Regression analysis

From Wikipedia, the free encyclopedia

In statistical modeling, **regression analysis** is a statistical process for estimating the relationships among variables. It includes many techniques for modeling and analyzing several variables, when the focus is on the relationship between a dependent variable and one or more independent variables (or 'predictors'). More specifically, regression analysis helps one understand how the typical value of the dependent variable (or 'criterion variable') changes when any one of the independent variables is varied, while the other independent variables are held fixed. Most commonly, regression analysis estimates the conditional expectation of the dependent variable given the independent variables – that is, the average value of the dependent variable when the independent variables are fixed. Less commonly, the focus is on a quantile, or other location parameter of the conditional distribution of the dependent variable given the independent variables. In all cases, the estimation target is a function of the independent variables called the **regression function**. In regression analysis, it is also of interest to characterize the variation of the dependent variable around the regression function which can be described by a probability distribution. A related but distinct approach is necessary condition analysis[1] (NCA), which estimates the maximum (rather than average) value of the dependent variable for a given value of the independent variable (ceiling line rather than central line) in order to identify what value of the independent variable is necessary but not sufficient for a given value of the dependent variable.

Regression analysis is widely used for prediction and forecasting, where its use has substantial overlap with the field of machine learning. Regression analysis is also used to understand which among the independent variables are related to the dependent variable, and to explore the forms of these relationships. In restricted circumstances, regression analysis can be used to infer causal relationships between the independent and dependent variables. However this can lead to illusions or false relationships, so caution is advisable;[2] for example, correlation does not imply causation.

Many techniques for carrying out regression analysis have been developed. Familiar methods such as linear regression and ordinary least squares regression are parametric, in that the regression function is defined in terms of a finite number of unknown parameters that are estimated from the data. Nonparametric regression refers to techniques that allow the regression function to lie in a specified set of functions, which may be infinite-dimensional.

The performance of regression analysis methods in practice depends on the form of the data generating process, and how it relates to the regression approach being used. Since the true form of the data-generating process is generally not known, regression analysis often depends to some extent on making assumptions about this process. These assumptions are sometimes testable if a sufficient quantity of data is available. Regression models for prediction are often useful even when the assumptions are moderately violated, although they may not perform optimally. However, in many applications, especially with small effects or questions of causality based on observational data, regression methods can give misleading results.[3][4]

In a narrower sense, regression may refer specifically to the estimation of continuous response variables, as opposed to the discrete response variables used in classification.[5] The case of a continuous output variable may be more specifically referred to as **metric regression** to distinguish it from related problems.[6]

Contents

History

The earliest form of regression was the method of least squares, which was published by Legendre in 1805,[7] and by Gauss in 1809.[8] Legendre and Gauss both applied the method to the problem of determining, from astronomical observations, the orbits of bodies about the Sun (mostly comets, but also later the then newly discovered minor planets). Gauss published a further development of the theory of least squares in 1821,[9] including a version of the Gauss–Markov theorem.

The term "regression" was coined by Francis Galton in the nineteenth century to describe a biological phenomenon. The phenomenon was that the heights of descendants of tall ancestors tend to regress down towards a normal average (a phenomenon also known as regression toward the mean).[10][11] For Galton, regression had only this biological meaning,[12][13] but his work was later extended by Udny Yule and Karl Pearson to a more general statistical context.[14][15] In the work of Yule and Pearson, the joint distribution of the response and explanatory variables is assumed to be Gaussian. This assumption was weakened by R.A. Fisher in his works of 1922 and 1925.[16][17][18] Fisher assumed that the conditional distribution of the response variable is Gaussian, but the joint distribution need not be. In this respect, Fisher's assumption is closer to Gauss's formulation of 1821.

In the 1950s and 1960s, economists used electromechanical desk calculators to calculate regressions. Before 1970, it sometimes took up to 24 hours to receive the result from one regression.[19]

Regression methods continue to be an area of active research. In recent decades, new methods have been developed for robust regression, regression involving correlated responses such as time series and growth curves, regression in which the predictor (independent variable) or response variables are curves, images, graphs, or other complex data objects, regression methods accommodating various types of missing data, nonparametric regression, Bayesian methods for regression, regression in which the predictor variables are measured with error, regression with more predictor variables than observations, and causal inference with regression.

Regression models

Regression models involve the following variables:

- The **unknown parameters**, denoted as $\boldsymbol{\beta}$, which may represent a scalar or a vector.
- The **independent variables**, $\mathbf{X}$.
- The **dependent variable**, Y.

In various fields of application, different terminologies are used in place of dependent and independent variables.

A regression model relates Y to a function of $\mathbf{X}$ and $\boldsymbol{\beta}$.

$$Y \approx f(\mathbf{X}, \boldsymbol{\beta})$$

The approximation is usually formalized as $E(Y \mid \mathbf{X}) = f(\mathbf{X}, \boldsymbol{\beta})$. To carry out regression analysis, the form of the function f must be specified. Sometimes the form of this function is based on knowledge about the relationship between Y and $\mathbf{X}$ that does not rely on the data. If no such knowledge is available, a flexible or convenient form for f is chosen.

Assume now that the vector of unknown parameters $\boldsymbol{\beta}$ is of length k. In order to perform a regression analysis the user must provide information about the dependent variable Y:

- If N data points of the form $(Y, \mathbf{X})$ are observed, where $N < k$, most classical approaches to regression analysis cannot be performed: since the system of equations defining the regression model is underdetermined, there are not enough data to recover $\boldsymbol{\beta}$.
- If exactly $N = k$ data points are observed, and the function f is linear, the equations $Y = f(\mathbf{X}, \boldsymbol{\beta})$ can be solved exactly rather than approximately. This reduces to solving a set of N equations with N unknowns (the elements of $\boldsymbol{\beta}$), which has a unique solution as long as the $\mathbf{X}$ are linearly independent. If f is nonlinear, a solution may not exist, or many solutions may exist.
- The most common situation is where $N > k$ data points are observed. In this case, there is enough information in the data to estimate a unique value for $\boldsymbol{\beta}$ that best fits the data in some sense, and the regression model when applied to the data can be viewed as an overdetermined system in $\boldsymbol{\beta}$.

In the last case, the regression analysis provides the tools for:

1. Finding a solution for unknown parameters $\boldsymbol{\beta}$ that will, for example, minimize the distance between the measured and predicted values of the dependent variable Y (also known as method of least squares).
2. Under certain statistical assumptions, the regression analysis uses the surplus of information to provide statistical information about the unknown parameters $\boldsymbol{\beta}$ and predicted values of the dependent variable Y.

Necessary number of independent measurements

Consider a regression model which has three unknown parameters, β_0, β_1, and β_2. Suppose an experimenter performs 10 measurements all at exactly the same value of independent variable vector $\mathbf{X}$ (which contains the independent variables X_1, X_2, and X_3). In this case, regression analysis fails to give a unique set of estimated values for the three unknown parameters; the experimenter did not provide enough information. The best one can do is to estimate the average value and the standard deviation of the dependent variable Y. Similarly, measuring at two different values of $\mathbf{X}$ would give enough data for a regression with two unknowns, but not for three or more unknowns.

If the experimenter had performed measurements at three different values of the independent variable vector $\mathbf{X}$, then regression analysis would provide a unique set of estimates for the three unknown parameters in $\boldsymbol{\beta}$.

In the case of general linear regression, the above statement is equivalent to the requirement that the matrix $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ is invertible.

Statistical assumptions

When the number of measurements, *N*, is larger than the number of unknown parameters, *k*, and the measurement errors ε_i are normally distributed then *the excess of information* contained in (*N* − *k*) measurements is used to make statistical predictions about the unknown parameters. This excess of information is referred to as the degrees of freedom of the regression.

Underlying assumptions

Classical assumptions for regression analysis include:

- The sample is representative of the population for the inference prediction.
- The error is a random variable with a mean of zero conditional on the explanatory variables.
- The independent variables are measured with no error. (Note: If this is not so, modeling may be done instead using errors-in-variables model techniques).
- The independent variables (predictors) are linearly independent, i.e. it is not possible to express any predictor as a linear combination of the others.
- The errors are uncorrelated, that is, the variance–covariance matrix of the errors is diagonal and each non-zero element is the variance of the error.
- The variance of the error is constant across observations (homoscedasticity). If not, weighted least squares or other methods might instead be used.

These are sufficient conditions for the least-squares estimator to possess desirable properties; in particular, these assumptions imply that the parameter estimates will be unbiased, consistent, and efficient in the class of linear unbiased estimators. It is important to note that actual data rarely satisfies the assumptions. That is, the method is used even though the assumptions are not true. Variation from the assumptions can sometimes be used as a measure of how far the model is from being useful. Many of these assumptions may be relaxed in more advanced treatments. Reports of statistical analyses usually include analyses of tests on the sample data and methodology for the fit and usefulness of the model.

Assumptions include the geometrical support of the variables.[20] Independent and dependent variables often refer to values measured at point locations. There may be spatial trends and spatial autocorrelation in the variables that violate statistical assumptions of regression. Geographic weighted regression is one technique to deal with such data.[21] Also, variables may include values aggregated by areas. With aggregated data the modifiable areal unit problem can cause extreme variation in regression parameters.[22] When analyzing data aggregated by political boundaries, postal codes or census areas results may be very distinct with a different choice of units.

Linear regression

In linear regression, the model specification is that the dependent variable, y_i is a linear combination of the *parameters* (but need not be linear in the *independent variables*). For example, in simple linear regression for modeling n data points there is one independent variable: x_i, and two parameters, β_0 and β_1:

straight line: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \dots, n.$

In multiple linear regression, there are several independent variables or functions of independent variables.

Adding a term in x_i^2 to the preceding regression gives:

parabola: $y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i,\ i = 1, \dots, n.$

This is still linear regression; although the expression on the right hand side is quadratic in the independent variable x_i, it is linear in the parameters β_0, β_1 and β_2.

In both cases, ε_i is an error term and the subscript i indexes a particular observation.

Returning our attention to the straight line case: Given a random sample from the population, we estimate the population parameters and obtain the sample linear regression model:

$$\widehat{y_i} = \widehat{\beta}_0 + \widehat{\beta}_1 x_i.$$

The residual, $e_i = y_i - \hat{y}_i$, is the difference between the value of the dependent variable predicted by the model, $\widehat{y_i}$, and the true value of the dependent variable, y_i. One method of estimation is ordinary least squares. This method obtains parameter estimates that minimize the sum of squared residuals, SSE,[23][24] also sometimes denoted RSS:

$$SSE = \sum_{i=1}^{n} e_i^2.$$

Minimization of this function results in a set of normal equations, a set of simultaneous linear equations in the parameters, which are solved to yield the parameter estimators, $\widehat{\beta}_0, \widehat{\beta}_1$.

In the case of simple regression, the formulas for the least squares estimates are

$$\widehat{\beta_1} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} \text{ and } \hat{\beta}_0 = \bar{y} - \widehat{\beta_1}\bar{x}$$

where $\bar{x}$ is the mean (average) of the x values and $\bar{y}$ is the mean of the y values.

Under the assumption that the population error term has a constant variance, the estimate of that variance is given by:

$$\hat{\sigma}_\varepsilon^2 = \frac{SSE}{n-2}.$$

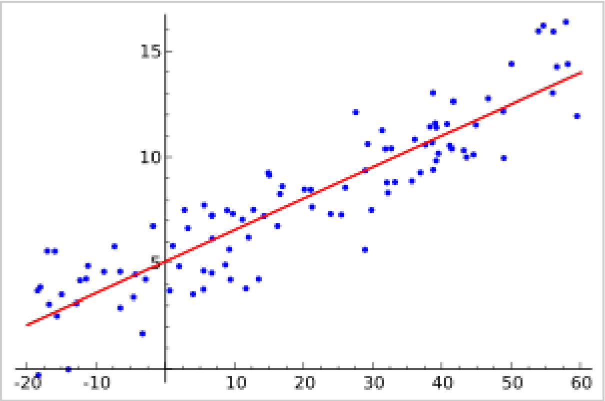

Illustration of linear regression on a data set.

This is called the mean square error (MSE) of the regression. The denominator is the sample size reduced by the number of model parameters estimated from the same data, (*n*-*p*) for *p* regressors or (*n*-*p*-1) if an intercept is used.[25] In this case, *p*=1 so the denominator is *n*-2.

The standard errors of the parameter estimates are given by

$$\hat{\sigma}_{\beta_0} = \hat{\sigma}_\varepsilon \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum(x_i - \bar{x})^2}}$$

$$\hat{\sigma}_{\beta_1} = \hat{\sigma}_\varepsilon \sqrt{\frac{1}{\sum(x_i - \bar{x})^2}}.$$

Under the further assumption that the population error term is normally distributed, the researcher can use these estimated standard errors to create confidence intervals and conduct hypothesis tests about the population parameters.

General linear model

In the more general multiple regression model, there are p independent variables:

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i,$$

where x_{ij} is the i^{th} observation on the j^{th} independent variable. If the first independent variable takes the value 1 for all i, $x_{i1} = 1$, then β_1 is called the regression intercept.

The least squares parameter estimates are obtained from p normal equations. The residual can be written as

$$\varepsilon_i = y_i - \hat{\beta}_1 x_{i1} - \cdots - \hat{\beta}_p x_{ip}.$$

The **normal equations** are

$$\sum_{i=1}^{n} \sum_{k=1}^{p} X_{ij} X_{ik} \hat{\beta}_k = \sum_{i=1}^{n} X_{ij} y_i,\ j = 1, \dots, p.$$

In matrix notation, the normal equations are written as

$$(\mathbf{X}^\top \mathbf{X}) \hat{\boldsymbol{\beta}} = \mathbf{X}^\top \mathbf{Y},$$

where the ij element of X is x_{ij}, the i element of the column vector Y is y_i, and the j element of $\hat{\beta}$ is $\hat{\beta}_j$. Thus X is $n \times p$, Y is $n \times 1$, and $\hat{\beta}$ is $p \times 1$. The solution is

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Y}.$$

Diagnostics

Once a regression model has been constructed, it may be important to confirm the goodness of fit of the model and the statistical significance of the estimated parameters. Commonly used checks of goodness of fit include the R-squared, analyses of the pattern of residuals and hypothesis testing. Statistical significance can be checked by an F-test of the overall fit, followed by t-tests of individual parameters.

Interpretations of these diagnostic tests rest heavily on the model assumptions. Although examination of the residuals can be used to invalidate a model, the results of a t-test or F-test are sometimes more difficult to interpret if the model's assumptions are violated. For example, if the error term does not have a normal distribution, in small samples the estimated parameters will not follow normal distributions and complicate inference. With relatively large samples, however, a central limit theorem can be invoked such that hypothesis testing may proceed using asymptotic approximations.

"Limited dependent" variables

The phrase "limited dependent" is used in econometric statistics for categorical and constrained variables.

The response variable may be non-continuous ("limited" to lie on some subset of the real line). For binary (zero or one) variables, if analysis proceeds with least-squares linear regression, the model is called the linear probability model. Nonlinear models for binary dependent variables include the probit and logit model. The multivariate probit model is a standard method of estimating a joint relationship between several binary dependent variables and some independent variables. For categorical variables with more than two values there is the multinomial logit. For ordinal variables with more than two values, there are the ordered logit and ordered probit models. Censored regression models may be used when the dependent variable is only sometimes observed, and Heckman correction type models may be used when the sample is not randomly selected from the population of interest. An alternative to such procedures is linear regression based on polychoric correlation (or polyserial correlations) between the categorical variables. Such procedures differ in the assumptions made about the distribution of the variables in the population. If the variable is positive with low values and represents the repetition of the occurrence of an event, then count models like the Poisson regression or the negative binomial model may be used instead.

Interpolation and extrapolation

Regression models predict a value of the *Y* variable given known values of the *X* variables. Prediction *within* the range of values in the dataset used for model-fitting is known informally as interpolation. Prediction *outside* this range of the data is known as extrapolation. Performing extrapolation relies strongly on the regression assumptions. The further the extrapolation goes outside the data, the more room there is for the model to fail due to differences between the assumptions and the sample data or the true values.

It is generally advised that when performing extrapolation, one should accompany the estimated value of the dependent variable with a prediction interval that represents the uncertainty. Such intervals tend to expand rapidly as the values of the independent variable(s) moved outside the range covered by the observed data.

For such reasons and others, some tend to say that it might be unwise to undertake extrapolation.[26]

However, this does not cover the full set of modelling errors that may be being made: in particular, the assumption of a particular form for the relation between *Y* and *X*. A properly conducted regression analysis will include an assessment of how well the assumed form is matched by the observed data, but it can only do so within the range of values of the independent variables actually available. This means that any extrapolation is particularly reliant on the assumptions being made about the structural form of the regression relationship. Best-practice advice here is that a linear-in-variables and linear-in-parameters relationship should not be chosen simply for computational convenience, but that all available knowledge should be deployed in constructing a regression model. If this knowledge includes the fact that the dependent variable cannot go outside a certain range of values, this can be made use of in selecting the model – even if the observed dataset has no values particularly near such bounds. The implications of this step of choosing an appropriate functional form for the regression can be great when extrapolation is considered. At a minimum, it can ensure that any extrapolation arising from a fitted model is "realistic" (or in accord with what is known).

Nonlinear regression

When the model function is not linear in the parameters, the sum of squares must be minimized by an iterative procedure. This introduces many complications which are summarized in Differences between linear and non-linear least squares

Power and sample size calculations

There are no generally agreed methods for relating the number of observations versus the number of independent variables in the model. One rule of thumb suggested by Good and Hardin is $N = m^n$, where N is the sample size, n is the number of independent variables and m is the number of observations needed to reach the desired precision if the model had only one independent variable.[27] For example, a researcher is building a linear regression model using a dataset that contains 1000 patients (N). If the researcher decides that five observations are needed to precisely define a straight line (m), then the maximum number of independent variables the model can support is 4, because

$$\frac{\log 1000}{\log 5} = 4.29.$$

Other methods

Although the parameters of a regression model are usually estimated using the method of least squares, other methods which have been used include:

- Bayesian methods, e.g. Bayesian linear regression
- Percentage regression, for situations where reducing *percentage* errors is deemed more appropriate.[28]
- Least absolute deviations, which is more robust in the presence of outliers, leading to quantile regression
- Nonparametric regression, requires a large number of observations and is computationally intensive
- Distance metric learning, which is learned by the search of a meaningful distance metric in a given input space.[29]

Software

All major statistical software packages perform least squares regression analysis and inference. Simple linear regression and multiple regression using least squares can be done in some spreadsheet applications and on some calculators. While many statistical software packages can perform various types of nonparametric and robust regression, these methods are less standardized; different software packages implement different methods, and a method with a given name may be implemented differently in different packages. Specialized regression software has been developed for use in fields such as survey analysis and neuroimaging.

See also

- Curve fitting
- Estimation Theory
- Forecasting
- Fraction of variance unexplained
- Function approximation
- Generalized linear models
- Kriging (a linear least squares estimation algorithm)
- Local regression
- Modifiable areal unit problem
- Multivariate adaptive regression splines
- Multivariate normal distribution
- Pearson product-moment correlation coefficient
- Prediction interval
- Regression validation
- Robust regression
- Segmented regression
- Signal processing
- Stepwise regression
- Trend estimation

References

1. Necessary Condition Analysis (http://www.erim.eur.nl/centres/necessary-condition-analysis/)
2. Armstrong, J. Scott (2012). "Illusions in Regression Analysis". *International Journal of Forecasting (forthcoming)*. **28**

2. Armstrong, J. Scott (2012). "Illusions in Regression Analysis". *International Journal of Forecasting (forthcoming)*. **28** (3): 689. doi:10.1016/j.ijforecast.2012.02.001.
3. David A. Freedman, *Statistical Models: Theory and Practice*, Cambridge University Press (2005)
4. R. Dennis Cook; Sanford Weisberg Criticism and Influence Analysis in Regression (http://links.jstor.org/sici?sici=0081-1750%281982%2913%3C313%3ACAIAIR%3E2.0.CO%3B2-3), *Sociological Methodology*, Vol. 13. (1982), pp. 313–361
5. Christopher M. Bishop (2006). *Pattern Recognition and Machine Learning*. Springer. p. 3. "Cases [...] in which the aim is to assign each input vector to one of a finite number of discrete categories, are called *classification* problems. If the desired output consists of one or more continuous variables, then the task is called *regression*."
6. Waegeman, Willem; De Baets, Bernard; Boullart, Luc (2008). "ROC analysis in ordinal regression learning". *Pattern Recognition Letters*. **29**: 1–9. doi:10.1016/j.patrec.2007.07.019.
7. A.M. Legendre. *Nouvelles méthodes pour la détermination des orbites des comètes* (https://books.google.com/books?id=FRcOAAAAQAAJ), Firmin Didot, Paris, 1805. “Sur la Méthode des moindres quarrés” appears as an appendix.
8. C.F. Gauss. *Theoria Motus Corporum Coelestium in Sectionibus Conicis Solem Ambientum*. (1809)
9. C.F. Gauss. *Theoria combinationis observationum erroribus minimis obnoxiae* (https://books.google.com/books?id=ZQ8OAAAAQAAJ&printsec=frontcover&dq=Theoria+combinationis+observationum+erroribus+minimis+obnoxiae&as_brr=3#v=onepage&q=&f=false). (1821/1823)
10. Mogull, Robert G. (2004). *Second-Semester Applied Statistics*. Kendall/Hunt Publishing Company. p. 59. ISBN 0-7575-1181-3.
11. Galton, Francis (1989). "Kinship and Correlation (reprinted 1989)". *Statistical Science*. Institute of Mathematical Statistics. **4** (2): 80–86. doi:10.1214/ss/1177012581. JSTOR 2245330.
12. Francis Galton. "Typical laws of heredity", Nature 15 (1877), 492–495, 512–514, 532–533. *(Galton uses the term "reversion" in this paper, which discusses the size of peas.)*
13. Francis Galton. Presidential address, Section H, Anthropology. (1885) *(Galton uses the term "regression" in this paper, which discusses the height of humans.)*
14. Yule, G. Udny (1897). "On the Theory of Correlation". *Journal of the Royal Statistical Society*. Blackwell Publishing. **60** (4): 812–54. doi:10.2307/2979746. JSTOR 2979746.
15. Pearson, Karl; Yule, G.U.; Blanchard, Norman; Lee,Alice (1903). "The Law of Ancestral Heredity". *Biometrika*. Biometrika Trust. **2** (2): 211–236. doi:10.1093/biomet/2.2.211. JSTOR 2331683.
16. Fisher, R.A. (1922). "The goodness of fit of regression formulae, and the distribution of regression coefficients". *Journal of the Royal Statistical Society*. Blackwell Publishing. **85** (4): 597–612. doi:10.2307/2341124. JSTOR 2341124.
17. Ronald A. Fisher (1954). *Statistical Methods for Research Workers* (Twelfth ed.). Edinburgh: Oliver and Boyd. ISBN 0-05-002170-2.
18. Aldrich, John (2005). "Fisher and Regression". *Statistical Science*. **20** (4): 401–417. doi:10.1214/088342305000000331. JSTOR 20061201.
19. Rodney Ramcharan. Regressions: Why Are Economists Obessessed with Them? (http://www.imf.org/external/pubs/ft/fandd/2006/03/basics.htm) March 2006. Accessed 2011-12-03.
20. N. Cressie (1996) Change of Support and the Modiable Areal Unit Problem. Geographical Systems 3:159–180.
21. Fotheringham, A. Stewart; Brunsdon, Chris; Charlton, Martin (2002). *Geographically weighted regression: the analysis of spatially varying relationships* (Reprint ed.). Chichester, England: John Wiley. ISBN 978-0-471-49616-8.
22. Fotheringham, AS; Wong, DWS (1 January 1991). "The modifiable areal unit problem in multivariate statistical analysis". *Environment and Planning A*. **23** (7): 1025–1044. doi:10.1068/a231025.
23. M. H. Kutner, C. J. Nachtsheim, and J. Neter (2004), "Applied Linear Regression Models", 4th ed., McGraw-Hill/Irwin, Boston (p. 25)
24. N. Ravishankar and D. K. Dey (2002), "A First Course in Linear Model Theory", Chapman and Hall/CRC, Boca Raton (p. 101)
25. Steel, R.G.D, and Torrie, J. H., *Principles and Procedures of Statistics with Special Reference to the Biological Sciences.*, McGraw Hill, 1960, page 288.
26. Chiang, C.L, (2003) *Statistical methods of analysis*, World Scientific. ISBN 981-238-310-7 - page 274 section 9.7.4 "interpolation vs extrapolation" (https://books.google.com/books?id=BuPNIbaN5v4C&lpg=PA274&dq=regression%20extrapolation&pg=PA274#v=onepage&q=regression%20extrapolation&f=false)
27. Good, P. I.; Hardin, J. W. (2009). *Common Errors in Statistics (And How to Avoid Them)* (3rd ed.). Hoboken, New Jersey: Wiley. p. 211. ISBN 978-0-470-45798-6.
28. Tofallis, C. (2009). "Least Squares Percentage Regression". *Journal of Modern Applied Statistical Methods*. **7**: 526–534. doi:10.2139/ssrn.1406472.
29. YangJing Long (2009). "Human age estimation by metric learning for regression problems" (PDF). *Proc. International Conference on Computer Analysis of Images and Patterns*: 74–82.

Further reading

- William H. Kruskal and Judith M. Tanur, ed. (1978), "Linear Hypotheses," *International Encyclopedia of Statistics*. Free Press, v. 1,

  Evan J. Williams, "I. Regression," pp. 523–41.
  Julian C. Stanley, "II. Analysis of Variance," pp. 541–554.

- Lindley, D.V. (1987). "Regression and correlation analysis," New Palgrave: A Dictionary of Economics, v. 4, pp. 120–23.
- Birkes, David and Dodge, Y., *Alternative Methods of Regression*. ISBN 0-471-56881-3
- Chatfield, C. (1993) "Calculating Interval Forecasts," *Journal of Business and Economic Statistics,* **11**. pp. 121–135.
- Draper, N.R.; Smith, H. (1998). *Applied Regression Analysis* (3rd ed.). John Wiley. ISBN 0-471-17082-8.
- Fox, J. (1997). *Applied Regression Analysis, Linear Models and Related Methods*. Sage
- Hardle, W., *Applied Nonparametric Regression* (1990), ISBN 0-521-42950-1
- Meade, N. and T. Islam (1995) "Prediction Intervals for Growth Curve Forecasts" (http://onlinelibrary.wiley.com/doi/10.1002/for.3980140502/abstract) *Journal of Forecasting,* **14**, pp. 413–430.
- A. Sen, M. Srivastava, *Regression Analysis — Theory, Methods, and Applications*, Springer-Verlag, Berlin, 2011 (4th printing).
- T. Strutz: *Data Fitting and Uncertainty (A practical introduction to weighted least squares and beyond)*. Vieweg+Teubner, ISBN 978-3-8348-1022-9.
- Malakooti, B. (2013). Operations and Production Systems with Multiple Objectives. John Wiley & Sons.

External links

- Hazewinkel, Michiel, ed. (2001), "Regression analysis", *Encyclopedia of Mathematics*, Springer, ISBN 978-1-55608-010-4
- Earliest Uses: Regression (http://jeff560.tripod.com/r.html) – basic history and references
- Regression of Weakly Correlated Data (http://www.vias.org/simulations/simusoft_regrot.html) – how linear regression mistakes can appear when Y-range is much smaller than X-range

Wikimedia Commons has media related to ***Regression analysis***.

Retrieved from "https://en.wikipedia.org/w/index.php?title=Regression_analysis&oldid=745068951"

Categories: Regression analysis | Statistical methods | Estimation theory | Actuarial science

---