International Journal of Multidisciplinary and Current Research ISSN: 2321-3124
Research Article Available at: http://ijmcr.com

Marginal Regression Analysis with Two Basic Parameterizations and the Dependency Ratio of Maximum Likelihood Estimation

[1]G. V. Arunamayi, [2]K.Nagendra Kumar, [3]K.V.R.Murthy

Department of Statistics, Andhra University, Visakhapatnam-530 003, Andhra Pradesh. India
Department of Engineering Mathematics, Andhra University,Visakhapatnam-530003, Andhra Pradesh. India

Accepted 10 Sept 2016, Available online 22 Sept 2016, **Vol.4 (Sept/Oct 2016 issue)**

Abstract

In this paper we have developed an Marginal Regression for a bi-variate response for untreated state, diabetes mellitus is recognized by chronic evolution of concentration of glucose in the blood (Hyperglycaemia). This is sometimes accompanied by symptoms of serve thirst, profuse urination, weight loss, and stupor culminating in coma and death in the absence of effective treatment. The underlying caves of diabetes are the defective production or action of the hormone insulin. We are tested through SPSS software by taking 200 samples and we determined this model by care processing, dependent variable encoding classification table Omnibus tests of model coefficients, and we also developed a model summary, Model if term removed variables in the equation and variables not in the equation. From above study we notice that explanatory variables age and DH are the significant variable as the before.

Keywords: *Marginal Regression Analysis,Binary Logistic Regression, logit model, Odds Ratio, Model validation, Hosmer and Lemeshow Test.*

Marginal Regression Analysis

Introduction

To study the relationship of Diabetic Mellitus and B.P, Age, DH,TD and therefore, the binary response is bivariate. Regression models for the one-dimensional marginal probabilities of the response which in corporate association are termed marginal probabilities of the response which in corporate association are termed marginal regression models.

The relation between three basic parameterisation for a multivariate binary random variable.

A new coefficient of association, the dependence ratio, which has a simple interpretation and a natural generalization to associations higher than second order.

Marginal regression models are fitted by Maximum likelihood for the present data and for longitudinal data described by [1]Fitzmaurice and Laird (1993). Three different association structures are new, being defined in terms of dependence ratios.

Review of literature

[2]Guangyong Zou and Allan Donner (2004) studied confidence Internal Estimation of the Interclass Correlation Coefficient for Binary Outcome Data. They obtained closed-from asymptotic variance formulae for three point estimator of the interclass correlation coefficient that may applied to binary outcome data arising in clusters of variable size. Their result include a special case those that have previously appeared in the literature (Fleiss and Cuzick, 1979, Applied Physiological Measurement 3, 537-542; Bloch and kraemer,1989, Biometrics 45,269-287; Altaye, Donner, and Klar,2001, Biometrics 57, 584-588). Simulation results indicate that confidence intervals based on the estimator proposed by Fleiss and Cuzick provide coverage levels close to nominal over a wide range of parameter combinations.

[3]Nicole H. Augustin, Stefan Lang, Monica Musio and Klaus von Wilpert (2007) jointly monitored a survey which was carried out in 1994 in the forests of Baden-Wurttemberg, a federal state in the south-western region of Germany. The survey is a part of large monitoring scheme that has been carried out since the 1980s at different spatial and temporal resolutions to observe the increase in forest damage. One indicator for tree vitality is tree defoliation to observe the increase in forest damage. Once indicator for tree vitality is tree defoliation, which is mainly caused by intrinsic factors, age and stand conditions, but also by biotic (e.g. insects) and a biotic stresses (e.g. industrial emissions). In the survey, needle loss of pine-trees and many potential covariates are recorded at about 580 grid points of a 4 km x 4 km grid. The aim is to identify a set of predictors for needle loss and to investigate the relationships between the needle

loss and the predictors. The response variable needle loss is recorded as a percentage in 5% steps estimated by eye using binoculars and categorized into healthy trees (10% or less), intermediate trees (10-25%) and damaged trees 25% or more). They used a Bayesian cumulative threshold model with non-linear functions of continuous variables and a random effect for spatial heterogeneity, for both the non-linear functions and the spatial random effect we use Bayesian versions of P-splines as priors. Our methods are novel in that it deals with several non-standard data requirements; the ordinal response variable (the categorized version of needle loss), non-linear effects of covariates, spatial heterogeneity and prediction with missing covariates. The model is a special case of models with a geoadditive or more generally structured additive predictor. Inference can be based on Markov chain Monte Carlo techniques or mixed model technology.

[4]Brain J. Reich, James S. Hodges, and Bradley P. Carlin(2007) analyzed attachment loss data that can be used as conditionally autoregressive (CAR) prior distribution that smooth fitted values toward neighboring values. However, it may be desirable to have more than one class of neighbor relation in the spatial structure, so that the different classes of neighbor relations can induce different degrees of smoothing,. Adequate modeling of the spatial structure may improve the monitoring of periodontal disease progression. They showed that the prior distribution on these parameters has little effect on the posterior of the fixed effects but has a marked influence on the posterior of both the random effects and the smoothing parameters. Their analysis of attachment loss data also suggests that the spatial structure itself varies between individuals.

[5]Bo Cai and David B. Suson (2006) generalized linear mixed model (GLMM), which extends the generalized linear model (GLM) to incorporate random effects characterizing heterogeneity among subjects, is widely used in analyzing correlated and longitudinal data. Although there is often interest in identifying the subset of predictors that have random effects, random effects selection can be challenging, particularly when outcome distributions are non-normal. They proposed a fully Bayesian approach to the problem of simultaneous selection of fixed and random effects in GLMMs. Integrating out of random effects induces a covariance structure on the multivariate outcome data, and an important problem that we also consider is that of covariance selection. Their approach relies on variable selection-type mixture priors for the components in a special Cholesky decomposition of the random effects covariance. A stochastic search MCMC algorithm is developed, which relies on Gibbs sampling with Taylor series expansions used to approximate intractable integrals. Simulated data examples are presented for different exponential family distributions, and the approach is applied to discrete survival data from a time-to-pregnancy study.

[6]Jerome A. Dupuis (2006) considered the problem of estimating the number of species of an animal community. It is assumed that it is possible to draw up a list of species liable to be present in this community. Data are collected from quadrate sampling. The parameterization enables us to incorporate prior information on the presence detect ability, and spatial density of species. Moreover the elaborated procedures to build the prior distributions on these parameters from information furnished by external data. A simulation study is carried out to examine the influence of different priors on the performances of our estimator.

The Two Basic Parameterisations

Consider a bi-variate binary response for a cluster of size 2, denoted by $Y=Y_1,Y_2$. There are 2^2 possible responses, termed cells. Denote the $1x2^2$ Vector of cell probabilities by π, where $P_r(Y=y)= \pi_y$ and $\sum \pi_y = 1$.

Let $d=2^2-1$ and define the 1xd vector of sufficient statistic $s(y)=(y_1,y_2,y_1y_2)$. The multinomial distribution over the cells with probabilities π is a member of the exponential family. The 1xd vector of canonical parameters is denoted by $\psi = (\psi_1, \psi_2, \psi_{12})$ indexing the canonical parameters by the products in s(y).

Random response vectors from different clusters are assumed independent. The contribution to the log likelihood from a single cluster is

$$l = \log(\pi_y) = s(y)\ \psi^T - K(\psi) \quad (1)$$

Where s(y) is the sufficient statistic for ψ. If follows from (1) that

$$\pi_y = \exp\left\{s(y)\ \psi^T - K(\psi)\right\}.$$

Where $K(\psi+z)-K(\psi)$ is the cumulate generating function of S(Y) and from $E\left(\frac{\partial^2 l}{\partial \psi}\right) = 0$ that $\frac{\partial k\psi}{\partial \psi} = E\left\{s(y)\right\}$.

Denote the 1xd vector of expected values of the sufficient statistic by

$$\phi = E\left\{s(y)\right\} = (\phi_1, \phi_2, \phi_{12}) \quad (2)$$

The vector ϕ is termed the mean parameter ([3]Barndorff Nielsen and Cox,1944,P.7) or alternatively the moment parameter. The component of ϕ are the marginal joint success probabilities of all orders, in particular of $\phi_{12} = \pi_{(11)}$. The mean parameters have an interpretation which does not depend on 2. The canonical parameters do not have a similar invariance. For examle, ψ_{12} is the conditional log odds ratio for the pair (Y_1,Y_2), conditional on $Y_1=Y_2=0$.

A canonical parameter with three indicates is the log of the ratio of two conditional odds ratios etc.

To derive the mapping from ϕ to π let y be a fixed cell. The random counter l_v(Y) is binary taking the value 1 if Y=y and else 0. Two different algebraic expressions for I_y(Y) are useful: (i) as a single product of 2 factors, either Y_v if $y_v = 1$ or $(1\text{-}Y_v)$ if $y_v = 0$;

ii) As a sum of products after expanding all $(1\text{-}Y_v)$ factors.

For example, if y=(0,1) then

$$I_y(Y) = (1-Y_1)Y_2 = Y_2 - Y_1Y_2 \quad (3)$$

Clearly $\Pr(Y = y) = \pi_y = E\{I_y(Y)\}$ and taking expectation of the sum-of-products from gives the explicit expression for π_y in terms of ϕ. For y= (0, 1) we find form (3) that

$$\pi_y = \pi(0,1) = \phi_2 - \phi_{12} \quad (4)$$

Generally, ? is an affine linear transformation of ϕ. P.W.F. Smith explores further the connections between the three basic parameterizations in his 1990 university of Lancaster.

Marginal Regression Using the Mean Parameterization

A Marginal regression model consists of two parts.

i) The marginal success probabilities, $\phi_{v,}$ are modelled by

ϕ_v = Pr(Y_v=1)h(X_g ;β) (v=1,2) --- (5)

Where h is a given function, X_g is a vector of p explanatory variables and β is a vector of p regression coefficients, constant over all units within and between clusters. Marginal logistic regression is a special case, where h is the inverse of the log it function

ii) The association structure of the q=2 components of Y=(Y_1,Y_2) is specified.

[6]Fitzmaurice, Larid and Rotnitzky (1993) review and compare several different approaches to modelling the association structure. See also cary, zeger the second – order or pair wise associations, but avoid assumptions about the full joint distribution. Likelihood-based inference is then unavailable. The approaches specifying the full joint distribution all propose likelihood inference, but vary in the parameterisation used. The choice of parameterization determines whether data from clusters of different size can be analysed.

Our approach is most akin to Fitzmaurice and Larid (1993). They specify the joint distribution by the mixed parameter ϕ = (ϕ_1,ϕ_2,ψ_{12}). One useful feature of ϕ is that that the regression coefficients and the association parameters are orthogonal (Barndorff -Nielsen and Cox, 1994, P.64). Data from clusters of different size cannot be analysed since the association parameters are canonical and thus conditional.

The Mean Parameterisation and the Dependence Ratio:

We specify the 2-dimensional bi-variate distribution by the pure mean parameterization ϕ = (ϕ_1,ϕ_2,ϕ_{12}) and model association using dependence ratios defined in terms of the mean parameters. Consider a bi-variate binary response, (Y_1,Y_2), with mean parameters (ϕ_1,ϕ_2, ϕ_{12}). The dependence ratio, τ_{12} and log dependence ratio, $\lambda_{12,}$ are defined by

$$\tau_{12} = \frac{\phi_{12}}{\phi_1\phi_2}, \lambda_{12} = \log(\tau_{12}) \quad (6)$$

So $(\tau_{12} - 1)$x100 indicates how many percent greater is the probability of both Y_1 and Y_2 being successes compared to what it would be under independence. Note that $\tau_{12} = 1 \Leftrightarrow Y_1 \perp Y_2$, and the $\max(0, \phi_1 + \phi_2 - 1) \le \phi_{12} \le \min(\phi_1, \phi_2)$ which induces corresponding bounds on τ_{12} and λ_{12}.The odds ratio and the correlation coefficient are monotone increasing functions of the dependence ratio, with the correlation coefficient proportions to $\tau_{12} - 1$.

If we exchange success and failure for one of the binary components say Y_2, so that we study (Y_1,1-Y_2), then the odds ratio and the correlation coefficient change in sing but not in absolute value. In contrast, τ_{12} is mapped onto $\frac{(1-\phi_2\tau_{12})}{1-\phi_2}$. Further, the correlation coefficient and the odds ratio for (1-Y_1, 1-Y_2).

Hence, using the dependence ratios, to specify the association results in different models depending on whether 1 models success or failure.

Let $\tau = (\tau_1, \tau_2, \tau_{12})$, avector of length d-2. Association structures of interest are specified by constraints on the elements of τ. We denote the vector of parameters specifying the association structure by α. Examples of association structures defined either purely by restrictions on τ or via a factorization and restrictions are presented in the analysis of the data sets. In principal regression of τ on explanatory variables is possible, but

not illustrated in this paper. See Ekhlom(1991) for models of ϕ in terms of explanatory variables and restrictions without the transformation to dependence ratio.

Horizontal and Vertical Homogeneity

Horizontal homogeneity assumes equal dependence for any equal-sized subset of units $\lambda_{12} = \lambda(q-1)q$

A motivation for vertical homogeneity is that

$$\lambda_{12} = \log(\phi_{12}) - \{\log(\phi_1) + \log(\phi_2)\} \quad (7)$$

Is a measure of the step fromY_1 and Y_2 to (Y_1,Y_2) and

$$\lambda_{123} - \lambda_{12} = \log(\phi_{123}) - \{\log(\phi_{12}) + \log(\phi_3)\},$$

Is a measure of the step form (Y_1,Y_2) and Y_3 to (Y_1,$Y_2$$Y_3$). Vertical homogeneity assumes that these two steps are equal, which implies $\lambda_{123} = 2\lambda_{12}$ induction vertically gives a.

$$\begin{aligned} &\lambda_{12} = \ldots\ldots = \lambda_{(q-1)q} = \lambda, \\ &\lambda_{123} = \ldots\ldots = \lambda_{(q-2)(q-1)q} = 2\lambda, \\ &\quad . \\ &\quad . \\ &\quad . \\ &\lambda_{1\ldots\ldots q} = (q-1)\lambda \end{aligned} \quad (8)$$

Homogeneous association is defined as both horizontal and vertical homogeneity.

Maximum Likelihood Estimation

Likelihood inference for β is straight forward because π is an affine linear transformation of ϕ consider a data set with n=4 clusters, where the units from different cluster are independent and many vary in number. We fit the regression model, assuming a given association structure with parameter α. A cluster of q=2 units provides a realization from a multinomial distribution with 2^3 cells. Estimates of (β, α) are obtained by using macros which fit nonlinear models in GLIM4, given by Ekholm and Green(1994); see also Francis Green and Payne(1993). The parameters β and α are not orthogonal, but the inverse of the Fisher information matrix for(β,α) is routinely computed. Not all association structures are compatible with all regression models. Trying to fit seriously wrong models can lead to negative fitted probabilities. Careful choice of model and initial values often help.

Logistic Regression

Table: I

Unweigthed Cases	a	N	Percent
Selected cases	Including in analysis Missing cases Total	200 0 200	100.0 .0 100.0
Unselected cases		0	.0
Total		200	100.0

a) If weight ids in effect, see classification table total number of cases

Observed		Predicted		
		FG		Percentage Corrected
		.00	1.00	
Step 0 FG	.00	0	50	.0
	1.00	1	150	100.0
Overall Percentage				75.0

b) Constant is included in the model.

c) The cut value is .500

The classification table from the SPSS output result summarizes the observed group and the predicted group classification. The overall correctly specified group percentage is 75.

Table: III Variables in the equation

	B	S.E	Wald	df	Sig	Exp(B)
Step 0 Constant	1.099	.163	45.261	1	.000	3.000

Table: IV Variables not in the equation

	Score	df	Sig.
Step Variables			
AGE	4.962	1	.026
0	1.506	1	.220
	.146	1	.702
TD	9.868	3	.020
Overall Statistics			

Block 1: Method = Back word Stepwise (Likelihood Ratio)

Table: V Omnibus Tests of Model Coefficients

	Chi-squere	df	Sig.
Step 1step	10.236	3	.017
Block	10.236	3	.017
Model	10.236	3	.017
Step 2 a step	-.251	1	.616
Block	9.985	2	.007
Model	9.985	2	.007

a) Negative chi-Squares value indicates that the Chi-Squares value has decreased from the previous step.

Table: VII Classification table

Observed		Predicted		
		FG		Percentage Corrected
		.00	1.00	
Step 1	FG			
	.00	2	48	4.0
	1.00	2	148	98.7
Overall Percentage				75.0
Step 2	FG	.00		
	.00	2	48	4.0
	1.00	2	148	98.7
Overall Percentage				75.0

a) The cut value is .500

The above table summarizes the observed group and the predicted group classification. The overall correctly specified group percentage is 75.

Table: VIII Variables in the equations

		B	S.E	Wald	df	Sig.	Exp(B)
Step 1[a]	AGE	-.047	.017	8.136	1	.004	.954
	DH	.112	.053	4.541	1	.033	1.119
	TD	-.210	.417	.254	1	.614	.811
	Constant	3.108	.847	13.460	1	.000	22.374
Step 2[a]	AGE	-.045	.016	7.991	1	.005	.956
	DH	.107	.051	4.349	1	.037	1.113
	Constant	2.971	.799	13.833	1	.000	19.521

a) Variable (Entered on step one: AGE,DH,TD).

The above table explains the AGE is significance at p-value<0.01 and DH is significance at p-value<0.05

Table: XI Model if Term Removed

Variable		Model Log likelihood	Change in -2log likelihood	df	Significance of the change
Step 1	AGE	-111.632	8.566	1	.003
	DH	-109.999	5.300	1	.021
	TD	-107.475	.251	1	.616
Step 2	AGE	-111.665	8.380	1	.004
	DH	-110.000	5.050	1	.025

a) Variable (Entered on step one:AGE,DH,TD).

The above table explains the AGE is significance at p-value<0.01 and DH is significance at p-value<0.05, TD are not-significant.

Table: X Variable not in the equation

	Score	df	Sig.
Step 2[a] variables			
TD	.254	1	.614
Overall Statistics	.254	1	.614

a) Variable removed on the step2:TD

Conclusion

In this paper we examined the suitability Marginal regression analysis of a bi- variate binary response values. A random sample of diabetes patients collected from King George Hospital, Visakhapatnam, were interviewed and the information on characteristics such AGE, fasting Blood glucose level, Income, Disease history(DH), whether the patient has Blood Pressure(BP), family history of the diseases, types of diseases, type of medicine were recorded for each patient. The linear regression, binary regression model and marginal regression were discussed their properties, estimation of parameters and testing aspects for the response variables such as fasting blood glucose levels (F.G) and the dichotomous response B.P are significant variables.

References

[1] Fitzmaurice,G.M.& Laird (1993).A likelihood- based method for analyzing longitudinal binary responses. Statist.Sci. 8, 284-309.

[2] Guangyong , Zon and Allan Donner(2004).In cluster randomized trials and observational studies the involve of a binary outcome,and introduce a BC a confidence interval for the ICC-2004;13:251-271.

[3] Nicole H.Augstin, Stefan Lang,Monica Musio and Klaus Von wilper-(2007)- A spatial model for the needle losses of pine-trees in the forests of Baden-ürttemberg: Appl.Statist.56,part-1,pp,29-50 an applicationof Bayesian structured additive regression

[5] Ekholm, A & Green,M(1994). Fitting non linear models in GLIM4 using numerical derivaties. GLIM Newslett. 23, 12-20.

[6] Jerome A. Dupuis -Bayesian Estimation of Species Richness from Quadrat Sampling. Data in the Presence of Prior Information Biometrics 62, 706–712- September 2006, DOI: 10.1111/j.1541-0420.2006.00524.x

[7] Francis, B, Green, M&Payne, C.(Ed)(1993). TheRelease4Manual. Oxford: Clarendon. Categorical responses.J.Am.Statist.Assoc. 89. 625-32.

[8] Fitzmaurice. Laird and Rotnitzky (1993).A likelihood- based method for analyzin longitudinal binary responses. Biometrica 80,141-15.

[9] Flesis, & Cuzick,1979, Applied Physiological Measurement 3, 537-542.

[10] 2007-Statistics under one umbrella is first Joint Statistical Submitting author.

[11] Jerome A. Dupuis (2006) consider the problem of estimating the number of species of an animal community. Eicholm Frances(1991)

[12] bloch AND Kramer, 1989, Biometrics 45, 269-287.

[13] Altaye, Donner and Klar 2001, Biometrics 57, 584-588

[14] Allcroft.D.J and Glasbey.C.A.(2003). A latent Gaussian markov random field model for spatiotemporal rainfall disaggregation. Appl.Statist. 52, 487-498.

[15] Agresti A.(1996), "An Introduction to Categorical Data Analysis" , wiley.

[16] David W. Hosmer and Stanley Lemeshow, "Applied Logistic Regression", Second Edition.

[17] Baggerly,K.A.(1998). Empirical likelihood as goodness-of-fit measure. Biomaterial 85, 535-47.

[18] Bianco,A.M. & Yohai, V.J.(1996). Robust estimation in the logistic regression model. In Robust Statistics, Data Analysis and Computer Intensive Methodes, Lecture Notes in Statistics, 109,Ed.H. Rieder, pp.17-34,Newyork, Springer-Verlag.

[19] Chib.S and Greenberg,E.(1998) Analysis of multivariate probit models. Biometrika 85, 347-361.

[20] Caroll R.J, Wang,S. & Wang,C.Y.(1995) Prospective analysis of logistic case-control studies. J.Am.Statist.Assoc. 90,157-69.

[21] Chen.M.H and Shao,Q.M(1999) Existence of Bayesian estimates for the polychotomous quantal response model. Annals of the institute of statistics and mathematics 51,637-656.

[22] Ging Qin and Biao Zhang (2003) Using Logistic Regression Procedures for Estimating receiver Operating Characteristic Curves; Biometrika,90,585-596

[23] Blotzheim, U.N. and Bauer, K.M.(1997) Pyrrhula, Gimpel,Dompfaff. In Handbuch der Vogel Mitteleuropas, Band 14/2, Teil 5, 1130-1181. Wiesbaden: Aula-verlag

[24] Kutner M.H; Nachtsheim C.J and Neter J. (2004),"Applied linear regression models (fourth edition)"

[25] Mitchell C. And Dayton (1992)," Logistic Regression Analysis",University of Maryland.World Health Organization (WHO). Obesity: preventing and managing the global epidemic. WHO Obesity Technical Report Series, No. 894. Geneva: World Health Organization, 2000.

[26] Jewell NP. (2004). Statistics for Epidemiology.New Yark, Chapman & Hall/CRC.