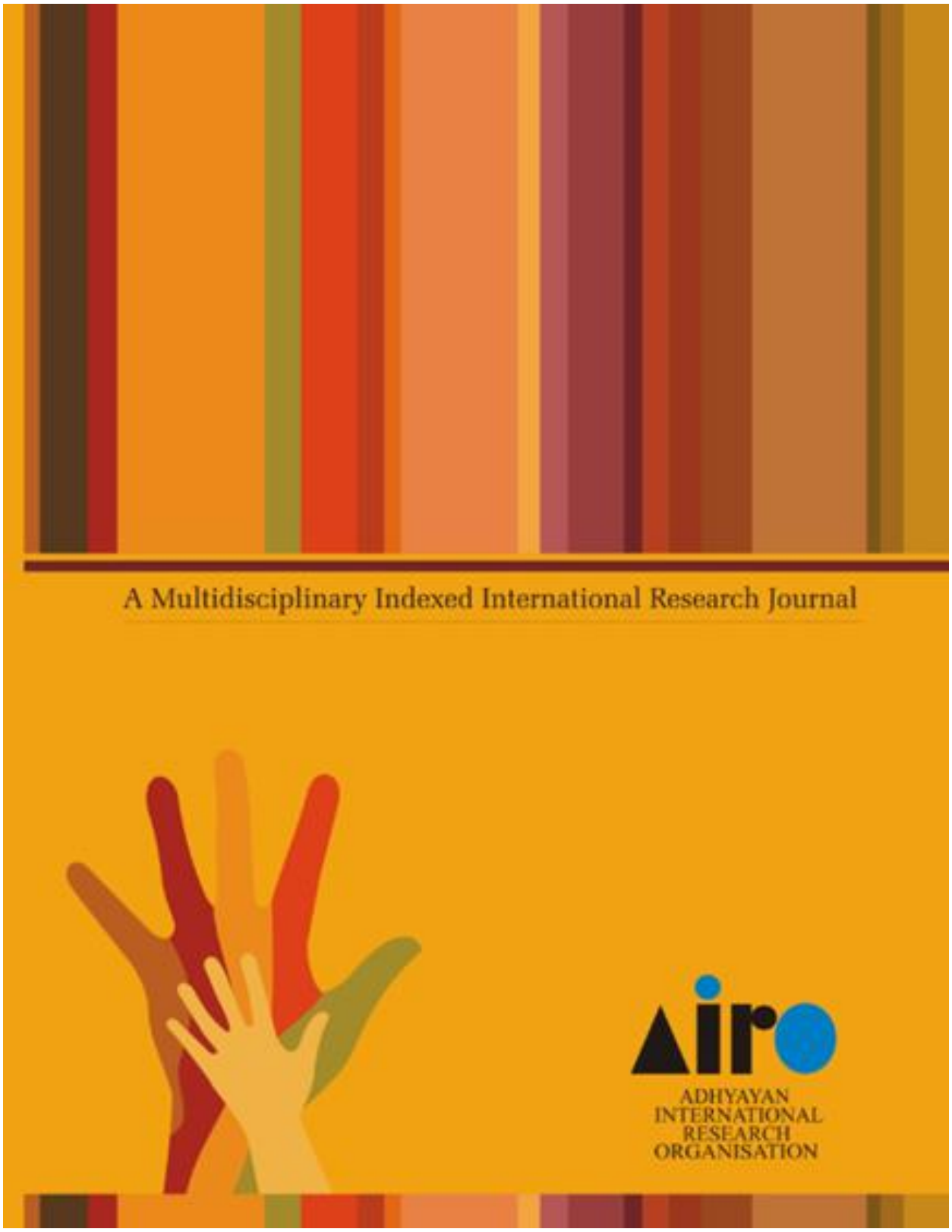
Airo International Research Journal
Volume XIII, ISSN: 2320-3714
September, 2017
Impact Factor 0.75 to 3.19
UGC Approval Number 63012



Airo International Research Journal
Volume XIII, ISSN: 2320-3714
September, 2017
Impact Factor 0.75 to 3.19
UGC Approval Number 63012

A NOVEL APPROACH FOR REGRESSION ANALYSIS WITH DIFFERENT MODELS

K. Chinnaiah

Research Scholar, Department of CSE, Sunrise University, Alwar, Rajasthan

Dr. Akash Saxena

Supervisor, Department of CSE, Sunrise University, Alwar, Rajasthan

Abstract:

Regression is a data mining function that predicts a number. Age, weight, distance, temperature, income, or sales could all be predicted using regression techniques. For example, a regression model could be used to predict children's height, given their age, weight, and other factors.[A]

A regression task begins with a data set in which the target values are known. For example, a regression model that predicts children's height could be developed based on observed data for many children over a period of time. The data might track age, height, weight, developmental milestones, family history, and so on. Height would be the target, the other attributes would be the predictors, and the data for each child would constitute a case.[A]

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'). 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.[B]

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.[B]

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.[B]***

Introduction:

Regression modeling has many applications in trend analysis, business planning, marketing, financial forecasting, time series prediction, biomedical and drug response modeling, and environmental modeling.[A]

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.[B]

Regression Models

Regression models involve the following variables:

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

Y=f(X,B)

The approximation is usually formalized as **E(Y/X)=f(X,B).** 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 relation between **Y**&**X** that does not rely on data. If no such knowledge is available, a flexible or convenient form of f is chosen.

Assume now that the vector of unknown parameters B 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,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 B.

- If exactly N=k data points are observed, and the function f is linear, the equations Y=f(X,B) can be solved exactly rather than approximately. This reduces to solving a set of N equations with N unknowns(elements of B),which has a unique solution as long as the 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 B 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 B.

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

1. Finding a solution for unknown parameters B 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 B and predicted values of the dependent variable Y.

Regression Work:

You do not need to understand the mathematics used in regression analysis to develop quality regression models for data mining. However, it is helpful to understand a few basic concepts.

The goal of regression analysis is to determine the values of parameters for a function that cause the function to best fit a set of data observations that you provide. The following equation expresses these relationships in symbols. It shows that regression is the process of estimating the value of a continuous target (*y*) as a function (*F*) of one or more predictors ($\mathbf{x}_1$, $\mathbf{x}_2$, ..., $\mathbf{x}_n$), a set of parameters (θ_1 , θ_2 , ..., θ_n), and a measure of error (*e*).

$$\boldsymbol{y = F(x,\theta) + e}$$

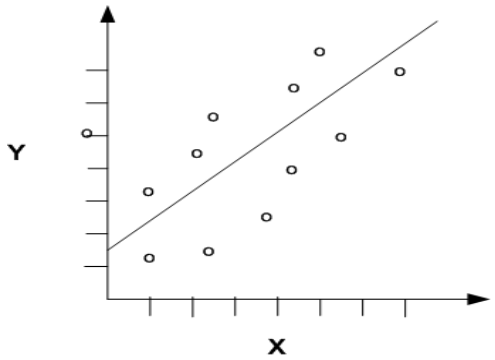
The process of training a regression model involves finding the best parameter values for the function that minimize a measure of the error, for example, the sum of squared errors.

TYPES OF REGRESSION

There are different families of regression functions and different ways of measuring the error.[A]

Linear Regression

The simplest form of regression to visualize is linear regression with a single predictor. A linear regression technique can be used if the relationship between *x* and *y* can be approximated with a straight line, as shown in Figure below.[A]



Linear regression is a common Statistical Data Analysis technique. It is used to determine the extent to which there is a linear relationship between a dependent variable and one or more independent variables. There are two types of linear regression, simple linear regression and multiple linear regression.[C]

In **simple linear regression** a single independent variable is used to predict the value of a dependent variable.

Or we can say that, **Simple regression** pertains to **one** dependent variable (y) and **one** independent variable (x):

$$y=f(x)$$

In **multiple linear regression** two or more independent variables are used to predict the value of a dependent variable. The difference between the two is the number of independent variables. In both cases there is only a single dependent variable.[C]or we can say that, **Multiple regression (multivariable regression)** pertains to **one** dependent variable and **multiple** independent variables:

$$y=f(x1,x2,...,xn)$$

Multivariate regression is a technique that estimates a single regression model with more than one outcome variable. When there is more than one predictor variable in a multivariate regression model, the model is a multivariate multiple regression.[I]

In Multivariate regression pertains to **multiple** dependent variables and **multiple** independent variables: y1,y2,...,ym=f(x1,x2,...,xn). You may encounter problems where both the dependent and independent variables are arranged as matrices of variables (e.g. y11,y12,...y11,y12,... and x11,x12,...x11,x12,...), so the expression may be written as Y=f(X), where capital letters indicate matrices.[H]

In a linear regression scenario with a single predictor ($y = \theta_2 x + \theta_1$), the regression parameters (also called coefficients) are:

The **slope** of the line (θ_2) — the angle between a data point and the regression line and

The ***y* intercept** (θ_1) — the point where *x* crosses the *y* axis (*x* = 0) [A]

Linear regression has many practical uses. Most applications fall into one of the following two broad categories:

- If the goal is prediction, or forecasting, or error reduction, linear regression can be used to fit a predictive model to an observed data set of *y* and *X* values. After developing such a model, if an additional value of *X* is then given without its accompanying value of *y*, the fitted model can be used to make a prediction of the value of *y*.

- Given a variable *y* and a number of variables X_1, ..., X_p that may be related to *y*, linear regression analysis can be applied to quantify the strength of the relationship between *y* and the X_j, to assess which X_j may have no relationship with *y* at all, and to identify which subsets of the X_j contain redundant information about *y*.[B]

In linear regression, the model specification is that the dependent variable, yi is a linear combination of the parameters. For example, in simple linear regression for modeling n data points there is one independent variable: xi, and two parameters, B0 and B1:

Straight line: yi = B0 + B1xi + ei, i=1,2,…….n.

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

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

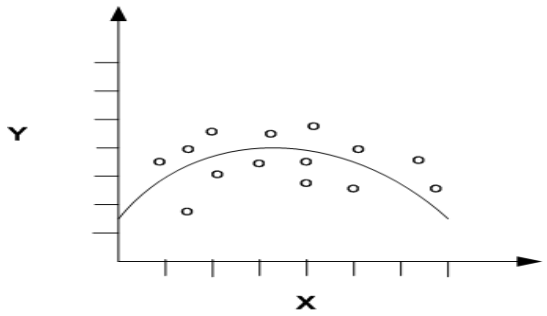
Parabola: yi = B0 + B1xi + B2xi^2 + ei, i = 1,2,…….n

This is still linear regression; although the expression on the right hand side is quadratic in the independent variable xi, it is linear in the parameters B0, B1, B2.

In both cases, ei is an error term and the subscript I indexes a particular observation.[B]

Non- Linear Regression

Often the relationship between *x* and *y* cannot be approximated with a straight line. In this case, a nonlinear regression technique may be used. Alternatively, the data could be preprocessed to make the relationship linear.[A]



Nonlinear regression is a form of regression analysis in which observational data are modeled by a function which is a nonlinear combination of the model parameters and depends on one or more independent variables. The data are fitted by a method of successive approximations.

Nonlinear regression is a statistical technique that helps describe nonlinear relationships in experimental data. Nonlinear regression models are generally assumed to be parametric, where the model is described as a nonlinear equation. Typically machine learning methods are used for non-parametric nonlinear regression.

Parametric nonlinear regression models the dependent variable (also called the response) as a function of a combination of nonlinear parameters and one or more independent variables (called predictors). The model can be univariate (single response variable) or multivariate (multiple response variables).

The parameters can take the form of an exponential, trigonometric, power, or any other nonlinear function. To determine the nonlinear parameter estimates, an iterative algorithm is typically used.[D]

$$\mathbf{Y=f(X,B) + e}$$

where,

B represents nonlinear parameter estimates to be computed and

e represents the error terms.

All of the models we have discussed thus far have been linear in the parameters (i.e., linear in the beta's). For example, polynomial regression was used to model curvature in our data by using higher-ordered values of the predictors. However, the final regression model was just a linear combination of higher-ordered predictors.[E]

Now we are interested in studying the **nonlinear regression** model:

$$\mathbf{Y=f(X,\beta)+\epsilon,}$$

where **X** is a vector of *p* predictors, β is a vector of *k* parameters, f(·) is some known regression function, and ϵ is an error term whose distribution may or may not be normal. Notice that we no longer necessarily have the dimension of the parameter vector simply one greater than the number of predictors. Some examples of nonlinear regression models are:

yi

yi

yi=eβ0+β1xi

1+eβ0+β1xi

=β0+β1xi1+β2eβ3xi+ϵi

yi=β0+(0.4−β0)e−β1(xi−5)+ϵi.

However, there are some nonlinear models which are actually called **intrinsically**

linear because they can be made linear in the parameters by a simple transformation. For example:

$$\mathbf{Y= \beta 0X}$$

$$\mathbf{\beta 1+X}$$

This function is nonlinear because it cannot be expressed as a linear combination of the two *{\displaystyle \beta }B*s.

can be rewritten as

$$1 = 1 + B1\ 1$$

Y B0 B0 X

= θ0+θ11

X,

which is linear in the transformed parameters θ0 and θ1. In such cases, transforming a model to its linear form often provides better inference procedures and confidence intervals, but one must be cognizant of the effects that the transformation has on the distribution of the errors.[E]

Popular algorithms for fitting a nonlinear regression include:

- Gauss-Newton algorithm
- Levenberg-Marquardt algorithm[F]
- Newton's Method

- **Newton's method**, a classical method based on a gradient approach but which can be computationally challenging and heavily dependent on good starting values.
- The **Gauss-Newton algorithm**, a modification of Newton's method that gives a good approximation of the solution that Newton's method should have arrived at but which is not guaranteed to converge.
- The **Levenberg-Marquardt method**, which can take care of computational difficulties arising with the other methods but can require a tedious search for the optimal value of a tuning parameter.[E]

Difference between Linear and Non Linear Regression

Linear Regression

A model is linear when each term is either a constant or the product of a parameter and a predictor variable. A linear equation is constructed by adding the results for each term. This constrains the equation to just one basic form:[G]

Response = constant + parameter * predictor + ... + parameter * predictor

$Y = b_o + b_1X_1 + b_2X_2 + ... + b_kX_k$

In statistics, a regression equation (or function) is linear when it is linear in the

parameters. While the equation must be linear in the parameters, you can transform the predictor variables in ways that produce curvature. For instance, you can include a squared variable to produce a U-shaped curve.

$$Y = b_o + b_1X_1 + b_2X_1^2$$

This model is still linear in the parameters *even though the predictor variable is squared*. You can also use log and inverse functional forms that are linear in the parameters to produce different types of curves.

Non Linear Regression

While a linear equation has one basic form, nonlinear equations can take many different forms. The easiest way to determine whether an equation is nonlinear is to focus on the term "nonlinear" itself. Literally, it's not linear. If the equation doesn't meet the criteria above for a linear equation, it's nonlinear.

That covers many different forms, which is why nonlinear regression provides the most flexible curve-fitting functionality.Thetas represent the parameters and X represents the predictor in the nonlinear functions. Unlike linear regression, these functions can have more than one parameter per predictor variable.[G]

Research methodology:

Logistic Regression

Logistic regression models a relationship between predictor variables and a categorical response variable. For example, we could use logistic regression to model the relationship between various measurements of a manufactured specimen (such as dimensions and chemical composition) to predict if a crack greater than 10 mils will occur (a binary variable: either yes or no). Logistic regression helps us estimate a probability of falling into a certain level of the categorical response given a set of predictors. We can choose from three types of logistic regression, depending on the nature of the categorical response variable:[E]

Binary Logistic Regression:

Used when the response is binary (i.e., it has two possible outcomes). The cracking example given above would utilize binary logistic regression. Other examples of binary responses could include passing or failing a test, responding yes or no on a survey, and having high or low blood pressure.

Nominal Logistic Regression:

Used when there are three or more categories with no natural ordering to the levels. Examples of nominal responses could include departments at a business (e.g., marketing, sales, HR), type of search engine used (e.g., Google, Yahoo!, MSN), and color (black, red, blue, orange).

Ordinal Logistic Regression:

Used when there are three or more categories with a natural ordering to the levels, but the ranking of the levels do not necessarily mean the intervals between them are equal. Examples of ordinal responses could be how students rate the effectiveness of a college course on a scale of 1-5, levels of flavors for hot wings, and medical condition (e.g., good, stable, serious, critical).[E]

The multiple **binary logistic regression model** is the following:

$$\pi=\frac{\exp(\beta 0+\beta 1X1+\ldots+\beta p-1Xp-1)}{1+\exp(\beta 0+\beta 1X1+\ldots+\beta p-1Xp-1)}$$

$$=\frac{\exp(X\beta)}{1+\exp(X\beta)}$$

$$=\frac{1}{1+\exp(-X\beta)}$$

where here π denotes a probability and *not* the irrational number 3.14....

- π is the probability that an observation is in a specified category of the binary *Y* variable, generally called the "success probability."
- Notice that the model describes the *probability of an event* happening as a function of *X* variables. For instance, it might provide estimates of the probability that an older person has heart disease.
- With the logistic model, estimates of π from equations like the one above will always be between 0 and 1. The reasons are:
  - The numerator $\exp(\beta 0+\beta 1X1+\ldots+\beta p-1Xp-1)$ must be positive, because it is a power of a positive value (*e*).
  - The denominator of the model is (1 + numerator), so the answer will always be less than 1.
  - With one *X* variable, the theoretical model for π has an elongated "S" shape (or sigmoidal shape) with asymptotes at 0 and 1, although in sample estimates we may not see this "S" shape if the range of the *X* variable is limited.

Conclusion:

Stepwise regression procedures are used in data mining, but are controversial. Several points of criticism have been made.

- The tests themselves are biased, since they are based on the same data.[15][16] Wilkinson and Dallal (1981)[17] computed percentage points of the multiple correlation coefficient by simulation and showed that a final regression obtained by forward selection, said

by the F-procedure to be significant at 0.1%, was in fact only significant at 5%.

- When estimating the degrees of freedom, the number of the candidate independent variables from the best fit selected may be smaller than the total number of final model variables, causing the fit to appear better than it is when adjusting the r^2 value for the number of degrees of freedom. It is important to consider how many degrees of freedom have been used in the entire model, not just count the number of independent variables in the resulting fit.[18]

- Models that are created may be over-simplifications of the real models of the data.[19]

Such criticisms, based upon limitations of the relationship between a model and procedure and data set used to fit it, are usually addressed by verifying the model on an independent data set, as in the PRESS procedure.

Critics regard the procedure as a paradigmatic example of data dredging, intense computation often being an inadequate substitute for subject area expertise. Additionally, the results of stepwise regression are often used incorrectly without adjusting them for the occurrence of model selection. Especially the practice of fitting the final selected model as if no model selection had taken place and reporting of estimates and confidence intervals as if least-squares theory were valid for them, has been described as a scandal.[7] Widespread incorrect usage and the availability of alternatives such as ensemble learning, leaving all variables in the model, or using expert judgement to identify relevant variables have led to calls to totally avoid stepwise model selection.[5]

References:

[1] Efroymson,M. A. (1960) "Multiple regression analysis," Mathematical Methods for Digital Computers, Ralston A. and Wilf,H. S., (eds.), Wiley, New York.

[2] Hocking, R. R. (1976) "The Analysis and Selection of Variables in Linear Regression," *Biometrics, 32.*

[3] Draper, N. and Smith, H. (1981) *Applied Regression Analysis, 2d Edition*, New York: John Wiley & Sons, Inc.

[4] SAS Institute Inc. (1989) *SAS/STAT User's Guide, Version 6, Fourth Edition, Volume 2*,Cary, NC: SAS Institute Inc.

[5] Flom, P. L. and Cassell, D. L. (2007) "Stopping stepwise: Why stepwise and similar selection methods are bad, and what you should use," NESUG 2007.

[6] Harrell, F. E. (2001) "Regression modeling strategies: With applications to linear models, logistic regression, and survival analysis," Springer-Verlag, New York.

[7] Chatfield, C. (1995) "Model uncertainty, data mining and statistical inference," J. R. Statist. Soc. A 158, Part 3, pp. 419–466.

[8] Efron, B. and Tibshirani, R. J. (1998) "An introduction to the bootstrap," Chapman & Hall/CRC

[9] Box–Behnken designs from a handbook on engineering statistics at NIST

[10] Efroymson, MA (1960) "Multiple regression analysis." In Ralston, A. and Wilf, HS, editors, *Mathematical Methods for Digital Computers.* Wiley.