Least Squares.

Hervé Abdi[1]
The University of Texas at Dallas

Introduction

The least square method—a very popular technique—is used to compute estimations of parameters and to fit data. It is one of the oldest techniques of modern statistics as it was first published in 1805 by the French mathematician Legendre in a now classic memoir. But this method is even older because it turned out that, after the publication of Legendre's memoir, Gauss, the famous German mathematician, published another memoir (in 1809) in which he mentioned that he had previously discovered this method and used it as early as 1795. A somewhat bitter anteriority dispute followed (a bit reminiscent of the Leibniz-Newton controversy about the invention of Calculus), which, however, did not diminish the popularity of this technique. Galton used it (in 1886) in his work on the heritability of size which laid down the foundations of correlation and (also gave the name) regression analysis. Both Pearson and Fisher, who did so much in the early development of statistics, used and developed it in different contexts (factor analysis for Pearson and experimental design for Fisher).

Nowadays, the least square method is widely used to find or estimate the numerical values of the parameters to fit a function to a set of data and to characterize the statistical properties of estimates. It exists with several variations: Its simpler version is called ordinary least squares (OLS), a more sophisticated version is called weighted least squares (WLS), which often performs better than OLS because it can modulate the importance of each observation in the final solution. Recent variations of the least square method are alternating least squares (ALS) and partial least squares (PLS).

Functional fit example: regression

The oldest (and still most frequent) use of OLS was linear regression, which corresponds to the problem of finding a line (or curve) that best fits a set of data. In the standard formulation, a set of N pairs of observations $\{Y_i, X_i\}$ is used to find a function giving the value of the dependent variable (Y) from the

[1]In: Lewis-Beck M., Bryman, A., Futing T. (Eds.) (2003). *Encyclopedia of Social Sciences Research Methods*. Thousand Oaks (CA): Sage.
Address correspondence to
Hervé Abdi
Program in Cognition and Neurosciences, MS: Gr.4.1,
The University of Texas at Dallas,
Richardson, TX 75083–0688, USA
E-mail: herve@utdallas.edu http://www.utdallas.edu/∼herve

values of an independent variable (X). With one variable and a linear function, the prediction is given by the following equation:

$$\hat{Y} = a + bX. \tag{1}$$

This equation involves two free parameters which specify the intercept (a) and the slope (b) of the regression line. The least square method defines the estimate of these parameters as the values which minimize the sum of the squares (hence the name *least squares*) between the measurements and the model (i.e., the predicted values). This amounts to minimizing the expression:

$$\mathcal{E} = \sum_i (Y_i - \hat{Y}_i)^2 = \sum_i \left[Y_i - (a + bX_i)\right]^2 \tag{2}$$

(where $\mathcal{E}$ stands for "error" which is the quantity to be minimized). This is achieved using standard techniques from calculus, namely the property that a quadratic (i.e., with a square) formula reaches its minimum value when its derivatives vanish. Taking the derivative of $\mathcal{E}$ with respect to a and b and setting them to zero gives the following set of equations (called the *normal* equations):

$$\frac{\partial \mathcal{E}}{\partial a} = 2Na + 2b\sum X_i - 2\sum Y_i = 0 \tag{3}$$

and

$$\frac{\partial \mathcal{E}}{\partial b} = 2b\sum X_i^2 + 2a\sum X_i - 2\sum Y_i X_i = 0\ . \tag{4}$$

Solving these 2 equations gives the least square estimates of a and b as:

$$a = M_Y - bM_X \tag{5}$$

(with M_Y and M_X denoting the means of X and Y) and

$$b = \frac{\sum (Y_i - M_Y)(X_i - M_X)}{\sum (X_i - M_X)^2}\ . \tag{6}$$

OLS can be extended to more than one independent variable (using matrix algebra) and to non-linear functions.

The geometry of least squares

OLS can be interpreted in a geometrical framework as an orthogonal projection of the data vector onto the space defined by the independent variable. The projection is orthogonal because the predicted values and the actual values are uncorrelated. This is illustrated in Figure 1, which depicts the case of two independent variables (vectors $\mathbf{x}_1$ and $\mathbf{x}_2$) and the data vector ($\mathbf{y}$), and shows that the error vector ($\mathbf{y}_1 - \hat{\mathbf{y}}$) is orthogonal to the least square ($\hat{\mathbf{y}}$) estimate which lies in the subspace defined by the two independent variables.

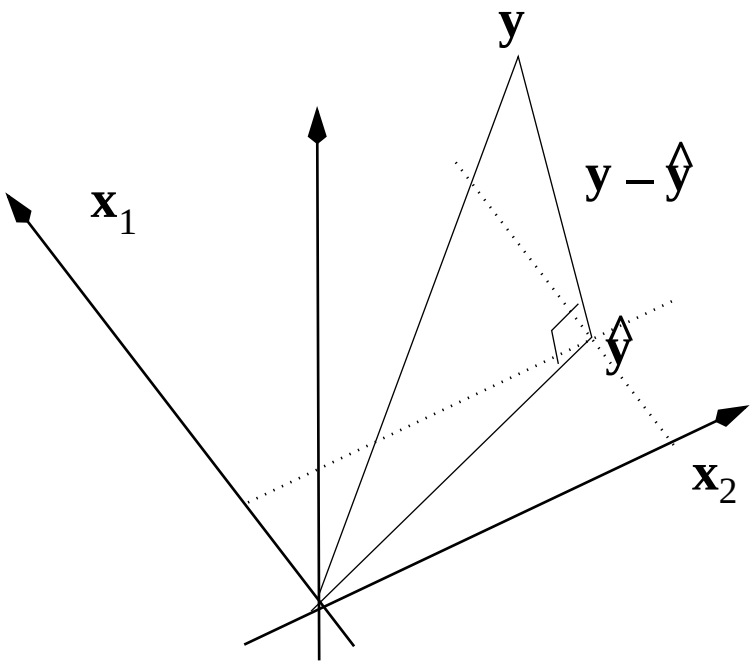

Figure 1: The least square estimate of the data is the orthogonal projection of the data vector onto the independent variable subspace.

Optimality of least square estimates

OLS estimates have some strong statistical properties. Specifically when (1) the data obtained constitute a random sample from a well-defined population, (2) the population model is linear, (3) the error has a zero expected value, (4) the independent variables are linearly independent, and (5) the error is normally distributed and uncorrelated with the independent variables (the so-called homoscedasticity assumption); then the OLS estimate is the *b*est *l*inear *u*nbiased *e*stimate often denoted with the acronym "BLUE" (the 5 conditions and the proof are called the Gauss-Markov conditions and theorem). In addition, when the Gauss-Markov conditions hold, OLS estimates are also MAXIMUM LIKELIHOOD estimates.

Weighted least squares

The optimality of OLS relies heavily on the homoscedasticity assumption. When the data come from different sub-populations for which an independent estimate of the error variance is available, a better estimate than OLS can be obtained using weighted least squares (WLS), also called generalized least squares (GLS). The idea is to assign to each observation a weight that reflects the uncertainty of the measurement. In general, the weight w_i, assigned to the ith observation, will be a function of the variance of this observation, denoted

σ_i^2. A straightforward weighting schema is to define $w_i = \sigma_i^{-1}$ (but other more sophisticated weighted schemes can also be proposed). For the linear regression example, WLS will find the values of a and b minimizing:

$$\mathcal{E}_w = \sum_i w_i (Y_i - \hat{Y}_i)^2 = \sum_i w_i \left[Y_i - (a + bX_i)\right]^2 \ . \tag{7}$$

Iterative methods: Gradient descent

When estimating the parameters of a nonlinear function with OLS or WLS, the standard approach using derivatives is not always possible. In this case, iterative methods are very often used. These methods search in a stepwise fashion for the best values of the estimate. Often they proceed by using at each step a linear approximation of the function and refine this approximation by successive corrections. The techniques involved are known as gradient descent and Gauss-Newton approximations. They correspond to nonlinear least squares approximation in numerical analysis and nonlinear regression in statistics. NEURAL NETWORKS constitutes a popular recent application of these techniques

**References*

[1] Bates, D.M. & Watts D.G. (1988). *Nonlinear regression analysis and its applications*. New York: Wiley

[2] Greene, W.H. (2002). *Econometric analysis*. New York: Prentice Hall.

[3] Nocedal J. & Wright, S. (1999). *Numerical optimization*. New York: Springer.

[4] Plackett, R.L. (1972). The discovery of the method of least squares. *Biometrika*, **59**, 239–251.

[5] Seal, H.L. (1967). The historical development of the Gauss linear model. *Biometrika*, **54**, 1–23.