Assessment of Different Link Functions for Modeling Binary Data to Derive Sound Inferences and Predictions

Falk Huettmann, Julia Linke

Geography Department, University of Calgary, Calgary AB T2N 1N4, Canada,
falk@ucalgary.ca

Abstract. Binary data are widely used for spatial modeling and when inferences and predictions are to be derived. If a Generalized Linear Model (GLM) is applied, logit functions are often used. Here we show alternatives to the traditional logit approach using probit and the complementary log log link functions. We present a software-based approach and two methods of assessing which link function performs best for inferences and for predictions. The first decision criterion is centered around the model deviance, e.g. relevant for inferences. The second criterion is based on predicting the findings back to the training data and then using the differences between expected and predicted values for known presences and absences as an indication of the fit. As an example we use Marbled Murrelet (*Brachyramphus marmoratus*) nesting habitat data derived from aerial telemetry and overlaid with GIS habitat layers (DEM and Forest Cover). This data set is large and carries inherent noise due to field data and a complex landscape; therefore it well covers the extremes of the fitted link functions. It is a representative example for a situation where the selection of a link function could affect the results. Findings indicate that for our data all three link functions behave similar, but logit link functions perform better than the cloclog and probit link functions when inferences as well as predictions are the study goals.

1 Introduction

Multiple regression techniques are widely used to assess wildlife-habitat relationships. Frequently, binary data are modeled with a variety of predictors, e.g. in a context of a Generalized Linear Model [1]. Such approaches mostly use the logit link function. Alternative link functions exist but are rarely considered and assessed. This situation is likely due to the general belief that the logit link function addresses the needs for most applications and thus is widely recommended [2]. The alternatives are seldom applied and tested although they could affect inferences and predictions in applications with larger data sets and which are distributed along the full gradient of the fitted link functions, e.g. at the extremes. The link functions are known to perform similar at the center, but differ at the lower and upper extremes. Besides the traditional logit link function, here we investigate the use of two alternatives, probit and cloglog, which has never been done for applications dealing with a large dat set of radio-telemetry wildlife data. Using field research data we present approaches for selecting the most appropriate link function when the emphasize is on inference

V. Kumar et al. (Eds.): ICCSA 2003, LNCS 2669, pp. 43-48, 2003.

and/or prediction. We use nesting and habitat data from the Marbled Murrelet (*Brachyramphus marmoratus*), a tree-nesting seabird species of international conservation concern, to investigate the effects that the correct link function has for large-scale wildlife conservation and landscape management decisions.

1.1. Background

A GLM with a logit link function assumes that data are truly binomial [3]. The fit of such functions can be assessed with a Hosmer-Lemeshow test [4,5]. However, such tests can be potentially biased and their validity is debated [5,6]; the true distribution remains unknown. A wide choice of alternative link functions exist, and besides logit here we use two other functions that are commonly suggested [1, 2, 5] and which can relatively easily be applied. The probit link function is based on the inverse normal function [7]. In contrast to the two previous link functions, the complementary log-log (cloglog) link function is not symmetric around p=0.5 [3]; this function is the inverse of the cumulative extreme-value function (also called Gompertz distribution). A forth link function is the log log function, but it is seldom used because its behaviour is inappropriate for most applications [1]; thus, it was excluded from this analysis. For more details on link function characteristics and relations see [1,3,7,8]. Our approach is based on existing SPLUS code [9], [2, 10], which we modified accordingly to address the outlined research questions.

2 Methods

Data. We used the existing data set for nesting locations of Marbled Murrelets in Desolation Sound, British Columbia [9]. Nesting locations were derived from aerial telemetry, and then overlaid in a GIS with a Digital Elevation Model (DEM) and with Forest Cover polygons [10]. Besides nest locations, this allows to derive random locations within available and comparable habitat (old-growth forest polygons) and to obtain a Resource Selection Function (RSF) [11].

Computations. We developed an SPLUS script, which allows, in an automated fashion, to compare 51 nest locations in old-growth forest with a set of available locations, randomly drawn from an overall pool of 5000 random locations (compare also with [9]). To these data, we applied a binomial GLM with slope and elevation as predictors, and used each time a different set of link functions, logit, probit and cloglog. Results for each GLM were stored externally and ready for external analysis as an ASCII file.

Assessing the link functions. The mean coefficients derived from the 1000 comparisons were used to derive a regression formula for each individual link function. The standard deviations (SD) for the coefficients were also computed. We decided to use 1000 comparisons since it is suggested to use an ‘infinite’ number of comparisons, and the findings don’t show a change beyond 1000 comparisons [11]. If the focus is on inference, we used the mean and standard deviation of the model

deviance for the 1000 models as a criterion for the fit (first criterion). Using the derived 'mean' regression formula for each of the three link function results, data were predicted back to themselves (training data) using formulas (1), (2) and (3). For the second criterion with an emphasize on prediction, we present the descriptive statistics of the predictions for the known absences and presences. Alternatively, the predictions could also be presented spatially in order to assess visually the impacts that the use of the individual link functions made on the predicted spatial distribution of nests.

$$\exp(\alpha+\beta1*\text{slope}+\beta2*\text{elevation})/1+\exp(\alpha+\beta1*\text{slope}+\beta2*\text{elevation}) = P \quad \textbf{(1 logit)}$$

$$1/(\text{sqrt}(2*1\text{^}2*3.141569))*\exp(-(\alpha+\beta1*\text{slope}+\beta2*\text{elevation})\text{^}2/2*1\text{^}2) = P \quad \textbf{(2 probit)}$$

$$1-\exp(-\exp(\alpha+\beta1*\text{slope}+\beta2*\text{elevation})) = P \quad \textbf{(3 cloglog)}$$

3 Results

The three mean regression formulas in table 1 show that the link functions affect the coefficients and the deviances of the fitted models. Using the first suggested criterion (emphasize on inference), the highest standard deviation for the model deviance is found for the link function clogloc and thus performs the least reliable; the model deviances for logit and probit behave identical. The intercept and the slope coefficients vary for each of the link functions, but they remain within their positive and negative ranges. The elevation coefficient is identical for the logit and probit link function, but differs for the clolog link function. Generally, the slope coefficients have a slightly higher standard deviation than the coefficients for elevation.

Table 1: Mean regression parameters for three link functions.

	Mean Intercept (SD)	**Mean Slope Coefficient (SD)**	**Mean Elevation Coefficient (SD)**	**Mean Model Deviance (SD)**
Logit	-0.56346 (0.31362)	0.16687 (0.06877)	-0.08149 (0.03118)	189.9673 (2.25170)
Probit	-0.34014 (0.19472)	0.09836 (0.04143)	-0.08149 (0.03118)	189.9673 (2.25170)
Cloglog	-0.84116 (0.23765)	0.14861 (0.05758)	-0.06659 (0.02417)	189.5573 (5.05914)

For the second suggested criterion (emphasize on prediction), depending on the link function used the predictions deviate from the known and expected presence and absence of the training data (table 2). The findings show a consistent trend and behavior among link functions in that presence predictions are higher than absence predictions. Logit and cloglog prediction values are on the same magnitude, but the probit link function generally has lower predicted values and does not indicate a truly statistical difference between predicted absence and presence, as the other two link functions indicate. The probit function has also inconsistent SDs (lowest for predicted absences and highest for predicted presences); it appears to perform poorly and as an 'outlier'. It should be kept in mind that these comparisons are based on different sample sizes for predicted absences (n=5000) and presences (n=51). Generally, the differences between predicted absences and predicted presences are relatively small and do not even reach one; this is likely due to the inherent noise within field- and GIS data for a very heterogeneous landscape. The predictions are biased towards lower prediction values. Using the model deviance as a decision criterion, the findings indicate that for the Marbled Murrelet nesting habitat data example the logit and probit link functions would fit the data best. Using the predictive performance as a criterion, the logit and cloglog function behaves best. Overall, it can be concluded that for the used data set, the logit link function performs well overall, but less so the cloglog and probit functions.

Table 2: Predictions for each of the link functions on training data (known presence and known absence)

	Logit		**Probit**		**Cloglog**	
Predictions	Known Absence n=5000	Known Presence n=51	Known Absence n=5000	Known Presence n=51	Known Absence n=5000	Known Presence n=51
Min	0.14358	0.20127	0.07340	0.12815	0.14684	0.19873
Mean	0.32648	0.35437	0.25164	0.27289	0.32575	0.35385
Median	0.32683	0.34764	0.25190	0.27892	0.31839	0.34891
Max	0.56826	0.60865	0.39758	0.39807	0.60173	0.65635
SD	0.07239	0.08732	0.06316	0.69907	0.07459	0.09423
LCI	0.32447	0.32981	0.24988	0.25322	0.32368	0.32935
UCI	0.32848	0.37893	0.25339	0.29255	0.32781	0.38236

4 Discussion

We investigated the effect of applying different link functions to the fitting of binary data in a GLM. Using the correct link function is important for inference and predictions, e. g. for Resource Selection Functions [11]. We use empirical field-based telemetry-and GIS-data which present a real-life example and which are crucial to investigate towards sound conservation and management decisions, based on true data. We used two methods to assess the quality of the link function: standard

deviation of the model deviance, and predictive accuracy. We did not assess how stepwise approaches (adding or dropping) for the individual predictors to the model vary with the link function since we believe that this approach is not providing relevant evidence for the selection of the most appropriate link function. For the data used, it was found that the smallest standard deviation of the model deviance occurred with the logit and probit link functions, suggesting either of these two functions to be used to model these binary data. For the predictions, we found that the logit and clogclog link functions performed best. Overall, our findings are in agreement with the assumption that the traditionally used logit function performs well overall [12.]. However, depending on the specific data this can change. Although our findings are pointing towards consistent results for the two assessment methods, we emphasize the conflict between using 'tight' regression coefficients vs. minimizing SD of predictions to select the best link function for a specific study and data set. Which of these two criteria are to be used depends on the overall context and goals of the individual study. Therefore, it is crucial that the project goals are specified in advance, e.g. (i) inference, (ii) prediction or (iii) both, so that the best link function is selected for the specific research purpose. In the absence of truly accepted fitting diagnostics and tests, our findings and approaches demonstrate an example of selecting the appropriate link functions for solid and sound inferences and predictions when using binary GLMs. Towards sound inference and predictions, it is recommended that an assessment as presented here be done for any application which uses binary data in a multiple regression and for an RSF approach, e.g. as the case for many modern GIS applications based on large data sets with inherent 'noise' from field research in complex landscapes.

References

1. McCullagh, P. and J. A. Nelder. Generalized Linear Models. Monographs on Statistics and Applied Probability 37. Chapman and Hall. (1989)

2. Venables, B. and B.D. Ripley. Modern Applied Statistics with S. Fourth Edition. Springer Verlag, New York. (2002)

3. Collet, D. Modelling Binary Data. Chapman & Hall. New York. (1991)

4. Hosmer, D.W. and S. Lemeshow. Goodness-of-fit tests for the multiple logistic regression model. Communications in Statistics – Theory and Methods: (1980) 1043-1068.

5. Hosmer, D.W. and S. Lemeshow. Applied Logistic Regression. Wiley & Sons (1989).

6. Harrell, F. E. Jr. Regression Modeling Strategies. Springer Series in Statistics. Springer-Verlag. New York. (2002)

7. Daganzo, C. Multinomial Probit: The Theory and its Application to Demand Forecasting. Economic Theory, Econometrics, and Mathematical Economics. Academic Press. (1979)

8. Menard. S. Applied Logistic Regression Analysis. Sage Publications (2001)

9. Mathsoft Inc. SPLUS. Professional Release 2. Seattle (2000)

10. Huettman, F. and Linke J. An automated method to derive habitat preferences of wildlife in GIS and telemetry studies: A flexible software tool and examples of its application. European Journal for Wildlife Research. (in press)

11. Huettmann, F., E. Cam, R.W. Bradley, L. Lougheed, L.M. Tranquilla, C., Lougheed, P. Yen, Y. Zharikov and F. Cooke. Breeding habitat selectivity by Marbled Murrelets in a fragmented old-growth forest landscape. Wildlife Monograph. (in review)

12. Manly, B.F., L. L. McDonald, D. L. Thomas, T. L. McDonald and W. P. Erickson. Resource Selections by Animals. Kluwer Academic Publishers, Netherlands. (2002)