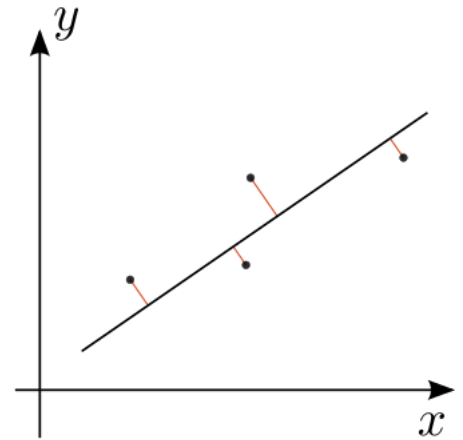# Total least squares

In applied statistics, **total least squares** is a type of errors-in-variables regression, a least squares data modeling technique in which observational errors on both dependent and independent variables are taken into account. It is a generalization of Deming regression and also of orthogonal regression, and can be applied to both linear and non-linear models.

The total least squares approximation of the data is generically equivalent to the best, in the Frobenius norm, low-rank approximation of the data matrix.[1]

Contents

The bivariate (Deming regression) case of total least squares. The red lines show the error in both *x* and *y*. This is different from the traditional least squares method which measures error parallel to the *y* axis. The case shown, with deviations measured perpendicularly, arises when *x* and *y* have equal variances.

Linear model

Background

In the least squares method of data modeling, the objective function, S,

$$S = \mathbf{r}^{\mathbf{T}}\mathbf{W}\mathbf{r},$$

is minimized, where r is the vector of residuals and W is a weighting matrix. In linear least squares the model contains equations which are linear in the parameters appearing in the parameter vector $\boldsymbol{\beta}$, so the residuals are given by

$$\mathbf{r} = \mathbf{y} - \mathbf{X}\boldsymbol{\beta}.$$

There are m observations in $\mathbf{y}$ and n parameters in $\boldsymbol{\beta}$ with $m>n$. $\mathbf{X}$ is a $m{\times}n$ matrix whose elements are either constants or functions of the independent variables, $\mathbf{x}$. The weight matrix $\mathbf{W}$ is, ideally, the inverse of the variance-covariance matrix $\mathbf{M}_y$ of the observations $\mathbf{y}$. The independent variables are assumed to be error-free. The parameter estimates are found by setting the gradient equations to zero, which results in the normal equations [note 1]

$$\mathbf{X^T W X}\boldsymbol{\beta} = \mathbf{X^T W y}$$

Allowing observation errors in all variables

Now, suppose that both $\mathbf{x}$ and $\mathbf{y}$ are observed subject to error, with variance-covariance matrices $\mathbf{M}_x$ and $\mathbf{M}_y$ respectively. In this case the objective function can be written as

$$S = \mathbf{r_x^T M_x^{-1} r_x} + \mathbf{r_y^T M_y^{-1} r_y}$$

where $\mathbf{r}_x$ and $\mathbf{r}_y$ are the residuals in $\mathbf{x}$ and $\mathbf{y}$ respectively. Clearly these residuals cannot be independent of each other, but they must be constrained by some kind of relationship. Writing the model function as $\mathbf{f(r_x,r_y,\boldsymbol{\beta})}$, the constraints are expressed by m condition equations.[2]

$$\mathbf{F} = \boldsymbol{\Delta}\mathbf{y} - \frac{\partial \mathbf{f}}{\partial \mathbf{r_x}}\mathbf{r_x} - \frac{\partial \mathbf{f}}{\partial \mathbf{r_y}}\mathbf{r_y} - \mathbf{X}\boldsymbol{\Delta}\boldsymbol{\beta} = \mathbf{0}$$

Thus, the problem is to minimize the objective function subject to the m constraints. It is solved by the use of Lagrange multipliers. After some algebraic manipulations,[3] the result is obtained.

$$\mathbf{X^T M^{-1} X}\boldsymbol{\Delta}\boldsymbol{\beta} = \mathbf{X^T M^{-1}}\boldsymbol{\Delta}\mathbf{y},$$

or alternatively $\mathbf{X^T M^{-1} X}\boldsymbol{\beta} = \mathbf{X^T M^{-1} y}$ Where $\mathbf{M}$ is the variance-covariance matrix relative to both independent and dependent variables.

$$\mathbf{M} = \mathbf{K_x M_x K_x^T} + \mathbf{K_y M_y K_y^T}; \ \mathbf{K_x} = -\frac{\partial \mathbf{f}}{\partial \mathbf{r_x}}, \ \mathbf{K_y} = -\frac{\partial \mathbf{f}}{\partial \mathbf{r_y}}$$

Example

When the data errors are uncorrelated, all matrices $\mathbf{M}$ and $\mathbf{W}$ are diagonal. Then, take the example of straight line fitting.

$$f(x_i, \beta) = \alpha + \beta x_i$$

in this case

$$M_{ii} = \sigma_{y,i}^2 + \beta^2 \sigma_{x,i}^2$$

showing how the variance at the ith point is determined by the variances of both independent and dependent variables and by the model being used to fit the data. The expression may be generalized by noting that the parameter β is the slope of the line.

$$M_{ii} = \sigma_{y,i}^2 + \left(\frac{dy}{dx}\right)_i^2 \sigma_{x,i}^2$$

An expression of this type is used in fitting pH titration data where a small error on x translates to a large error on y when the slope is large.

Algebraic point of view

First of all it is necessary to note that the TLS problem does not have a solution in general, which was already shown in 1980.[4] The following considers the simple case where a unique solution exists without making any particular assumptions.

The computation of the TLS using singular value decomposition is described in standard texts.[5] We can solve the equation

$$XB \approx Y$$

for B where X is m-by-n and Y is m-by-k. [note 2]

That is, we seek to find B that minimizes error matrices E and F for X and Y respectively. That is,

$$\text{argmin}_{E,F} \|[E\ F]\|_F, \qquad (X+E)B = Y + F$$

where $[E\ F]$ is the augmented matrix with E and F side by side and $\|\cdot\|_F$ is the Frobenius norm, the square root of the sum of the squares of all entries in a matrix and so equivalently the square root of the sum of squares of the lengths of the rows or columns of the matrix.

This can be rewritten as

$$[(X+E)\ (Y+F)]\begin{bmatrix} B \\ -I_k \end{bmatrix} = 0.$$

where I_k is the $k \times k$ identity matrix. The goal is then to find $[E\ F]$ that reduces the rank of $[X\ Y]$ by k. Define $[U][\Sigma][V]^*$ to be the singular value decomposition of the augmented matrix $[X\ Y]$.

$$[X\ Y] = [U_X\ U_Y]\begin{bmatrix} \Sigma_X & 0 \\ 0 & \Sigma_Y \end{bmatrix}\begin{bmatrix} V_{XX} & V_{XY} \\ V_{YX} & V_{YY} \end{bmatrix}^* = [U_X\ U_Y]\begin{bmatrix} \Sigma_X & 0 \\ 0 & \Sigma_Y \end{bmatrix}\begin{bmatrix} V_{XX}^* & V_{YX}^* \\ V_{XY}^* & V_{YY}^* \end{bmatrix}$$

where V is partitioned into blocks corresponding to the shape of X and Y.

Using the Eckart–Young theorem, the approximation minimising the norm of the error is such that matrices U and V are unchanged, while the k-smallest singular values are replaced with zeroes. That is, we want

$$[(X+E)\ (Y+F)] = [U_X\ U_Y]\begin{bmatrix} \Sigma_X & 0 \\ 0 & 0_{k\times k} \end{bmatrix}\begin{bmatrix} V_{XX} & V_{XY} \\ V_{YX} & V_{YY} \end{bmatrix}^*$$

so by linearity,

$$[E\ F] = -[U_X\ U_Y]\begin{bmatrix} 0_{n\times n} & 0 \\ 0 & \Sigma_Y \end{bmatrix}\begin{bmatrix} V_{XX} & V_{XY} \\ V_{YX} & V_{YY} \end{bmatrix}^*.$$

We can then remove blocks from the U and Σ matrices, simplifying to

$$[E\ F] = -U_Y \Sigma_Y \begin{bmatrix} V_{XY} \\ V_{YY} \end{bmatrix}^* = -[X\ Y] \begin{bmatrix} V_{XY} \\ V_{YY} \end{bmatrix} \begin{bmatrix} V_{XY} \\ V_{YY} \end{bmatrix}^*.$$

This provides E and F so that

$$[(X + E)\ (Y + F)] \begin{bmatrix} V_{XY} \\ V_{YY} \end{bmatrix} = 0.$$

Now if V_{YY} is nonsingular, which is not always the case (note that the behavior of TLS when V_{YY} is singular is not well understood yet), we can then right multiply both sides by $-V_{YY}^{-1}$ to bring the bottom block of the right matrix to the negative identity, giving[6]

$$[(X + E)\ (Y + F)] \begin{bmatrix} -V_{XY} V_{YY}^{-1} \\ -V_{YY} V_{YY}^{-1} \end{bmatrix} = [(X + E)\ (Y + F)] \begin{bmatrix} B \\ -I_k \end{bmatrix} = 0,$$

and so

$$B = -V_{XY} V_{YY}^{-1}.$$

A naive GNU Octave implementation of this is:

```
function B = tls(X,Y)

[m n]   = size(X);            % n is the width of X (X is m by n)
Z       = [X Y];             % Z is X augmented with Y.
[U S V] = svd(Z,0);          % find the SVD of Z.
VXY     = V(1:n,1+n:end);     % Take the block of V consisting of the first n rows and the n+1 to last column
VYY     = V(1+n:end,1+n:end); % Take the bottom-right block of V.
B       = -VXY/VYY;

end
```

The way described above of solving the problem, which requires that the matrix V_{YY} is nonsingular, can be slightly extended by the so-called *classical TLS algorithm*.[7]

Computation

The standard implementation of classical TLS algorithm is available through Netlib (http://www.netlib.org/vanhuffel/index.html), see also.[8][9] All modern implementations based, for example, on solving a sequence of ordinary least squares problems, approximate the matrix B (denoted X in the literature), as introduced by Van Huffel and Vandewalle. It is worth noting, that this B is, however, *not the TLS solution* in many cases.[10][11]

Non-linear model

For non-linear systems similar reasoning shows that the normal equations for an iteration cycle can be written as

$$\mathbf{J}^{\mathrm{T}} \mathbf{M}^{-1} \mathbf{J} \Delta \boldsymbol{\beta} = \mathbf{J}^{\mathrm{T}} \mathbf{M}^{-1} \Delta \mathbf{y}.$$

Geometrical interpretation

When the independent variable is error-free a residual represents the "vertical" distance between the observed data point and the fitted curve (or surface). In total least squares a residual represents the distance between a data point and the fitted curve measured along some direction. In fact, if both variables are measured in the same units and the errors on both variables are the same, then the residual represents the shortest distance between the data point and the fitted curve, that is, the residual vector is perpendicular to the tangent of the curve. For this reason, this type of regression is sometimes called *two dimensional Euclidean regression* (Stein, 1983)[12] or *orthogonal regression*.

Scale invariant methods

A serious difficulty arises if the variables are not measured in the same units. First consider measuring distance between a data point and the curve – what are the measurement units for this distance? If we consider measuring distance based on Pythagoras' Theorem then it is clear that we shall be adding quantities measured in different units, and so this leads to meaningless results. Secondly, if we rescale one of the variables e.g., measure in grams rather than kilograms, then we shall end up with different results (a different curve). To avoid this problem of incommensurability it is sometimes suggested that we convert to dimensionless variables—this may be called normalization or standardization. However there are various ways of doing this, and these lead to fitted models which are not equivalent to each other. One approach is to normalize by known (or estimated) measurement precision thereby minimizing the Mahalanobis distance from the points to the line, providing a maximum-likelihood solution; the unknown precisions could be found via analysis of variance.

In short, total least squares does not have the property of units-invariance—i.e. it is not scale invariant. For a meaningful model we require this property to hold. A way forward is to realise that residuals (distances) measured in different units can be combined if multiplication is used instead of addition. Consider fitting a line: for each data point the product of the vertical and horizontal residuals equals twice the area of the triangle formed by the residual lines and the fitted line. We choose the line which minimizes the sum of these areas. Nobel laureate Paul Samuelson proved in 1942 that, in two dimensions, it is the only line expressible solely in terms of the ratios of standard deviations and the correlation coefficient which (1) fits the correct equation when the observations fall on a straight line; (2) exhibits scale invariance, and (3) exhibits invariance under interchange of variables.[13] This line has been rediscovered in different disciplines and is variously known as standardised major axis (Ricker 1975, Warton et al., 2006),[14][15] the reduced major axis, the geometric mean functional relationship (Draper and Smith, 1998),[16] least products regression, diagonal regression, line of organic correlation, and the least areas line. Tofallis (2002)[17] has extended this approach to deal with multiple variables.

See also

- Deming regression, a special case with two predictors and independent errors.
- Errors-in-variables model
- Linear regression
- Least squares

Notes

1. An alternative form is $\mathbf{X^T W X \Delta\beta = X^T W \Delta y}$, where $\mathbf{\Delta\beta}$ is the parameter shift from some starting estimate of $\boldsymbol{\beta}$ and $\mathbf{\Delta y}$ is the difference between $\mathbf{y}$ and the value calculated using the starting value of $\boldsymbol{\beta}$

2. The notation $XB \approx Y$ is used here to reflect the notation used in the earlier part of the article. In the computational literature the problem has been more commonly presented as $AX \approx B$, i.e. with the letter X used for the n-by-k matrix of unknown regression coefficients.

References

1. I. Markovsky and S. Van Huffel, *Overview of total least squares methods.* Signal Processing, vol. 87, pp. 2283–2302, 2007. preprint (http://eprints.ecs.soton.ac.uk/13855/1/tls_overview.pdf)

2. W.E. Deming, Statistical Adjustment of Data, Wiley, 1943

3. Gans, Peter (1992). *Data Fitting in the Chemical Sciences* (http://www.wiley.com/WileyCDA/WileyTitle/productCd-0471934127.html). Wiley. ISBN 9780471934127. Retrieved 4 December 2012.

4. G. H. Golub and C. F. Van Loan, An analysis of the total least squares problem. Numer. Anal., 17, 1980, pp. 883–893.

5. Golub, Gene H.; Van Loan, Charles F. (1996). *Matrix Computations* (3rd ed.). The Johns Hopkins University Press. pp 596.

6. Björck, Ake (1996) *Numerical Methods for Least Squares Problems*, Society for Industrial and Applied Mathematics. ISBN 978-0898713602

7. S. Van Huffel and J. Vandewalle (1991) *The Total Least Squares Problems: Computational Aspects and Analysis*. SIAM Publications, Philadelphia PA.

8. S. Van Huffel, Documented Fortran 77 programs of the extended classical total least squares algorithm, the partial singular value decomposition algorithm and the partial total least squares algorithm, Internal Report ESAT-KUL 88/1, ESAT Lab., Dept. of Electrical Engineering, Katholieke Universiteit Leuven, 1988.

9. S. Van Huffel, The extended classical total least squares algorithm, J. Comput. Appl. Math., 25, pp. 111–119, 1989.

10. M. Plešinger, The Total Least Squares Problem and Reduction of Data in AX ≈ B. Doctoral Thesis, TU of Liberec and Institute of Computer Science, AS CR Prague, 2008. Ph.D. Thesis

11. I. Hnětynková, M. Plešinger, D. M. Sima, Z. Strakoš, and S. Van Huffel, The total least squares problem in AX ≈ B. A new classification with the relationship to the classical works. SIMAX vol. 32 issue 3 (2011), pp. 748–770.

12. Stein, Yaakov J. "Two Dimensional Euclidean Regression" (http://www.dspcsp.com/pubs/euclreg.pdf) (PDF).

13. Samuelson, Paul A. (1942). "A Note on Alternative Regressions". *Econometrica*. The Econometric Society. **10** (1): 80–83. doi:10.2307/1907024 (https://doi.org/10.2307%2F1907024). JSTOR 1907024 (https://www.jstor.org/stable/1907024).

14. Ricker, W. E. (1975). "A note concerning Professor Jolicoeur's Comments". *Journal of the Fisheries Research Board of Canada*. **32**: 1494–1498. doi:10.1139/f75-172 (https://doi.org/10.1139%2Ff75-172).

15. Warton, David I.; Wright, Ian J.; Falster, Daniel S.; Westoby, Mark (2006). "Bivariate line-fitting methods for allometry". *Biological Reviews*. Wiley. **81** (2): 259–291. doi:10.1017/S1464793106007007 (https://doi.org/10.1017%2FS1464793106007007).

16. Draper, NR and Smith, H. *Applied Regression Analysis*, 3rd edition, pp. 92–96. 1998

17. Tofallis, Chris (2002). "Model Fitting for Multiple Variables by Minimising the Geometric Mean Deviation". In Van Huffel, Sabine; Lemmerling, P. *Total Least Squares and Errors-in-Variables Modeling: Analysis, Algorithms and Applications*. Dordrecht: Kluwer Academic Publ. ISBN 978-1402004766. SSRN 1077322 (https://ssrn.com/abstract=1077322).

Others

- I. Hnětynková, M. Plešinger, D. M. Sima, Z. Strakoš, and S. Van Huffel, *The total least squares problem in AX ≈ B. A new classification with the relationship to the classical works.* SIMAX vol. 32 issue 3 (2011), pp. 748–770. Available as a preprint (ftp://ftp.sam.math.ethz.ch/pub/sam-reports/reports/reports2010/2010-38.pdf).

- M. Plešinger, *The Total Least Squares Problem and Reduction of Data in AX ≈ B.* Doctoral Thesis, TU of Liberec and Institute of Computer Science, AS CR Prague, 2008. Ph.D. Thesis (http://www.fp.tul.cz/~plesinger/my_publications/doctoral_thesis/thesis.pdf)

- C. C. Paige, Z. Strakoš, *Core problems in linear algebraic systems.* SIAM J. Matrix Anal. Appl. 27, 2006, pp. 861–875.

- S. Van Huffel and P. Lemmerling, *Total Least Squares and Errors-in-Variables Modeling: Analysis, Algorithms and Applications*. Dordrecht, The Netherlands: Kluwer Academic Publishers, 2002.

- S. Jo and S. W. Kim, *Consistent normalized least mean square filtering with noisy data matrix.* IEEE Trans. Signal Processing, vol. 53, no. 6, pp. 2112–2123, Jun. 2005.
- R. D. DeGroat and E. M. Dowling, *The data least squares problem and channel equalization.* IEEE Trans. Signal Processing, vol. 41, no. 1, pp. 407–411, Jan. 1993.
- S. Van Huffel and J. Vandewalle, *The Total Least Squares Problems: Computational Aspects and Analysis.* SIAM Publications, Philadelphia PA, 1991.
- T. Abatzoglou and J. Mendel, *Constrained total least squares*, in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP'87), Apr. 1987, vol. 12, pp. 1485–1488.
- P. de Groen *An introduction to total least squares*, in Nieuw Archief voor Wiskunde, Vierde serie, deel 14, 1996, pp. 237–253 arxiv.org (http://arxiv.org/pdf/math.RA/9805076/).
- G. H. Golub and C. F. Van Loan, *An analysis of the total least squares problem.* SIAM J. on Numer. Anal., 17, 1980, pp. 883–893.
- Perpendicular Regression Of A Line (http://www.mathpages.com/home/kmath110.htm) at MathPages
- A. R. Amiri-Simkooei and S. Jazaeri *Weighted total least squares formulated by standard least squares theory*,in Journal of Geodetic Science, 2 (2): 113–124, 2012 [1] (http://engold.ui.ac.ir/~amiri/JGS_Amiri_Jazaeri_2012.pdf).

This page was last edited on 19 January 2018, at 11:27.