Least Squares Estimation

Sara A. van de Geer

Volume 2, pp. 1041–1045

in

Encyclopedia of Statistics in Behavioral Science

ISBN-13: 978-0-470-86080-9
ISBN-10: 0-470-86080-4

Editors

Brian S. Everitt & David C. Howell

Least Squares Estimation

The method of least squares is about estimating parameters by minimizing the squared discrepancies between observed data, on the one hand, and their expected values on the other (*see* **Optimization Methods**). We will study the method in the context of a regression problem, where the variation in one variable, called the response variable Y, can be partly explained by the variation in the other variables, called covariables X (*see* **Multiple Linear Regression**). For example, variation in exam results Y are mainly caused by variation in abilities and diligence X of the students, or variation in survival times Y (*see* **Survival Analysis**) are primarily due to variations in environmental conditions X. Given the value of X, the best prediction of Y (in terms of mean square error – *see* **Estimation**) is the mean $f(X)$ of Y given X. We say that Y is a function of X plus noise:

$$Y = f(X) + \text{noise}.$$

The function f is called a regression function. It is to be estimated from sampling n covariables and their responses $(x_1, y_1), \ldots, (x_n, y_n)$.

Suppose f is known up to a finite number $p \le n$ of parameters $\beta = (\beta_1, \ldots, \beta_p)'$, that is, $f = f_\beta$. We estimate β by the value $\hat{\beta}$ that gives the best fit to the data. The least squares estimator, denoted by $\hat{\beta}$, is that value of b that minimizes

$$\sum_{i=1}^{n} (y_i - f_b(x_i))^2, \tag{1}$$

over all possible b.

The least squares criterion is a computationally convenient measure of fit. It corresponds to **maximum likelihood estimation** when the noise is normally distributed with equal variances. Other measures of fit are sometimes used, for example, least absolute deviations, which is more robust against **outliers**. (*See* **Robust Testing Procedures**).

Linear Regression. Consider the case where f_β is a linear function of β, that is,

$$f_\beta(X) = X_1\beta_1 + \cdots + X_p\beta_p. \tag{2}$$

Here $(X_1, \ldots, X_p)$ stand for the observed variables used in $f_\beta(X)$.

To write down the least squares estimator for the linear regression model, it will be convenient to use matrix notation. Let $\mathbf{y} = (y_1, \ldots, y_n)'$ and let $\mathbf{X}$ be the $n \times p$ data matrix of the n observations on the p variables

$$\mathbf{X} = \begin{pmatrix} \mathbf{x}_{1,1} & \cdots & \mathbf{x}_{1,p} \\ \vdots & \ldots & \vdots \\ \mathbf{x}_{n,1} & \cdots & \mathbf{x}_{n,p} \end{pmatrix} = (\, \mathbf{x}_1 \quad \ldots \quad \mathbf{x}_p \,), \tag{3}$$

where $\mathbf{x}_j$ is the column vector containing the n observations on variable j, $j = 1, \ldots, n$. Denote the squared length of an n-dimensional vector $\mathbf{v}$ by $\|\mathbf{v}\|^2 = \mathbf{v}'\mathbf{v} = \sum_{i=1}^{n} v_i^2$. Then expression (1) can be written as

$$\|\mathbf{y} - \mathbf{X}b\|^2,$$

which is the squared distance between the vector $\mathbf{y}$ and the linear combination b of the columns of the matrix $\mathbf{X}$. The distance is minimized by taking the *projection* of $\mathbf{y}$ on the space spanned by the columns of $\mathbf{X}$ (see Figure 1).

Suppose now that $\mathbf{X}$ has full column rank, that is, no column in $\mathbf{X}$ can be written as a linear combination of the other columns. Then, the least squares estimator $\hat{\beta}$ is given by

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\,\mathbf{X}'\mathbf{y}. \tag{4}$$

The Variance of the Least Squares Estimator. In order to construct **confidence intervals** for the components of $\hat{\beta}$, or linear combinations of these components, one needs an estimator of the covariance

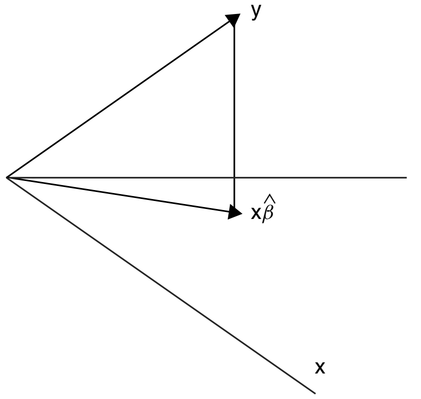

Figure 1 The projection of the vector $\mathbf{y}$ on the plane spanned by $\mathbf{X}$

matrix of $\hat{\beta}$. Now, it can be shown that, given $\mathbf{X}$, the covariance matrix of the estimator $\hat{\beta}$ is equal to

$$(\mathbf{X}'\mathbf{X})^{-1}\sigma^2.$$

where σ^2 is the variance of the noise. As an estimator of σ^2, we take

$$\hat{\sigma}^2 = \frac{1}{n-p}\|\mathbf{y} - \mathbf{X}\hat{\beta}\|^2 = \frac{1}{n-p}\sum_{i=1}^{n}\hat{e}_i^2, \tag{5}$$

where the $\hat{e}_i$ are the residuals

$$\hat{e}_i = y_i - \mathbf{x}_{i,1}\hat{\beta}_1 - \cdots - \mathbf{x}_{i,p}\hat{\beta}_p. \tag{6}$$

The covariance matrix of $\hat{\beta}$ can, therefore, be estimated by

$$(\mathbf{X}'\mathbf{X})^{-1}\hat{\sigma}^2.$$

For example, the estimate of the variance of $\hat{\beta}_j$ is

$$\hat{\text{var}}(\hat{\beta}_j) = \tau_j^2\hat{\sigma}^2,$$

where τ_j^2 is the jth element on the diagonal of $(\mathbf{X}'\mathbf{X})^{-1}$. A confidence interval for β_j is now obtained by taking the least squares estimator $\hat{\beta}_j \pm$ a margin:

$$\hat{\beta}_j \pm c\sqrt{\hat{\text{var}}(\hat{\beta}_j)}, \tag{7}$$

where c depends on the chosen confidence level. For a 95% confidence interval, the value $c = 1.96$ is a good approximation when n is large. For smaller values of n, one usually takes a more conservative c using the tables for the student distribution with $n - p$ degrees of freedom.

Numerical Example. Consider a regression with constant, linear and quadratic terms:

$$f_\beta(X) = \beta_1 + X\beta_2 + X^2\beta_3. \tag{8}$$

We take $n = 100$ and $x_i = i/n$, $i = 1, \ldots, n$. The matrix $\mathbf{X}$ is now

$$\mathbf{X} = \begin{pmatrix} 1 & x_1 & x_1^2 \\ \vdots & \vdots & \vdots \\ 1 & x_n & x_n^2 \end{pmatrix}. \tag{9}$$

This gives

$$\mathbf{X}'\mathbf{X} = \begin{pmatrix} 100 & 50.5 & 33.8350 \\ 50.5 & 33.8350 & 25.5025 \\ 33.8350 & 25.5025 & 20.5033 \end{pmatrix},$$
$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{pmatrix} 0.0937 & -0.3729 & 0.3092 \\ -0.3729 & 1.9571 & -1.8189 \\ 0.3092 & -1.8189 & 1.8009 \end{pmatrix}. \tag{10}$$

We simulated n independent standard normal random variables $e_1, \ldots, e_n$, and calculated for $i = 1, \ldots, n$,

$$y_i = 1 - 3x_i + e_i. \tag{11}$$

Thus, in this example, the parameters are

$$\begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} = \begin{pmatrix} 1 \\ -3 \\ 0 \end{pmatrix}. \tag{12}$$

Moreover, $\sigma^2 = 1$. Because this is a simulation, these values are known.

To calculate the least squares estimator, we need the values of $\mathbf{X}'\mathbf{y}$, which, in this case, turn out to be

$$\mathbf{X}'\mathbf{y} = \begin{pmatrix} -64.2007 \\ -52.6743 \\ -42.2025 \end{pmatrix}. \tag{13}$$

The least squares estimate is thus

$$\hat{\beta} = \begin{pmatrix} 0.5778 \\ -2.3856 \\ -0.0446 \end{pmatrix}. \tag{14}$$

From the data, we also calculated the estimated variance of the noise, and found the value

$$\hat{\sigma}^2 = 0.883. \tag{15}$$

The data are represented in Figure 2. The dashed line is the true regression $f_\beta(x)$. The solid line is the estimated regression $f_{\hat{\beta}}(x)$.

The estimated regression is barely distinguishable from a straight line. Indeed, the value $\hat{\beta}_3 = -0.0446$ of the quadratic term is small. The estimated variance of $\hat{\beta}_3$ is

$$\hat{\text{var}}(\hat{\beta}_3) = 1.8009 \times 0.883 = 1.5902. \tag{16}$$

Using $c = 1.96$ in (7), we find the confidence interval

$$\beta_3 \in -0.0446 \pm 1.96\sqrt{1.5902} = [-2.5162, 2.470]. \tag{17}$$

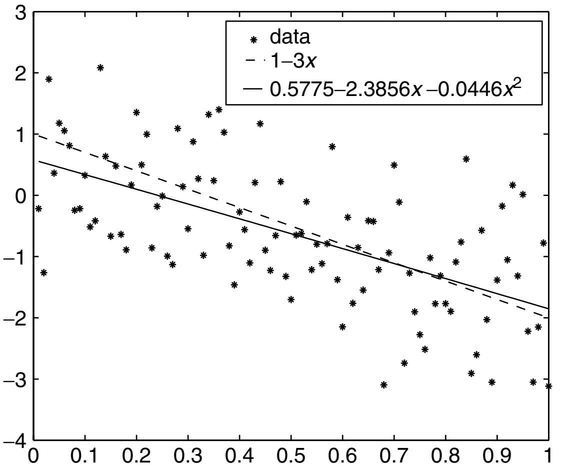

Figure 2 Observed data, true regression (dashed line), and least squares estimate (solid line)

Thus, β_3 is not significantly different from zero at the 5% level, and, hence, we do not reject the hypothesis H_0: $\beta_3 = 0$.

Below, we will consider general test statistics for testing hypotheses on β. In this particular case, the test statistic takes the form

$$T^2 = \frac{\hat{\beta}_3^2}{\hat{\text{var}}(\hat{\beta}_3)} = 0.0012. \qquad (18)$$

Using this test statistic is equivalent to the above method based on the confidence interval. Indeed, as $T^2 < (1.96)^2$, we do not reject the hypothesis $H_0 : \beta_3 = 0$.

Under the hypothesis $H_0 : \beta_3 = 0$, we use the least squares estimator

$$\begin{pmatrix} \hat{\beta}_{1,0} \\ \hat{\beta}_{2,0} \end{pmatrix} = (\mathbf{X}_0'\mathbf{X}_0)^{-1}\mathbf{X}_0'\mathbf{y} = \begin{pmatrix} 0.5854 \\ -2.4306 \end{pmatrix}. \qquad (19)$$

Here,

$$\mathbf{X}_0 = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}. \qquad (20)$$

It is important to note that setting β_3 to zero changes the values of the least squares estimates of β_1 and β_2:

$$\begin{pmatrix} \hat{\beta}_{1,0} \\ \hat{\beta}_{2,0} \end{pmatrix} \neq \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix}. \qquad (21)$$

This is because $\hat{\beta}_3$ is correlated with $\hat{\beta}_1$ and $\hat{\beta}_2$. One may verify that the correlation matrix of $\hat{\beta}$ is

$$\begin{pmatrix} 1 & -0.8708 & 0.7529 \\ -0.8708 & 1 & -0.9689 \\ 0.7529 & -0.9689 & 1 \end{pmatrix}.$$

Testing Linear Hypotheses. The testing problem considered in the numerical example is a special case of testing a linear hypothesis $H_0 : A\beta = 0$, where A is some $r \times p$ matrix. As another example of such a hypothesis, suppose we want to test whether two coefficients are equal, say $H_0 : \beta_1 = \beta_2$. This means there is one restriction $r = 1$, and we can take A as the $1 \times p$ row vector

$$A = (1, -1, 0, \ldots, 0). \qquad (22)$$

In general, we assume that there are no linear dependencies in the r restrictions $A\beta = 0$. To test the linear hypothesis, we use the statistic

$$T^2 = \frac{\|\mathbf{X}\hat{\beta}_0 - \mathbf{X}\hat{\beta}\|^2/r}{\hat{\sigma}^2}, \qquad (23)$$

where $\hat{\beta}_0$ is the least squares estimator under $H_0 : A\beta = 0$. In the numerical example, this statistic takes the form given in (18). When the noise is normally distributed, critical values can be found in a table for the F distribution with r and $n - p$ degrees of freedom. For large n, approximate critical values are in the table of the χ^2 distribution with r degrees of freedom.

Some Extensions

Weighted Least Squares. In many cases, the variance σ_i^2 of the noise at measurement i depends on x_i. Observations where σ_i^2 is large are less accurate, and, hence, should play a smaller role in the estimation of β. The *weighted* least squares estimator is that value of b that minimizes the criterion

$$\sum_{i=1}^{n} \frac{(y_i - f_b(x_i))^2}{\sigma_i^2}.$$

overall possible b. In the linear case, this criterion is numerically of the same form, as we can make the change of variables $\tilde{y}_i = y_i/\sigma_i$ and $\tilde{\mathbf{x}}_{i,j} = \mathbf{x}_{i,j}/\sigma_i$.

The minimum χ^2-estimator (*see* **Estimation**) is an example of a weighted least squares estimator in the context of density estimation.

Nonlinear Regression. When f_β is a nonlinear function of β, one usually needs iterative algorithms to find the least squares estimator. The variance can then be approximated as in the linear case, with $\dot{f}_{\hat{\beta}}(x_i)$ taking the role of the rows of **X**. Here, $\dot{f}_\beta(x_i) = \partial f_\beta(x_i)/\partial\beta$ is the row vector of derivatives of $f_\beta(x_i)$. For more details, see e.g. [4].

Nonparametric Regression. In nonparametric regression, one only assumes a certain amount of smoothness for f (e.g., as in [1]), or alternatively, certain qualitative assumptions such as monotonicity (see [3]). Many nonparametric least squares procedures have been developed and their numerical and theoretical behavior discussed in literature. Related developments include estimation methods for models where the number of parameters p is about as large as the number of observations n. The *curse of dimensionality* in such models is handled by applying various complexity regularization techniques (see e.g., [2]).

References

[1] Green, P.J. & Silverman, B.W. (1994). *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach*, Chapman & Hall, London.

[2] Hastie, T., Tibshirani, R. & Friedman, J. (2001). *The Elements of Statistical Learning. Data Mining, Inference and Prediction*, Springer, New York.

[3] Robertson, T., Wright, F.T. & Dykstra, R.L. (1988). *Order Restricted Statistical Inference*, Wiley, New York.

[4] Seber, G.A.F. & Wild, C.J. (2003). *Nonlinear Regression*, Wiley, New York.

Sara A. van de Geer