Journal of Applied Statistics

ISSN: 0266-4763 (Print) 1360-0532 (Online) Journal homepage: http://www.tandfonline.com/loi/cjas20

A comparison of ordinal logistic regression models using Classical and Bayesian approaches in an analysis of factors associated with diabetic retinopathy

K. Vaitheeswaran, M. Subbiah, R. Ramakrishnan & T. Kannan

To cite this article: K. Vaitheeswaran, M. Subbiah, R. Ramakrishnan & T. Kannan (2016): A comparison of ordinal logistic regression models using Classical and Bayesian approaches in an analysis of factors associated with diabetic retinopathy, Journal of Applied Statistics, DOI: 10.1080/02664763.2016.1140725

To link to this article: http://dx.doi.org/10.1080/02664763.2016.1140725

Published online: 03 Feb 2016.

JOURNAL OF APPLIED STATISTICS, 2016
http://dx.doi.org/10.1080/02664763.2016.1140725

Taylor & Francis
Taylor & Francis Group

A comparison of ordinal logistic regression models using Classical and Bayesian approaches in an analysis of factors associated with diabetic retinopathy

K. Vaitheeswaran[a,b], M. Subbiah[c], R. Ramakrishnan[d] and T. Kannan[e]

[a]Manonmaniam Sundaranar University, Tirunelveli, Tamil Nadu, India; [b]National Center for Disease Informatics and Research (ICMR), Bangalore, Karnataka, India; [c]Department of Mathematics, L.N Government College, Ponneri, Tamil Nadu, India; [d]National Institute of Epidemiology (ICMR), Ayapakkam, Chennai, Tamil Nadu, India; [e]National Institute for Research in Tuberculosis (ICMR), Chetpet, Chennai, Tamil Nadu, India

ABSTRACT

Estimating the risk factors of a disease such as diabetic retinopathy (DR) is one of the important research problems among bio-medical and statistical practitioners as well as epidemiologists. Incidentally many studies have focused in building models with binary outcomes, that may not exploit the available information. This article has investigated the importance of retaining the ordinal nature of the response variable (e.g. severity level of a disease) while determining the risk factors associated with DR. A generalized linear model approach with appropriate link functions has been studied using both Classical and Bayesian frameworks. From the result of this study, it can be observed that the ordinal logistic regression with probit link function could be more appropriate approach in determining the risk factors of DR. The study has emphasized the ways to handle the ordinal nature of the response variable with better model fit compared to other link functions.

ARTICLE HISTORY
Received 26 April 2014
Accepted 7 January 2016

KEYWORDS
AIC; diabetic retinopathy; DIC; link functions; ordinal logistic regression

1. Introduction

Diabetic retinopathy (DR) disease is the major component of Type II Diabetes Mellitus that is being studied by biomedical, applied statistics, and epidemiology researchers in identifying the risk factors associated with DR. Many such studies include ordinal response variable in developing regression models [1–7,9,11,12,14–16,18,19,22–26,28] or converting interval measurements into categorical scales with two or more categories in estimating the effect of covariates involved in the study. Further, it has been observed that scale of measurement of the response variable plays a major role in the choice or building a model [11] and hence, preserving the ordinal nature of the response variable is advantageous than methods for nominal data or binary models [4,22,28].

More precisely, pooling of some groups of the response variable in order to have binary form might tend to have loss of information that reduces statistical power in the results. Alternatively, methods such as ordinal logistic regression can be chosen to investigate the

CONTACT K. Vaitheeswaran ✉ vaitheeswaran81@gmail.com

effect and significance of predictor variables on all levels of ordered response variable. In such attempt like any other generalized linear model (GLM), there is a need to study the impact of plausible link functions together with estimation of parameters and model fit diagnostics.

The present work has identified a scope to have a statistical study based on GLM approach for handling ordinal response variable with more than two categories. In particular, this study has focused on determining the risk factors associated with DR using a data set [30] from Sankara Nethralaya-Diabetic Retinopathy Epidemiology and Molecular Genetic Study (SN-DREAMS I) that has an objective to assess the severity of DR which is recorded on a 5-level ordinal scale.

The paper has been organized as follows, Section 2 deals with the details of underlying regression model and a description about the data; Statistical analysis and result have been presented in Section 3 and Section 4 provides discussion and recommendations.

2. Materials and methods

GLM extends the linear modelling framework so as to include response variables which are not normally distributed in general; that is the variables that may be proportion, count, binary, multinomial, and ordered or unordered.

A GLM involves:

(1) Random component, a vector of observed data $y = (y_1, y_2, \dots y_n)$.
(2) Systematic component, p-predictors in a matrix form X and coefficients β to form a linear predictor $X\beta$.
(3) A link function $g(\mu)$ of mean μ of the response variable that links the random and systematic components.

The present work has considered modelling categorical response variables with K ($K > 2$) levels. Let $\pi_1(X_i), \dots, \pi_K(X_i)$ denote response probabilities at values for a set of explanatory variables. Then assuming one of the K levels as a reference category the cumulative probabilities of the remaining $K-1$ categories are

$$\theta_k(X_i) = p(Y \leq k/X_i) = \pi_1(X_i) + \cdots + \pi_k(X_i), \quad k = 1, 2, \dots, K-1.$$

Then the proportional odds model (POM) can be expressed as

$$\gamma_k(X_i) = \gamma_k \exp(-\beta^{\mathrm{T}} X) \quad k = 1, 2, \dots, K-1,$$

where $\gamma_k = (\theta_k(X))/(1 - \theta_k(X))$, the odds for the event $Y \leq k$ and β is a vector of parameters. Also the assumption for POM is that ratio of the odds $\dfrac{F_k(X_i)}{F_k(X_j)} = \exp(-\beta^{\mathrm{T}}(X_j - X_i))$ is constant across response categories.

The problem of interest is to estimate the parameters β and summarize the odds ratio for predictor obtained from separate binary logistic regressions [20]. As an illustration, if an ordinal response has six levels then five logits will be modelled with two groups having level 1 in one group and levels 2–6 in another group; subsequently at each step one level from the second group is shifted to first group [13]. Further, this study has limited to three most widely applied link function that are presented in the following table though there is no consensus in choosing a specific link function that fits well for a given scenario. However, few studies [10] have indicated the merits of particular link function when higher categories are more probable.

Function	Form
Logit	$\log\left(\frac{\theta_k(X_i)}{1-\theta_k(X_i)}\right)$
Probit	$\Phi^{-1}[\theta_k(X_i)]$
Complementary log–log	$\log[-\log(1-\theta_k(X_i))]$

In this study POM has been applied for the data set from SN-DREAMS I [30] which has aimed to identify risk factors associated with DR. The data set has a total of 5999 individuals from Chennai, India; of these, 1414 subjects with diabetes are included for data analysis in the study [27] that forms the basis for the present statistical investigation. The response variable DR is ordinal in nature with five levels (0 = No DR, 1 = Mild non-proliferative DR, 2 = Moderate non-proliferative DR, 3 = Severe non-proliferative DR and 4 = Proliferative DR) and 21 predictors are included in the model that have following characteristics

Variables	Nature	Details
Gender	Categorical	0 = female, 1 = male
Age	Continuous	
Duration of Diabetes Mellitus	Continuous	
History of diabetic status	Categorical	0 = newly detected diabetes, 1 = known diabetes
Family of history of diabetic status	Categorical	0 = no, 1 = yes
Physical activity status	Categorical	0 = sedentary work, 1 = moderate worker, 2 = heavy worker
Socio-economic status	Categorical	0 = low income group, 1 = middle income group, 2 = high income group
Waist circumference	Continuous	
Hip circumference	Continuous	
Body mass index	Continuous	
User of insulin	Categorical	0 = non-user, 1 = user
Presence of neuropathy	Categorical	0 = absent, 1 = present
Presence of hypertension	Categorical	0 = absent, 1 = present
Presence of nephropathy	Categorical	0 = absent, 1 = present
Smoker	Categorical	0 = non-smoker, 1 = smoker
Alcohol user	Categorical	0 = non-user, 1 = user
Glycosylated haemoglobin	Continuous	
Serum total cholesterol	Continuous	
Serum HDL cholesterol	Continuous	
Serum triglycerides	Continuous	
Haemoglobin	Continuous	

This data set provides a scope to identify the significant predictors for DR based on POM in Classical and Bayesian statistical paradigms. Statistical significance has been set at 5% in Classical methods so that p-value less than 0.05 is considered to have statistical significance. Assessment of models is based on Akaike information criterion (AIC) and deviance information criterion (DIC) for Classical and Bayesian models, respectively.

3. Result

Table 1 provides descriptive statistics for each of the 21 predictors used in this work; categorical variables are presented as number (percentage) of respondents in a specific category of that variable and continuous variables are given as mean ± standard deviation.

The estimated value of the predictors in ordinal regression model using Classical procedure with three link functions are presented in Table 2; Bayesian estimates are found to be similar in numerical form and hence the estimates are not made available in Table 2, nevertheless this study has considered for comparison and related conclusions. It can be observed that eight predictors have $p < 0.05$ when logit and clog–log link function are used; whereas, one more predictor (Presence of neuropathy) is found to be significant under probit link function. Bold face values in Table 2 indicate statistically significant predictors. Also, in terms of direction of the estimates, all significant predictors except age and haemoglobin are positive.

Further, Table 3 provides the values of AIC and DIC; smallest value indicates the better fit model obtained through ordinal regression method using three link functions. In the case of AIC, probit provides the least value (1669.97) when compared to logit (1678.95) and clog–log (1697.85) link functions. Similarly, the smallest value of DIC through Bayesian ordinal regression model can be found under probit (1668.48) when compared to logit (1677.17) and clog–log (1696.31) link functions.

Table 1. Descriptive statistics of predictors that have been used in the present study.

	DR severity level				
Predictors	Level 0 (1159)	Level 1 (127)	Level 2 (88)	Level 3 (18)	Level 4 (22)
Gender (male)	592 (51.1)	80 (63.0)	52 (59.1)	13 (72.2)	13 (59.1)
Age (in years)	56.1 ± 10.2	56.8 ± 9.6	57.0 ± 8.1	57.7 ± 7.8	60.6 ± 9.2
Duration of diabetes mellitus (years)	4.7 ± 5.8	8.0 ± 6.5	10.4 ± 6.7	10.3 ± 6.9	12.2 ± 5.8
History of diabetic status (known diabetes)	926 (79.9)	115 (90.6)	86 (97.7)	17 (94.4)	22 (100.0)
Family history of diabetic status	679 (58.6)	83 (65.4)	57 (64.8)	11 (61.1)	15 (68.2)
Physical activity status					
Sedentary	581 (50.1)	62 (48.8)	47 (53.4)	13 (72.2)	16 (72.7)
Moderately active	543 (46.9)	60 (47.2)	38 (43.2)	5 (27.8)	6 (27.3)
Heavy worker	35 (3.0)	5 (3.9)	3 (3.4)	0 (0.0)	0 (0.0)
Socio-economic status					
Lower	136 (11.7)	15 (11.8)	5 (5.7)	1 (5.6)	1 (4.5)
Middle	806 (69.5)	92 (72.4)	67 (76.1)	14 (77.8)	16 (72.7)
Upper	217 (18.7)	20 (15.7)	16 (18.2)	3 (16.7)	5 (22.5)
Waist circumference (cm)	91.5 ± 9.7	90.2 ± 10.2	90.9 ± 10.6	89.8 ± 11.3	90.5 ± 8.8
Hip circumference (cm)	101.2 ± 10.6	98.5 ± 10.6	99.5 ± 10.6	96.5 ± 8.2	97.7 ± 9.7
Body mass index (kg/m^2)	25.6 ± 4.0	24.3 ± 4.2	24.7 ± 4.2	23.2 ± 3.6	23.8 ± 3.8
User of insulin	32 (2.8)	12 (9.4)	11 (12.5)	3 (16.7)	10 (45.5)
Presence of neuropathy	198 (17.1)	20 (15.7)	26 (29.5)	9 (50.0)	11 (50.0)
Presence of hypertension	733 (63.2)	89 (70.1)	51 (58.0)	13 (72.2)	16 (72.7)
Presence of nephropathy	171 (14.8)	33 (26.0)	34 (38.6)	12 (66.7)	14 (63.6)
Smoker	224 (19.3)	29 (22.8)	14 (15.9)	6 (33.3)	4 (18.2)
Alcohol user	244 (21.1)	33 (26.0)	23 (26.1)	7 (38.9)	3 (13.6)
Glycosylated haemoglobin (g%)	7.9 ± 2.1	8.9 ± 2.2	9.7 ± 2.5	9.2 ± 2.1	9.9 ± 2.4
Serum total cholesterol (mg/dl)	186.4 ± 39.5	181.6 ± 43.0	192.9 ± 49.2	205.7 ± 43.9	178.5 ± 46.4
Serum HDL cholesterol (mg/dl)	39.0 ± 9.9	39.6 ± 10.5	40.1 ± 10.7	40.4 ± 8.8	43.0 ± 16.6
Serum triglycerides(mg/dl)	154.9 ± 103.0	144.5 ± 87.9	155.7 ± 84.3	176.6 ± 92.8	119.8 ± 75.4
Haemoglobin (g/dl)	13.8 ± 1.5	13.9 ± 1.8	13.6 ± 1.6	13.7 ± 1.9	12.8 ± 2.0

Note: HDL, high density lipoprotein.

Table 2. Estimates of the risk factors for DR using classical ordinal regression model with different link functions.

	Ordinal regression model with								
	Probit link function			logit link function			Complementary log–log link function		
Predictors	Estimate	SE	*p*	Estimate	SE	*p*	Estimate	SE	*P*
Gender	0.310	0.143	**0.030**	0.592	0.261	**0.024**	0.457	0.225	**0.042**
Age (in years)	−0.012	0.005	**0.021**	−0.027	0.009	**0.005**	−0.025	0.008	**0.003**
Duration of diabetes mellitus (years)	0.044	0.007	< **0.0001**	0.079	0.012	< **0.0001**	0.059	0.009	< **0.0001**
History of diabetic status	0.341	0.149	**0.022**	0.719	0.303	**0.018**	0.748	0.279	**0.007**
Family history of diabetic status	0.095	0.088	0.276	0.193	0.163	0.237	0.122	0.138	0.378
Physical activity status	0.030	0.080	0.706	0.073	0.147	0.621	0.046	0.123	0.712
Socio-economic status	−0.026	0.080	0.744	−0.061	0.148	0.682	−0.075	0.125	0.546
Waist circumference (cm)	0.002	0.008	0.790	0.003	0.014	0.812	0.007	0.112	0.565
Hip circumference (cm)	0.005	0.008	0.559	0.009	0.015	0.508	0.007	0.013	0.569
Body mass index (kg/m^2)	−0.029	0.021	0.162	−0.058	0.039	0.133	−0.063	0.033	0.057
User of insulin	0.682	0.152	< **0.0001**	1.207	0.264	< **0.0001**	0.842	0.198	< **0.0001**
Presence of neuropathy	0.206	0.104	**0.047**	0.334	0.191	0.080	0.229	0.159	0.150
Presence of hypertension	0.049	0.089	0.578	0.095	0.165	0.566	0.121	0.140	0.388
Presence of nephropathy	0.146	0.095	< **0.0001**	1.082	0.169	< **0.0001**	0.822	0.139	< **0.0001**
Smoker	−0.080	0.122	0.510	−0.128	0.222	0.563	−0.086	0.186	0.643
Alcohol user	0.146	0.123	0.234	0.229	0.223	0.304	0.181	0.188	0.337
Glycosylated haemoglobin (g %)	0.118	0.018	< **0.0001**	0.208	0.033	< **0.0001**	0.160	0.025	< **0.0001**
Serum total cholesterol (mg/dl)	0.001	0.001	0.309	0.002	0.002	0.275	0.002	0.002	0.290
Serum HDL cholesterol (mg/dl)	0.004	0.004	0.396	0.006	0.008	0.405	0.008	0.007	0.225
Serum triglycerides(mg/dl)	−0.001	−0.001	0.560	−0.001	0.001	0.562	−0.001	0.001	0.421
Haemoglobin (g/dl)	−0.780	0.030	**0.009**	−0.140	0.055	**0.011**	−0.118	0.047	**0.013**

Note: HDL, high density lipoprotein.

Table 3. Model fit indices for Classical and Bayesian ordinal regression models with different link functions.

	Ordinal regression model		Bayesian ordinal regression model
Ordinal model with	Log likelihood	AIC	DIC
Probit	**−809.98**	**1669.97**	**1668.48**
Logit	−814.47	1678.95	1677.17
Clog–log	−823.92	1697.85	1696.31

Note: Bold values indicate better fit.

4. Discussion

Statistical investigation of real time problem always warrant careful assessment of assumptions and requirements of underlying models. GLM that accommodates different kind of response variables, also possesses such thoughtful applications [4,8,28]. The present study has focused on the different aspects of regression modelling in GLM framework to determine the significant risk factors associated with DR. In particular, emphasize has been given to the scale of measurement of response variables in terms of its ordinal and polytomous nature while fitting the model. Another aspect of the study is the comparison of three link functions which are essential components of a GLM. These tasks are carried out using Classical and Bayesian inferential procedures exploiting the availability of computing facilities [9,17,29] in such model fit studies.

The present study has highlighted the negligible differences in the estimates of parameters between the Classical and Bayesian ordinal regression models with three different link functions. Further, appreciable similarity in estimates for small probabilities among the three link function is quite visible and such behaviour is supplementing earlier studies [23,21] and present study has noticed such observations in the case of model fit statistics also, yet decisions are usually made based on such diagnostic tools

Present analysis of DR data has emphasized the choice of the model with link functions and the nature of response variable that is usually pertinent to research problems. Rich statistical literature and available computing power provide ample scope to utilize most appropriate techniques that are quite relevant to specific practical applications. In conclusion, the ordinal regression model (both Classical and Bayesian) with probit link function has been found to be more appropriate in determination of significant factors associated with DR.

Acknowledgements

Authors would like to thank the Editor and referees for the valuable comments and suggestions which helped to improve the article.

Disclosure statement

No potential conflict of interest was reported by the authors.

References

[1] A. Agresti, *Tutorial on modeling ordered categorical response data*, Psychol. Bull. 105 (1989), pp. 290–301.

[2] A. Agresti, An introduction to Categorical Data Analysis, 2nd ed., Wiley-Interscience, New York, NY, 2007.

[3] C.V. Ananth and D.G. Kleinbaum, *Regression models for ordinal data: A review of methods and applications*, Int. J. Epidemiol. 26 (1997), pp. 1323–1333.

[4] J.A. Anderson, *Regression and ordered categorical variables (with discussion)*, J. R. Stat. Soc. B 46 (1984), pp. 1–30.

[5] B. Armstrong and M. Sloan, *Ordinal regression models for epidemiologic data*, Am. J. Epidemiol. 129 (1989), pp. 191–204.

[6] D. Ashby, S.J. Pocock, and A.G. Shaper, *Ordered polytomous regression: An example relating serum biochemistry and haematology to alcohol consumption*, Appl. Stat. 35 (1986), pp. 289–301.

[7] D. Ashby, C.R. West, and D. Ames, *The ordered logistic regression model in psychiatry: Rinsing prevalence of dementia in old people's homes*, Stat. Med. 8 (1989), pp. 1317–1326.

[8] R. Bender and A. Benner, *Calculating ordinal regression models in SAS and S-plus*, Biom. J. 42 (2000), pp. 677–699.

[9] R. Bender and U. Grouven, *Using binary logistic regression models for ordinal data with non-proportional odds*, J. Clin. Epidemiol. 51 (1998), pp. 809–816.

[10] J. Cheng, Z. Wang, and G. Pollastri, *A neural network approach to ordinal regression*, J. Comp-eleceng. 1 (2007), pp. 1028–1036.

[11] C. Cox, *Location scale cumulative odds models for ordinal data: A generalized non-linear model approach*, Stat. Med. 14 (1995), pp. 1191–1203.

[12] C. Cox, *Multinomial regression models based on continuation ratios*, Stat. Med. 7 (1997), pp. 435–441.

[13] M.J. Gameroff, Using the Proportional Odds Model for Health-related Outcomes: Why, When, and How with Various SAS Procedures, New York State Psychiatric Institute, New York, NY, 2005, pp. 205–230.

[14] S. Greenland, *An application of logistic models to the analysis of ordinal responses*, Biom. J. 27 (1985), pp. 189–197.

[15] S. Greenland, *Alternative models for ordinal logistic regression*, Stat. Med. 13 (1994), pp. 1665–1677.

[16] C. Greenwood and V. Farewell, *A comparison of regression models for ordinal data in an analysis of transplant-kidney function*, Canad. J. Statist. 16 (1988), pp. 325–335.

[17] F.E. Harell Jr, *Designs and functions for bio-statistical/epidemiologic modeling, testing, estimation, validation, graphs, and prediction. Functions available on the Web in the StatLib repositary of tatistical software*. Available at http://www.lib.stat.edu/S/Harrell/ [last cited on 1998a].

[18] T.J. Haste, J.L. Botha, and M. Schnitzler, *Regression with an ordered categorical response*, Stat. Med. 8 (1989), pp. 785–794.

[19] W. Holtbrugge and M.A. Schumacher, *Comparison of regression models for the analysis of ordered categorical data*, Appl. Stat. 40 (1991), pp. 249–259.

[20] D.W. Hosmer and S. Lemeshow, Applied Logistic Regression, John Wiley & Sons, NewYork, 1989.

[21] E. Laara and J.N. Mathews, *The equivalence of two models for ordinal data*, Biometrika 72 (1985), pp. 206–207.

[22] J. Lee, *Cumulative logit modeling for ordinal response variables: Applications of biomedical research*, Comput. Appl. Biosci. 8 (1992), pp. 555–562.

[23] P. McCullagh and J.A. Nelder, Generalized Linear Models, Chapman and Hall, London, 1989.

[24] S. Menard, *Applied Logistic Regression Analysis*, 2nd ed. Sage Publications, Thousand Oaks, CA. Series: Quantitative Applications in the Social Sciences, Vol. 106, 2002, pp. 1–128.

[25] A.A. O'Connell, *Logistic Regression Models for Ordinal Response Variables*, Sage Publications, Thousand Oaks, CA. Quantittive Applications in the Social Sciences, Vol. 146, 2005, pp. 1–140.

[26] B. Peterson, E. Frank, and J.R. Harrell, *Partial proportional odds model for ordinal response variables*, Appl. Stat. 39 (1990), pp. 205–217.

[27] R. Rajiv, K.R. Padmaja, and R.R. Sudhir, *Prevalence of diabetic retinopathy in India(SN-DREAMS Report-2)*, AAOP 116 (2009), pp. 311–318.

[28] S.C. Scott, M.S. Goldberg, and N.E. Mayo, *Statistical assessment of ordinal outcome in comparative studies*, J. Clin. Epidemiol. 50 (1997), pp. 45–55.

[29] N. Shadi, *The analysis of Bayesian probit regression of binary and polychotomous response data*, IJE Trans. B Appl. 20 (2007), pp. 237–248.

[30] A. Swati, R. Rajiv, and G.P. Pradeep, *Sankara Nethralaya-diabetic retinopathy epidemiology and molecular genetic study (SN-DREAMS I): Study design and research methodology*, Ophthalmic Epidemiol. 12 (2004), pp. 143–153.