REVIEW ARTICLE

Linear Regression Analysis

Part 14 of a Series on Evaluation of Scientific Publications

by Astrid Schneider, Gerhard Hommel, and Maria Blettner

SUMMARY

Background: Regression analysis is an important statistical method for the analysis of medical data. It enables the identification and characterization of relationships among multiple factors. It also enables the identification of prognostically relevant risk factors and the calculation of risk scores for individual prognostication.

Methods: This article is based on selected textbooks of statistics, a selective review of the literature, and our own experience.

Results: After a brief introduction of the uni- and multivariable regression models, illustrative examples are given to explain what the important considerations are before a regression analysis is performed, and how the results should be interpreted. The reader should then be able to judge whether the method has been used correctly and interpret the results appropriately.

Conclusion: The performance and interpretation of linear regression analysis are subject to a variety of pitfalls, which are discussed here in detail. The reader is made aware of common errors of interpretation through practical examples. Both the opportunities for applying linear regression analysis and its limitations are presented.

► Cite this as:
Schneider A, Hommel G, Blettner M: Linear regression analysis—part 14 of a series on evaluation of scientific publications. Dtsch Arztebl Int 2010; 107(44): 776–82. DOI: 10.3238/arztebl.2010.0776

Depatrtment of Medical Biometrics, Epidemiology, and Computer Sciences, Johannes Gutenberg University, Mainz, Germany: Dipl. Math. Schneider, Prof. Dr. rer. nat. Hommel, Prof. Dr. rer. nat. Blettner

The purpose of statistical evaluation of medical data is often to describe relationships between two variables or among several variables. For example, one would like to know not just whether patients have high blood pressure, but also whether the likelihood of having high blood pressure is influenced by factors such as age and weight. The variable to be explained (blood pressure) is called the dependent variable, or, alternatively, the response variable; the variables that explain it (age, weight) are called independent variables or predictor variables. Measures of association provide an initial impression of the extent of statistical dependence between variables. If the dependent and independent variables are continuous, as is the case for blood pressure and weight, then a correlation coefficient can be calculated as a measure of the strength of the relationship between them *(Box 1)*.

Regression analysis is a type of statistical evaluation that enables three things:

- Description: Relationships among the dependent variables and the independent variables can be statistically described by means of regression analysis.
- Estimation: The values of the dependent variables can be estimated from the observed values of the independent variables.
- Prognostication: Risk factors that influence the outcome can be identified, and individual prognoses can be determined.

Regression analysis employs a model that describes the relationships between the dependent variables and the independent variables in a simplified mathematical form. There may be biological reasons to expect a priori that a certain type of mathematical function will best describe such a relationship, or simple assumptions have to be made that this is the case (e.g., that blood pressure rises linearly with age). The best-known types of regression analysis are the following *(Table 1)*:

- Linear regression,
- Logistic regression, and
- Cox regression.

The goal of this article is to introduce the reader to linear regression. The theory is briefly explained, and the interpretation of statistical parameters is illustrated with examples. The methods of regression analysis are comprehensively discussed in many standard textbooks (1–3).

Cox regression will be discussed in a later article in this journal.

Methods

Linear regression is used to study the linear relationship between a dependent variable Y (blood pressure) and one or more independent variables X (age, weight, sex).

The dependent variable Y must be continuous, while the independent variables may be either continuous (age), binary (sex), or categorical (social status). The initial judgment of a possible relationship between two continuous variables should always be made on the basis of a scatter plot (scatter graph). This type of plot will show whether the relationship is linear *(Figure 1)* or nonlinear *(Figure 2)*.

Performing a linear regression makes sense only if the relationship is linear. Other methods must be used to study nonlinear relationships. The variable transformations and other, more complex techniques that can be used for this purpose will not be discussed in this article.

Univariable linear regression

Univariable linear regression studies the linear relationship between the dependent variable Y and a single independent variable X. The linear regression model describes the dependent variable with a straight line that is defined by the equation Y = a + b × X, where a is the y-intersect of the line, and b is its slope. First, the parameters a and b of the regression line are estimated from the values of the dependent variable Y and the independent variable X with the aid of statistical methods. The regression line enables one to predict the value of the dependent variable Y from that of the independent variable X. Thus, for example, after a linear regression has been performed, one would be able to estimate a person's weight (dependent variable) from his or her height (independent variable) *(Figure 3)*.

The slope b of the regression line is called the regression coefficient. It provides a measure of the contribution of the independent variable X toward explaining the dependent variable Y. If the independent variable is continuous (e.g., body height in centimeters), then the regression coefficient represents the change in the dependent variable (body weight in kilograms) per unit of change in the independent variable (body height in centimeters). The proper interpretation of the regression coefficient thus requires attention to the units of measurement. The following example should make this relationship clear:

In a fictitious study, data were obtained from 135 women and men aged 18 to 27. Their height ranged from 1.59 to 1.93 meters. The relationship between height and weight was studied: weight in kilograms was the dependent variable that was to be estimated from the independent variable, height in centimeters. On the basis of the data, the following regression line was determined: Y= –133.18 + 1.16 × X, where X is

BOX 1

Interpretation of the correlation coefficient (r)

Spearman's coefficient:

Describes a monotone relationship

A monotone relationship is one in which the dependent variable either rises or sinks continuously as the independent variable rises.

Pearson's correlation coefficient:

Describes a linear relationship

Interpretation/meaning:

Correlation coefficients provide information about the strength and direction of a relationship between two continuous variables. No distinction between the explaining variable and the variable to be explained is necessary:

- r = ± 1: perfect linear and monotone relationship. The closer r is to 1 or –1, the stronger the relationship.
- r = 0: no linear or monotone relationship
- r < 0: negative, inverse relationship (high values of one variable tend to occur together with low values of the other variable)
- r > 0: positive relationship (high values of one variable tend to occur together with high values of the other variable)

Graphical representation of a linear relationship:

Scatter plot with regression line

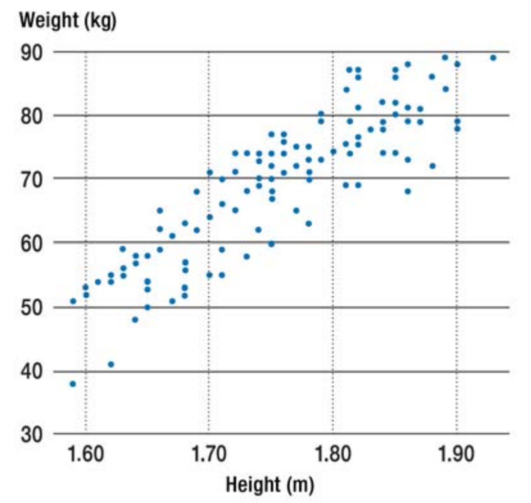
A negative relationship is represented by a falling regression line (regression coefficient b < 0), a positive one by a rising regression line (b > 0).

TABLE 1

Regression models

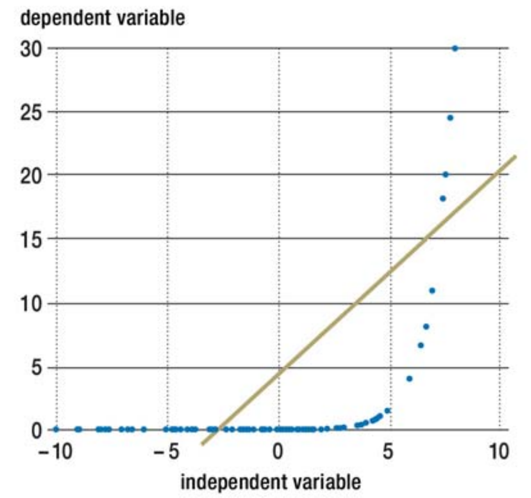
	Application	Dependent variables	Independent variables
Linear regression	Description of a linear relationship	Continuous (weight, blood pressure)	Continuous and/or categorical
Logistic regression	Prediction of the probability of belonging to groups (outcome: yes/no)	Dichotomous (success of treatment: yes/no)	
Proportional hazard regression (Cox regression)	Modeling of survival data	Survival time (time from diagnosis to event)	
Poisson regression	Modeling of counting processes	Counting data: whole numbers representing events in temporal sequence (e.g., the number of times a woman gave birth over a certain period of time)	

FIGURE 1



A scatter plot showing a linear relationship

FIGURE 2



A scatter plot showing an exponential relationship. In this case, it would not be appropriate to compute a coefficient of determination or a regression line

height in centimeters and Y is weight in kilograms. The y-intersect a = –133.18 is the value of the dependent variable when X = 0, but X cannot possibly take on the value 0 in this study (one obviously cannot expect a person of height 0 centimeters to weigh negative 133.18 kilograms). Therefore, interpretation of the constant is often not useful. In general, only values within the range of observations of the independent variables should be used in a linear regression model; prediction of the value of the dependent variable becomes increasingly inaccurate the further one goes outside this range.

The regression coefficient of 1.16 means that, in this model, a person's weight increases by 1.16 kg with each additional centimeter of height. If height had been measured in meters, rather than in centimeters, the regression coefficient b would have been 115.91 instead. The constant a, in contrast, is independent of the unit chosen to express the independent variables. Proper interpretation thus requires that the regression coefficient should be considered together with the units of all of the involved variables. Special attention to this issue is needed when publications from different countries use different units to express the same variables (e.g., feet and inches vs. centimeters, or pounds vs. kilograms).

Figure 3 shows the regression line that represents the linear relationship between height and weight.

For a person whose height is 1.74 m, the predicted weight is 68.50 kg (y = –133.18 + 115.91 × 1.74 m). The data set contains 6 persons whose height is 1.74 m, and their weights vary from 63 to 75 kg.

Linear regression can be used to estimate the weight of any persons whose height lies within the observed range (1.59 m to 1.93 m). The data set need not include any person with this precise height. Mathematically it is possible to estimate the weight of a person whose height is outside the range of values observed in the study. However, such an extrapolation is generally not useful.

If the independent variables are categorical or binary, then the regression coefficient must be interpreted in reference to the numerical encoding of these variables. Binary variables should generally be encoded with two consecutive whole numbers (usually 0/1 or 1/2). In interpreting the regression coefficient, one should recall which category of the independent variable is represented by the higher number (e.g., 2, when the encoding is 1/2). The regression coefficient reflects the change in the dependent variable that corresponds to a change in the independent variable from 1 to 2.

For example, if one studies the relationship between sex and weight, one obtains the regression line Y = 47.64 + 14.93 × X, where X = sex (1 = female, 2 = male). The regression coefficient of 14.93 reflects the fact that men are an average of 14.93 kg heavier than women.

When categorical variables are used, the reference category should be defined first, and all other categories are to be considered in relation to this category.

The coefficient of determination, r^2, is a measure of how well the regression model describes the observed data *(Box 2)*. In univariable regression analysis, r^2 is simply the square of Pearson's correlation coefficient. In the particular fictitious case that is described above, the coefficient of determination for the relationship between height and weight is 0.785. This means that 78.5% of the variance in weight is due to height. The remaining 21.5% is due to individual variation and might be explained by other factors that were not taken into account in the analysis, such as eating habits, exercise, sex, or age.

In formal terms, the null hypothesis, which is the hypothesis that $b = 0$ (no relationship between variables, the regression coefficient is therefore 0), can be tested with a t-test. One can also compute the 95% confidence interval for the regression coefficient (4).

Multivariable linear regression

In many cases, the contribution of a single independent variable does not alone suffice to explain the dependent variable Y. If this is so, one can perform a multivariable linear regression to study the effect of multiple variables on the dependent variable.

In the multivariable regression model, the dependent variable is described as a linear function of the independent variables X_i, as follows: $Y = a + b1 \times X1 + b2 \times X_2 +\ldots+ b_n \times X_n$. The model permits the computation of a regression coefficient b_i for each independent variable X_i *(Box 3)*.

Just as in univariable regression, the coefficient of determination describes the overall relationship between the independent variables X_i (weight, age, body-mass index) and the dependent variable Y (blood pressure). It corresponds to the square of the multiple correlation coefficient, which is the correlation between Y and $b_1 \times X_1 + \ldots + b_n \times X_n$.

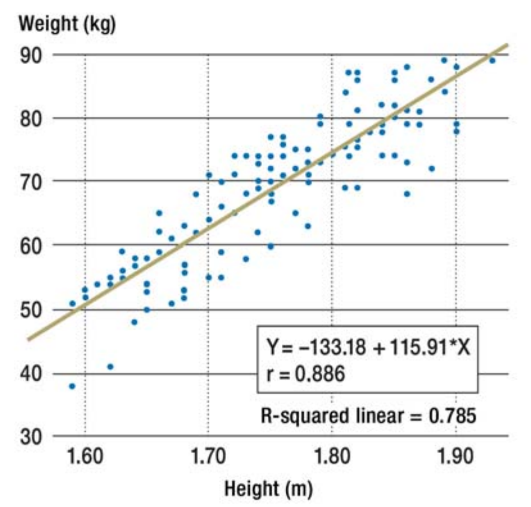
It is better practice, however, to give the corrected coefficient of determination, as discussed in *Box 2*. Each of the coefficients b_i reflects the effect of the corresponding individual independent variable X_i on Y, where the potential influences of the remaining independent variables on X_i have been taken into account, i.e., eliminated by an additional computation. Thus, in a multiple regression analysis with age and sex as independent variables and weight as the dependent variable, the adjusted regression coefficient for sex represents the amount of variation in weight that is due to sex alone, after age has been taken into account. This is done by a computation that adjusts for age, so that the effect of sex is not confounded by a simultaneously operative age effect *(Box 4)*.

In this way, multivariable regression analysis permits the study of multiple independent variables at the same time, with adjustment of their regression coefficients for possible confounding effects between variables.

Multivariable analysis does more than describe a statistical relationship; it also permits individual prognostication and the evaluation of the state of health of a given patient. A linear regression model can be used, for instance, to determine the optimal values for respiratory function tests depending on a person's age, body-mass index (BMI), and sex. Comparing a patient's measured respiratory function with these computed optimal values yields a measure of his or her state of health.

Medical questions often involve the effect of a very large number of factors (independent variables). The goal of statistical analysis is to find out which of these factors truly have an effect on the dependent variable. The art of statistical evaluation lies in finding the variables that best explain the dependent variable.

FIGURE 3



A scatter plot and the corresponding regression line and regression equation for the relationship between the dependent variable body weight (kg) and the independent variable height (m).
r = Pearsons's correlation coefficient
R-squared linear = coefficient of determination

One way to carry out a multivariable regression is to include all potentially relevant independent variables in the model (complete model). The problem with this method is that the number of observations that can practically be made is often less than the model requires. In general, the number of observations should be at least 20 times greater than the number of variables under study.

Moreover, if too many irrelevant variables are included in the model, overadjustment is likely to be the result: that is, some of the irrelevant independent variables will be found to have an apparent effect, purely by chance. The inclusion of irrelevant independent variables in the model will indeed allow a better fit with the data set under study, but, because of random effects, the findings will not generally be applicable outside of this data set (1). The inclusion of irrelevant independent variables also strongly distorts the determination coefficient, so that it no longer provides a useful index of the quality of fit between the model and the data *(Box 2)*.

In the following sections, we will discuss how these problems can be circumvented.

The selection of variables

For the regression model to be robust and to explain Y as well as possible, it should include only independent variables that explain a large portion of the variance in Y. Variable selection can be performed so that only such independent variables are included (1).

Variable selection should be carried out on the basis of medical expert knowledge and a good understanding of biometrics. This is optimally done as a collaborative

BOX 2

Coefficient of determination (R-squared)

Definition:

Let

- n be the number of observations (e.g., subjects in the study)
- $\hat{y}_i$ be the estimated value of the dependent variable for the i^{th} observation, as computed with the regression equation
- y_i be the observed value of the dependent variable for the i^{th} observation
- $\bar{y}$ be the mean of all n observations of the dependent variable

The coefficient of determination is then defined as follows:

$$r^2 = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} = \frac{\text{explained variance}}{\text{overall variance}} = \frac{\text{explained variation}}{\text{overall variation}}$$

→ r^2 is the fraction of the overall variance that is explained. The closer the regression model's estimated values $\hat{y}_i$ lie to the observed values y_i, the nearer the coefficient of determination is to 1 and the more accurate the regression model is.

Meaning: In practice, the coefficient of determination is often taken as a measure of the validity of a regression model or a regression estimate. It reflects the fraction of variation in the Y-values that is explained by the regression line.

Problem: The coefficient of determination can easily be made artificially high by including a large number of independent variables in the model. The more independent variables one includes, the higher the coefficient of determination becomes. This, however, lowers the precision of the estimate (estimation of the regression coefficients b_i).

Solution: Instead of the raw (uncorrected) coefficient of determination, the corrected coefficient of determination should be given: the latter takes the number of explanatory variables in the model into account. Unlike the uncorrected coefficient of determination, the corrected one is high only if the independent variables have a sufficiently large effect.

effort of the physician-researcher and the statistician. There are various methods of selecting variables:

Forward selection

Forward selection is a stepwise procedure that includes variables in the model as long as they make an additional contribution toward explaining Y. This is done iteratively until there are no variables left that make any appreciable contribution to Y.

Backward selection

Backward selection, on the other hand, starts with a model that contains all potentially relevant independent variables. The variable whose removal worsens the prediction of the independent variable of the overall set of independent variables to the least extent is then removed from the model. This procedure is iterated until no dependent variables are left that can be removed without markedly worsening the prediction of the independent variable.

BOX 3

Regression line for a multivariable regression

$Y= a + b_1 \times X_1 + b_2 \times X_2 + ...+ b_n \times X_n$,

where

Y = dependent variable

X_i = independent variables

a = constant (y-intersect)

b_i= regression coefficient of the variable X_i

Example: regression line for a multivariable regression $Y = –120.07 + 100.81 \times X_1 + 0.38 \times X_2 + 3.41 \times X_3$,

where

X_1 = height (meters)

X_2 = age (years)

X_3 = sex (1 = female, 2 = male)

Y = the weight to be estimated (kg)

Stepwise selection

Stepwise selection combines certain aspects of forward and backward selection. Like forward selection, it begins with a null model, adds the single independent variable that makes the greatest contribution toward explaining the dependent variable, and then iterates the process. Additionally, a check is performed after each such step to see whether one of the variables has now become irrelevant because of its relationship to the other variables. If so, this variable is removed.

Block inclusion

There are often variables that should be included in the model in any case—for example, the effect of a certain form of treatment, or independent variables that have already been found to be relevant in prior studies. One way of taking such variables into account is their block inclusion into the model. In this way, one can combine the forced inclusion of some variables with the selective inclusion of further independent variables that turn out to be relevant to the explanation of variation in the dependent variable.

The evaluation of a regression model requires the performance of both forward and backward selection of variables. If these two procedures result in the selection of the same set of variables, then the model can be considered robust. If not, a statistician should be consulted for further advice.

Discussion

The study of relationships between variables and the generation of risk scores are very important elements of medical research. The proper performance of regression analysis requires that a number of important factors should be considered and tested:

1. Causality

Before a regression analysis is performed, the causal relationships among the variables to be considered must be examined from the point of view of their content and/or temporal relationship. The fact that an independent variable turns out to be significant says nothing about causality. This is an especially relevant point with respect to observational studies (5).

2. Planning of sample size

The number of cases needed for a regression analysis depends on the number of independent variables and of their expected effects (strength of relationships). If the sample is too small, only very strong relationships will be demonstrable. The sample size can be planned in the light of the researchers’ expectations regarding the coefficient of determination (r^2) and the regression coefficient (b). Furthermore, at least 20 times as many observations should be made as there are independent variables to be studied; thus, if one wants to study 2 independent variables, one should make at least 40 observations.

BOX 4

Two important terms

- **Confounder** (in non-randomized studies): an independent variable that is associated, not only with the dependent variable, but also with other independent variables. The presence of confounders can distort the effect of the other independent variables. Age and sex are frequent confounders.
- **Adjustment:** a statistical technique to eliminate the influence of one or more confounders on the treatment effect. Example: Suppose that age is a confounding variable in a study of the effect of treatment on a certain dependent variable. Adjustment for age involves a computational procedure to mimic a situation in which the men and women in the data set were of the same age. This computation eliminates the influence of age on the treatment effect.

BOX 5

What special points require attention in the interpretation of a regression analysis?

1. How big is the study sample?
2. Is causality demonstrable or plausible, in view of the content or temporal relationship of the variables?
3. Has there been adjustment for potential confounding effects?
4. Is the inclusion of the independent variables that were used justified, in view of their content?
5. What is the corrected coefficient of determination (R-squared)?
6. Is the study sample homogeneous?
7. In what units were the potentially relevant independent variables reported?
8. Was a selection of the independent variables (potentially relevant independent variables) performed, and, if so, what kind of selection?
9. If a selection of variables was performed, was its result confirmed by a second selection of variables that was performed by a different procedure?
10. Are predictions of the dependent variable made on the basis of extrapolated data?

3. Missing values

Missing values are a common problem in medical data. Whenever the value of either a dependent or an independent variable is missing, this particular observation has to be excluded from the regression analysis. If many values are missing from the dataset, the effective sample size will be appreciably diminished, and the sample may then turn out to be too small to yield significant findings, despite seemingly adequate advance planning. If this happens, real relationships can be overlooked, and the study findings may not be generally applicable. Moreover, selection effects can be expected in such cases. There are a number of ways to deal with the problem of missing values (6).

4. The data sample

A further important point to be considered is the composition of the study population. If there are subpopulations within it that behave differently with respect to the independent variables in question, then a real effect (or the lack of an effect) may be masked from the analysis and remain undetected. Suppose, for instance, that one wishes to study the effect of sex on weight, in a study population consisting half of children under age 8 and half of adults. Linear regression analysis over the entire population reveals an effect of sex on weight. If, however, a subgroup analysis is performed in which children and adults are considered separately, an effect of sex on weight is seen only in adults, and not in children. Subgroup analysis should only be performed if the subgroups have been predefined, and the questions already formulated, before the data analysis begins; furthermore, multiple testing should be taken into account (7, 8).

5. The selection of variables

If multiple independent variables are considered in a multivariable regression, some of these may turn out to be interdependent. An independent variable that would be found to have a strong effect in a univariable regression model might not turn out to have any appreciable effect in a multivariable regression with variable selection. This will happen if this particular variable itself depends so strongly on the other independent variables that it makes no additional contribution toward explaining the dependent variable. For related reasons, when the independent variables are mutually dependent, different independent variables might end up being included in the model depending on the particular technique that is used for variable selection.

Overview

Linear regression is an important tool for statistical analysis. Its broad spectrum of uses includes relationship description, estimation, and prognostication. The technique has many applications, but it also has prerequisites and limitations that must always be considered in the interpretation of findings *(Box 5).*

Conflict of interest statement
The authors declare that they have no conflict of interest as defined by the guidelines of the International Committee of Medical Journal Editors.

Manuscript submitted on 11 May 2010, revised version accepted on 14 July 2010.

Translated from the original German by Ethan Taub, MD

REFERENCES

1. Fahrmeir L, Kneib T, Lang S: Regression – Modelle, Methoden und Anwendungen. 2nd edition. Berlin, Heidelberg: Springer 2009
2. Bortz J: Statistik für Human-und Sozialwissenschaftler. 6th edition. Heidelberg: Springer 2004.
3. Selvin S: Epidemiologic Analysis. Oxford University Press 2001.
4. Bender R, Lange S: Was ist ein Konfidenzintervall? Dtsch Med Wschr 2001; 126: T41.
5. Sir Bradford Hill A: The environment and disease: Association or Causation? Proc R Soc Med 1965; 58: 295–300.
6. Carpenter JR, Kenward MG: Missing Data in Randomised Controlled Trials: A practical guide. Birmingham, Alabama: National Institute for Health Research; 2008. http://www.pcpoh.bham.ac.uk/publichealth/methodology/projects/RM03_JH17_MK.shtml. Publication RM03/JH17/MK.
7. EMEA: Poiints to consider on multiplicity issues in clinical trials; www.emea.europa.eu/pdfs/human/ewp/090899en.pdf
8. Horn M, Vollandt R: Multiple Tests und Auswahlverfahren. Stuttgart: Gustav Fischer Verlag 1995.

Corresponding author
Prof. Dr. rer. nat. Maria Blettner
Department of Medical Biometrics, Epidemiology, and Computer Sciences
Johannes Gutenberg University
Obere Zahlbacher Str. 69
55131 Mainz
Germany