Journal of Statistical and Econometric Methods, vol.1, no.2, 2012, 1-8
ISSN: 2241-0384 (print), 2241-0376 (online)
Scienpress Ltd, 2012

Modeling for Regressing Variables

Norzima Zulkifli[1], Shahryar Sorooshian[2,*] and Alireza Anvari[3]

Abstract

Regression modeling is an approach to modeling the relationship between endogenous and one or more explanatory exogenous variables. This paper reviews the bias of this modeling, as well as history, purpose, assumptions, and steps. Then the paper follows with Classification of regression analysis in to 8 classes. Specifically, linear and non linear regression is discussed by details. We end the paper with application of regression modeling. We find this paper as a statistical modeling guide for scholars.

Mathematics Subject Classification: 62J05
Keywords: statistical modeling, linear regression, non linear regression, functional relationship

[1] Department of Mechanical and Manufacturing Engineering, University Putra Malaysia, Malaysia.
[2] Business School, Taylor's University, Malaysia.
[*] Corresponding author, e-mail: sorooshian@gmail.com
[3] Department of Industrial Management, Gachsaran Branch, Islamic Azad University, Gachsaran, Iran.

Article Info: *Received* : December 16, 2011. *Revised* : April 25, 2012
Published online : July 30, 2012

1 Introduction

Regression as a term was first coined by a biologist Francis Galton in the 19th century when he attempted to describe a biological phenomenon that the heights of descendants of tall ancestors tend to regress down towards the normal average (Draper, 1998 and Fox, 1997). The earliest form of regression analysis was the method of least squares 1st published in 1805 by Legendre and in 1809 by Gauss (Chatfield, 1993). Both Legendre and Gauss used the method to determine the position of orbits of the planets in the solar system. Regression analysis is defined as a statistical tool meant for the investigation of the relationship between variables (Zou et al., 2003). It is a mathematical process of using data points on observations to find the *line* of *best fit* through the data points in order to make estimates and predictions about the behaviour of the variables. This line of best fit may be linear (straight) or curvilinear to some mathematical formula. Regression analysis includes any statistical techniques for modelling and analyzing several variables, when the focus on the relationship is between dependent variable and one or more independent variables. Usually the researcher seeks to ascertain the causal effect of one variable upon another variable. After fitting the model, the statistical significance of the estimated relationships that the true relationship is close to the estimated relationship is assessed.

2 Purpose of regression analysis

The purpose of regression analysis is to analyze relationships among variables. The analysis is carried out through the estimation of a relationship $y = f(x_1,x_2, \ldots,x_n)$ and the result serve the following two purposes:

i. Answer the question of how much y changes with changes in each of the x's $(x_1,x_2, \ldots,x_n)$.
ii. Forecast or predict the value of y based on the values of the x's.

3 Assumptions in regression analysis

For ordinary regression estimates to have good properties the following assumptions also called Gauss – Markov assumptions need to be met (Zou et al., 2003) :

i. The sample is representative of the population for inference predictions.
ii. The errors (also called noise or disturbance) is a random variable with expected value of zero, ei(u,1) = 0.
iii. The independent variables are measured accurately with no errors.
iv. The independent variables (predictors) are linearly independent i.e. multi-colinearity is absent.

v. Error variance is the same for each observation (homoscedastic).
vi. The errors are not auto-correlated i.e. errors associated with different observations are uncorrelated.

4 Steps of regression analysis

To carry out any regression analysis, the following steps as outlined by Samprit and Ali (2006) need to be observed:

i. Statement of the problem
ii. Selection of potentially relevant variables
iii. Data collection
iv. Model specification
v. Choice of fitting method
vi. Model fitting
vii. Model validation and criticism

5 Classification of regression analysis

The classes of regression analysis and the conditions for their use as given by Samprit and Ali (2006) are as tabulated in Table 1.

Table 1: Types of regression

Type of regression	conditions
Univariate	only one quantitative response variable
Multivariate	Two or more quantitative response variables
Simple	Only one independent variable
Multiple	Two or more independent variables
Linear	All parameters enter the equation linearly, possibly after transformation

Nonlinear	The relationship between response and some of the independent variables is non linear or some of the parameters appear nonlinearly but no transformation is possible to make the parameters appear linearly.
Analysis of variance	Some predictors are quantitative variables
Analysis of covariance	Some predictors are quantitative variables and others are qualitative variables.
Logistic	The response variable is qualitative

5.1 Linear regression

The purpose of simple regression analysis is to evaluate the relative impact of a predictor variable on a particular outcome. Linear regression model may be *simple* (when the model contains only one independent variable, X_i) or *multiple* (when the model contains two or more explanatory variables $X_1, X_2, ----,X_n$). The term linear means that the regression parameters a, and b enter the model in a linear fashion. Thus the expression $y = a + bx^2 + ei$ is a linear model even though the relationship between y and x is quadratic. A simple linear regression model is expressed as:

$$y = a + bx + ei$$

where,

$$b = \frac{\sum (y - Y)(x - X)}{\sum (x - X)^2} \quad \text{and} \quad a = y - bx$$

Where *a* is the intercept (on the y axis), and *b* is the slope of the regression line. The random error term *ei* is assumed to be uncorrelated, normally distributed with a mean of 0 and constant variance. Thus, the word *line* (linear, independent, normal, equal variance) summarizes these requirements (Zou et al., 2003). The significance of the slope of the regression line is determined from the t-statistic or using the F-ratio instead of t-statistic. The t-statistic for the significance of the slope is essentially a test to determine if the regression model (equation) is usable. If the slope is significantly different than zero, then we can use the regression model to predict the dependent variable for any value of the independent variable. In testing the hypothesis for regression parameters,

H_o: $a = 0$, and H_1: $a \neq 0$ and

H_o: $b = 0$, and H_1: $b \neq 0$ / the corresponding t-statistics to use are given as:

$$t_a = \frac{a}{s.e.(a)},$$

where $s.e.(a)$ = standard error of a.

$$t_b = \frac{b}{s.e.(b)},$$

where $s.e.(b)$ = standard error of b.

The critical value for the above tests is $t_{n-2,a/2}$. If $t_a > t_{n-2,a/2}$ the result is highly significant and H_o is rejected.
Similarly for $t_b > t_{n-2,a/2}$ will result in rejecting H_o, else H_1 is accepted.
Confidence interval is computed using the following equations as given by Chatfield (1993):

$$a = \pm t_{n-2,a/2} \times s.e.(a)$$

$$b = \pm t_{n-2,a/2} \times s.e.(b)$$

A multiple linear regression model may be in the following format:

$$y = a + b_1 x_1 + b_2 x_2$$

$$y = a + b_1 x + b_2 x^2$$

where $a, b_1, b_2, \ldots, b_n$ are the regression parameters which are estimated using the following expressions:

$$b_1 = \frac{(SSx_2.SSx_1 y - SSx_1 x_2.SSx_2 y)}{SSx_1.SSx_2 - (SSx_1 x_2)^2}$$

$$b_2 = \frac{(SSx_1.SSx_2 y - SSx_1 x_2.SSx_1 y)}{SSx_1.SSx_2 - (SSx_1 x_2)^2}$$

$$a = y - b_1 x_1 - b_2 x_2$$

The general multiple linear regression model is expressed as

$$y = a + b_1 x_1 + b_2 x_2 + \cdots + b_n x_n + ei$$

For multiple linear regression, F-ratio is used to test the hypothesis:

H_o: $b_1 = b_2 = 0$ and H_A : some $b \neq 0$

If F-cal > F(k, n-k-1), the null hypothesis is rejected and H_A accepted that some $b \neq 0$ (i.e. y is linearly related to x_1 and x_2). F-cal is obtained thus:

ANOVA

Source	df	SS	MS	F
Model	k	$b_1 \Sigma x_1 y + - - - + b_k \Sigma x_k$		
Error	n – k – 1	$\Sigma y^2 – (b_1 \Sigma x_1 y + - - - + b_k \Sigma x_k)$	S^2	
Total	n	Σy^2		

$$\text{Fcal} = \frac{MS\ \text{Regressed}}{MS\ \text{Error}} = R^2$$

5.2 Coefficient of determination (R^2)

The coefficient of determination in a regression model, also known as the R-square statistic (R^2), measures the proportion of variability in the response that is explained by the regressor variables (McDonald, 2009). R^2 statistics measures how well the model explains the data. It is interpreted directly as the proportion of variance in the dependent variable that is accounted for by the regression equation. For example an R^2 of .89 means that 89% of the variance in the observed values of the dependent variable is explained by the model, and 11% of those differences remains unexplained in the error (residual) term.

In a linear regression model with intercept, R^2 is defined as 1 minus the ratio of residual variability, mathematically expressed as:

$$R^2 = 1 - \frac{\text{SSE}}{\text{SST}}$$

Where SSE is the residual (error) sum of squares and SST is the total sum of squares corrected for the mean. The adjusted R^2 statistic is an alternative to R^2 that takes into account the number of parameters in the model. This statistic is calculated as

$$\text{ADJRSQ} = 1 - \frac{n-i}{n-p}(1-R^2)$$

where n is the number of observations used to fit the model, p is the number of parameters in the model (including the intercept), and i is 1 if the model includes an intercept term, and 0 otherwise. Adjusted R^2 *does not have the same interpretation as* R^2. As such, care must be taken in interpreting and reporting

this statistic. Adjusted R^2 is particularly useful in future selection stage of model building.

5.3 Data transformation

Data transformation to achieve normality may be applied when one or more variables are not normally distributed (Carroll and Ruppert, 1988), with the view to normalize them. Transformation is as well done to correct for hetero scedasiticy, non-linearity and outliers. There are several methods for data transformations (square, square root, log, reciprocal, arcsine etc.) however, the method to adopt depends on the one that gives the best result i.e. transformation whose distribution is most normal. Some nonlinear regression problems can be moved to a linear domain by a suitable transformation of the model formulation. For example, consider the nonlinear regression problem (ignoring the error):

$$y = ae^{bx}.$$

If we take a logarithm of both sides, it becomes

$$\ln(y) = \ln(a) + bx$$

Suggesting that estimation of the unknown parameters by a linear regression of $\ln(y)$ on *x*, a computation that does not require iterative optimization.

5.4 Nonlinear regression

In statistics, Nonlinear regression, is the problem of fitting a model $y = f(x,t) + c$ to measured x, y data, where f is a nonlinear function of x with parameter t. The parameters appear as functions, such as b^2, e^{bx} and so forth which cannot be expressed in linear form. For example the expression:

$$V = \frac{V_m(S)}{K_m + S}$$

this can be written as:

$$f(x,b) = \frac{b_1 X}{b_2 + X}$$

Where b_1 is the parameter V_m, b_2 is the parameter K_m and [*S*] is the independent variable, *x*. This function is nonlinear because it cannot be expressed as a linear combination of the b_s. Other examples of nonlinear functions include exponential, logarithmic, trigonometric, power, Gaussian functions, and

Lorentziancurves (Seber and Wild, 1989). In nonlinear regression, data are fitted by a method of successive approximations.

6 Conclusion

Linear regression is a widely used statistical tool for providing functional relationship among variables. Its areas of applications include: economics, finance, law, meteorology, medicine, physics, chemistry, biological, education, sports, history, behavioural and social sciences. It ranks as one of the most important statistical tools used in these disciplines.

Although regression analysis as a statistical tool for making predictions and forecast is widely used in both biological, engineering, behavioral and social sciences, the technique has one major conceptual limitation. That is the model is only able to ascertain relationships between variables, but never tell with certainty about the underlying causal mechanism. No matter how strong a relationship is demonstrated in regression analysis, it is never interpreted as causation (as in the correlation analysis). As a precautionary measure against major errors, model developed using regression technique should not be used to predict or estimate outside the range of values of the independent variable of the sample.

References

[1] R.J. Carroll and D. Ruppert, *Transformation and weighting in regression*, New York, NY: Chapman & Hall, 2-61, 1988.

[2] C. Chatfield, Calculating Interval Forecasts, *Journal of Business and Economic Statistics*, **11**, (1993), 121-135.

[3] N.R. Draper and H. Smith, *Applied Regression Analysis*, Wiley Series in Probability and Statistics, 1998.

[4] J. Fox, *Applied Regression Analysis*, Linear Models and Related Methods, Sage, 1997.

[5] J.H. McDonald, *Handbook of Biological Statistics* (2nd ed.), Sparky House Publishing, Baltimore, Maryland, 2009.

[6] C. Samprit and S.H. Ali, *Regression analysis by example*, John Wiley and Sons, New York, 10-50, 2006.

[7] G.A.F. Seber and C.J. Wild, *Nonlinear Regression*, John Wiley and Sons, New York, 88 -100, 1989.

[8] K.H. Zou, T.M.D. Kemal and G.S. Stuart, Correlation and simple linear regression 1 Statistical Concepts Series. J. Radiology, **227**, (2003), 617-628.