Linear least squares

Linear least squares (**LLS**) is the least squares approximation of linear functions to data. It is a set of formulations for solving statistical problems involved in linear regression, including variants for ordinary (unweighted), weighted, and generalized (correlated) residuals. Numerical methods for linear least squares include inverting the matrix of the normal equations and orthogonal decomposition methods.

Contents

Main formulations

The three main linear least squares formulations are:

- **Ordinary least squares** (OLS) is the most common estimator. OLS estimates are commonly used to analyze both experimental and observational data.

  The OLS method minimizes the sum of squared residuals, and leads to a closed-form expression for the estimated value of the unknown parameter vector β:

  $$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{y},$$

  where $\mathbf{y}$ is a vector whose *i*th element is the *i*th observation of the dependent variable, and $\mathbf{X}$ is a matrix whose *ij* element is the *i*th observation of the *j*th independent variable. (Note: $(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}$ is the Moore–Penrose inverse.) The estimator is unbiased and consistent if the errors have finite variance and are uncorrelated with the regressors:[1]

  $$\operatorname{E}[\,\mathbf{x}_i\varepsilon_i\,] = 0,$$

where $\mathbf{x}_i$ is the transpose of row *i* of the matrix $\mathbf{X}$. It is also efficient under the assumption that the errors have finite variance and are homoscedastic, meaning that E[ε_i^2|$\mathbf{x}_i$] does not depend on *i*. The condition that the errors are uncorrelated with the regressors will generally be satisfied in an experiment, but in the case of observational data, it is difficult to exclude the possibility of an omitted covariate *z* that is related to both the observed covariates and the response variable. The existence of such a covariate will generally lead to a correlation between the regressors and the response variable, and hence to an inconsistent estimator of **β**. The condition of homoscedasticity can fail with either experimental or observational data. If the goal is either inference or predictive modeling, the performance of OLS estimates can be poor if multicollinearity is present, unless the sample size is large.

- **Weighted least squares** (WLS) are used when heteroscedasticity is present in the error terms of the model.
- **Generalized least squares** (GLS) is an extension of the OLS method, that allows efficient estimation of *β* when either heteroscedasticity, or correlations, or both are present among the error terms of the model, as long as the form of heteroscedasticity and correlation is known independently of the data. To handle heteroscedasticity when the error terms are uncorrelated with each other, GLS minimizes a weighted analogue to the sum of squared residuals from OLS regression, where the weight for the i^{th} case is inversely proportional to var(ε_i). This special case of GLS is called "weighted least squares". The GLS solution to estimation problem is

  $$\hat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathsf{T}}\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{X}^{\mathsf{T}}\boldsymbol{\Omega}^{-1}\mathbf{y},$$

  where **Ω** is the covariance matrix of the errors. GLS can be viewed as applying a linear transformation to the data so that the assumptions of OLS are met for the transformed data. For GLS to be applied, the covariance structure of the errors must be known up to a multiplicative constant.

Alternative formulations

Other formulations include:

- **Iteratively reweighted least squares** (IRLS) is used when heteroscedasticity, or correlations, or both are present among the error terms of the model, but where little is known about the covariance structure of the errors independently of the data.[2] In the first iteration, OLS, or GLS with a provisional covariance structure is carried out, and the residuals are obtained from the fit. Based on the residuals, an improved estimate of the covariance structure of the errors can usually be obtained. A subsequent GLS iteration is then performed using this estimate of the error structure to define the weights. The process can be iterated to convergence, but in many cases, only one iteration is sufficient to achieve an efficient estimate of *β*.[3][4]
- **Instrumental variables** regression (IV) can be performed when the regressors are correlated with the errors. In this case, we need the existence of some auxiliary *instrumental variables* $\mathbf{z}_i$ such that E[$\mathbf{z}_i\varepsilon_i$] = 0. If **Z** is the matrix of instruments, then the estimator can be given in closed form as

  $$\hat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathsf{T}}\mathbf{Z}(\mathbf{Z}^{\mathsf{T}}\mathbf{Z})^{-1}\mathbf{Z}^{\mathsf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{Z}(\mathbf{Z}^{\mathsf{T}}\mathbf{Z})^{-1}\mathbf{Z}^{\mathsf{T}}\mathbf{y}.$$

  Optimal instruments regression is an extension of classical IV regression to the situation where E[ε_i | $\mathbf{z}_i$] = 0.
- **Total least squares** (TLS)[5] is an approach to least squares estimation of the linear regression model that treats the covariates and response variable in a more geometrically symmetric manner than OLS. It is one approach to handling the "errors in variables" problem, and is also sometimes used even when the covariates are assumed to be error-free.

In addition, **percentage least squares** focuses on reducing percentage errors, which is useful in the field of forecasting or time series analysis. It is also useful in situations where the dependent variable has a wide range without constant variance, as here the larger residuals at the upper end of the range would dominate if OLS were used. When the percentage or relative error is normally distributed, least squares percentage regression provides maximum likelihood estimates. Percentage regression is linked to a multiplicative error model, whereas OLS is linked to models containing an additive error term.[6]

In constrained least squares, one is interested in solving a linear least squares problem with an additional constraint on the solution.

Objective function

In OLS (i.e., assuming unweighted observations), the optimal value of the objective function is found by substituting the optimal expression for the coefficient vector:

$$S=\mathbf{y}^{\mathrm{T}}(\mathbf{I}-\mathbf{H})^{\mathrm{T}}(\mathbf{I}-\mathbf{H})\mathbf{y}=\mathbf{y}^{\mathrm{T}}(\mathbf{I}-\mathbf{H})\mathbf{y},$$

where $\mathbf{H}=\mathbf{X}(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathsf{T}}$, the latter equality holding since $(\mathbf{I}-\mathbf{H})$ is symmetric and idempotent. It can be shown from this[7] that under an appropriate assignment of weights the expected value of S is $m - n$. If instead unit weights are assumed, the expected value of S is $(m-n)\sigma^2$, where σ^2 is the variance of each observation.

If it is assumed that the residuals belong to a normal distribution, the objective function, being a sum of weighted squared residuals, will belong to a chi-squared (χ^2) distribution with $m - n$ degrees of freedom. Some illustrative percentile values of χ^2 are given in the following table.[8]

$m-n$	$\chi^2_{0.50}$	$\chi^2_{0.95}$	$\chi^2_{0.99}$
10	9.34	18.3	23.2
25	24.3	37.7	44.3
100	99.3	124	136

These values can be used for a statistical criterion as to the goodness of fit. When unit weights are used, the numbers should be divided by the variance of an observation.

For WLS, the ordinary objective function above is replaced for a weighted average of residuals.

Discussion

In statistics and mathematics, **linear least squares** is an approach to fitting a mathematical or statistical model to data in cases where the idealized value provided by the model for any data point is expressed linearly in terms of the unknown parameters of the model. The resulting fitted model can be used to summarize the data, to predict unobserved values from the same system, and to understand the mechanisms that may underlie the system.

Mathematically, linear least squares is the problem of approximately solving an overdetermined system of linear equations **A x** = **b**, where **b** is not an element of the column space of the matrix **A**. The approximate solution is realized as an exact solution to **A x** = **b'**, where **b'** is the projection of **b** onto the column space of **A**. The best approximation is then that which minimizes the sum of squared differences between the data values and their corresponding modeled values. The approach is called *linear* least squares since the assumed function is linear in the parameters to be estimated. Linear least squares problems are convex and have a closed-form solution that is unique, provided that the number of data points used for fitting equals or exceeds the number of unknown parameters, except in special degenerate situations. In contrast, non-linear least squares problems generally must be solved by an iterative procedure, and the problems can be non-convex with multiple optima for the objective function. If prior distributions are available, then even an underdetermined system can be solved using the Bayesian MMSE estimator.

In statistics, linear least squares problems correspond to a particularly important type of statistical model called linear regression which arises as a particular form of regression analysis. One basic form of such a model is an ordinary least squares model. The present article concentrates on the mathematical aspects of linear least squares problems, with discussion of the formulation and interpretation of statistical regression models and statistical inferences related to these being dealt with in the articles just mentioned. See outline of regression analysis for an outline of the topic.

Properties

If the experimental errors, $\boldsymbol{\epsilon}$, are uncorrelated, have a mean of zero and a constant variance, $\boldsymbol{\sigma}$, the Gauss–Markov theorem states that the least-squares estimator, $\hat{\boldsymbol{\beta}}$, has the minimum variance of all estimators that are linear combinations of the observations. In this sense it is the best, or optimal, estimator of the parameters. Note particularly that this property is independent of the statistical distribution function of the errors. In other words, *the distribution function of the errors need not be a normal distribution*. However, for some probability distributions, there is no guarantee that the least-squares solution is even possible given the observations; still, in such cases it is the best estimator that is both linear and unbiased.

For example, it is easy to show that the arithmetic mean of a set of measurements of a quantity is the least-squares estimator of the value of that quantity. If the conditions of the Gauss–Markov theorem apply, the arithmetic mean is optimal, whatever the distribution of errors of the measurements might be.

However, in the case that the experimental errors do belong to a normal distribution, the least-squares estimator is also a maximum likelihood estimator.[9]

These properties underpin the use of the method of least squares for all types of data fitting, even when the assumptions are not strictly valid.

Limitations

An assumption underlying the treatment given above is that the independent variable, *x*, is free of error. In practice, the errors on the measurements of the independent variable are usually much smaller than the errors on the dependent variable and can therefore be ignored. When this is not the case, total least squares or more generally errors-in-variables models, or *rigorous least squares*, should be used. This can be done by adjusting the weighting scheme to take into account errors on both the dependent and independent variables and then following the standard procedure.[10][11]

In some cases the (weighted) normal equations matrix $X^{\mathrm{T}}X$ is ill-conditioned. When fitting polynomials the normal equations matrix is a Vandermonde matrix. Vandermonde matrices become increasingly ill-conditioned as the order of the matrix increases. In these cases, the least squares estimate amplifies the measurement noise and may be grossly inaccurate. Various regularization techniques can be applied in such cases, the most common of which is called ridge regression. If further information about the parameters is known, for example, a range of possible values of $\hat{\boldsymbol{\beta}}$, then various techniques can be used to increase the stability of the solution. For example, see constrained least squares.

Another drawback of the least squares estimator is the fact that the norm of the residuals, $\|\mathbf{y} - X\hat{\boldsymbol{\beta}}\|$ is minimized, whereas in some cases one is truly interested in obtaining small error in the parameter $\hat{\boldsymbol{\beta}}$, e.g., a small value of $\|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\|$. However, since the true parameter $\boldsymbol{\beta}$ is necessarily unknown, this quantity cannot be directly minimized. If a prior probability on $\hat{\boldsymbol{\beta}}$ is known, then a Bayes estimator can be used to minimize the mean squared error, $E\left\{\|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\|^2\right\}$. The least squares method is often applied when no prior is known. Surprisingly, when several parameters are being estimated jointly, better estimators can be constructed, an effect known as Stein's phenomenon. For example, if the measurement error is Gaussian, several estimators are known which dominate, or outperform, the least squares technique; the best known of these is the James–Stein estimator. This is an example of more general shrinkage estimators that have been applied to regression problems.

Applications

- Polynomial fitting: models are polynomials in an independent variable, *x*:
  - Straight line: $f(x, \boldsymbol{\beta}) = \beta_1 + \beta_2 x$.[12]
  - Quadratic: $f(x, \boldsymbol{\beta}) = \beta_1 + \beta_2 x + \beta_3 x^2$.
  - Cubic, quartic and higher polynomials. For regression with high-order polynomials, the use of orthogonal polynomials is recommended.[13]
- Numerical smoothing and differentiation — this is an application of polynomial fitting.
- Multinomials in more than one independent variable, including surface fitting
- Curve fitting with B-splines [10]
- Chemometrics, Calibration curve, Standard addition, Gran plot, analysis of mixtures

Uses in data fitting

The primary application of linear least squares is in data fitting. Given a set of *m* data points $y_1, y_2, \dots, y_m$, consisting of experimentally measured values taken at *m* values $x_1, x_2, \dots, x_m$ of an independent variable (x_i may be scalar or vector quantities), and given a model function $y = f(x, \boldsymbol{\beta})$, with $\boldsymbol{\beta} = (\beta_1, \beta_2, \dots, \beta_n)$, it is desired to find the parameters β_j such that the model function "best" fits the data. In linear least squares, linearity is meant to be with respect to parameters β_j, so

$$f(x, \boldsymbol{\beta}) = \sum_{j=1}^{n} \beta_j \varphi_j(x).$$

Here, the functions φ_j may be **nonlinear** with respect to the variable **x**.

Ideally, the model function fits the data exactly, so

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

for all $i = 1, 2, \dots, m$. This is usually not possible in practice, as there are more data points than there are parameters to be determined. The approach chosen then is to find the minimal possible value of the sum of squares of the residuals

$$r_i(\boldsymbol{\beta}) = y_i - f(x_i, \boldsymbol{\beta}),\ (i = 1, 2, \dots, m)$$

so to minimize the function

$$S(\boldsymbol{\beta}) = \sum_{i=1}^{m} r_i^2(\boldsymbol{\beta}).$$

After substituting for r_i and then for f, this minimization problem becomes the quadratic minimization problem above with

$$X_{ij} = \varphi_j(x_i),$$

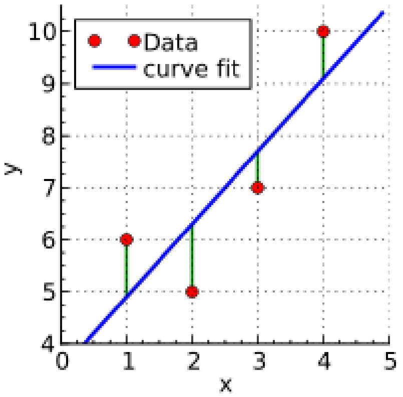
and the best fit can be found by solving the normal equations.

Example

As a result of an experiment, four (x, y) data points were obtained, $(1,6)$, $(2,5)$, $(3,7)$, and $(4,10)$ (shown in red in the diagram on the right). We hope to find a line $y = \beta_1 + \beta_2 x$ that best fits these four points. In other words, we would like to find the numbers β_1 and β_2 that approximately solve the overdetermined linear system:

$$\begin{aligned} \beta_1 + 1\beta_2 + r_1 &= 6 \\ \beta_1 + 2\beta_2 + r_2 &= 5 \\ \beta_1 + 3\beta_2 + r_3 &= 7 \\ \beta_1 + 4\beta_2 + r_4 &= 10 \end{aligned}$$

of four equations in two unknowns in some "best" sense.



A plot of the data points (in red), the least squares line of best fit (in blue), and the residuals (in green).

r represents the residual, at each point, between the curve fit and the data:

$$\begin{aligned} r_1 &= 6 - (\beta_1 + 1\beta_2) \\ r_2 &= 5 - (\beta_1 + 2\beta_2) \\ r_3 &= 7 - (\beta_1 + 3\beta_2) \\ r_4 &= 10 - (\beta_1 + 4\beta_2) \end{aligned}$$

The least squares approach to solving this problem is to try to make the sum of the squares of these residuals as small as possible; that is, to find the minimum of the function:

$$\begin{aligned} S(\beta_1, \beta_2) &= r_1^2 + r_2^2 + r_3^2 + r_4^2 \\ &= [6 - (\beta_1 + 1\beta_2)]^2 + [5 - (\beta_1 + 2\beta_2)]^2 + [7 - (\beta_1 + 3\beta_2)]^2 + [10 - (\beta_1 + 4\beta_2)]^2 \\ &= 4\beta_1^2 + 30\beta_2^2 + 20\beta_1\beta_2 - 56\beta_1 - 154\beta_2 + 210 \end{aligned}$$

The minimum is determined by calculating the partial derivatives of $S(\beta_1, \beta_2)$ with respect to β_1 and β_2 and setting them to zero:

$$\begin{aligned} \frac{\partial S}{\partial \beta_1} &= 0 = 8\beta_1 + 20\beta_2 - 56 \\ \frac{\partial S}{\partial \beta_2} &= 0 = 20\beta_1 + 60\beta_2 - 154. \end{aligned}$$

This results in a system of two equations in two unknowns, called the normal equations, which when solved give:

$$\beta_1 = 3.5$$
$$\beta_2 = 1.4$$

and the equation $y = 3.5 + 1.4x$ is the line of best fit. The residuals, that is, the differences between the y values from the observations and the y predicated variables by using the line of best fit, are then found to be $1.1,$ $-1.3,$ $-0.7,$ and 0.9 (see the diagram on the right). The minimum value of the sum of squares of the residuals is $S(3.5, 1.4) = 1.1^2 + (-1.3)^2 + (-0.7)^2 + 0.9^2 = 4.2.$

More generally, one can have n regressors x_j, and a linear model

$$y = \beta_0 + \sum_{j=1}^{n} \beta_j x_j.$$

Using a quadratic model

Importantly, in "linear least squares", we are not restricted to using a line as the model as in the above example. For instance, we could have chosen the restricted quadratic model $y = \beta_1 x^2$. This model is still linear in the β_1 parameter, so we can still perform the same analysis, constructing a system of equations from the data points:

$$6 = \beta_1 (1)^2$$
$$5 = \beta_1 (2)^2$$
$$7 = \beta_1 (3)^2$$
$$10 = \beta_1 (4)^2$$

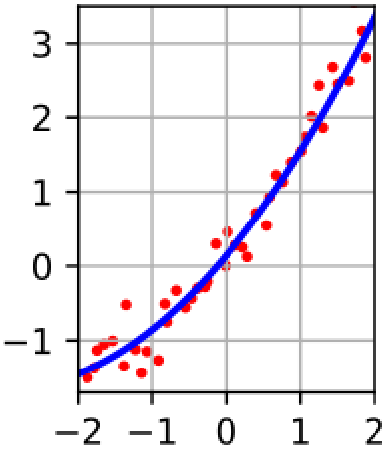
The partial derivatives with respect to the parameters (this time there is only one) are again computed and set to 0:

$$\frac{\partial S}{\partial \beta_1} = 0 = 708\beta_1 - 498$$

and solved

$$\beta_1 = 0.703$$

leading to the resulting best fit model $y = 0.703x^2$.



The result of fitting a quadratic function $y = \beta_1 + \beta_2 x + \beta_3 x^2$ (in blue) through a set of data points (x_i, y_i) (in red). In linear least squares the function need not be linear in the argument x, but only in the parameters β_j that are determined to give the best fit.

See also

- Line-line intersection#Nearest point to non-intersecting lines, an application
- Line fitting
- Nonlinear least squares
- Regularized least squares
- Simple linear regression
- Partial least squares regression

- Linear function

References

1. Lai, T.L.; Robbins, H.; Wei, C.Z. (1978). "Strong consistency of least squares estimates in multiple regression" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC392707). *PNAS*. **75** (7): 3034–3036. Bibcode:1978PNAS...75.3034L (https://ui.adsabs.harvard.edu/abs/1978PNAS...75.3034L). doi:10.1073/pnas.75.7.3034 (https://doi.org/10.1073%2Fpnas.75.7.3034). JSTOR 68164 (https://www.jstor.org/stable/68164). PMC 392707 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC392707). PMID 16592540 (https://pubmed.ncbi.nlm.nih.gov/16592540).
2. del Pino, Guido (1989). "The Unifying Role of Iterative Generalized Least Squares in Statistical Algorithms" (https://doi.org/10.1214%2Fss%2F1177012408). *Statistical Science*. **4** (4): 394–403. doi:10.1214/ss/1177012408 (https://doi.org/10.1214%2Fss%2F1177012408). JSTOR 2245853 (https://www.jstor.org/stable/2245853).
3. Carroll, Raymond J. (1982). "Adapting for Heteroscedasticity in Linear Models" (https://doi.org/10.1214%2Faos%2F1176345987). *The Annals of Statistics*. **10** (4): 1224–1233. doi:10.1214/aos/1176345987 (https://doi.org/10.1214%2Faos%2F1176345987). JSTOR 2240725 (https://www.jstor.org/stable/2240725).
4. Cohen, Michael; Dalal, Siddhartha R.; Tukey, John W. (1993). "Robust, Smoothly Heterogeneous Variance Regression". *Journal of the Royal Statistical Society, Series C*. **42** (2): 339–353. JSTOR 2986237 (https://www.jstor.org/stable/2986237).
5. Nievergelt, Yves (1994). "Total Least Squares: State-of-the-Art Regression in Numerical Analysis". *SIAM Review*. **36** (2): 258–264. doi:10.1137/1036055 (https://doi.org/10.1137%2F1036055). JSTOR 2132463 (https://www.jstor.org/stable/2132463).
6. Tofallis, C (2009). "Least Squares Percentage Regression" (https://digitalcommons.wayne.edu/cgi/viewcontent.cgi?article=1466&context=jmasm). *Journal of Modern Applied Statistical Methods*. **7**: 526–534. doi:10.2139/ssrn.1406472 (https://doi.org/10.2139%2Fssrn.1406472). SSRN 1406472 (https://ssrn.com/abstract=1406472).
7. Hamilton, W. C. (1964). *Statistics in Physical Science* (https://archive.org/details/statisticsinphys0000hami). New York: Ronald Press.
8. Spiegel, Murray R. (1975). *Schaum's outline of theory and problems of probability and statistics*. New York: McGraw-Hill. ISBN 978-0-585-26739-5.
9. Margenau, Henry; Murphy, George Moseley (1956). *The Mathematics of Physics and Chemistry* (https://archive.org/details/mathematicsofphy0002marg). Princeton: Van Nostrand.
10. Gans, Peter (1992). *Data fitting in the Chemical Sciences*. New York: Wiley. ISBN 978-0-471-93412-7.
11. Deming, W. E. (1943). *Statistical adjustment of Data*. New York: Wiley.
12. Acton, F. S. (1959). *Analysis of Straight-Line Data*. New York: Wiley.
13. Guest, P. G. (1961). *Numerical Methods of Curve Fitting*. Cambridge: Cambridge University Press.

Further reading

- Bevington, Philip R.; Robinson, Keith D. (2003). *Data Reduction and Error Analysis for the Physical Sciences*. McGraw-Hill. ISBN 978-0-07-247227-1.

External links

- Least Squares Fitting – From MathWorld (http://mathworld.wolfram.com/LeastSquaresFitting.html)

- Least Squares Fitting-Polynomial – From MathWorld (http://mathworld.wolfram.com/LeastSquaresFittingPolynomial.html)

Retrieved from "https://en.wikipedia.org/w/index.php?title=Linear_least_squares&oldid=998292342"