SCITECH
RESEARCH ORGANISATION

Volume 12, Issue 4
Published online: November 03, 2017|

Journal of Progressive Research in Mathematics
www.scitecresearch.com/journals

A Comparison Study of Linear and Nonlinear Regression Models

Taghreed Abdul-Razek Abdul-Motaleb Al-Said, Ph.D.
A lecturer of Statistics at AL AZHAR University, Faculty of Commerce,
Department of Statistics, Cairo, Egypt.
Assistant professor of Statistics at King Abdul-Aziz University, Faculty of Science,
Department of Statistics.

Abstract

Regression analysis is an important statistical tool for analyzing the relationships between dependent, and independent variables. The main goal of regression analysis is determine, and estimate parameters of a function that describe the best fit for a given data sets. There are many linear types of regression analysis models such as simple and multiple regression models. Also, there are the non-linear regression analyses such as binary and multinomial logistic regression models. This research at first, introduced many types of such models. Second, estimates the parameters of the models by using the maximum likelihood estimation, and the least square estimation methods. Also, it introduces some criteria for evaluating methods. Two suitable applications on two different data sets are conducted, and useful results are concluded.

Keywords: linear regression models; logistic regression models; ordinary least-square, Wald test, R-squared test.

1. Introduction

Regression analysis is the widely used statistical tool for understanding relationships among variables. It is used when there is a continuous dependent variable which could predict by independent variables. If the dependent variable is dichotomous, logistic regression is the reasonable model in this case. Regression analysis can be used in many applications such as medical, education, and many other applications

This research concentrated with many regression models such as the linear regression models, the simple and multiple regression models. Also logistic regression models will be included in the research in a comparison between linear and nonlinear models in two suitable applications. The definitions and details of regression models will be stated in section (2). Section (3) presents many evaluating criteria can be used with these models. Two suitable applications of these models and analysis will be stated and discussed in section (4). The conclusions and recommendations of the study will be in section (5), and finally at the end of the study there is the references section in section (6).

2. The Regression Models:

This section has details of many regression models. It has the linear regression models, the simple and multiple regression models. Also it has the nonlinear regression models.

2.1 The linear regression models:

Seber (1977) defined linear regression analysis, LRA, as a common technique of estimating the relationship between any two random variables, the explanatory variable X, and the dependent variable Y such as height and weight, income and intelligence quotient, ages of husband and wife. Bates and Watts (1988) mentioned that LRA as a powerful methodology for analyzing data and used for describing the relation between the predictor variables. A researcher often has a mathematical expression which relates the response and the predictor variables, and these models are usually nonlinear in the parameters. In such cases, linear regression techniques tend to be more complexity and has not validity. Chatterjee and Hadi (2006) defined regression analysis, RA as a conceptually simple method for investigating functional relationship among variables. The simple relationship among dependent and explanatory variables can be defined as follows:

$$Y= f(X_1, X_2, \dots, X_p) + \varepsilon \qquad (2.1)$$

where a random error representing the discrepancy in the approximation is assumed to be ε. It accounts the failure of the model to fit the data exactly. The function$f(X_1, X_2, \dots, X_p)$ describes the relationship between the dependent variable Y, and the explanatory variables $X_1, X_2, \dots, X_p$.

Hutcheson and Moutinho (2011) defined simple linear regression, SLR model as a relationship between a continuous response variable Y, and a continuous explanatory variable X may be represented by using a line of best fit where Y is predicted at least to some extent by X. When the relationship is linear, it may be represented mathematically using a straight line equation. The regression coefficient describes the change in Y that is associated with a unit change in X. This line is frequently computed using the least square procedure.

Dayton (1992) defined multiple linear regression, MLR models a linear combination of a set of predictors, error and the dependent variable. The relation for an outcome variable Y, and a set of p prediction variables $X_1, X_2, \dots, X_p$ has the following form:

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon$$

$$= \alpha + \sum_{j=1}^{p} \beta_j X_j + \varepsilon \qquad (2.2)$$

whereα is the Y-intercept, the expected value of Y when all X's are set equal to 0, and a multiple regression coefficients areβ_j. The expected change in Y per unit change in X_j assuming all other X's are held constant. The error of prediction is ε. If error is omitted, the model represented the expected as predictedvalue of Y as follows:

$$E(Y|X_1, X_2, \dots, X_p) = \hat{Y}$$

$$= \alpha + \sum_{j=1}^{p} \beta_j X_j \qquad (2.3)$$

The multiple regression analysis is applicable when the outcome variable Y is continuous. It is not appropriate for situation in which Y is dichotomous, categorical, or discrete.

Hosmer and Lemeshow (2000) mentioned that in any regression problem, the key quantity is the mean value of the outcome variable given the value of the independent variables. This quantity is called the conditional mean. It also known as conditional expected value, or conditional expectation. It is the expected value of a real random variable with respect to a conditional probability distribution.

An analysis of variance table, partitions the total sum of squared deviations of observations about their means into two parts: the sum of squared deviations of observations about regression line SSE, or residual sum of squares, and the sum of squares of the predicted values based on the regression model about the mean of the dependent variable SSR, or due regression sum of squares.

2.2 The nonlinear regression models:

Ratkowsky (1983) defined the nonlinear regression, NLR models as follows:

$$Y_t = X_t{}^{\theta} + \epsilon_t \qquad (2.4)$$

Where the response variables are Y_t for t=1,2,…,n ,the parameter to be estimated isθ. The predictors areX_t, andunobservable random error term whose values are unknown and assume to have zero mean value are ϵ_t.

The nonlinear regression models different from linear regression models not only in biased parameters of the least square estimators, or the non-normality distributed of the response variable. But also, the variances exceeding the minimum possible variance, over-dispersion phenomena. In additional, the nonlinearity of the relationship between the response variable and the predictors. The purpose of linear regression is to find values for the slope/s, and intercept to define the line that comes closest to the data. Nonlinear regression models are more general than linear regression, and can fit any data by defining Y as a function of X with one or more parameters. The values of those parameters generate the closed curve to the data.

Cox and Snell (1989) mentioned that logistic distribution has primary reasons for choosing it for solving nonlinear regression models because of its extremely flexible and easily used function. Also, it lends a clinically meaningful interpretation.

Dayton (1992) defined logistic regression, LOR model as an extend technique of multiple regression analysis to study situations in which the outcome variable Y is categorical or dichotomous. It does not model this outcome variable directly. It is based on probabilities associated with the values of Y. The logistic regression model is defined as follows:

$$\log\left(\frac{\pi_j(x_i)}{\pi_k(x_i)}\right) = \beta_{0j} + \beta_{1j}x_{1i} + \beta_{2j}x_{2i} + \cdots + \beta_{pj}x_{pi} \qquad (2.5)$$

$$j=1,2,\ldots,(k-1),\ i=1,2,\ldots,n$$

Since all the π's add to unity, this reduces to:

$$\log(\pi_j(x_i))=\frac{\exp(\beta_{0j}+\beta_{1j}x_{1i}+\beta_{2j}x_{2i}+\cdots+\beta_{pj}x_{pi})}{1+\sum_{j=1}^{k-1}\exp(\beta_{0j}+\beta_{1j}x_{1i}+\beta_{2j}x_{2i}+\cdots+\beta_{pj}x_{pi})} \quad (2.6)$$

wherethe multiple regression coefficient are$\beta_{0j},\beta_{1j},\ldots,\beta_{pj}$, andthe predictors that often called independent variables are$X_1,X_2,\ldots,X_p$.

Hosmer and Lemeshow (2000) defined logistic regression as a method that describe an integral component of any data analysis concerned with describing the relationship between a response variable, and one or more explanatory variables. Logistic regression model is different from the linear regression model where the outcome variable in logistic regression is binary, dichotomous, or categorical. Also, the difference between logistic, and a linear regression is effected both in the choice of parametric model, and in the assumptions.

Pohlmann and Leitner (2003) defined Logistic regression, LOR, as the most frequently used statistical procedures in social science research, and medical researchers. In this technique events are coded as binary variables with a value of 1 that represent the occurrence of target outcome, and a value zero represents its absence. If the research does indicate a certain order importance of the predictor variables, then a sequential logistic regression is the appropriate statistic to use.

Chatterjee and Hadi (2006) commented that logistic regression model can be extended to situations where the response variable assumes more than two values. In a study of the choice of mode of transportation to work, the response variable may be private automobile, car pool, public transport, bicycle, or walking. The response falls into five categories .There is no natural ordering of the categories. The researcher might want to analyze how the choice is related to factors such as age, sex, income, distance traveled, etc. The resulting model can be analyzed by using slightly modified methods that were used in analyzing the dichotomous outcomes. This method is called the multinomial (polytomous) logistic regression.

Raghavendra and Srivatsa (2011) introduced the performance of logistic regression, and neural network models on publicly available medical datasets. An attempt was made to evaluate logistic regression, and neural network model with sensitivity analysis. The classification accuracy was used to measure the performance of both models. From the experimental the neural network model was the sensitivity analysis, and gave more efficient results.

3. Evaluating Criterions for Regression Models

There are many criteria for evaluating regression models. This section introduced many of these criterions.

3.1 The multiple R- square measure:

The multiple R-square, $R^2$is the famous measure of the goodness of fit for the fitted regression line to a given set data. It describes the amount of variation that explained by the model. R-squared can also be interpreted as the proportionate reduction in error in estimating the dependent variable when knowing the independents variables. The R^2 reflects the number of errors made when using the regression model to guess the value of the dependent variable, in

ratio to the total errors made when using only the dependent mean as the basis for estimating all cases. It can be estimated as follows:

$$R^2 = (1 - \frac{SSE}{SST}) \tag{3.1}$$

where the error sum of squares is SSE that equals the squared sum of $(Y_i - \widehat{Y}_i)$. The actual value of Y for the i th case and the regression prediction for case iis $\widehat{Y}_i$. The total sum of squares, SST can be estimated by summing$(Y_i - \overline{Y})^2$. R-square ranges from 0 to 1,where 0 reflects no variation in the dependent variable is explained by the independent variables, and 1 reflects all the variation in the dependent variable explained by the independent variables. [Frank, etal (2007)].

3.2 The adjusted R-squared measure

The adjusted coefficient of determination of a multiple linear regression model or the adjusted R-Squared measure is defined in terms of the coefficient of determination as follows:

$$R^2_{adj} = 1 - (1 - R^2)\frac{(n-1)}{(n-p-1)} \tag{3.2}$$

where the number of observations in the data set is n, and the number of independent variables is p. Adjusted R-square statistic is used instead of the R-square because of added significantly variables which contribute to the model and raise the statistic value. It is often used to compare several models and reflects the best model. [Frank, et al. (2007)]

This R^2 test of logistic regression model is like other regression models. It tries to measure the strength of association of the model. The values of this test are between 0 and 1. It is the most common and considerable measure of indication the strength of association. Various R-squared statistics have been proposed for logistic regression to quantify the extent binary response that predicted by a given logistic regression model, and covariates. [Frank, et al. (2007)]

3.3 The Pearson chi-square statistic

The standard test for assessing goodness-of-fit of logistic regression models, is the Pearson chi-square, $\mathcal{X}^2_p$, statistic. It can be calculated as follows:

$$\mathcal{X}^2_p = \sum_{i=1}^{j} \frac{(y_i - m_i\hat{\pi}_i)^2}{m_i\hat{\pi}_i(1-\hat{\pi}_i)} \tag{3.3}$$

where the Pearson residual term is $\frac{(y_i - m_i\hat{\pi}_i)}{\sqrt{m_i\hat{\pi}_i(1-\hat{\pi}_i)}}$, $\sum m_i = n$ is the number of subjects, i=1,…,j . The number of distinct values of observed x is denoted by j, and $\hat{\pi}_i$ is the maximum likelihood of π_i(conditional mean). [Hosmer and Lemeshow (2000)].

3.4 The deviance residual statistic

The deviance residual, D statistic can be defined as follows:

$$d(Y_i, \hat{\pi}_i) = \pm\left\{2\left[Y_i \ln\left(\frac{Y_i}{m_i\hat{\pi}_i}\right) + (m_i - Y_i)\ln\left(\frac{m_i - Y_i}{m_i(1-\hat{\pi}_i)}\right)\right]\right\}^{1/2} \quad (3.4)$$

where the sign is the same as the sign of $(y_i - m_i\hat{\pi}_i)$,$\sum m_i = n$ is the number of subjects, i=1,…,j . The number of distinct values of observed x is denoted by j, and $\hat{\pi}_i$ is the maximum likelihood of π_i(conditional mean). The distribution of the statistics is chi-square with (j-p-1) degrees-of-freedom, where j is the number of covariate patterns, and p is the number of predictor variables in the model. [Hosmer and Lemeshow (2000)]Pearson statistic and deviance rely on comparing observed $Y_i$andpredicted $(m_i\hat{\pi}_i)$ values, and should be large if the model does not fit the data well. [Kuss (2002)].

3.5 The Wald test

The Wald test is used to test the significance for logistic regression coefficients. The null hypothesis, H_0: $\beta_j = 0$, against the alternative H_1: $\beta_j \neq 0$,and the statistic has the following form:

$$Z = \frac{b_j}{s_j} \quad (3.6)$$

wheres_j is the estimated standard error for the estimated coefficient b_j. SPSS and SAS packages report $\chi^2 = Z^2$ and label there values as Wald statistic.The test hypothesis is simultaneously for all partial logistic regression coefficient is 0, i.e. ,H_0: $\beta_j = 0$ for all j. This test is labeled in SPSS as "Model Chi-Square". [Dayton (1992)]

3.6 Cox and Snell R-square measure

The ratio of the likelihoods reflects the improvement of the full model over the intercept model (the smaller of the ratio, the greater of the improvement). Consider the conditional probability of the dependent variable given the independent variables is L (M). If there are N observations in the dataset, then L(M) is the product of N such probabilities. Thus, taking the n^{th} root of the product L(M)provides an estimate of the likelihood of each Y value. Cox and Snell's present the R-squared as a transformation of the $-2ln\left[\frac{L(M_{intercept})}{L(M_{full})}\right]$statistic that is used to determine the convergence of a logistic regression.

Cox and Snell's pseudo R-squared has a maximum value that is not 1. If the full model predicts the outcome perfectly and has a likelihood of 1, Cox and Snell's is then

$1 - L(M_{intercept})^{2/N}$, which is less than one. The statistics has the following form:

$$R^2 = 1 - \left\{\frac{L(M_{intercept})}{L(M_{full})}\right\}^{\frac{2}{N}} \quad (3.15)$$

For a normal generalized linear model, R-Square formula has a maximum of one, but for logistic regression its maximum is 0.75 or lowers [Baguley (2012)].

4. The Applications

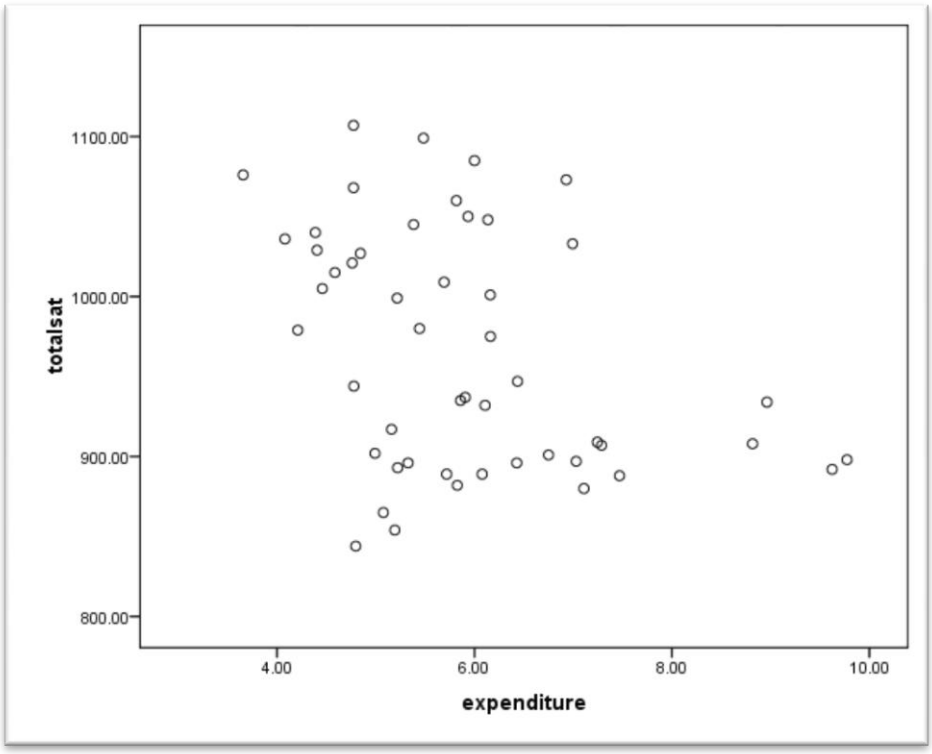
This section has two applications of simple, multiple and logistic regression models as an important comparison between the suggested three models.

4.1 The SAT Data Set

The first dataset is the SAT data. It is extracted from the 1997 Digest of Education Statistics of 50 subjects. This teaching case analyzing the relationship between public school expenditures, and academic performance as measured by the SAT. The data in Table (A.2) in the appendix contains eight columns. The name of state will be in the first column, the current expenditure per pupil in average daily attendance in public elementary and secondary schools will be in second column, the average teacher ratio in public elementary and secondary schools in the third column, estimated average annual salary of teachers in public elementary and secondary schools in the fourth column, percentage of all eligible students taking the SAT in fifth column, average verbal SAT score in the sixth column, average math SAT score in the seventh column, and average total score on the SAT in eighth column.

4.1.1 The descriptive of the SAT data

A scatterplot will be drawn to exposure the relationship between the public school expenditures (the independent variable), and the academic performance (the dependent variable).Figure (4.1) shows the relationship between the two variables. It appears to be negative, that shows that score highest on the SAT on average spend less money per student. It is not a logical result, so the second variable will be added to the model. The new model contains the academic performance as measured by SAT (the dependent variable), and the public school expenditures, estimated average annual salary of teachers in public elementary and secondary schools, and percentage of all eligible students taking the SAT expenditures (the independent variables)



Figure(4.1): Scatterplot of the SAT set data of the simple linear regression

4.1.2 The estimation of the simple and multiple model parameters of the SAT set data

The SAT data set will be analyzed by using SPSS package, and the results are seated in Table (4.1). It shows the estimate values of the model parameters for the data set of the simple regression model, and multiple regression model respectively. For the simple linear regression, $\widehat{B}_0$ is the intercept estimate that equal to 1089. The standard error of estimate, S.E, for $\widehat{B}_0$ is equal to 44.39. It indicate how badly the prediction in the unit of this variable. Also, the P-value of the $\widehat{B}_0$ is equal to 0.0001. It means that there is enough evidence to support the hypothesis that $\widehat{B}_0$ is effect on the model with confidence interval level equal to 95%.

Table (4.1): Estimation of the simple and multiple linear model parameter

Data	Types of analysis	Parameters	Estimation of parameters	S.E	P-value
The SAT set data	Simple Linear regression	$\widehat{B}_0$	1089	44.39	.0001
		$\widehat{B}_1$	-20.892	7.328	.006
	Multiple Linear regression	$\widehat{B}_0$	1035	50.316	.0001
		$\widehat{B}_1$	-2.851	.215	.0001
		$\widehat{B}_2$	11.01	4.452	.017
		$\widehat{B}_3$	- 2.03	2.207	.363

The result of the simple linear fitted regression model using the ML estimation method, where x1 is the public school expenditures will be as follows:

$$\widehat{Y} = 1089 - 20.892\ x1$$

The fitted model of the multiple linear regression will be as follows:

$$\widehat{Y} = 1035 - 2.85\ x1 + 11.01\ x2 - 2.03\ x3$$

where x_1 percentage of all eligible students taking the SAT expenditures, x2 public school expenditures, and x3 estimated average annual salary of teachers in public elementary and secondary school.

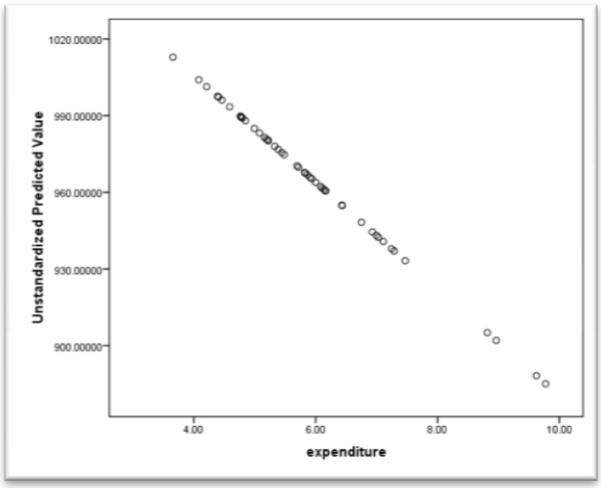
4.1.3 The comparative criteria of the SAT data set

In linear regression model, the multiple R-Squared and adjusted R-squared criterion are used to explain how the independent variables interpret the model. Table (4.2) has the coefficient of determination in R-Square for the simple linear regression, R^2 equals 0.145 which means that 14% of the total variation in Y can be explained by the explanatory variable x1 and the other 86% remains unexplained. On the other hand, the adjusted R-square in multiple linear regression equals0.811 which means that 81% of the total variation in y can be explained by the explanatory variables X_1, X_2, X_3 , where the other 19% remains unexplained.

Table (4.2): The coefficient of determination of The SAT set data

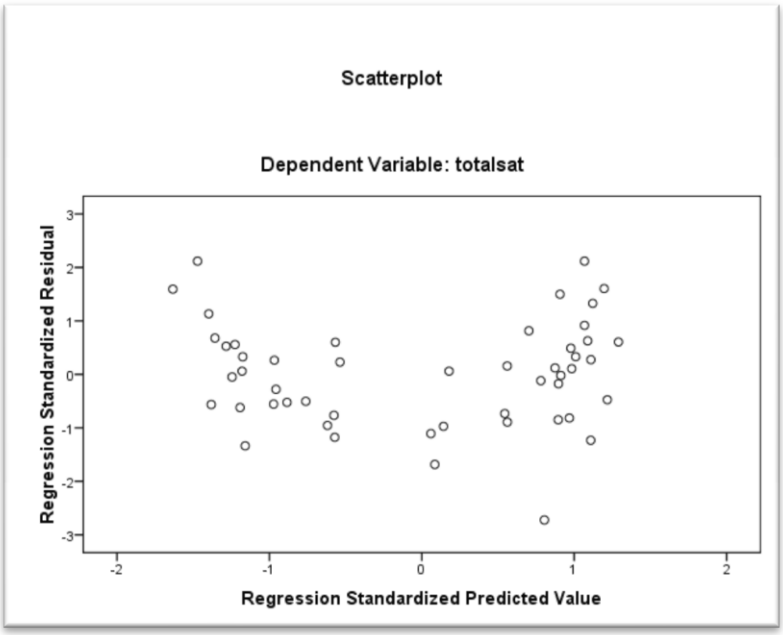
The SAT set data	
R-Squared of simple regression model	Adjusted R-Squared of the multiple regression model
0.145	0.811

Figure (4.2) shows the scatterplot for the fitted model of the simple regression model of the first data set. It reveals a decreasing relationship between the prediction values (dependent variable) , and the public school expenditures (independent variable).



Figure(4.2): Scatterplot of the fitted simple regression model

Figure (4.3) shows the scatterplot for the residual of the multiple regression model. It has the U-shape, that means the polynomial term or the logarithmic transformation will produce a best fitting nonlinear regression equation.



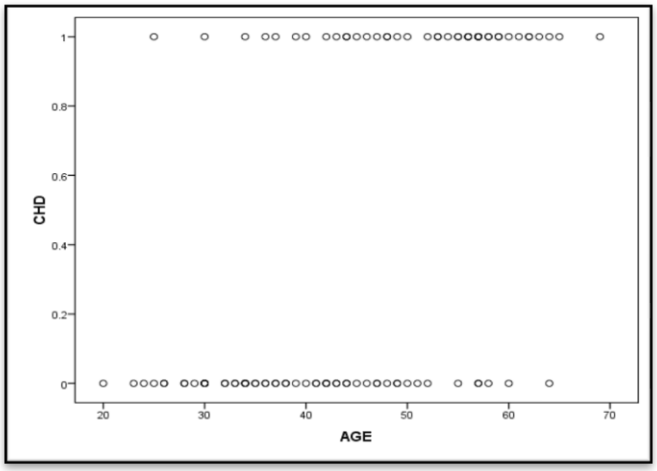
Figure(4.3): Scatterplot for the residual of the multiple regression model

4.2 The Coronary Heart Disease data set

The second data set, coronary heart disease data set, collected from the Hosmer and Lemeshow (2000) of 100 subjects to detected evidence of significant coronary heart disease (CHD). The data in Table (A. 2) in the appendix lists age in years (AGE) in the first column, and presence or absence of evidence of significant coronary heart disease (CHD) in the second column. Also, the table contains an age group variable (AGE in groups)in the third column that suggests groups. In the fourth column, the percent of CHD. The outcome (dependent) variable is (CHD), which is binary and coded with a value (0) to indicate CHD is absent, and (1) to indicate the present of the disease in the individual. The study is interested on exploring the relationship between age and the presence or absence of CHD.

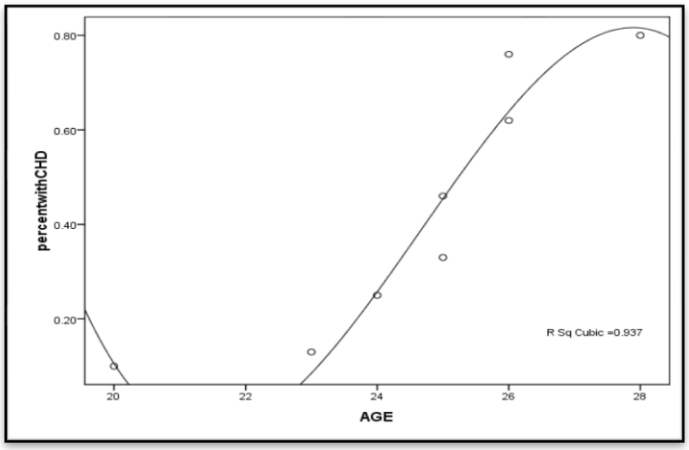
4.2.1 Descriptive of the coronary heart disease data set

In the coronary heart disease data set, scatterplot of the outcome (CHD) versus the independent variable (AGE) where plotted and described in Figure (4.4) that shows the nature and strength of the relationship between the outcome and the independent variable. The scatterplot shows all points fall on approximately two parallel lines representing the absence of CHD (y=0) and the presence of CHD (y=1). It does not provide clear picture of the relationship between CHD and age. Another problem is the variability in CHD at all ages is large. This makes it difficult to describe that relationship between age and CHD.



Figure(4.4): Scatterplot of the coronary heart disease data set

In Figure (4.5) solve the problem by using categories of the age group variable (AGE) using eight categories. For each age group, the frequency of occurrence of each outcome as the mean (or proportion with CHD) is present.



Figure(4.5): Scatterplot of the categories of the coronary heart disease data set

4.2.2 The estimation of the logistic model parameters of the coronary heart disease data set

The coronary heart disease data set will be analyzed by using SPSS package, and the results are seated in Table (4.3). It shows the estimate values of the logistic regression model parameters for the data set. The $\widehat{B}_0$ is the intercept estimate which equals to -5.309. The standard error of estimate S.E, for $\widehat{B}_0$ equals to 1.134. It indicates badly prediction in this variable. Also, the P-value of the $\widehat{B}_0$ is equal to 0.0001. It means, there is enough evidence to support the hypothesis that $\widehat{B}_0$ is effect on the model with confidence interval level equal to 95%, and Exp(B) is equal 1.116. It is used to model a relationship in which a constant change in the independent variable gives the same proportional change (i.e. percentage increase or decrease) in the dependent variable and named the odds.

Table (4.3): Estimation of the binary logistic regression model parameters

		Parameters	Estimation of parameters	S.E	P-value	Exp(B)
the coronary heart disease data set	Logistic regression	$\widehat{B}_0$	-5.309	1.134	.0001	1.117
		$\widehat{B}_1$	0.111	0.024	0.0001	0.005

The logistic regression model of the data set will be as follows :

$$\hat{\pi}(x) = \frac{e^{-5.309+0.111\times AGE}}{1+e^{-5.309+0.111\times AGE}}$$

Table (4.3) shows the estimation of the parameters $\widehat{B}_0$ and $\widehat{B}_1$ as -5.309 and 0.111 respectively. The estimation values distracted from SPSS Package and using the ML estimation method. Since $\widehat{B}_1 = 0.111 > 0$, the estimated probability of CHD increases as AGE level increases.If AGE changes by one unit, then the odds change multiplicatively by 1.117, where

$$odds = \frac{\hat{\pi}(x)}{1-\hat{\pi}(x)} = \exp(-5.309 + 0.111 \times AGE);$$

$$= \exp(-5.903) * \exp(0.111 \times \text{AGE})\,;$$

And,

$$\text{logit}[\hat{\pi}(x)] = -5.309 + 0.111 \times \text{AGE}.$$

4.2.3 The comparative criteria of the coronary heart disease data set

In logistic regression model, the criteria Cox and Snell R-Squared can be interpreted like multiple R-Squared, in Table (4.5) the coefficient of determination using Cox and Snell, R-Squared equals 0.254 and 25.4%, the outcome variable y can be explained by the relationship between Y and predictor .

Table (4.4): The coefficient of determination of the coronary heart disease data set

The coronary heart disease data set
Cox and Snell R-Square
.2540

Figure (4.6) shows the scatterplot for the fitted model of the logistic regression model of the coronary heart disease data set. The shape of Figure (4.6) is S-shape, When $\beta > 0$, $\pi(x)$ increases as x increases.

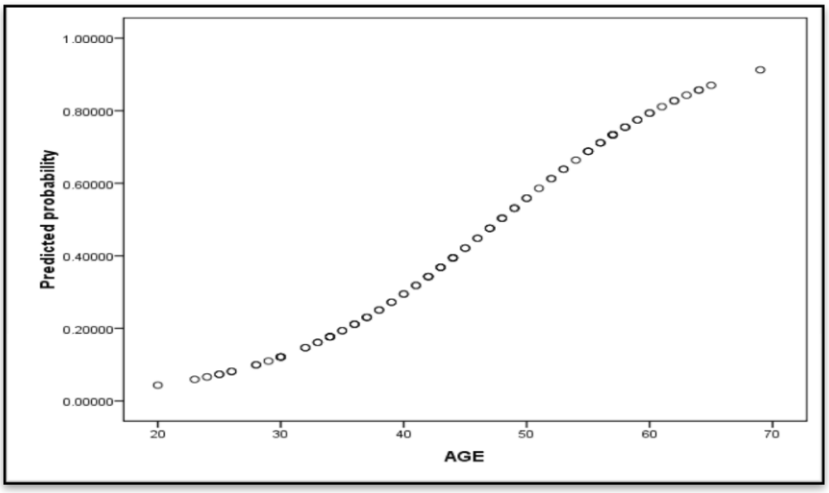


Figure (4.6): Scatterplot of the fitted binary logistic regression model

5. Conclusions and Recommendations

There are many types of regression analysis models. This research interested in comparing two frequently use linear regression analysis models and one of common use model in non-linear regression analysis models. The linear regression analysis models are simple linear regression and multiple linear regression models. The non-linear regression analysis is the logistic regression model. The linear regression models are applied on a suggested data set extracted from the 1997 Digest of Education Statistics of 50 subjects. This case analyze the linear relationship between public school expenditures, and academic performance as measured by the SAT. The second data set collected from the Hosmer and Lemeshow (2000) of 100 subjects to detected evidence of significant coronary heart disease (CHD).

The regression models of the two data sets are distracted using two estimating techniques for estimating the parameters of the suggested models, the maximum likelihood estimation, and least square estimation method. Also, some criteria for evaluation models are reviewed and applied, the Pearson Chi-Square statistic, the Deviance Residual statistic, the Wald test, Cox and Snell R-Square. The results show that it is very important to graph data set to decide which type of models suitable to use. Also, if the dependent variable is binary or categorical the logistic model is the suitable use model to describe and analyze the data. It is very important to use suitable evaluating criterion to determine which strong explanatory variable that explain and associate the relationship and also determining the suitable model can be used. It is suitable to try using the Bayesian approach for estimating regression parameters and applying regression models in solving experimental design problems. There are alternative models can be used with categorical data than the logistic regression model

References:

[1] Bates, D., and Watts, D. (1988). Nonlinear Regression Analysis and Its Application. The United States of America, John Wiley & Sons, Inc.

[2] Chatterjee, S., and Hadi, A. (2006). Regression Analysis by Example. The United States of America, John Wiley & Sons, Inc.

[3] Cox, D., and Snell, E. (1989). Analysis of Binary Data. London, Chapman & Hall, Inc.

[4] Frank, P., Weil, R., Wager, M., and Hughes, C. (2007). Litigation Services Handbook: The Role of The Financial Expert. Canada, John Wiley & Sons, Inc.

[5] Hosmer, D., and Lemeshow, S. (2000). Applied Logistic Regression. Canada, John Wiley & Sons, Inc.

[6] Hutcheson, G., and Moutinho, L. (2011). Ordinary least square regression. The SAGE Dictionary of Quantitative Management Research, PP. 224-228

[7] Kuss, O. (2002). Global Goodness-of-Fit Tests in Logistic Regression with Sparse Data. Germany, John Wiley & Sons, Ltd.

[8] Pohlmann, J., and Leitner, D. (2003). A Comparison of ordinary least squares and logistic regression. Ohio Journal of Science, vol.103, no.5, pp.118-125.

[9] Raghavendra, B.K., and Srivatsa, S.K. (2011). Evaluation of logistic regression and neural network model with sensitivity analysis on medical datasets. International Journal of Computer Science and Security, vol.5, no.5, pp.504-511.

[10] Ratkowsky, D. (1983). Nonlinear regression Model: A unified Practical approach. New York, Marcel Dekker, Inc.

[11] Seber, G. (1977). Linear Regression Analysis. Canada, John Wiley & Sons, Inc.

List of Sites

Baguley, T. (2012). Pseudo-R2 and Related Measures. Online Supplement 4 to serious stats:

A guide to advanced statistics for the behavioral sciences.http://www.palgrave.com/psychology/ Baguley /students/supplements/9780230_577183_04_sup04.pdf [Accessed: December 9,2012]

Dayton, C.M. (1992).Logistic Regression Analysis.http:// bus.utk.edu /stat /datamining / Logistic%20Regression%20Analysis%20(Dayton). pdf [Accessed: November 12,2012].

Appendix

Table (A.1): The SAT Data Set

State	expenditure	teacher	Salary	Taking	Sat1	Sat2	Total sat
"Alabama"	4.405	17.2	31.144	8.0	491.0	538.0	1029.0
"Alaska"	8.963	17.6	47.951	47.0	445.0	489.0	934.0
"Arizona"	4.778	19.3	32.175	27.0	448.0	496.0	944.0
"Arkansas	4.459	17.1	28.934	6.0	482.0	523.0	1005.0
"Californ	4.992	24.0	41.078	45.0	417.0	485.0	902.0
"Colorado	5.443	18.4	34.571	29.0	462.0	518.0	980.0
"Connecti	8.817	14.4	50.045	81.0	431.0	477.0	908.0
"Delaware	7.03	16.6	39.076	68.0	429.0	468.0	897.0
"Florida"	5.718	19.1	32.588	48.0	420.0	469.0	889.0
"Georgia"	5.193	16.3	32.291	65.0	406.0	448.0	854.0
"Hawaii"	6.078	17.9	38.518	57.0	407.0	482.0	889.0
"Idaho"	4.21	19.1	29.783	15.0	468.0	511.0	979.0
"Illinois	6.136	17.3	39.431	13.0	488.0	560.0	1048.0
"Indiana"	5.826	17.5	36.785	58.0	415.0	467.0	882.0
"Iowa"	5.483	15.8	31.511	5.0	516.0	583.0	1099.0
"Kansas"	5.817	15.1	34.652	9.0	503.0	557.0	1060.0
"Kentucky	5.217	17.0	32.257	11.0	477.0	522.0	999.0
"Louisian	4.761	16.8	26.461	9.0	486.0	535.0	1021.0
"Maine"	6.428	13.8	31.972	68.0	427.0	469.0	896.0
"Maryland	7.245	17.0	40.661	64.0	430.0	479.0	909.0
"Massachu	7.287	14.8	40.795	80.0	430.0	477.0	907.0
"Michigan	6.994	20.1	41.895	11.0	484.0	549.0	1033.0
"Minnesot	6.0	17.5	35.948	9.0	506.0	579.0	1085.0
"Mississi	4.08	17.5	26.818	4.0	496.0	540.0	1036.0
"Missouri	5.383	15.5	31.189	9.0	495.0	550.0	1045.0
"Montana"	5.692	16.3	28.785	21.0	473.0	536.0	1009.0
"Nebraska	5.935	14.5	30.922	9.0	494.0	556.0	1050.0
"Nevada"	5.16	18.7	34.836	30.0	434.0	483.0	917.0
"New Ham	5.859	15.6	34.72	70.0	444.0	491.0	935.0
"New Jers	9.774	13.8	46.087	70.0	420.0	478.0	898.0
"New Mexi	4.586	17.2	28.493	11.0	485.0	530.0	1015.0
"New York	9.623	15.2	47.612	74.0	419.0	473.0	892.0
"North Ca	5.077	16.2	30.793	60.0	411.0	454.0	865.0
"North Da	4.775	15.3	26.327	5.0	515.0	592.0	1107.0
"Ohio"	6.162	16.6	36.802	23.0	460.0	515.0	975.0
"Oklahoma	4.845	15.5	28.172	9.0	491.0	536.0	1027.0
"Oregon"	6.436	19.9	38.555	51.0	448.0	499.0	947.0
"Pennsylv	7.109	17.1	44.51	70.0	419.0	461.0	880.0
"Rhode Is	7.469	14.7	40.729	70.0	425.0	463.0	888.0
"South Ca	4.797	16.4	30.279	58.0	401.0	443.0	844.0
"South Da	4.775	14.4	25.994	5.0	505.0	563.0	1068.0
"Tennesse	4.388	18.6	32.477	12.0	497.0	543.0	1040.0
"Texas"	5.222	15.7	31.223	47.0	419.0	474.0	893.0
"Utah"	3.656	24.3	29.082	4.0	513.0	563.0	1076.0

"Vermont"	6.75	13.8	35.406	68.0	429.0	472.0	901.0
"Virginia	5.327	14.6	33.987	65.0	428.0	468.0	896.0
"Washingt	5.906	20.2	36.151	48.0	443.0	494.0	937.0
"West Vir	6.107	14.8	31.944	17.0	448.0	484.0	932.0
"Wisconsi	6.93	15.9	37.746	9.0	501.0	572.0	1073.0
"Wyoming"	6.16	14.9	31.285	10.0	476.0	525.0	1001.0
"Alabama"	4.405	17.2	31.144	8.0	491.0	538.0	1029.0

Table (A.2): The Coronary Heart Disease data set

AGE	CHD	AGE in groups	Percent with CHD
20.0	0.0	1.0	0.1
23.0	0.0	1.0	
24.0	0.0	1.0	
25.0	0.0	1.0	
25.0	1.0	1.0	
26.0	0.0	1.0	
26.0	0.0	1.0	
28.0	0.0	1.0	
28.0	0.0	1.0	
29.0	0.0	1.0	
30.0	0.0	2.0	0.13
30.0	0.0	2.0	
30.0	0.0	2.0	
30.0	0.0	2.0	
30.0	0.0	2.0	
30.0	1.0	2.0	
32.0	0.0	2.0	
32.0	0.0	2.0	
33.0	0.0	2.0	
33.0	0.0	2.0	
34.0	0.0	2.0	
34.0	0.0	2.0	
34.0	1.0	2.0	
34.0	0.0	2.0	
34.0	0.0	2.0	
35.0	0.0	3.0	0.25
35.0	0.0	3.0	
36.0	0.0	3.0	
36.0	1.0	3.0	
36.0	0.0	3.0	
37.0	0.0	3.0	
37.0	1.0	3.0	
37.0	0.0	3.0	
38.0	0.0	3.0	
38.0	0.0	3.0	
39.0	0.0	3.0	
39.0	1.0	3.0	

40.0	0.0	4.0	
40.0	1.0	4.0	
41.0	0.0	4.0	
41.0	0.0	4.0	
42.0	0.0	4.0	
42.0	0.0	4.0	
42.0	0.0	4.0	
42.0	1.0	4.0	0.33
43.0	0.0	4.0	
43.0	0.0	4.0	
43.0	1.0	4.0	
44.0	0.0	4.0	
44.0	0.0	4.0	
44.0	1.0	4.0	
44.0	1.0	4.0	
45.0	0.0	5.0	
45.0	1.0	5.0	
46.0	0.0	5.0	
46.0	1.0	5.0	
47.0	0.0	5.0	
47.0	0.0	5.0	
47.0	1.0	5.0	0.46
48.0	0.0	5.0	
48.0	1.0	5.0	
48.0	1.0	5.0	
49.0	0.0	5.0	
49.0	0.0	5.0	
49.0	1.0	5.0	
50.0	0.0	6.0	
50.0	1.0	6.0	
51.0	0.0	6.0	
52.0	0.0	6.0	
52.0	1.0	6.0	0.62
53.0	1.0	6.0	
53.0	1.0	6.0	
54.0	1.0	6.0	
55.0	0.0	7.0	
55.0	1.0	7.0	
55.0	1.0	7.0	
56.0	1.0	7.0	
56.0	1.0	7.0	
56.0	1.0	7.0	
57.0	0.0	7.0	0.76
57.0	0.0	7.0	
57.0	1.0	7.0	
57.0	1.0	7.0	
57.0	1.0	7.0	
57.0	1.0	7.0	
58.0	0.0	7.0	

58.0	1.0	7.0	
58.0	1.0	7.0	
59.0	1.0	7.0	
59.0	1.0	7.0	
60.0	0.0	8.0	0.8
60.0	1.0	8.0	
61.0	1.0	8.0	
62.0	1.0	8.0	
62.0	1.0	8.0	
63.0	1.0	8.0	
64.0	0.0	8.0	
64.0	1.0	8.0	
65.0	1.0	8.0	
69.0	1.0	8.0	