Imperial Journal of Interdisciplinary Research (IJIR)
Vol-3, Issue-11, 2017
ISSN: 2454-1362, http://www.onlinejournal.in

Modelling Logistic Regression using Multivariable Fractional Polynomials

Dina Omer[1] & Abdullah Bashir Musa [2]
[1] University of Gezira, faculty of mathematics and computer, Department of statistic, Madani, Sudan
[2] University of Dammam, college of computer sciences and information technology, department of computer Sciences, Dammam, Saudi Arabia kingdom

Abstract: *Logistic regression (LR) is a well-known statistical method that has been used widely to build a covariates variables with a binary outcome in many applications of statistic, machine learning, and bioinformatics. Since logistic regression model is a member of a general linear models (GLMs), when building a logistic regression model continuous or categorical covariates are assumed to be linearly associated with outcome variable in logit transformation function. However, it is not always the situation and the assumption may be violated and the relation is nonlinear in some cases which results in inaccurate model. Multivariate fractional polynomial (MFP) modeling is a flexible method to expose such non-linear associations. For the continuous covariates MFP can be used when investigators want to preserve continuous nature of covariates and suspect that the relationship is non-linear. The article aims to apply MFP methods for building Logistic regression model by using R package. Since Continuous variables are often encountered medical studies which are often used to assess risk or prognosis or to select a therapy consequently, in this paper a breast cancer data with 15 covariates variables is used to determine the covariates variables that influence breast cancer. Statistical comparison is used to compare logistic regression model with MFP logistic regression. The study stated that using of MFP logistic regression yield more accurate model and improving logistic regression classification accuracy.*

Keyword*: Logistic regression (LR), Multivariate fractional polynomial (MFP), continuous covariates, non-linear associations, general linear models (GLMs), breast cancer, R package.*

1. Introduction

Predictive models are used in a variety of medical domains for diagnostic and prognostic tasks. These models are built from ‘‘experience’’, which constitutes data acquired from actual cases. The data can be preprocessed and expressed in a set of rules, such as it is often the case in knowledge-based expert systems, or serve as training data for statistical and machine learning models. Among the options in the latter category, the most popular models in medicine are logistic regression (LR) [1].To relate a response variable to continuous covariates variables, a suitable general linear models (GLMs) is required. General linear models assumed linear relation between the covariates variable and the associated transformation function, simple and popular approach is to assume a linear relation, but the linearity assumption may be questionable [2]. To avoid this strong assumption, researchers often apply cut points to categorize the continuous variable, implying regression models with step functions. This simplifies the analysis and interpretation of results.

Fractional multivariate polynomials (FMP) as a useful extension of polynomial regression and as a sensible way to model the relationship (Royston and Sauerbrei 2008). Use of a suitable function selection procedure (FSP) gives a simple way to check whether a linear function (our default) is adequate or whether a non-linear FP function improves the fit of the data substantially. Instance by investigating the effect of age as a prognostic factor for breast cancer, we illustrate how conclusions depend strongly on the manner in which the continuous variable is analyzed.

Logistic regression (LR) [3–6] is a well-known statistical modeling method that origins belong to statistic community, recently has been used widely in a variety of applications including document classification [7] and bioinformatics [8–10]. Logistic regression is particularly appropriate for models involving disease state (healthy/diseased), decision making (yes/no), or mortality (dead, living). It is widely used in binary classification problems in applied sciences such as medicine, biology and epidemiology [11,24]. It has been widely applied due to its simplicity and great interpretability. LR has an additional advantage that the extension to the multi-class case is well described. However this study is focused only on binary classes.

Multivariable Fractional Polynomial (MFP) method is such a method that it lets software to determine whether an explanatory variable is

important for the model, and its functional form. MFP can be used when detectives want to preserve continuous nature of covariates and suspect that the relationship is non-linear. There are two components in the procedure: (I) backward elimination of covariates that are statistically in significant; and (II) iterative examination of the scale of all continuous covariates. Therefore, two significance levels are needed; α1, for the elimination and addition of a covariates, and α2 for the purpose of significance of fractional transformation of continuous covariates [12].Several studies have applied fractional polynomial with linear regression ; W. Sauerbrei and C. Meier-Hirmer [13] applied fractional polynomial with multivariable regression using SAS ,STATA and R , Harald Binder, Willi Sauerbrei [14] proposed a comparison study between splines and fractional polynomials for multivariable model building with continuous covariates concluded that MFP performs better than splines on several criteria . Since Multivariable fractional polynomial (MFP) models are commonly used in medical , several medical studies have applied MFP; Royston1 , W. Sauerbreim [15], Build Multivariable Regression Models with Continuous Covariates in Clinical Epidemiology With an Emphasis on Fractional Polynomials, W. Sauerbrei, P. Royston[16], Build multivariable prognostic and diagnostic models: transformation of the predictors by using fractional polynomials , Tim P. Morris, Ian R. White [17], proposed study that Combine fractional polynomial model with multiple imputation; Royston1 , W. Sauerbreim [18] , Build Multivariable Regression Models with Continuous Covariates in Clinical Epidemiology With an Emphasis on Fractional Polynomials, W Sauerbrei, P Royston, H Binder [19], used fractional polynomials for selection of important variables and determination of functional form for continuous predictors in multivariable model.

To the best of our knowledge, the only study that applied logistic regression with fractional polynomials was proposed by B Silke, J Kellett, T Rooney[20] who used fractional polynomial with logistic regression for improving medical admissions risk system. This study was focusing on the paper' medical issue while our study, this study is mainly focuses on applying of MFP with logistic regression from statistic point of view.

This paper proposes using of MFP which would be accompany the classical, usually method that use for building the logistic regression model. An important advantage of using this approach is to cover the nonlinearity between the covariates variables and the logit transformation function which might be occurs in some cases, this would yield a more accurate model. Goodness of fit have been compute to compare of building logistic regression model with and without MFP. Many classification measures have been computed such as accuracy, spasticity, sensitivity and area under the curve (AUC).

2. Fractional polynomial

Fractional Polynomial (MFP) [21, 22] method is such a method that it lets software to determine whether an explanatory variable is important for the model, and its functional form. MFP can be used when detectives want to preserve continuous nature of covariates and suspect that the relationship is non-linear. There are two components in the procedure: (I) backward elimination of covariates that are statistically in significant; and (II) iterative examination of the scale of all continuous covariates. Therefore, we want two significance levels α1, for the elimination and addition of a covariates, and α2 for the purpose of significance of fractional transformation of continuous covariates [12].

The first cycle is to build a multivariable model by all potential explanatory covariates. Alternatively, variables through P<0.25 or 0.2 in unavailable analysis can be combined into the initial model. This is also the starting model for purpose full selection of covariates. All dichotomous and categorical variables are not subject to Fractional Polynomial (FP) transformation and are modeled via one degree of freedom. They are tested for their contribution to the model using α1 by Wald test. Continuous variables are modeled using closed test to observe whether they should be saved or distant using α1, and whether transformation should be performed using α2. The closed test initiates through comparing the best fitting second-degree fractional polynomial (FP2) with null model. The term is dropped if the test is non-significant. Then the best-fitting FP2 is compared with the linear term. Linear term is adopted if the test is non-significant. Or else we continue to compare the best fitting FP2 to the best-fitting FP1. If the test is significant the best fitting FP2 is accepted. Otherwise the best-fitting FP1 is adopted. The second cycle begins with a fit of the model covering significant covariates, either in original or polynomial transformed form. All covariates are then examined in descending order of significance for their inclusion, exclusion and possible transformation. The procedure stops when two conversation steps contain the same covariates with the same FP transformations.

Closed test algorithm for choosing a fractional polynomial model with maximum permitted degree of 2 for a single continuous predictor [13,23] has been used in this study. The first step is to determine whether a predictor should be included in a model is to compare models with and without FP2. If FP2 model is not better than null model, the predictor is dropped. Otherwise, we continue to compare FP2

with linear model. If FP2 is not better than linear one, we choose linear model. Otherwise, we continue to compare FP2 with FP1. If FP2 is not better than FP1, the FP1 model is chosen. Otherwise the FP2 model is chosen.

FP of a certain degree contains many terms, depending on the number of powers allowed. By convention, powers are selected from the collection (–2, –1, –0.5, 0, 0.5, 1, 2, 3), where 0 denotes the log transformation. FP3 is usually not required, it been presented here for better understanding of fractional polynomial term. FP2 is the most complex and it is compared to the null model. If FP2 is not well than null by statistical test, linear and FP1 of the variable are unlikely to be important to the model. Therefore, the variable is excluded from the model [22].

The method of fractional polynomials can be used with a multivariable logistic regression model, but for sake of simplicity we describe the procedure using a model with a single continuous covariate the logit that is linear in the covariate is.

$$g(x,\beta) = \beta_0 + x\beta_1 \qquad (1)$$

Where β denotes the vector of model coefficients one way to generalize this function is to specify it as

$$g(x,\beta) = \beta_0 + \sum_{j=1}^{J} F_j(x)\beta_j \quad (2)$$

The functions $F_j(x)$ are a particular type of power function the value of the first function is $F_1(x) = x^{P_1}$, P_1could by any number. Royston and Altman (1994) propose restricting the power to be among those in the set S = {-2,-1,-0.5, 0, 0.5, 1, 2, 3} where $P_1 = 0$ denotes the log of the variable.

The remaining functions are defined as

$$F_j(x) = \begin{cases} x^{P_j} , & P_j \neq P_{j-1} \\ F_{j-1}(x)\ln(x) & P_j = P_{j-1} \end{cases}$$

For j=2,…,J and restricting powers to those in S the model is quadratic in x when J=2 with $P_1 = 1$ and $P_2 = 2$ again, we could allow the covariate to enter the model with any number of functions J, but in most applied stings an adequate transformation may be found if we use J=1 or 2.

Implementation of the method requires for J=1 fitting 8 models the best model is the one with the largest log likelihood. The process is repeated with J=2 by fitting 36 models obtained from the distinct pairs of powers, and the best model is again the one with the largest log likelihood.

Let $l(1)$ denote the log likelihood for the linear model, that is J=1 and $P_1 = 1$ and $l(P_1)$ denote the log likelihood for the best J=1 model and $l(P_1, P_2)$ denote the log likelihood for the best J=2 model.

The partial likelihood ratio test comparing the linear model to the best J=1 model

$$G(1, P_1) = -2\{l(1) - l(P_1)\} \qquad (3)$$

Is approximately distributed as chi-square with 1 degree of freedom under the null hypothesis of linearity in x the partial likelihood ratio test comparing the best J=1 model to the best J=2 model

$$G(P_1, (P_1, P_2)) = -2\{l(P_1) - l(P_1, P_2)\} \quad (4)$$

Is approximately distributed as chi-square with 2 degree of freedom under the null hypothesis that the second function is equal to zero similarly [7]

3. Logistic regression

Logistic regression (LR) [1, 2] is a well-known statistical approach to model dichotomous (binary) data; logistic regression is a member of generalized linear models. In logistic regression, a single outcome variable y_i , where i=1,…,n ,each y_i takes only two values 0 or 1 (but not both), so it follows a Bernoulli Probability density function $p(y_i) = \pi_i^{\,y_i}(1-\pi_i)^{1-y_i}$: that takes the value 1 with probability p_i and 0 with probability (1 - p_i). Our interest is in y_i = 1 with the interest probability p_i, which varies over the observations as an inverse logistic function of a vector x_i, which includes a constant (x_0) and k explanatory variables (x_1,…, x_k). Its function can be given as follows: $Y_i \sim Bernoulli(\pi_i)$

$$P(Y_i = 1) = \pi_i = \frac{\exp(X\beta)}{1+\exp(X\beta)} = \frac{1}{1+\exp(-X\beta)} \qquad (5)$$

Where $\beta = (\beta_0, \beta_1')$ *is a* $(k+1)\times 1$ vector that contains the parameters that need to be estimated, β_0 is an intercept term corresponding to x_0 and $\beta$is (k × 1) vector with elements corresponding to the explanatory variables. The odd ratio of y = 1 is p(y = 1)/(1 – p (y = 1) = pi/(1 - pi).By using this odd ratio; the following transformation can be obtained

$$logit(Y_i = 1) = \log(odd) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta X_i \quad (6)$$

The above logit function can be expressed in matrix form as follows:

$$\text{logit } P(Y_i = 1) = X\beta \qquad (7)$$

The importance of the transformation in (13) is that it has many of the desirable properties of the linear regression model. The logit is linear in the parameters vector b. These parameters will be estimated using the maximum likelihood function. The maximum likelihood function of Bernoulli density function is $L(\pi_i|y_i) = \pi_i^{\,y_i}(1-\pi_i)^{1-y_i}$. By assuming independence over the observations, the maximum

likelihood function for y = y_1; . . .; y_n can be written as follows:

$$L(\beta|y) = \prod_{i=1}^{N} \pi_i^{y_i}(1-\pi_i)^{1-y_i} \qquad (8)$$

By taking the logarithm, the log-likelihood will be

$$L(\beta|y) = \sum_{i=1}^{n} [Y_i \log \pi_i + (1-Y_i)\log(1-\pi_i)] \quad (9)$$

After estimating the parameters, the significance of each of these parameters will be assessed by comparing the observed values of the response variable to the predicted values obtained from the model with and without the variable in the model. In logistic regression this comparison is based on the log likelihood function defined in (15). This can be obtained by using the following statistic:

$$G = -2\ln\left[\frac{likelihood\ without\ the\ variable}{likelihood\ with\ the\ variable}\right] \qquad (10)$$

This statistic will be compared with $\chi^2_{(\alpha,1)}$ to test the hypothesis whether the parameter is equal to zero or not, if G> $\chi^2_{(\alpha,1)}$, then the parameter is not significant and should be deleted from the model. There are several selection procedures used to construct the best fitting model such as forward selection which looks at each explanatory variable individually and selects the single explanatory variable that fits the data the best on its own as the first variable included in the model, among the remaining variables the one that adds the most is included. This is repeated until none of the remaining variables will add significantly. Backward selection starts with a model that contains all of the explanatory variables, and then a variable that, if removed, would cause the smallest change in the overall fit of the model is removed.

This continues until all variables in the model are significant. For assessing the goodness-of-fit for the model, there are several goodness-of-fit tests that can be obtained by comparing the overall difference between the observed and fitted values. Among these tests Pearson ChiSquare χ^2 and Deviance D test are used the most. Suppose the number of the covariate patterns is j, let j\n, let m_i denote the number of (y_i =1) among these patterns. The Pearson statistic is defined as follows:

$$\chi^2 = \sum_{i=1}^{n} \frac{(y_i - m_i\hat{\pi}_i)^2}{m_i\hat{\pi}_i(1-\hat{\pi}_i)} \qquad (11)$$

And the residual deviance statistic is defined as follows:

$$D = -2\sum\left[\left(y_i \ln\left(\frac{y_i}{m_i\hat{\pi}_i}\right)\right) + (m_i - y_i)\ln\left(\frac{m_i - y_i}{m_i(1-\hat{\pi}_i)}\right)\right] \quad (12)$$

These statistics are rely on the principle of comparing observed y_i to predicted $m_i\hat{\pi}_i$ values and they should be small if the model fits the data well.

These two statistics are compared to the value of $\chi^2_{(\alpha,n-k-1)}$ to judge their statistical significance. These statistics are used when j<n. Their results are invalid when $j \geq n$ [1, 23]. In this case there are other alternative statistics that can be used, such as Osius and Rojek statistic, Farrington statistic and Hosmer–Lemeshow statistic. The predicted label for the logistic regression model will equal to 1 if $\hat{\pi}_i$ is greater than or equal to some threshold (the default is 0.5), as shown below:

$$if\ \left(P(y=1)\right) \geq 0.5 \quad the\ instance \in class\ (y=1)$$
$$if\ \left(P(y=1)\right) < 0.5 \quad the\ instance \in class\ (y=0)$$

4. Materials and methods

4.1. The data set

The data set that consist of 200 instances been collected from people those are doing physical examination breast cancer at Radiation and Isotopes Center Khartoum, Sudan, The center is sponsored and ensured by (WHO, IAEA), it is the third one in Africa after Cairo and South Africa centers

The questionnaire has 14 questions that are used to build a model for positive/negative breast cancer test, the sample contain 100 of people with positive breast cancer test and 100 with negative breast negative cancer test.

The summary description of the data is given in table.1.

Table 1- summary description of the Breast Cancer dataset of the Breast Cancer dataset

Variable	Description variable	Variable type	Codes/Values
X_1	Age	Continuous	
X_2	Weight	Continuous	
X_3	The age at starting periods	Continuous	
X_4	Marital status	Binary	0=No, married 1 = married
X_5	Having children	Multinomial	0 = No 1 = Yes 2= No married
X_6	Breast feeding	Multinomial	1 = Yes 0 = No 2=No married
X_7	The age after first child	Multinomial	$1 = \geq 30$ $0 = < 30$ 2= No married
X_8	The use of contraceptives	Binary	1 = Yes 0 = No
X_9	Hormone therapy after menopause	Multinomial	1 = Yes 0 = No 2= not to reach menopause
X_{10}	The age at going through menopause	Multinomial	$1 = \geq 55$ $0 = < 55$ 2= not to reach menopause
X_{11}	Having a family history of breast cancer	Binary	1 = Yes 0 = No
X_{12}	Have certain benign Breast Conditions	Binary	1 = Yes 0 = No
X_{13}	Having radiation	Binary	1 = Yes 0 = No
X_{14}	Drinking alcohol and tobacco smoke	Binary	1 = Yes 0 = No

4.2 design of experiment

Coefficient				
	Estimate	**Std. Error**	**z value**	**Pr(>|z|)**
Intercept	0.2916495	4.4749145	0.065	0.948035
X_1	0.0008851	0.0197403	0.045	0.964238
X_2	0.0173840	0.0202244	0.860	0.390034
X_3	-0.0670583	0.1363251	-0.492	0.622790
X_4	1.5543332	1.5227820	1.021	0.307387
X_5	-0.8277958	0.7582412	-1.092	0.274951
X_6	0.1611508	1.4445927	0.112	0.911177
X_7	0.1531333	0.5909776	0.259	0.795544

X_8	1.1539065	0.7021182	1.643	0.100287
X_9	-1.6615149	0.6321344	-2.628	0.008578 **
X_{10}	0.1851722	0.5690915	0.325	0.744892
X_{11}	1.7003278	0.6570896	2.588	0.009663 **
X_{12}	3.0718469	1.0589131	2.901	0.003720 **
X_{13}	2.4413057	0.7147450	3.416	0.000636 ***
X_{14}	-0.6998345	1.3144823	-0.532	0.594447

The Classical Logistic Regression (LR) method has been used to build the logistic regression model, needed. The backward selection procedure was used with 0.05 as the default significance level to build the model, the goodness fitting of the model has been computed by computing the residual deviance statistic, the other method is applying logistic regression with the Fractional polynomials (FP) which is used to determine whether a continuous covariates variable is important for the model, and its functional form. As flow; the closed test procedure, which first conclude the best fitting second degree polynomial by choosing the powers transformation p and q from the above-mentioned set.

The MFP function under R package is use for modelling the effect of continuous variables on the outcome in regression models and to find the best values of power among the combinations of the powers p and q that maximizes the likelihood ratio or equivalently that which minimizes the deviance. MFP uses sequential and closed testing selection procedures for three continuous variables. The final model is selected after two cycles. The first cycle begins by including all covariates into the model and their FP functions are examined. The best fitting FP functions while the second Cycle is the last cycle where the model converges. Transformation of each variable

Similarly the goodness of model fitting has been checked by computing. The residual deviance statistic

Several metrics have been used for the comparison between using logistic regression and companying Fractional polynomials (FP) with logistic regression for analyzing the data including accuracy, Sensitivity, Specificity, Area under curve (AUC) and Receiver Operating Characteristics (ROC) analysis

5. The Results and discussion

5.1 The Results

The results of the implementation of LR on R-package version (R i386 3.2.5) for building the model is shown in Table 2

Table 2. Results of fitting the LR to the breast cancer dataset

Significant .Codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

The Null deviance with 199 degree of freedom is 277.26 while the Residual deviance with 185 degree of freedom is 128.71 and the AIC is158.71

From table.1, the significant factor that effect the breast cancer are X_9 Hormone therapy after menopause, X_{11} family history of breast, X_{12} benign tumor and X_{13} having radiation)

The deviance of the best-fitting FP2 model, the best-fitting FP1 model, the linear model, and the null model is reported in Table 3. along with the results of the test comparing different FP models

For obtaining the best fit fractional polynomial and the transformation, the RA2 selection algorithm is used, the results shown that the FP2 transformation with P = 2; the second fractional polynomial transformation is the best-fitting FP transformation.

Table.3 Fractional polynomial models for the of Breast Cancer data

Model	Deviance	Power	Step	Comparison	Deviance difference
FP2	89.96	(-2,-0.5),(-2,2),(3,3)	1	FP2 versus null	187.3
FP1	123.75	-2	2	FP2 versus linear	38.75
Linear	128.71	1	3	FP2 versus FP1	33.79
Null	277.26				

Many model comparisons have been conducted, comparing of best-fitting FP2 against null model which represented the test of nonlinearity, comparing best fitting FP2 against linear model and comparing first degree (FP1) and second degree (FP2) transformations, the results of Deviance analysis is shown in table 4.

Table 4: Analysis of Deviance

	df	Deviance	AIC
<none>		101.53	125.53
-X_8	1	104.77	126.77
-I((X_2/100)^-2)	1	111.23	133.23
-I((X_2/100)^2)	1	111.59	133.59

-X_{11}	1	112.84	134.84
-I((X_2/10)^3 * log((X_2/10)))	1	114.64	136.65
-I((X_1/100)^-0.5)	1	115.11	137.11
-I((X_2/10)^3)	1	116.04	138.04
-X_{12}	1	116.70	138.70
-X_{13}	1	117.39	139.39
-I((X_1/100)^-2)	1	124.14	146.14
-X_9	1	130.58	152.58

The accuracy, sensitivity, specificity and the area under the curve have been computed for logistic regression model and logistic regression model with fractional polynomial, the results is shown in table 5. The ROC analysis for the logistic regression model and the logistic regression model with fractional polynomial are depicted in figure 2 and figure 3respectively.

Table.5 the metrics results for LR and LR-FP models

Measures	Accuracy	Sensitivity	Specificity	AUC
LR-FP	92	95	89	95.35
LR	85.5	81	90	89.29

Figure1. The ROC analysis for logistic regression model

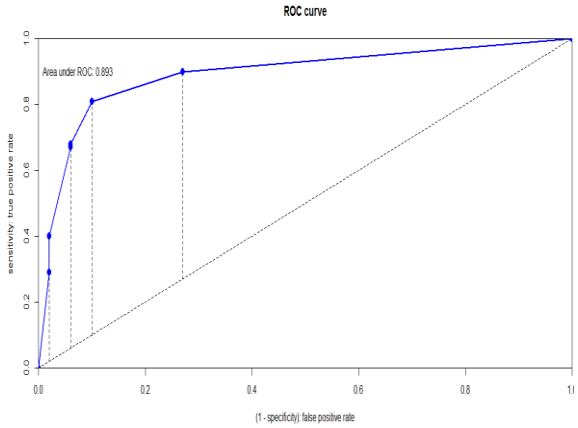
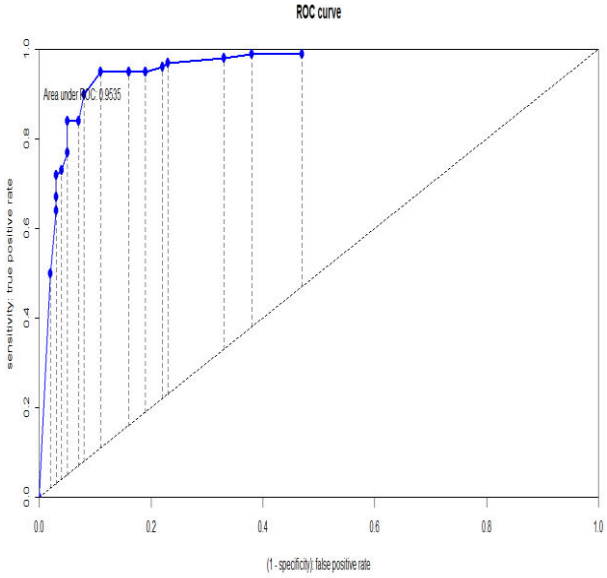


Figure 2. The ROC analysis of the logistic regression with fractional polynomial model



5.2. Discussion

This study investigate the effectiveness of applying Fractional polynomials (FP) for continuous variable in improving the quality of the logistic regression model. A comparison between applying logistic regression model and applying FP with logistic regression have been done. The results of building the logistic regression model is shown in table 1 while the result of building FP with logistic regression model is shown in table 2, in addition several classification measures have been use for the comparison; accuracy, sensitivity, Specificity and area under the ROC curves are computed, the results are shown in table 4.for more accurate comparison the ROC curve analysis of logistic regression and logistic regression with Fractional polynomials have been depicted in figure 3 and figure 4 respectively .
All results demonstrated that using FPM is fit the data better than using LRM.
The final FP model that best descripe to the relationship between factors and the risk of having the Brest Cancer disease is :

$$logit(\pi) = \beta_0 + \beta_1\left(\frac{X_1}{100}\right)^{-2} + \beta_2\left(\frac{X_1}{100}\right)^{-0.5} + \beta_3\left(\frac{X_2}{100}\right)^{-2} + \beta_4\left(\frac{X_2}{100}\right)^{2} + \beta_5\left(\frac{X_3}{10}\right)^{3} + \beta_6\left(\frac{X_3}{10}\right)^{3} * log\left(\left(\frac{X_3}{10}\right)\right) + \beta_7 X_{12} + \beta_8 X_{11} + \beta_9 X_9 + \beta_{10} X_8 + \beta_{11} X_{13}$$

where X_1denotes age, X_2 weight , X_3 age at starting periods, X_{12} benign tumor, X_{11} family history of breast, X_9 Hormone therapy after menopause, X_8 use of contraceptives, X_{13} having radiation.

The point estimates of the model parameters are respectevely

$$-13.7837 \quad -0.9572 \quad 14.9975$$
$$1.9399 \quad 12.4450 \quad -8.0136 \quad 10.2826$$
$$4.4229 \quad 2.3040 \quad -1.8360 \quad 1.4577$$
$$2.6290.$$

6. Conclusion

This study applied fractional polynomial logistic regression model to analysis a breast cancer data, applying of fractional polynomial with logistic regression result in accurate model with increasing the classification performance.

The study stated that the FP model is a flexible technique in detecting the predictive effect of continuous variables. This method enhances the ability to evaluate the predictive ability of variables and improves model adequate and classification performance.

This paper proposed combining of FP and LR to create a hybrid method for Breast cancer prediction.

The study shows the best model for patients with breast cancer includes these risk factors are age, weight , age at starting periods, benign tumor, family history of breast, Hormone therapy after menopause, use of contraceptives and having radiation.

Finally, the study conclude that combining of FP with logistic regression model is very efficient method in detecting the predictive effect of continuous variable that resulted in improving model adequate and increasing classification performance.

Acknowledgments

I wish to thank master degree students of the departments of computer Sciences at college of computer sciences and information technologies for their encouragement, useful discussions, and interest.

References

1. Dreiseitl, Stephan, and Lucila Ohno-Machado. "Logistic regression and artificial neural network classification models: a methodology review." Journal of biomedical informatics 35.5 (2002): 352-359
2. Harrell, Frank. Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis. Springer, 2015.
3. Hosmer DW, Lemeshow S (2000) Applied logistic regression. Wiley series in probability and statistics, 2nd edn. Wiley, New York.
4. Menard S (2002) Applied logistic regression analysis, 2nd edn. Sage publications Inc, UK
5. Neter J, Kutner MH, Nachtsheim CJ, Wasserman W (1996) Applied linear statistical models, 4th edn. Irwin, Chicago.
6. Ryan TP (2008) Modern regression methods, 2nd edn. Wiley, New York
7. Brzezinski JR Knafl GJ (1999) Logistic regression modeling for context-based classification. In: Proceedings tenth international workshop on database and expert systems applications 1999, pp 755–759. doi:10.1109/DEXA.1999.795279
8. Liao JG, Chin K-V (2007) Logistic regression for disease classification using microarray data: model selection in a large p and small n case. Bioinformatics 23(15):1945–1951
9. Sartor MA, Leikauf GD, Medvedovic Lrpath M (2008) A logistic regression approach for identifying enriched biological groups in gene expression data. Bioinformatics 25(2):211–217
10. Asgary MP, Jahandideh S, Abdolmaleki P, Kazemnejad A (2007) Analysis and identification of b-turn types using multinomial logistic regression and artificial neural network. Bioinformatics 23(23):3125–3130.
11. Musa, Abdallah Bashir. "A comparison of ℓ1-regularizion, PCA, KPCA and ICA for dimensionality reduction in logistic regression." International Journal of Machine Learning and Cybernetics 5.6 (2014): 861-873.
12. Royston, J. P. "mfpa: Extension of mfp using the ACD covariate transformation for enhanced parametric multivariable modeling." The Stata Journal 16.1 (2016): 72-87.
13. Sauerbrei, Willi, et al. "Multivariable regression model building by using fractional polynomials: description of SAS, STATA and R programs." Computational Statistics & Data Analysis 50.12 (2006): 3464-3485.
14. Binder, Harald, Willi Sauerbrei, and Patrick Royston. "Comparison between splines and fractional polynomials for multivariable model building with continuous covariates: a simulation study with continuous response." Statistics in Medicine 32.13 (2013): 2262-2277.
15. Royston, P., and W. Sauerbrei. "Building multivariable regression models with continuous covariates in clinical epidemiology with an emphasis on

fractional polynomials." Methods Archive 44.4 (2005): 561-571.

16. Sauerbrei, Willi, and Patrick Royston. "Building multivariable prognostic and diagnostic models: transformation of the predictors by using fractional polynomials." Journal of the Royal Statistical Society: Series A (Statistics in Society) 162.1 (1999): 71-94.
17. Morris, Tim P., et al. "Combining fractional polynomial model building with multiple imputation." Statistics in medicine 34.25 (2015): 3298-3317.
18. Royston, P., and W. Sauerbrei. "Building multivariable regression models with continuous covariates in clinical epidemiology with an emphasis on fractional polynomials." Methods Archive 44.4 (2005): 561-571.
19. Sauerbrei, Willi, Patrick Royston, and Harald Binder. "Selection of important variables and determination of functional form for continuous predictors in multivariable model building." Statistics in medicine 26.30 (2007): 5512-5528.
20. Silke, B., et al. "An improved medical admissions risk system using multivariable fractional polynomial logistic regression modelling." QJM: An International Journal of Medicine 103.1 (2010): 23-32.
21. Royston, Patrick, and Willi Sauerbrei. Multivariable model-building: a pragmatic approach to regression anaylsis based on fractional polynomials for modelling continuous variables. Vol. 777. John Wiley & Sons, 2008.
22. Royston, Patrick, and Douglas G. Altman. "Regression using fractional polynomials of continuous covariates: parsimonious parametric modelling." Applied statistics (1994): 429-467.S
23. Ambler, Gareth, and Patrick Royston. "Fractional polynomial model selection procedures: investigation of Type I error rate." Journal of Statistical Computation and Simulation 69.1 (2001): 89-108.
24. Musa, A.B. Int. J. Mach. Learn. & Cyber. (2013) 4: 13. https://doi.org/10.1007/s13042-012-0068-x