DOI: http://dx.doi.org/10.18203/2394-6040.ijcmph20162539

Review Article

Regression technique: model to predict causal relationship between variables

Gladius Jennifer Hirudayaraj[1]*, Bagavan Das[2]

[1]Department of Community Medicine, Karpaga Vinayaga Institute of Medical Sciences and Research Centre, Chinna Kolambakkam, Palayanoor, Kancheepuram District, Tamilnadu, India
[2]School of Public Health, SRM University, Chennai, Tamilnadu, India

Received: 11 June 2016
Accepted: 02 July 2016

***Correspondence:**
Dr. Gladius Jennifer Hirudayaraj,
E-mail: gladiusjennifer@gmail.com

ABSTRACT

Medical research is aim to quantify the disease magnitude and establish association between the study variables. Regression is the technique, which will not only find the correlation but also predict how much are the strength of relationship between variables. This article aims, to discuss various types of regression techniques such as Linear Regression, Multiple Regression, Logistic Regression, Meta regression and spatial regression and Regression Imputation with assumptions and models. This article is to sensitize doctors and post-graduate medical students about this useful analytical technique.

Keywords: Regression, MLR, Logistic Regression, Meta regression, Spatial regression

INTRODUCTION

Medical Research are mostly designed to find the relationship between the variables or to measure co variation between them and some researcher would like to establish the cause and effect relationship in terms of one variable on other. So, the variables are classified as Independent or explored variables, Dependent or explained variables and usually denoted as X and Y. For example age, sex, education, occupation are independent variables where bmi, obesity, depression are dependent variables. To find the relationship or to measure the co variation, correlation is appropriate technique. Correlation is used to assess the relationship between two or more continuous variables. It ranges from -1 to 1, usually denoted as r and always expressed in percentages. The scatter diagrams explain the relationship in graphical like positive and perfect positive correlations, no correlation, negative and perfect negative correlation. Karl Pearson, the British biometrician developed correlation; it is based three assumptions i.e., the variables X and Y should be normally distributed, linear and homogeneous. To find the strength or function of the relationship between these variables, Regression Models are appropriate.[1]

Sir Francis Galton developed technique called Regression[1], to predict or estimate the value of the response variable from known values of one or more independent variables.

There are many types of Regression Models available in medical research. They are Simple Linear Regression (SLR), Multiple Linear Regression (MLR), Logistic Regression (LR), Nominal / Ordinal Logistic Regression. There are also some other Regression Techniques which are used in the field of Imputation, Systematic Reviews and Spatial Patterns. They are Regression Imputation, Meta Regression and Spatial Regression, Bayesian Regression, Autoregressive Models etc.

Simple linear regression

Simple Linear Regression is used when there is single continuous independent variable and single continuous response variable. It is done to know the tendency of one variable to change with other.

Assumptions

- The observations are independent
- Variables are normally distributed
- Relationship between variables is linear

The linear Regression Model: Y = a + b X,

Where Y is Dependent Variable, X is Independent variable, b is Slope and its calculated as, b = cov (x,y) / var(x)

a = Y - b * X; where, a = intercept for which X = 0 the slope b will explain that for each unit change in x, y increase.

Example: A pediatric registrar has measured the pulmonary anatomical dead space (in ml) and height (in cm) of 15 children.[2]

HT: 110 116 124 129 131 138 142 150 153 155 156 159 164 168 174

Space: 44 31 43 45 56 79 57 56 58 92 78 64 88 112 101

$\sum xy = 150605$, $\sigma x = 19.36$ $\bar{y} = 66.93$, $\bar{x} = 144.6$,

b=1.03, a = -82.4 ➔ regression line y = -82.4 + 1.03 x

Multiple linear regression model

It attempts to determine a formula that can describe how elements in a vector of variables respond simultaneously to changes in others. The dependent variable (Y) is continuous variable which follows normal distribution. Independent variables (x_1, x_2….x_n) are both continuous and categorical variables. Like linear regression a - is intercept and b_1, b_2…b_n are slopes for corresponding independent variables. Multiple linear Model is given by, $Y = a + b_1x_1 + b_2x_2 + \ldots + b_{nxn}$. For example, to predict cholesterol level (Y) with predictors age (x_1), sex (x_2) and weight (x_3)

Here Y is continuous dependent variable follows normal distribution, the independent variables either continuous (age, weight) or categorical (sex). The regression line will be Cholesterol level (Y) = a + b_1 (Age) + b_2 (Sex) + b_3 (Weight).

Logistic regression model

When the response variable is binary (Nominal/categorical), π (x) the probability that Y equals one for a given value of X. the usual model is the logistic regression model, a non-linear model with a sigmodal shape. The change in the probability that Y equals one for a given change in X is greatest for values of X near the middle of its range, rather than for values at the extremes. The error term is not normally distributed followed binomial distribution. It can be extended to multiple predictor variables. There are various types of logistic regressions: Multiple logistic regression models, Ordinal Logistic Regression etc.

The logit model derived as follows:

$P\ (Y = 1/X) = e^{(a + bx)} / 1 + e^{(a + bx)}$

$1 - P\ (Y = 1/X) = 1 – \{e^{(a + bx)} / 1 + e^{(a + bx)}\}$

$\text{Odds} = P / 1 - P = e^{(a + bx)}$

Logit = Log Odds = Ln (P / 1 – P) = a + bx

Where, a – Intercept, b – slope, b = Ln (Odds Ratio)

Example

Graft rejection status and marrow cell dose data for 68 aplastic anemia patients[2]

Graft rejection Marrow cell dose (108 cells/kg)

Here b = 1.83 and a =-1.95

Table 1: Logistic regression model.

	< 3.0	>3.0	Total
Yes	17	4	21
No	19	28	47
Total	36	32	68

Regression imputation

Imputation is defined as follows: to substitute missing values with certain fabricated values. It is a process of explicitly or implicitly substituting data for incomplete or inconsistent items in survey records. Regression imputation is defined as replacing the missing values with predicted values from the estimated regression model. This model permits more auxiliary variables to be used and it employs a linear additive model. There are two types of techniques to replace missing values. They are simple linear regression model and random regression model. Simple linear imputation estimates the missing value by fitting a regression line. Random regression imputation method imputes values directly from the estimated regression line. Random residual errors can be added to the regression prediction to provide dispersion about the regression line. The only difference between the simple and random imputation method is error term.

In simple linear regression imputation for model Y on X ➔ $Y = \hat{\beta}_0 + \hat{\beta}_1 X + \varepsilon$,the missing values is replaced by the equation $Y^* = \hat{\beta}_0 + \hat{\beta}_1 X$ when the independent variable has no missing values. If the independent value has the missing value, it is calculated by $\hat{X} = \bar{X}$. Then imputed value is calculated by using the formula $Y^* = \hat{\beta}_0 + \hat{\beta}_1 \bar{\bar{X}}$.[3]

Random Regression Imputation imputes values directly from the estimated regression line. Regression residual errors can be added to the regression prediction to provide dispersion about the regression line. This method imputes missing value $Y = Y^* = \bar{Y} + \bar{\varepsilon}$ where $\bar{\varepsilon} = Y - \hat{\beta}_0 + \hat{\beta}_1 X$. if independent variable X is missing $\hat{X} = \bar{X}$. The only difference between simple linear regression imputation and random regression imputation is ε.[3] Random is better than simple. SLR imputation underestimates the outcome values. Whereas random regression imputation, estimates are more or less similar to before imputation.

Meta regression model

Multiple regressions are used in primary studies to assess the relationship between subject-level covariates and an outcome. We can also use meta-regression in meta-analysis to assess the relationship between study level covariates and effect size. Meta-regression may be performed under the fixed-effect or the random effects model, but in most cases the latter is appropriate. In addition to testing the impact of covariates for statistical significance, it is also important to quantify the magnitude of their relationship with effect size.

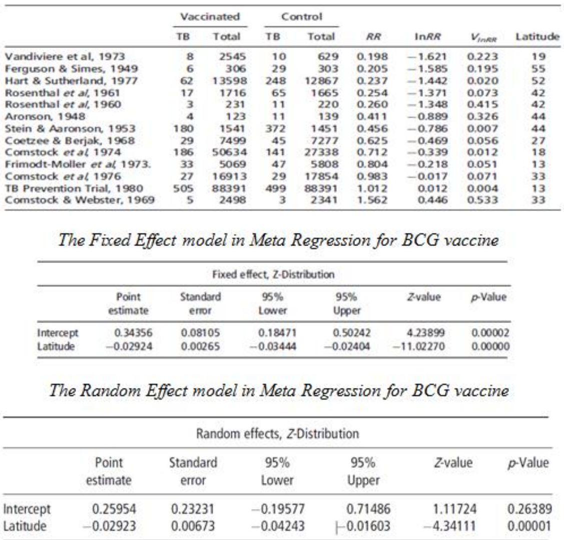
True effect size is the effect size in the underlying population and is the effect size that we would observe if the study had infinitely large sample size. Observed effect size is the effect size that is actually observed.[4] The Fixed effect model assumes that there is one true effect size which underlies all the studies in the analysis, and that all differences in observed effects are due to sampling error. We denote the true effect size by theta. If each study had an infinite sample size, then the sampling error would be zero and the observed effect for each study would be the same as the true effect. In plots, the observed effects would exactly coincide with true effects. The Random effect model assumes that the true treatment effects in the individual studies may be different from each other. There is no single number to estimate in the meta analysis but a distribution of numbers. It assumes that these different true effects are normally distributed. It estimates the mean and standard deviation of the different effects. However, in special cases multiple meta-regressions that include both the fixed and random-effects model can be used. If we use a fixed-effect model within subgroups and also across subgroups, the analysis is called a fixed-effects analysis. If we use a random-effects model within subgroups and a fixed-effect model across subgroups (the approach that we generally advocate), the model is called a mixed-effects model. We have the further possibility of assuming random effects both within and across subgroups; such a model is called a random-effects (or fully random-effects) model.

Meta regression model

Example

Colditz et al, and Berkey et al showed how meta-regression could be used in an attempt to explain some of the variance in treatment effects of BCG vaccine for tuberculosis.

The data collected for Meta-analysis given below (Figure 1).[4]



	Vaccinated		Control					
	TB	Total	TB	Total	*RR*	ln*RR*	V_{lnRR}	Latitude
Vandiviere et al, 1973	8	2545	10	629	0.198	−1.621	0.223	19
Ferguson & Simes, 1949	6	306	29	303	0.205	−1.585	0.195	55
Hart & Sutherland, 1977	62	13598	248	12867	0.237	−1.442	0.020	52
Rosenthal *et al*, 1961	17	1716	65	1665	0.254	−1.371	0.073	42
Rosenthal *et al*, 1960	3	231	11	220	0.260	−1.348	0.415	42
Aronson, 1948	4	123	11	139	0.411	−0.889	0.326	44
Stein & Aaronson, 1953	180	1541	372	1451	0.456	−0.786	0.007	44
Coetzee & Berjak, 1968	29	7499	45	7277	0.625	−0.469	0.056	27
Comstock *et al*, 1974	186	50634	141	27338	0.712	−0.339	0.012	18
Frimodt-Moller *et al*, 1973.	33	5069	47	5808	0.804	−0.218	0.051	13
Comstock *et al*, 1976	27	16913	29	17854	0.983	−0.017	0.071	33
TB Prevention Trial, 1980	505	88391	499	88391	1.012	0.012	0.004	13
Comstock & Webster, 1969	5	2498	3	2341	1.562	0.446	0.533	33

The Fixed Effect model in Meta Regression for BCG vaccine

Fixed effect, Z-Distribution						
	Point estimate	Standard error	95% Lower	95% Upper	*Z*-value	*p*-Value
Intercept	0.34356	0.08105	0.18471	0.50242	4.23899	0.00002
Latitude	−0.02924	0.00265	−0.03444	−0.02404	−11.02270	0.00000

The Random Effect model in Meta Regression for BCG vaccine

Random effects, *Z*-Distribution						
	Point estimate	Standard error	95% Lower	95% Upper	*Z*-value	*p*-Value
Intercept	0.25954	0.23231	−0.19577	0.71486	1.11724	0.26389
Latitude	−0.02923	0.00673	−0.04243	−0.01603	−4.34111	0.00001

Figure 1: Meta regression model.

Spatial regression model

Linear regression for spatial variables is a technique that can be used to model the broad-scale (first-order) spatial trend of a dataset where the outcome variable is continuously distributed. The following are the key assumptions behind this type of regression analysis.[6]

- For all values of μ there must be a corresponding value of *X*

- The value of μ at any point is not affected by the value of at any other point (independence).

- The relationship between μ and *X* should be approximately linear (linearity).

- The variance of μ about the estimated regression line is equal for all values of X.

- The residuals μ are normally distributed with a mean of zero (normality).

The linear regression model for spatially auto correlated variables

The term $X\boldsymbol{\beta}$ is often referred to as the *mean structure*, *large-scale variation*, or *trend*, to distinguish this variation from the variation in the residual vector that defines the *small-scale variation* in the data and determines the *stochastic dependence structure* or *residual autocorrelation* in the data. The residual process adjusts the model for any residual spatial variation remaining in the data after accounting for covariate effects. Two methods which are often used to estimate the parameters of a semi variogram model, and equivalently, the parameters in E($\boldsymbol{\theta}$), under the general linear model are: iteratively reweighted generalized least squares (IRWGLS) and maximum likelihood (ML).[6,7]

Spatial auto regression model

A structure mirroring the time-series literature makes wide use of *autoregressive* models wherein we regress the current observation on observed values of all or, more commonly, a subset of other observations. In time series, the "other" observations occur in the recent past; in the spatial setting, they occur nearby. Just as the term autocorrelation reflects self-correlation, the term autoregressive reflects self-regression. Through such regressions, we incorporate spatial similarity by treating observations of the outcome variable at other locations as additional covariates in the model with associated parameters defining spatial association, rather than building an explicit parametric model of the covariance function of the error terms. The autoregressive model induces a particular covariance structure for the joint distribution of variables, but we typically do not fit the covariance directly. Instead, the autoregressive model itself defines this covariance for us. Simultaneous and conditional autoregressive models are the types of spatial auto regressive models. Spatial autoregressive models incorporate spatial dependence through the use of spatial proximity measures. These measures provide a flexible modeling tool for geographical data. An alternative approach is provided by geo statistics, where the semi variogram based on inter-centroid distances quantifies the spatial autocorrelation in the data. These two different approaches, and the differences in the results obtained with them, lead us to wonder to what extent the choice of spatial dependence measure has on conclusions. Conceptually, given the differences in the spatial proximity measures, the impact of the choice of spatial proximity measure is likely to be great, as it will result in very different neighborhood structures and thus allow much different interactions among the data.[6,7]

- It is better to use any reasonable method for modeling spatial autocorrelation than to assume that the data are independent.

- The choice of spatial dependence model can greatly affect both the estimates from regression models and their standard errors. Thus, exploratory spatial analysis is important since it can provide valuable information concerning the spatial relationships among the data that can be used to choose a spatial dependence model substantiated by the data. It is also wise to compare results from several different spatial dependence models and to try to understand their differences.

- Accounting for population heterogeneity in geographically aggregated data is very important. Many of the tools for inference with spatial data assume second-order stationary and thus may give misleading conclusions when applied to data based on units with differing population sizes or with different spatial support.

- The principle of parsimony is paramount; we should choose the simplest model that both adequately explains the variation in our data and facilitates an interpretation that is consistent with our knowledge about the people, places, and processes we are studying.[6,7]

Example: Rebeca Ramis Prieto et al used Bayesian regression to predict variables of municipal mortality due to hematological neoplasias and Muhammad NA et al used Spatial Ordinal Logistic Regression and The Principal Component to Predict Poverty Status of Districts in Java Island.[5]

CONCLUSION

Regression models are unique technique to predict variables. Apart from these above mentioned models, there are also many models available in regression which can be used in multivariate analysis such as Poisson Regression, Bayesian regression Generalized Estimation models and so on. This article may be the eye opener for early stage medical researchers and students.

Funding: No funding sources
Conflict of interest: None declared
Ethical approval: Not required

REFERENCES

1. Gupta SC, Kapoor VK. Fundamentals of Mathematical Statistics. 9th ed, Sultan Chand and Sons.
2. Murthy S, NandaKumar BS, Shivaraj NS, Gautham MS, Pruthvish S. Epidemiological research methods and biostatistics lecture notes on epidemiology and biostatistics.

3. Jinn JH. The effect of different imputation methods on analytical statistics of simple linear regression. Interstat. 2002.
4. Borens MT, Hedges LV, Higgins JPT, Rothstein HR. Introduction to Meta-Analysis. 2009 John Wiley and Sons, Ltd; ISBN:978-0-470-05724-7.
5. Modeling Spatial Ordinal Logistic Regression and The Principal Component to Predict Poverty Status of Districts. In: Java Island Muhammad NA, Tuti Purwaningsih SS. International Journal of Statistics and Applications. 2013;3(1):1-8
6. Lance AW, Carol AG. Applied Spatial Statistics for Public Health Data. John Wiley and Sons inc.
7. Gaetan C, Guyon X. Spatial Statistics and Modeling. Springer publication.

Cite this article as: Gladius JH, Das B. Regression technique: model to predict causal relationship between variables. Int J Community Med Public Health 2016:3:1981-5.