|  |  |  |  |
| --- | --- | --- | --- |
| |  |  | | --- | --- | | **4.1.4.1.** | **Linear Least Squares Regression** | | |
| *Modeling Workhorse* | Linear least squares regression is by far the most widely used modeling method. It is what most people mean when they say they have used "regression", "linear regression" or "least squares" to fit a model to their data. Not only is linear least squares regression the most widely used modeling method, but it has been adapted to a broad range of situations that are outside its direct scope. It plays a strong underlying role in many other modeling methods, including the other methods discussed in this section: [nonlinear least squares regression](https://www.itl.nist.gov/div898/handbook/pmd/section1/pmd142.htm), [weighted least squares regression](https://www.itl.nist.gov/div898/handbook/pmd/section1/pmd143.htm) and[LOESS](https://www.itl.nist.gov/div898/handbook/pmd/section1/pmd144.htm). |
| *Definition of a Linear Least Squares Model* | Used directly, with an [appropriate data set](https://www.itl.nist.gov/div898/handbook/pmd/section3/pmd32.htm), linear least squares regression can be used to fit the data with any function of the form  *f*(*x*⃗ ;*β*⃗ )=*β*0+*β*1*x*1+*β*2*x*2+…  in which   1. each explanatory variable in the function is multiplied by an unknown parameter, 2. there is at most one unknown parameter with no corresponding explanatory variable, and 3. all of the individual terms are summed to produce the final function value.   In statistical terms, any function that meets these criteria would be called a "linear function". The term "linear" is used, even though the function may not be a straight line, because if the unknown parameters are considered to be variables and the explanatory variables are considered to be known coefficients corresponding to those "variables", then the problem becomes a system (usually overdetermined) of linear equations that can be solved for the values of the unknown parameters. To differentiate the various meanings of the word "linear", the linear models being discussed here are often said to be "linear in the parameters" or "statistically linear". |
| *Why "Least Squares"?* | Linear least squares regression also gets its name from the way the estimates of the unknown parameters are computed. The "method of least squares" that is used to obtain parameter estimates was independently developed in the late 1700's and the early 1800's by the mathematicians Karl Friedrich Gauss, Adrien Marie Legendre and (possibly) Robert Adrain [[Stigler (1978)]](https://www.itl.nist.gov/div898/handbook/pmd/section7/pmd7.htm#Stigler,%20S.M.%20(1978)) [[Harter (1983)]](https://www.itl.nist.gov/div898/handbook/pmd/section7/pmd7.htm#Harter,%20H.L.%20(1983))[[Stigler (1986)]](https://www.itl.nist.gov/div898/handbook/pmd/section7/pmd7.htm#Stigler,%20S.M.%20(1986)) working in Germany, France and America, respectively. In the least squares method the unknown parameters are estimated by minimizing the sum of the squared deviations between the data and the model. The minimization process reduces the overdetermined system of equations formed by the data to a sensible system of *p*,(where *p* is the number of parameters in the functional part of the model) equations in *p* unknowns. This new system of equations is then solved to obtain the parameter estimates. To learn more about how the method of least squares is used to estimate the parameters, see [Section 4.4.3.1](https://www.itl.nist.gov/div898/handbook/pmd/section4/pmd431.htm). |
| *Examples of Linear Functions* | As just mentioned above, linear models are not limited to being straight lines or planes, but include a fairly wide range of shapes. For example, a simple quadratic curve,  *f*(*x*;*β*⃗ )=*β*0+*β*1*x*+*β*11*x*2,  is linear in the statistical sense. A straight-line model inlog(*x*),  *f*(*x*;*β*⃗ )=*β*0+*β*1ln(*x*),  or a polynomial in sin(*x*),  *f*(*x*;*β*⃗ )=*β*0+*β*1sin(*x*)+*β*2sin(2*x*)+*β*3sin(3*x*),  is also linear in the statistical sense because they are linear in the parameters, though not with respect to the observed explanatory variable, *x*. |
| *Nonlinear Model Example* | Just as models that are linear in the statistical sense do not have to be linear with respect to the explanatory variables, nonlinear models can be linear with respect to the explanatory variables, but not with respect to the parameters. For example,  *f*(*x*;*β*⃗ )=*β*0+*β*0*β*1*x*  is linear in *x*, but it cannot be written in the general form of a linear model presented [above](https://www.itl.nist.gov/div898/handbook/pmd/section1/pmd141.htm#def). This is because the slope of this line is expressed as the product of two parameters. As a result, nonlinear least squares regression could be used to fit this model, but linear least squares cannot be used. For further examples and discussion of nonlinear models see the next section, [Section 4.1.4.2](https://www.itl.nist.gov/div898/handbook/pmd/section1/pmd142.htm). |
| *Advantages of Linear Least Squares* | Linear least squares regression has earned its place as the primary tool for process modeling because of its effectiveness and completeness. |
|  | Though there are types of data that are better described by functions that are nonlinear in the parameters, many processes in science and engineering are well-described by linear models. This is because either the processes are inherently linear or because, over short ranges, any process can be well-approximated by a linear model. |
|  | The estimates of the unknown parameters obtained from linear least squares regression are the optimal estimates from a broad class of possible parameter estimates under the usual assumptions used for process modeling. Practically speaking, linear least squares regression makes very efficient use of the data. Good results can be obtained with relatively small data sets. |
|  | Finally, the theory associated with linear regression is well-understood and allows for construction of different types of easily-interpretable statistical intervals for predictions, calibrations, and optimizations. These statistical intervals can then be used to give clear answers to scientific and engineering questions. |
| *Disadvantages of Linear Least Squares* | The main disadvantages of linear least squares are limitations in the shapes that linear models can assume over long ranges, possibly poor extrapolation properties, and sensitivity to outliers. |
|  | Linear models with nonlinear terms in the predictor variables curve relatively slowly, so for inherently nonlinear processes it becomes increasingly difficult to find a linear model that fits the data well as the range of the data increases. As the explanatory variables become extreme, the output of the linear model will also always more extreme. This means that linear models may not be effective for extrapolating the results of a process for which data cannot be collected in the region of interest. Of course extrapolation is potentially dangerous regardless of the model type. |
|  | Finally, while the method of least squares often gives optimal estimates of the unknown parameters, it is very sensitive to the presence of unusual data points in the data used to fit a model. One or two outliers can sometimes seriously skew the results of a least squares analysis. This makes [model validation](https://www.itl.nist.gov/div898/handbook/pmd/section4/pmd44.htm), [especially with respect to outliers](https://www.itl.nist.gov/div898/handbook/pmd/section4/pmd445.htm#nppi), critical to obtaining sound answers to the questions motivating the construction of the model. |

The method of **least squares** is a standard approach in [regression analysis](https://en.wikipedia.org/wiki/Regression_analysis) to approximate the solution of [overdetermined systems](https://en.wikipedia.org/wiki/Overdetermined_system), i.e., sets of equations in which there are more equations than unknowns. "Least squares" means that the overall solution minimizes the sum of the squares of the residuals made in the results of every single equation.

The most important application is in [data fitting](https://en.wikipedia.org/wiki/Curve_fitting). The best fit in the least-squares sense minimizes *the sum of squared* [*residuals*](https://en.wikipedia.org/wiki/Errors_and_residuals_in_statistics) (a residual being: the difference between an observed value, and the fitted value provided by a model). When the problem has substantial uncertainties in the [independent variable](https://en.wikipedia.org/wiki/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](https://en.wikipedia.org/wiki/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](https://en.wikipedia.org/wiki/Ordinary_least_squares) and [nonlinear least squares](https://en.wikipedia.org/wiki/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](https://en.wikipedia.org/wiki/Regression_analysis); it has a [closed-form solution](https://en.wikipedia.org/wiki/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.

[Polynomial least squares](https://en.wikipedia.org/wiki/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](https://en.wikipedia.org/wiki/Exponential_family) and mild conditions are satisfied, least-squares estimates and [maximum-likelihood](https://en.wikipedia.org/wiki/Maximum_likelihood) estimates are identical.[[1]](https://en.wikipedia.org/wiki/Least_squares#cite_note-1) The method of least squares can also be derived as a [method of moments](https://en.wikipedia.org/wiki/Method_of_moments_(statistics)) estimator.

The following discussion is mostly presented in terms of [linear](https://en.wikipedia.org/wiki/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](https://en.wikipedia.org/wiki/Fisher_information)), the least-squares method may be used to fit a [generalized linear model](https://en.wikipedia.org/wiki/Generalized_linear_model).

The least-squares method is usually credited to [Carl Friedrich Gauss](https://en.wikipedia.org/wiki/Carl_Friedrich_Gauss) (1795),[[2]](https://en.wikipedia.org/wiki/Least_squares#cite_note-brertscher-2) but it was first published by [Adrien-Marie Legendre](https://en.wikipedia.org/wiki/Adrien-Marie_Legendre) (1805).

### **Context**[[edit](https://en.wikipedia.org/w/index.php?title=Least_squares&action=edit&section=2)]

The method of least squares grew out of the fields of [astronomy](https://en.wikipedia.org/wiki/Astronomy) and [geodesy](https://en.wikipedia.org/wiki/Geodesy), as scientists and mathematicians sought to provide solutions to the challenges of navigating the Earth's oceans during the [Age of Exploration](https://en.wikipedia.org/wiki/Age_of_Exploration). 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:[[4]](https://en.wikipedia.org/wiki/Least_squares#cite_note-stigler-4)

* 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](https://en.wikipedia.org/wiki/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](https://en.wikipedia.org/wiki/Tobias_Mayer) while studying the [librations](https://en.wikipedia.org/wiki/Libration) of the moon in 1750, and by [Pierre-Simon Laplace](https://en.wikipedia.org/wiki/Pierre-Simon_Laplace) in his work in explaining the differences in motion of [Jupiter](https://en.wikipedia.org/wiki/Jupiter) and [Saturn](https://en.wikipedia.org/wiki/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](https://en.wikipedia.org/wiki/Roger_Joseph_Boscovich) in his work on the shape of the earth in 1757 and by [Pierre-Simon Laplace](https://en.wikipedia.org/wiki/Pierre-Simon_Laplace) for the same problem in 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](https://en.wikipedia.org/wiki/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](https://en.wikipedia.org/wiki/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**[[edit](https://en.wikipedia.org/w/index.php?title=Least_squares&action=edit&section=3)]



[Carl Friedrich Gauss](https://en.wikipedia.org/wiki/Carl_Friedrich_Gauss)

The first clear and concise exposition of the method of least squares was published by [Legendre](https://en.wikipedia.org/wiki/Adrien-Marie_Legendre) in 1805.[[5]](https://en.wikipedia.org/wiki/Least_squares#cite_note-5) 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. The value of Legendre's method of least squares was immediately recognized by leading astronomers and geodesists of the time.

In 1809 [Carl Friedrich Gauss](https://en.wikipedia.org/wiki/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. 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](https://en.wikipedia.org/wiki/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](https://en.wikipedia.org/wiki/Arithmetic_mean) is indeed the best estimate of the location parameter by changing both the [probability density](https://en.wikipedia.org/wiki/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' method](https://en.wikipedia.org/wiki/Gauss%27_method) came when it was used to predict the future location of the newly discovered asteroid [Ceres](https://en.wikipedia.org/wiki/Ceres_(dwarf_planet)). On 1 January 1801, the Italian astronomer [Giuseppe Piazzi](https://en.wikipedia.org/wiki/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](https://en.wikipedia.org/wiki/Kepler%27s_laws_of_planetary_motion) of planetary motion. The only predictions that successfully allowed Hungarian astronomer [Franz Xaver von Zach](https://en.wikipedia.org/wiki/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](https://en.wikipedia.org/wiki/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](https://en.wikipedia.org/wiki/Gauss%E2%80%93Markov_theorem).

The idea of least-squares analysis was also independently formulated by the American [Robert Adrain](https://en.wikipedia.org/wiki/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.[[6]](https://en.wikipedia.org/wiki/Least_squares#cite_note-6)

## Limitations[[edit](https://en.wikipedia.org/w/index.php?title=Least_squares&action=edit&section=5)]

This regression formulation considers only residuals in the dependent variable (but the alternative [total least squares](https://en.wikipedia.org/wiki/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](https://en.wikipedia.org/wiki/Regression_analysis) that leads to fitting by least squares there is an implicit assumption that errors in the [independent variable](https://en.wikipedia.org/wiki/Independent_variable) are zero or strictly controlled so as to be negligible. When errors in the [independent variable](https://en.wikipedia.org/wiki/Independent_variable) are non-negligible, [models of measurement error](https://en.wikipedia.org/wiki/Errors-in-variables_models) can be used; such methods can lead to [parameter estimates](https://en.wikipedia.org/wiki/Parameter_estimation), [hypothesis testing](https://en.wikipedia.org/wiki/Hypothesis_testing) and [confidence intervals](https://en.wikipedia.org/wiki/Confidence_interval) that take into account the presence of observation errors in the independent variables.[[7]](https://en.wikipedia.org/wiki/Least_squares#cite_note-7) An alternative approach is to fit a model by [total least squares](https://en.wikipedia.org/wiki/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.

### **Differences between linear and nonlinear least squares**[[edit](https://en.wikipedia.org/w/index.php?title=Least_squares&action=edit&section=9)]

* The model function, *f*, in LLSQ (linear least squares) is a linear combination of parameters of the form {\displaystyle f=X\_{i1}\beta \_{1}+X\_{i2}\beta \_{2}+\cdots }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 {\displaystyle \beta ^{2},e^{\beta x}}\beta^2, e^{\beta x} and so forth. If the derivatives {\displaystyle \partial f/\partial \beta \_{j}}\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.
* Algorithms for finding the solution to a NLLSQ problem require initial values for the parameters, LLSQ does not.
* Like LLSQ, solution algorithms for NLLSQ often require that the Jacobian can be calculated. 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.
* In NLLSQ non-convergence (failure of the algorithm to find a minimum) is a common phenomenon whereas the LLSQ is globally concave so non-convergence is not an issue.
* NLLSQ is usually an iterative process. The iterative process 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](https://en.wikipedia.org/wiki/Gauss%E2%80%93Seidel) 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.

LINEAR REGRESSION

In [statistics](https://en.wikipedia.org/wiki/Statistics), **linear regression** is a [linear](https://en.wikipedia.org/wiki/Linearity) approach to modelling the relationship between a scalar response (or [dependent variable](https://en.wikipedia.org/wiki/Dependent_variable)) and one or more [explanatory variables](https://en.wikipedia.org/wiki/Explanatory_variable) (or [independent variables](https://en.wikipedia.org/wiki/Independent_variable)). The case of one explanatory variable is called [simple linear regression](https://en.wikipedia.org/wiki/Simple_linear_regression). For more than one explanatory variable, the process is called **multiple linear regression**.[[1]](https://en.wikipedia.org/wiki/Linear_regression#cite_note-Freedman09-1) This term is distinct from [multivariate linear regression](https://en.wikipedia.org/wiki/Multivariate_linear_regression), where multiple correlated dependent variables are predicted, rather than a single scalar variable.[[2]](https://en.wikipedia.org/wiki/Linear_regression#cite_note-2)

In linear regression, the relationships are modeled using [linear predictor functions](https://en.wikipedia.org/wiki/Linear_predictor_function) whose unknown model [parameters](https://en.wikipedia.org/wiki/Parameters) are [estimated](https://en.wikipedia.org/wiki/Estimation_theory) from the [data](https://en.wikipedia.org/wiki/Data). Such models are called [linear models](https://en.wikipedia.org/wiki/Linear_model).[[3]](https://en.wikipedia.org/wiki/Linear_regression#cite_note-3) Most commonly, the [conditional mean](https://en.wikipedia.org/wiki/Conditional_expectation) of the response given the values of the explanatory variables (or predictors) is assumed to be an [affine function](https://en.wikipedia.org/wiki/Affine_transformation) of those values; less commonly, the conditional [median](https://en.wikipedia.org/wiki/Median) or some other [quantile](https://en.wikipedia.org/wiki/Quantile) is used. Like all forms of [regression analysis](https://en.wikipedia.org/wiki/Regression_analysis), linear regression focuses on the [conditional probability distribution](https://en.wikipedia.org/wiki/Conditional_probability_distribution) of the response given the values of the predictors, rather than on the [joint probability distribution](https://en.wikipedia.org/wiki/Joint_probability_distribution) of all of these variables, which is the domain of [multivariate analysis](https://en.wikipedia.org/wiki/Multivariate_analysis).

Linear regression was the first type of regression analysis to be studied rigorously, and to be used extensively in practical applications.[[4]](https://en.wikipedia.org/wiki/Linear_regression#cite_note-4) This is because models which depend linearly on their unknown parameters are easier to fit than models which are non-linearly related to their parameters and because the statistical properties of the resulting estimators are easier to determine.

Linear regression has many practical uses. Most applications fall into one of the following two broad categories:

* If the goal is prediction, or forecasting, or error reduction,[[*clarification needed*](https://en.wikipedia.org/wiki/Wikipedia:Please_clarify)] linear regression can be used to fit a predictive model to an observed [data set](https://en.wikipedia.org/wiki/Data_set)of values of the response and explanatory variables. After developing such a model, if additional values of the explanatory variables are collected without an accompanying response value, the fitted model can be used to make a prediction of the response.
* If the goal is to explain variation in the response variable that can be attributed to variation in the explanatory variables, linear regression analysis can be applied to quantify the strength of the relationship between the response and the explanatory variables, and in particular to determine whether some explanatory variables may have no linear relationship with the response at all, or to identify which subsets of explanatory variables may contain redundant information about the response.

Linear regression models are often fitted using the [least squares](https://en.wikipedia.org/wiki/Least_squares) approach, but they may also be fitted in other ways, such as by minimizing the "lack of fit" in some other [norm](https://en.wikipedia.org/wiki/Norm_(mathematics)) (as with [least absolute deviations](https://en.wikipedia.org/wiki/Least_absolute_deviations) regression), or by minimizing a penalized version of the least squares [cost function](https://en.wikipedia.org/wiki/Loss_function) as in [ridge regression](https://en.wikipedia.org/wiki/Ridge_regression)(*L*2-norm penalty) and [lasso](https://en.wikipedia.org/wiki/Lasso_(statistics)) (*L*1-norm penalty). Conversely, the least squares approach can be used to fit models that are not linear models. Thus, although the terms "least squares" and "linear model" are closely linked, they are not synonymous.

# What is Linear Regression?

Linear regression is a basic and commonly used type of predictive analysis. The overall idea of regression is to examine two things: (1) does a set of predictor variables do a good job in predicting an outcome (dependent) variable? (2) Which variables in particular are significant predictors of the outcome variable, and in what way do they–indicated by the magnitude and sign of the beta estimates–impact the outcome variable? These regression estimates are used to explain the relationship between one dependent variable and one or more independent variables. The simplest form of the regression equation with one dependent and one independent variable is defined by the formula y = c + b\*x, where y = estimated dependent variable score, c = constant, b = regression coefficient, and x = score on the independent variable.

Naming the Variables. There are many names for a regression’s dependent variable. It may be called an outcome variable, criterion variable, endogenous variable, or regressand. The independent variables can be called exogenous variables, predictor variables, or regressors.

Three major uses for regression analysis are (1) determining the strength of predictors, (2) forecasting an effect, and (3) trend forecasting.

First, the regression might be used to identify the strength of the effect that the independent variable(s) have on a dependent variable. Typical questions are what is the strength of relationship between dose and effect, sales and marketing spending, or age and income.

Second, it can be used to forecast effects or impact of changes. That is, the regression analysis helps us to understand how much the dependent variable changes with a change in one or more independent variables. A typical question is, “how much additional sales income do I get for each additional $1000 spent on marketing?”

Third, regression analysis predicts trends and future values. The regression analysis can be used to get point estimates. A typical question is, “what will the price of gold be in 6 months?”

There are several types of linear regression analyses available to researchers.

* Simple linear regression
* 1 dependent variable (interval or ratio), 1 independent variable (interval or ratio or dichotomous)

* [Multiple linear regression](http://www.statisticssolutions.com/data-analysis-plan-multiple-linear-regression/)
* 1 dependent variable (interval or ratio) , 2+ independent variables (interval or ratio or dichotomous)

* [Logistic regression](http://www.statisticssolutions.com/data-analysis-plan-logistic-regression/)
* 1 dependent variable (dichotomous), 2+ independent variable(s) (interval or ratio or dichotomous)

* [Ordinal regression](http://www.statisticssolutions.com/data-analysis-plan-ordinal-regression/)
* 1 dependent variable (ordinal), 1+ independent variable(s) (nominal or dichotomous)

* [Multinominal regression](http://www.statisticssolutions.com/data-analysis-plan-multinominal-logistic-regression/)
* 1 dependent variable (nominal), 1+ independent variable(s) (interval or ratio or dichotomous)

* [Discriminant analysis](http://www.statisticssolutions.com/discriminant-analysis-independent-variables/)
* 1 dependent variable (nominal), 1+ independent variable(s) (interval or ratio)

When selecting the model for the analysis, an important consideration is model fitting. Adding independent variables to a linear regression model will always increase the explained variance of the model (typically expressed as R²). However, overfitting can occur by adding too many variables to the model, which reduces model generalizability. Occam’s razor describes the problem extremely well – a simple model is usually preferable to a more complex model. Statistically, if a model includes a large number of variables, some of the variables will be statistically significant due to chance alone.

# **Linear Regression for Machine Learning**

by [**Jason Brownlee**](https://machinelearningmastery.com/author/jasonb/) on March 25, 2016 in **[Machine Learning Algorithms](https://machinelearningmastery.com/category/machine-learning-algorithms/)**

Linear regression is perhaps one of the most well known and well understood algorithms in statistics and machine learning.

In this post you will discover the linear regression algorithm, how it works and how you can best use it in on your machine learning projects. In this post you will learn:

* Why linear regression belongs to both statistics and machine learning.
* The many names by which linear regression is known.
* The representation and learning algorithms used to create a linear regression model.
* How to best prepare your data when modeling using linear regression.

You do not need to know any statistics or linear algebra to understand linear regression. This is a gentle high-level introduction to the technique to give you enough background to be able to use it effectively on your own problems.

## **Isn’t Linear Regression from Statistics?**

Before we dive into the details of linear regression, you may be asking yourself why we are looking at this algorithm.

Isn’t it a technique from statistics?

Machine learning, more specifically the field of predictive modeling is primarily concerned with minimizing the error of a model or making the most accurate predictions possible, at the expense of explainability. In applied machine learning we will borrow, reuse and steal algorithms from many different fields, including statistics and use them towards these ends.

As such, linear regression was developed in the field of statistics and is studied as a model for understanding the relationship between input and output numerical variables, but has been borrowed by machine learning. It is both a statistical algorithm and a machine learning algorithm.

Next, let’s review some of the common names used to refer to a linear regression model.

## **Many Names of Linear Regression**

When you start looking into linear regression, things can get very confusing.

The reason is because linear regression has been around for so long (more than 200 years). It has been studied from every possible angle and often each angle has a new and different name.

Linear regression is a **linear model**, e.g. a model that assumes a linear relationship between the input variables (x) and the single output variable (y). More specifically, that y can be calculated from a linear combination of the input variables (x).

When there is a single input variable (x), the method is referred to as **simple linear regression**. When there are **multiple input variables**, literature from statistics often refers to the method as multiple linear regression.

Different techniques can be used to prepare or train the linear regression equation from data, the most common of which is called **Ordinary Least Squares**. It is common to therefore refer to a model prepared this way as Ordinary Least Squares Linear Regression or just Least Squares Regression.

Now that we know some names used to describe linear regression, let’s take a closer look at the representation used.

## **Linear Regression Model Representation**

[Linear regression](https://en.wikipedia.org/wiki/Linear_regression) is an attractive model because the representation is so simple.

The representation is a linear equation that combines a specific set of input values (x) the solution to which is the predicted output for that set of input values (y). As such, both the input values (x) and the output value are numeric.

The linear equation assigns one scale factor to each input value or column, called a coefficient and represented by the capital Greek letter Beta (B). One additional coefficient is also added, giving the line an additional degree of freedom (e.g. moving up and down on a two-dimensional plot) and is often called the intercept or the bias coefficient.

For example, in a simple regression problem (a single x and a single y), the form of the model would be:

y = B0 + B1\*x

In higher dimensions when we have more than one input (x), the line is called a plane or a hyper-plane. The representation therefore is the form of the equation and the specific values used for the coefficients (e.g. B0 and B1 in the above example).

It is common to talk about the complexity of a regression model like linear regression. This refers to the number of coefficients used in the model.

When a coefficient becomes zero, it effectively removes the influence of the input variable on the model and therefore from the prediction made from the model (0 \* x = 0). This becomes relevant if you look at regularization methods that change the learning algorithm to reduce the complexity of regression models by putting pressure on the absolute size of the coefficients, driving some to zero.

Now that we understand the representation used for a linear regression model, let’s review some ways that we can learn this representation from data.



What is Linear Regression?

Photo by [Estitxu Carton](https://www.flickr.com/photos/bichuas/3961559679/), some rights reserved.

## **Linear Regression Learning the Model**

Learning a linear regression model means estimating the values of the coefficients used in the representation with the data that we have available.

In this section we will take a brief look at four techniques to prepare a linear regression model. This is not enough information to implement them from scratch, but enough to get a flavor of the computation and trade-offs involved.

There are many more techniques because the model is so well studied. Take note of Ordinary Least Squares because it is the most common method used in general. Also take note of Gradient Descent as it is the most common technique taught in machine learning classes.

### **1. Simple Linear Regression**

With simple linear regression when we have a single input, we can use statistics to estimate the coefficients.

This requires that you calculate statistical properties from the data such as means, standard deviations, correlations and covariance. All of the data must be available to traverse and calculate statistics.

This is fun as an exercise in excel, but not really useful in practice.

### **2. Ordinary Least Squares**

When we have more than one input we can use Ordinary Least Squares to estimate the values of the coefficients.

The [Ordinary Least Squares](https://en.wikipedia.org/wiki/Ordinary_least_squares) procedure seeks to minimize the sum of the squared residuals. This means that given a regression line through the data we calculate the distance from each data point to the regression line, square it, and sum all of the squared errors together. This is the quantity that ordinary least squares seeks to minimize.

This approach treats the data as a matrix and uses linear algebra operations to estimate the optimal values for the coefficients. It means that all of the data must be available and you must have enough memory to fit the data and perform matrix operations.

It is unusual to implement the Ordinary Least Squares procedure yourself unless as an exercise in linear algebra. It is more likely that you will call a procedure in a linear algebra library. This procedure is very fast to calculate.

### **3. Gradient Descent**

When there are one or more inputs you can use a process of optimizing the values of the coefficients by iteratively minimizing the error of the model on your training data.

This operation is called [Gradient Descent](https://en.wikipedia.org/wiki/Gradient_descent) and works by starting with random values for each coefficient. The sum of the squared errors are calculated for each pair of input and output values. A learning rate is used as a scale factor and the coefficients are updated in the direction towards minimizing the error. The process is repeated until a minimum sum squared error is achieved or no further improvement is possible.

When using this method, you must select a learning rate (alpha) parameter that determines the size of the improvement step to take on each iteration of the procedure.

Gradient descent is often taught using a linear regression model because it is relatively straightforward to understand. In practice, it is useful when you have a very large dataset either in the number of rows or the number of columns that may not fit into memory.

### **4. Regularization**

There are extensions of the training of the linear model called regularization methods. These seek to both minimize the sum of the squared error of the model on the training data (using ordinary least squares) but also to reduce the complexity of the model (like the number or absolute size of the sum of all coefficients in the model).

Two popular examples of regularization procedures for linear regression are:

* [Lasso Regression](https://en.wikipedia.org/wiki/Lasso_(statistics)): where Ordinary Least Squares is modified to also minimize the absolute sum of the coefficients (called L1 regularization).
* [Ridge Regression](https://en.wikipedia.org/wiki/Tikhonov_regularization): where Ordinary Least Squares is modified to also minimize the squared absolute sum of the coefficients (called L2 regularization).

These methods are effective to use when there is collinearity in your input values and ordinary least squares would overfit the training data.

Now that you know some techniques to learn the coefficients in a linear regression model, let’s look at how we can use a model to make predictions on new data.

## **Making Predictions with Linear Regression**

Given the representation is a linear equation, making predictions is as simple as solving the equation for a specific set of inputs.

Let’s make this concrete with an example. Imagine we are predicting weight (y) from height (x). Our linear regression model representation for this problem would be:

y = B0 + B1 \* x1

or

weight =B0 +B1 \* height

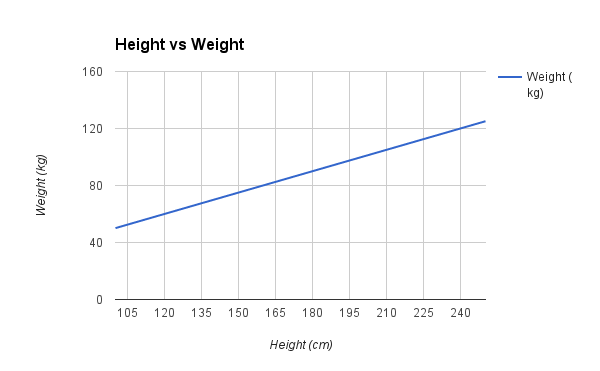
Where B0 is the bias coefficient and B1 is the coefficient for the height column. We use a learning technique to find a good set of coefficient values. Once found, we can plug in different height values to predict the weight.

For example, lets use B0 = 0.1 and B1 = 0.5. Let’s plug them in and calculate the weight (in kilograms) for a person with the height of 182 centimeters.

weight = 0.1 + 0.05 \* 182

weight = 91.1

You can see that the above equation could be plotted as a line in two-dimensions. The B0 is our starting point regardless of what height we have. We can run through a bunch of heights from 100 to 250 centimeters and plug them to the equation and get weight values, creating our line.



Sample Height vs Weight Linear Regression

Now that we know how to make predictions given a learned linear regression model, let’s look at some rules of thumb for preparing our data to make the most of this type of model.

## **Preparing Data For Linear Regression**

Linear regression is been studied at great length, and there is a lot of literature on how your data must be structured to make best use of the model.

As such, there is a lot of sophistication when talking about these requirements and expectations which can be intimidating. In practice, you can uses these rules more as rules of thumb when using Ordinary Least Squares Regression, the most common implementation of linear regression.

Try different preparations of your data using these heuristics and see what works best for your problem.

* **Linear Assumption**. Linear regression assumes that the relationship between your input and output is linear. It does not support anything else. This may be obvious, but it is good to remember when you have a lot of attributes. You may need to transform data to make the relationship linear (e.g. log transform for an exponential relationship).
* **Remove Noise**. Linear regression assumes that your input and output variables are not noisy. Consider using data cleaning operations that let you better expose and clarify the signal in your data. This is most important for the output variable and you want to remove outliers in the output variable (y) if possible.
* **Remove Collinearity**. Linear regression will over-fit your data when you have highly correlated input variables. Consider calculating pairwise correlations for your input data and removing the most correlated.
* **Gaussian Distributions**. Linear regression will make more reliable predictions if your input and output variables have a Gaussian distribution. You may get some benefit using transforms (e.g. log or BoxCox) on you variables to make their distribution more Gaussian looking.
* **Rescale Inputs**: Linear regression will often make more reliable predictions if you rescale input variables using standardization or normalization.

See the [Wikipedia article on Linear Regression](https://en.wikipedia.org/wiki/Linear_regression#Assumptions) for an excellent list of the assumptions made by the model. There’s also a great list of assumptions on the [Ordinary Least Squares Wikipedia article](https://en.wikipedia.org/wiki/Ordinary_least_squares#Assumptions).

## **Further Reading**

There’s plenty more out there to read on linear regression. Start using it before you do more reading, but when you want to dive deeper, below are some references you could use.

In [statistics](https://en.wikipedia.org/wiki/Statistics), **ordinary least squares** (**OLS**) is a type of [linear least squares](https://en.wikipedia.org/wiki/Linear_least_squares) method for estimating the unknown [parameters](https://en.wikipedia.org/wiki/Statistical_parameter) in a [linear regression](https://en.wikipedia.org/wiki/Linear_regression) model. OLS chooses the parameters of a [linear function](https://en.wikipedia.org/wiki/Linear_function) of a set of [explanatory variables](https://en.wikipedia.org/wiki/Explanatory_variable) by the principle of [least squares](https://en.wikipedia.org/wiki/Least_squares): minimizing the sum of the squares of the differences between the observed [dependent variable](https://en.wikipedia.org/wiki/Dependent_variable) (values of the variable being predicted) in the given [dataset](https://en.wikipedia.org/wiki/Dataset) and those predicted by the linear function.

Geometrically, this is seen as the sum of the squared distances, parallel to the axis of the dependent variable, between each data point in the set and the corresponding point on the regression surface – the smaller the differences, the better the model fits the data. The resulting [estimator](https://en.wikipedia.org/wiki/Statistical_estimation) can be expressed by a simple formula, especially in the case of a [simple linear regression](https://en.wikipedia.org/wiki/Simple_linear_regression) (single regressor) on the right-hand side.

The OLS estimator is [consistent](https://en.wikipedia.org/wiki/Consistent_estimator) when the [regressors](https://en.wikipedia.org/wiki/Regressors) are [exogenous](https://en.wikipedia.org/wiki/Exogenous), and [optimal in the class of linear unbiased estimators](https://en.wikipedia.org/wiki/Best_linear_unbiased_estimator) when the [errors](https://en.wikipedia.org/wiki/Statistical_error) are [homoscedastic](https://en.wikipedia.org/wiki/Homoscedastic) and [serially uncorrelated](https://en.wikipedia.org/wiki/Autocorrelation). Under these conditions, the method of OLS provides [minimum-variance mean-unbiased](https://en.wikipedia.org/wiki/UMVU) estimation when the errors have finite [variances](https://en.wikipedia.org/wiki/Variance). Under the additional assumption that the errors are [normally distributed](https://en.wikipedia.org/wiki/Normal_distribution), OLS is the [maximum likelihood estimator](https://en.wikipedia.org/wiki/Maximum_likelihood_estimator).

OLS is used in fields as diverse as [economics](https://en.wikipedia.org/wiki/Economics) ([econometrics](https://en.wikipedia.org/wiki/Econometrics)), [political science](https://en.wikipedia.org/wiki/Political_science), [psychology](https://en.wikipedia.org/wiki/Psychology) and [engineering](https://en.wikipedia.org/wiki/Engineering) ([control theory](https://en.wikipedia.org/wiki/Control_theory) and [signal processing](https://en.wikipedia.org/wiki/Signal_processing)).