



BioTechnology

An Indian Journal

FULL PAPER

BTAIJ, 8(3), 2013 [292-297]

Modeling continuous health risk factors by using fractional polynomials transformation

Dalin Zeng¹, Youquan Xu^{2*}, Xiancun Hu³

¹School of Economic Management, Tongji University, Shanghai, 200092, (CHINA)

²School of Management Engineering, Shandong Jianzhu University, Jinan, Shandong, 250101, (CHINA)

³School of Architecture and Building, Deakin University, Australia, (CHINA)

E-mail: zdllw8102@163.com; yqxu@sdjzu.edu.cn; huxiancun@163.com

ABSTRACT

Assessing health risk factors often involves observational data that with heterogeneous patient background that information is limited to researchers. Thus forms a curvature, sometimes in complicate shape, on the continuous predictive variables in fitting models. Fractional Polynomials (FP) transformation provide a flexible model fitting framework account such curvature. In this paper we assess the utilization of FP model fitting framework in health risk factor assessment.

© 2013 Trade Science Inc. - INDIA

KEYWORDS

Health risk;
Factors;
Fractional polynomials;
Transformation.

INTRODUCTION

Regression analysis can be used to determine significant predictors of a response, as well as for predicting the response variable from a set of covariates (factors). In many situations, researchers would like to determine a relationship between response variable (Y) and predictive variables (covariates, X_1, X_2, \dots, X_m) so that they can assess the effects of predictors on response as well as predict the mean response based on a certain set of values of predictors.

Through ordinary least squares (linear regression only) or maximum likelihood, the model parameters could be estimated from a set of observed values of Y s and corresponding X s. Through such fitted model, a mean response of Y could be estimated according to values of X s.

However, in modeling health outcomes, research-

ers often deal with observational data that with limited patient background information. These heterogeneity forms curvature shape, non-linearity between outcome and predictors, in continuous risk factors (e.g Age, Year of Disease Durations, etc.) that creates difficulties in fitting models in regression analyses. Fractional Polynomials (FP) transformation provides a flexible yet powerful solution.

FRACTIONAL POLYNOMIAL REGRESSION MODELS

When the relationship between a response and covariates is not linear, researchers may fit nonlinear polynomial models. However, in many cases, even polynomial models may not fit well or need to fit in a much higher order (power³) because of the limited shapes of low order polynomials. However, higher power terms

fit poorly in extreme values, which limit their application. On the other hand, conventional polynomial models usually include lower order terms and/or linear terms when a higher order term appears. This restrains the flexibility in choosing polynomial terms because linear and quadratic functions are limited in their range of curve shapes, whereas cubic and higher order curves often produce undesirable edge effects and waves^[1].

Royston et al.^[1] introduced Fractional Polynomials (FP) models that can be incorporated into any regression model to provide a flexible approach for modeling nonlinear relationships between a response and continuous covariates^[1,2].

The Fractional Polynomials regression model for a given continuous variable X is defined as:

$$\phi_m(X; \beta, \mathbf{p}) = \beta_0 + \sum_{j=1}^m \beta_j X^{(p_j)}$$

where (1) m is a positive integer; (2) p_1, p_2, \dots, p_m are real-valued vector of powers with $p_1 < p_2 < \dots < p_m$; (3) $\beta_1, \beta_2, \dots, \beta_m$ are real-value coefficients; (4)

$$X^{(p_j)} = \begin{cases} X^{(p_j)}; & p_j \neq 0 \\ \ln X; & p_j = 0 \end{cases}; \text{ (5) } p \text{ are all possible } m\text{-}$$

tuples from $\mathcal{P} = \{-2, -1, -1/2, 0, 1/2, 1, 2, \max(3, m)\}$.

FPs are a family of regression models that use a subset of powers on the covariates from a defined set of powers (suggested $[-2, -1, -1/2, 0, 1/2, 1, 2, 3, \dots]$, where 0 means using logarithm)^[1]. Note that in the definition of Fraction Polynomials regression model above the mean response can be Y or a link function transformed variable. The following two examples are both Fractional Polynomials regression models.

$$(1) \phi_m(X; \beta, \{-0.5\}) = \beta_0 + \beta_1 X^{-1/2}$$

$$(2) \phi_m(X; \beta, \{-1, 0.5\}) = \beta_0 + \beta_1 X^{-1} + \beta_2 X^{1/2}$$

Although many powers and power combinations can be chosen, in practice fewer than two powers from the defined set of powers for each predictive variable are sufficient in fitting most practical models. Meanwhile, we generally do not need to consider powers that are more than 3 or less than -2 since such powers do not provide much benefit compared to their complexity.

The degree of a Fractional Polynomials model is

the maximum number of polynomial terms used on a predictor in a fitted model. Here we denote $\phi(m, p)$ denotes a Fractional Polynomials model, where m is the degree of the model, and p is the set of Fractional Polynomial terms listed inside braces.

Under this style of notation, the above two examples of FP models on X are denoted as:

$$(1) \phi(1, \{-0.5\})$$

$$(2) \phi(2, \{-1, 0.5\})$$

Because many combinations of powers may be used to fit a given set of data, it is critical to determine a best fitting model. For such determination, all possible models with different combinations of powers are fitted using maximum likelihood.

We further denote $D(m, p)$ denotes the deviance of a Fractional Polynomials Regression model $\phi(m, p)$, where $D(m, p) = -2 \log$ -likelihood of the model (in Maximum Likelihood estimation framework or MLE) or Sum of Squares (in Ordinary Least Squares estimation framework or OLS). Throughout the paper, we only demonstrate $D(m, p)$ in MLE. The optimum model determination on OLS is similar and will be left to the reader.

Under this style of notation, deviances of the above two examples of FP models on X are denoted as:

$$(1) D(1, \{-0.5\})$$

$$(2) D(2, \{-1, 2\})$$

In addition, we denote D_{\max} the maximum deviance difference from all Fractional Polynomials Regression models to the model with only linear term. $D_{\max} = \max(D(1, \{1\}) - D(m, p))$, for all possible p on a certain m .

For a given m , the best power vector $\tilde{p} = (\tilde{p}_1, \tilde{p}_2, \dots, \tilde{p}_m)$ is the one associated with the model with the highest likelihood (the lowest deviance.). Thus \tilde{p} may be regarded as the maximum likelihood estimator of p over the restricted parameter space based on \mathcal{P} . The value $D(m, p) - D(m, \tilde{p})$ has an approximate χ^2 distribution with m degree of freedom for large n . In determining the degree of FP, Royston suggested using $\Delta D = D(m, \tilde{p}) - D(m+1, \tilde{p}) > \chi^2_{2; 0.90} (=4.7)$ as a rule for preferring models with degree $m+1$ to those with degree m . Also suggested is the benchmark significant level of test $\alpha=0.1$ ^[1,2].

In practice, since usually only models up to degree

FULL PAPER

two are considered when working with FP models, it is convenient to use the deviance of linear first order term ($D(1, \{1\})$) as a baseline. The maximum deviance difference comparing the fit of a model with that of a straight line ($p = 1$) is distributed approximately, for large n , as a χ^2 distribution with 1 or 3 degrees of freedom for first- or second-degree models, respectively. Also suggested is the benchmark significant level of test $\alpha=0.05$ [2].

Under such conditions, we can recursively fit models with all possible combinations of powers (up to $\binom{8}{1} + \binom{8}{2} = 36$). Those models with deviance greater than first order FP $\phi(1, \{1\})$ become candidate FPs. We then find the model that has maximum deviance difference D_{max} , and finally determine if $D_{max} > \chi^2_{\alpha=0.05}(1$ or 3 df for first or second degree FP, respectively). If the test is not significant (p -value > 0.05) then the $\phi(1, \{1\})$ model is sufficient. Otherwise the model with deviance D_{max} will be the optimal FPs model^[1,2].

APPLICATION IN HEALTH RISK FACTORS ASSESSMENT

We apply the FP transformation in a Logistic regression model using the data similar to Brodie et al.^[3], to estimate the effect of Door-to-Balloon Time (DBT) on Survival of ST-elevated Myocardial Infarction Patient under Primary Percutaneous Coronary Intervention. We compare our results to the publications, using a Cox Proportional Hazard model, and receive similar conclusions. Our purpose is to demonstrate the esti-

mation framework and model fitting process rather than to find out the health outcome implications of the risk factors.

DATA AND RESULTS

Study data is a randomly selected sample from studies described in^[4]. Summary Statistics are in TABLE 1.

We follow the same methodology and model determination described above. All model estimations are carried and compute by programs authors developed in statistical software R. First we assess the First Degree Fractional Polynomials (FDFP) fitting model, FP applies only on DBT, estimation results are shown in TABLE 2. The best fitted FP is $\phi(1, \{-2\})$, with Deviance Difference of 9.75, a p -value of χ^2 at 0.001792 (p -value of χ^2 column in the table). In the TABLE and what follows, we only show p -values when they are statistically significant at 5% level.

Next we assessed the Second Degree FPs (SDFP) fitting model, estimations are shown in TABLE 3. From

TABLE 1 : Summary statistics of outcomes and predictors.

Measurements	Possible Values (unit)	Role in model
Death in hospital not lab	(0,1) 0 = alive, 1 = dead	Response
Door-to-balloon Time	Continuous positive value (hour)	Predictor
Age of the patient	Continuous positive value (year)	Predictor
Sex of the patient	0 = male, 1 = female	Predictor
Prior Myocardial infarction	0 = No, 1 = Yes	Predictor
Presence of Diabetes	0 = No, 1 = Yes	Predictor
Current or former Smoker	1 = current, 2 = Former	Predictor

TABLE 2 : Logistic regression with fractional polynomials on dbt, first degree of fractional polynomials.

FP	Power	p -value of coefficient	Survival (Odds Ratio)			$D(I, p)$	Deviances			DF	p -value of χ^2
			Point	Wald Confidence (95%)			Drop-in-Deviance (Likelihood Ratio)	Deviance Difference $D(1, \{1\}) - D(I, p)$			
				Lower	Upper						
$\phi(1, \{3\})$	3.00	0.5545	1	1	1	1058.979	334.3671	-3.106	1		
$\phi(1, \{2\})$	2.00	0.799	1	0.997	1.002	1059.316	334.05	-3.443	1		
$\phi(1, \{1\})$	1.00	0.0483	0.95	0.902	1	1055.873	337.4728	0	1		
$\phi(1, \{0.5\})$	0.50	0.005	0.684	0.524	0.892	1052.071	341.2756	3.802	1	0.051191	
$\phi(1, \{0\})$	0.00	0.001	0.629	0.477	0.83	1048.765	344.5812	7.108	1	0.007674	
$\phi(1, \{-0.5\})$	-0.50	0.0007	5.263	2.026	13.67	1046.897	346.4492	8.976	1	0.002736	
$\phi(1, \{-1\})$	-1.00	0.0008	3.354	1.651	6.811	1046.124	347.2223	9.749	1	0.001794	
$\phi(1, \{-2\})$	-2.00	0.0028	2.34	1.341	4.084	1046.12	347.2238	9.751	1	0.001792	

TABLE 3 : Logistic regression with fractional polynomials, dbt, second degree of fractional polynomials (SDFP)

Powers			Deviances				Survival (Odds Ratio by p)			Survival (Odds Ratio by q)			
FP	p	q	D(m, p)	Drop in Deviance (Likelihood Ratio)	Deviance Difference D(1, {1}) -D(m, p)	DF	p-value of χ^2	Point	Wald Confidence (95%)		Point	Wald Confidence (95%)	
									Lower	Upper		Lower	Upper
$\phi(1, \{1\})$	1	.	1055.873	337.473	0	1		0.95	0.902	1			
$\phi(2, \{1, 3\})$	1	3	1044.812	348.534	11.061	3		0.831	0.748	0.923	1	1	1.001
$\phi(2, \{1, 2\})$	1	2	1044.985	348.361	10.888	3		0.77	0.664	0.892	1.011	1.002	1.019
$\phi(2, \{1, 0.5\})$	1	0.5	1046.09	347.256	9.783	3		1.28	1.037	1.579	0.206	0.073	0.581
$\phi(2, \{1, 0\})$	1	0	1046.304	347.042	9.569	3		1.088	0.975	1.215	0.435	0.251	0.756
$\phi(2, \{1, -0.5\})$	1	-0.5	1046.278	347.068	9.595	3		1.032	0.952	1.12	7.776	1.939	31.182
$\phi(2, \{1, -1\})$	1	-1	1046.091	347.255	9.782	3		1.006	0.939	1.079	3.511	1.47	8.385
$\phi(2, \{1, -2\})$	1	-2	1045.734	347.612	10.139	3		0.981	0.924	1.041	2.186	1.223	3.907
$\phi(1, \{3\})$	3	.	1058.979	334.367	-3.106	1		1	1	1			
$\phi(2, \{3, 2\})$	3	2	1046.324	347.022	9.549	3		1.001	1	1.002	0.976	0.961	0.992
$\phi(2, \{3, 0.5\})$	3	0.5	1044.376	348.97	11.497	3		1	1	1	0.495	0.341	0.717
$\phi(2, \{3, 0\})$	3	0	1044.234	349.112	11.639	3	0.008728	1	1	1	0.541	0.393	0.745
$\phi(2, \{3, -0.5\})$	3	-0.5	1044.316	349.03	11.557	3	0.009066	1	1	1	6.673	2.41	18.477
$\phi(2, \{3, -1\})$	3	-1	1044.527	348.819	11.346	3		1	1	1	3.662	1.761	7.615
$\phi(2, \{3, -2\})$	3	-2	1045.271	348.075	10.602	3		1	1	1	2.395	1.361	4.215
$\phi(1, \{2\})$	2	.	1059.316	334.03	-3.443	1		1	0.997	1.002			
$\phi(2, \{2, 0.5\})$	2	0.5	1045.026	348.32	10.847	3		1.005	1.001	1.01	0.432	0.276	0.678
$\phi(2, \{2, 0\})$	2	0	1045.175	348.171	10.698	3		1.003	1	1.007	0.514	0.36	0.735
$\phi(2, \{2, -0.5\})$	2	-0.5	1045.329	348.017	10.544	3		1.002	0.999	1.005	7.04	2.39	20.741
$\phi(2, \{2, -1\})$	2	-1	1045.487	347.859	10.386	3		1.001	0.998	1.004	3.678	1.73	7.819
$\phi(2, \{2, -2\})$	2	-2	1046.04	347.306	9.833	3		1	0.998	1.003	2.371	1.344	4.182
$\phi(1, \{0.5\})$	0.5	.	1052.071	341.275	3.802	1		0.684	0.524	0.892			
$\phi(2, \{0.5, 0\})$	0.5	0	1046.654	346.692	9.219	3		2.116	0.743	6.028	0.303	0.106	0.871
$\phi(2, \{0.5, -0.5\})$	0.5	-0.5	1046.498	346.848	9.375	3		1.201	0.675	2.136	8.989	1.279	63.166
$\phi(2, \{0.5, -1\})$	0.5	-1	1046.123	347.223	9.75	3		0.995	0.645	1.534	3.323	1.147	9.629
$\phi(2, \{0.5, -2\})$	0.5	-2	1045.238	348.108	10.635	3		0.85	0.61	1.185	1.999	1.082	3.692
$\phi(1, \{0\})$	0	.	1048.765	344.581	7.108	1	0.007674	0.629	0.477	0.83			
$\phi(2, \{0, -0.5\})$	0	-0.5	1046.53	346.816	9.343	3		1.421	0.447	4.52	16.325	0.343	777.587
$\phi(2, \{0, -1\})$	0	-1	1046.121	347.225	9.752	3		0.982	0.519	1.857	3.224	0.698	14.881
$\phi(2, \{0, -2\})$	0	-2	1045.085	348.261	10.788	3		0.81	0.544	1.206	1.813	0.904	3.635
$\phi(1, \{-0.5\})$	-0.5	.	1046.897	346.449	8.976	1	0.002736	5.263	2.026	13.67			
$\phi(2, \{-0.5, -1\})$	-0.5	-1	1046.12	347.226	9.753	3		0.873	0.012	61.711	3.687	0.176	77.159
$\phi(2, \{-0.5, -2\})$	-0.5	-2	1045.315	348.031	10.558	3		2.217	0.402	12.235	1.668	0.698	3.986
$\phi(1, \{-1\})$	-1	.	1046.124	347.222	9.749	1		3.354	1.651	6.811			
$\phi(2, \{-1, -2\})$	-1	-2	1045.652	347.694	10.221	3		1.899	0.312	11.55	1.533	0.429	5.481
$\phi(1, \{-2\})$	-2	.	1046.122	347.224	9.751	1	0.001792	2.34	1.341	4.084			

the table we can see the best fitted FP is $\phi(1, \{-0.5\})$, with Deviance Difference of 8.976, a p-value of χ^2 at 0.002736. As shown in the same table, even the best SDFP is not better than FDFP (also shown in TABLE 3.) So FDFP is better when only FPS on DBT is ap-

plied.

We then apply FPs to both DBT and Age, however making the two continuous variables in the risk factors. Shown in TABLE 4 is the SDFP on Age with DBT set on $\phi(1, \{-2\})$, then best FPs on DBT we

FULL PAPER

found in earlier step. As we can see in the table, we cannot find any significant transformation of FP. We also evaluate different combinations of FPs on DBT and Age but fail to find a statistically significant FP

transformation on Age, respectively (results not shown here). Since no significant FDFP on Age when FDFP (-2) present in DBT, we did not go further degree FPs on Age.

TABLE 4 : Logistic regression with fractional polynomials on age, with dbt⁻², first degree fractional polynomials on age

FP	Power	p-value of coefficient	Survival(Odds Ratio)			D(I, p)	Deviances			DF	p-value of χ^2
			Point	Wald Confidence (95%)			Drop-in-Deviance (Likelihood Ratio)	Deviance Difference D(1, {1}) -D(I, p)			
				Lower	Upper						
$\phi(1, \{3\})$	3.00	0.0001	1	1	1	1049.691	343.6552	-3.5686	1		
$\phi(1, \{2\})$	2.00	0.0001	0.999	0.999	1	1046.909	346.4375	-0.7863	1		
$\phi(1, \{1\})$	1.00	0.0001	0.935	0.921	0.949	1046.122	347.2238	0	1		
$\phi(1, \{0.5\})$	0.50	0.0001	0.341	0.266	0.436	1046.649	346.6971	-0.5267	1		
$\phi(1, \{0\})$	0.00	0.0001	0.014	0.005	0.039	1047.863	345.4828	-1.741	1		
$\phi(1, \{-0.5\})$	-0.50	0.0001	-	-	-	1049.802	343.5445	-3.6793	1		
$\phi(1, \{-1\})$	-1.00	0.0001	-	-	-	1052.48	340.8659	-6.3579	1		
$(1, \{-2\})$	-2.00	0.0001	-	-	-	1060.01	333.3388	-13.885	1		

Thus we reach our best fitting model: -2 transformation on DBT but linear in Age (no FP). Results are shown in TABLE 5. From the results we can see that

the longer Door-to-Balloon Time (DBT) the lower survival chance, Age follows the same pattern. The results are similar to^[4] in a survival analysis framework.

TABLE 5 : Summary of the final fitted model

Covariate	DF	Maximum Likelihood Estimates				Odds ratio		
		Estimate	Error	χ^2	Pr> χ^2	Estimates	95% Wald's Lower	CI Upper
Intercept	1	10.5952	0.6171	294.810	<.0001			
Door-to-balloon Time ⁻²	1	-0.3345	0.102	10.750	0.0010	0.629	0.477	0.830
Age of the patient	1	-0.067	0.0077	75.581	<.0001	0.935	0.921	0.949
Sex of the patient	1	-0.0510	0.1786	0.0816	0.7752	0.950	0.670	1.349
Prior Myocardial infarction	1	-2.5321	0.1877	181.955	<.0001	0.079	0.055	0.115
Presence of Diabetes	1	-0.3227	0.2047	2.4869	0.1148	0.724	0.485	1.082
OLDMI	1	-0.0454	0.2120	0.0459	0.8304	0.956	0.631	1.448
OLDCABGS	1	-0.5228	0.3199	2.6707	0.1022	0.593	0.317	1.110
VESLADZ	1	-0.6346	0.1689	14.1210	0.0002	0.530	0.381	0.738

ACKNOWLEDGEMENTS

This work was financially supported by the National Natural Science Foundation of China (71072046).

REFERENCES

[1] R.Royston, D.G.Altman; Regression using Fractional Polynomial of Continuous Covariates: Parsimonious Parametric Modeling, Applied Statistics, **43(3)**, 429-467 (1994).

[2] R.Royston, G.Ambler, W.Sauerbrei; The Use of fractional polynomial to model continuous risk variables in epidemiology, International Journal of Epidemiology, **28**, 964-974 (1999).

[3] P.McCullagh, J.A.Nelder; Generalized Linear Model, 2nd Edition, CRC Press, 26-32 (1989).

[4] B.R.Brodie, C.Hansen, T.D.Stuckey, S.J.Richter et al.; Door-to-Balloon Time With Primary Percutaneous Coronary Intervention for Acute Myocardial Infarction Impacts Late Cardiac Mortality in High Risk Patients and Patients Presenting Early After the Onset of Symptoms, Journal of the American College of Cardiology, **47(2)**, 289-295 (2006).

- [5] R.Royston, D.G.Altman; Regression using Fractional Polynomial of Continuous Covariates: Parsimonious Parametric Modeling, *Applied Statistics*, **43(3)**, 429-467 (1994).
- [6] R.Royston, G.Ambler, W.Sauerbrei; The Use of fractional polynomial to model continuous risk variables in epidemiology, *International Journal of Epidemiology*, **28**, 964-974 (1999).
- [7] P.McCullagh, J.A.Nelder; *Generalized Linear Model*, 2nd Edition, CRC Press, 26-32 (1989);
- [8] B.R.Brodie, C.Hansen, T.D.Stuckey, S.J.Richter et al.; Door-to-Balloon Time With Primary Percutaneous Coronary Intervention for Acute Myocardial Infarction Impacts Late Cardiac Mortality in High Risk Patients and Patients Presenting Early After the Onset of Symptoms, *Journal of the American College of Cardiology*, **47(2)**, 289-295 (2006).