</th>
<th>s.e.</th>
<th>(p)-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>United States</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>Britain</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
</tr>
<tr>
<td>Japan</td>
<td>-0.0417</td>
<td>0.0176</td>
<td>0.0192</td>
</tr>
</tbody>
</table>
```
Kappa indices
------------------
Country | Kappa
---------|--------
United States | 0.55
Britain | 0.70
Japan | 0.51

Let us consider now the "homogeneous row and column effects" specification of the pattern of association between father’s class and son’s class. To specify this pattern within unidiff, we have

```
. unidiff obs, row(father) col(son) lay(country) effect(add) pattern(hrce)  
> extra(diag1-diag35) nodispext
```

Analysis of differences in two-way associations

Table structure
-----------------------------
Name | Label | N of categories
-----|-------|------------------
Row | father | Father’s occupational class | 5
Column | son | Son’s occupational class | 5
Layer | country | Country | 3

Model specification
-----------------------------
Layer effect: additive
R-C association pattern: homogeneous row & column effects
Additional variables: diag1 diag2 diag3 diag4 diag5 diag6 diag7 diag8 diag9 diag10 diag11 diag12 diag13 diag14 diag15 diag16 diag17 diag18 diag19 diag20 diag21 diag22 diag23 diag24 diag25 diag26 diag27 diag28 diag29 diag30 diag31 diag32 diag33 diag34 diag35

Goodness-of-fit statistics
-----------------------------
Model | N | df | X2 | p | G2 | p | rG2 | BIC | DI
-----|---|---|----|---|----|---|-----|-----|-----
Cond. indep. | 28887 | 48 | 6659.6 | 0.00 | 5591.5 | 0.00 | 0.0 | 5098.5 | 16.0
Null effect | 28887 | 29 | 125.5 | 0.00 | 107.0 | 0.00 | 96.1 | -190.9 | 1.3
Additive effect | 28887 | 27 | 117.2 | 0.00 | 98.4 | 0.00 | 98.2 | -179.0 | 1.2

Beta parameters
-----------------------------
Country | estimate | s.e. | p-value
--------|---------|------|--------
United States | 0.0000 | 0.0000 | 0.0000
Britain | -0.0025 | 0.0148 | 0.8657
Japan | -0.0524 | 0.0178 | 0.0034

R-C association parameters
-----------------------------
Variable | Label | estimate | s.e. | p-value
----------|-------|---------|------|--------
rc_rc2 | Row-Column effect 2 | 0.1012 | 0.0329 | 0.0021
rc_rc3 | Row-Column effect 3 | 0.2729 | 0.0250 | 0.0000
rc_rc4 | Row-Column effect 4 | 0.3333 | 0.0258 | 0.0000
rc_rc5 | Row-Column effect 5 | 0.3333 | 0.0227 | 0.0000

Kappa indices
------------------
Country | Kappa
---------|--------
United States | 0.62
Britain | 0.77
Japan | 0.62

The multiplicative version of the uniform layer effect model with full interaction pattern of the $R$ by $C$ association and “blocked” diagonal cells can be estimated by

```
. unidiff obs, row(father) col(son) lay(country) effect(mult) pattern(fi)  
> extra(diag1-diag35) nodispext
```
Iteration 1: deviance = 53.4755  
Iteration 2: deviance = 35.3391  
Iteration 3: deviance = 0.6855  
Iteration 4: deviance = 0.0206  
Iteration 5: deviance = 0.0099  
Iteration 6: deviance = 0.0000

Analysis of differences in two-way associations

Table structure

<table>
<thead>
<tr>
<th>Name</th>
<th>Label</th>
<th>N. of categories</th>
</tr>
</thead>
<tbody>
<tr>
<td>Row</td>
<td>father</td>
<td>5</td>
</tr>
<tr>
<td>Column</td>
<td>son</td>
<td>5</td>
</tr>
<tr>
<td>Layer</td>
<td>country</td>
<td>3</td>
</tr>
</tbody>
</table>

Model specification

Layer effect: multiplicative  
R-C association pattern: full interaction  
Additional variables: diag11 diag12 diag13 diag14 diag15 diag21 diag22 diag23 diag24 diag25 diag31 diag32 diag33 diag34 diag35

Goodness-of-fit statistics

<table>
<thead>
<tr>
<th>Model</th>
<th>N</th>
<th>df</th>
<th>X2</th>
<th>p</th>
<th>G2</th>
<th>p</th>
<th>rG2</th>
<th>BIC</th>
<th>DI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cond. indep.</td>
<td>26987</td>
<td>48</td>
<td>6659.6</td>
<td>0.00</td>
<td>5891.5</td>
<td>0.00</td>
<td>99.4</td>
<td>-159.7</td>
<td>0.9</td>
</tr>
<tr>
<td>Null effect</td>
<td>26987</td>
<td>22</td>
<td>36.2</td>
<td>0.00</td>
<td>36.2</td>
<td>0.03</td>
<td>99.4</td>
<td>-159.7</td>
<td>0.9</td>
</tr>
<tr>
<td>Multipl. effect</td>
<td>26987</td>
<td>20</td>
<td>30.7</td>
<td>0.06</td>
<td>30.9</td>
<td>0.06</td>
<td>99.4</td>
<td>-174.5</td>
<td>0.7</td>
</tr>
</tbody>
</table>

Phi parameters (layer scores)

<table>
<thead>
<tr>
<th>Country</th>
<th>Raw</th>
<th>Scaled 1</th>
<th>Scaled 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>United States</td>
<td>1.7025</td>
<td>1.0000</td>
<td>0.6064</td>
</tr>
<tr>
<td>Britain</td>
<td>1.7703</td>
<td>1.0398</td>
<td>0.6305</td>
</tr>
<tr>
<td>Japan</td>
<td>1.3605</td>
<td>0.7991</td>
<td>0.4845</td>
</tr>
</tbody>
</table>

Psi parameters (R-C association scores)

<table>
<thead>
<tr>
<th>Father's occupational class</th>
<th>Son's occupational class</th>
<th>UpNonM</th>
<th>LowNon</th>
<th>UpManu</th>
<th>LowMan</th>
<th>Farm</th>
</tr>
</thead>
<tbody>
<tr>
<td>UpNonManual</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td></td>
</tr>
<tr>
<td>LowNonManual</td>
<td>0.00</td>
<td>0.69</td>
<td>0.51</td>
<td>0.54</td>
<td>0.37</td>
<td></td>
</tr>
<tr>
<td>UpManual</td>
<td>0.00</td>
<td>0.48</td>
<td>1.14</td>
<td>0.99</td>
<td>0.64</td>
<td></td>
</tr>
<tr>
<td>LowManual</td>
<td>0.00</td>
<td>0.57</td>
<td>1.14</td>
<td>1.35</td>
<td>0.73</td>
<td></td>
</tr>
<tr>
<td>Farm</td>
<td>0.00</td>
<td>0.64</td>
<td>1.29</td>
<td>1.55</td>
<td>2.93</td>
<td></td>
</tr>
</tbody>
</table>

Kappa indices

<table>
<thead>
<tr>
<th>Country</th>
<th>Kappa</th>
</tr>
</thead>
<tbody>
<tr>
<td>United States</td>
<td>0.55</td>
</tr>
<tr>
<td>Britain</td>
<td>0.70</td>
</tr>
<tr>
<td>Japan</td>
<td>0.51</td>
</tr>
</tbody>
</table>

As we can see, the output includes two new items:

- A table reporting the maximum likelihood estimates of the $\phi$ parameters (layer scores). Three series of $\phi$ parameters are reported: raw estimates, estimates rescaled so that $\phi_r = 1$, and estimates rescaled so that $\sum_{k=1}^{K} \phi_k^2 = 1$ (see Xie 1992, 382).
- A table reporting the maximum likelihood estimates of the $\psi$ parameters (R by C association scores).
Example 2

To illustrate the estimation of the linear additive uniform layer effect model, in this second example I make use of the sixteen-country social mobility data originally assembled by Hazelrigg and Garnier (1976) and subsequently analyzed by several researchers (Grusky and Hauser 1984, Xie 1992). It is a $3 \times 3 \times 16$ contingency table that cross-classifies father’s occupational class (the row variable), son’s occupational class (the column variable), and country (the layer variable). As in the previous example, the pattern of association between father’s class and son’s class is assumed to be constant across countries. The purpose of the analysis is to estimate the effect exerted by some country-level variables on the strength of that association. Following Hauser and Grusky (1988), four variables have been selected: degree of economic development (measured as per capita energy consumption in tons of coal), degree of social democracy (measured as percentage of seats in the national legislature held by social democratic parties), a dummy variable indicating countries belonging to the Eastern block, and a dummy variable indicating Asian countries (for details, see Hauser and Grusky 1988).

To begin, let us use `unidiff` to replicate Xie’s (1992) application of the multiplicative model to the sixteen-country social mobility data:

```
    . use example2.dta, clear
    . describe
    Contains data from example2.dta
    obs: 144
    vars: 8 27 Dec 1999 10:09
    size: 2,736 (%8.5% of memory free)

    1. obs int 9.0g Observed cell frequencies
    2. father byte 9.0g class Father’s occupational class
    3. son byte 9.0g class Son’s occupational class
    4. country byte 13.0g country Country
    5. develop float 9.0g Economic development index
    6. socdem float 9.0g Social democracy index
    7. east byte 9.0g Eastern block country
    8. asia byte 9.0g Asian country

    Sorted by:
    . label list
      class:
      1 NonManual
      2 Manual
      3 Farm
      country:
      1 Australia
      2 Belgium
      3 France
      4 Hungary
      5 Italy
      6 Japan
      7 Philippines
      8 Spain
      9 United States
     10 West Germany
     11 West Malaysia
     12 Yugoslavia
     13 Denmark
     14 Finland
     15 Norway
     16 Sweden
```

```
    . unidiff obs, row(father) col(son) lay(country) effect(mult) pattern(fi)
    Iteration 1:  deviance = 378.0271
    Iteration 2:  deviance = 67.2765
    Iteration 3:  deviance = 6.9561
    Iteration 4:  deviance = 0.5449
    Iteration 5:  deviance = 0.0430
    Iteration 6:  deviance = 0.0029
    Iteration 7:  deviance = 0.0010
    Iteration 8:  deviance = 0.0000
```
### Analysis of differences in two-way associations

#### Table structure

<table>
<thead>
<tr>
<th>Name</th>
<th>Label</th>
<th>N. of categories</th>
</tr>
</thead>
<tbody>
<tr>
<td>Row father</td>
<td>Father’s occupational class</td>
<td>3</td>
</tr>
<tr>
<td>Column son</td>
<td>Son’s occupational class</td>
<td>3</td>
</tr>
<tr>
<td>Layer country</td>
<td>Country</td>
<td>16</td>
</tr>
</tbody>
</table>

#### Model specification

- **Layer effect:** multiplicative
- **R-C association pattern:** full interaction
- **Additional variables:** none

#### Goodness-of-fit statistics

<table>
<thead>
<tr>
<th>Model</th>
<th>$N$</th>
<th>$df$</th>
<th>$X^2$</th>
<th>$p$</th>
<th>$G^2$</th>
<th>$p$</th>
<th>$rG^2$</th>
<th>BIC</th>
<th>DI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cond. indep.</td>
<td>113556</td>
<td>64</td>
<td>43389.7</td>
<td>0.00</td>
<td>42970.0</td>
<td>0.00</td>
<td>0.0</td>
<td>4225.0</td>
<td>25.6</td>
</tr>
<tr>
<td>Full effect</td>
<td>113556</td>
<td>60</td>
<td>1327.3</td>
<td>0.00</td>
<td>1328.8</td>
<td>0.00</td>
<td>96.9</td>
<td>630.4</td>
<td>3.7</td>
</tr>
</tbody>
</table>

#### Phi parameters (layer scores)

<table>
<thead>
<tr>
<th>Country</th>
<th>Raw</th>
<th>Scaled 1</th>
<th>Scaled 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Australia</td>
<td>6.7022</td>
<td>1.0000</td>
<td>0.2170</td>
</tr>
<tr>
<td>Belgium</td>
<td>9.1682</td>
<td>1.3679</td>
<td>0.2988</td>
</tr>
<tr>
<td>France</td>
<td>8.6107</td>
<td>1.2848</td>
<td>0.2788</td>
</tr>
<tr>
<td>Hungary</td>
<td>7.3955</td>
<td>1.1333</td>
<td>0.2459</td>
</tr>
<tr>
<td>Italy</td>
<td>9.2504</td>
<td>1.3802</td>
<td>0.2995</td>
</tr>
<tr>
<td>Japan</td>
<td>7.1213</td>
<td>1.0625</td>
<td>0.2306</td>
</tr>
<tr>
<td>Philippines</td>
<td>7.3397</td>
<td>1.0951</td>
<td>0.2376</td>
</tr>
<tr>
<td>Spain</td>
<td>9.1393</td>
<td>1.3636</td>
<td>0.2999</td>
</tr>
<tr>
<td>United States</td>
<td>7.3490</td>
<td>1.0865</td>
<td>0.2379</td>
</tr>
<tr>
<td>West Germany</td>
<td>6.8884</td>
<td>1.0233</td>
<td>0.2230</td>
</tr>
<tr>
<td>West Malaysia</td>
<td>6.1284</td>
<td>0.9144</td>
<td>0.1984</td>
</tr>
<tr>
<td>Yugoslavia</td>
<td>6.9040</td>
<td>1.0301</td>
<td>0.2235</td>
</tr>
<tr>
<td>Denmark</td>
<td>8.6761</td>
<td>1.2945</td>
<td>0.2809</td>
</tr>
<tr>
<td>Finland</td>
<td>6.9970</td>
<td>1.0440</td>
<td>0.2265</td>
</tr>
<tr>
<td>Norway</td>
<td>6.0621</td>
<td>0.9045</td>
<td>0.1983</td>
</tr>
<tr>
<td>Sweden</td>
<td>8.5000</td>
<td>1.2682</td>
<td>0.2752</td>
</tr>
</tbody>
</table>

#### Psi parameters (R-C association scores)

<table>
<thead>
<tr>
<th>Father’s occupational class</th>
<th>Son’s occupational class</th>
</tr>
</thead>
<tbody>
<tr>
<td>NonManual</td>
<td>Manual</td>
</tr>
<tr>
<td>NonManual</td>
<td>0.00</td>
</tr>
<tr>
<td>Manual</td>
<td>0.00</td>
</tr>
<tr>
<td>Farm</td>
<td>0.00</td>
</tr>
</tbody>
</table>

#### Kappa indices

<table>
<thead>
<tr>
<th>Country</th>
<th>Kappa</th>
</tr>
</thead>
<tbody>
<tr>
<td>Australia</td>
<td>0.61</td>
</tr>
<tr>
<td>Belgium</td>
<td>0.83</td>
</tr>
<tr>
<td>France</td>
<td>0.78</td>
</tr>
<tr>
<td>Hungary</td>
<td>0.69</td>
</tr>
<tr>
<td>Italy</td>
<td>0.84</td>
</tr>
<tr>
<td>Japan</td>
<td>0.64</td>
</tr>
<tr>
<td>Philippines</td>
<td>0.66</td>
</tr>
<tr>
<td>Spain</td>
<td>0.83</td>
</tr>
<tr>
<td>United States</td>
<td>0.66</td>
</tr>
<tr>
<td>West Germany</td>
<td>0.62</td>
</tr>
<tr>
<td>West Malaysia</td>
<td>0.55</td>
</tr>
<tr>
<td>Yugoslavia</td>
<td>0.62</td>
</tr>
</tbody>
</table>
In his analysis, Xie (1992) explores the effect exerted by four country-level variables (partially different from ours) on the strength of the father-son association by computing zero-order correlation coefficients between those variables and the $\phi$ parameters estimated by the multiplicative model. Alternatively, we can estimate the effect of country-level explanatory variables by means of the linear additive uniform layer effect model.

```
.unidiff obs, row(father) col(son) lay(country) effect(addlin) pattern(fi)
> scores(develop socdem east asia)
```

Analysis of differences in two-way associations

Table structure

<table>
<thead>
<tr>
<th>Name</th>
<th>Label</th>
<th>N. of categories</th>
</tr>
</thead>
<tbody>
<tr>
<td>Row</td>
<td>father</td>
<td>Father’s occupational class</td>
</tr>
<tr>
<td>Column</td>
<td>son</td>
<td>Son’s occupational class</td>
</tr>
<tr>
<td>Layer</td>
<td>country</td>
<td>Country</td>
</tr>
</tbody>
</table>

Model specification

Layer effect: linear additive
Layer score variables: develop socdem east asia
R-C association pattern: full interaction
Additional variables: none

Goodness-of-fit statistics

<table>
<thead>
<tr>
<th>Model</th>
<th>N</th>
<th>df</th>
<th>$X^2$</th>
<th>p</th>
<th>$G^2$</th>
<th>p</th>
<th>$rG^2$</th>
<th>BIC</th>
<th>DI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cond. indep.</td>
<td>113556</td>
<td>64</td>
<td>43380.7</td>
<td>0.00</td>
<td>42970.0</td>
<td>0.00</td>
<td>0.0</td>
<td>42225.0</td>
<td>25.6</td>
</tr>
<tr>
<td>Null effect</td>
<td>113556</td>
<td>60</td>
<td>1327.3</td>
<td>0.00</td>
<td>1326.8</td>
<td>0.00</td>
<td>0.0</td>
<td>96.9</td>
<td>630.4</td>
</tr>
<tr>
<td>Lin. add. effect</td>
<td>113556</td>
<td>56</td>
<td>1014.6</td>
<td>0.00</td>
<td>1005.2</td>
<td>0.00</td>
<td>0.0</td>
<td>97.7</td>
<td>353.4</td>
</tr>
</tbody>
</table>

Beta parameters

<table>
<thead>
<tr>
<th>Variable</th>
<th>Label</th>
<th>estimate</th>
<th>s.e.</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>develop</td>
<td>Economic development index</td>
<td>-0.0068</td>
<td>0.0037</td>
<td>0.1165</td>
</tr>
<tr>
<td>socdem</td>
<td>Social democracy index</td>
<td>-0.0961</td>
<td>0.0005</td>
<td>0.0000</td>
</tr>
<tr>
<td>east</td>
<td>Eastern block country</td>
<td>0.2025</td>
<td>0.0329</td>
<td>0.0000</td>
</tr>
<tr>
<td>asia</td>
<td>Asian country</td>
<td>-0.2327</td>
<td>0.0220</td>
<td>0.0000</td>
</tr>
</tbody>
</table>

Layer scores

<table>
<thead>
<tr>
<th>Country</th>
<th>develop</th>
<th>socdem</th>
<th>east</th>
<th>asia</th>
</tr>
</thead>
<tbody>
<tr>
<td>Australia</td>
<td>4.80</td>
<td>39.50</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Belgium</td>
<td>4.73</td>
<td>34.90</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>France</td>
<td>2.95</td>
<td>11.60</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Hungary</td>
<td>2.81</td>
<td>0.00</td>
<td>1.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Italy</td>
<td>1.79</td>
<td>18.60</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Japan</td>
<td>1.78</td>
<td>35.70</td>
<td>0.00</td>
<td>1.00</td>
</tr>
<tr>
<td>Philippines</td>
<td>0.21</td>
<td>0.00</td>
<td>0.00</td>
<td>1.00</td>
</tr>
<tr>
<td>Spain</td>
<td>1.02</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>United States</td>
<td>9.20</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>West Germany</td>
<td>4.23</td>
<td>37.40</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>West Malaysia</td>
<td>0.36</td>
<td>8.20</td>
<td>0.00</td>
<td>1.00</td>
</tr>
<tr>
<td>Yugoslavia</td>
<td>1.19</td>
<td>0.00</td>
<td>1.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Denmark</td>
<td>4.17</td>
<td>44.20</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Finland</td>
<td>2.69</td>
<td>26.50</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Norway</td>
<td>3.59</td>
<td>49.30</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Sweden</td>
<td>4.51</td>
<td>48.30</td>
<td>0.00</td>
<td>0.00</td>
</tr>
</tbody>
</table>

R-C association parameters
As we can see from the table reporting the $\beta$ parameters, the strength of the association between father’s class and son’s class decreases as both economic development and social democracy increase. Moreover, the father-son association is, *coeteris paribus*, stronger in the countries belonging to the Eastern block and weaker in the Asian countries. It is important to stress that, given the bad fit of the model, these results should be considered only for pedagogic purposes.

### Saved Results

`unidiff` saves in `r()`:

Scalars

- `r(di)`: dissimilarity index
- `r(r22)`: reduction in $G^2$
- `r(G2)`: $G^2$
- `r(X2)`: $X^2$
- `r(N)`: number of observations
- `r(ncol)`: number of cells
- `r(ncells)`: number of categories of column variable
- `r(ncells)`: number of categories of row variable
- `r(nlay)`: number of categories of layer variable
- `r(df)`: residual degrees of freedom
- `r(mcells)`: number of cells
- `r(mcol)`: number of categories of column variable

Macros

- `r(cellvar)`: name of variable containing the observed cell frequencies
- `r(rowvar)`: name of row variable
- `r(colvar)`: name of column variable
- `r(layvar)`: name of layer variable
- `r(effect)`: type of uniform layer effect
- `r(pattern)`: type of $R$ by $C$ association pattern
- `r(extra)`: list of extra variables

### Methods and Formulas

Let $f_{ijk}$ denote the observed frequency in cell $(i, j, k)$, $F_{ijk}$ denote the expected frequency in cell $(i, j, k)$ under the fitted model, $N$ denote the total number of observations, and $df$ denote the residual degrees of freedom under the fitted model. The Pearson chi-squared statistic is

$$X^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K} \frac{(f_{ijk} - E_{ijk})^2}{E_{ijk}}$$

The likelihood-ratio chi-squared statistic is

$$G^2 = 2 \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K} f_{ijk} \log \left( \frac{f_{ijk}}{E_{ijk}} \right)$$
The percent reduction in $G^2$ is

$$rG^2 = (1 - \frac{G^2_{M1}}{G^2_{M0}}) \times 100$$

where $G^2_{M1}$ denotes the likelihood-ratio chi-squared statistic associated with the fitted model, and $G^2_{M0}$ denotes the likelihood-ratio chi-squared statistic associated with the conditional independence model.

The Bayesian information criterion is

$$BIC = G^2 - df \times \log(N)$$

The dissimilarity index is

$$\Delta = \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K} \frac{|f_{ijk} - F_{ijk}|}{2N} \times 100$$

The raw total interaction effects estimated with option `lambda(rawlog)` are

$$\Lambda_{tijk} = \begin{cases} 
\lambda_{ij}^{RC} + \sum_{t=1}^{T} \xi_t x_{tijk}, & \text{for the null model} \\
\lambda_{ij}^{RC} + ij \beta_k + \sum_{t=1}^{T} \xi_t x_{tijk}, & \text{for the additive model} \\
\lambda_{ij}^{RC} + \sum_{v=1}^{V} i j S_{vk} \beta_v + \sum_{t=1}^{T} \xi_t x_{tijk}, & \text{for the linear additive model} \\
\psi_{ij} \phi_k + \sum_{t=1}^{T} \xi_t x_{tijk}, & \text{for the multiplicative model}
\end{cases}$$

where the two-way terms $\lambda_{ij}^{RC}$ and $\psi_{ij}$ are parameterized according to the selected pattern of the $R$ by $C$ association.

The standardized total interaction effects estimated with option `lambda(stdlog)` satisfy the following conditions:

$$\sum_{t=1}^{T} \hat{\Lambda}_{tijk} = 0, \quad j = 1, \ldots, J, k = 1, \ldots, K$$

$$\sum_{j=1}^{J} \hat{\Lambda}_{tijk} = 0, \quad i = 1, \ldots, I, k = 1, \ldots, K$$

The layer-specific kappa indices are

$$\kappa_k = \sqrt{\sum_{i=1}^{I} \sum_{j=1}^{J} \frac{\hat{\Lambda}_{tijk}^2}{IJ}}, \quad k = 1, \ldots, K$$

The structural shift parameters estimated with option `shd(log)` are

$$\log(\alpha_{jk}) = (\lambda_{ij}^C + \lambda_{jk}^{CL}) - (\lambda_{ij}^C + \lambda_{jk}^{CL}), \quad j = 1, \ldots, J, k = 1, \ldots, K$$

where $\log(\alpha_{11k}) = 0$ for identification purposes.

The structural distances estimated with option `shd(log)` are

$$\log(\alpha_{(j/j^*)k}) = \log(\alpha_{jk}) - \log(\alpha_{j^*k}), \quad j, j^* = 1, \ldots, J, k = 1, \ldots, K$$

The mean structural distances estimated with option `shd(log)` are

$$\log(\alpha_{(j/j^*)[k]}) = \frac{\sum_{j^*=1}^{J} \log(\alpha_{(j/j^*)[k]})}{J - 1}, \quad j = 1, \ldots, J, k = 1, \ldots, K$$

The overall structural effects estimated with option `shd(log)` are

$$\log(\alpha_k) = \frac{\sum_{j=1}^{J} \sum_{j^*=1}^{J} \log(|\alpha_{(j/j^*)[k]}|)}{(J \times J) - J}, \quad k = 1, \ldots, K$$
References


Roger Newson, Guy’s, King’s and St Thomas’ School of Medicine, London, UK, roger.newson@kcl.ac.uk

Abstract: Rank order or so-called nonparametric methods are in fact based on population parameters, which are zero under the null hypothesis. Two of these parameters are Kendall’s $\tau_a$ and Somers’ $D$, the parameter tested by a Wilcoxon rank-sum test. Confidence limits for these parameters are more informative than $p$-values alone, for three reasons. Firstly, confidence intervals show that a high $p$-value does not prove a null hypothesis. Secondly, for continuous data, Kendall’s $\tau_a$ can often be used to define robust confidence limits for Pearson’s correlation by Greiner’s relation. Thirdly, we can define confidence limits for differences between two Kendall’s $\tau_a$’s or Somers’ $D$’s, and these are informative, because a larger Kendall’s $\tau_a$ or Somers’ $D$ cannot be secondary to a smaller one. The program `somerst` calculates confidence intervals for Somers’ $D$ or Kendall’s $\tau_a$, using jackknife variances. There is a choice of transformations, including Fisher’s $z$, Daniels’ arcsine, Greiner’s $\rho$, and the $z$-transform of Greiner’s $\rho$. A `cluster` option is available. The estimation results are saved as for a model fit, so that differences can be estimated using `lincom`.

Keywords: Somers’ $D$, Kendall’s $\tau_a$, rank correlation, rank-sum test, Wilcoxon test, confidence intervals, nonparametric methods.

Syntax

```
somersd varlist [weight] [if exp] [in range], clustervarname level(#) tsum tdist 
transf(transformation_name) 
```

where `transformation_name` is one of

- `iden` | `z` | `asin` | `rho` | `zrho`

`fweights`, `iweights` and `pweights` are allowed.

Description

`somersd` calculates the nonparametric statistics Somers’ $D$ (corresponding to rank-sum tests) and Kendall’s $\tau_a$, with confidence limits. Somers’ $D$ or $\tau_a$ is calculated for the first variable of `varlist` as a predictor of each of the other variables in `varlist`, with estimates and jackknife variances and confidence intervals output and saved in `e()` as if for the parameters of a model fit. It is possible to use `lincom` to output confidence limits for differences between the population Somers’ $D$ or Kendall’s $\tau_a$ values.
The matched-pairs rank-sum test represents the case where there are paired data and a two-sample rank-sum test represents the case where two subpopulations. Kendall’s \( \tau \) and Somers’ \( D \) are binary variables indicating membership of one of two subpopulations. Kendall’s \( \tau \) is usually tested on “continuous” data, using the corresponding sample statistics \( \hat{\tau}_{XY} \) and \( \hat{D}_{XY} \). These are commonly known as nonparametric statistics, even though confidence intervals to \( p \)-values alone:

1. Nonstatisticians often quote a nonsignificant result for a nonparametric test and argue as if they have “proved” a null hypothesis, when a confidence interval would show a wide range of other hypotheses which also fit the data.

2. In the case of continuous bivariate data, there is a correspondence between Kendall’s \( \tau \) and the more familiar Pearson’s correlation coefficient \( \rho \), known as Greiner’s relation (Kendall 1970). This states that

\[
\rho = \sin \left( \frac{\pi}{2} \tau \right)
\]

and holds if the joint distribution of \( X \) and \( Y \) is bivariate normal. Under this relation, Kendall’s \( \tau \)-values of 0, \( \pm \frac{1}{2} \), and \( \pm 1 \) correspond to Pearson’s correlations of 0, \( \pm \frac{1}{2} \), \( \pm 1 \) and \( \pm 1 \), respectively. A similar correspondence is likely to hold in a wider range of continuous bivariate distributions (Kendall 1949, Newson 1987).

3. Kendall’s \( \tau \) has the desirable property that a larger \( \tau \) cannot be secondary to a smaller \( \tau \), that is, if a positive \( \tau_{XY} \) is caused entirely by a monotonic positive relationship of both variables with a third variable \( W \), then \( \tau_{WX} \) and \( \tau_{WY} \) must
both be greater than \( \tau_{XY} \). If we can show that \( \tau_{XY} - \tau_{YW} > 0 \) (or, equivalently, that \( D_{YX} - D_{YW} > 0 \), then this implies that the correlation between \( X \) and \( Y \) is not caused entirely by the influence of \( W \).

To understand the third point, assume that trivariate data points \((W_i, X_i, Y_i)\) are sampled independently from a common population, with discrete probability mass function \( f_{W,X,Y}(\cdot, \cdot, \cdot) \) and marginal probability mass function \( f_{W,X}(\cdot, \cdot) \). Define the conditional expectation

\[
Z(w_1, x_1, w_2, x_2) = E[\text{sign}(Y_2 - Y_1) | W_1 = w_1, X_1 = x_1, W_2 = w_2, X_2 = x_2]
\]

(4)

for any \( w_1 \) and \( w_2 \) in the range of \( W \)-values and any \( x_1 \) and \( x_2 \) in the range of \( X \)-values. If we state that the positive relationship between \( X_i \) and \( Y_i \) is caused entirely by a monotonic positive relationship between both variables and \( W_i \), then that is equivalent to stating that

\[
Z(w_1, x_1, w_2, x_2) \geq 0
\]

whenever \( w_1 \leq w_2 \) and \( x_2 \leq x_1 \). However, the difference between the two \( \tau_a \) coefficients is

\[
\tau_{YW} - \tau_{XY} = 2 \left\{ \sum_w \sum_{x_2 < x_1} f_{W,X}(w, x_1)f_{W,X}(w, x_2)Z(w, x_1, x, x_2) + 2 \sum_w \sum_{x_1 < w} f_{W,X}(w_1, x)f_{W,X}(w_2, x)Z(w_1, x, w_2, x) + 4 \sum_{w_1 < w_2} \sum_{x_2 < x_1} f_{W,X}(w_1, x_1)f_{W,X}(w_2, x_2)Z(w_1, x_1, w_2, x_2) \right\}
\]

(6)

This difference must be nonnegative whenever the inequality (5) applies. In particular, if the distribution of the \( W_i \) and \( X_i \) is nearly continuous, then the difference (6) will be dominated by the third term, representing discordant \((W_i, X_i)\)-pairs. The difference between \( \tau_a \)-values will then be determined by the ordering of the \( Y \)-values when the larger of two \( W \)-values is associated with the smaller of two \( X \)-values.

We now define the formulas for estimating \( \tau_{XY}, D_{YX} \) and their differences. We assume the general case where the observations are clustered, which becomes the familiar unclustered case when there is one observation per cluster. Suppose there are \( n \) clusters, and the \( i \)th cluster contains \( m_i \) observations. Define \( w_{hi}, X_{hi} \) and \( Y_{hi} \) to be the importance weight, \( X \)-value and \( Y \)-value, respectively, for the \( i \)th observation of the \( i \)th cluster. (Like most estimation commands, somerad treats \( fweights \) and \( pweights \) as importance weights, and treats \( fweights \) as if they denoted a number of identical observations.) Define

\[
v_{hijk} = \begin{cases} w_{hi}w_{jk}, & h \neq j \\ 0, & h = j \end{cases}
\]

(7)

and any other sums over any other indices. Given that the clusters are sampled independently from a common population of clusters, we can define

\[
V = E[v_{h,j}], \quad T_{XY} = E[t_{h,j}^{(XY)}]
\]

(9)

for all \( h \neq j \). (In the terminology of Hoeffding (1948), these quantities are regular functionals of the cluster population distribution, and the expressions inside the square brackets are kernels of these regular functionals.) The quantities we really want to estimate are Kendall’s \( \tau_a \) and Somers’ \( D \), defined respectively by

\[
\tau_{XY} = T_{XY}/V, \quad D_{YX} = T_{XY}/T_{XX} = \tau_{XY}/\tau_{XX}
\]

(10)

(These are equal to the familiar formulas (1) and (2) if each cluster contains one observation with an importance weight of one.) To estimate these, we use the jackknife method of Arvesen (1969) on the regular functionals (9) and use appropriate Taylor polynomials. The functionals \( V \) and \( T_{XY} \) are estimated by the Hoeffding (1948) \( U \)-statistics

\[
\hat{V} = \frac{v_{...,}}{n(n-1)}, \quad \hat{T}_{XY} = \frac{t_{X,...}^{(XY)}}{n(n-1)}
\]

(11)
and the respective jackknife pseudovalues corresponding to the $h$th cluster are given by
\[
\psi^{(V)}_h = (n - 1)^{-1} v_\ldots - (n - 2)^{-1} [v_\ldots - 2v_{h\ldots}]
\]
\[
\psi^{(XY)}_h = (n - 1)^{-1} t^{(XY)}_\ldots - (n - 2)^{-1} [t^{(XY)}_\ldots - 2t^{(XY)}_{h\ldots}]
\]
(12)

`somersd` calculates correlation measures for a single variable $X$ with a set of $Y$-variate $(Y^{(1)}, \ldots, Y^{(p)})$. It calculates, in the first instance, the covariance matrix for $\hat{V}$, $\hat{T}_{XX}$, and $\hat{T}_{XY(i)}$ for $1 \leq i \leq p$. This is done using the jackknife influence matrix $\mathcal{T}$, which has $n$ rows labeled by the cluster subscripts, and $p + 2$ columns labeled (in Stata fashion) by the names $V$, $X$, and $Y^{(i)}$ for $1 \leq i \leq p$. It is defined by
\[
\mathcal{T}[h,V] = \psi^{(V)}_h - \hat{V}, \quad \mathcal{T}[h,X] = \psi^{(X)}_h - \hat{T}_{XX}, \quad \mathcal{T}[h,Y^{(i)}] = \psi^{(Y^{(i)})}_h - \hat{T}_{XY^{(i)}}
\]
(13)
The jackknife covariance matrix is then equal to
\[
\hat{C} = [n(n - 1)]^{-1} \mathcal{T}'\mathcal{T}
\]
(14)
The estimates for Kendall’s $\tau_a$ and Somers’ $D$ are defined by
\[
\widehat{\tau}_{XY} = \hat{T}_{XY}/\hat{V}, \quad \widehat{D}_{XY} = \hat{T}_{XY}/\hat{T}_{XX}
\]
(15)
and the covariance matrices are defined using Taylor polynomials. In the case of Somers’ $D$, we define the $p \times (p + 2)$ matrix of estimated derivatives $\hat{\Gamma}^{(D)}$, whose rows are labeled by the names $Y^{(1)}, \ldots, Y^{(p)}$, and whose columns are labeled by $V, X, Y^{(1)}, \ldots, Y^{(p)}$. This matrix is defined by
\[
\hat{\Gamma}^{(D)}[Y^{(i)}, X] = \frac{\partial \hat{D}_{XY}}{\partial \hat{T}_{XX}} = \frac{\hat{T}_{XY}}{\hat{T}_{XX}^2}
\]
\[
\hat{\Gamma}^{(D)}[Y^{(i)}, Y^{(i)}] = \frac{\partial \hat{D}_{XY}}{\partial \hat{T}_{XY}} = \frac{1}{\hat{T}_{XX}}
\]
(16)
all other entries being zero. In the case of Kendall’s $\tau_a$, we define a $(p + 1) \times (p + 2)$ matrix of estimated derivatives $\hat{\Gamma}^{(\tau)}$, whose rows are labeled by $X, Y^{(1)}, \ldots, Y^{(p)}$, and whose columns are labeled by $V, X, Y^{(1)}, \ldots, Y^{(p)}$. This matrix is defined by
\[
\hat{\Gamma}^{(\tau)}[X, V] = \frac{\partial \hat{\tau}_{XX}}{\partial \hat{V}} = \frac{\hat{T}_{XX}}{\hat{V}^2}
\]
\[
\hat{\Gamma}^{(\tau)}[X, X] = \frac{\partial \hat{\tau}_{XX}}{\partial \hat{T}_{XX}} = \frac{1}{\hat{V}}
\]
\[
\hat{\Gamma}^{(\tau)}[Y^{(i)}, V] = \frac{\partial \hat{\tau}_{XY}}{\partial \hat{V}} = \frac{\hat{T}_{XY}}{\hat{V}^2}
\]
\[
\hat{\Gamma}^{(\tau)}[Y^{(i)}, Y^{(i)}] = \frac{\partial \hat{\tau}_{XY^{(i)}}}{\partial \hat{T}_{XY^{(i)}}} = \frac{1}{\hat{V}}
\]
(17)
all other entries again being zero. The estimated dispersion matrices of the Somers’ $D$ and $\tau_a$ estimates are therefore $\hat{C}^{(D)}$ and $\hat{C}^{(\tau)}$, respectively, defined by
\[
\hat{C}^{(D)} = \hat{\Gamma}^{(D)} \hat{C} \hat{\Gamma}^{(D)\top}, \quad \hat{C}^{(\tau)} = \hat{\Gamma}^{(\tau)} \hat{C} \hat{\Gamma}^{(\tau)\top}
\]
(18)
The `transform` option offers a choice of transformations. Since these are available both for Somers’ $D$ and for Kendall’s $\tau_a$, we will denote the original estimate as $\theta$ (which can stand for $D$ or $\tau$) and the transformed estimate as $\zeta$. They are summarized below, together with their derivatives $d\zeta/d\theta$ and their inverses $\theta(\zeta)$.

<table>
<thead>
<tr>
<th><code>transform</code></th>
<th>Transform name</th>
<th>$\zeta(\theta)$</th>
<th>$d\zeta/d\theta$</th>
<th>$\theta(\zeta)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>iden</td>
<td>Untransformed</td>
<td>$\theta$</td>
<td>1</td>
<td>$\zeta$</td>
</tr>
<tr>
<td>z</td>
<td>Fisher’s $z$</td>
<td>$\text{arctanh}(\theta)$</td>
<td>$(1-\theta^2)^{-1/2}$</td>
<td>$\text{tanh}(\zeta)$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\frac{1}{2}\log([1+\theta]/[1-\theta])$</td>
<td></td>
<td>$\exp(2\zeta)-1/[\exp(2\zeta)+1]$</td>
</tr>
<tr>
<td>asin</td>
<td>Daniels’ arcsine</td>
<td>$\text{arcsin}(\theta)$</td>
<td>$(1-\theta^2)^{-1/2}$</td>
<td>$\sin(\zeta)$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$\frac{1}{2}\text{cos}(\theta)^2$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>rho</td>
<td>Greiner’s $\rho$</td>
<td>$\text{sin}(\frac{\theta}{2})$</td>
<td>$(2/\pi)\text{arcsin}(\zeta)$</td>
<td>$\frac{1}{2}\text{cos}(\frac{\theta}{2})$</td>
</tr>
<tr>
<td>zrho</td>
<td>Greiner’s $\rho$</td>
<td>$\text{arctanh}[\text{sin}(\frac{\theta}{2})]$</td>
<td>$(2/\pi)\text{arcsin}(\text{tanh}(\zeta))$</td>
<td></td>
</tr>
</tbody>
</table>
If trans is specified, then somersd displays and saves the transformed estimates and their estimated covariance, instead of the untransformed versions. If \( \tilde{C}(\hat{\theta}) \) is the covariance matrix for the untransformed estimates given by (18), and \( \tilde{F}(\tilde{C}) \) is the diagonal matrix whose diagonal entries are the \( d\zeta/d\theta \) estimates specified in the table, then the transformed parameter and its covariance matrix are

\[
\tilde{\zeta} = \zeta(\hat{\theta}), \quad \tilde{C}(\tilde{\zeta}) = \tilde{F}(\tilde{C}) \tilde{C}(\theta) \tilde{F}(\tilde{C}),
\]

(19)

Fisher’s z-transform was originally recommended for the Pearson correlation coefficient by Fisher (1921) (see also Gayen 1951), but Edwards (1995) recommended it specifically for Somers’ \( D \) on the basis of simulation studies. Daniels’ arcsine was suggested as a normalizing transform in Daniels and Kendall (1947). If \( \text{transf}(z) \) or \( \text{transf}(\text{asin}) \) is specified, then somersd prints asymmetric confidence intervals for the untransformed \( \zeta \) or \( \theta \) values, calculated from symmetric confidence intervals for the transformed parameters using the inverse function \( \theta(z) \). (This feature corresponds to the eform option of other estimation commands.) Greiner’s \( r \) (Kendall 1970) is based on the relation (3), and is designed to estimate the Pearson correlation coefficient corresponding to the measured \( \zeta \). If \( \text{transf}(z\rho) \) is specified, somersd prints asymmetric confidence intervals for Greiner’s \( r \), using the inverse \( z \)-transform on symmetric confidence intervals for the \( z \)-transformed Greiner’s \( r \).

**Example 1**

In the auto data, we compare US cars with foreign cars regarding weight and fuel efficiency. First, we use ranksum to give significance tests without confidence intervals:

```
. ranksum mpg, by(foreign)
Two-sample Wilcoxon rank-sum (Mann-Whitney) test

<table>
<thead>
<tr>
<th></th>
<th>obs</th>
<th>rank sum</th>
<th>expected</th>
</tr>
</thead>
<tbody>
<tr>
<td>Domestic</td>
<td>52</td>
<td>1688.5</td>
<td>1950</td>
</tr>
<tr>
<td>Foreign</td>
<td>22</td>
<td>1096.5</td>
<td>825</td>
</tr>
</tbody>
</table>

combined | 74  | 2775.00  | 2775.00  |
unadjusted variance | 7150.00
adjustment for ties | -.56.95
adjusted variance | 7113.05

Ho: mpg(foreign==Domestic) = mpg(foreign==Foreign)
z = -3.101
Prob > |z| = 0.0019
```

```
. ranksum weight, by(foreign)
Two-sample Wilcoxon rank-sum (Mann-Whitney) test

<table>
<thead>
<tr>
<th></th>
<th>obs</th>
<th>rank sum</th>
<th>expected</th>
</tr>
</thead>
<tbody>
<tr>
<td>Domestic</td>
<td>52</td>
<td>2379.5</td>
<td>1950</td>
</tr>
<tr>
<td>Foreign</td>
<td>22</td>
<td>395.5</td>
<td>825</td>
</tr>
</tbody>
</table>

combined | 74  | 2775.00  | 2775.00  |
unadjusted variance | 7150.00
adjustment for ties | -1.06
adjusted variance | 7148.94

Ho: weight(foreign==Domestic) = weight(foreign==Foreign)
z = 5.680
Prob > |z| = 0.0000
```

We note that American cars are typically heavier and travel fewer miles per gallon than foreign cars. For confidence intervals, we use somersd:

```
. somersd foreign mpg weight
Somers’ D Transformation: Untransformed
Valid observations: 74

|     | Jackknife | Coef. | Std. Err. | z    | P>|z|  | 96% Conf. Interval |
|-----|-----------|-------|-----------|------|------|------------------|
|mpg | .135146   | .0032485 | -9.020 | 0.000 | -91.0363 | -.56771          |
|weight | - .7506741 | .0032485 | -9.020 | 0.000 | -91.0363 | -.56771          |
```

We see that, given a randomly-chosen foreign car and a randomly-chosen American car, the foreign car is 46% more likely to travel more miles per gallon than the American car than vice versa, with confidence limits from 19% to 72% more
likely. However, being foreign seems to be more reliable as a negative predictor of weight than as a positive predictor of “fuel efficiency”. We can use \texttt{lincom} to define confidence limits for the difference:

\begin{verbatim}
  . lincom -mpg - mpg = 0.0
  ( 1)  - mpg - weight = 0.0

  foreign |     Coef.  Std. Err.     z   P>|z|   [95% Conf. Interval]
  -------------+---------------------------------+-----------------------------
       (1) |  3.321     0.001       3.321  0.001       0.1298677    0.4670449
  -------------+---------------------------------+-----------------------------

  foreign |     Coef.  Std. Err.     z   P>|z|   [95% Conf. Interval]
  -------------+---------------------------------+-----------------------------
       (1) |  3.2858    0.4018       8.202  0.000      2.496961    4.074639
  -------------+---------------------------------+-----------------------------

  The difference between Somers’ $D$-values is positive. This indicates that, if there are two cars, one heavier and consuming fewer gallons per mile, the other lighter and consuming more gallons per mile, then the second is more likely to be foreign. So maybe 1970’s American cars were not as wasteful as some people think, and were, if anything, more fuel-efficient for their weight than non-American cars at the time. Figure 1 illustrates this graphically. Data points are domestic cars (“D”) and foreign cars (“F”). A regression analysis could show the same thing, but Somers’ $D$ shows it in stronger terms, without contentious assumptions such as linearity. (On the other hand, a regression model is more informative if its assumptions are true, so the two methods are mutually complementary.)

\begin{figure}[h]
  \centering
  \includegraphics[width=\textwidth]{figure1.png}
  \caption{Applying \texttt{somersd} to the \texttt{auto} data.}
  \end{figure}

The confidence intervals for such high values of Somers’ $D$ would probably be more reliable if we used the $z$-transform, recommended by Edwards (1995). The results of this are as follows:

\begin{verbatim}
  . somersd foreign mpg weight, tran(z)
  Somers’ $D$ Transformation: Fisher’s $z$
  Valid observations: 74
  ---------------------------------+---------------------------------+-----------------------------
  |     Coef.  Std. Err.     z   P>|z|   [95% Conf. Interval]
  ---------------------------------+---------------------------------+-----------------------------
  foreign |  0.4937249  0.1708551    2.890  0.004      0.188551    0.8288947
  ---------------------------------+---------------------------------+-----------------------------
  mpg | -0.9749561  0.1908547   -5.108  0.000      -1.349024    -0.6008786
  ---------------------------------+---------------------------------+-----------------------------

  95% CI for untransformed Somers’ $D$
  Somers’$D$  Minimum  Maximum
  mpg  0.45716783  0.51753197  0.67972072
  weight  -0.73057413  -0.67361282  -0.53768968
  . lincom -weight -mpg
  ( 1)  - mpg - weight = 0.0
  ---------------------------------+---------------------------------+-----------------------------
  |     Coef.  Std. Err.     z   P>|z|   [95% Conf. Interval]
  ---------------------------------+---------------------------------+-----------------------------
       (1) |  3.2858    0.4018       8.202  0.000      2.496961    4.074639
  ---------------------------------+---------------------------------+-----------------------------

  Note that \texttt{somersd} gives not only symmetric confidence limits for the $z$-transformed Somers’ $D$ estimates, but also the more informative asymmetric confidence limits for the untransformed Somers’ $D$ estimates (corresponding to the \texttt{eform} option). The asymmetric confidence limits for the untransformed estimates are closer to zero than the symmetric confidence limits for the untransformed estimates in the previous output, and are probably more realistic. The output to \texttt{lincom} gives confidence limits
\end{verbatim}
for the difference between $z$-transformed Somers’ $D$ values. This difference is expressed in $z$-units, but must, of course, be in the same direction as the difference between untransformed Somers’ $D$ values. The conclusions are similar.

**Example 2**

In this example, we demonstrate Kendall’s $\tau_a$ by comparing weight (pounds) and displacement (cubic inches) as predictors of fuel efficiency (miles per gallon). We first use `ktau` to carry out significance tests with no confidence limits:

```
. ktau mpg weight
  Number of obs =  74
  Kendall/'s tau/-a =  -0.6857
  Kendall/'s tau/-b =  -0.7059
  Kendall/'s score =  -1.852
```

Test of Ho: mpg and weight independent

```
  Pr > |z| = 0.0000 (continuity corrected)
```

We then use `somersd` (with the `taua` option and the $z$-transform) to compute the same statistics with confidence limits. Note that `somersd` also outputs the $\tau_a$ of mpg with mpg, which is simply the probability that two independently sampled mpg-values are not equal.

```
. somersd mpg weight displ/,taua tr/(z/)
```

We can use `lincom` to compare the two predictors and test whether smaller and heavier cars travel fewer miles per gallon than larger and lighter cars. This seems to be the case, as `weight` is a more negative predictor of `mpg` than `displ`:

```
. lincom weight/-displ
( 1)  weight - displ = 0.0
```

We demonstrate the `cluster` option using the variable `manuf`, equal to the first word of `make`, and used in [U] 23.11 Obtaining robust variance estimates to denote manufacturer. This analysis assumes that we are sampling from the population of car manufacturers rather than the population of car models. The results are as follows:

```
. somersd mpg weight displ/,taua tr/(z/) cluster(manuf)
```

(standard errors adjusted for clustering on manuf)
Note that, in contrast to the case of most estimation commands, the `cluster' option affects the estimates as well as their standard errors. This is because the clustered estimates are calculated only from between-cluster comparisons, in this case pairs of car models from different manufacturers.

Suppose that we are writing for an audience more familiar with Pearson’s correlation than with Kendall’s \( \tau_a \). To estimate the Pearson correlations corresponding to our \( \tau_a \) coefficients, we use the \( \text{zrho} \) transform. The results are as follows:

```
. somersd mpg weight displ, taua trans(zrho)
Valid observations: 74

|   | Coef. | Std. Err. | z    | P>|z| | [95% Conf. Interval] |
|---|-------|-----------|------|------|---------------------|
| mpg | 3.17952 | 1.458796 | 21.796 | 0.000 | 2.893692 - 3.465439 |
| weight | -1.378273 | .1475561 | -9.341 | 0.000 | -1.667478 - -1.089069 |
| displ | -1.108383 | .158893 | -6.979 | 0.000 | -1.420262 - -0.7974132 |
```

The \(-59\%\) \( \tau_a \) between displacement and fuel efficiency (from the unclustered output) is seen to correspond to a more impressive \(-80\%\) Pearson correlation. The estimated Greiner’s \( \rho \) is probably less likely to be oversensitive to outliers than the usual Pearson coefficient.

### Saved Results

`somersd` saves in `e()`:

- **Scalars**
  - `e(N)` number of observations
  - `e(N_cluster)` number of clusters
- **Macros**
  - `e(cmd)` `somersd`
  - `e(par)` parameter label in output
  - `e(par_label)` name of parameter
  - `e(vartype)` covariance estimation method
  - `e(transf)` transformation specified by `transf`
- **Matrices**
  - `e(b)` coefficient vector
  - `e(V)` variance–covariance matrix of the estimators
- **Functions**
  - `e(sample)` marks estimation sample

Note that (confusingly) `e(depvar)` is the \( X \)-variable, or predictor variable, in the conventional terminology for defining Somers’ \( D \). `somersd` is also different from most estimation commands in that its results are not designed to be used by `predict`. 
References


Cumulative index for STB-49–STB-54

<table>
<thead>
<tr>
<th>Announcements</th>
</tr>
</thead>
<tbody>
<tr>
<td>STB-49 2</td>
</tr>
<tr>
<td>STB-50 2</td>
</tr>
<tr>
<td>STB-53 2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Stata Corporation updates</th>
</tr>
</thead>
<tbody>
<tr>
<td>STB-50 34</td>
</tr>
<tr>
<td>STB-54 2</td>
</tr>
<tr>
<td>STB-54 4</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Data Management</th>
</tr>
</thead>
<tbody>
<tr>
<td>STB-49 2</td>
</tr>
<tr>
<td>STB-52 2</td>
</tr>
<tr>
<td>STB-51 2</td>
</tr>
<tr>
<td>STB-51 2</td>
</tr>
<tr>
<td>STB-51 2</td>
</tr>
<tr>
<td>STB-53 3</td>
</tr>
<tr>
<td>STB-49 2</td>
</tr>
<tr>
<td>STB-49 6</td>
</tr>
<tr>
<td>STB-50 3</td>
</tr>
<tr>
<td>STB-51 2</td>
</tr>
<tr>
<td>STB-49 7</td>
</tr>
<tr>
<td>STB-50 3</td>
</tr>
<tr>
<td>STB-50 5</td>
</tr>
<tr>
<td>STB-50 9</td>
</tr>
<tr>
<td>STB-51 3</td>
</tr>
<tr>
<td>STB-51 5</td>
</tr>
<tr>
<td>STB-52 2</td>
</tr>
<tr>
<td>STB-52 2</td>
</tr>
<tr>
<td>STB-54 7</td>
</tr>
<tr>
<td>STB-52 8</td>
</tr>
<tr>
<td>STB-53 6</td>
</tr>
<tr>
<td>STB-54 8</td>
</tr>
<tr>
<td>STB-54 16</td>
</tr>
</tbody>
</table>
[gr] Graphics
STB-49 8  gr34.2  Drawing Venn diagrams
STB-54 17 gr34.3  An update to drawing Venn diagrams
STB-49 8  gr36  An extension of for, useful for graphics commands
STB-49 10 gr37  Cumulative distribution function plots
STB-50 17 gr38  Enhancement to the hilite command
STB-51 7  gr39  3D surface plots
STB-51 10 gr40  A simple contour plot
STB-51 12 gr41  Distribution function plots
STB-51 16 gr42  Quantile plots, generalized
STB-54 19 gr43  Overlaying graphs

[ip] Instruction on Programming
STB-52 9  ip18.1 Update to resample
STB-50 20 ip28 Automatically sorting by subgroup
STB-52 10 ip29 Metadata for user-written contributions to the Stata programming language
STB-54 21 ip29.1 Metadata for user-written contributions to the Stata programming language: extensions

[os] Operating system, hardware, & interprogram communication
STB-51 19 os17 Command-name registration at www.stata.com

[sbe] Biostatistics & Epidemiology
STB-49 12 sbe27 Assessing confounding effects in epidemiological studies
STB-49 15 sbe28 Meta-analysis of p-values
STB-50 21 sbe29 Generalized linear models: extensions to the binomial family
STB-51 24 sbe30 Improved confidence intervals for odds ratios
STB-52 12 sbe31 Exact confidence intervals for odds ratios from case–control studies
STB-54 23 sbe32 Automated outbreak detection from public health surveillance data

[sg] General Statistics
STB-53 17 sg35.2 Robust tests for the equality of variances: update to Stata 6
STB-49 17 sg64.1 Update to pwcorrs
STB-51 27 sg67.1 Update to univar
STB-49 17 sg81.1 Multivariable fractional polynomials: update
STB-50 25 sg81.2 Multivariable fractional polynomials: update
STB-54 25 sg84.2 Concordance correlation coefficient: update for Stata 6
STB-49 23 sg97.1 Revision of outreg
STB-49 23 sg107.1 Generalized Lorenz curves and related graphs
STB-49 24 sg111 A modified likelihood-ratio test command
STB-49 25 sg112 Nonlinear regression models involving power or exponential functions of covariates
STB-50 26 sg112.1 Nonlinear regression models involving power or exponential functions of covariates: update
STB-50 26 sg113 Tabulation of modes
STB-50 27 sg114 rglm - Robust variance estimates for generalized linear models
STB-51 28 sg115 Bootstrap standard errors for indices of inequality
STB-51 32 sg116 Hotdeck imputation
STB-54 26 sg116.1 Update to hotdeck imputation
STB-51 34 sg117 Robust standard errors for the Foster–Greer–Thorbecke class of poverty indices
STB-51 37 sg118 Partitions of Pearson’s $\chi^2$ for analyzing two-way tables that have ordered columns
STB-52 16 sg119 Improved confidence intervals for binomial proportions
STB-52 19 sg120 Receiver Operating Characteristic (ROC) analysis
STB-53 18 sg120.1 Two new options added to roccomp command
STB-54 26 sg120.2 Correction to roccomp command
STB-52 34 sg121 Seemingly unrelated estimation and the cluster-adjusted sandwich estimator
STB-52 47 sg122 Truncated regression
STB-52 52 sg123 Hodges–Lehmann estimation of a shift in location between two populations
STB-53 19 sg124 Interpreting logistic regression in all its forms
STB-53 29 sg125 Automatic estimation of interaction effects and their confidence intervals
STB-53 31 sg126 Two-parameter log-gamma and log-inverse Gaussian models
STB-53 32 sg127 Summary statistics for estimation sample
Some programs for growth estimation in fisheries biology
Generalized linear latent and mixed models
Box–Cox regression models
On the manipulability of Wald tests in Box–Cox regression models
Analysis of variance from summary statistics
Sequential and drop one term likelihood-ratio tests
Model selection using the Akaike information criterion

Analysis of multiple failure-time data with Stata

Bivariate Granger causality test

Update to random allocation of treatments to blocks
Random allocation of treatments balanced in blocks: update

Cumulative index for STB-43–STB-48
STB categories and insert codes

Inserts in the STB are presently categorized as follows:

**General Categories:**
- *an* announcements
- *cc* communications & letters
- *dm* data management
- *dt* datasets
- *gr* graphics
- *in* instruction
- *ip* instruction on programming
- *os* operating system, hardware, & interprogram communication
- *qs* questions and suggestions
- *tt* teaching
- *zz* not elsewhere classified

**Statistical Categories:**
- *sbe* biostatistics & epidemiology
- *sed* exploratory data analysis
- *sg* general statistics
- *smv* multivariate analysis
- *snp* nonparametric methods
- *sqc* quality control
- *sqv* analysis of qualitative variables
- *srdr* robust methods & statistical diagnostics
- *ssa* survival analysis
- *ssi* simulation & random numbers
- *ssv* survey sampling
- *sss* social science & psychometrics
- *stx* time-series, econometrics
- *sxd* experimental design
- *szz* not elsewhere classified

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 (979–845–3144) 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 xyz.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.

<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Address: P.O. Box 1169 17100 NAZERATH-ELLIT</td>
<td>Address: P.O. Box 2222</td>
</tr>
<tr>
<td>Israel</td>
<td>PRIMROSE 1416</td>
</tr>
<tr>
<td>Phone: +972 (0)6 6100101</td>
<td>Phone: +27-11-8286169</td>
</tr>
<tr>
<td>Fax: +972 (0)6 6554254</td>
<td>Fax: +27-11-8221377</td>
</tr>
<tr>
<td>Email: <a href="mailto:assc@netvision.net.il">assc@netvision.net.il</a></td>
<td>Email: <a href="mailto:iem@hot.co.za">iem@hot.co.za</a></td>
</tr>
<tr>
<td>Countries served: Israel</td>
<td>Countries served: South Africa, Botswana, Lesotho, Namibia, Mozambique, Swaziland, Zimbabwe</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Company: Axon Technology Company Ltd</th>
<th>Company: MercoStat Consultores</th>
</tr>
</thead>
<tbody>
<tr>
<td>Address: 9F, No. 259, Sec. 2 Ho-Ping East Road</td>
<td>Address: 9 de junio 1389 CP 11400 MONTEVIDEO</td>
</tr>
<tr>
<td>TAIPEI 106</td>
<td>Uruguay</td>
</tr>
<tr>
<td>Taiwan</td>
<td></td>
</tr>
<tr>
<td>Phone: +886-(0)2-27045535</td>
<td>Phone: 598-2-613-7905</td>
</tr>
<tr>
<td>Fax: +886-(0)2-27541785</td>
<td>Fax: Same</td>
</tr>
<tr>
<td>Email: <a href="mailto:hank@axon.axon.com.tw">hank@axon.axon.com.tw</a></td>
<td>Email: <a href="mailto:mercost@adinet.com.uy">mercost@adinet.com.uy</a></td>
</tr>
<tr>
<td>Countries served: Taiwan</td>
<td>Countries served: Uruguay, Argentina, Brazil, Paraguay</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Company: Chips Electronics</th>
<th>Company: Metrika Consulting</th>
</tr>
</thead>
<tbody>
<tr>
<td>Address: Lokasari Plaza 1st Floor Room 82 Jalan Mangga Besar Raya No. 82</td>
<td>Address: Mosstorpsvagen 48 183 30 Taby STOCKHOLM</td>
</tr>
<tr>
<td>JAKARTA</td>
<td>Sweden</td>
</tr>
<tr>
<td>Indonesia</td>
<td></td>
</tr>
<tr>
<td>Phone: 62 - 21 - 600 66 47</td>
<td>Phone: +46-708-163128</td>
</tr>
<tr>
<td>Fax: 62 - 21 - 600 66 47</td>
<td>Fax: +46-8-7924747</td>
</tr>
<tr>
<td>Email: <a href="mailto:puyuh23@indo.net.id">puyuh23@indo.net.id</a></td>
<td>Email: <a href="mailto:sales@metrika.se">sales@metrika.se</a></td>
</tr>
<tr>
<td>Countries served: Indonesia</td>
<td>URL: <a href="http://www.metrika.se">http://www.metrika.se</a></td>
</tr>
<tr>
<td>Countries served: Indonesia</td>
<td>Countries served: Sweden, Baltic States, Denmark, Finland, Iceland, Norway</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Company: Dittrich &amp; Partner Consulting</th>
<th>Company: Ritme Informatique</th>
</tr>
</thead>
<tbody>
<tr>
<td>Address: Kieler Strasse 17 5. floor</td>
<td>Address: 34, boulevard Haussmann 75009 Paris</td>
</tr>
<tr>
<td>D-42697 Solingen Germany</td>
<td>France</td>
</tr>
<tr>
<td>Phone: +49 2 12 / 26 066 - 0</td>
<td>Phone: +33 (0)1 42 46 00 42</td>
</tr>
<tr>
<td>Fax: +49 2 12 / 26 066 - 66</td>
<td>Fax: +33 (0)1 42 46 00 33</td>
</tr>
<tr>
<td>Email: <a href="mailto:sales@dpc.de">sales@dpc.de</a></td>
<td>Email: <a href="mailto:info@ritme.com">info@ritme.com</a></td>
</tr>
<tr>
<td>URL: <a href="http://www.dpc.de">http://www.dpc.de</a></td>
<td>URL: <a href="http://www.ritme.com">http://www.ritme.com</a></td>
</tr>
<tr>
<td>Countries served: Germany, Austria, Italy</td>
<td>Countries served: France, Belgium, Luxembourg</td>
</tr>
</tbody>
</table>

(List continued on next page)
<table>
<thead>
<tr>
<th>Company: Scientific Solutions S.A.</th>
<th>Company: Timberlake Consulting S.L.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Address: Avenue du Général Guisan, 5</td>
<td>Address: Calle Mendez Nunez, 1, 3</td>
</tr>
<tr>
<td>CH-1009 Pully/Lausanne Switzerland</td>
<td>41011 Sevilla Spain</td>
</tr>
<tr>
<td>Phone: 41 (0)21 711 15 20</td>
<td>Phone: +34 (9) 5 422 0648</td>
</tr>
<tr>
<td>Fax: 41 (0)21 711 15 21</td>
<td>Fax: +34 (9) 5 422 0648</td>
</tr>
<tr>
<td>Email: <a href="mailto:info@scientific-solutions.ch">info@scientific-solutions.ch</a></td>
<td>Email: <a href="mailto:timberlake@zoom.es">timberlake@zoom.es</a></td>
</tr>
<tr>
<td>Countries served: Switzerland</td>
<td>Countries served: Spain</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Address: Doormanstraat 19</td>
<td>Address: Praceta Raúl Brandao, n°1, 1ºE</td>
</tr>
<tr>
<td>5151 GM Drunen Netherlands</td>
<td>2720 ALFRAGIDE Portugal</td>
</tr>
<tr>
<td>Phone: +31 416-378 125</td>
<td>Phone: +351 (0)1 471 73 47</td>
</tr>
<tr>
<td>Fax: +31 416-378 385</td>
<td>Fax: +351 (0)1 471 73 47</td>
</tr>
<tr>
<td>Email: <a href="mailto:info@smitconsult.nl">info@smitconsult.nl</a></td>
<td>Email: <a href="mailto:timberlake.co@mail.telepac.pt">timberlake.co@mail.telepac.pt</a></td>
</tr>
<tr>
<td>URL: <a href="http://www.smitconsult.nl">http://www.smitconsult.nl</a></td>
<td></td>
</tr>
<tr>
<td>Countries served: Netherlands</td>
<td>Countries served: Portugal</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Address: 249 Eramosa Road West</td>
<td>Kadikoy 81320 ISTANBUL Turkey</td>
</tr>
<tr>
<td>Moorooduc VIC 3933 Australia</td>
<td></td>
</tr>
<tr>
<td>Phone: +61 (0)3 5978 8329</td>
<td>Phone: +90 (216) 414 19 58</td>
</tr>
<tr>
<td>Fax: +61 (0)3 5978 8623</td>
<td>Fax: +30 (216) 366 89 23</td>
</tr>
<tr>
<td>Email: <a href="mailto:sales@survey-design.com.au">sales@survey-design.com.au</a></td>
<td>Email: <a href="mailto:info@unidost.com">info@unidost.com</a></td>
</tr>
<tr>
<td>Countries served: Australia, New Zealand</td>
<td>Countries served: Turkey</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Company: Timberlake Consultants</th>
<th>Company: Vishvas Marketing-Mix Services</th>
</tr>
</thead>
<tbody>
<tr>
<td>Address: Unit B3 Broomsleigh Bus. Park</td>
<td>Address: C\O S. D. Wamorkar “Prashant” Vishnu Nagar, Naupada</td>
</tr>
<tr>
<td>Worsley Bridge Road LONDON SE26 5BN United Kingdom</td>
<td>THANE - 400602 India</td>
</tr>
<tr>
<td>Phone: +44 (0)208 697 3377</td>
<td>Phone: +91-251-440087</td>
</tr>
<tr>
<td>Fax: +44 (0)208 697 3388</td>
<td>Fax: +91-22-5378552</td>
</tr>
<tr>
<td>Email: <a href="mailto:info@timberlake.co.uk">info@timberlake.co.uk</a></td>
<td>Email: <a href="mailto:vishvas@vsnl.com">vishvas@vsnl.com</a></td>
</tr>
<tr>
<td>URL: <a href="http://www.timberlake.co.uk">http://www.timberlake.co.uk</a></td>
<td></td>
</tr>
<tr>
<td>Countries served: United Kingdom, Eire</td>
<td>Countries served: India</td>
</tr>
</tbody>
</table>