Group covariates assessment on real-life Diabetes patients by Fractional Polynomials: a study based on Logistic Regression Modeling

Muhammad Noman Sohail[1, *], Jiadong Ren[1], Musa Uba Muhammad[1], Tahir Rizwan[2], Wasim Iqbal[3], Shake Ibna Abir[1], Muhammad Irshad[1], Musavir Bilal[1]

[1]Department of Information Science & Technology, Yanshan University, Qinhuangdao, Hebei, China. [2]Department of Controls & Automation, Shanghai Jiao Tong University, Shanghai, China. [3]Department of Economics & Management, Yanshan University, Qinhuangdao, Hebei, China

Received: April 16, 2019; **accepted:** May 20, 2019.

The advanced approach to modeling the logistic regression with fractional polynomials is applied in place of the traditional linear predictors to group the continuous covariates for the healthcare dataset. The real-life data obtained from the 500 of diabetic patients in northwestern Nigeria. The statistical modeling and predictions of finding the group covariates analytically based on the patients' variables, age and the occupation, by the theories of "Royston and Altman" and "Royston and Sauerbrei". The algorithm in terms of the selection for key factors with the properties congregates at φ (3, 3) with the deviance ratio of 113.00 and the log likelihood assessment of -56.50 for the model results of patients' age. For the patients' occupation, the algorithm for the key factors with extensive outcomes converged at φ (-2, 3) with the deviance ratio of 111.36 and the log likelihood assessment of -56.43. The analysis modeling approach for the second standard method with the fractional polynomials provides the excellent results on the healthcare dataset to investigate the diabetic status. The method is also sufficient for the metadata of different disease because it produces the minimum deviance and maximum log likelihood values.

Keywords: algorithm; deviance; diabetes mellitus; fractional polynomials; group covariates; implications; logistic regression; log likelihood; odd ratios.

Financial support: This project was supported by NSFC "Natural Science Foundation of Hebei Province, China" under the grants of 61572420, 61472341, and 61772449.

[*]**Corresponding author:** Muhammad Noman Sohail, Department of Information Science & Technology, Yanshan University, Qinhuangdao, Hebei, China. Phone: +86 150 3237 0085. Email: mn.sohail@stumail.ysu.edu.cn.

Introduction

In the immense rising field of inclusive studies like medicines, preventive healthcare, and many countless others, it became very significant to accurately predict the binary variable response or in the required probability of frequent occurrence of basic values. Therefore, in the binary supervised learning, it is an ultimate goal to predict, how to carefully distinguish between the two privileged classes of 0 and 1 or in variable classes X and Y on the base of predictor variables like (k), which is called the covariates [1]. The modeling of logistic regression (LR) was used in statistics from several years and recently it became the widespread study on objects in the environment of machine learning. The statistical tool strengthens the desires of modeling the probabilities of privileged classes and observations by linear functions of key variables. The approach in flattening the relationship of variables is under on calculations. And

sometimes these relationships are however defined unknown because the study of statistics is involved in the comparative analysis of one or more variables data. The multiples of LR analysis were used to smoothen the model for datasets, which may involve the linear terms of covariates. By the studies, the conscious choices of developing an appropriate model are based on the ordinary linear effects, but still, the linearity postulations are debatable [2, 3]. According to Royston and Sauerbrei et al. the primary active users for the multiple regressions or covariance analysis was adopting the linear terms of covariates in datasets [4, 5]. Incoherent words for each, covariate (X) appears prominent in the specific models as the linear term of βX. The specific model used for extended quadratic terms if the curving relation by variables X and Y have expected.

In studies of practical applications, the preferred choice for the linear quadratic was concluded with cubic and higher order polynomials, which were used rarely [4]. It has proven that the conventional lower level polynomials do not properly fit on the dataset instead of the higher order polynomials to experience an issue in dynamic ranges of X according to the Royston and Altman et al. [6]. They have stated that the standard models with more than one X variables are difficult in accurately valuing the divine powers. It is unnecessary to evaluate the intellectual powers because the gleaming surface is usually flat sometimes and on maximum, although Y variable is not being linear in X^p. Though LR model is useful, it still drives the difficulties on the assessment and analysis of classification to accurately measure the assumptions of linear effects of covariates of real problems.

Various researchers present the work on data analysis, classification, and forecast predictions upon the datasets related to various modern industries such as Tingley et al. introduced the Bayesian regression methodology for inferior dimensional functions with amicable relations and horizontal effects, affectionately called BWISE [7]. From the study, the specific model of Bayesian regression suited well to smooth the data in case of standard deviations from linearity. Another simplified approach for the multiple nonlinear regression methods is MARS [8], while the specific number of one-dimensional spline functions proved the essential functionality. In MARS, the essential functions become added during the learning process by involving modern techniques called sequential forward selection. From another point of view, LR has enriched considerable popularity in the modern machine learning environments to forecast the standard models for predictions due to the intimate relations with established techniques such as Support Vector Machine (SVM), AdaBoost M1, Artificial Neural Networks (ANN), and others [9–11]. Sohail et al. resembled the SVM and LR to diminish the loss functions in datasets and showed that the loss functions of LR can have a good approach by seeking the SVMN multiple knots of support vector machine [12]. Ayyildiaz et al. mentioned the possible scenario of SVM and LR by comparing their loss function for group covariates [13]. Gong Wei et al. explained the considerable disparity of LR relation with efficient AdaBoost algorithms to stabilize the probability scattering model [14]. Many other researchers showed the different ways in the assessment of group covariates. They used LR, Neural Networks, and others. The studies show some other approaches to astound the dimensionality curse, which represents the sum of dimensional functions [15–17].

The objective of this revised study [18] is to obtain the assessment of group covariates on healthcare dataset by linear regression modeling through the fractional polynomial approach. This study utilizes the model theories of “Royston and Altman” and “Royston and Sauerbrei” to test on the healthcare dataset, instead of the traditional linear predictor of linear regression to group the continuous covariates. The healthcare record of diabetes patients was obtained from northwestern Nigeria. Only three variables were examined to find the correlation between them

including "AGE, Occupation, and Diabetes Mellitus Types," Initially, the polynomial fractional regression (PFR) model was considered for analysis of the group continuous covariates by estimating the power terms to the small-predefined set of unique integers and non-integer values on the statistical package platform of STATA (12.0). Moreover, STATA analyses were used on the R statistical programming platform (5.3.1) to present the canonical correlation status assessment between the three selected variables.

Materials and Methods

Data collection

The dataset of diabetic patients obtained from northwestern Nigeria includes healthy and unhealthy diets patient. All procedures performed in this study were in accordance with the ethical standards of the institutional research committee and with the 1964 official Helsinki declaration and its more recent amendments or comparable ethical standards.

Model analytical platform

The proposed model framework is shown in the Figure 1, which includes three stages of data collection, attribute selection, and modeling. Stage one describes about the collection of data. Stage two presents the process of features selection. Stage three shows the modeling phase.

Data preprocessing

The collection of real-life patients' data was obtained by questionnaire designs and verbal interviews held by consulting medical specialists from 1,257 diabetic patients (healthy and unhealthy from age of 15 to 87 years old with 104 variables from Abdullahi Wase Specialist Hospital and Ajingi General Hospital from 2017 to 2018). For the correlation analysis in studies, we used three major attributes "Age, Occupation, and Diabetes Mellitus Type". Out of 1,257 records, 757 records were removed because of missing values in variables. 500 records were utilized in the analysis.

Dataset classification is a modern technique, which is used to correctly assign the class labels to a dataset. There are two significant types of standard classifications [9-11]. Supervised learning is the dataset which the class label is known in advance. Training data is the set of accurate records which ordinarily have in common multi attributes including predefined class. In this scenario, the model is built up as the training dataset, while the model is used to assign the privileged classes to the testing dataset. Mostly in diabetes data and other disease data, the datasets are based on supervised learning because either the reliable data is gathered manually like in this research or from the artificial repository databases with the class is defined as tested positive or tested negative. Data pre-processing has the quality of effect on prediction results. In other words, pre-processing plays a major role. In our model, we used optimization method to the dataset by analyzing each attribute for our variables. Because the number of dependencies has little relations, we changed the nominal attributes to numeric attribute with value 0 as 'No' and 1 as 'Yes' for the diabetes patients. The occupation status of patients has converted into four sets of numeric values. Hence, the complexity of data was reduced. After the above processing, the values of the dataset were changed into values of 0 and 1 by equation (1), where x' is the average value and s representing the standard deviation for variables.

$$values = \frac{value - x'}{s} \quad (1)$$

The variables, which were considered to conduct experiment in this study by LR, are Age, Diabetes Type, and Occupation as shown in Figure 2. These variables are dependent by nature as default and independent in values. STATA (12.0) was utilized firstly to check the group covariates according to "Royston and Altman" theory. In

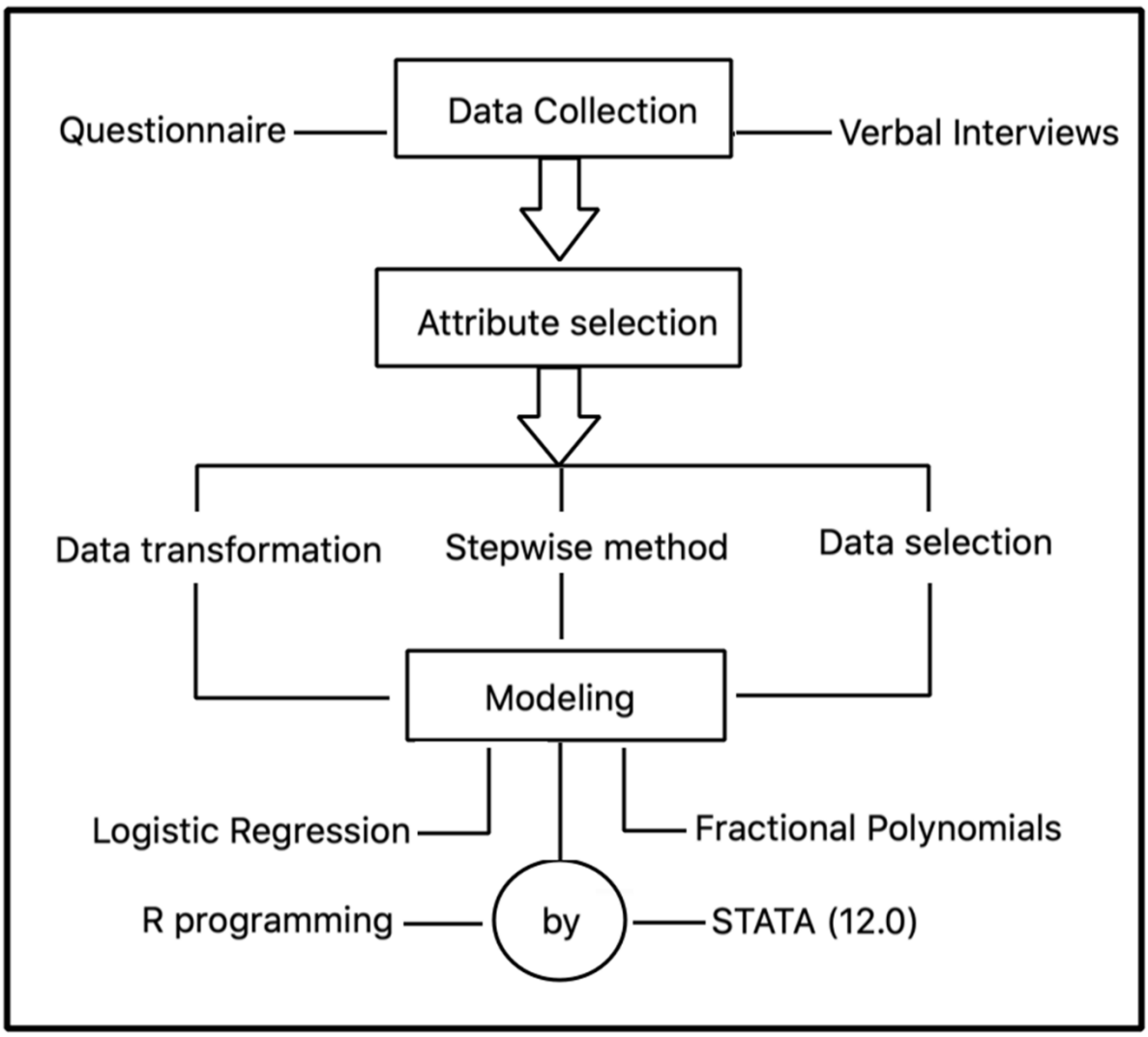

Figure 1. The flow structure of the methodology.

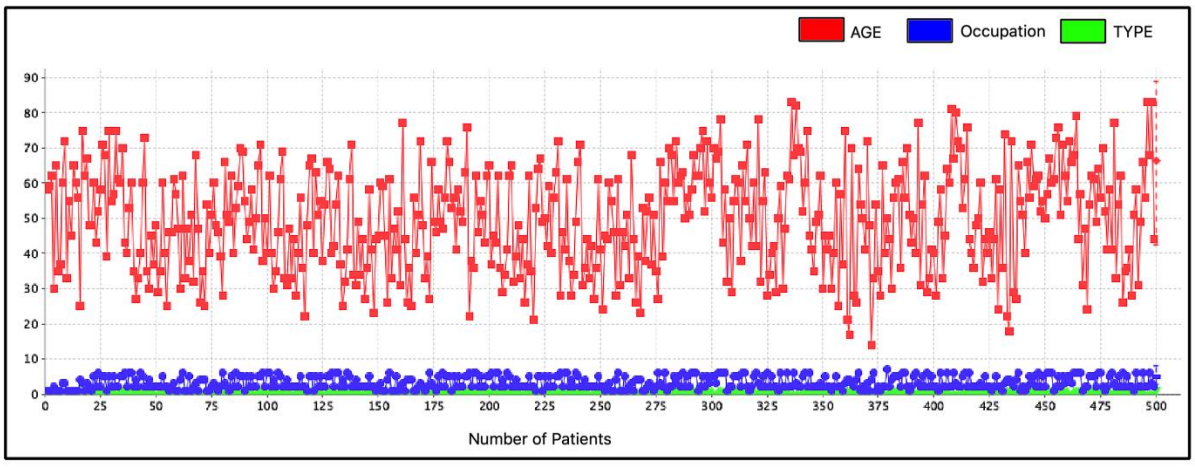

Figure 2. The attribute detail utilized in the research.

addition, we used the R programming package for the presentation of the variable correlations because R production and illustration is easy to understand for the readers.

Model competencies

It is extremely important to measure the model precisely to prove the moral validity and operational capabilities. The assessments observed as an evidence to promote the acceptance and usability ratio of model. The model adequacy and capability can be measured by R^2 as the complex coefficient of accurate determination. The determination of measurement accountability is the considerable number of variations in the accurate data by LR model. In addition, we used R^2-adjusted as necessary adjustment of coefficient determination, R^2-press as the predictions error rate in sufficient sum of squares, and R^2-predicition as accurate predictions of coefficient determination. Royston and Altman et al. tested the deviance and likelihood methodology in terms of selecting the fractional polynomial models. The result showed that the small deviance with the large log likelihood values proved the best fit to the results [6]. They prolonged the deviance method to deviance the two fractional polynomials (FP) in terms of gain (G). Moreover, they delivered that if the standard model presents the large gain, it means it is the best fit on the dataset for results.

Initially, our work adopts the model adequacy method for deviance, gain (G), and logs likelihood functions because it receives straightforward interpretations by the help of Stata software. Furthermore, model demonstrates the correlation between selected variables obtained by R programming for the more proper understanding.

Logistic regression

The way to generalize the linear model GLM is LR, which concerned with the binary model reactions. The multiple LR models emphasize with the covariates $x_1, x_2, x_3 \ldots x_n$ to measure the probability p for the binary events of interest y by the equation (2).

$$logit(p) = \frac{p}{1-p} = \beta_0 + \sum_{j=1}^{n} \beta_j x_j \quad (2)$$

Where P/(1-P) is the odd events, if in case the y has the values of 0 and 1 for an event and non-event than y adopts the Bernoulli distribution method to measure the parameters of probabilities with values of p. Let's consider that the values of binary outcome variable are y and vector ($x = 1, x_1, x_2, x_3 \ldots x_n$) of the covariant for each individual N, while we have the hypothesis for the input vectors contains the perpetual value 1 for interception. We code the two classification classes as 0 and 1 represented by y_i, while y_i has the values 1 and 0 for the first and second class. Let's represent the conditional probability p, which is associated with the first class and represent with the equation (3).

$$logit(p) = \ln\left(\frac{p}{1-p}\right) = f(x,\beta) = \beta^t x \quad (3)$$

Where β represents the coefficient vector model with values ($\beta_0, \beta_1, \beta_2 \ldots \beta_n$) and β^t represents as a transpose vector. This simple calculation in the above equation shows the occurrence of probability in functions of covariates for an event of nonlinear as represented in equation (4). The function of conditional probability is known by the optimal decision as shown in equation (5).

$$p(x;\beta) = \frac{e^{\beta^t x}}{1+e^{\beta^t x}} \quad (4)$$

$$r(x) = \sin\{\ln(\frac{p}{1-p})\} \quad (5)$$

Fractional polynomials

The functions in FP are flexible as compared to the standard polynomial functions [4, 5]. The FP has the desire to cut down the points and used

by the regression models in order to keep fitting the non-linear functions. In this study with diabetes dataset, FP was defined with the degree m by equation (6).

$$\emptyset_m(x;\xi;p) = \xi_0 + \sum_{j=1}^{m} \xi_j x^{p_j} \qquad (6)$$

Where m is the positive integer, p is the values as (p_1, p_2, p_3, … p_m) containing the powers in the real vector values ($p_1 < \cdots < p_m$), and ξ is the coefficients of real values which are (ξ_0, ξ_1, ξ_2, … ξ_m). The round brackets have an approach which brings the alteration of Box-Tidwell [6]. Royston and Altman mentioned the extension of above equation in terms of equal powers, which are (m > 1) and ($p_i = p_j$) in order to perform the distant indices for at least one pair (i, j), while (1≤ i and j ≤ m). For if m = 2, (i, j) = (1, 2), p will be indicated with the (p_1, p_1). As we have the equation (7) for the FP of degree 1 instead of 2, the p_2 will have the limit as compared to the p_1, which is mentioned in equations (8) and (9) and is proved by the equation (10). Equation (10) indicates the family curve with three parameters for the dataset, while m > 2 and ($p_1 = \cdots = p_m$), which is articulated as an equation (11) to specify the subjective powers of ($p_1 \le p_2 \le \cdots \le p_m$). We have set $d_0(x) = 1$, $p_0 = 0$, which has to be expressed in extended definition as shown in equation (12), where values of j equal to (1 … m) and are expressed in equation (13).

$$\phi_2(x;\ \xi;p) = \xi_0 + (\xi_1 + \xi_2)x^{p_1} \qquad (7)$$

$$\frac{\xi_0+\xi_1 x^{p_1}(x^{p_2-p_1}-1)}{p_2-p_1} \qquad (8)$$

$$\xi_0 + \xi_1 x^{p_1} + \xi_2 x^{p_1}\log x\, It_{p_2-p_1} x^{(p_2-p_1)-1} = x^{-1} = \log x \qquad (9)$$

$$\xi_0 + \xi_1 x^{p_1} + \xi_2 x^{p_1}\log x \qquad (10)$$

$$\xi_0 + \xi_1 x^{p_1} + \sum_{j=2}^{m} \xi_j x^{p_1}\log x^{j-1} \qquad (11)$$

$$\phi_m(x;\ \xi,p) = \sum_{j=0}^{m} \xi_j H_j(x) \qquad (12)$$

$$H_j(x) = \begin{cases} x^{p_j} & (if\ p_j \neq p_{j-1}) \\ H_{j-1}(x)\log x & (if\ p_j = p_{j-1}) \end{cases} \qquad (13)$$

The repeated relationship for H_j (x) in relation with H_{j-1} (x) is the depiction of functional parts, when $p_j = p_{j-1}$ and compares the fractional polynomials straight for the diabetes dataset. In addition, the H (x) can be represented as a vector function as (H_0, H_1, H_2, … H_m), which is the most appropriate classification of polynomial fractions of m degree.

Results and Discussion

The deviance measurement evaluates the capability of methodology in the dataset and models. In our proposed situation, the parameters of μ has to be expressed as the log likelihood which ratio is represented by the equation (14).

$$D * (y\mu') = -2(\log(\mu';y) - \log(\mu'_{max};y) \qquad (14)$$

Where log (μ'; y) represents the log likelihood of the proposed method and log (μ'_{max}; y) expresses the log likelihood of the maximum capacity of the accomplished model. In terms of generalizing linear model (GLM), their deviance can be scaled and measured as the equation (15).

$$D * (y;\ \mu') = \frac{1}{\phi} D(y;\ \mu') \qquad (15)$$

Where D ∗ (y; μ') represents the residual deviance of methodology and evolves the sum of discrete deviance assistances. φ is uttered as scattering parameter.

Table 1. The outcomes and consideration assessment of patients' variable "Age and Occupation" by diabetic status.

Variables	Odd ratios (OR)	Error rate	P- value	95% Confident interval	
				Lower bound	Upper bound
Grouped Age 1	0.78	0.011	0.286	0.58	1.17
Grouped Age 2	1.00	0.100	0.215	0.875	1.20
Occupation status	0.406	0.100	0.011 *	0.386	0.78
Constant	2.173	0.57	0.001 *	1.13	4.03 [1]

[1]The results of LR by utilizing FP method and OR of the algorithm for group covariates of patients "Age and Occupation" with prestige of diabetes. When age should be static at the level of 5% in the connotation, occupation with the coefficient of 0.78 will be noteworthy with the probability value of 0.011. The algorithm for selective variables with significance has effect to be covered with the deviance for the standard model with moral values of 113.00. This deviance value possesses the most divine powers for the grouped age model, which is precisely (3, 3) and represents the intellectual power of covariates in the standard model at which the algorithm has been covered with the log likelihood ratio of -56.50 and x^2 moral value of 8.11 with probability value of 0.03.

According to Royston and Altman [11], the power vector ($p\sim = (p_1\sim, p_2\sim, p_3\sim, \ldots p_m\sim)$ for m is associated with the method of uppermost likelihood and is also correspondent with the inferior deviance D. In that case, the $p\sim$ can be acted as the concentrated estimate likelihood MLE of p over the constrained parameters of plot, which is centered by S. And in the model, the convenient use of deviance D (1, 1) is concomitant with the conservative model line of φ_1 (x; 1), which is (m = 1, p = 1) heroic as the base lines of reporting deviance of other replicas. Hence, the gain (G) for the model can be expressed on a dataset as the deviance for φ_1 (x; 1) and is represented as the equation (16), where G cites the alterations of two alterable degrees deviances. The larger gain (G) specifies the finest fit in the method.

$$G = G(m,p) = D(1,1) - D(m,p) \qquad (16)$$

Assessment of "Age and Occupation"

The diabetic status of patients with the set of age and occupation were evaluated for coefficient ratios, standard errors, and probability values with the interval ratios. Moreover, the deviance power and log likelihood of the grouped aged model were covered by the used algorithm. The resulted values and ratios are described in Table 1.

In this model, about 2.6 subtracted from the status of occupation, which comes as (x^3 - 8.76) and (x^3 log x - 6.36), has been designated from the age in order to improve the scaling ratio of the coefficients of regression with odd ratios (OR), while x acts as age. The improved model for the assessment of "Age and Occupation" with status of diabetes is as equation (17), where D acts as diabetic status and x represents the grouped age.

$$D = -14.52 + 0.78x^3 + 1.0x^3 \qquad (17)$$

Assessment of "Occupation and Age"

The diabetic status of patients with the set of patients' occupation and age were evaluated for coefficient ratios, standard errors, and probability values with the interval ratios. The deviance and log likelihood of the grouped aged model has been covered by the algorithm. The resulted values and ratios are described in Table 2.

In this model, about 2.1 subtracted from the status of age, which comes as (x^2 - 0.126) and (x^3 - 16.288), has been designated from the

Table 2. The outcomes and consideration assessment of patients "Occupation and Age" by privileged diabetic status.

Variables	Odd ratios (OR)	Error rate	P- value	95% Confident interval	
				Lower bound	Upper bound
Occupation status 1	0.611	7.75	0.026 *	0.531	70.32
Occupation status 2	0.983	0.004	0.432	0.967	1.00
Grouped Age	1.101	0.276	0.63	0.613	1.876
Constant	2.376	0.718	0.002 *	1.24	4.380 [2]

[2]The results of LR by FP method and OR of the algorithm for group covariates of patients "Occupation and Age" with e prestige of diabetes. When the occupation should be static at the level of 5% in the connotation, the age with the coefficient of 0.611 will be noteworthy with the probability value of 0.026. The algorithm for selective factors with significance has effect to be covered with the deviance for the standard model. This deviance value possesses the most divine powers of (-2, 3), while (-2) indicates the FP power, which is the best fit on our diabetes dataset. The algorithm covers the deviance of 111.36 for the extensive assortment of moral considerations with the effects for the standard model. The deviance value (111.36) maintains the powers for the grouped age model, which is (-2, 3) and it accurately represents the power of covariates in the model at which the algorithm has been covered with the log likelihood ratio of -56.43 and x^2 moral value of 8.62 with polynomial value of 0.01.

occupation in order to improve the scaling ratio of the coefficients of regression with OR, while x acts as status of occupation. The improved model for the assessment of "Occupation and Age" with status of diabetes is as equation (18), where D acts as diabetic status.

$$D = -15.99 + 0.61x^{-2} + 0.983x^{3} \qquad (18)$$

We used R programming tool kit [19] to present the direct correlation graphically based on the results achieved by STATA. Canonical correlation analysis was adopted between the three variables in this study and was shown in Figure 3.

In the practical terms of medial solicitations, modern scholars used to compartmentalize the continuous covariates such as discussed in the related work section to erstwhile the molding analyses. From the statistical view, this eradicates the requirements of linearity postulations and allows individuals to clarify the results in simple and ingenious ways. Besides, classification can be resulted as the lack of powers and information to hold linearity rather than the inception association concerns. By the comprehensive analysis and comparison of the intellectual abilities of statistical methods to find out the direct correlation for continuous covariates on datasets, it shows that the FP is the best modeling technique to cope with the "linear and polynomial" effects with the maximum potential of properties and results. FP does not escalate the inaccuracy of type one. The standard deviation remains comfortably for the significant key factor which has been used to predict the correlation of infectious diseases in the different datasets of medical health diagnosis. According to the study conducted in Australia [20], the LR model was best fitted to the FP to figure out the covariates relations of variables to predict the risk analysis. In this study, we have selected the variables of diabetic patients to find the correlations in the patient's age and occupation. We have determined the association of the model fitted for correlations in grouped polynomial covariates on the diabetes dataset. Although the tractability of the FP model produces profound effect in over fitting with outcomes, it reflects the medical philosophies and acquaintance, and is accentuated. To prevent the conflicting results, it is significant to achieve the adequate consistency of primary focus. In our consistent findings, two successful models of "Royston and

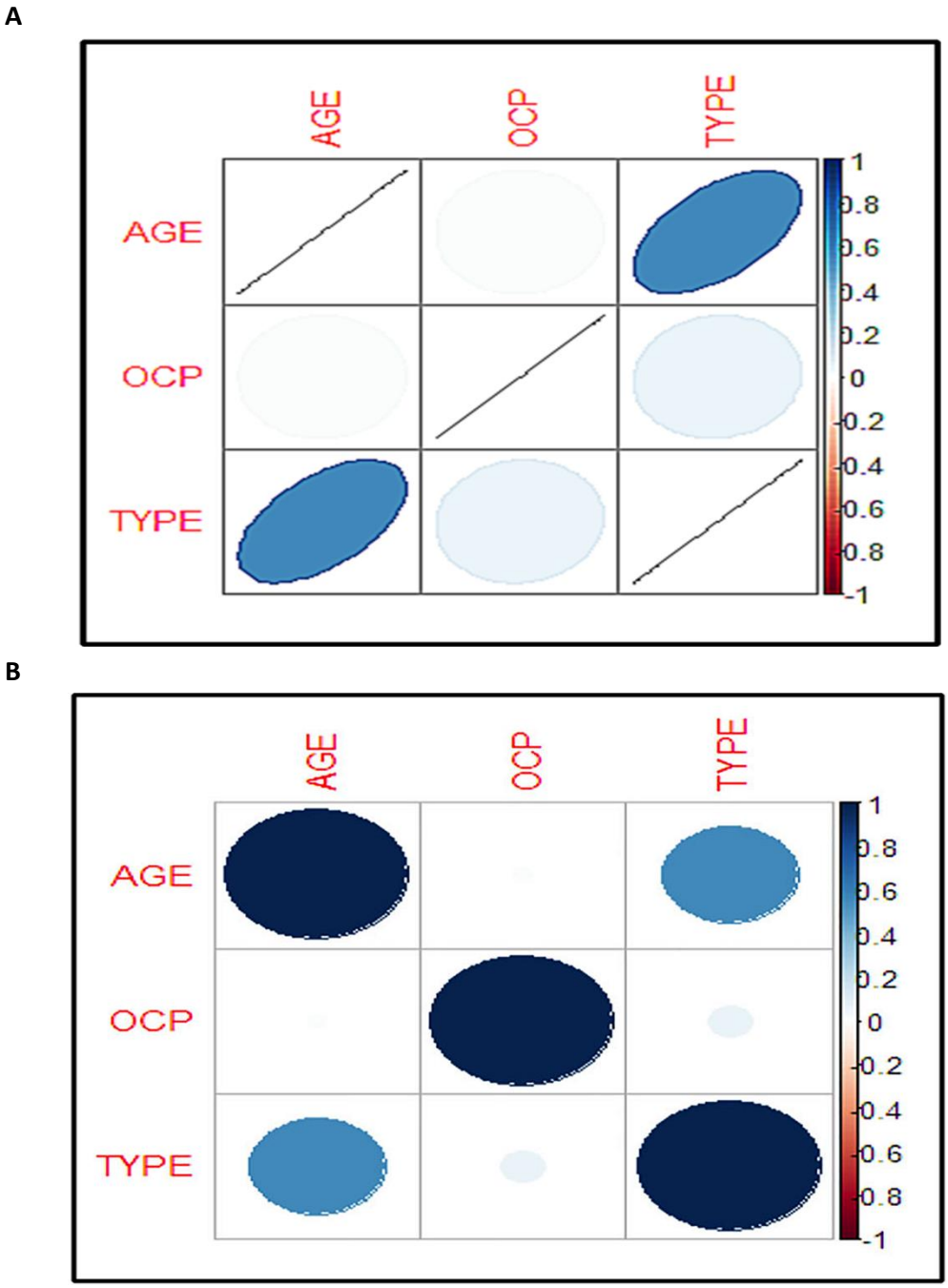

Figure 3. The correlation between the selected variable of diabetic patients "AGE, OCP, and TYPE". The results using R toolkit indicates that there is a complex relationship between Diabetes Type and patients Age of approximately 0.6, and a relationship between Diabetes Type and patients Occupation status of approximately 0.1. There is no direct relationship between the patients Age and Occupation.

Altman" and "Royston and Sauerbrei" for the FP have sufficiently shown the desire results.

In conclusion, statistical methods and specific models of this study disclosed the relationship among the variables according to the most encyclopedic knowledge. Principally, the objective of this study is to analyze the best-fit LR method and model with the possible help of FP fractional polynomials of grouped continuous

covariates. By comparing the tailored models and the approach in term of possible associations among the variables, FP shows the best-fit method comparing to that of the conventional polynomials as shown in the functional equations (17) and (18). In the analogous manners, the model in equation (18) has produced the most superior approach. The results proved that the explained equation in results obtained the less deviance of 111.36 and maximum log likelihood values of -56.43. After all, it showed the gain (G) 1.31 by converging at φ (-2, 3).

Acknowledgement

We adore prompting our appreciations to NSFC and Yanshan University, China to accompany us in this research.

References

1. Guyon I, De AM. 2003. An Introduction to Variable and Feature Selection André Elisseeff. J Mach Learn Res. 3:1157–1182.
2. Silverman BW. 2018. Density Estimation for Statistics and Data Analysis. 1st ed. Routledge. doi:10.1201/9781315140919.
3. Aggarwal CC. 2015. Data Mining: The Textbook. doi:10.1007/978-3-319-14142-8.
4. Royston P, Sauerbrei W. 2008. Multivariable model-building: a pragmatic approach to regression analysis based on fractional polynomials for modelling continuous variables. John Wiley.
5. Royston P, Sauerbrei W. 2008. Interactions between treatment and continuous covariates: a step toward individualizing therapy. J Clin Oncol. 26:1397–1399. doi:10.1200/JCO.2007.14.8981.
6. Royston P, Altman DG. 1994. Regression Using Fractional Polynomials of Continuous Covariates: Parsimonious Parametric Modelling. Appl Stat. 43:429. doi:10.2307/2986270.
7. Tingley D, Buzsski G. 2018. Transformation of a Spatial Map Across the Hippocampal-Lateral Septal Circuit. SSRN Electron J. 47. doi:10.2139/ssm.3155563.
8. Lewis PAW, Stevens JG. 1991. Nonlinear Modeling of Time Series Using Multivariate Adaptive Regression Splines (MARS). J Am Stat Assoc. 86:864–877. doi:10.1080/01621459.1991.10475126.
9. Sohail MN, Ren J, Uba MM, Irshad MI, Musavir B, Abir SI, Wasim I, Usman A, Tahir R, Anthony JV. 2018. Why only data mining? a pilot study on inadequacy and domination of data mining technology. Int J Recent Sci Res. 9:29066–29073. doi:10.24327/ijrsr.2018.0910.2787.
10. Sohail MN, Jiadong R, Uba MM, Irshad M. 2019. A Comprehensive Looks at Data Mining Techniques Contributing to Medical Data Growth: A Survey of Researcher Reviews, Springer, Singapore. pp. 21–26. doi:10.1007/978-981-10-8944-2_3.
11. Sohail N, Jiadong R, Uba M, Irshad M, Khan A. 2018. Classification and cost benefit Analysis of Diabetes mellitus Dominance. Int J Comput Sci Netw Secur. 18:29–35.
12. Sohail MN, Jiadong R, Irshad M, Uba MM, Abir SI. 2018. Data mining techniques for Medical Growth: A Contribution of Researcher reviews. Int J Comput Sci Netw Secur. 18:5–10.
13. Ayyıldız E, Purutçuoğlu V, Weber GW. 2018. Loop-based conic multivariate adaptive regression splines is a novel methodfor advanced construction of complex biological networks. Eur J Oper Res. 270:852–861. doi:10.1016/J.EJOR.2017.12.011.
14. Gong W. 2018. Regression Techniques Used in Hydrometeorology. Handb. Hydrometeorol. Ensemble Forecast., Berlin, Heidelberg: Springer Berlin Heidelberg. pp. 1–27. doi:10.1007/978-3-642-40457-3_63-1.
15. Sohail MN, Ren J, Uba MM, Irshad M, Musavir B, Usman A, et al. 2018. Forecast Regression analysis for Diabetes Growth: An inclusive data mining approach. Int J Adv Res Comput Eng Technol. 7:715–721.
16. Sohail MN, Ren J, Muhammad MU. 2019. A Euclidean Group Assessment on Semi-Supervised Clustering for Healthcare Clinical Implications Based on Real-Life Data. Int J Environ Res Public Heal. 16:1581. doi:10.3390/IJERPH16091581.
17. Sohail MN, Ren J, Muhammad MU, Chauhdary ST, Arshad J, Verghese AJ, et al. 2019. An Accurate Clinical Implication Assessment for Diabetes Mellitus Prevalence Based on a Study from Nigeria. Process. 7:289. doi:10.3390/PR7050289.
18. Muhammad MU, Asiribo OE, Noman SM. 2017. Application of Logistic Regression Modeling Using Fractional Polynomials of Grouped Continuous Covariates. Niger Stat Soc Ed Proc 1st Int Conf. 1:144–147.
19. Gentleman R. 2008. R Programming for Bioinformatics. vol. 13. 1st ed. Chapman and Hall/CRC. doi:10.1201/9781420063684.
20. Zhu Y, Imamura M, Nikovski D, Keogh E. 2018. Introducing time series chains: a new primitive for time series data mining. Knowl Inf Syst. 2018:1–27. doi:10.1007/s10115-018-1224-8.