WIKIPEDIA

Linear regression

In statistics, **linear regression** is a linear approach to modelling the relationship between a scalar response and one or more explanatory variables (also known as dependent and independent variables). The case of one explanatory variable is called *simple linear regression*; for more than one, the process is called **multiple linear regression**.[1] This term is distinct from multivariate linear regression, where multiple correlated dependent variables are predicted, rather than a single scalar variable.[2]

In linear regression, the relationships are modeled using linear predictor functions whose unknown model parameters are estimated from the data. Such models are called linear models.[3] Most commonly, the conditional mean of the response given the values of the explanatory variables (or predictors) is assumed to be an affine function of those values; less commonly, the conditional median or some other quantile is used. Like all forms of regression analysis, linear regression focuses on the conditional probability distribution of the response given the values of the predictors, rather than on the joint probability distribution of all of these variables, which is the domain of multivariate analysis.

Linear regression was the first type of regression analysis to be studied rigorously, and to be used extensively in practical applications.[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, forecasting, or error reduction, linear regression can be used to fit a predictive model to an observed 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 approach, but they may also be fitted in other ways, such as by minimizing the "lack of fit" in some other norm (as with least absolute deviations regression), or by minimizing a penalized version of the least squares cost function as in ridge regression (L^2-norm penalty) and lasso (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.

Contents

Introduction
- Assumptions
- Interpretation

Extensions
- Simple and multiple linear regression
- General linear models
- Heteroscedastic models
- Generalized linear models
- Hierarchical linear models
- Errors-in-variables
- Others

Estimation methods
- Least-squares estimation and related techniques
- Maximum-likelihood estimation and related techniques
- Other estimation techniques

Applications
- Trend line
- Epidemiology
- Finance
- Economics
- Environmental science
- Machine learning

History

See also

References
- Citations
- Sources

Further reading

External links

Introduction

Given a data set $\{y_i,\, x_{i1}, \ldots, x_{ip}\}_{i=1}^n$ of n statistical units, a linear regression model assumes that the relationship between the dependent variable y and the p-vector of regressors $\mathbf{x}$ is linear. This relationship is modeled through a *disturbance term* or *error variable* ε — an unobserved random variable that adds "noise" to the linear relationship between the dependent variable and regressors. Thus the model takes the form

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i = \mathbf{x}_i^{\mathsf{T}} \boldsymbol{\beta} + \varepsilon_i, \qquad i = 1, \ldots, n,$$

where $^{\mathrm{T}}$ denotes the transpose, so that $\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta}$ is the inner product between vectors $\mathbf{x}_i$ and $\boldsymbol{\beta}$.

Often these n equations are stacked together and written in matrix notation as

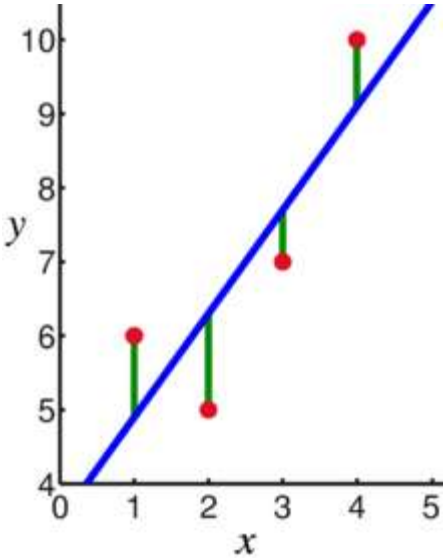
$$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix},$$

$$X = \begin{pmatrix} \mathbf{x}_1^{\mathsf{T}} \\ \mathbf{x}_2^{\mathsf{T}} \\ \vdots \\ \mathbf{x}_n^{\mathsf{T}} \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix},$$

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}, \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$



In linear regression, the observations (**red**) are assumed to be the result of random deviations (**green**) from an underlying relationship (**blue**) between a dependent variable (*y*) and an independent variable (*x*).

Some remarks on notation and terminology:

- $\mathbf{y}$ is a vector of observed values y_i $(i = 1, \ldots, n)$ of the variable called the *regressand*, *endogenous variable*, *response variable*, *measured variable*, *criterion variable*, or *dependent variable*. This variable is also sometimes known as the *predicted variable*, but this should not be confused with *predicted values*, which are denoted $\hat{\mathbf{y}}$. The decision as to which variable in a data set is modeled as the dependent variable and which are modeled as the independent variables may be based on a presumption that the value of one of the variables is caused by, or directly influenced by the other variables. Alternatively, there may be an operational reason to model one of the variables in terms of the others, in which case there need be no presumption of causality.
- X may be seen as a matrix of row-vectors $\mathbf{x}_i$ or of *n*-dimensional column-vectors X_j, which are known as *regressors*, *exogenous variables*, *explanatory variables*, *covariates*, *input variables*, *predictor variables*, or *independent variables* (not to be confused with the concept of independent random variables). The matrix X is sometimes called the design matrix.
  - Usually a constant is included as one of the regressors. In particular, $\mathbf{x}_{i0} = 1$ for $i = 1, \ldots, n$. The corresponding element of $\boldsymbol{\beta}$ is called the *intercept*. Many statistical inference procedures for linear models require an intercept to be present, so it is often included even if theoretical considerations suggest that its value should be zero.
  - Sometimes one of the regressors can be a non-linear function of another regressor or of the data, as in polynomial regression and segmented regression. The model remains linear as long as it is linear in the parameter vector $\boldsymbol{\beta}$.
  - The values x_{ij} may be viewed as either observed values of random variables X_j or as fixed values chosen prior to observing the dependent variable. Both interpretations may be appropriate in different cases, and they generally lead to the same estimation procedures; however different approaches to asymptotic analysis are used in these two situations.
- $\boldsymbol{\beta}$ is a $(p + 1)$-dimensional *parameter vector*, where β_0 is the intercept term (if one is included in the model—otherwise $\boldsymbol{\beta}$ is *p*-dimensional). Its elements are known as *effects* or *regression coefficients* (although the latter term is sometimes reserved for the *estimated* effects). Statistical estimation and

inference in linear regression focuses on $\boldsymbol{\beta}$. The elements of this parameter vector are interpreted as the partial derivatives of the dependent variable with respect to the various independent variables.

- $\boldsymbol{\varepsilon}$ is a vector of values ε_i. This part of the model is called the *error term*, *disturbance term*, or sometimes *noise* (in contrast with the "signal" provided by the rest of the model). This variable captures all other factors which influence the dependent variable *y* other than the regressors **x**. The relationship between the error term and the regressors, for example their correlation, is a crucial consideration in formulating a linear regression model, as it will determine the appropriate estimation method.

Fitting a linear model to a given data set usually requires estimating the regression coefficients $\boldsymbol{\beta}$ such that the error term $\boldsymbol{\varepsilon} = \mathbf{y} - \boldsymbol{X\beta}$ is minimized. For example, it is common to use the sum of squared errors $||\boldsymbol{\varepsilon}||_2^2$ as a measure of $\boldsymbol{\varepsilon}$ for minimization.

Example. Consider a situation where a small ball is being tossed up in the air and then we measure its heights of ascent h_i at various moments in time t_i. Physics tells us that, ignoring the drag, the relationship can be modeled as

$$h_i = \beta_1 t_i + \beta_2 t_i^2 + \varepsilon_i,$$

where β_1 determines the initial velocity of the ball, β_2 is proportional to the standard gravity, and ε_i is due to measurement errors. Linear regression can be used to estimate the values of β_1 and β_2 from the measured data. This model is non-linear in the time variable, but it is linear in the parameters β_1 and β_2; if we take regressors $\mathbf{x}_i = (x_{i1}, x_{i2}) = (t_i, t_i^2)$, the model takes on the standard form

$$h_i = \mathbf{x}_i^{\mathsf{T}} \boldsymbol{\beta} + \varepsilon_i.$$

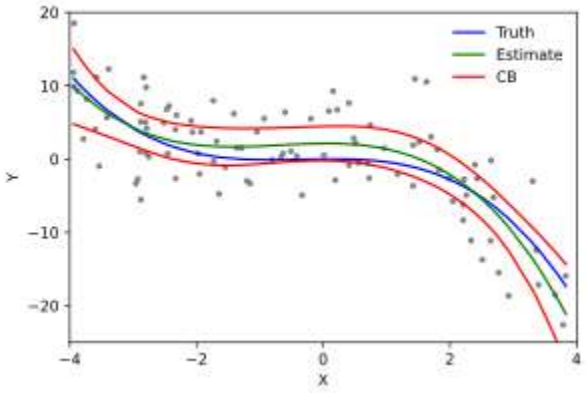
Assumptions

Standard linear regression models with standard estimation techniques make a number of assumptions about the predictor variables, the response variables and their relationship. Numerous extensions have been developed that allow each of these assumptions to be relaxed (i.e. reduced to a weaker form), and in some cases eliminated entirely. Generally these extensions make the estimation procedure more complex and time-consuming, and may also require more data in order to produce an equally precise model.

The following are the major assumptions made by standard linear regression models with standard estimation techniques (e.g. ordinary least squares):

- **Weak exogeneity**. This essentially means that the predictor variables *x* can be treated as fixed values, rather than random variables. This means, for example, that the predictor variables are assumed to be error-free—that is, not contaminated with measurement errors. Although this assumption is not realistic in many settings, dropping it leads to significantly more difficult errors-in-variables models.
- **Linearity**. This means that the mean of the response variable is a linear combination of the parameters (regression coefficients) and the predictor variables. Note that this assumption is much less restrictive than it may at first seem. Because the predictor variables are treated as fixed values (see above), linearity is really only a restriction on the parameters. The predictor variables themselves can be arbitrarily transformed, and in fact multiple copies of the same underlying predictor variable can be added, each one transformed differently. This technique is used, for example, in polynomial regression, which uses linear regression to fit the response variable as an arbitrary polynomial function (up to a given rank) of a predictor variable. With this much flexibility,

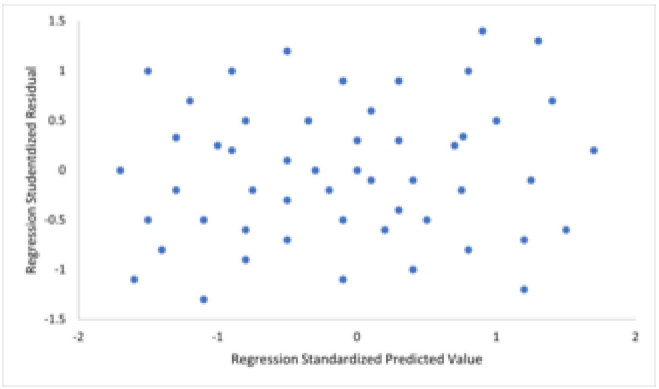
models such as polynomial regression often have "too much power", in that they tend to overfit the data. As a result, some kind of regularization must typically be used to prevent unreasonable solutions coming out of the estimation process. Common examples are ridge regression and lasso regression. Bayesian linear regression can also be used, which by its nature is more or less immune to the problem of overfitting. (In fact, ridge regression and lasso regression can both be viewed as special cases of Bayesian linear regression, with particular types of prior distributions placed on the regression coefficients.)



Example of a cubic polynomial regression, which is a type of linear regression. Although *polynomial regression* fits a nonlinear model to the data, as a statistical estimation problem it is linear, in the sense that the regression function E(*y* | *x*) is linear in the unknown parameters that are estimated from the data. For this reason, polynomial regression is considered to be a special case of multiple linear regression.

- **Constant variance** (a.k.a. **homoscedasticity**). This means that the variance of the errors does not depend on the values of the predictor variables. Thus the variability of the responses for given fixed values of the predictors is the same regardless of how large or small the responses are. This is often not the case, as a variable whose mean is large will typically have a greater variance than one whose mean is small. For example, a person whose income is predicted to be $100,000 may easily have an actual income of $80,000 or $120,000—i.e., a standard deviation of around $20,000—while another person with a predicted income of $10,000 is unlikely to have the same $20,000 standard deviation, since that would imply their actual income could vary anywhere between −$10,000 and $30,000. (In fact, as this shows, in many cases—often the same cases where the assumption of normally distributed errors fails—the variance or standard deviation should be predicted to be proportional to the mean, rather than constant.) The absence of homoscedasticity is called heteroscedasticity. In order to check this assumption, a plot of residuals versus predicted values (or the values of each individual predictor) can be examined for a "fanning effect" (i.e., increasing or decreasing vertical spread as one moves left to right on the plot). A plot of the absolute or squared residuals versus the predicted values (or each predictor) can also be examined for a trend or curvature. Formal tests can also be used; see Heteroscedasticity. The presence of heteroscedasticity will result in an overall "average" estimate of variance being used instead of one that takes into account the true variance structure. This leads to less precise (but in the case of ordinary least squares, not biased) parameter estimates and biased standard errors, resulting in misleading tests and interval estimates. The mean squared error for the model will also be wrong. Various estimation techniques including weighted least squares and the use of heteroscedasticity-consistent standard errors can handle heteroscedasticity in a quite general way. Bayesian linear regression techniques can also be used when the variance is assumed to be a function of the mean. It is also possible in some cases to fix the problem by applying a transformation to the response variable (e.g., fitting the logarithm of the response variable using a linear regression model, which implies that the response variable itself has a log-normal distribution rather than a normal distribution).
- **Independence of errors**. This assumes that the errors of the response variables are uncorrelated with each other. (Actual statistical independence is a stronger condition than mere lack of correlation and is often not needed, although it can be exploited if it is known to hold.) Some methods such as generalized least squares are capable of handling correlated errors, although they typically require significantly more data unless some sort of regularization is used to bias the model towards assuming uncorrelated errors. Bayesian linear regression is a general way of handling this issue.
- **Lack of perfect multicollinearity** in the predictors. For standard least squares estimation methods, the design matrix *X* must have full column rank *p*; otherwise perfect multicollinearity exists in the

predictor variables, meaning a linear relationship exists between two or more predictor variables. This can be caused by accidentally duplicating a variable in the data, using a linear transformation of a variable along with the original (e.g., the same temperature measurements expressed in Fahrenheit and Celcius), or including a linear combination of multiple variables in the model, such as their mean. It can also happen if there is too little data available compared to the number of parameters to be estimated (e.g., fewer data points than regression coefficients). Near violations of this assumption, where predictors are highly but not perfectly correlated, can reduce the precision of parameter estimates (see Variance inflation factor). In the case of perfect multicollinearity, the parameter vector $\boldsymbol{\beta}$ will be non-identifiable—it has no unique solution. In such a case, only some of the parameters can be identified (i.e., their values can only be estimated within some linear subspace of the full parameter space $\mathbf{R}^p$). See partial least squares regression. Methods for fitting linear models with multicollinearity have been developed,[5][6][7][8] some of which require additional assumptions such as "effect sparsity"—that a large fraction of the effects are exactly zero. Note that the more computationally expensive iterated algorithms for parameter estimation, such as those used in generalized linear models, do not suffer from this problem.



To check for violations of the assumptions of lineaarity, constant variance, and independence of errors within a linear regression model, the residuals are typically plotted against the predicted values (or each of the individual predictors). An apparently random scatter of points about the horizontal midline at 0 is ideal, but cannot rule out certain kinds of violations such as autocorrelation in the errors or their correlation with one or more covariates.

Beyond these assumptions, several other statistical properties of the data strongly influence the performance of different estimation methods:

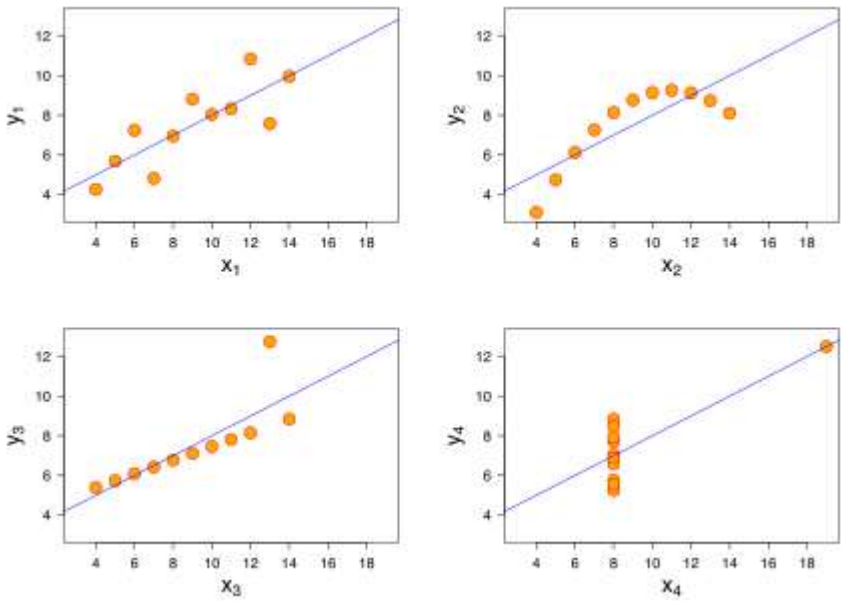
- The statistical relationship between the error terms and the regressors plays an important role in determining whether an estimation procedure has desirable sampling properties such as being unbiased and consistent.
- The arrangement, or probability distribution of the predictor variables **x** has a major influence on the precision of estimates of $\boldsymbol{\beta}$. Sampling and design of experiments are highly developed subfields of statistics that provide guidance for collecting data in such a way to achieve a precise estimate of $\boldsymbol{\beta}$.

Interpretation

A fitted linear regression model can be used to identify the relationship between a single predictor variable x_j and the response variable y when all the other predictor variables in the model are "held fixed". Specifically, the interpretation of β_j is the expected change in y for a one-unit change in x_j when the other covariates are held fixed—that is, the expected value of the partial derivative of y with respect to x_j. This is sometimes called the *unique effect* of x_j on y. In contrast, the *marginal effect* of x_j on y can be assessed using a correlation coefficient or simple linear regression model relating only x_j to y; this effect is the total derivative of y with respect to x_j.

Care must be taken when interpreting regression results, as some of the regressors may not allow for marginal changes (such as dummy variables, or the intercept term), while others cannot be held fixed (recall the example from the introduction: it would be impossible to "hold t_i fixed" and at the same time change the value of t_i^2).

It is possible that the unique effect can be nearly zero even when the marginal effect is large. This may imply that some other covariate captures all the information in x_j, so that once that variable is in the model, there is no contribution of x_j to the variation in y. Conversely, the unique effect of x_j can be large while its marginal effect is nearly zero. This would happen if the other covariates explained a great deal of the variation of y, but they mainly explain variation in a way that is complementary to what is captured by x_j. In this case, including the other variables in the model reduces the part of the variability of y that is unrelated to x_j, thereby strengthening the apparent relationship with x_j.



The data sets in the Anscombe's quartet are designed to have approximately the same linear regression line (as well as nearly identical means, standard deviations, and correlations) but are graphically very different. This illustrates the pitfalls of relying solely on a fitted model to understand the relationship between variables.

The meaning of the expression "held fixed" may depend on how the values of the predictor variables arise. If the experimenter directly sets the values of the predictor variables according to a study design, the comparisons of interest may literally correspond to comparisons among units whose predictor variables have been "held fixed" by the experimenter. Alternatively, the expression "held fixed" can refer to a selection that takes place in the context of data analysis. In this case, we "hold a variable fixed" by restricting our attention to the subsets of the data that happen to have a common value for the given predictor variable. This is the only interpretation of "held fixed" that can be used in an observational study.

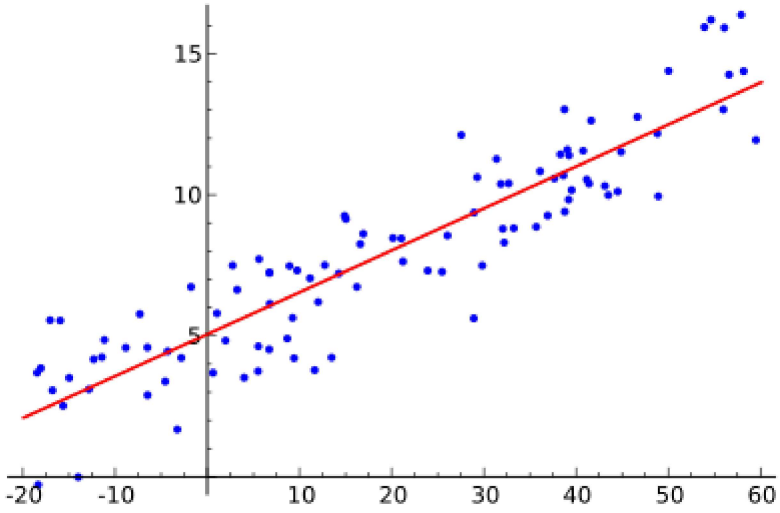
The notion of a "unique effect" is appealing when studying a complex system where multiple interrelated components influence the response variable. In some cases, it can literally be interpreted as the causal effect of an intervention that is linked to the value of a predictor variable. However, it has been argued that in many cases multiple regression analysis fails to clarify the relationships between the predictor variables and the response variable when the predictors are correlated with each other and are not assigned following a study design.[9] Commonality analysis may be helpful in disentangling the shared and unique impacts of correlated independent variables.[10]

Extensions

Numerous extensions of linear regression have been developed, which allow some or all of the assumptions underlying the basic model to be relaxed.

Simple and multiple linear regression

The very simplest case of a single scalar predictor variable x and a single scalar response variable y is known as *simple linear regression*. The extension to multiple and/or vector-valued predictor variables (denoted with a capital X) is known as *multiple linear regression*, also known as *multivariable linear regression* (not to be confused with *multivariate linear regression* [11]).



Example of simple linear regression, which has one independent variable

Multiple linear regression is a generalization of simple linear regression to the case of more than one independent variable, and a special case of general linear models, restricted to one dependent variable. The basic model for multiple linear regression is

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_p X_{ip} + \epsilon_i$$

for each observation $i = 1, \ldots, n$.

In the formula above we consider n observations of one dependent variable and p independent variables. Thus, Y_i is the i^{th} observation of the dependent variable, X_{ij} is i^{th} observation of the j^{th} independent variable, $j = 1, 2, \ldots, p$. The values β_j represent parameters to be estimated, and ε_i is the i^{th} independent identically distributed normal error.

In the more general multivariate linear regression, there is one equation of the above form for each of $m > 1$ dependent variables that share the same set of explanatory variables and hence are estimated simultaneously with each other:

$$Y_{ij} = \beta_{0j} + \beta_{1j} X_{i1} + \beta_{2j} X_{i2} + \ldots + \beta_{pj} X_{ip} + \epsilon_{ij}$$

for all observations indexed as $i = 1, \ldots, n$ and for all dependent variables indexed as $j = 1, \ldots,$ m.

Nearly all real-world regression models involve multiple predictors, and basic descriptions of linear regression are often phrased in terms of the multiple regression model. Note, however, that in these cases the response variable y is still a scalar. Another term, *multivariate linear regression*, refers to cases where y is a vector, i.e., the same as *general linear regression*.

General linear models

The general linear model considers the situation when the response variable is not a scalar (for each observation) but a vector, $\mathbf{y}_i$. Conditional linearity of $E(\mathbf{y} \mid \mathbf{x}_i) = \mathbf{x}_i^{\mathsf{T}} B$ is still assumed, with a matrix B replacing the vector $\boldsymbol{\beta}$ of the classical linear regression model. Multivariate analogues of ordinary least squares (OLS) and generalized least squares (GLS) have been developed. "General linear models" are also called "multivariate linear models". These are not the same as multivariable linear models (also called "multiple linear models").

Heteroscedastic models

Various models have been created that allow for heteroscedasticity, i.e. the errors for different response variables may have different variances. For example, weighted least squares is a method for estimating linear regression models when the response variables may have different error variances, possibly with correlated errors. (See also Weighted linear least squares, and Generalized least squares.) Heteroscedasticity-consistent standard errors is an improved method for use with uncorrelated but potentially heteroscedastic errors.

Generalized linear models

Generalized linear models (GLMs) are a framework for modeling response variables that are bounded or discrete. This is used, for example:

- when modeling positive quantities (e.g. prices or populations) that vary over a large scale—which are better described using a skewed distribution such as the log-normal distribution or Poisson distribution (although GLMs are not used for log-normal data, instead the response variable is simply transformed using the logarithm function);
- when modeling categorical data, such as the choice of a given candidate in an election (which is better described using a Bernoulli distribution/binomial distribution for binary choices, or a categorical distribution/multinomial distribution for multi-way choices), where there are a fixed number of choices that cannot be meaningfully ordered;
- when modeling ordinal data, e.g. ratings on a scale from 0 to 5, where the different outcomes can be ordered but where the quantity itself may not have any absolute meaning (e.g. a rating of 4 may not be "twice as good" in any objective sense as a rating of 2, but simply indicates that it is better than 2 or 3 but not as good as 5).

Generalized linear models allow for an arbitrary *link function*, g, that relates the mean of the response variable(s) to the predictors: $\boldsymbol{E(Y) = g^{-1}(XB)}$. The link function is often related to the distribution of the response, and in particular it typically has the effect of transforming between the $(-\infty, \infty)$ range of the linear predictor and the range of the response variable.

Some common examples of GLMs are:

- Poisson regression for count data.
- Logistic regression and probit regression for binary data.
- Multinomial logistic regression and multinomial probit regression for categorical data.
- Ordered logit and ordered probit regression for ordinal data.

Single index models allow some degree of nonlinearity in the relationship between x and y, while preserving the central role of the linear predictor $\beta' x$ as in the classical linear regression model. Under certain conditions, simply applying OLS to data from a single-index model will consistently estimate β up to a proportionality constant.[12]

Hierarchical linear models

Hierarchical linear models (or *multilevel regression*) organizes the data into a hierarchy of regressions, for example where A is regressed on B, and B is regressed on C. It is often used where the variables of interest have a natural hierarchical structure such as in educational statistics, where students are nested

in classrooms, classrooms are nested in schools, and schools are nested in some administrative grouping, such as a school district. The response variable might be a measure of student achievement such as a test score, and different covariates would be collected at the classroom, school, and school district levels.

Errors-in-variables

Errors-in-variables models (or "measurement error models") extend the traditional linear regression model to allow the predictor variables X to be observed with error. This error causes standard estimators of β to become biased. Generally, the form of bias is an attenuation, meaning that the effects are biased toward zero.

Others

- In Dempster–Shafer theory, or a linear belief function in particular, a linear regression model may be represented as a partially swept matrix, which can be combined with similar matrices representing observations and other assumed normal distributions and state equations. The combination of swept or unswept matrices provides an alternative method for estimating linear regression models.

Estimation methods

A large number of procedures have been developed for parameter estimation and inference in linear regression. These methods differ in computational simplicity of algorithms, presence of a closed-form solution, robustness with respect to heavy-tailed distributions, and theoretical assumptions needed to validate desirable statistical properties such as consistency and asymptotic efficiency.

Some of the more common estimation techniques for linear regression are summarized below.

Least-squares estimation and related techniques

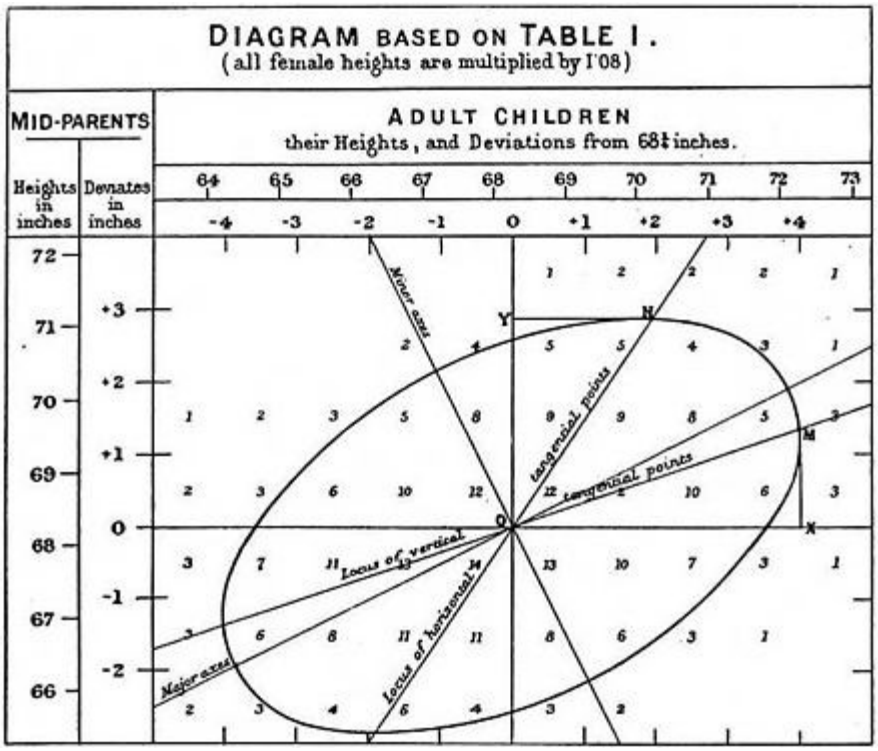
Assuming that the independent variable is $\overrightarrow{x_i} = \left[x_1^i, x_2^i, \ldots, x_m^i\right]$ and the model's parameters are $\vec{\beta} = [\beta_0, \beta_1, \ldots, \beta_m]$, then the model's prediction would be

$$y_i \approx \beta_0 + \sum_{j=1}^{m} \beta_j \times x_j^i.$$

If $\overrightarrow{x_i}$ is extended to $\overrightarrow{x_i} = \left[1, x_1^i, x_2^i, \ldots, x_m^i\right]$ then y_i would become a dot product of the parameter and the independent variable, i.e.

$$y_i \approx \sum_{j=0}^{m} \beta_j \times x_j^i = \vec{\beta} \cdot \overrightarrow{x_i}.$$

In the least-squares setting, the optimum parameter is defined as such that minimizes the sum of mean squared loss:



Francis Galton's 1886[13] illustration of the correlation between the heights of adults and their parents. The observation that adult children's heights tended to deviate less from the mean height than their parents suggested the concept of "regression toward the mean", giving regression its name. The "locus of horizontal tangential points" passing through the leftmost and rightmost points on the ellipse (which is a level curve of the bivariate normal distribution estimated from the data) is the OLS estimate of the regression of parents' heights on children's heights, while the "locus of vertical tangential points" is the OLS estimate of the regression of children's heights on parent's heights. The major axis of the ellipse is the TLS estimate.

$$\hat{\vec{\beta}} = \underset{\vec{\beta}}{\arg\min}\, L\left(D, \vec{\beta}\right) = \underset{\vec{\beta}}{\arg\min} \sum_{i=1}^{n} \left(\vec{\beta} \cdot \vec{x_i} - y_i\right)^2$$

Now putting the independent and dependent variables in matrices $\boldsymbol{X}$ and $\boldsymbol{Y}$ respectively, the loss function can be rewritten as:

$$\begin{aligned} L\left(D, \vec{\beta}\right) &= \|X\vec{\beta} - Y\|^2 \\ &= \left(X\vec{\beta} - Y\right)^{\mathsf{T}} \left(X\vec{\beta} - Y\right) \\ &= Y^{\mathsf{T}}Y - Y^{\mathsf{T}}X\vec{\beta} - \vec{\beta}^{\mathsf{T}}X^{\mathsf{T}}Y + \vec{\beta}^{\mathsf{T}}X^{\mathsf{T}}X\vec{\beta} \end{aligned}$$

As the loss is convex the optimum solution lies at gradient zero. The gradient of the loss function is (using Denominator layout convention):

$$\begin{aligned}\frac{\partial L\left(D,\vec{\beta}\right)}{\partial\vec{\beta}} &= \frac{\partial\left(Y^{\mathsf{T}}Y - Y^{\mathsf{T}}X\vec{\beta} - \vec{\beta}^{\mathsf{T}}X^{\mathsf{T}}Y + \vec{\beta}^{\mathsf{T}}X^{\mathsf{T}}X\vec{\beta}\right)}{\partial\vec{\beta}} \\ &= -2X^{\mathsf{T}}Y + 2X^{\mathsf{T}}X\vec{\beta}\end{aligned}$$

Setting the gradient to zero produces the optimum parameter:

$$\begin{aligned}-2X^{\mathsf{T}}Y + 2X^{\mathsf{T}}X\vec{\beta} &= 0 \\ \Rightarrow X^{\mathsf{T}}Y &= X^{\mathsf{T}}X\vec{\beta} \\ \Rightarrow \vec{\hat{\beta}} &= \left(X^{\mathsf{T}}X\right)^{-1}X^{\mathsf{T}}Y\end{aligned}$$

Note: To prove that the $\hat{\beta}$ obtained is indeed the local minimum, one needs to differentiate once more to obtain the Hessian matrix and show that it is positive definite. This is provided by the Gauss–Markov theorem.

Linear least squares methods include mainly:

- Ordinary least squares
- Weighted least squares
- Generalized least squares

Maximum-likelihood estimation and related techniques

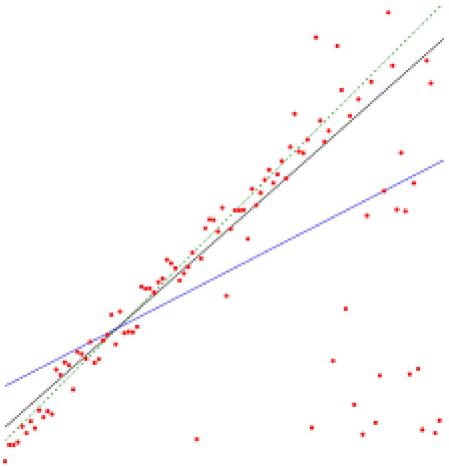
- **Maximum likelihood estimation** can be performed when the distribution of the error terms is known to belong to a certain parametric family f_θ of probability distributions.[14] When f_θ is a normal distribution with zero mean and variance θ, the resulting estimate is identical to the OLS estimate. GLS estimates are maximum likelihood estimates when ε follows a multivariate normal distribution with a known covariance matrix.
- **Ridge regression**[15][16][17] and other forms of penalized estimation, such as **Lasso regression**,[5] deliberately introduce bias into the estimation of β in order to reduce the variability of the estimate. The resulting estimates generally have lower mean squared error than the OLS estimates, particularly when multicollinearity is present or when overfitting is a problem. They are generally used when the goal is to predict the value of the response variable y for values of the predictors x that have not yet been observed. These methods are not as commonly used when the goal is inference, since it is difficult to account for the bias.
- **Least absolute deviation** (LAD) regression is a robust estimation technique in that it is less sensitive to the presence of outliers than OLS (but is less efficient than OLS when no outliers are present). It is equivalent to maximum likelihood estimation under a Laplace distribution model for ε.[18]
- **Adaptive estimation**. If we assume that error terms are independent of the regressors, $\varepsilon_i \perp \mathbf{x}_i$, then the optimal estimator is the 2-step MLE, where the first step is used to non-parametrically estimate the distribution of the error term.[19]

Other estimation techniques

- **Bayesian linear regression** applies the framework of Bayesian statistics to linear regression. (See also Bayesian multivariate linear regression.) In particular, the regression coefficients β are assumed

to be random variables with a specified prior distribution. The prior distribution can bias the solutions for the regression coefficients, in a way similar to (but more general than) ridge regression or lasso regression. In addition, the Bayesian estimation process produces not a single point estimate for the "best" values of the regression coefficients but an entire posterior distribution, completely describing the uncertainty surrounding the quantity. This can be used to estimate the "best" coefficients using the mean, mode, median, any quantile (see quantile regression), or any other function of the posterior distribution.

- **Quantile regression** focuses on the conditional quantiles of y given X rather than the conditional mean of y given X. Linear quantile regression models a particular conditional quantile, for example the conditional median, as a linear function $\beta^T x$ of the predictors.
- **Mixed models** are widely used to analyze linear regression relationships involving dependent data when the dependencies have a known structure. Common applications of mixed models include analysis of data involving repeated measurements, such as longitudinal data, or data obtained from cluster sampling. They are generally fit as parametric models, using maximum likelihood or Bayesian estimation. In the case where the errors are modeled as normal random variables, there is a close connection between mixed models and generalized least squares.[20] Fixed effects estimation is an alternative approach to analyzing this type of data.
- **Principal component regression** (PCR)[7][8] is used when the number of predictor variables is large, or when strong correlations exist among the predictor variables. This two-stage procedure first reduces the predictor variables using principal component analysis then uses the reduced variables in an OLS regression fit. While it often works well in practice, there is no general theoretical reason that the most informative linear function of the predictor variables should lie among the dominant principal components of the multivariate distribution of the predictor variables. The partial least squares regression is the extension of the PCR method which does not suffer from the mentioned deficiency.
- **Least-angle regression**[6] is an estimation procedure for linear regression models that was developed to handle high-dimensional covariate vectors, potentially with more covariates than observations.
- The **Theil–Sen estimator** is a simple robust estimation technique that chooses the slope of the fit line to be the median of the slopes of the lines through pairs of sample points. It has similar statistical efficiency properties to simple linear regression but is much less sensitive to outliers.[21]
- Other robust estimation techniques, including the **α-trimmed mean** approach, and **L-, M-, S-, and R-estimators** have been introduced.



Comparison of the Theil–Sen estimator (black) and simple linear regression (blue) for a set of points with outliers.

Applications

Linear regression is widely used in biological, behavioral and social sciences to describe possible relationships between variables. It ranks as one of the most important tools used in these disciplines.

Trend line

A **trend line** represents a trend, the long-term movement in time series data after other components have been accounted for. It tells whether a particular data set (say GDP, oil prices or stock prices) have increased or decreased over the period of time. A trend line could simply be drawn by eye through a set

of data points, but more properly their position and slope is calculated using statistical techniques like linear regression. Trend lines typically are straight lines, although some variations use higher degree polynomials depending on the degree of curvature desired in the line.

Trend lines are sometimes used in business analytics to show changes in data over time. This has the advantage of being simple. Trend lines are often used to argue that a particular action or event (such as training, or an advertising campaign) caused observed changes at a point in time. This is a simple technique, and does not require a control group, experimental design, or a sophisticated analysis technique. However, it suffers from a lack of scientific validity in cases where other potential changes can affect the data.

Epidemiology

Early evidence relating tobacco smoking to mortality and morbidity came from observational studies employing regression analysis. In order to reduce spurious correlations when analyzing observational data, researchers usually include several variables in their regression models in addition to the variable of primary interest. For example, in a regression model in which cigarette smoking is the independent variable of primary interest and the dependent variable is lifespan measured in years, researchers might include education and income as additional independent variables, to ensure that any observed effect of smoking on lifespan is not due to those other socio-economic factors. However, it is never possible to include all possible confounding variables in an empirical analysis. For example, a hypothetical gene might increase mortality and also cause people to smoke more. For this reason, randomized controlled trials are often able to generate more compelling evidence of causal relationships than can be obtained using regression analyses of observational data. When controlled experiments are not feasible, variants of regression analysis such as instrumental variables regression may be used to attempt to estimate causal relationships from observational data.

Finance

The capital asset pricing model uses linear regression as well as the concept of beta for analyzing and quantifying the systematic risk of an investment. This comes directly from the beta coefficient of the linear regression model that relates the return on the investment to the return on all risky assets.

Economics

Linear regression is the predominant empirical tool in economics. For example, it is used to predict consumption spending,[22] fixed investment spending, inventory investment, purchases of a country's exports,[23] spending on imports,[23] the demand to hold liquid assets,[24] labor demand,[25] and labor supply.[25]

Environmental science

Linear regression finds application in a wide range of environmental science applications. In Canada, the Environmental Effects Monitoring Program uses statistical analyses on fish and benthic surveys to measure the effects of pulp mill or metal mine effluent on the aquatic ecosystem.[26]

Machine learning

Linear regression plays an important role in the subfield of artificial intelligence known as machine learning. The linear regression algorithm is one of the fundamental supervised machine-learning algorithms due to its relative simplicity and well-known properties.[27]

History

Least squares linear regression, as a means of finding a good rough linear fit to a set of points was performed by Legendre (1805) and Gauss (1809) for the prediction of planetary movement. Quetelet was responsible for making the procedure well-known and for using it extensively in the social sciences.[28]

See also

- Analysis of variance
- Blinder–Oaxaca decomposition
- Censored regression model
- Cross-sectional regression
- Curve fitting
- Empirical Bayes methods
- Errors and residuals
- Lack-of-fit sum of squares
- Line fitting
- Linear classifier
- Linear equation
- Logistic regression
- M-estimator
- Multivariate adaptive regression splines
- Nonlinear regression
- Nonparametric regression
- Normal equations
- Projection pursuit regression
- Response modeling methodology
- Segmented linear regression
- Stepwise regression
- Structural break
- Support vector machine
- Truncated regression model

References

Citations

1. David A. Freedman (2009). *Statistical Models: Theory and Practice*. Cambridge University Press. p. 26. "A simple regression equation has on the right hand side an intercept and an explanatory variable with a slope coefficient. A multiple regression e right hand side, each with its own slope coefficient"
2. Rencher, Alvin C.; Christensen, William F. (2012), "Chapter 10, Multivariate regression – Section 10.1, Introduction", *Methods of Multivariate Analysis* (https://books.google.com/books?id=0g-PAuKub3QC&pg=PA19), Wiley Series in Probability and Statistics, **709** (3rd ed.), John Wiley & Sons, p. 19, ISBN 9781118391679.
3. Hilary L. Seal (1967). "The historical development of the Gauss linear model". *Biometrika*. **54** (1/2): 1–24. doi:10.1093/biomet/54.1-2.1 (https://doi.org/10.1093%2Fbiomet%2F54.1-2.1). JSTOR 2333849 (https://www.jstor.org/stable/2333849).
4. Yan, Xin (2009), *Linear Regression Analysis: Theory and Computing* (https://books.google.com/books?id=MjNv6rGv8NIC&pg=PA1), World Scientific, pp. 1–2, ISBN 9789812834119, "Regression analysis ... is probably one of the oldest topics in mathematical statistics dating back to about two hundred years ago. The earliest form of the linear regression was the least squares method, which was published by Legendre in 1805, and by Gauss in 1809 ... Legendre and Gauss both applied the method to the problem of determining, from astronomical observations, the orbits of bodies about the sun."