LMS (ALL, LTS, MOST, PRINT, SILENT, SUBSETS=*number of random subsets*, TERSE)
dependent variable   list of independent variables;

Function:

LMS computes least median of squares regression. This is a very robust procedure that is useful for outlier detection. It is the highest possible "breakdown" estimator, which means that up to 50% of the data can be replaced with bad numbers and it will still yield a consistent estimate. Proper standard errors (such as asymptotically normal) for LMS coefficients are not known at present.

Usage:

To estimate by least median of squares in TSP, use the LMS command just like the OLSQ command. For example,

```
LMS CONS,C,GNP ;
```

estimates the consumption function. The PRINT option can be used to print any outliers, or you can define them yourself using @RES.

Options:

ALL/**NOALL** uses all possible observation subsets (see **Method**), even if there are over one million of them.

LTS/**NOLTS** computes the Least Trimmed Squares estimates, which minimize the sum of squared residuals, from the smallest up to the median (instead of LMS, which minimizes just the squared median residual). Usually the LTS and LMS estimates are fairly close to each other.

MOST/**NOMOST** uses all possible subsets, unless the number of subsets is one million or more (in which case random subsets are used).

PRINT/**NOPRINT** prints better subsets (as progress towards a minimum is made), and the final outliers.

SILENT/**NOSILENT** suppresses all printed output.

SUBSETS= number of random subsets to use. The default is 500 to 3000. If the number of possible subsets is less than the number of random subsets, all possible subsets will be evaluated systematically.

TERSE/**NOTERSE** suppresses all printed output except the table of coefficient estimates and the value of the objective function.

Output:

The usual regression output is printed and stored (see OLSQ for a table). The number of possible subsets and the best subset found are also printed.

LMS

Method:

The LMS estimator minimizes the square (or the absolute value) of the median residual with respect to the coefficient vector b:

$$\min_{b} \ \underset{i}{median} |y_i - X_i b|$$

Clearly this ignores the sizes of the largest residuals in the sample (i.e. those whose absolute values are larger than the median), so it will be robust to the presence of any extreme data points (outliers).

If there are K independent variables (excluding the constant term), LMS will consider many different subsets of K observations each. An exact fit regression line is computed for each subset, and residuals are computed for the remaining observations using these coefficients. The residuals are essentially sorted to find the median, and a slight adjustment is made to allow for the constant term. If K or the number of observations is large, the number of subsets could be very large (and sorting time could be lengthy), so random subsets will usually be considered in this case. This is controlled with the ALL, MOST, and SUBSETS= options.

The result is not necessarily the global Least Median of Squares optimum, but a feasible close approximation to it, with the same properties for outlier detection. Least Trimmed Squares (LTS) also has about the same properties.

The LMS estimator occasionally produces a non-unique estimate of the coefficient vector b; TSP reports the number of non-unique subsets in this case.

An extremely rough estimate of the variance-covariance of the estimated coefficients is computed with an OLS-type formula:

$$\square\square\hat{b}) = \sigma^2(X'X)^{-1} \text{ , where } \sigma^2 \text{ deletes the largest residuals.}$$

This estimate is not asymptotically normal, and is likely an underestimate, so it should not be used for serious hypothesis testing. It all depends on how the outliers are generated by the underlying model.

The code used in TSP was adapted from Rousseeuw’s Progress program, obtainable from his web page.

References:

Rousseeuw, P.J., "Least Median of Squares Regression," **Journal of the American Statistical Association** 79 (1984), pp. 871-880.

Rousseeuw, P.J., and Leroy, A.M., **Robust Regression and Outlier Detection**, Wiley, (1987).

Rousseeuw, P.J., and Wagner, J., "Robust Regression with a distributed intercept using Least Median of Squares," **Computational Statistics and Data Analysis** 17 (1994), pp.66-68.

http://win-www.uia.ac.be/u/statis/ (Rousseeuw / Progress)