INTERNATIONAL JOURNAL OF UNANI AND INTEGRATIVE MEDICINE

E-ISSN: 2616-4558
P-ISSN: 2616-454X
IJUIM 2018; 2(1): 01-05
Received: 01-11-2017
Accepted: 02-12-2017

Anirban Goswami
Investigator Statistics, Regional Research Institute of Unani Medicine, Patna, Under CCRUM, Ministry of Ayush, India

Correspondence
Anirban Goswami
Investigator Statistics, Regional Research Institute of Unani Medicine, Patna, Under CCRUM, Ministry of Ayush, India

Utilization of regression analysis in clinical research

Anirban Goswami

Abstract
Regression analysis is important statistical tools for assessing the mathematical relationships among the variables in clinical research. Clinical research is increasingly based on the empirical studies and the results of these are usually presented and analyzed with statistical methods. Therefore discuss frequently used regression analysis for different type of data set to find the casual relationship among the variables under different assumptions. The regression analysis is a technique, that is appropriate to understand the association between one independent (or predictor) variable and one continuous dependent (or outcome) variables. Regression analysis may provide a researcher with an equation for a graph so that he can make predictions about research data. It is therefore an advantage for any physician or researcher he/she is familiar with the frequently used regression model, as this is the only way he or she can evaluate the regression model in scientific publications and thus correctly interpret their findings.

Keywords: Clinical research, regression analysis, model

Introduction
Clinical research aims to advance medical knowledge by studying people, either through direct interaction or through the collection and analysis of treatment, blood, tissues, or other samples. Simply put, it involves human participants and helps translate basic research (done in labs) into new treatments and information to benefit patients. Clinical research are conducted to collect and recorded data on each subject, such as the patient's demographic characteristics, disease related risk factors, medical history, biochemical markers, pathological history, medical therapies, and outcome or endpoint data at different time points. This data may be continuous or discrete or dichotomous. Understanding that the types/assumptions of data are more important as they determine which method of data analysis is to be use and how to report the results [1]. For the assessment of the safety, efficacy, and / or the mechanism of action of an investigational medicinal product, or new drug or device that is in development. Clinical Research involves a cycle of events starting from pre-clinical animal testing to various stages of drug development before they are introduced in the market for mass consumption.

In clinical research, patient's and investigator's responses to treatments can be documented according to the occurrence of some meaningful and well-defined event such as death, infection, or cure of a certain disease, any serious adverse events, biochemical and pathological findings and to determine the casual relationship of them through regression analysis. Regression analysis determines the relationship of an independent variable (such as bone mineral density) on a dependent variable (such as ageing) with the statistical assumption that all other variables remain fixed [2]. The regression analysis might be useful in clinical data analysis include the modeling of blood pressure response (y) on the dose of a new antihypertensive drug (x), cholesterol level (y) on patient's age (x), pain relief (y) on time after dosing with an anti-inflammatory treatment (x), or degree of wound healing (y) on the baseline surface area of a burn wound (x). The calculation of the relationship between dependent variable and independents variables may be linear or non-linear [3].

The differential use of regression for statistical summarization of relationships between the outcome and predictors, for statistical adjustment of the observed relationship of the outcome to one or more key predictors, and for statistical prediction of the outcome from a set of candidate predictors. In clinical research that evaluate the relative contribution of various factors to a single binary outcome, such as the presence or absence of death or disease, most often employ the method of logistic regression [4]. Logistical regression models are often used to adjust for covariates when the primary outcome of the study is event rate, and the Cox proportional hazards regression model for trials with time to an event as the endpoint.

Similarly for the longitudinal data when the response variable is non-normal are not nearly as comprehensive as for normally distributed, in this satiation nonparametric or semi-parametric regression is suitable to find the relationship between dependent variable and independent variable [5].

Regression models have some basic assumption and validate these assumptions after fitted regression model. The regression validation is the process of deciding whether the numerical results quantifying hypothesized relationships between variables, obtained from regression analysis, are acceptable as descriptions of the data. This validity of conclusions generally drawn from model-based analyses relies on the assumption that the model is correctly specified, that is, the assumption that the statistical model accurately represents the true data generating distribution. Moreover, in clinical research and studies involving biological systems in general, due to the complexity of relationships between variables, regression models may fail to accurately represent define the true relationships between these variables. So, it may not even be possible to detect when a model is incorrectly specified, since for the sample sizes available in many applications, diagnostics of model fit have good power.

Regression Analysis used in Clinical Research

Multiple Linear Regression

Multiple linear regression is the most common form of linear regression analysis. Multiple linear regression analysis used to assess the linear association between two or more independent variables and a single continuous dependent variable. Generally it is used to explain the linear relationship between one continuous dependent variable and two or more independent variables. Linear regression gives the straight line equation that best describes it and enables to the prediction of one variable from the other. There are four major assumptions for multiple linear regression as (1) regression residuals must be normally distributed, (2) a linear relationship is assumed between the dependent variable and the independent variables, (3) the residuals are homoscedastic and approximately rectangular-shaped and (4) absence of multicollinearity is assumed in the model, meaning that the independent variables are not too highly correlated. For example in clinical research, suppose 100 women attending an antenatal clinic took part in a study to identify variables associated with birth weight of the child with the eventual aim to predicting women ‘at risk’ of having a low birth weight baby. As per result of multiple regression, shows that birth weight was significantly related to age of mother, height of mother, parity, period of gestation and family income [6].

Nonlinear Regression

In clinical research, linear regression or multiple linear regressions are not appropriate for all situations. There are many situations where the response variable and the independent variables are related through a known nonlinear function. Nonlinear regression models are generally assumed to be parametric, where the model is described as a nonlinear equation. The model can be univariate (single response variable) or multivariate (multiple response variables). For example, polynomial or growth regression model is used to model curvature in research data by using higher-ordered values of the predictors. However, the nonlinear regression model was just a linear combination of higher-ordered predictors. The nonlinear regression is used to estimate the parameters in a nonlinear model without performing hypothesis tests. The usual assumption about the normality of the residuals is not needed. Instead, the main assumption needed is that the data may be well represented by the model. For example, polynomial regression model used to identify and study circadian rhythms in ten mildly hypertensive patients both untreated and after single-dose treatment with different categories of antihypertensive agents of the ambulatory blood pressure monitoring (ABPM) data [7].

Linear Mixed Effects Regression

It is an extension of the general linear regression model, in which factors and covariates are assumed to have a linear relationship to the dependent variable. The mixed-effects model consists of two parts, fixed effects and random effects. Fixed-effects terms are usually the conventional linear regression part, and the random effects are associated with individual experimental units drawn at random from a population. The random effects have prior distributions whereas fixed effects do not. In mixed effect regression models add at least one random variable to a linear or generalized linear model. The random variables of a mixed model add the assumption that observations within a level, the random variable groups, are correlated. It is designed to address this correlation and do not cause a violation of the independence of observations assumption from the underlying model, e.g. linear or generalized linear. The assumption is relaxed to observations are independent of the other observations except where there is correlation specified by the random variable groups. For example in clinical research, it is used to estimate the effect of factors such as treatment while taking into account that observation on the same subjects are correlated. Where subject were randomized to receive either a new chewable tablet formulation of carbamezepine consider as treatment A and standard carbamezepine tablet consider as treatment B. After a four week washout period, subjects crossed over to the other treatment and blood sample were collected over 2 days following a 200 mg does to derive concentration time curves for each subjects and each formulation [8].

Multiple Logistic Regression

Logistic regression analysis is similar to linear regression analysis except that the outcome is dichotomous (e.g., success/failure or yes/no or died/lived). Logistic regression analysis refers to the regression application with one dichotomous outcome or dependent variable and one independent variable; multiple logistic regression analysis applies when there is a single dichotomous outcome and more than one independent variable. Multiple logistic regression assumes that requires the (1) dependent variable to be binary and ordinal logistic regression requires the dependent variable to be ordinal, (2)observations to be independent of each other. In other words, the observations should not come from repeated measurements or matched data, (3) no multicollinearity among the independent variables and (4) linearity of independent variables and log odds. although this analysis does not require the dependent and independent variables to be related linearly, it requires that the independent variables are linearly related to the log odds. For example in clinical research, to identify the

significant lifestyle risk factors of hypertension. In this study, lifestyle factors such as BMI, tobacco chewing, alcoholism and laziness were determined as highly associated with the hypertension (p-value <0.05) where as smoking have found non-significant (p-value>0.05) result [9].

Multinomial Logistic Regression

Multinomial logistic regression is a simple extension of binary logistic regression that allows for more than two categories of the dependent or outcome variable. It is used to model nominal outcome variables, in which the log odds of the outcomes are modeled as a linear combination of the predictor variables. Multinomial logistic regression does have assumptions, such as the assumption of independence among the dependent variable choices, outcome follows a categorical distribution which is linked to the covariates via a link function as in ordinary logistic regression, independence of observational units and linear relation between covariates and (link-transformed) expectation of the outcome [4]. For example in clinical research, to analyzed the effect on LDL cholesterol, model the ending cholesterol level as a function of treatment using the beginning level as a covariate, where subjects to a control and treatment groups assigned by the randomly and treatment subjects ate cereal containing psyllium daily [4].

Poisson Regression

It is a form of regression analysis used to model the count data. It assumes the response variable has a poisson distribution [10] and assumes the logarithm of its expected value can be modeled by a linear combination of unknown parameters. Generally, poisson regression is used to predict a dependent variable that consists of "count data" with one or more independent variables. Basic five assumption of this regression as dependent variable should be consists of count data, distribution of counts (conditional on the model) follow a Poisson distribution, one or more independent variables, which can be measured on a continuous, ordinal or nominal/dichotomous scale, each observation is independent of the other observations and mean and variance of the model are identical. In a clinical research, patients with superficial bladder tumours which were removed. Three treatment, placebo, pyridoxine pills or instillation of a chemotherapeutic agent, thiotepa were assigned by randomly. At subsequent monthly follow-up visits any tumous noticed were removed and treatment was continued and to determine the effect of treatment on the frequency of tumor recurrence, defined as the presence of at least one tumor at a follow-up visit.

Negative Binomial Regression

It is a type of generalized linear model in which the dependent variable is a count of the number of times an event occurs i.e. the dependent variable follows the negative binomial distribution. Negative binomial regression is a form of generalized Poisson regression which loosens the restrictive assumption that the variance is equal to the mean made by the Poisson model. The traditional negative binomial (NB) regression model, commonly known as NB with order of parameter 2, is based on the Poisson-gamma mixture distribution [11,4]. Negative binomial regression model assume the conditional means are not equal to the conditional variances. This inequality is captured by estimating a dispersion parameter) that is held constant in a Poisson model. Thus, the Poisson model is actually nested in the negative binomial model. For example, negative binomial regression to analyze hypoglycemic events only compares the counts of hypoglycemic events during a specified period and comparing hypoglycemic event rates between treatment groups at different time periods to understand the trend over time [12].

Geometric Regression

The special case of negative binomial regression is Geometric regression in which the dispersion parameter is set to one [13]. It is similar as multiple regression except that the dependent variable is an observed count that follows by the geometric distribution. The geometric distribution is a probability distribution where 1st occurrence of any event can be modelled. In the generalized linear model for the geometric distribution can contribute significantly to exhibit many important facts associated with the 1st occurrence of any event. Geometric regression is a generalization of Poisson regression which loosens the restrictive assumption that the variance is equal to the mean made by the Poisson model. Basic assumption of this regression as dependent variable should be consists of count data, distribution of counts (conditional on the model) follow a geometric distribution, one or more independent variables, which can be measured on a continuous, ordinal or nominal/dichotomous scale, each observation is independent of the other observations. Geometric regression used to determine the association of biochemical parameters and personal characteristics with the Nuzj(concoction) appearance day in cases of Lymphatic Filariasis.

Loglinear Regression

The loglinear regression is one of the specialized cases of generalized linear models for poisson distributed data. Log-linear regression assumed that the discrete variables to be nominal, but these models can be adjusted to deal with ordinal and matched data. Log-linear models are more general than logit regression models, but some log-linear models have direct correspondence to logit regression models. It is an extension of the two-way contingency table where the conditional relationship between two or more discrete, categorical variables is analyzed by taking the natural logarithm of the cell frequencies within a contingency table [14].

Quantile regression

Linear regression approach based on Least Squares method focuses on the conditional mean of the response variable for the given set of independent variables. There are certain attractive features of mean regression which make it popular and practically useful. First, the least squares method is easy for computation and interpretation. Second, classical regression assumes Gaussian distribution on noise and homoscedasticity in variance which helps to develop attractive statistical theory for the estimator of model parameters. When these assumptions are violated, linear regression estimates are not valid. Quantile regression method overcomes the drawbacks of linear regression and can be applied when the data is skewed and equal variance assumptions are violated. Quantile regression does not rely on the assumptions of normality or homoscedastic errors to model the conditional percentiles. For eample in clinical

research, to find out the probability of mothers having low birth weight babies, linear regression model based on the average birth weight as a function of different predictors will leave the lower birth weight categories which is not correct as it losses the important information due to the skewness in the distribution. Quantile regression used in these scenarios as it includes both the lower, middle and upper quantiles in predicting the probabilities of lower birth weight [15].

Ridge Regression

In a linear regression, least squares estimation isn't defined at all when the number of predictors exceeds the number of observations; It doesn't differentiate "important" from "less-important" predictors in a model, so it includes all of them. This leads to over fitting a model and failure to find unique solutions. Least squares also has issues dealing with multicollinearity in data. Ridge Regression is a technique for analyzing multiple regression data that suffer from multicollinearity. When multicollinearity occurs, least squares estimates are unbiased, but their variances are large so they may be far from the true value. By adding a degree of bias to the regression estimates, ridge regression reduces the standard errors [16]. For example, generally genetic markers are obtained using SNP arrays, next-generation sequencing technologies and imputation. However, SNPs typed can be highly correlated due to linkage disequilibrium among them, and standard multiple regression techniques fail with these data sets due to their high dimensionality and correlation structure. In this case, Ridge regression is one such technique which does not perform variable selection, instead estimating a regression coefficient for each predictor variable. So that, it is therefore desirable to obtain an estimate of the significance of each ridge regression coefficient [17].

Nonparametric Regression

The usual assumption of nonparametric regression relaxes of linearity and enables you to uncover relationships between the independent variables and the dependent variable. In traditional parametric regression models, the functional form of the model is specified before the model is fit to data, and the object is to estimate the parameters of the model. In the nonparametric regression, in contrast, the object is to estimate the regression function directly without specifying its form explicitly [14]. In the nonparametric regression technique that combines both regression splines and model selection methods. It does not assume parametric model forms and does not require specification of knot values for constructing regression spline terms. Instead, it constructs spline basis functions in an adaptive way by automatically selecting appropriate knot values for different variables and obtains reduced models by applying model selection techniques. The generalized additive model procedure provides powerful tools for nonparametric regression and smoothing [18]. Nonparametric regression relaxes the usual assumption of linearity and enables you to uncover relationships between the independent variables and the dependent variable that might otherwise be missed. For example of respiratory disorder, to estimate of regression coefficients, standard errors and standardized statistics for the treatment, centre, sex, age and baseline respiratory status effects, for comparison, and find the association of them [5].

Semi-parametric Regression

Semi-parametric regression analysis is a combination of parametric and nonparametric regression technique. It is used in situations where the fully nonparametric regression may not perform well or when the researcher wants to use a parametric model but the functional form with respect to a subset of the independent variables or the density of the errors is not known. Semi-parametric models contain a parametric component, they rely on parametric assumptions and may be misspecified and inconsistent, just like a fully parametric model. It can be formulated using a linear mixed model. Generally semi-paramedic regression used in longitudinal data when the response is non-normal are not as comprehensive as for normally distribute. For example, it is used to identify the effects of smoking, age and geographical location with lung cancer levels [19, 5].

Cox Regression

It is a popular mathematical model used for analyzing the survival data on one or more predictors or independent variables. It provides an estimate of the hazard ratio and its confidence interval. Cox regression is considered a 'semi-parametric' procedure because the baseline hazard function, does not required any specification of the probability distribution of the survival times [20]. Two main assumption of cox regression, first and foremost is the issue of non-informative censoring: to satisfy this assumption, the design of the underlying study must ensure that the mechanisms giving rise to censoring of individual subjects are not related to the probability of an event occurring and secondly in the Cox model is that of proportional hazards: in a regression typesetting this means that the survival curves for two strata must have hazard functions that are proportional over time (i.e. constant relative hazard) [21]. Cox regression used to examine the effect of a single or multiple independent variables with treatment groups in a randomized controlled clinical trial to compare two treatment for prostate cancer. Where two groups took as 1 mg of diethylstilbestrol or placebo daily by month, and their survival time was recorded in months. The variable Treatment indicates which treatment was received, Time is the time from the beginning of the trials to death or end of the trial, status indicate that the subject died or whether survived is right censored and Age, Haem, Size, Gleason were observed [22].

Conclusion

The regression analysis or model is used to analyze or find the casual relationship of the different type of clinical data in different situations and nature of the data set. Different regression model used to different satiations and every model has some assumptions. Before using the regression model in clinical research we need to check the assumptions, type of the study or objective of an experiment and data structure. Most of these models play a very important role to getting appropriate and desired result in clinical research, to make the decision on the objectives. Researchers / Physicians are helpful to used regression analysis or model to determine results from experiments, clinical research of medicine and symptoms of diseases. The use of regression model in medicine provides generalizations for the public to better understand their risks for certain diseases, links between certain behaviors of diseases, effectiveness of drug(s) and to significant finding of experimental objectives.

References

1. Wang D, Bakhai A. Clinical Trials-A Practical Guide to Design, Analysis, and Reporting. Remedica Publishing, USA, 2006.
2. Draper NR, Smith H. Applied regression analysis. Wiley-Interscience, 1998.
3. Faraway JJ. Linear Models with R.CHAPMAN & HALL/CRC, 2009.
4. Agresti A. Categorical Data Analysis.2nd Ed. John Wiley & Sons, New Jersey, 2002.
5. Davis CS. Semi-parametric and non-parametric methods for the analysis of repeated measurements with applications to clinical trials. Stat Med. 1991; 10(12):1959-80.
6. Sarmukaddam SB. Clinical Biostatistics. New Age International Publishers, 1st Edt. 2014, 172.
7. Zwinderman AH, Cleophas TA, Cleophas TJ, van der Wall EE. Polynomial analysis of ambulatory blood pressure measurements. Neth Heart J. 2001; 9(2):68-74.
8. Everitt B, Rabe-Hesketh S. Analysis of Medical Data Using S-PLUS. Springr, New York, 2001.
9. Sultana S, Urooj S, Goswami A. Association of Lifestyle Risk Factors with Incidence of Hypertension. International Journal of Current Research 2017; 9(12): 62725-62729.
10. Kianifard F, Gallo PP. Poisson regression analysis in clinical research. J Biopharm Stat. 1995; 5(1):115-29.
11. Venables WN, Ripley BD. Modern Applied Statistics with S, Fourth Edition. New York: Springer, 2002.
12. Wang M, Luo J, Fu H, Qu Y. Piecewise Negative Binomial Regression in Analyzing Hypoglycemic Events with Missing Observations. J Biomet Biostat 2014; 5(3):195.
13. Zeileis A, Kleiber C, Jackman S. Regression Models for Count Data in R. Journal of Statistical Software. 2008; 27(8):1-25. (https://www.jstatsoft.org/article/view/v027i08)).
14. McCullagh P, Nelder J A. Generalized Linear Models. 2nd Edt. Chapman and Hall/CRC, 1989.
15. EDITOR IJSMI. Application of Quantile Regression in Clinical Research: An Overview with the Help of R and SAS Statistical Package. International Journal of Statistics and Medical Informatics. 2017, 2(1).
16. Montgomery DC, Peck EA, Vining GG. Introduction to Linear Regression Analysis. 3rd Edt., John Wiley, New York, 2001.
17. Cule Erika, Vineis P, Iorio MD. Significance testing in ridge regression for genetic data. BMC Bioinformatics 2011; 12:372.
18. Wood SN. Generalized Additive Models: An introduction with R. Chapman and Hall/CRC, 2006.
19. Ruppert D, Wand MP, Carroll RJ. Semiparametric Regression. Cambridge University Press, 2003.
20. Kleinbaum DG. Survival Analysis: A Self-Learning Text. Springer, New York, 1996.
21. Cox DR, Oakes D. Analysis of survival data. London, England: Chapman and Hall, 2001.
22. Everitt B, Rabe-Hesketh S. Analysis of Medical Data Using S-PLUS. Springr, New York, 2001.