



# New and Improved Features

VERSION 5.0

Clint Cummins and Bronwyn H. Hall

© 2004 by TSP International

All Rights Reserved

## TSP Version 5.0

### New and Improved Features – Summary

#### 1. General

- dashed lists
- long names
- faster SORT

#### 2. Printing and graphics

- better formatted PRINT output
- new graphics options for density curves on histograms
- better treatment of missing and panel data in PLOT and GRAPH
- lines with symbols via SYMBOL option

#### 3. Major enhancements

- many new panel data features (random and fixed effects estimation for PROBIT, fixed effects estimation for LSQ, 3SLS, GMM, FIML, AR1, one and two-way random effects via ML, etc.)
- block diagonal HS and autocorrelation consistent standard errors for panel data
- Interval regression procedure
- Kernel density or regression estimation
- Censored quantile regression estimation
- new nonlinear second derivative approximations (numeric second derivatives, based on numeric or analytic first derivatives), new tuning parameters
- Common factor test in AR1; root moduli and extended sample ACF for Box-Jenkins routines
- Different instruments for different equations in GMM (easier specification)

#### 4. Minor enhancements

- too many to list (see below)

## TSP Version 5.0 as of June 2004

This documentation describes the new features in TSP 5.0 as well as changes made to TSP since the initial release of TSP 4.5 in June 1999.

The changes are described below in four sections: 1) syntax and general changes; 2) printing and graphics; 3) new or substantially enhanced procedures; and 4) bug fixes and minor enhancements).

### 1. Syntax and general enhancements:

**dashed lists** - Leading zeros in list names are now expanded better.

A01-A05 expands to A01 A02 A03 A04 A05 (formerly to A1 A2 A3 A4 A5).

A01-A10 expands to A01 A02 ... A09 A10 (formerly to A1 A2 ... A9 A10).

**long names** - it is now possible to have more than 9,999 long names in a program or in a databank. The new upper limit is 9,999,999. Of course, you usually will need a special large version of TSP to get even 9999 long names, and it can be slow to use this many different variable names.

The error message for lack of identification in linear models has been improved. If you start with 1000 observations and 10 variables, but you only have 3 observations after dropping missing values, the message now says "#vars < #obs after dropping missing values".

Now, when an error message is greater than the buffer, as much of the message is printed as possible (as opposed to none at all).

The 2 changes above will print as much warning message as possible, when a large number of variables have missing values in a regression. Before this, no variables were listed if there were "too many" to fit in the message buffer.

**TSP/GiveWin** - jump from an output file error message to the corresponding line in the input file. Just double-click on the line just above the \*\*\* ERROR message, which has the filename and physical line number for the input file.

Example: if output is to [c:\program files\tsp 4.5\illus45.out](#) and there is an error on physical line number 22, the following new line will appear in the output file:

```
C:\program files\tsp 4.5\illus45.out (22): ERROR
```

After reading the error message explanation on the lines below, double-click on this line, and you'll jump back to line 22 of the input file, so you can make any needed corrections (to the command which starts on line 22 or to previous commands). This should make the edit/run cycle a bit easier. At present, TSP will remember the physical line numbers for the first 2000 commands in the main batch input file. For command errors in other INPUT files or in commands beyond the first 2000, use the old methods of checking the "COMMAND" number (formerly called a "line" number; not a physical line number) in the PROGRAM listing.

A faster sorting algorithm is used in **SORT**, **MSD(ALL)**, **REGOPT SWILK**, **GRAPH** (hybrid QuickSort). There is a noticeable speed difference for 1000 or more observations and this algorithm also works well on presorted series. Timings on Pentium-166 for sorting (in seconds):

**Random series**

Observations	Old (seconds)	New (seconds)
1,000	0.22	0.00
10,000	20.7	0.17
100,000	2085.6	1.4

**Presorted series**

Observations	Old (seconds)	New (seconds)
1,000	0.00	0.05
10,000	0.00	0.05
100,000	0.28	0.88

**2. Printing and graphics enhancements:**

**TSP/GiveWin** makes simpler printed output for series listings.

In interactive mode, when PRINT is used on a series with more than 21 observations, it no longer inserts the title of the series after every 21 lines. (TSP thought there was a page break).

In batch mode, all page headers have been eliminated from the default output, since these are supplied by the Givewin output formatting feature.

New graphics options for the HIST procedure in TSP/Givewin:

**CDF/NOCDF** - include a normal QQ Plot below the histogram

**DENSITY/NODENSITY** - superimpose a smoothed density on the histogram

**HIST/NOHIST** - include a bar type histogram (the default)

**NORMAL/NONORMAL** - superimpose a normal density on the histogram

**STANDARD/NOSTANDARD** - use standardized data.

**TITLE=** 'title string' to label plot [the default is no title]

**WINDOW=** 'window name' to label window [the default is 'TSP Graphics']

In addition to the graphics version of HIST, the regular text output of HIST is produced when the PRINT option is on. @HIST and @HISTVAL are stored in either case. The SILENT option suppresses graphics and text output (@HIST and @HISTVAL will still be stored). The NOPREV option will suppress just the graphical window. The text version ignores the graphics options (like

DENSITY). The graphics version will not run if old options like NBINS= or DISCRETE are used (so use the NOPREV option with these, or else you will get an error message).

**GRAPH:** Missing values are now handled independently for each X,Y series pair (pairwise deletion rather than requiring all variables to be present). There is a new option **SORT/NOSORT**, which controls the ordering of the data on the X-axis so that the line connecting the points will not cross itself.

**PLOT** and **GRAPH:** When the **FREQ** is **PANEL** (stacked panel data), the plot lines have a break between individuals. **SMPL** gaps for any **FREQ** also appear as a break in the plot line. (Previously, a character-style plot was used if there were any **SMPL** gaps).

**PLOT (SYMBOL)** or **GRAPH (SYMBOL)** draws lines with symbols.

### 3. New or substantially improved procedures:

#### Linear procedures

Heteroskedastic and autocorrelation-consistent standard error estimates for panel data models are now computed using the new **HCOMEGA** option, available in **OLSQ**, **PANEL**, and **2SLS (INST)**. The form of the option is **HCOMEGA=BLOCK/DIAGONAL,HCTYPE=0/1/2/3**. This option specifies the form of the  $\text{OMEGA} = E[u u']$  matrix used in computing robust standard errors. When **FREQ(PANEL)** is in effect, the default is **HCOMEGA=BLOCK**, which implies a block diagonal matrix (each block corresponding to a unit or individual); this yields SEs that are consistent to both heteroskedasticity and autocorrelation. The **HCTYPE=0/1** option controls whether or not a degrees of freedom correction is applied as well.

When **FREQ(PANEL)** is not in effect, **HCOMEGA=DIAGONAL** is the default, which yields SEs robust to heteroskedasticity.

#### Nonlinear procedures

##### (**ML Proc, BJEST, SOLVE(METHOD=FLPOW), HITER=D, HITER=F**)

A new option **EPSMIN=[.0001]** specifies the minimum parameter change ( $\varepsilon$ ) for numeric first derivatives. TSP evaluates the log likelihood at  $b_i \pm \varepsilon$  (and  $b_i \pm 2\varepsilon$  when **GRAD=C4**), where  $\varepsilon = \max(.001b_i, \text{EPSMIN})$ . Because  $\varepsilon$  appears in the denominator of the numeric derivative, it must be nonzero. It may be useful to use **EPSMIN = 10\*\*(-6)** or less to make more accurate derivatives when the value of  $b_i$  is less than 0.1. If you attempt to use **EPSMIN=0**, **EPSMIN** will be set to **10\*\*(-30)**, which may cause an overflow.

The **EPSMIN** option is also used to control the numeric step size when computing **HCOV=C** and **HCOV=U** (see below). When you have parameters with values smaller than **10\*\*(-5)** in magnitude, it will be helpful to use **EPSMIN** with a value somewhat smaller than your smallest parameter value. Otherwise, too large a stepsize will be used, and small parameters will appear to have zero standard errors.

**SBIC** is now computed from **LOGL** using the number of identified parameter (**@NCID**) in the model instead of the number of supplied parameters (**@NCOEF**). That is, the parameters with zero standard errors are not counted.

**Iteration output:**

More digits are now printed for F= and FNEW= . The new default is to print 11 digits (vs. 5), which means small changes in F and FNEW during the final iterations are usually visible. An exception occurs when F or FNEW is less than .01 (in absolute value). In this case only 7 digits are printed, because this happens when an exact solution with F=0 is being found. The number of digits printed can be changed with OPTIONS SIGNIF=n.

**New approximations to second derivatives (BJEST, ML, and FIML):**

1. HITER/HCOV=U - nUmeric second derivatives, using equations (25.3.23) and (25.3.26) in Abramovitz and Stegun. For commands which don't have analytic second derivatives available, numeric second derivatives provide a very close approximation to the true Hessian. The drawback is that it is relatively slow, requiring  $2K^2$  function evaluations for a model with K parameters. HITER=U yields quadratic convergence during iterations, which can be faster than HITER=F if the number of parameters is less than 7 or so. HCOV=U provides standard errors which are more reliable than BFGS (HCOV=F, which often produces "false zero" standard errors). They match HCOV=N standard errors to 3-5 digits in the tests we've performed. BJEST(EXACTML) and ML PROC have been changed so that HCOV=U is now the default. HITER=F remains the default for these, but HITER=U can be chosen as an option.

2. HITER/HCOV=C - disCrete Hessian, which is a numeric difference of analytic first derivatives. This is even more accurate than HCOV=U in terms of matching HCOV=N results -- it's usually good to 6+ digits in standard errors. It requires 2K derivative evaluations (times K values each), for a model with K parameters. Unfortunately it is very specialized -- it is really only useful in a command which has analytic first derivatives but not analytic second derivatives. The most important such command in TSP is FIML.

HITER=C provides quadratic convergence, close to the optimum, which can be dramatically better than HITER=G. However, we have left HITER=G as the default, because Calzolari and Panattoni found it tends to perform better when the starting values are far from the optimum. HCOV=C provides generally smaller standard errors than the default HCOV=B. Calzolari and Panattoni found that HCOV=B is usually closer to the true small sample distribution of the parameters. ML(HITER=C,HCOV=C) may also be helpful, for complicated models with a very large number of parameters. Sometimes using HITER=N for these models takes a very long time, because ML's algorithms for simplifying the second derivative code are rather primitive. So HITER=C may be worth a try for such models.

**ACTFIT**

Changes and percent changes versions of Theil's 1966 U-statistic are now printed.

Output has been improved and the SILENT/TERSE options added to suppress the output.

@R, @R2, @RMSE, @MSE, @MAE, @ME, @RMSPE, @MSPE, @MAPE, @MPE, @BETA, %BETA1, @U66, @U66P, @FBIAS, @FDVAR, @FDCOV, @FDB1, @FRES are now stored (as scalars).

**AR1(OBJFN=GLS)**

A common factor test has been added (stored as @COMFAC). This test is a likelihood ratio test of AR(1) versus OLS with the lagged dependent and lagged right hand side variables added. It is similar to doing the Wald test for the nonlinear in the parameters restrictions implied by an AR(1) model. COMFAC is computed only when using the GLS method for estimation or when there is a lagged dependent variable on the right hand side. [The test is not well-defined for the default maximum likelihood method for AR(1), because there is no comparable treatment of the first observation for the unconstrained OLS model.]

### **AR1(REI)**

does random effects estimation using ML. It is similar to PANEL(REI), but with an added AR(1) component. The procedure follows Baltagi and Li, *Journal of Econometrics*, 1991, and uses analytic second derivatives for quadratic convergence and accurate t-statistics for all parameters (including RHO and RHO\_I, the intraclass correlation coefficient, which can be negative).

### **BJEST**

The ROOTS (or PRINT) options now print the moduli of the roots below the real/imaginary root values. Roots and moduli are stored as column vectors for each lag polynomial (of order 2 or higher), using the following names:

@ARRTRE AR RooTs REal parts (NAR x 1)  
 @ARRTIM AR RooTs IMaginary parts (NAR x 1)  
 @ARRTMO AR RooTs MOduli (NAR x 1)  
 @MARTRE MA RooTs REal parts (NMA x 1)  
 @MARTIM MA RooTs IMaginary parts (NMA x 1)  
 @MARTMO MA RooTs MOduli (NMA x 1)  
 @SARRTRE Seasonal AR RooTs REal parts (NSAR x 1)  
 and so forth.....

This allows the user to check for unit roots (for moving average models only, unless stationarity checking is turned off) and near unit roots (minimum moduli close to 1)

The period in which backforecasts are started has been changed to use all available data when the model is a pure MA. This conforms to the way it is done in Box and Jenkins(1976).

### **BJIDENT**

New **ESACF,NAR=nar,NMA=nma,[BARTLETT]/NOBART,PRINT/[NOPRINT]** options compute the extended sample ACF of Tsay and Tiao, *JASA*, March 1984, pp.84-96. This can be useful for identifying stationary and nonstationary ARMA models. The upper left vertex of a triangle of zeroes in the Indicator matrix identifies the order of the ARMA model. The zeros correspond to nonsignificant autocorrelations. Here is an example:

**BJIDENT(ESACF,NAR=5,NMA=8) CHEM; ? (Box-Jenkins Series C)**

The result is the following matrix:

MA

```

AR      0 1 2 3 4 5 6 7 8
0      9 9 9 9 9 9 9 9 1
1      9 9 9 9 9 1 1 0 0
2      0 0 0 0 0 0 0 0 0
3      9 0 0 0 0 0 0 0 0
4      9 9 0 0 0 0 0 0 0
5      9 9 9 0 0 0 0 0 0

```

A triangle of zeroes with upper left vertex at (2,0) is seen; this indicates an ARMA(2,0) model.

Output normally includes the ESACF correlations, their p-values, and a table of Indicators. These are stored as @ESACF, %ESACF, and @ESACFI. The PRINT option will print a table of AR coefficient estimates labelled @PHI (these are stored as @PHI). Normally, BJIDENT will produce its usual tables of autocorrelations and partial autocorrelations before printing the ESACF results. To avoid printing the other results, use the options NLAG=0, NLAGP=0. If just the ESACF option is used, with NAR and NMA not specified, the default values of NLAG (20) and NLAGP (10) will be used for AR and NMA.

The NOBARTLETT option can be specified to use simply  $1/(T-p-q)$  for the asymptotic variance of the ESACF (instead of using lower order autocorrelations); NOBARTLETT will result in smaller p-values at higher MA orders, and a few more significant autocorrelations. ESACF has been tested with Box-Jenkins series A and C, reproducing exactly the correlations in the tables of Tsay and Tiao (1984).

## CDF

The **CHISQ** option now allows non-integer DF (degrees of freedom), which is useful for evaluating the incomplete gamma function.

The **F** option now allows non-integer DF1 and DF2 (degrees of freedom), which is useful for evaluating the incomplete beta function.

## CONVERT

More than one series can be converted in a single command. Each series is replaced by its converted version. To store them under different names, either use the new = old form of the command, or use the COPY command to save the old series under the new names before running CONVERT.

The **MAP=mapseries,[SMPL]/NOSMPL** option computes SUM (default) or AVERAGE from old series to new, using a MAP of pointers. This is helpful for aggregating grouped data, such as industries, states, or individuals with panel data. The rows of the map correspond to the rows in the old series. The values in the map correspond to the rows of the new series. Zero values mean the observation is not mapped. The SMPL option is the default when MAP is used, and it puts the map and old series under the control of the current SMPL, while the new output series will be **FREQ N**, starting at observation 1. If NOSMPL is used, the traditional CONVERT method is used, where the old and map series are used at their maximum defined lengths, and the current **FREQ/SMPL** are only used to determine the **FREQ** and starting point of the new series. With NOSMPL, the map and old series must be defined over exactly the same set of observations. The map cannot contain any

missing values or contain only zeroes. The old series can contain missing values; they will result in missing values in the new series if the observations are mapped. If no observations of the old series are mapped to a given element of the new series, the element is given the value zero (for both sum and average). The length of the new series is equal to the maximum value in the map series.

### Databank routines:

**DBCOPY creates LOAD instead of READ commands, for better compatibility with SIMPC.**

**DBLIST(SILENT), SHOW(SILENT) SERIES;** - store @RNMS silently

**DOT** loops can now be nested up to 10 deep (so there can be up to ten dots in a name).

**EQSUB** has a new option [LAGS]/NOLAGS that controls substitution for the dependent variable name. EQSUB(NOLAGS) substitutes the equation only for the unlagged appearances of the variable name in a model.

### FORM

FRMLs can now be created after a VAR estimation, with the PARAM, COEFPR, and VARPR options.

Example:

```
VAR Y1 Y2 | C T; FORM(VARPR=H) EQ1 EQ2;
```

creates

```
FRML EQ1 Y1 = HY1_0 + HY1_T*T;
FRML EQ2 Y2 = HY2_0 + HY2_T*T;
```

with HY1\_0, HY1\_T, HY2\_0 and HY2\_T params set to estimated values from the VAR.

The SUM option has been added. This is useful when creating a sum with a lot of terms or an arbitrary number of terms that is not known in advance. It can be used as part of a log likelihood for ML, when summing across the T dimension in panel data, because T does not have to be known in advance.

Example:

```
FORM(SUM) EQ S X1-X3;
```

creates

```
FRML EQ S = X1+X2+X3;
```

**GMM (INST=(list1 | list2 | ... | listG)) eq1-eqG;** is a new syntax that specifies different instruments for each equation (like the MASK=matrix option, which still works). Unlike the MASK implementation, GMM uses a smaller COVOC matrix in this case, with only the orthogonality conditions specified, resulting in substantial savings in time and memory. The instrument names are printed for each equation, unless the TERSE or SILENT options are used.

**IN/OUT** - The limit of 16 different databanks in a run has been removed. There is still a limit of 8 active IN and 8 active OUT databanks at any particular point in a run.

**INTERVAL** – new procedure for interval regression. This is very similar to ordered probit with known bound values between the categories. It is also similar to 2-limit Tobit, but when  $y^*$  is between the known lower and upper bounds, the actual value of  $y^*$  is not observed. Interval regression can be used when the dependent variable is in a known range, but the actual value has been censored for confidentiality, which is often done for income or housing cost in surveys.

### **KERNEL**

This new procedure computes a kernel density or kernel regression. With a single argument, a Gaussian kernel density of  $x$  is computed and stored in `@DENSITY`. With two arguments, a Gaussian kernel regression of  $y$  on  $x$  is computed and the smoothed values of  $y$  are stored in `@FIT`.

### **LAD**

New options (**LOWER=series or UPPER=series**) compute a Censored Quantile Regression, using the BRCENS algorithm by Bernd Fitzenberger. No VCOV is computed at present, nor is a unique solution checked for. The option **NBOOT=** can be used to control the number of bootstrap replications for standard error computation.

The Machado & Santos-Silva (modified Glejser) heteroskedasticity test has replaced the LMHET test. This test is done by regressing weighted absolute values of the residuals on the RHS variables.

`@UNIQUE=0` is stored if the solution is probably not unique. `@IFCONV=0` is stored if there is a loss of precision in the simplex iterations.

**LIST** now allows a numeric list in decreasing order. Example:

```
LIST (first=4,last=1,prefix=x) xs;
```

Creates a list with elements X4, X3, X3, and X1 (in that order).

**LIST (SUFFIX=)** adds a suffix to the created names rather than a prefix.

**LOGIT** now handles missing values. When there are multiple records per case, one record with missing values will drop the entire case.

The SUFFIX option is used to give short names to the alternatives. These names are used in 4 places: in the initial table of frequencies for  $Y$ , as coefficient names for multinomial variables, as labels for  $dP/dZ$ , and as the suffixes for conditional variable names (when there is one observation per case).

SUFFIX does not imply COND. For example, say  $Y$  is coded 1,2,3.

```
LOGIT (COND,NCHOICE=3) Y XA XB;
```

will use the conditional variables XA1,XA2,XA3,XB1,XB2,XB3.

```
LOGIT (COND,NCHOICE=3,SUFFIX=(CAR,BUS,RT)) Y XA XB;
```

will use the conditional variables XACAR,XABUS,XART,XBCAR,XBBUS,XBRT.

The SUFFIX names need to be in the proper order, relative to the values of  $Y$ ; in the above example,  $Y=1$  is CAR,  $Y=2$  is BUS, and  $Y=3$  is RT. If some of the alternatives are never chosen, be sure to use SUFFIX or NCHOICE= to make sure the full set of conditional variables is used (corresponding to all available alternatives).

**LSQ/3SLS/GMM/FIML:**

An option for additive individual Fixed Effects (FEI) has been added to these procedures (only available when the equations are linear). Checks are made for linearity in the parameters, and linearity in endogenous variables for FIML. It is assumed the equations are linear in all variables. Individual means are removed from all variables, including instruments. The estimated fixed effects are not computed at present.

**LMS** - The coefficients have been reordered so that the constant C remains in the position specified by the user. Prior to this, C was always put last.

**MODEL (DONGALLO)** - This is a new option that orders each simultaneous block for a near-minimal feedback set. It prints an F next to the feedback variables, and a block summary. The Don-Gallo ordering is sometimes useful with Gauss-Seidel, and could be even more useful when SOLVE handles the feedback sets.

**MSD (ALL,WEIGHT=w)** - The weight is now used for computing the quartiles.

**OPTIONS:**

**LIMWMISS**= [10] is a new option for controlling the number of warning messages printed for missing values. Using **OPTIONS LIMWMISS=0** is preferred to **OPTIONS LIMWARN=0**, because **LIMWARN** will suppress other types of warnings which may still be informative.

**OPTIONS [ARGSUB]/NOARGSUB** controls substitution of actual arguments for formal arguments, for commands executed within a PROC. **OPTIONS NOARGSUB**; turns off the substitution. **NOARGSUB** is helpful if the PROC has **LOCAL** variables which have the same names as global variables that are passed as arguments to the PROC (it prevents the local variables from being used instead of the PROC arguments). Here is an example:

```
PROC FOO X;
LOCAL Y;
PRINT X;
ENDPROC;
SET Y=123; FOO Y;
```

fails with **ARGSUB**, because the Proc tries to print the local variable Y.

With **NOARGSUB**, the PROC prints X correctly. With **ARGSUB**, you would get to see the actual argument name (Y) when it is printed, if there was no conflict with a local variable. Thus when using **ARGSUB** (the default), it is advisable to choose local variable names which are unlikely to match global variables that might be used as arguments.

**PANEL:**

Titles for the different estimators within **PANEL** have been changed slightly and underlined with **====** in the output. A summary table of all the estimated ML models with **LOGL**, **SBIC** and title is printed. The “Ahrens-Pincus” measure of the degree of unbalancedness is printed.

**BYID** option prints and stores **@LOGL** and **@SBIC** (the coefficients are still not printed by default).

**TOTAL** option prints and stores **@LOGLT**, **@SBICT**, **@AICT**.

**WITHIN** option prints and stores @LOGLW, @SBICW, @AICW.

**REIT** option does two-way random effects ML estimation, with balanced or unbalanced data, using the method of P. Davis (2002).

**REI,[ALL]/NOALL,nonlinear\_options**) does Random Effects Individual (one-way) ML estimation. (VARCOMP is the GLS version of this model). The variable which indexes the individual is given in a `FREQ(PANEL,ID=individual_var,TIME=time_var)` command. The data is assumed sorted by ID. If you want to estimate a model with random Time or Group effects, and no standard Individual effects, you can use `FREQ(PANEL,ID=time_or_group_variable)`, but you need to make sure the data is sorted by this variable.

The parameters `RHO_I` and `SIGMA2` are estimated. `SIGMA2` is the residual variance and `RHO_I` is the ratio of the off-diagonal error covariances to `SIGMA2`. `RHO_I=0` is the same as the `TOTAL` estimator. `RHO_I` is bounded between  $-1/(TI\_MAX-1)$  and 1, where `TI_MAX` is the largest number of observations for any individual in the dataset. These bounds make sure that the overall residual covariance matrix (Omega) is positive definite. Note that `RHO_I` can be negative, which would correspond to negative off-diagonal covariances. `REI` does a grid search over 22 values of `RHO_I`, to locate any multiple local optima. All local optima are refined by iteration, any multiple optima are noted, and the best one is printed. The existence of multiple optima for this model was explored by Nerlove and Maddala in *Econometrica*, March 1971.

The grid search was tested with multiple optima obtained from Nerlove's 1971 data generation process (including cases where the global optimum has `RHO_I` negative and others where the global optimum has `RHO_I` positive). If you supply starting values in @START (including `RHO_I` and `SIGMA2`), the grid search will be skipped. The Log likelihood parameterization is based on Nerlove's spectral decomposition of the residual covariance matrix (in terms of its eigenvalues). As usual, analytic first and second derivatives are used for accuracy and fast/reliable convergence.

After the usual table of coefficients (including `RHO_I` and `SIGMA2`) and standard errors is printed, a table of alternative variance components and their standard errors is printed. These are:

`S2_I` = variance of `e_i`

`S2_IT` = variance of `e_it` (overall residual `u_it = e_i + e_it`)

They are related to the main parameters:

$SIGMA2 = S2\_I + S2\_IT$

$RHO\_I = S2\_I/SIGMA2$

Nonlinear options such as `MAXIT` and `TOL` may be useful. `HITER=N` and `HCOV=N` are used and `HITER=B` is not implemented.

The main variables are stored with suffix `REI`: @LOGLREI, @SBICREI, @COEFREI, @SESREI, @VCOVREI, @RNMSREI, etc. The alternative variance components are stored with suffix `VI`: @COEFVI, @SESVI, @VCOVVI, etc. The `NOALL` option can be used to turn off all regression models.

So if you only want to see the `REI` results, use `PANEL(NOALL,REI)`. At present, `NOREI` is the default, but `REI` will become the default when the next version of TSP is released.

**REIT** [Random Effects Individual and Time (Two-way) ML estimation]

The variables which index Individual and Time are given in `FREQ(PANEL, ID=individual_var, TIME=time_var)` command. The data is assumed sorted by ID. `time_var` is a non-negative integer variable which indexes the second variance component. It does not have to index time periods; it can index groups, so that models with nested structures can be estimated. The group could be the same value in all observations for an individual, which would be a nested value, or it can have a partially nested or non-nested (true time) structure.

The parameters `RHO_I`, `RHO_T` and `SIGMA2` are estimated. `SIGMA2` is the residual variance. `RHO_I` is the ratio of the off-diagonal error covariances (for a given individual) to `SIGMA2`. `RHO_T` is the ratio of the off-diagonal error covariances (for a given time) to `SIGMA2`. `RHO_I=RHO_T=0` is the same as the TOTAL (OLS) estimator. Note that `RHO_I` and/or `RHO_T` can be negative, which would correspond to negative off-diagonal covariances.

REIT does not at present do a grid search, so it does not detect multiple local optima. Instead it simply starts from the REI estimates (if the REI option is used, which is recommended), with `RHO_T` started at .05. The user may override this by supplying starting values in an `@START` vector. Multiple optima do exist for this model, and a grid search will probably be added soon.

The Log likelihood parameterization is based on P. Davis (2002), with simplifications from the 3-component model to the 2-component model. During iterations, the parameters `R_I=S2_I/S2_IT`, `R_T=S2_T/S2_IT` and `S2_IT` are used, because this yields the simplest form for the analytic derivatives.

It is assumed that `I (N)` is the largest dimension of the data, and there may be thousands of individuals, so no `NxN` matrices are created. A few `TxT` and `TxN` matrices are created, but usually these are quite small.

Nonlinear options such as `MAXIT` and `TOL` may be useful. `HITER=N` and `HCOV=N` are used and `HITER=B` is not really possible.

After the usual table of coefficients (including `RHO_I`, `RHO_T` and `SIGMA2`) and standard errors is printed, a table of alternative variance components and their standard errors is printed. These are:

`S2_I` = variance of `e_i`

`S2_T` = variance of `e_t`

`S2_IT` = variance of `e_it` (overall residual `u_it = e_i + e_t + e_it`)

They are related to the main parameters:

$SIGMA2 = S2_I + S2_T + S2_IT$

$RHO_I = S2_I / SIGMA2$

$RHO_T = S2_T / SIGMA2$

Nonlinear options such as `MAXIT` and `TOL` may be useful. `HITER=N` and `HCOV=N` are used and `HITER=B` is not implemented.

The main variables are stored with suffix REIT: `@LOGLREIT`, `@SBICREIT`, `@COEFREIT`, `@SESREIT`, `@VCOVREIT`, `@RNMSREIT`, etc. The alternative variance components are stored with suffix VIT: `@COEFVIT`, `@SESVIT`, `@VCOVVIT`, etc. The `NOALL` option can be used to turn off all regression models.

So if you only want to see the REI and REIT results, use

PANEL(NOALL,REI,REIT) .

At present, NOREIT is the default, but REIT will become the default when the next version of TSP is released.

PANEL(REI,REIT) prints and stores the Ahrens-Pincus Unbalancedness measures (@APUI for I dimension - always, and @APUT for T dimension - if REIT is used). These are described in the Davis (2002) article and elsewhere.

**PARAM** - If 2 consecutive numbers are found in the arguments, issue a (single) warning, instead of an error message. Ignore ( ) \* \*\* in the argument list. This is useful for pasting results from a results table back into a PARAM command, to restart iterations from previous estimates.

### PROBIT:

**FEI, FEPRINT** computes estimates for a model with Fixed Effects for Individuals.

FREQ(PANEL) must be in effect. Hundreds or thousands of fixed effects can be handled, for balanced or unbalanced data. A very efficient iteration algorithm is used to estimate the fixed effects. The FEPRINT option prints a table of the fixed effects (AI1, AI2, ...), their standard errors, and t-statistics. Fixed effects are also stored in the series @AI. Individuals with all zeroes or all ones for the dependent variable are allowed, although their data is not informative for the slope coefficients. The fixed effects estimator of the probit model is known to have a finite-T bias, but the size and direction of the bias are not known.

**REI** computes estimates of a Random Effects for Individuals. FREQ(PANEL) must be in effect. The model estimated is

$$Y_{it} = b X_{it} + u_i + e_{it}, u_i \sim N(0, \sigma_u^2), e_{it} \sim N(0, \sigma^2).$$

The normalization  $\sigma_u^2 + \sigma^2 = 1$  is used. Note that some other packages (notably Limdep and Stata) use the normalization  $\sigma = 1$ . This is less convenient, because the slope estimates are not normalized comparably with the results from the usual PROBIT command (unless  $\sigma_u = 0$ ). The parameter  $\rho = \sigma_u^2 / (\sigma_u^2 + \sigma^2)$  is estimated, with  $0 < \rho < 1$ . The integral is computed with Hermite quadrature, using a default 20 points. For a benchmark with bivariate probit for the case T=2, see <http://www.stanford.edu/~clint/bench/probre2.tsp> .

This command is still under development. The marginal effects and other diagnostics for REI have not been completed yet. There is also the matter of choosing the number of points in the Hermite quadrature adaptively, to make sure the LogL is accurate. The NHERMITE=[20/last]) option controls the number of points in Hermite quadrature. The default number of points is 20. Higher values (up to 92 or higher) may be required for accuracy if  $\rho$  is relatively high. If the NHERMITE option is used, the new value is remembered and is used in subsequent commands unless it is changed again.

### RANDOM:

An error message is now printed if the supplied covariance matrix VCOV is not positive definite.

The options MEAN=, \STDEV=, and VAR= can now take the names of series as arguments instead of scalar names or values. This feature is most useful for drawing Poisson or negative binomial random variables with different means in each observation.

The NOREPLACE, DRAW= option can now be used to draw from a series or matrix without replacement. For example, to draw a set of 5 cards from a deck of 52:

```
SMPL 1,52;
TREND OBS; SUITE = 1 + INT((OBS-1)/13);
TREND(PER=13) NUMBER;
MMAKE CARDS SUITE NUMBER;
SMPL 1,5;
RANDOM(DRAW=CARDS,NOREPL) SUITE NUMBER;
```

Similarly, to permute or "shuffle" a series of residuals:

```
SMPL 1,100;
OLSQ Y C X1 X2;
RANDOM(DRAW=@RES,NOREPL) U;
```

### READ:

**FORMAT=FREE** If the the number of values found in the file does not match the expected number of observations times number of variables, an improved message is printed, and the program tries to guess the values of Nobs or Nvars which match the number of values read. Trailing ? comments in free format files are now handled correctly.

**FORMAT=STATA** now reads Stata version 7 .dta files.

**FORMAT=EXCEL** now handles all current versions of Excel files (5, 7/95, 97/98/2000/2002). Reading the Excel 97 and higher files directly may be helpful, because these files can have more than 16384 rows (up to 65536 rows), and files with this many rows cannot be easily saved to the Excel 4 format (which TSP has long been able to read).

### REGOPT:

The diagnostic regressions (for CHOW, CHOWHET, RESET2, etc.) are now run in the default precision (usually NOFAST). Previously, the diagnostics were run always in FAST precision to save time (FAST is about 3 times faster than NOFAST for each regression). The precision can make a difference for some diagnostics like RESET2 in some models with a large number of RHS variables.

A Chow test that is robust to simple heteroskedasticity has been added (the @CHOWHET option). This is the MAC2 test from Thursby, *Journal of Econometrics*, 1992.

A new option QLAGS=k turns on output for Q statistics up to k lags.

### SHOW

PANEL "frequency" is now displayed when FREQ(PANEL) is in effect.

The equation output shows the number of arguments and number of operations, so the size of the equation can be assessed.

A list of all series stored as @RNMS, which makes it easier to do operations on all the series in a run or in a databank, such as writing them to a spreadsheet file.

### SIML:

All exogenous variables with missing values are now listed instead of just the first.

Only a single iteration is used for any linear model. Previously SIML took more than one iteration unless the Jacobian was also time-invariant..

When the PRNRES option is on, SIML now prints the Jacobian for the first 2 time periods.

### **SOLVE**

All exogenous variables with missing values are now listed instead of just the first. The labelling of model blocks has been improved.

The iteration count for Gauss-Seidel has been added to the output.

**SUR** - When SUR(WNAME=<matrix>) is used for Minimum Distance estimation, a "MINIMUM DISTANCE ESTIMATION" label is printed, and the residual covariance matrices and individual "equation" statistics are not printed, because there is just a single observation (on the PI matrix of reduced form coefficients).

**TOBIT**

New options (LOWER=lowerlimit,UPPER=upperlimit) allow both lower and upper censoring in the same model. The lowerlimit and upperlimit can be scalars or series. The default is LOWER=0, which is the same as before.

**2SLS**

A pseudo F-statistic for zero slopes is now printed and stored.

**FEI** option computes 2SLS or LIML estimates with Fixed Effects for Individuals. These are computed by removing individual means from all variables and instruments. The **FREQ(PANEL)** command must be in effect prior to using the FEI option, so that the individuals are well defined. If the **FEPRINT** option is on, a table of the estimated fixed effects, SEs, t-stats, and p-values will be printed after the usual table for the slope coefficients. Estimated fixed effects are stored in the series **@AI**, and in the vector **@COEFAI** (with corresponding **@SESAI**, **@TAI**, **%TAI**). **C** (intercept) variables on the RHS or in the instrument list are removed prior to estimation, since they are not identified. **2SLS(FEI,ROBUST)** yields robust SEs for the slope, but non-robust SEs for the fixed effects (at present). If the model reduces to ordinary least squares because there are no RHS endogenous variables, no calculations are done, because FEI is not (currently) available in OLS (this type of model can be estimated using the **PANEL** command). The **WEIGHT** option is not available with FEI.

**2SLS, LSQ(INST=)** now stores and prints a test of overidentifying restrictions (**@FOVERID**) when the number of instruments **NX** is greater than the number of right hand side variables **NZ**. Essentially, this is a test for whether the excluded instruments enter the right hand side of the equation with nonzero coefficients (i.e. whether their exclusion significantly degrades the fit). The formula for the test is

$$F_{\text{overid}} = @PHI/(@S2*(NX-NZ))$$

**VAR, LSQ, FIML, SAMPSEL** now use  $\log(\text{NOB} * \text{NEQ})$  when computing **@SBIC** for multiequation models. This is the correct degrees of freedom which counts the total number of residuals and makes it possible to use SBIC to compare multiequation models like SUR with stacked panel models.

**WRITE(FORMAT=FREE):**

The values written in free format are more concise, especially if they are integers. Double and single precision variables are written to all significant digits.

Each observation now starts on a new line. Rounding has been cleaned up and minimal spacing is used to produce more compact output.

**4. Bug fixes and minor enhancements since June 1999:**

**ANALYZ** An error message is given when no **FRML** argument(s) are supplied. The correct degrees of freedom is computed when the restrictions are collinear (subtract one df for each collinear restriction). The correct degrees of freedom for the F-test when there are gaps in the sample is computed. Lagged variables are allowed in equations (series can be used when there is a single observation in the current sample, which permits the computation of standard errors for various predicted functions of the data).

**ARI** The IFTSCS option (time series-cross section) is turned on only when `FREQ = PANEL`, but not when there are simply gaps in the sample.

**BJEST** Changes to avoid floating overflow in the root check, which can occur in some models/datasets. BJEST now remembers the EXACTML and NSPAN options from command to command, just like it remembers the other options. That is, they become the default in later BJEST/BJFRCST commands. An error message is given if parameter values are supplied but the START keyword was omitted.

**BJFRCST** An error message is given if the forecast series is missing at the origin date. The length of @FIT has been corrected when the series has been differenced. Changes to avoid zero divide if S value is left out or zero when the NORETRIEVE option is used.

Correct parsing for `CD=X`; `TITLE=2`; `INPUT=5`; so that they are GENRs and not implied commands.

**CDF** The p-value for options BIVNORM,UPTAIL is corrected; in this case it is not the same as  $1 - P(\text{lowtail})$ .

**CNORM2** The derivatives of this function at  $(x,y,\rho)$  when  $\rho$  is a constant have been corrected.

**COMPRESS** Occasional loss of data or invalid data due to compress when SELECT is in effect has been fixed.

**COPY, RENAME, or DELETE** Program crash that occurred when these commands were applied to a LOCAL variable in a PROC is now avoided.

**DIFFER** When applied to an unnormalized equation, it now creates and unnormalized resulting equation. When the PRINT option is used with long names, the "w.r.t." label is displayed correctly.

**DIVIND** When there are zero quantities, the derived index for the period prior to the period when the number of goods changes has been corrected (the series are spliced together).

**DOC** The ADD option now works properly; previously it did not save the old documentation.

**DOT** When used within a PROC, and a COMPRESS occurred before the second use of the PROC, the second use did not work correctly; this is now fixed. An error message is given when lagged values are used as DOT arguments (such as `X(-1)`).

**FIML** The dependent variable in a FRML can now be a LIST with a single element or a PROC argument. The model is no longer considered linear when some of the equation(s) contain non-differentiable operations (such as logical or comparison operations). Convergence cannot be guaranteed in a single iteration if the derivative doesn't exist.

**FORM** Forming equations after a VAR command now works correctly when there are CONSTs or variable names.

**FRML** Parsing when there is a comma after the equation name has been corrected (a comma is allowed, as elsewhere in TSP).

**FREQ(PANEL)** There are problems if `SMPL k,NT`; is used, where  $k$  is not 1. Program now tries to restore to `SMPL 1,NT`; if a SELECT command is given. TSP now refuses to execute the command if a `SMPL 1,NT`; cannot be obtained. Using lags with the panel frequency now works correctly: formerly, when a SELECT command removed some observations from the middle of an individual

series, the later observations for that individual were sometimes improperly set to missing for lagged variables.

**GENR** Parsing of this command when there are options and an equation name instead of an equation has now been corrected. Parsing when there is a subscripted list as an equation name has been corrected (such as GENR listfrml(1) y;). A PC only error in processing a statement such as GENR Y = (-1)\*\*4 has been corrected.

**LCNORM** This function now uses a more accurate approximation when the argument is less than -37.5.

**LIST(DELETE)** When used on a nonexistent series, just give a warning rather than an error.

## LOGIT

The procedure now allows for 11 (instead of 6) digits when printing the number of choices.

## LSQ

The dependent variable in the equation(s) to be estimated can now be a LIST or PROC argument. The option NODROPMISS now works even if there are no instruments. When there is a single equation, estimation is allowed even if the number of observations NOB equals the number of parameters to be estimated NCOEF. This will usually (but not always, if model is nonlinear) result in a perfect fit. The option HITER=D (Davidon-Fletcher-Powell method) is now allowed, although HCOV=D is not recommended, because the estimated variance matrix will be badly scaled when the objective function is not the log likelihood.

**MAT** For all procedures, including MAT, double precision values larger than  $1D+37$  are now set to missing when they are stored to a single precision series, and a numeric warning is printed, in order to avoid a crash. If double precision storage is needed, use the DOUBLE option on the initial OPTIONS command in the run. Parsing for equations such as  $Y = (X=2)$  is now corrected so that the equation is interpreted as  $Y = X .EQ. 2$ . The matrix rank function RANK() now correctly handles an M by N matrix with  $M > N$  and rank less than N. The treatment of subscripts on the right hand side of a MAT equation has been corrected. Scalars on the right hand side of a MAT equation now preserve their type (PARAM or CONST). An equation of the type  $S = (A*B)'(A*B)$ ; now stores S as symmetric type matrix. An error message now results if a FRML is used as an argument to the matrix procedure. When the SER function is used with a single observation sample, the results is still stored as a series, not as a scalar. A bug associated with removing arguments needed for later matrix operations in the same command when running out of memory has been corrected.

**MFORM** An error message is given if the TYPE option is enclosed in quotes, such as TYPE="sym"). An error message is given if the MFORM command specifies a matrix that is too big (more than the memory limit)

**ML PROC** A bug involving variables named "U" or "F" has been fixed. A bug in parsing the SILENT option has been corrected.

**MSD** When this command was used with a combination of the ALL and WEIGHT options, the weighted mean, etc. were incorrect The SILENT option for MSD is now truly silent and the TERSE option suppresses the computation of Skewness and Kurtosis.

**OLSQ** The combination of the SILENT and WEIGHT options turns off the display of weighted/unweighted header output.

**PANEL** The variance-covariance matrix label is not printed when the SILENT option is used.

**POISSON** A bug that caused the procedures to crash in some cases when there were missing values in the RHS variables has been corrected.

**PLOT/GRAPH** In the DOS/Win and GiveWin versions (with graphics), long or lagged variable names are now labelled correctly. In the DOS/Win version, when the DEV= option is used the program looks for the graphics drivers in the default install directory.

**PLOTS CUSUM** CUSUM and CUSUMSQ plots now handle missing observations correctly.

**RANDOM** A possible crash with the GAMMA option after many millions of draws have been made has been corrected.

**READ** The program no longer suggests number of observations or variables when reading a matrix instead of series. When reading .WK3 files, the value of zero when in a TREAL cell is now read correctly, instead of as 2. Missing values in Stata v6 double variables are now read correctly. Byte variables in Stata .dta files written on Sun computers that have negative values are now read correctly.

**REGOPT** A possible crash due to a negative square root during the computation of recursive residuals has been avoided. A possible problem with the BPHET options when there are many missing values in the original regression is now fixed. %LMHET (the p-value for the LM heteroskedasticity statistic is no longer lost when REGOPT(NOCALC) DW or AUTO is used

**SET** This command (rather than GENR) is now implied for statements like  $y(i2,i2) = 22$ ;

**SHOW Proc** Junk characters are no longer printed when the list of arguments is longer than 80 characters.

**SIML** The model is no longer considered linear when some of the equation(s) contain non-differentiable operations (such as logical or comparison operations). Convergence cannot be guaranteed in a single iteration if the derivative doesn't exist. A warning is issued when an endogenous variable does not appear in the model (in any of the equations). The SILENT option now turns off the printing of the simulated results (NOPRNSIM is implied).

## **SMPL**

An occasional crash when the second SMPL argument was a series has been avoided.

## **SOLVE**

There was an error when one or more endogenous variables appear on the right hand side only with a lag in a recursive model; the results are now correct. For example, this can happen when simulating a GARCH model. The SILENT option now turns off the printing of the simulated results (NOPRNSIM is implied).

**SORT** The procedure now handles mixed DOUBLE and NODOUBLE series correctly (previously, it converted them to the current type, possibly losing precision when DOUBLE was converted to NODOUBLE). The command SORT X X (a user error) did not yield the correct ordering for X; this is now fixed.

**TITLE** This command is now parsed correctly when there is an option or when the argument is a . (period)..

**WRITE** When writing a file with labels (the `FORMAT=LABEL` option) in interactive mode, do not pause the screen. For TSP/GiveWin and Win32 TSP the missing values written to Excel and Lotus files have been corrected. For free format (`FORMAT=FREE`) files, missing values in scalars and matrices, as well as series, are now written as . (instead of the value `-0.1099512E+13`). Writing variables of different types in a single free format statement is now allowed.