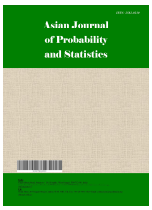

Asian Journal of Probability and Statistics

9(4): 29-47, 2020; Article no.AJPAS.62913
ISSN: 2582-0230

On the Study of Clear Causal Risk Factors of Diabetes Mellitus Using Multiple Regression

Ibrahim Abubakar Sadiq[1,2*] and Kode Komali[2]

[1]Department of Statistics, Ahmadu Bello University, Zaria, Nigeria.
[2]Department of Mathematics and Statistics, Mewar University, Rajasthan, India.

Authors' contributions

This work was carried out in collaboration between both the authors. Author IAS designed the study, performed the statistical analysis, wrote the protocol, managed the literature searches, managed the analyses of the study and wrote the first draft of the manuscript. Author KK managed the analyses of the study. Both the authors read and approved the final manuscript.

Article Information

DOI: 10.9734/AJPAS/2020/v9i430233
Editor(s):
(1) Dr. Ali Akgül, Siirt University, Turkey.
Reviewers:
(1) Oluchukwu Chukwuemeka Asogwa, Alex Ekwueme Federal University Ndufu-Alike Ikwo, Nigeria.
(2) Charles Okechukwu Aronu, Chukwuemeka Odumegwu Ojukwu University, Nigeria.
Complete Peer review History: http://www.sdiarticle4.com/review-history/62913

Study Protocol

Received: 06 September 2020
Accepted: 10 November 2020
Published: 25 November 2020

Abstract

The incurable lingering metabolic syndrome of diabetes mellitus is an up-surging global tricky with tremendous physical, social, mental, economics and health undesired ramifications. Three hundred and ninety four diabetic patients were measured on 4 baseline variable age (years), sex (Male=1 and Female=2), body mass index (kg/m^2) and blood pressure (mmHg). Blood sugar concentration (mg/dl) represented the response variable. The basic objective of this study is to verify the clear causal risk factors of diabetes. Both Multiple Linear Regression and Stepwise Regression techniques were applied on the data and the analysis showed that Body Mass Index (kg/m^2) and Blood Pressure (mmHg) are the clearest risk factors of diabetes. This justification served the same purpose in the procedure of variables selection used.

Keywords: Multiple linear regression; stepwise regression techniques; diabetes; baseline variable; risk factor; SPSS.

**Corresponding author: E-mail: abubakarsadiq463@gmail.com;*

1 Introduction

The incurable lingering metabolic syndrome of diabetes mellitus is an up-surging global tricky with tremendous physical, social, mental, economics and health undesired ramifications. It was estimated globally in 2010 that about 285 million people (approximately 6.4% of the adult population) suffered from this syndrome. This figure was estimated to upsurge to 430 million in the nonappearance of better regulator or medical remedy. Diabetes causal factors are different for each type of diabetes disorder. In this our study we examine four factors to discover which most contribution as clear fundamental risk factors for diabetes. The factors are Age (in years), Sex (Male=1 and Female=2), Body Mass Index (kg/m^2) and Blood Pressure (mmHg). The scope of this study is restricted to the application of regression analysis and stepwise techniques on Diabetes and its clear causal risk factors. This study will serve and function as a reference for subsequent researchers who want to carry out studies related to the topic. It will also serve as a reference point to support the Government in its tireless efforts to control the outbreak of disease and providing adequate drug for the diastase [1,2]. This study is structured and targeted to use regression analysis to determine what the clear causal risk factors are for Diabetes with specific study plan design. It is requisite selecting the statistical most significant variables as of a limited group of independent variables. Generally, fitting a regression model of the relationship between our mentioned response variable and independent variables is considered. Verification and statistical examination which among these variables is the clearest causal risk factor for diabetes.

A regression analysis is simple academically applied statistical tools and techniques used for investigating mathematical relationship between variables [3]. It is usually conventional method used to build a mathematical model to associate response variables to predictor variables [4]. A Multiple Linear Regression is one and only of the greatest frequently applied data mining practices and can be responsible for insightful information in circumstances where the severe assumptions allied with multiple linear regression methods remain bump into [5]. A multiple linear regression is a very resourceful device and can be employed to just about any advances, schemes, or field of study [6]. Considerably vast availability has existed in published on the subject of this similar topic, and reader of interest can refer to [7,8] for more details knowledge.

The most fundamental step in emerging a suitable multiple linear regression model is by choosing a process of model construction and setting up the best criteria of the model [9]. The familiarized stepwise regression, is usually applied to model building. A stepwise regression is a programmed technique that selects the important statistical substantial variables as of a limited group of explanatory or independent variables. The three different ways for executing the stepwise regression techniques include mixed, backward and forward selection. The mixed selection techniques is the greatest and strong statistically type of stepwise regression which is conglomeration of the backward and forward techniques as indicated in [10,11].

Equally noted from [10] that the model justification is the concluding step in a regression modeling structure practice. Additionally, it remained highlighted in that for which there exist three leading approaches attached for with model validation includes, collection of new observations to authenticate the existing model and the aforementioned likelihood [12]. Distinction of existing outcomes through other hypothetical standards, experimental and simulation of effects [13]. Practice of a cross-validation representable sample to verify and evaluate the predictive power of the contemporary model [14].

The methodology of cross-validation are employed to measure the strength and certainty for prediction of the build regression models, meanwhile, a positive quantity of the data values are detached from the model building procedure for instance fourteen registered observation, and formerly practices the built model to estimate their calculated values [15]. The universal rule of thumb trendy regression model architecting is to employ eighty per cent (80%) of the data set for the improvement of the working out model and the outstanding twenty per cent (20%) main for justification of the model as illustrious in [10]. The validation histories can be carefully chosen at random beginning the whole data set and or in some the instances of time series data, required, the justification set may be able to be the greatest in progress 20%. Satisfactory, many

regression model are plausible hypothesised to produce reasonable estimates which are close to the actual data significance. Numerous available statistic are used to help in the measuring of predictive power of the built statistical regression model. The vast well known statistic is the root mean square residuals [11]. The statistic is work out by computing the square root of the residuals sum squares for the awaiting histories at odds of the matching degrees of freedom. The lesser the root mean square residuals probability values point out better and healthier the model predictability. Additional equivalent model endorsement indicator is the classical empirical coefficient (R^2 or adjusted R^2) statistic [16]. These figures is calculated for the awaiting histories trial, and provides certain insightful into the predictability of the model. Through interpretations, greater R^2 values are desired, which mean, the R^2 statistic signposts the total of variation described by the independent variables in the regression model [17,18].

2 Materials and Methods

The data used for this research is a secondary data sourced from a study directed by [19] "Least Angle Regression.". Three hundred and ninety four (394) diabetic patients were measured on 4 baseline variables [19]. There are five (5) variables in the model, one is dependent variable and four (4) others are independent variables. The dependent variable $y = BSC(mg/dl)$. The others four independent variables are $x_1 = Age$, $x_2 = Sex(m=1, f=2)$, $x_3 = BMI(kg/m^2)$ and $x_4 = BP(mmHg)$. The sex variable where categorically coded as $(Male = 1, Female = 2)$. We use multiple linear regression and stepwise techniques in modelling the sourced data and equally the reduced model.

2.1 Parameter estimation in multiple regression model

The general method of least squares was used in parameter estimation of regression coefficients of multiple regression model by the below given equation [20,17].

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik} + \varepsilon_i \tag{1}$$

Suppose that we have an $n > k$ observations are presented, and suppose that x_{ij} denote the i^{th} observation of variable x_j .The observations of the data set are could assumed $(x_{i1}, x_{i2}, ..., x_{ik}, y_i)$, where $i = 1, 2, ..., n$ and $n > k$. It traditional to present the multiple regression data in a table here upon such, we cut it short, each observation of our data $(x_{i1}, x_{i2}, ..., x_{ik}, y_i)$, satisfies the equation model in (1) above or

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik} + \varepsilon_i \\ &= \beta_0 + \sum_{i=1}^{n} \varepsilon_i^{\,2} = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{i=1}^{n} \beta_j x_{ij} \right)^2, \quad \begin{matrix} i = 1, 2, ..., n \\ j = 1, 2, ..., k \end{matrix} \end{aligned} \tag{2}$$

The least square function is

$$L = \sum_{i=1}^{n} \varepsilon_i^{\,2} = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{k} \beta_{ij} x_{ij} \right)^2 \tag{3}$$

We minimize L with respect to $\beta_0, \beta_1, ..., \beta_k$. The least square estimates of $\beta_0, \beta_1, ..., \beta_k$ need to satisfy

$$\frac{\partial L}{\partial \hat{\beta}_0} = -2\sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \sum_{j=1}^{k} \hat{\beta}_j x_{ij} \right) = 0 \tag{4}$$

$$\frac{\partial L}{\partial \hat{\beta}_k} = -2\sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \sum_{j=1}^{k} \hat{\beta}_j x_{ij} \right) x_{ij} = 0 \tag{5}$$

Simplifying equation (4) and (5), we get the normal equations of the least squares.

$$\begin{aligned}
&n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_{i1} + \hat{\beta}_2 \sum_{i=1}^{n} x_{i2} + \ldots + \hat{\beta}_k \sum_{i=1}^{n} x_{ik} = \sum_{i=1}^{n} y_i \\
&\hat{\beta}_0 \sum_{i=1}^{n} x_{i1} + \hat{\beta}_1 \sum_{i=1}^{n} x_{i1}^2 + \hat{\beta}_2 \sum_{i=1}^{n} x_{i1}x_{i2} + \ldots + \hat{\beta}_k \sum_{i=1}^{n} x_{i1}x_{ik} = \sum_{i=1}^{n} x_{i1}y_i \\
&\cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \\
&\hat{\beta}_0 \sum_{i=1}^{n} x_{ik} + \hat{\beta}_1 \sum_{i=1}^{n} x_{ik}x_{i1} + \hat{\beta}_2 \sum_{i=1}^{n} x_{ik}x_{i2} + \ldots + \hat{\beta}_k \sum_{i=1}^{k} x_{ik}^2 = \sum_{i=1}^{n} x_{ik}y_i
\end{aligned} \tag{6}$$

Suppose that we have k independent variables with n observations, $(x_{i1}, x_{i2}, \ldots, x_{ik}, y_i)$, $i = 1, 2, \ldots, n$ and that the model relating the independent variables to the response variable is

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} + \varepsilon_i \tag{7}$$

This model is a system of n equations that can be expressed in matrix notation as

$$y = X\beta + \varepsilon \tag{8}$$

Where

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \qquad X = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1n} \\ 1 & x_{21} & x_{22} & \cdots & x_{2n} \\ 1 & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nn} \end{pmatrix} \qquad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} \text{ and } \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix} \tag{9}$$

Generally, y is a vector $(n \times 1)$ of the observations and X is a matrix of $(n \times p)$ is an $(n \times p)$ of the independent variables, β is a vector of $(p \times 1)$ for the regression coefficients, and ε is an $(n \times 1)$ vector of random errors.

We obtain the vector of least squares estimators, $\hat{\beta}$, that minimizes the sum of square of deviation between observed and predicted y (error), we use the equation

$$\hat{\beta} = (X'X)^{-1}(X'y) \tag{10}$$

Is the required least squares estimator for β in the regression model.

By using matrix notation, the estimate is obtained as

The matrix notation of the fitted model is $\hat{y} = X\hat{\beta}$ and the fitted regression model is

$$\begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_k \end{pmatrix} = \begin{pmatrix} n & \sum_{i=1}^{n} x_{i1} & \sum_{i=1}^{n} x_{i2} & \cdots & \sum_{i=1}^{n} x_{ik} \\ \sum_{i=1}^{n} x_{i1} & \sum_{i=1}^{n} x_{i1}^2 & \sum_{i=1}^{n} x_{i1}x_{i2} & \cdots & \sum_{i=1}^{n} x_{i1}x_{ik} \\ \vdots & \vdots & \vdots & & \vdots \\ \sum_{i=1}^{n} x_{ik} & \sum_{i=1}^{n} x_{ik}x_{i1} & \sum_{i=1}^{n} x_{ik}x_{i2} & \cdots & \sum_{i=1}^{n} x_{ik}^2 \end{pmatrix}^{-1} \begin{pmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_{i1}y_i \\ \vdots \\ \sum_{i=1}^{n} x_{ik}y_i \end{pmatrix} \tag{11}$$

$$\hat{y} = \hat{\beta}_0 + \sum_{j=1}^{k} \hat{\beta}_j x_{ij}, \quad \begin{matrix} i = 1, 2, ..., n \\ j = 1, 2, ..., k \end{matrix} \tag{12}$$

2.2 Analysis of variance (ANOVA) on regression

The model adequacy is tested using analysis of variance (ANOVA) which is summarized in below table.

Table1. ANOVA for Testing Significance of Regression in Multiple Regression

Source of variation	Sum of Squares	Degrees of freedom	Mean square	F_{cal}
Regression	$SS_R = \hat{\beta}'(X'y) - n\bar{y}^2$	k	$MS_R = SS_R / k$	MS_R / MS_E
Residual (error)	$SS_E = SS_T - SS_R$	$n-k-1$	$MS_E = SS_E / n-k-1$	
Total	$SS_T = y'y - n\bar{y}^2$	$n-1$		

Where k = number of parameters, n = number of observations

The null hypothesis that the model is insignificant is rejected if F_{cal} is greater than $F_{\alpha(k),(n-p)}$

Estimating δ^2

There is essentially additional unknown parameter in our regression model, δ^2 (the variation of the error term ε) [21,22]. The residuals $e_i = y_i - \hat{y}_i$ are used to obtain an estimate of δ^2. The sum of square of the residual often called the error sum of square is

$$SS_E = \sum_{I=1}^{n} e_i^2 = \sum_{i=1}^{n} \left(y_i - \hat{y}_i \right)^2 \tag{13}$$

We can clearly displays that the expected value of the error sum of square is given by

$$E\left(SS_E\right)=\left(n-p\right)\ \delta^2 \tag{14}$$

$$\hat{\delta}^2=\frac{\sum_{i=1}^{n}e_i^2}{n-p}-\frac{SS_E}{n-p} \tag{15}$$

2.3 The coefficient of multiple determination (R^2)

Since the proportion of disparity of the dependent variable which exactly explained by the explanatory variables in regression analysis is termed as the coefficient of determination and is given by

$$R^2=\frac{SS_R}{SS_T}=1-\frac{SS_E}{SS_T} \tag{16}$$

$$R^2_{adj}=1-\frac{SS_E/\left(n-k-1\right)}{SS_T/\left(n-1\right)} \tag{17}$$

3 Results and Discussion

3.1 Interpretation of coefficients of determination

The result of $\mathbf{R^2}$ implies that 38.4% of Diabetes explained by variation shown that age, sex, BP and BMI are the clearer causal risk factors while the remaining 62% is due to other factors not considered. Also The result of **Adjusted $\mathbf{R^2}$** implies that 37.8% of Diabetes explained by variation shown that age, sex, blood pressure and body mass index are the clearer causal risk factors while the remaining 62% is due to other factors not considered.

Table 2. The model summary

Model	R	R Square	Adjusted R Square	Estimate of Standard Error
1	.620[a]	.384	.378	60.77111

a. Predictors: (Constant), Blood Pressure, Sex, Age, Body Mass Index

Table 3. The ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	893586.334	4	223396.583	60.490	.000[b]
	Residual	1432933.748	388	3693.128		
	Total	2326520.081	392			

a. Dependent Variable: Diabetes

b. Predictors: (Constant), Blood Pressure, Sex, Age, Body Mass Index

3.2 Interpretations of ANOVA

The output in Table 3 displays the results of a fitted multiple linear regression model which describe the relationship between Diabetes and 4 independent variables Age, Sex, Body Mass Index and Blood Pressure.

Since the Probability value in Table 3 is lesser than 0.05, it is clearly that, exists a statistically significant relationship between the variables at the 95.0% confidence level. The R-Squared statistic indicates that the model as fitted explains 38.4% of the variability in Diabetes. Our adjusted R-squared result is more appropriate for comparing models with distinct quantities of independent variables, however is approximately 37.8%. Our estimated standard indicates the residual standard deviation to be 60.77111. This value can be used to construct prediction limits for new observations. In determining whether the model can be simplified, notice that the highest P-value on the independent variables is 0.714, belonging to AGE. Meanwhile the probability value is slightly greater than 0.05, the statistic is not statistically significant at the 95.0% or higher confidence level. Consequently, you should consider removing AGE from the model.

Table 4. The Coefficients[a]

		B	Std. Error	Beta		
1	(Constant)	-193.290	24.821		-7.787	.000
	Age	.091	.249	.016	.367	.714
	Sex	-9.333	6.326	-.061	-1.475	.141
	Body Mass Index	8.617	.762	.490	11.309	.000
	Blood Pressure	1.348	.258	.239	5.225	.000

a. Dependent Variable: Diabetes

$$y = -193.290 + 0.091x_1 - 9.333x_2 + 8.617x_3 + 1.348x_4 \quad (18)$$

The regression equation above can be used to estimate or predict the Diabetes (y) based on known Age (x_1), Sex (x_2), Body Mass Index (x_3) and Blood Pressure (x_4) values.

3.3 Regression Equation Result Discussion

The equation (18) and Table 4 represent the regression equation which can be interpreted thus. Any increase in age, diabetes will increase by 0.091, while other variables Sex, Body Mass Index and Blood Pressure remained constants. Any increase in sex, diabetes will decrease by 9.333, while other variables Age, Body Mass Index and Blood Pressure remained constants. Any increase in body mass index the diabetes will increase by 8.617, while other variables Age, Sex and Blood Pressure remained constants and any increase in blood pressure, diabetes will increase by 1.348, while other variables Age, Sex and Body Mass Index remained constants.

Table 5. The reduced model summary

Model	R	R Square	Adjusted R Square	Estimate of Standard Error
1	.579[a]	.335	.334	62.88168
2	.617[b]	.381	.377	60.78919

a. Predictors: (Constant), Body Mass Index
b. Predictors: (Constant), Body Mass Index, Blood Pressure

3.4 Interpretation of reduce and non-reduced model coefficient of determination

The result of $\mathbf{R^2}$ above implies that at first stage 33.5% of Diabetes explained by variation shown that body mass index is the clear causal risk factor while the remaining 66.5% is due to other factors not considered. At the second stage, 38.1% of Diabetes explained by variation shown that body mass index and the blood pressure are the clearest causal risk factors while the remaining 61.9% is due to other factors not considered. Also the result of adjusted $\mathbf{R^2}$ above implies that at first stage 33.4% of Diabetes explained by variation shown that body mass index is the clear causal risk factor while the remaining 66.6% is due to other factors not considered. At the second stage 37.7% of Diabetes explained by variation shown that body mass index

and the blood pressure are the clearest causal risk factors while the remaining 62.3% is due to other factors not considered.

Table 6. The reduced ANOVA[a]

Model		Sum of Squares	DF	Mean Square	F	Sig.
1	Regression	780464.563	1	780464.563	197.381	.000[b]
	Residual	1546055.518	391	3954.106		
	Total	2326520.081	392			
2	Regression	885343.272	2	442671.636	119.792	.000[c]
	Residual	1441176.809	390	3695.325		
	Total	2326520.081	392			

a. Dependent Variable: Diabetes
b. Predictors: (Constant), Body Mass Index
c. Predictors: (Constant), Body Mass Index, Blood Pressure

3.5 Interpretation

The production of the above Table 6 displays the results of fitting a multiple linear regression model to describe the relationship between Diabetes and 4 independent variables. Since the probability value in the above ANOVA table is lesser than 0.05, thus there exist a statistical significant association among the variables at 95.0% confidence level. Also the coefficient of determination figure at first step indicates that the fitted model explains 33.5% of the variability in Diabetes. The adjusted coefficient of determination figure is an indication of suitability on comparing models with dissimilar quantities of explanatory variables record about 33.4%. The estimate of the standard error is a confirmations that 62.88168 the standard variability residuals. The value is used to construct prediction limits for new observations. At second step the R-Squared statistic indicates that the reduced fitted model accounts for 38.1% of the variability on the response variable (Diabetes). However, adjusted coefficient of determination figure is an indication of suitability on comparing models with dissimilar quantities of explanatory variables record about 37.7% using stepwise techniques. The estimate of the standard error is a confirmations that 60.78919 the standard variability residuals. Also this value can be used to construct prediction limits for new observations for the reduced model. In determining whether the model can be simplified, notice that the highest P-value on the independent variables is 0.000, belonging to BP. Meanwhile the probability value is lesser than 0.05, that term is statistically significant at the 95.0% confidence level. Consequently, we probably rich a final step of stepwise analysis. For this reason, further removal any variables from the model is having zero possibility statistically.

Table 7. The Coefficients[a] of the reduced model

Model		Unstandardized Coefficients		Standardized Coefficients	T-statistic	Sig.
		B	Std. Error	Beta		
1	(Constant)	-116.526	19.371		-6.015	.000
	Body Mass Index	10.193	.726	.579	14.049	.000
2	(Constant)	-198.229	24.205		-8.190	.000
	Body Mass Index	8.629	.760	.490	11.348	.000
	Blood Pressure	1.298	.244	.230	5.327	.000

a. Dependent Variable: Diabetes

4 Overall Discussion

The standard multiple regression technique was employed the results described the relationship between Diabetes and 4 independent variables, Sex, Age, Body Mass Index and Blood Pressure. Subsequently the

probability value from Table 3 is lesser than 0.05 level of confidence, for this reason we discovered that, exist is a significant statistical associations among the study variables at the 95.0% confidence level. The R-Squared statistic indicates that the model as fitted explains 38.4% of the variability in Diabetes. Our adjusted R-squared result is more appropriate for comparing models with distinct quantities of independent variables, however is approximately 37.8%. Our estimated standard indicates the residual standard deviation to be 60.77111. This value can be used to construct prediction limits for new observations. In determining whether the model can be simplified, notice that the highest P-value on the independent variables is 0.714, belonging to AGE. Then the probability value is higher than 0.05, such higher value is not under statistical conclusion possibility from 95.0% or higher confidence level. Consequently, it is considerable efficient facts for removing AGE from the model. The fitted multiple regression model is presented by the below equation as earlier discovered from equation (18) in of similar kinds.

$$y = -193.290 + 0.091x_1 - 9.333x_2 + 8.617x_3 + 1.348x_4 \tag{19}$$

Furthermore, our regression model in equation (18) or (19) can surely be applicable in controlling and predicting the response variable (Diabetes) on basis of well acknowledged value of the four study independent variables referred. Thus, our built regression equation can be interpreted to mean any increase of age (in years), diabetes will increase by 0.091mg/dl while others independent variables remained constant. Any change in sex (male or female), diabetes will decrease by 9.333mg/dl while remaining others independent variables remained constant. Any increase in BMI (kg/m^2), the diabetes will increase by 8.617mg/dl while others independent variables remained constant. Any increase in BP (mmHg), diabetes will increase by 1.348mg/dl while others independent variables remained constant.

A stepwise regression technique was employed to selects most statistically significant variable. Also the coefficient of determination figure at first step indicates that the fitted model explains 33.5% of the variability in Diabetes. The adjusted coefficient of determination figure is an indication of suitability on comparing models with dissimilar quantities of explanatory variables record about 33.4%. The estimate of the standard error is a confirmations that 62.88168 the standard variability residuals. The value is used to construct prediction limits for new observations. At second step the R-Squared statistic indicates that the reduced fitted model accounts for 38.1% of the variability on the response variable (Diabetes). However, adjusted coefficient of determination figure is an indication of suitability on comparing models with dissimilar quantities of explanatory variables record about 37.7% using stepwise techniques. The estimate of the standard error is a confirmations that 60.78919 the standard variability residuals. Also this value can be used to construct prediction limits for new observations for the reduced model. In determining whether the model can be simplified, notice that the highest P-value on the independent variables is 0.000, belonging to BP. Meanwhile the probability value is lesser than 0.05, that term is statistically significant at the 95.0% confidence level. Consequently, we probably rich a final step of stepwise analysis. For this reason, further removal any variables from the model is having zero possibility statistically. **Thus the reduced model was:**

$$y = -198.229 + 8.629x_3 + 1.298x_4 \tag{20}$$

5 Conclusion

Diabetes is emerging as a major global health problem with the number of people living with diabetes expected to rise to 380 million by 2025. Approximately 10% of this population will have T1DM characterised by the progressive loss of b cells and complete insulin deficiency. The remaining 90% of the population will have T2DM characterised by insulin resistance and impaired insulin secretion. Although current management and treatment strategies are able to help patients with diabetes, new efficient treatments are needed. This study demonstrates that multiple linear regression and stepwise regression can be used to assess predictor variables influencing the risk of diabetes in adult patients.394 diabetes patients were measured on 4 baseline variables these are age, BP, BMI and gender. In general, the multiple regression and

stepwise techniques reveals that the fundamental findings of this paper was clear indication on BMI and BP among others are the most risk factors for diabetes.

Competing Interests

Authors have declared that no competing interests exist.

References

[1] Turlach B, Venables W. Simultaneous variable selection. Technometrics. 2005;3(7):349-363.

[2] Akgül A. Reproducing kernel Hilbert space method based on reproducing kernel functions for investigating boundary layer flow of a Powell–Eyring non-Newtonian fluid. Journal of Taibah University for Science. 2019;13(1):858-863.

[3] Harell H. Multivariate prognostic models, issues in developing models, evaluating assumptions and the adequacy and measuring and reducing errors. Statiscal Medical. 1996;15(10):361-387.

[4] Hesterverg T. Least angle and L1 penalized regression: A review. Statistical Survey. 2008;2(3):61-93.

[5] Hoerl AE, Kennard R. Ridge regression. Bias Estimation for Non Orthogonol Problems. 1970;44:55-67.

[6] Hurvich C, Tsai C. The impact of model selection on inference in linear regression. American Statistician. 1990;44:214-217.

[7] Legendre MA. Nouvelles methodes pour la determination desorbites des cometes. Paris: Kendall Publishing; 1833.

[8] Adams L. Acomputer experiment to evaluate regression analysis. Proceedings of the statistical computing. 1990;55-62.

[9] Andre N, Young TM. Online monitoring of the buffer capacity of practical board furnish by near-infrared spectroscopy. Applied spectroscopy. 2006;3(10):1204-1209.

[10] Beer DG, Kardia SL. Gene-expression profiles predict survival of patients with lung adenocarcinoma. National Medical. 2002;8(2):816-824.

[11] Bobelstad HM. Predicting survival from micro-array data a comparative study. Bioinformatics. 2007;23(12):2080-2087.

[12] Breiman L, Friedman J. Predicting multiple responses in multiple linear regression (with discussion). Journal of the royal Statiatics Society, Series B. 1997;59:3-54.

[13] Gauss CF. Theoria motus corporum coelestium in sectionibus conicis. Solem ambientum: Blackwell Publishing; 1777-1855.

[14] Daper NR, Smith H. Applied regression analysis, second edition. New York: John Wiley and Sons; 1981.

[15] DCC. Management of type 2 Diabetes: Evolving strategies for the treatment of patients with type 2 Diabetes. Metabolism. 2008;2(4):1-23.

[16] Kutner MH, Nachtsheim CJ, Neter J, Li W. Applied linear statistical models (fifth edition). New York. PP.300-321.: McGraw-Hill; 2004.

[17] Montgomery DC, Runger GC. Applied Statistics and Probability for Engineers (3rd ed.). (W. Anderson, & J. Welter, Eds.) New York: John Wiley & Sons, Inc; 2003.

[18] Roecker E. Prediction error and its estimation for subset-selection models. Technometrics. 1991;33: 459-468.

[19] Efron et al. Least Angle Regrssion. New York: John Wiley and Sons; 2004.

[20] Efroymson MA. Multiple regression analysis. New York: John Wiley and Sons; 1960.

[21] Myers RH. Classical and modern regression with applications. Boston: FWS-Kent Publishing Company; 1990.

[22] Neter J, Kunter MH. Applied linear regression models third edition. Chicago: Illinois/Irwin; 1996.

Appendix

S/NO.	AGE	SEX(Male=2,Female=1)	BMI	BP	Diabetes
S/N/1	**59**	2	32.1	101	151
S/N/2	**48**	1	21.6	87	75
S/N/3	**72**	2	30.5	93	141
S/N/4	**24**	1	25.3	84	206
S/N/5	**50**	1	23	101	135
S/N/6	**23**	1	22.6	89	97
S/N/7	**36**	2	22	90	138
S/N/8	**66**	2	26.2	114	63
S/N/9	**60**	2	32.1	83	110
S/N/10	**29**	1	30	85	310
S/N/11	**22**	1	18.6	97	101
S/N/12	**56**	2	28	85	69
S/N/13	**53**	1	23.7	92	179
S/N/14	**50**	2	26.2	97	185
S/N/15	**61**	1	24	91	118
S/N/16	**34**	2	24.7	118	171
S/N/17	**47**	1	30.3	109	166
S/N/18	**68**	2	27.5	111	144
S/N/19	**38**	1	25.4	84	97
S/N/20	**41**	1	24.7	83	168
S/N/21	**35**	1	21.1	82	68
S/N/22	**25**	2	24.3	95	49
S/N/23	**25**	1	26	92	68
S/N/24	**61**	2	32	103.67	245
S/N/25	**31**	1	29.7	88	184
S/N/26	**30**	2	25.2	83	202
S/N/27	**19**	1	19.2	87	137
S/N/28	**42**	1	31.9	83	85
S/N/29	**63**	1	24.4	73	131

S/NO.	AGE	SEX(Male=2,Female=1)	BMI	BP	Diabetes
S/N/30	**67**	2	25.8	113	283
S/N/31	**32**	1	30.5	89	129
S/N/32	**42**	1	20.3	71	59
S/N/33	**58**	2	38	103	341
S/N/34	**57**	1	21.7	94	87
S/N/35	**53**	1	20.5	78	65
S/N/36	**62**	2	23.5	80.33	102
S/N/37	**52**	1	28.5	110	265
S/N/38	**46**	1	27.4	78	276
S/N/39	**48**	2	33	123	252
S/N/40	**48**	2	27.7	73	90
S/N/41	**50**	2	25.6	101	100
S/N/42	**21**	1	20.1	63	55
S/N/43	**32**	2	25.4	90.33	61
S/N/44	**54**	1	24.2	74	92
S/N/45	**61**	2	32.7	97	259
S/N/46	**56**	2	23.1	104	53
S/N/47	**33**	1	25.3	85	190
S/N/48	**27**	1	19.6	78	142
S/N/49	**67**	2	22.5	98	75
S/N/50	**37**	2	27.7	93	142
S/N/51	**58**	1	25.7	99	155
S/N/52	**65**	2	27.9	103	225
S/N/53	**34**	1	25.5	93	59
S/N/54	**46**	1	24.9	115	104
S/N/55	**35**	1	28.7	97	182
S/N/56	**37**	1	21.8	84	128
S/N/57	**37**	1	30.2	87	52
S/N/58	**41**	1	20.5	80	37
S/N/59	**60**	1	20.4	105	170
S/N/60	**66**	2	24	98	170
S/N/61	**29**	1	26	83	61
S/N/62	**37**	2	26.8	79	144
S/N/63	**41**	2	25.7	83	52
S/N/64	**39**	1	22.9	77	128
S/N/65	**67**	2	24	83	71
S/N/66	**36**	2	24.1	112	163
S/N/67	**46**	2	24.7	85	150
S/N/68	**60**	2	25	89.67	97
S/N/69	**59**	2	23.6	83	160
S/N/70	**53**	1	22.1	93	178
S/N/71	**48**	1	19.9	91	48
S/N/72	**48**	1	29.5	131	270
S/N/73	**66**	2	26	91	202
S/N/74	**52**	2	24.5	94	111
S/N/75	**52**	2	26.6	111	85
S/N/76	**46**	2	23.5	87	42
S/N/77	**40**	2	29	115	170

S/NO.	AGE	SEX(Male=2,Female=1)	BMI	BP	Diabetes
S/N/78	**22**	1	23	73	200
S/N/79	**50**	1	21	88	252
S/N/80	**20**	1	22.9	87	113
S/N/81	**68**	1	27.5	107	143
S/N/82	**52**	2	24.3	86	51
S/N/83	**44**	1	23.1	87	52
S/N/84	**38**	1	27.3	81	210
S/N/85	**49**	1	22.7	65.33	65
S/N/86	**61**	1	33	95	141
S/N/87	**29**	2	19.4	83	55
S/N/88	**61**	1	25.8	98	134
S/N/89	**34**	2	22.6	75	42
S/N/90	**36**	1	21.9	89	111
S/N/91	**52**	1	24	83	98
S/N/92	**61**	1	31.2	79	164
S/N/93	**43**	1	26.8	123	48
S/N/94	**35**	1	20.4	65	96
S/N/95	**27**	1	24.8	91	90
S/N/96	**29**	1	21	71	162
S/N/97	**64**	2	27.3	109	150
S/N/98	**41**	1	34.6	87.33	279
S/N/99	**49**	2	25.9	91	92
S/N/100	**48**	1	20.4	98	83
S/N/101	**53**	1	28	88	128
S/N/102	**53**	2	22.2	113	102
S/N/103	**23**	1	29	90	302
S/N/104	**65**	2	30.2	98	198
S/N/105	**41**	1	32.4	94	95
S/N/106	**55**	2	23.4	83	53
S/N/107	**22**	1	19.3	82	134
S/N/108	**56**	1	31	78.67	144
S/N/109	**54**	2	30.6	103.33	232
S/N/110	**59**	2	25.5	95.33	81
S/N/111	**60**	2	23.4	88	104
S/N/112	**54**	1	26.8	87	59
S/N/113	**25**	1	28.3	87	246
S/N/114	**54**	2	27.7	113	297
S/N/115	**55**	1	36.6	113	258
S/N/116	**40**	2	26.5	93	229
S/N/117	**62**	2	31.8	115	275
S/N/118	**65**	1	24.4	120	281
S/N/119	**33**	2	25.4	102	179
S/N/120	**53**	1	22	94	200
S/N/121	**35**	1	26.8	98	200
S/N/122	**66**	1	28	101	173
S/N/123	**62**	2	33.9	101	180
S/N/124	**50**	2	29.6	94.33	84
S/N/125	**47**	1	28.6	97	121

S/NO.	AGE	SEX(Male=2,Female=1)	BMI	BP	Diabetes
S/N/126	**47**	2	25.6	94	161
S/N/127	**24**	1	20.7	87	99
S/N/128	**58**	2	26.2	91	109
S/N/129	**34**	1	20.6	87	115
S/N/130	**51**	1	27.9	96	268
S/N/131	**31**	2	35.3	125	274
S/N/132	**22**	1	19.9	75	158
S/N/133	**53**	2	24.4	92	107
S/N/134	**37**	2	21.4	83	83
S/N/135	**28**	1	30.4	85	103
S/N/136	**47**	1	31.6	84	272
S/N/137	**23**	1	18.8	78	85
S/N/138	**50**	1	31	123	280
S/N/139	**58**	2	36.7	117	336
S/N/140	**55**	1	32.1	110	281
S/N/141	**60**	2	27.7	107	118
S/N/142	**41**	1	30.8	81	317
S/N/143	**60**	2	27.5	106	235
S/N/144	**40**	1	26.9	92	60
S/N/145	**57**	2	30.7	90	174
S/N/146	**37**	1	38.3	113	259
S/N/147	**40**	2	31.9	95	178
S/N/148	**33**	1	35	89	128
S/N/149	**32**	2	27.8	89	96
S/N/150	**35**	2	25.9	81	126
S/N/151	**55**	1	32.9	102	288
S/N/152	**49**	1	26	93	88
S/N/153	**39**	2	26.3	115	292
S/N/154	**60**	2	22.3	113	71
S/N/155	**67**	2	28.3	93	197
S/N/156	**41**	2	32	109	186
S/N/157	**44**	1	25.4	95	25
S/N/158	**48**	2	23.3	89.33	84
S/N/159	**45**	1	20.3	74.33	96
S/N/160	**47**	1	30.4	120	195
S/N/161	**46**	1	20.6	73	53
S/N/162	**36**	2	32.3	115	217
S/N/163	**34**	1	29.2	73	172
S/N/164	**53**	2	33.1	117	131
S/N/165	**61**	1	24.6	101	214
S/N/166	**37**	1	20.2	81	59
S/N/167	**33**	2	20.8	84	70
S/N/168	**68**	1	32.8	105.67	220
S/N/169	**49**	2	31.9	94	268
S/N/170	**48**	1	23.9	109	152
S/N/171	**55**	2	24.5	84	47
S/N/172	**43**	1	22.1	66	74
S/N/173	**60**	2	33	97	295

S/NO.	AGE	SEX(Male=2,Female=1)	BMI	BP	Diabetes
S/N/174	**31**	2	19	93	101
S/N/175	**53**	2	27.3	82	151
S/N/176	**67**	1	22.8	87	127
S/N/177	**61**	2	28.2	106	237
S/N/178	**62**	1	28.9	87.33	225
S/N/179	**60**	1	25.6	87	81
S/N/180	**42**	1	24.9	91	151
S/N/181	**38**	2	26.8	105	107
S/N/182	**62**	1	22.4	79	64
S/N/183	**61**	2	26.9	111	138
S/N/184	**61**	2	23.1	113	185
S/N/185	**53**	1	28.6	88	265
S/N/186	**28**	2	24.7	97	101
S/N/187	**26**	2	30.3	89	137
S/N/188	**30**	1	21.3	87	143
S/N/189	**50**	1	26.1	109	141
S/N/190	**48**	1	20.2	95	79
S/N/191	**51**	1	25.2	103	292
S/N/192	**47**	2	22.5	82	178
S/N/193	**64**	2	23.5	97	91
S/N/194	**51**	2	25.9	76	116
S/N/195	**30**	1	20.9	104	86
S/N/196	**56**	2	28.7	99	122
S/N/197	**42**	1	22.1	85	72
S/N/198	**56**	2	28.7	99	122
S/N/199	**42**	1	22.1	85	72
S/N/200	**62**	2	26.7	115	129
S/N/201	**34**	1	31.4	87	142
S/N/202	**60**	1	22.2	104.67	90
S/N/203	**64**	1	21	92.33	158
S/N/204	**39**	2	21.2	90	39
S/N/205	**71**	2	26.5	105	196
S/N/206	**48**	2	29.2	110	222
S/N/207	**79**	2	27	103	277
S/N/208	**40**	1	30.7	99	99
S/N/209	**49**	2	28.8	92	196
S/N/210	**51**	1	30.6	103	202
S/N/211	**57**	1	30.1	117	155
S/N/212	**59**	2	24.7	114	77
S/N/213	**51**	1	27.7	99	191
S/N/214	**74**	1	29.8	101	70
S/N/215	**67**	1	26.7	105	73
S/N/216	**49**	1	19.8	88	49
S/N/217	**57**	1	23.3	88	65
S/N/218	**56**	2	35.1	123	263
S/N/219	**52**	2	29.7	109	248
S/N/220	**69**	1	29.3	124	296
S/N/221	**37**	1	20.3	83	214

S/NO.	AGE	SEX(Male=2,Female=1)	BMI	BP	Diabetes
S/N/222	**24**	1	22.5	89	185
S/N/223	**55**	2	22.7	93	78
S/N/224	**36**	1	22.8	87	93
S/N/225	**42**	2	24	107	252
S/N/226	**21**	1	24.2	76	150
S/N/227	**41**	1	20.2	62	77
S/N/228	**57**	2	29.4	109	208
S/N/229	**20**	2	22.1	87	77
S/N/230	**67**	2	23.6	111.33	108
S/N/231	**34**	1	25.2	77	160
S/N/232	**41**	2	24.9	86	53
S/N/233	**38**	2	33	78	220
S/N/234	**51**	1	23.5	101	154
S/N/235	**52**	2	26.4	91.33	259
S/N/236	**67**	1	29.8	80	90
S/N/237	**61**	1	30	108	246
S/N/238	**67**	2	25	111.67	124
S/N/239	**56**	1	27	105	67
S/N/240	**64**	1	20	74.67	72
S/N/241	**58**	2	25.5	112	257
S/N/242	**55**	1	28.2	91	262
S/N/243	**62**	2	33.3	114	275
S/N/244	**57**	2	25.6	96	177
S/N/245	**20**	2	24.2	88	71
S/N/246	**53**	2	22.1	98	47
S/N/247	**32**	2	31.4	89	187
S/N/248	**41**	1	23.1	86	125
S/N/249	**60**	1	23.4	76.67	78
S/N/250	**26**	1	18.8	83	51
S/N/251	**37**	1	30.8	112	258
S/N/252	**45**	1	32	110	215
S/N/253	**67**	1	31.6	116	303
S/N/254	**34**	2	35.5	120	243
S/N/255	**50**	1	31.9	78.33	91
S/N/256	**71**	1	29.5	97	150
S/N/257	**57**	2	31.6	117	310
S/N/258	**49**	1	20.3	93	153
S/N/259	**35**	1	41.3	81	346
S/N/260	**41**	2	21.2	102	63
S/N/261	**70**	2	24.1	82.33	89
S/N/262	**52**	1	23	107	50
S/N/263	**60**	1	25.6	78	39
S/N/264	**62**	1	22.5	125	103
S/N/265	**44**	2	38.2	123	308
S/N/266	**28**	2	19.2	81	116
S/N/267	**58**	2	29	85	145
S/N/268	**39**	2	24	89.67	74
S/N/269	**34**	2	20.6	98	45

S/NO.	AGE	SEX(Male=2,Female=1)	BMI	BP	Diabetes
S/N/270	**65**	1	26.3	70	115
S/N/271	**66**	2	34.6	115	264
S/N/272	**51**	1	23.4	87	87
S/N/273	**50**	2	29.2	119	202
S/N/274	**59**	2	27.2	107	127
S/N/275	**52**	1	27	78.33	182
S/N/276	**69**	2	24.5	108	241
S/N/277	**53**	1	24.1	105	66
S/N/278	**47**	2	25.3	98	94
S/N/279	**52**	1	28.8	113	283
S/N/280	**39**	1	20.9	95	64
S/N/281	**67**	2	23	70	102
S/N/282	**59**	2	24.1	96	200
S/N/283	**51**	2	28.1	106	265
S/N/284	**23**	2	18	78	94
S/N/285	**68**	1	25.9	93	230
S/N/286	**44**	1	21.5	85	181
S/N/287	**60**	2	24.3	103	156
S/N/288	**52**	1	24.5	90	233
S/N/289	**38**	1	21.3	72	60
S/N/290	**61**	1	25.8	90	219
S/N/291	**68**	2	24.8	101	80
S/N/292	**28**	2	31.5	83	68
S/N/293	**65**	2	33.5	102	332
S/N/294	**69**	1	28.1	113	248
S/N/295	**51**	1	24.3	85.33	84
S/N/296	**29**	1	35	98.33	200
S/N/297	**55**	2	23.5	93	55
S/N/298	**34**	2	30	83	85
S/N/299	**67**	1	20.7	83	89
S/N/300	**49**	1	25.6	76	31
S/N/301	**55**	2	22.9	81	129
S/N/302	**59**	2	25.1	90	83
S/N/303	**53**	1	33.2	82.67	275
S/N/304	**48**	2	24.1	110	65
S/N/305	**52**	1	29.5	104.33	198
S/N/306	**69**	1	29.6	122	236
S/N/307	**60**	2	22.8	110	253
S/N/308	**46**	2	22.7	83	124
S/N/309	**51**	2	26.2	101	44
S/N/310	**67**	2	23.5	96	172
S/N/311	**49**	1	22.1	85	114
S/N/312	**46**	2	26.5	94	142
S/N/313	**47**	1	32.4	105	109
S/N/314	**75**	1	30.1	78	180
S/N/315	**28**	1	24.2	93	144
S/N/316	**65**	2	31.3	110	163
S/N/317	**42**	1	30.1	91	147

S/NO.	AGE	SEX(Male=2,Female=1)	BMI	BP	Diabetes
S/N/318	**51**	1	24.5	79	97
S/N/319	**53**	2	27.7	95	220
S/N/320	**54**	1	23.2	110.67	190
S/N/321	**73**	1	27	102	109
S/N/322	**54**	1	26.8	108	191
S/N/323	**42**	1	29.2	93	122
S/N/324	**75**	1	31.2	117.67	230
S/N/325	**55**	2	32.1	112.67	242
S/N/326	**68**	2	25.7	109	248
S/N/327	**57**	1	26.9	98	249
S/N/328	**48**	1	31.4	75.33	192
S/N/329	**61**	2	25.6	85	131
S/N/330	**69**	1	37	103	237
S/N/331	**38**	1	32.6	77	78
S/N/332	**45**	2	21.2	94	135
S/N/333	**51**	2	29.2	107	244
S/N/334	**71**	2	24	84	199
S/N/335	**57**	1	36.1	117	270
S/N/336	**56**	2	25.8	103	164
S/N/337	**32**	2	22	88	72
S/N/338	**50**	1	21.9	91	96
S/N/339	**43**	1	34.3	84	306
S/N/340	**54**	2	25.2	115	91
S/N/341	**31**	1	23.3	85	214
S/N/342	**56**	1	25.7	80	95
S/N/343	**44**	1	25.1	133	216
S/N/344	**57**	2	31.9	111	263
S/N/345	**64**	2	28.4	111	178
S/N/346	**43**	1	28.1	121	113
S/N/347	**19**	1	25.3	83	200
S/N/348	**71**	2	26.1	85	139
S/N/349	**50**	2	28	104	139
S/N/350	**59**	2	23.6	73	88
S/N/351	**57**	1	24.5	93	148
S/N/352	**49**	2	21	82	88
S/N/353	**41**	2	32	126	243
S/N/354	**25**	2	22.6	85	71
S/N/355	**52**	2	19.7	81	77
S/N/356	**34**	1	21.2	84	109
S/N/357	**42**	2	30.6	101	272
S/N/358	**28**	2	25.5	99	60
S/N/359	**47**	2	23.3	90	54
S/N/360	**32**	2	31	100	221
S/N/361	**43**	1	18.5	87	90
S/N/362	**59**	2	26.9	104	311
S/N/363	**53**	1	28.3	101	281
S/N/364	**60**	1	25.7	103	182
S/N/365	**54**	2	36.1	115	321

S/NO.	AGE	SEX(Male=2,Female=1)	BMI	BP	Diabetes
S/N/366	**35**	2	24.1	94.67	58
S/N/367	**49**	2	25.8	89	262
S/N/368	**58**	1	22.8	91	206
S/N/369	**36**	2	39.1	90	233
S/N/370	**46**	2	42.2	99	242
S/N/371	**44**	2	26.6	99	123
S/N/372	**46**	1	29.9	83	167
S/N/373	**54**	1	21	78	63
S/N/374	**63**	2	25.5	109	197
S/N/375	**41**	2	24.2	90	71
S/N/376	**28**	1	25.4	93	168
S/N/377	**19**	1	23.2	75	140
S/N/378	**61**	2	26.1	126	217
S/N/379	**48**	1	32.7	93	121
S/N/380	**54**	2	27.3	100	235
S/N/381	**53**	2	26.6	93	245
S/N/382	**48**	1	22.8	101	40
S/N/383	**53**	1	28.8	111.67	52
S/N/384	**29**	2	18.1	73	104
S/N/385	**62**	1	32	88	132
S/N/386	**50**	2	23.7	92	88
S/N/387	**58**	2	23.6	96	69
S/N/388	**55**	2	24.6	109	219
S/N/389	**54**	1	22.6	90	72
S/N/390	**36**	1	27.8	73	201
S/N/391	**63**	2	24.1	111	110
S/N/392	**47**	2	26.5	70	51
S/N/393	**51**	2	32.8	112	277
S/N/394	**42**	1	19.9	76	63
S/N/395	**37**	2	23.6	94	118

Peer-review history:
The peer review history for this paper can be accessed here (Please copy paste the total link in your browser address bar)
http://www.sdiarticle4.com/review-history/62913