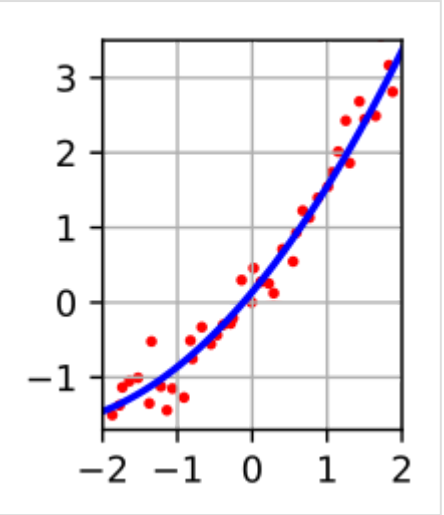
Least squares

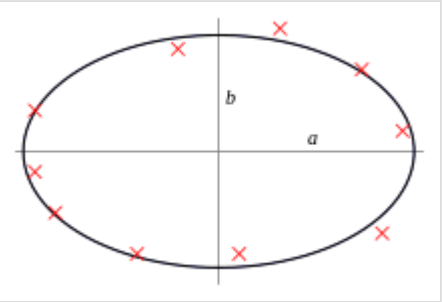
The method of **least squares** is a parameter estimation method in regression analysis based on minimizing the sum of the squares of the residuals (a residual being the difference between an observed value and the fitted value provided by a model) made in the results of each individual equation.



The result of fitting a set of data points with a quadratic function

The most important application is in data fitting. When the problem has substantial uncertainties in the independent variable (the x variable), then simple regression and least-squares methods have problems; in such cases, the methodology required for fitting errors-in-variables models may be considered instead of that for least squares.

Least squares problems fall into two categories: linear or ordinary least squares and nonlinear least squares, depending on whether or not the residuals are linear in all unknowns. The linear least-squares problem occurs in statistical regression analysis; it has a closed-form solution. The nonlinear problem is usually solved by iterative refinement; at each iteration the system is approximated by a linear one, and thus the core calculation is similar in both cases.



Conic fitting a set of points using least-squares approximation

Polynomial least squares describes the variance in a prediction of the dependent variable as a function of the independent variable and the deviations from the fitted curve.

When the observations come from an exponential family with identity as its natural sufficient statistics and mild-conditions are satisfied (e.g. for normal, exponential, Poisson and binomial distributions), standardized least-squares estimates and maximum-likelihood estimates are identical.[1] The method of least squares can also be derived as a method of moments estimator.

The following discussion is mostly presented in terms of linear functions but the use of least squares is valid and practical for more general families of functions. Also, by iteratively applying local quadratic approximation to the likelihood (through the Fisher information), the least-squares method may be used to fit a generalized linear model.

The least-squares method was officially discovered and published by Adrien-Marie Legendre (1805),[2] though it is usually also co-credited to Carl Friedrich Gauss (1809),[3][4] who contributed significant theoretical advances to the method,[4] and may have also used it in his earlier work in 1794 and 1795.[5][4]

History

Founding

The method of least squares grew out of the fields of astronomy and geodesy, as scientists and mathematicians sought to provide solutions to the challenges of navigating the Earth's oceans during the Age of Discovery. The accurate description of the behavior of celestial bodies was the key to enabling ships to sail in open seas, where sailors could no longer rely on land sightings for navigation.

The method was the culmination of several advances that took place during the course of the eighteenth century:[6]

- The combination of different observations as being the best estimate of the true value; errors decrease with aggregation rather than increase, perhaps first expressed by Roger Cotes in 1722.
- The combination of different observations taken under the *same* conditions contrary to simply trying one's best to observe and record a single observation accurately. The approach was known as the method of averages. This approach was notably used by Tobias Mayer while studying the librations of the Moon in 1750, and by Pierre-Simon Laplace in his work in explaining the differences in motion of Jupiter and Saturn in 1788.
- The combination of different observations taken under *different* conditions. The method came to be known as the method of *least absolute deviation*. It was notably performed by Roger Joseph Boscovich in his work on the shape of the Earth in 1757 and by Pierre-Simon Laplace for the same problem in 1789 and 1799.
- The development of a criterion that can be evaluated to determine when the solution with the minimum error has been achieved. Laplace tried to specify a mathematical form of the probability density for the errors and define a method of estimation that minimizes the error of estimation. For this purpose, Laplace used a symmetric two-sided exponential distribution we now call Laplace distribution to model the error distribution, and used the sum of absolute deviation as error of estimation. He felt these to be the simplest assumptions he could make, and he had hoped to obtain the arithmetic mean as the best estimate. Instead, his estimator was the posterior median.

The method


Carl Friedrich Gauss

The first clear and concise exposition of the method of least squares was published by Legendre in 1805.[7] The technique is described as an algebraic procedure for fitting linear equations to data and Legendre demonstrates the new method by analyzing the same data as Laplace for the shape of the Earth. Within ten years after Legendre's publication, the method of least squares had been adopted as a standard tool in astronomy and geodesy in France, Italy, and Prussia, which constitutes an extraordinarily rapid acceptance of a scientific technique.[6]

In 1809 Carl Friedrich Gauss published his method of calculating the orbits of celestial bodies. In that work he claimed to have been in possession of the method of least squares since 1795.[8] This naturally led to a priority dispute with Legendre. However, to Gauss's credit, he went beyond Legendre and succeeded in connecting the method of least squares with the principles of probability and to the normal distribution. He had managed to complete Laplace's program of specifying a mathematical form of the probability density for the observations, depending on a finite number of unknown parameters, and define a method of estimation that minimizes the error of estimation. Gauss showed that the arithmetic mean is indeed the best estimate of the location parameter by changing both the probability density and the method of estimation. He then turned the

problem around by asking what form the density should have and what method of estimation should be used to get the arithmetic mean as estimate of the location parameter. In this attempt, he invented the normal distribution.

An early demonstration of the strength of Gauss's method came when it was used to predict the future location of the newly discovered asteroid Ceres. On 1 January 1801, the Italian astronomer Giuseppe Piazzi discovered Ceres and was able to track its path for 40 days before it was lost in the glare of the Sun. Based on these data, astronomers desired to determine the location of Ceres after it emerged from behind the Sun without solving Kepler's complicated nonlinear equations of planetary motion. The only predictions that successfully allowed Hungarian astronomer Franz Xaver von Zach to relocate Ceres were those performed by the 24-year-old Gauss using least-squares analysis.

In 1810, after reading Gauss's work, Laplace, after proving the central limit theorem, used it to give a large sample justification for the method of least squares and the normal distribution. In 1822, Gauss was able to state that the least-squares approach to regression analysis is optimal in the sense that in a linear model where the errors have a mean of zero, are uncorrelated, and have equal variances, the best linear unbiased estimator of the coefficients is the least-squares estimator. This result is known as the Gauss–Markov theorem.

The idea of least-squares analysis was also independently formulated by the American Robert Adrain in 1808. In the next two centuries workers in the theory of errors and in statistics found many different ways of implementing least squares.[9]

Problem statement

The objective consists of adjusting the parameters of a model function to best fit a data set. A simple data set consists of n points (data pairs) (x_i, y_i), $i = 1, \ldots, n$, where x_i is an independent variable and y_i is a dependent variable whose value is found by observation. The model function has the form $f(x, \boldsymbol{\beta})$, where m adjustable parameters are held in the vector $\boldsymbol{\beta}$. The goal is to find the parameter values for the model that "best" fits the data. The fit of a model to a data point is measured by its residual, defined as the difference between the observed value of the dependent variable and the value predicted by the model:

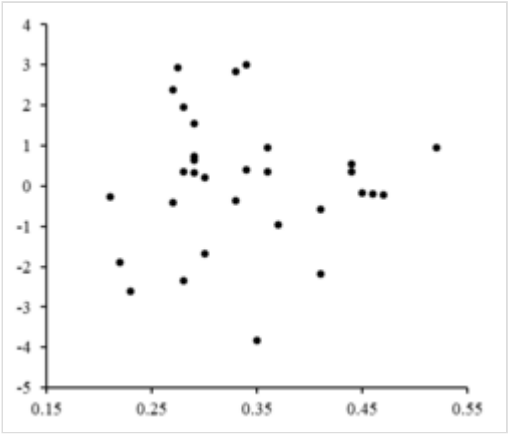
$$r_i = y_i - f(x_i, \boldsymbol{\beta}).$$

The least-squares method finds the optimal parameter values by minimizing the sum of squared residuals, S:[10]

$$S = \sum_{i=1}^{n} r_i^2.$$

In the simplest case $f(x_i, \boldsymbol{\beta}) = \beta$ and the result of the least-squares method is the arithmetic mean of the input data.

An example of a model in two dimensions is that of the straight line. Denoting the y-intercept as β_0 and the slope as β_1, the model function is given by $f(x, \boldsymbol{\beta}) = \beta_0 + \beta_1 x$. See linear least squares for a fully worked out example of this model.

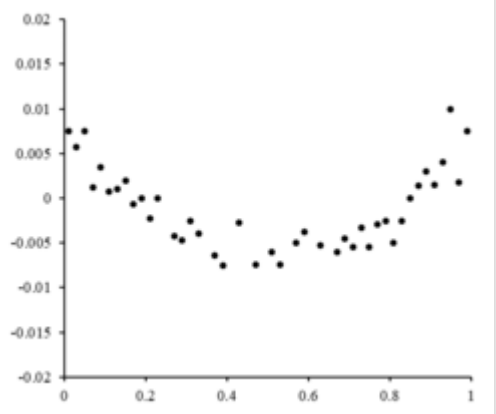


The residuals are plotted against corresponding x values. The random fluctuations about $r_i = 0$ indicate a linear model is appropriate.

A data point may consist of more than one independent variable. For example, when fitting a plane to a set of height measurements, the plane is a function of two independent variables, *x* and *z*, say. In the most general case there may be one or more independent variables and one or more dependent variables at each data point.

To the right is a residual plot illustrating random fluctuations about $r_i = 0$, indicating that a linear model $(Y_i = \alpha + \beta x_i + U_i)$ is appropriate. U_i is an independent, random variable.[10]

If the residual points had some sort of a shape and were not randomly fluctuating, a linear model would not be appropriate. For example, if the residual plot had a parabolic shape as seen to the right, a parabolic model $(Y_i = \alpha + \beta x_i + \gamma x_i^2 + U_i)$ would be appropriate for the data. The residuals for a parabolic model can be calculated via $r_i = y_i - \hat{\alpha} - \hat{\beta} x_i - \hat{\gamma} x_i^2$.[10]



The residuals are plotted against the corresponding x values. The parabolic shape of the fluctuations about $r_i = 0$ indicates a parabolic model is appropriate.

Limitations

This regression formulation considers only observational errors in the dependent variable (but the alternative total least squares regression can account for errors in both variables). There are two rather different contexts with different implications:

- Regression for prediction. Here a model is fitted to provide a prediction rule for application in a similar situation to which the data used for fitting apply. Here the dependent variables corresponding to such future application would be subject to the same types of observation error as those in the data used for fitting. It is therefore logically consistent to use the least-squares prediction rule for such data.
- Regression for fitting a "true relationship". In standard regression analysis that leads to fitting by least squares there is an implicit assumption that errors in the independent variable are zero or strictly controlled so as to be negligible. When errors in the independent variable are non-negligible, models of measurement error can be used; such methods can lead to parameter estimates, hypothesis testing and confidence intervals that take into account the presence of observation errors in the independent variables.[11] An alternative approach is to fit a model by total least squares; this can be viewed as taking a pragmatic approach to balancing the effects of the different sources of error in formulating an objective function for use in model-fitting.

Solving the least squares problem

The minimum of the sum of squares is found by setting the gradient to zero. Since the model contains *m* parameters, there are *m* gradient equations:

$$\frac{\partial S}{\partial \beta_j} = 2 \sum_i r_i \frac{\partial r_i}{\partial \beta_j} = 0, \ j = 1, \ldots, m,$$

and since $r_i = y_i - f(x_i, \boldsymbol{\beta})$, the gradient equations become

$$-2 \sum_i r_i \frac{\partial f(x_i, \boldsymbol{\beta})}{\partial \beta_j} = 0, \ j = 1, \ldots, m.$$

The gradient equations apply to all least squares problems. Each particular problem requires particular expressions for the model and its partial derivatives.[12]

Linear least squares

A regression model is a linear one when the model comprises a linear combination of the parameters, i.e.,

$$f(x, \boldsymbol{\beta}) = \sum_{j=1}^{m} \beta_j \phi_j(x),$$

where the function ϕ_j is a function of x.[12]

Letting $X_{ij} = \phi_j(x_i)$ and putting the independent and dependent variables in matrices X and Y, respectively, we can compute the least squares in the following way. Note that D is the set of all data.[12][13]

$$L(D, \boldsymbol{\beta}) = \|Y - X\boldsymbol{\beta}\|^2 = (Y - X\boldsymbol{\beta})^{\mathsf{T}}(Y - X\boldsymbol{\beta}) = Y^{\mathsf{T}}Y - Y^{\mathsf{T}}X\boldsymbol{\beta} - \boldsymbol{\beta}^{\mathsf{T}}X^{\mathsf{T}}Y + \boldsymbol{\beta}^{\mathsf{T}}X^{\mathsf{T}}X\boldsymbol{\beta}$$

The gradient of the loss is:

$$\frac{\partial L(D, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \frac{\partial \left(Y^{\mathsf{T}}Y - Y^{\mathsf{T}}X\boldsymbol{\beta} - \boldsymbol{\beta}^{\mathsf{T}}X^{\mathsf{T}}Y + \boldsymbol{\beta}^{\mathsf{T}}X^{\mathsf{T}}X\boldsymbol{\beta}\right)}{\partial \boldsymbol{\beta}} = -2X^{\mathsf{T}}Y + 2X^{\mathsf{T}}X\boldsymbol{\beta}$$

Setting the gradient of the loss to zero and solving for $\boldsymbol{\beta}$, we get:[13][12]

$$-2X^{\mathsf{T}}Y + 2X^{\mathsf{T}}X\boldsymbol{\beta} = 0 \Rightarrow X^{\mathsf{T}}Y = X^{\mathsf{T}}X\boldsymbol{\beta}$$

$$\hat{\boldsymbol{\beta}} = \left(X^{\mathsf{T}}X\right)^{-1}X^{\mathsf{T}}Y$$

Non-linear least squares

There is, in some cases, a closed-form solution to a non-linear least squares problem – but in general there is not. In the case of no closed-form solution, numerical algorithms are used to find the value of the parameters $\boldsymbol{\beta}$ that minimizes the objective. Most algorithms involve choosing initial values for the parameters. Then, the parameters are refined iteratively, that is, the values are obtained by successive approximation:

$$\beta_j{}^{k+1} = \beta_j{}^{k} + \Delta\beta_j,$$

where a superscript *k* is an iteration number, and the vector of increments $\Delta\beta_j$ is called the shift vector. In some commonly used algorithms, at each iteration the model may be linearized by approximation to a first-

order Taylor series expansion about $\boldsymbol{\beta}^k$:

$$\begin{aligned} f(x_i, \boldsymbol{\beta}) &= f^k(x_i, \boldsymbol{\beta}) + \sum_j \frac{\partial f(x_i, \boldsymbol{\beta})}{\partial \beta_j} \left(\beta_j - {\beta_j}^k \right) \\ &= f^k(x_i, \boldsymbol{\beta}) + \sum_j J_{ij}\, \Delta\beta_j. \end{aligned}$$

The Jacobian **J** is a function of constants, the independent variable *and* the parameters, so it changes from one iteration to the next. The residuals are given by

$$r_i = y_i - f^k(x_i, \boldsymbol{\beta}) - \sum_{k=1}^{m} J_{ik}\, \Delta\beta_k = \Delta y_i - \sum_{j=1}^{m} J_{ij}\, \Delta\beta_j.$$

To minimize the sum of squares of r_i, the gradient equation is set to zero and solved for $\Delta\beta_j$:

$$-2 \sum_{i=1}^{n} J_{ij} \left(\Delta y_i - \sum_{k=1}^{m} J_{ik}\, \Delta\beta_k \right) = 0,$$

which, on rearrangement, become *m* simultaneous linear equations, the **normal equations**:

$$\sum_{i=1}^{n} \sum_{k=1}^{m} J_{ij} J_{ik}\, \Delta\beta_k = \sum_{i=1}^{n} J_{ij}\, \Delta y_i \qquad (j = 1, \ldots, m).$$

The normal equations are written in matrix notation as

$$\left(\mathbf{J}^\mathsf{T} \mathbf{J}\right) \Delta\boldsymbol{\beta} = \mathbf{J}^\mathsf{T} \Delta\mathbf{y}.$$

These are the defining equations of the Gauss–Newton algorithm.

Differences between linear and nonlinear least squares

- The model function, *f*, in LLSQ (linear least squares) is a linear combination of parameters of the form $f = X_{i1}\beta_1 + X_{i2}\beta_2 + \cdots$ The model may represent a straight line, a parabola or any other linear combination of functions. In NLLSQ (nonlinear least squares) the parameters appear as functions, such as $\beta^2, e^{\beta x}$ and so forth. If the derivatives $\partial f / \partial \beta_j$ are either constant or depend only on the values of the independent variable, the model is linear in the parameters. Otherwise the model is nonlinear.
- Need initial values for the parameters to find the solution to a NLLSQ problem; LLSQ does not require them.
- Solution algorithms for NLLSQ often require that the Jacobian can be calculated similar to LLSQ. Analytical expressions for the partial derivatives can be complicated. If analytical expressions are impossible to obtain either the partial derivatives must be calculated by

numerical approximation or an estimate must be made of the Jacobian, often via finite differences.
- Non-convergence (failure of the algorithm to find a minimum) is a common phenomenon in NLLSQ.
- LLSQ is globally concave so non-convergence is not an issue.
- Solving NLLSQ is usually an iterative process which has to be terminated when a convergence criterion is satisfied. LLSQ solutions can be computed using direct methods, although problems with large numbers of parameters are typically solved with iterative methods, such as the Gauss–Seidel method.
- In LLSQ the solution is unique, but in NLLSQ there may be multiple minima in the sum of squares.
- Under the condition that the errors are uncorrelated with the predictor variables, LLSQ yields unbiased estimates, but even under that condition NLLSQ estimates are generally biased.

These differences must be considered whenever the solution to a nonlinear least squares problem is being sought.[12]

Example

Consider a simple example drawn from physics. A spring should obey Hooke's law which states that the extension of a spring *y* is proportional to the force, *F*, applied to it.

$$y = f(F, k) = kF$$

constitutes the model, where *F* is the independent variable. In order to estimate the force constant, *k*, we conduct a series of *n* measurements with different forces to produce a set of data, $(F_i, y_i),\ i = 1, \ldots, n$, where y_i is a measured spring extension.[14] Each experimental observation will contain some error, ε, and so we may specify an empirical model for our observations,

$$y_i = kF_i + \varepsilon_i.$$

There are many methods we might use to estimate the unknown parameter *k*. Since the *n* equations in the *m* variables in our data comprise an overdetermined system with one unknown and *n* equations, we estimate *k* using least squares. The sum of squares to be minimized is

$$S = \sum_{i=1}^{n} (y_i - kF_i)^2.$$[12]

The least squares estimate of the force constant, *k*, is given by

$$\hat{k} = \frac{\sum_i F_i y_i}{\sum_i F_i^2}.$$

We assume that applying force ***causes*** the spring to expand. After having derived the force constant by least squares fitting, we predict the extension from Hooke's law.

Uncertainty quantification

In a least squares calculation with unit weights, or in linear regression, the variance on the *j*th parameter, denoted $\operatorname{var}(\hat{\beta}_j)$, is usually estimated with

$$\operatorname{var}(\hat{\beta}_j) = \sigma^2 \left(\left[X^\mathsf{T} X \right]^{-1} \right)_{jj} \approx \hat{\sigma}^2 C_{jj},$$
$$\hat{\sigma}^2 \approx \frac{S}{n-m}$$
$$C = \left(X^\mathsf{T} X \right)^{-1},$$

where the true error variance σ^2 is replaced by an estimate, the reduced chi-squared statistic, based on the minimized value of the residual sum of squares (objective function), S. The denominator, $n - m$, is the statistical degrees of freedom; see effective degrees of freedom for generalizations.[12] C is the covariance matrix.

Statistical testing

If the probability distribution of the parameters is known or an asymptotic approximation is made, confidence limits can be found. Similarly, statistical tests on the residuals can be conducted if the probability distribution of the residuals is known or assumed. We can derive the probability distribution of any linear combination of the dependent variables if the probability distribution of experimental errors is known or assumed. Inferring is easy when assuming that the errors follow a normal distribution, consequently implying that the parameter estimates and residuals will also be normally distributed conditional on the values of the independent variables.[12]

It is necessary to make assumptions about the nature of the experimental errors to test the results statistically. A common assumption is that the errors belong to a normal distribution. The central limit theorem supports the idea that this is a good approximation in many cases.

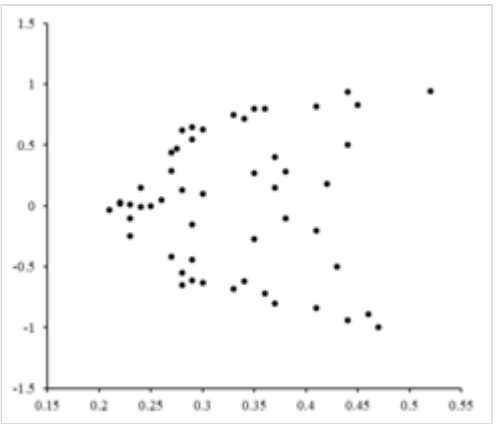
- The Gauss–Markov theorem. In a linear model in which the errors have expectation zero conditional on the independent variables, are uncorrelated and have equal variances, the best linear unbiased estimator of any linear combination of the observations, is its least-squares estimator. "Best" means that the least squares estimators of the parameters have minimum variance. The assumption of equal variance is valid when the errors all belong to the same distribution.[15]
- If the errors belong to a normal distribution, the least-squares estimators are also the maximum likelihood estimators in a linear model.

However, suppose the errors are not normally distributed. In that case, a central limit theorem often nonetheless implies that the parameter estimates will be approximately normally distributed so long as the sample is reasonably large. For this reason, given the important property that the error mean is independent of the independent variables, the distribution of the error term is not an important issue in regression analysis. Specifically, it is not typically important whether the error term follows a normal distribution.

Weighted least squares

A special case of generalized least squares called **weighted least squares** occurs when all the off-diagonal entries of Ω (the correlation matrix of the residuals) are null; the variances of the observations (along the covariance matrix diagonal) may still be unequal (heteroscedasticity). In simpler terms, heteroscedasticity is

when the variance of Y_i depends on the value of x_i which causes the residual plot to create a "fanning out" effect towards larger Y_i values as seen in the residual plot to the right. On the other hand, homoscedasticity is assuming that the variance of Y_i and variance of U_i are equal.[10]



"Fanning Out" Effect of Heteroscedasticity

Relationship to principal components

The first principal component about the mean of a set of points can be represented by that line which most closely approaches the data points (as measured by squared distance of closest approach, i.e. perpendicular to the line). In contrast, linear least squares tries to minimize the distance in the y direction only. Thus, although the two use a similar error metric, linear least squares is a method that treats one dimension of the data preferentially, while PCA treats all dimensions equally.

Relationship to measure theory

Notable statistician Sara van de Geer used empirical process theory and the Vapnik–Chervonenkis dimension to prove a least-squares estimator can be interpreted as a measure on the space of square-integrable functions.[16]

Regularization

Tikhonov regularization

In some contexts a regularized version of the least squares solution may be preferable. Tikhonov regularization (or ridge regression) adds a constraint that $\|\beta\|_2^2$, the squared ℓ_2-norm of the parameter vector, is not greater than a given value to the least squares formulation, leading to a constrained minimization problem. This is equivalent to the unconstrained minimization problem where the objective function is the residual sum of squares plus a penalty term $\alpha\|\beta\|_2^2$ and α is a tuning parameter (this is the Lagrangian form of the constrained minimization problem).[17]

In a Bayesian context, this is equivalent to placing a zero-mean normally distributed prior on the parameter vector.

Lasso method

An alternative regularized version of least squares is Lasso (least absolute shrinkage and selection operator), which uses the constraint that $\|\beta\|_1$, the L_1-norm of the parameter vector, is no greater than a given value.[18][19][20] (One can show like above using Lagrange multipliers that this is equivalent to an unconstrained minimization of the least-squares penalty with $\alpha\|\beta\|_1$ added.) In a Bayesian context, this is equivalent to placing a zero-mean Laplace prior distribution on the parameter vector.[21] The optimization problem may be solved using quadratic programming or more general convex optimization methods, as well as by specific algorithms such as the least angle regression algorithm.

One of the prime differences between Lasso and ridge regression is that in ridge regression, as the penalty is increased, all parameters are reduced while still remaining non-zero, while in Lasso, increasing the penalty will cause more and more of the parameters to be driven to zero. This is an advantage of Lasso over ridge regression, as driving parameters to zero deselects the features from the regression. Thus, Lasso automatically selects more relevant features and discards the others, whereas Ridge regression never fully discards any features. Some feature selection techniques are developed based on the LASSO including Bolasso which bootstraps samples,[22] and FeaLect which analyzes the regression coefficients corresponding to different values of α to score all the features.[23]

The L^1-regularized formulation is useful in some contexts due to its tendency to prefer solutions where more parameters are zero, which gives solutions that depend on fewer variables.[18] For this reason, the Lasso and its variants are fundamental to the field of compressed sensing. An extension of this approach is elastic net regularization.

See also

- Least-squares adjustment
- Bayesian MMSE estimator
- Best linear unbiased estimator (BLUE)
- Best linear unbiased prediction (BLUP)
- Gauss–Markov theorem
- L_2 norm
- Least absolute deviations
- Least-squares spectral analysis
- Measurement uncertainty
- Orthogonal projection
- Proximal gradient methods for learning
- Quadratic loss function
- Root mean square
- Squared deviations from the mean

References

1. Charnes, A.; Frome, E. L.; Yu, P. L. (1976). "The Equivalence of Generalized Least Squares and Maximum Likelihood Estimates in the Exponential Family". *Journal of the American Statistical Association*. **71** (353): 169–171. doi:10.1080/01621459.1976.10481508 (https://doi.org/10.1080%2F01621459.1976.10481508).
2. Mansfield Merriman, "A List of Writings Relating to the Method of Least Squares"
3. Bretscher, Otto (1995). *Linear Algebra With Applications* (3rd ed.). Upper Saddle River, NJ: Prentice Hall.
4. Stigler, Stephen M. (1981). "Gauss and the Invention of Least Squares" (https://doi.org/10.1214%2Faos%2F1176345451). *Ann. Stat.* **9** (3): 465–474. doi:10.1214/aos/1176345451 (https://doi.org/10.1214%2Faos%2F1176345451).
5. Plackett, R.L. (1972). "The discovery of the method of least squares" (https://hedibert.org/wp-content/uploads/2016/08/plackett1972-thediscoveryofthemethodofleastsquares.pdf) (PDF). *Biometrika*. **59** (2): 239–251.

6. Stigler, Stephen M. (1986). *The History of Statistics: The Measurement of Uncertainty Before 1900* (https://archive.org/details/historyofstatist00stig). Cambridge, MA: Belknap Press of Harvard University Press. ISBN 978-0-674-40340-6.
7. Legendre, Adrien-Marie (1805), *Nouvelles méthodes pour la détermination des orbites des comètes* (https://books.google.com/books?id=FRcOAAAAQAAJ) [*New Methods for the Determination of the Orbits of Comets*] (in French), Paris: F. Didot, hdl:2027/nyp.33433069112559 (https://hdl.handle.net/2027%2Fnyp.33433069112559)
8. "The Discovery of Statistical Regression" (https://priceonomics.com/the-discovery-of-statistical-regression/). *Priceonomics*. 2015-11-06. Retrieved 2023-04-04.
9. Aldrich, J. (1998). "Doing Least Squares: Perspectives from Gauss and Yule". *International Statistical Review*. **66** (1): 61–81. doi:10.1111/j.1751-5823.1998.tb00406.x (https://doi.org/10.1111%2Fj.1751-5823.1998.tb00406.x). S2CID 121471194 (https://api.semanticscholar.org/CorpusID:121471194).
10. *A modern introduction to probability and statistics: understanding why and how*. Dekking, Michel, 1946-. London: Springer. 2005. ISBN 978-1-85233-896-1. OCLC 262680588 (https://www.worldcat.org/oclc/262680588).
11. For a good introduction to error-in-variables, please see Fuller, W. A. (1987). *Measurement Error Models*. John Wiley & Sons. ISBN 978-0-471-86187-4.
12. Williams, Jeffrey H. (Jeffrey Huw), 1956- (November 2016). *Quantifying measurement: the tyranny of numbers*. Morgan & Claypool Publishers, Institute of Physics (Great Britain). San Rafael [California] (40 Oak Drive, San Rafael, CA, 94903, USA). ISBN 978-1-68174-433-9. OCLC 962422324 (https://www.worldcat.org/oclc/962422324).
13. Rencher, Alvin C.; Christensen, William F. (2012-08-15). *Methods of Multivariate Analysis* (https://books.google.com/books?id=0g-PAuKub3QC&pg=PA19). John Wiley & Sons. p. 155. ISBN 978-1-118-39167-9.
14. Gere, James M. (2013). *Mechanics of materials*. Goodno, Barry J. (8th ed.). Stamford, Conn.: Cengage Learning. ISBN 978-1-111-57773-5. OCLC 741541348 (https://www.worldcat.org/oclc/741541348).
15. Hallin, Marc. "Gauss-Markov Theorem" (https://onlinelibrary.wiley.com/doi/10.1002/9780470057339.vnn102). *Wiley Online Library*. Encyclopedia of Environmetrics. Retrieved 18 October 2023.
16. van de Geer, Sara (June 1987). "A New Approach to Least-Squares Estimation, with Applications" (https://doi.org/10.1214%2Faos%2F1176350362). *Annals of Statistics*. **15** (2): 587–602. doi:10.1214/aos/1176350362 (https://doi.org/10.1214%2Faos%2F1176350362). S2CID 123088844 (https://api.semanticscholar.org/CorpusID:123088844).
17. van Wieringen, Wessel N. (2021). "Lecture notes on ridge regression". arXiv:1509.09169 (https://arxiv.org/abs/1509.09169). {{cite journal}}: Cite journal requires |journal= (help)
18. Tibshirani, R. (1996). "Regression shrinkage and selection via the lasso". *Journal of the Royal Statistical Society, Series B*. **58** (1): 267–288. JSTOR 2346178 (https://www.jstor.org/stable/2346178).
19. Hastie, Trevor; Tibshirani, Robert; Friedman, Jerome H. (2009). *The Elements of Statistical Learning* (https://web.archive.org/web/20091110212529/http://www-stat.stanford.edu/~tibs/ElemStatLearn/) (second ed.). Springer-Verlag. ISBN 978-0-387-84858-7. Archived from the original (http://www-stat.stanford.edu/~tibs/ElemStatLearn/) on 2009-11-10.
20. Bühlmann, Peter; van de Geer, Sara (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer. ISBN 9783642201929.

21. Park, Trevor; Casella, George (2008). "The Bayesian Lasso". *Journal of the American Statistical Association*. **103** (482): 681–686. doi:10.1198/016214508000000337 (https://doi.org/10.1198%2F016214508000000337). S2CID 11797924 (https://api.semanticscholar.org/CorpusID:11797924).
22. Bach, Francis R (2008). "Bolasso" (http://dl.acm.org/citation.cfm?id=1390161). *Proceedings of the 25th international conference on Machine learning - ICML '08*. pp. 33–40. arXiv:0804.1302 (https://arxiv.org/abs/0804.1302). Bibcode:2008arXiv0804.1302B (https://ui.adsabs.harvard.edu/abs/2008arXiv0804.1302B). doi:10.1145/1390156.1390161 (https://doi.org/10.1145%2F1390156.1390161). ISBN 9781605582054. S2CID 609778 (https://api.semanticscholar.org/CorpusID:609778).
23. Zare, Habil (2013). "Scoring relevancy of features based on combinatorial analysis of Lasso with application to lymphoma diagnosis" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3549810). *BMC Genomics*. **14** (Suppl 1): S14. doi:10.1186/1471-2164-14-S1-S14 (https://doi.org/10.1186%2F1471-2164-14-S1-S14). PMC 3549810 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3549810). PMID 23369194 (https://pubmed.ncbi.nlm.nih.gov/23369194).

Further reading

- Björck, Å. (1996). *Numerical Methods for Least Squares Problems*. SIAM. ISBN 978-0-89871-360-2.
- Kariya, T.; Kurata, H. (2004). *Generalized Least Squares*. Hoboken: Wiley. ISBN 978-0-470-86697-9.
- Luenberger, D. G. (1997) [1969]. "Least-Squares Estimation" (https://books.google.com/books?id=lZU0CAH4RccC&pg=PA78). *Optimization by Vector Space Methods*. New York: John Wiley & Sons. pp. 78–102. ISBN 978-0-471-18117-0.
- Rao, C. R.; Toutenburg, H.; et al. (2008). *Linear Models: Least Squares and Alternatives* (https://books.google.com/books?id=3LK9JoGEyN4C). Springer Series in Statistics (3rd ed.). Berlin: Springer. ISBN 978-3-540-74226-5.
- Van de moortel, Koen (April 2021). "Multidirectional regression analysis" (https://www.researchgate.net/publication/350838636).
- Wolberg, J. (2005). *Data Analysis Using the Method of Least Squares: Extracting the Most Information from Experiments*. Berlin: Springer. ISBN 978-3-540-25674-8.

External links

- Media related to Least squares at Wikimedia Commons