# Binomial Additive Modeling for Nonlinear Relationships

## Doğrusal Olmayan İlişkiler İçin Binomial Toplamsal Modelleme

Ahmet SEZER,[a]
Betül KAN,[a]
Berna YAZICI[a]

[a]Department of Statistics,
Anadolu University Science Faculty,
Eskişehir

Geliş Tarihi/*Received:* 27.05.2011
Kabul Tarihi/*Accepted:* 16.06.2011

Yazışma Adresi/*Correspondence:*
Betül KAN
Anadolu University Science Faculty,
Department of Statistics, Eskişehir,
TÜRKİYE/TURKEY
bkan@anadolu.edu.tr

**ABSTRACT Objective:** In most of the fields, researcher misapplies linear models, although there are more complex models needed. In this study generalized additive modeling of a real life data set is taken into account for binomial response case. **Material and Methods:** Generalized additive models (GAM) have become an elegant and practical option in model building. Those models represent an extension of generalized linear models (GLM) with a linear predictor involving a sum of smooth functions of covariates. GAMs allow for rather flexible specification of the dependence of the response on the covariates by specifying the model in terms of smoothing functions besides linear components. The advantage in that type of modeling is that the forms of the explanatory variables are not predetermined unlike in linear regression modeling but are constructed according to information derived from the data. This provides significant advantage for the nonparametric modeling over the linear regression modeling. **Results:** The models that are constructed for different components are interpreted and compared using Un-Biased Risk Estimator criterion. The model which best explains the structure of the dataset using splines in term of UBRE is given.

**Key Words:** Linear models; models, statistical; models, theoretical

**ÖZET Amaç:** Pek çok alanda araştırmacılar daha karmaşık regresyon modelleri oluşturmaları gereken durumlarda, yanlışlıkla doğrusal modelleme yapmaktadırlar. Bu çalışmada binomial yanıt değişkeni olduğu durumda gerçek bir veri seti için genelleştirilmiş toplamsal modelleme ele alınmıştır. **Gereç ve Yöntemler:** Genelleştirilmiş toplamsal modeller model kurmada iyi ve pratik bir seçenektir. Bu modeller, bağımsız değişkenlerin düzgün fonksiyonlarının toplamını içeren bir doğrusal tahminciyle oluşturulmuş genelleştirilmiş doğrusal modellerin genişletilmiş halidir. Genelleştirilmiş toplamsal modeller doğrusal bileşenlerin yanısıra düzeltilmiş fonksiyonların oluşturduğu terimleri de içererek, bağımsız değişkenler ile bağımlı değişkenlerin belirlenmesinde daha esnek bir yapıya sahiptirler. Bu modellemenin bir avantajı açıklayıcı değişkenlerin şekli doğrusal regresyonda olduğu gibi önceden belirlenmek yerine veriden alınan bilgiye göre oluşturulur. Bu durum nonparametrik modellemenin doğrusal regresyonla modellemeye göre belirgin bir avantajıdır. Sonuçlar: Farklı bileşenler için kurulan modeller yorumlanmış ve yansız risk tahmincisi kriterine göre karşılaştırılmılştır. Veri setinin yapısını UBRE skorlarına gore en iyi açıklayan model verilmiştir.

**Anahtar Kelimeler:** Doğrusal modeller; modeller, istatistiksel; modeller, teorik

**Turkiye Klinikleri J Biostat 2011;3(2):84-8**

Generalized additive models play a special role for several reasons. First, they are natural extension of the generalized linear models. Second they can be used for time dependent observations. And most importantly semiparametric extension is available.The choice of the degre-

e of smoothness is integrated with variable selection in a computationally efficient manner using criteria such as GCV or UBRE.

Additive regression models were introduced in the early eighties, and extended by the work of Buja et al. and Hastie & Tibshirani.[1,2] Stone and Newey studied properties of estimators.[3,4] Opsomer and Ruppert and Opsomer studied asymptotic properties of the backfitting estimator in detail.[5,6] Model selection methods for GAMs are still under study. However, Generalized Cross Validation (GCV) and Un-Biased Risk Estimator (UBRE) are commonly used in recent studies.[7,8]

In this study, occurrence of fire will be considered as a random phenomenon and a generalized additive model with several explanatory variables will be fitted. It is clear that fires depend on local conditions such as: elevation, wind speed, precipitation, air humidity, topography and other variables. Although the meteorological variables are important there are some other variables might be important, too. Our aim is to obtain a model which best explains the structure of the dataset using splines in terms of UBRE.

The forest fire happens because of two main reasons, the human error or the nature process. This study will concentrate on meteorological variables other than the human error. Indeed, most of the studies found that weather during the fire season is the most significant factor and rainfall has a modest negative relationship with the burnt area. Indeed, forest fire is one of the most dangerous hazard for the human being. Because of that, it deserves considerable attention to prevent this disaster on time. Knowing that what cause the significant fires will play essential role for the necessary prevention plans.

Since the fires are one of the major environmental concerns in all over the world, the risk analysis of probability of occurrence a forest fires have almost been interesting field for researchers. Preisler, Haase and Sackett developed a stochastic model for temperature profiles recorded at four depths beneath the soil during a large prescribed burn study and they assessed the temporal fit of the data to particular solutions of the heat equation using random effects model.[9] Ballart and Riba examined the relation between government measures, human participation, climate variables and forest fires.[10] Miller and Yool compared the sizes of simulated fire areas resulting from fuels maps.[11] Diaz-Delgado, Lloret and Pons estimated fire frequency in Catalonia for the last quarter of the 20th Century from historical burned area maps fitting a Weibull distribution to the observed proportion of fire intervals.[12]

Dayananda and Mandallaz and Ye employed Poisson distribution for the number of fires.[13] Martell mentioned that Symington showed that in the Parry Sound district of Ontario, the negative binomial distribution fit the historical fire occurrence data better than the Poisson distribution.[14] In another approach Poulin and Costello used the logistic distribution.[15] Brillinger, Preisler and Benoit studied probabilistic risk assessment using generalized mixed model.[16]

## GENERALIZED LINEAR MODELS

Let $y_1,..., y_n$ denote n independent observations on a response. In the general linear model we assume that Yi has a normal distribution with mean $\mu_i$ and $\sigma^2$.

GLM relaxes the assumption of normality by supposing the responses are drawn independently from a one parameter exponential family distribution, with density or probability function

$$p(y_i;\theta_i,\varphi) = \exp\left( \frac{y_i\theta_i - b(\theta_i)}{\varphi} + c(y_i,\varphi) \right) \quad (1)$$

Here, $\square_i$ is the natural parameter of the exponential family, $\varphi$ is the scale parameter. The form of the distribution in Eq. (1) is determined by the functions b and c. In GLM the second assumption on the mean is that it follows a linear model. A transformation of the mean, $g(\mu_i)$ is given by Eq. (2):

$$\eta_i = g(\mu_i) = \mathbf{x}_i'\beta \ . \quad (2)$$

and $\eta_i$ is called the linear predictor. The function $g(\mu_i)$ is called the link function. Link functions include of several examples: Identity, log, logit, pro-

bit and reciprocal. Since the link function is one to one, it can be invertible as

$$\mu_i = g^{-1}(\mathbf{x}_i^{'}\beta). \quad (3)$$

Therefore, the quantity $\eta_i$ is much simpler than the model for $\mu_i$. Here it should be noted that the transformation is not related to $y_i$ but its expected value of $\mu_i$.

Nelder and Wedderburn proposed Fisher scoring as a general evaluation of $\hat{\beta}$.[17] Each iteration of Fisher scoring method for numerical evaluation of the maximum likelihood estimation is weighted least squares regression. So that, the estimates of parameters $\hat{\beta}$, are obtained by iteratively re-weighted least squares in GLM.

## GENERALIZED ADDITIVE MODELS

A generalized additive model is a linear predictor involving a sum of smooth functions of covariates. The model is given in Eq. (4)

$$g(\mu_i) = \boldsymbol{X}_i^{*}\theta + f_1(x_{1i}) + f_2(x_{2i}) + f_3(x_{3i}, x_{4i}) + \ldots \quad (4)$$

Here, $\boldsymbol{X}_i^{*}$ is the $i$th row of the model matrix for any strictly parametric model components, $i$=1,…,n. is the corresponding parameter vector. $\mu_i = \mathrm{E}(Y_i)$ and $Y_i$ has some exponential family distribution. Additive models are attractive as they provide effective dimension and great flexibility in modeling.[2] For the choice of the degree of the smoothness in GAM, unbiased risk estimation (UBRE) is used.

GAM is fitted by backfitting algorithm. It can be represented using penalized regression splines and the appropriate degree of smoothness for the $f_j$ can be estimated by UBRE.[18] The fitting objective amounts to minimizing

$$\mathrm{SS}(\beta) = \|y - \boldsymbol{X}\beta\|^2 + \sum_j \lambda_j \beta' S_j \beta \quad (5)$$

with respect to $\beta$. Here, $S_j$ is called the penalty matrix and $\lambda$ is smoothing parameter. For numerical results, penalized iterative reweighted least squares method is prepared. UBRE is used for smoothing parameter selection and it is defined as in Eq. (6)

$$V_u(\lambda) = \|y - Ay\|^2 / n - \sigma^2 + 2\mathrm{tr}(A)\sigma^2 / n \quad (6)$$

A indicates the influence matrix. Hence, it seems reasonable to choose smoothing parameters which minimize the UBRE, when $\sigma^2$ is known.[19]

## APPLICATION

Variable selection is an important issue in both parametric and nonparametric in nature. However, nonparametric framework is more challenging than a parametric approach because of the lack of underlying assumptions that makes it difficult to define a general test approach for variable selection. The occurrence of fire is considered as a random phenomen and a generalized additive model with several explanatory variables is fitted. The response variable is defined as (burnt area) the binomial variable. For some specific value, we defined that burnt area is big enough to say that the fire is large and for other case it is insignificant.

There are several fire weather index systems constructed such as Canadian Forest Fire Index System, The French Index (RN), the Spanish Index (ICONA), etc to combine the affects of several variables.[20] Fire Weather Index (FWI) is the first part of the Canadian Forest Fire Danger Rating System introduced into New Zeeland in 1980. The FWI is based on weather readings taken at noon Standard time and rates fire danger at the mid afternoon peak from 2 to 4pm. Weather readings required are: air temperature (in the shade), relative humidity (in the shade), wind speed (at 10 meters above ground for an average over 10 minutes), rainfall (for the previous 24 hours). The FWI has six components: fine fuel moisture code (FFMC), duff moisture code (DMC), drought code (DC), initial spread index (ISI), build up index, fire weather index.

FFMC: this is a numerical rating of the moisture content of surface litter and other cured fine fuels. The FFMC rating is on a scale of 0 to 99. Any figure above 70 is high, and above 90 is extreme.

DMC is a numerical rating of the average moisture content of loosely compacted organic layers of moderate depth. The code indicates the depth that fire will burn in moderate duff layers and me-

dium size woody material. The DMC rating is dry if more than 30 or intensive burning will occur in the duff and medium fuels if above 40.

DC is a numerical rating of the moisture content of deep, compact, organic layers. It shows the likelihood of fire involving the deep duff layers and large lags. The DC rating of 200 is high, and more than 300 is extreme which indicates that fire will involve deep sub-surface and heavy fuels.

ISI indicates the rate fire will spread in its early stages. It is calculated from the FFMC rating and the wind factor. The rating starts from 0 to 10 which mean a high rate.

The information of the data used is listed below:

Temperature (Temp) - in Celcius degress (2.2 - 33.30)

Relative Humidity (RH) - in % (15.0 - 100)

Wind Speed (Wind) - Wind speed in km/h (0.40 - 9.40)

FFMC - FFMC index from the FWI system (18.7 - 96.20)

DMC - DMC index from the FWI system (1.1 - 291.3)

DC - DC index from the FWI system (7.9 - 860.6)

ISI - ISI index from the FWI system (0.0 - 56.10)

The data set is described in detail in http://archive.ics.uci.edu/ml/datasets/Forest+Fires (Table 1).[20]

## CONCLUSION

We obtained a model which best explains the structure of the dataset using splines in term of UBRE. Because of concurvity we expect to have just one of the index variables in the model. UBRE scores suggest that smooth function of FFMC and linear form of the temperature (model 15) found significant factors on the burnt area. Same UBRE score obtained when both FFMC and Temperature were included in the model as smooth function (model 16). Prediction power of the models may be tested to decide the form of the temperature for the best model.

We believe that the affect of the relative humidity and wind speed has been covered by the FFMC index. In this study it is shown that GAMs can be successfully used for the nonlinear relationship besides with some linear predictors for a bi-

**TABLE 1:** Selection of models via UBRE.

| Models | FFMC | DMC | DC | ISI | Temp | Wind | RH | UBRE |
|---|---|---|---|---|---|---|---|---|
| 1 | p | | | | | | | 0.172 |
| 2 | | | s (insig) | | p | s (insig) | | 0.140 |
| 3 | | | s (insig) | | p | | | 0.121 |
| 4 | | | | | p | s (insig) | | 0.129 |
| 5 | | | | | | p (insig) | | 0.158 |
| 6 | | | s (insig) | | p | | s (insig) | 0.136 |
| 7 | | | s (insig) | | p | s (insig) | s (insig) | 0.156 |
| 8 | | | s (insig) | | s (insig) | s (insig) | | 0.124 |
| 9 | s | | | | p | | s | 0.034 |
| 10 | s | | | | p | s(insig) | | 0.032 |
| 11 | s(insig) | | | p(insig) | p | | | 0.032 |
| 12 | s(insig) | | | s(insig) | p | | | 0.021 |
| 13 | | | | s(insig) | p | | | 0.016 |
| 14 | | | | | p(insig) | | s(insig) | 0.125 |
| 15 | s | | | | p | | | 0.012* |
| 16 | s | | | | s | | | 0.012* |
| 17 | s | s | s | | p | | | -0.05 |

p: parametric linear component; s: smooth function ; insig: insigficant

nomial response variable. We should emphasize that the studies with small $R^2$ (determination of coefficient) value is commonly obtained in many fields. This might be the indication of some important variables are excluded from the model or some more complex models are needed. Generalized additive models are handy tools to overcome these problems.

## REFERENCES


1. Buja A, Hastie T, Tibshirani R. Linear smoothers and additive models (with discussion). Ann Statist 1989;17(2):453-555.
2. Hastie TJ, Tibshirani, RJ. Additive models. Generalized Additive Models. 1st ed. New York: Chapman and Hall; 1990. p.82-103.
3. Stone CJ. The use of polynomial splines and their tensor products in multivariate function estimation. Ann Statist 1994;22(1):118-71.
4. Newey WK. Convergence rates and asymptotic normality for series estimators. J Econom 1997;79(1):147-68.
5. Opsomer JD, Ruppert D. Fitting a bivariate additive model by local polynomial regression. Ann Statist 1997;25(1):186-211.
6. Opsomer JD. Asymptotic properties of backfitting estimators. J Multivar Anal 2000; 73(2)166-79.
7. Craven P, Wahba G. Smoothing noisy data with spline functions. Numer Math 1979;31(4): 377-403.
8. Wahba G. Estimating the smoothing parameter. Spline Models for Observational Data. 1st ed. Pennsylvania: SIAM; 1990. p 45-75.
9. Preisler HK, Haase SM, Sackett SS. Modeling and risk assessment for soil temperatures beneath prescribed forest fires. Environ Ecol Stat 2000;7(3):239-54.
10. Ballart X, Biba C. Forest fires: evaluation of government measures. Policy Sciences 2002;35(4):361-77.
11. Miller JD, Yool SR. Modeling fire semi-desert grassland/oak woodland: the spatial implications. Ecol Model 2002;153(3):229-45.
12. Diaz-Delgado R, Lloret F, Pons X. Statistical analysis of fire frequency models for catalonia (NE Spain, 1975-1998) based on fire scar maps from Landsat MSS data. Int J Wildland Fire 2004;13(1):89-99.
13. Mandallaz D, Ye R. Prediction of forest fires with Poisson models. Can J For Res 1997; 27(10):1685-94.
14. Martell DL, Otukol S, Stocks SL. A logistic model for predicting daily people-caused forest fire occurrence in Ontario. Canad J Forest Res 1987;17(5):394-401.
15. Poulin-Costello M. People-caused forest fire prediction using poisson and logistic regression. Master's thesis. Victoria: Dept of Math and Stat University of Victoria. 1993. p.116.
16. Brillinger DR, Preisler HK, Benoit JW. Probabilistic risk assessment for wildfires. Environmetrics 2006;17(6):623-33.
17. Nelder JA, Wedderburn RWM. Generalized linear models. J Royal Statis Soc 1972;135(A): 370-84.
18. Marx BD, Eilers PHC. Direct generalized additive modeling with penalized likelihood. Comput Statist&Data Anal 1998;28(2):193-209.
19. Wood SN. Introducing GAMs. Generalized Additive Models: An Introduction with R.1st ed. London: CRC Press 2006. p.121-220.
20. Cortez P, Morais A. A data mining approach to predict forest fires using meteorological data. In: Neves J, Santos MF, Machado J, eds. New Trends in Artificial Intelligence. Proceedings of the 13th EPIA Portuguese Conference on Artificial Intelligence. Lisbon: ICEIS; 2007. p.512-23.