Application of Logistic Regression Modeling Using Fractional Polynomials of Grouped Continuous Covariates

M. U. Muhammad[1]; O. E. Asiribo[2]; M. N. Sohail[3]

[1]Department of Statistics,
Kano State University of Science and Technology,
Wudil, Nigeria.
e-mail: musaubamuhammad@gmail.com[1];

[2]Department of Statistics,
Federal University of Agriculture Abeokuta,
Abeokuta, Nigeria.

[3]College of Information and Technology,
Yanshan University Qinhuangdao,
Peoples' Republic of China,
China.
e-mail: mn.sohail@stumail.ysu.edu.cn[3]

Abstract **—--In this paper, we studied a logistic regression modeling using fractional polynomials of grouped continuous covariates. Instead of the usual linear predictor, a fractional polynomial model was fitted to model the hypertensive status of diabetic patient on their ages (x_1) and occupation status (x_2) as covariates. It was observed that for (x_1) the algorithm for the selection of factors with significant effect converged at $\phi(3, 3)$ with model terms deviance of 113.74 and Log-likelihood value of -56.51. For (x_2) the algorithm for selection of factors with significant effect converged at $\phi(-2, 3)$, with model terms deviance of 112.40 and Log-likelihood value of -56.20. This implies that, the second model which is a fractional polynomial model fit the data better. Also, the model is adequate enough since it produces less deviance and larger Log-likelihood values.**

Keywords-*Algorithm, fractional polynomials, logistic regression.*

I. INTRODUCTION

Sometimes in statistics, the interest might be on a simple approximation for smoothening relationships between variables and such relationships maybe known but complicated or completely unknown. Statistics studies include the collection and analysis of data on one or more variables. Often multiple regression analyses are used to model such data sets which may include only linear terms in the covariate(s). In most applications, the choice of the model building is based on simple linear effect modeling approach, but the linearity assumption may be questionable.

Royston and Sauerbrei (2008) stated that, most users of multiple regression or analysis of covariance with such data sets include only linear terms in the covariate(s). In other words each covariate(X) appears in the model as a term of the form βX. If the curvature in the relationship between the response variable(Y) and the X is suspected, the model maybe extended to include a quadratic term. In most applications, a choice is made between linear and quadratic, with cubic or higher order polynomials being rarely used.

It has been recognized that conventional low order polynomials do not always fit the data well. Higher order polynomials tend to fit the data more closely but may fit badly at the extremes of the observed range of X (Royston and Altman, 1994). They also, stated that for models with more than one X-variable, there is considerable difficulty in estimating the powers reliably. They believed that estimation of the precise power(s) is unnecessary because the likelihood surface is usually nearly flat near maximum, even though Y may not be linear in X^p.

In this paper the fractional polynomial regression model given by Royston and Altman model will be considered. By fitting a logistic regression model using fractional polynomials of grouped continuous covariates

whose power terms are restricted to a small predefined set of integer and non-integer values.

Model Adequacy

Model testing is commonly used to prove the validity of a model and the tests are typically presented as evidences to promote its acceptance and usability (Sterman, 2002).
Model adequacy can be checked using R^2 - the "coefficient of determination" which is a measure of the amount of variation in the data accounted for by the regression model. Also, one may use Adjusted-R^2 (adjusted coefficient of determination), Prediction Error Sum Of Squares (PRESS) and R^2 - Prediction (prediction coefficient of determination).

Royston and Altman (1994) used the deviance and likelihood methods in selecting the best fractional polynomial model where a model with a small deviance or large log-likelihood value is best. They extended the deviance method to deviance difference which is the difference in deviance of two fractional polynomials of varying degrees and they term this difference as gain (G). A model produces largest gain is considerable the best fit.

In this work, the model adequacy methods to be adopted are deviance difference or gain (G) and log-likelihood function because these measures are adequate and simple to interpret.

II. Materials and Methods

The data used for this research were obtained from the Health Records Department, Federal Medical Center, Birnin-kudu, Jigawa State, which were on 104 patients who came for a checkup for their Hypertensive status (Positive or Negative) from October 2016 to February 2017. The variables considered in the study were Age, hypertension status and occupational status of the patients. Hypertensive status (Positive or Negative) is the dependent variable while age and occupation are the independent variables. STATA 12.0 was used for analysis in this work.

A. Logistic Regression

Logistic regression, a special case of a generalized linear model (GLM) and is concerned with modeling binary responses. The multiple linear logistic regression model with covariates $x_1, x_2, ..., x_k$ asserts that the probability p of occurrence of a binary event y of interest. For example "death" in a case-control study may be represented by

$$LogitP = \frac{p}{1-p} = \beta_0 + \sum_{j=1}^{k} \beta_j x_j \qquad (1)$$

where; $\frac{p}{1-p}$ is known as the odds of an event. Suppose y takes the values 1 for an event and 0 for a nonevent, hence y has a Bernoulli distribution with probability parameter (and expected value) p.

B. Fractional Polynomials

A fractional polynomial of degree m is defined to be the function

$$\phi_m(X;\xi,p) = \xi_0 + \sum_{j=1}^{m} \xi_j X^{(p_j)} \qquad (2)$$

where m is a positive integer, $\mathbf{p} = (p_1, p_2, ...p_m)$ is a real-valued vector of powers with $p_1 < \ldots < p_m$ and $\xi = (\xi_0, \xi_1, ..., \xi_m)$ are real valued-coefficients. The round bracket notation signifies the Box-Tidwell transformation (Royston and Altman, 1994),

Royston and Altman (1994) gave an extension of equation (3) to the case of equal powers, that is $m > 1$ and $p_i = p_j$ for at least one pair of distinct indices (i, j), $1 \le i, j \le m$. For $m = 2$, $(i, j) = (1, 2)$ and $\mathbf{p} = (p_1, p_1)$, we have

$$\phi_2(X;\xi,p) = \xi_0 + (\xi_1 + \xi_2)X^{(p_1)} \qquad (3)$$

a fractional polynomial of degree 1, not 2. Hence, the limit as p_2 tends to p_1

$$\xi_0 + \xi_1 X^{(p_1)} + \xi_2 X^{(p_1)}(X^{(p_2-p_1)} - 1)/\, p_2 - p_1 \qquad (4)$$

will be reduced to

$$\xi_0 + \xi_1 X^{(p_1)} + \xi_2 X^{(p_1)} \ln X \qquad (5)$$

$$\lim_{p_2 - p_1} X^{(p_2-p_1)-1} = X^{-1} = \ln X$$

Hence, equation (4) becomes (5) as

$$\xi_0 + \xi_1 X^{(p_1)} + \xi_2 X^{(p_1)} \ln X$$

which is a three parameter family of curves. The generalization of equation (5) for $m>2$ and $p_1 = \ldots. = p_m$, can be expressed as

$$\xi_0 + \xi_1 X^{(p_1)} + \sum_{j=2}^{m} \xi_j X^{(p_1)} (\ln X)^{j-1} \qquad (6)$$

For arbitrary powers $p_1 \le p_2 \le .... \le p_m$, we set $H_0(\mathbf{X}) = 1$, $p_0 = 0$ and combine definition (2) with expression (6) to obtain an extended definition

$$\phi_m(X;\xi,p) = \sum_{j=0}^{m} \xi_j H_j(X) \qquad (7)$$

where for $j = 1, \ldots, m$

$$H_j(X) = \begin{cases} X^{(p_j)} & \text{if } p_j \neq p_{j-1}, \\ H_{j-1}(X) \ln X & \text{if } p_j = p_{j-1}. \end{cases} \quad (8)$$

The recurrence relation in equation (8) for H_j (**X**) in terms of H_{j-1}(X) when $p_j = p_{j-1}$ is a representation of the functional part of the equation (6) and makes computer evaluation of fractional polynomials straightforward. H_j(X) can be written as a vector function **H** (**X**) = (H_0, H_1, . . .,H_m). Expressions (7) and (8) are the full (and most concise) definition of a fractional polynomial of degree *m*. "Royston and Altman (1994)".

III. Performance Measure

Deviance Measures of Model Fitness

The deviance *D* is one of the measures of assessing the adequacy of model fit. The log-likelihood can be expressed in terms of the mean parameter μ and the log-likelihood ratio which is the scaled deviance expressed as

$$\mathbf{D}^*(y\ \mu) = -2\left(l(\mu; y) - l(\mu_{\max}; y)\right) \quad (9)$$

where, $l(\mu; y)$ is the log-likelihood under the model; $l(\mu_{\max}; y)$ is the log-likelihood under the maximum achievable (saturated) model.

For generalized linear models, the scaled deviance can be expressed as

$$D^*(y; \mu) = \frac{1}{\phi} D(y; \mu) \quad (10)$$

where, D(y; μ) is the residual deviance for the model and it's the sum of individual deviance contributions and ϕ is the dispersion parameter.

Royston and Altman (1994) stated that for a given *m*, the best power vector $\tilde{\mathbf{p}} = (\tilde{p}_1, \ldots, \tilde{p}_m)$ is that associated with the model with the highest likelihood or equivalently with the lowest deviance *D*. Thus $\tilde{\mathbf{p}}$ may be regarded as the maximum likelihood estimate (MLE) of **p** over the restricted parameter space based on *S*.

Also, it is convenient to use the deviance *D*(1, 1) associated with the straight line model ϕ_1 (**X**; 1) that is *m* = 1, *p* =1 as a base line for reporting the deviances of other models. Hence, a *gain G* for a model can be defined on a set of data as the deviance for ϕ_1 (**X**; 1) minus that for the model in question:

$$G = G(m, \mathbf{p}) = D(1, 1) - D(m, \mathbf{p}) \quad (11)$$

G is the difference between two deviances of varying degrees. The larger gain indicates a better fit.

IV. Results And Discussion

Table 1: Effect and Parameter Estimation for Age and Occupation on Patients' Hypertensive Status.

Variables	Odd ratio	Std. Error	P-value	95% Conf. Interval
Age group1	0.89097	0.11119	0.355	0.69763 – 1.13789
Age group2	1.09509	0.10023	0.321	0.91526 – 1.31027
Occup. Status	0.60509	0.12293	0.013*	0.40635 –0.90104
Constant	2.37507	0.67923	0.002*	1.35596 – 4.16013

Deviance: 113.74 Best Powers of Grouped Aged models fit: 3, 3. *The values "3 and 3" represent the power of covariates in the model at which the algorithm converges.*

Log likelihood = -56.51

χ^2-value = 8.13

P – value = 0.043

Approximately 2.6 was subtracted from occupation status, (x^3 – 8.835),(x^3 lnx – 6.417) from age, to improve the scaling of the regression coefficients "(Odd Ratio), Where x is age".

Table 1 present the result of a logistic regression using fractional polynomial model. And the odd ratio selection algorithm for the covariate (age and occupation) on patients' hypertensive status, when age was fixed at 5% level of significance, occupation with coefficient of 0.89097 is significant with P-value of 0.013. The algorithm for selection variables with significant effect converged with the deviance for the model with all terms of *113.74*.

Hence the model is

$$H = -14.5191 + 0.8919X^3 1.0950X^3 \quad (12)$$

where, *H* is hypertension and *x* is grouped age

Table 2: Effect and Parameter Estimation for Occupation and Age on Patients' Hypertensive Status.

Variables	Odd ratio	Std. Error	P-value	95% Conf. Interval
Occup. Status 1	0.63022	7.77871	0.036*	0.56089 – 70.81322
Occup. Status 2	0.99427	0.00758	0.451	0.97953 – 1.00923
Age group	1.12318	0.29005	0.653	0.67708 – 1.86320
Constant	2.48086	0.75093	0.003*	1.37074 – 4.49005

Deviance: 112.40 Best Powers of Grouped Aged models fit: -2, 3. *The values "-2 and 3" represent the powers of covariates in the model at which the algorithm converges*. Log likelihood = -56.20

χ^2-value = 8.76

P – value = 0.036

Approximately 2.1 was subtracted from age, $(x^{-2}$– 0.148),$(x^3$ – 17.498) from occupation status, to improve the scaling of the regression coefficients "(odd ratio), Where x is occupation status".

Table 2 presents the result of a logistic regression using fractional polynomial. The odd ratio selection algorithm for the covariate (occupation and age) on patients' hypertensive status, when the occupation was fixed, at 5% level of significance occupation with a coefficient of 0.63022 is significant with a p-value of 0.036. The best power for the covariates was -2 and 3. The best power for the covariates was "-2 and 3" of which (-2) is a fractional polynomial power, this indicates that fractional polynomials fit our data best. We do not suggest the fractional polynomial models should supplant existing method; rather, they should be seen as a convenient. The algorithm for selection of factors with significant effect converged with the deviance for the model with all terms of *112.40*.

Hence the model is

$$H = -15.9878 + 0.6303X^{-2} + 0.9943X^{3} \quad (13)$$

where, H is hypertension and x is occupation status.

V. Conclusion

Statistical models approximate the relationships between variables. In general, the main objective of this research is to fit a logistic regression model using fractional polynomials of grouped continuous covariates. From the fitted models, terms in the models give a good description of the relationship between the variables. As we have seen from expression (12) and (13), the fractional polynomials gave better fit than conventional polynomials. In the same vein, the model expressed in (13) described as the best fit. Hence it produces less deviance of 112.40 and higher Log-likelihood value of – (56.20). Thus, it converged at $\phi(-2,3)$, and produced a gain G of 1.34.

References

[1] Kleinbaum, D. G. and Klein, M. (2010). Logistic Regression. *A Self-Learning Text, 3rd ed. Springer.*

[2] Royton, P. and Altman, D. G. (1994). Regression using fractional polynomials of continuous covariates: parsimonious parametric modeling (with discussion). *Applied Statistics*, 43(3):429–467.

[3] Royston P, Sauerbrei W. (2008). Multivariable Model-Building: *A Pragmatic Approach to Regression Analysis Based on Fractional Polynomials for Continuous Variables*. Wiley: New York.

[4] Sterman, J. D (2002). "All models are wrong: Reflections on Becoming a system Scientist. *System Dynamics review* 18, pp. 501-531