# Flexible Logistic Models with Correlated Variables

## Korelasyonlu Değişkenler ile Esnek Lojistik Modeller

Betül KAN KILINÇ,[a]
Mustafa ÇAVUŞ[a]

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

Geliş Tarihi/*Received:* 15.06.2015
Kabul Tarihi/*Accepted:* 13.08.2015

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

**ABSTRACT Objective:** The objective of this work is to examine the correlations among covariates involved in logistic models. Strong correlations among explanatory variables can cause unstable regression coefficients. Thus, a common problem occurred in multiple linear or logistic regression is investigated by a flexible method. **Material and Methods:** Generalized additive models (GAMs) can provide flexible nonparametric modeling of response by using the additive function of the predictor variables. Due to the binary form of the response and the flexibility in additive models, generalized additive logistic regression was used for estimation in this study. Various regression models were developed using a different number of correlated predictor variables and a binary outcome in order to evaluate and compare the model performance in terms of Akaike Information Criteria (AIC). **Results:** The change in the basis made little difference to the AIC values at the same correlation parameter. The models where smooth terms were represented by cubic regression splines were more often tended to give smaller AIC values than the models where other splines were used. **Conclusion:** Additive logistic regression with high correlated explanatory variables produced reasonable results in terms of AIC for all sample sizes when the consistency was provided between the structure of the explanatory variables with the response variable and correlation parameter.

**Key Words:** Nonparametric; regression; correlation

**ÖZET Amaç:** Bu çalışmada amaçlanan konu aralarında bağlantı olan açıklayıcı değişkenler içeren lojistik regresyon modellerini incelemektir. Açıklayıcı değişkenler arasındaki güçlü ilişkiler regresyon modeli katsayılarının güvenilmez olmasına neden olmaktadır. Bu yüzden, çoklu doğrusal ya da lojistik regresyon modellerinde yaygın olan bu problem esnek bir metot yardımıyla incelenmiştir. **Gereç ve Yöntemler:** Genelleştirilmiş Toplamsal Modeller (GAMs) bağımsız değişkenlerin toplamsal fonksiyonları yardımıyla yanıt değişkenini daha esnek bir şekilde modellemeye çalışan parametrik olmayan bir yöntemdir. Yanıt değişkeninin ikili yapıda olması ve toplamsal modellerin sağladığı esneklik nedeniyle bu çalışmada modelin tahminlemesi için Genelleştirilmiş Toplamsal Lojistik Modeller (GALMs) kullanılmıştır. Farklı sayıda bağlantılı açıklayıcı değişkenler ve ikili yanıt değişkeni kullanılarak çeşitli regresyon modelleri oluşturulmuştur ve bu modellerin performansı Akaike Bilgi Kriteri (AIC) yardımıyla değerlendirilmiştir. **Bulgular:** Korelasyon parametresinin aynı değerlerinde taban fonksiyonundaki değişim AIC değerinde çok fazla değişiklik yaratmamıştır. Düzeltme terimlerinin kübik eğri fonksiyon olarak tanımlandığı modellerden elde edilen AIC değerleri, düzeltme terimlerini oluşturmada kullanılan diğer eğrilerin oluşturduğu modellerden elde edilen AIC değerlerinden daha küçüktür. **Sonuç:** Toplamsal lojistik regresyon modeli, yüksek korelasyonlu açıklayıcı değişkenlerin ikili değerler alan yanıt değişkenini açıklamada AIC ölçütüne göre iyi sonuçlar üretmiştir.

**Anahtar Kelimeler:** Parametrik olmayan; regresyon; korelasyon

**Turkiye Klinikleri J Biostat 2015;7(2):90-5**

doi: 10.5336/biostatic.2015-46786

Generalized Additive Models (GAMs) was introduced by Hastie and Tibshirani in 1986.[1] They allow the linear relationships in multiple regression to be examined in a wider class of nonlinear relationships between predictor variables and the response. The use of GAM has considerably flexible effect in statistical modeling. GAMs can be extended to the logistic case as well. The generalized additive model was studied for logistic cases by Casetti in 1972.[2] Wrigley and Dunn (1986) showed the importance of fitting of smooth functions to the detection of nonlinearities. Unfortunately, the method they describe may lead to false results when there is more than nonlinearity present in the data.[3] Jones and Wrigley studied a more detailed nature of the logistic, binary-response case within the penalized least squares, cubic splines, and iteratively reweighted least squares estimation in 1995.[4] This paper examines the relationship between the correlated predictor variables and the binary response within additive models and provides an opportunity to see how additive logistic regression handles with correlated variables in the model.

## METHODS

In this section we briefly explained the theoretical part of the study. Firstly, additive models, some of the regression splines, and additive logistic model are explained in details.

### ADDITIVE MODELS AND PENALIZED REGRESSION

Consider the additive model structure as given in Eq. (1)

$$Y_i = f_j(x_i) + f_j(z_i) + f_j(k_i) + \varepsilon_i \ , \ x_i, z_i, k_i \epsilon[0,1], \quad i = 1,2,\dots,n, j = 1,2,3 \qquad (1)$$

where $f_j$ are smooth functions and the $\varepsilon_i$ are i.i.d N($0,\sigma^2$) random variables. $Y_i$ is the response variable. The additive model can be represented using penalized regression splines and estimated by penalized least squares.[5]

Penalized regression fitting curve is more flexible than classical least squares regression whereby the relationship between response and explanatory variables can be explained better. For the curve construction, penalized regression model has an extra term which is the penalty term:

$$\sum_{i=1}^{n}[Y_i - f(x_i)]^2 + \lambda \int_a^b [f''(x)]^2\, dx \qquad (2)$$

where $\lambda \int_a^b [f''(x)]^2\, dx$ is penalty term with smoothing parameter $\lambda$ represents the rate of change between residual error and local variation.[6] Low levels of the smoothing parameter are formed the regression curve more roughness and high levels of them are formed the regression curve less roughness. $\lambda \to \infty$ leads to a straight line estimate for $f$, while $\lambda = 0$ results in an un-penalized regression line estimate.[5] Penalized regression defines the $f$ function as regression curve by minimizing the Eq. (2).

### SMOOTH FUNCTIONS

Smooth functions, are also known as splines, used for explaining the response variable with explanatory variables. In the following sections, some common smooth functions are explained such as cubic splines, thin plate splines, and thin plate shrinkage splines.

#### Cubic Splines (CS)

CS can be described as basically a curve which is constructed from sections of different cubic polynomials. Consider an interval $[a, b]$ which can be defined as construction of $[x_i, x_{i+1}]$ where any $x_i$ is named as knots. A function $f$ defined on $[a, b]$ is a cubic spline if two conditions are satisfied. These conditions are

- On each of the intervals $[x_i, x_{i+1}]$, $f$ must be a cubic polynomial.
- On each knots, $f$, its first and second derivatives must be continuous on the whole of $[a, b]$.

An example of cubic spline can be shown as below

$$f(x) = \alpha_1(x - x_i)^3 + \alpha_2(x - x_i)^2 + \alpha_3(x - x_i) + \alpha_4 \quad \text{where } x_i \leq x \leq x_{i+1} \qquad (3)$$

$\alpha_1, \alpha_2, \alpha_3, \alpha_4$ are constant which obtain the shape of the spline. The disadvantage of this kind of spline is that there are many parameters which must be estimated.[6]

#### Thin Plate Splines (TPS)

TPS is introduced to geometric design by Duchon in 1977.[7] TPS is thought as the generalization of smoothing splines to more than one dimension. In this type of smoothing splines, there are at least two dimensional explanatory variables in data set.

TPS provides some advantages to researchers such as no obligatory for choosing knot positions or basis functions and also TPS allows the researcher some flexibility to select the order of derivative used in the measure of function wiggliness. As well as the advantages, it has heavy computational cost according the other type splines because of the number of unknown parameters. Instead thin plate regression splines are used based on the idea of truncating the space of wiggly components of the thin plate spline. They also avoid the problem of knot placement and are very cheap to compute.[5]

### Thin Plate Shrinkage Splines (TPSS)

TPSS is another type of smooth functions. It is used when the dimension of the estimation problem is so high. Model selection with GAMs, it would be more convenient if smooth functions could be zeroed by adjustment of smoothing parameters. There is two way to this adjustment

- Adding an extra penalty
- Adding a small multiple of the identity matrix to the penalty matrix

First one would open up the possibility of penalizing the smooth components of a function more than the wiggly components. Otherwise, the penalty shrinks all parameters to zero if its associated smoothing parameter is large enough by the second way.

$$S \rightarrow S + \epsilon I \tag{4}$$

Here, $I$ is identity matrix and $\in$ is small multiple constant. If $\in$ is small enough, the identity part of the penalty have almost no impact when a function is wiggly. It becomes close to completely smooth, the identity component become important and start shrinking the parameters towards zero.[5]

## GENERALIZED ADDITIVE MODELS

GAMs were originally developed by Hastie and Tibshrani to blend the properties of Generalized Linear Models with additive models.[8] GAMs are more flexible than the other regression approaches Consider the model where response variable follows any exponential family distribution as below

$$g[E(Y_i)] = f_0 + f_1(x) + f_2(x_2) + \cdots + f_k(x_k) = f_0 + \sum_{i=1} f_k(x_k) \tag{5}$$

where $Y_i$ is response variable which is distributed as any exponential family distribution, $x_i$ are explanatory variables, g is the link function and $f_i$ are smooth functions of the model.

## GENERALIZED ADDITIVE LOGISTIC MODEL

Generalized Additive Logistic Model (GALM) can be considered as a mixture of additive and logistic models. In this model, the probability of possible outcome of response variable which is explained by smooth functions of explanatory variables. The expected value of binary response variable is $E(x) = P(Y = 1|X)$ to the explanatory variables in the regression model with logit function as in Eq. (6):

$$log\left(\frac{E(x)}{1 - E(x)}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k \tag{6}$$

The GALM replaces each linear term by a smooth function form as in Eq. (7):

$$log\left(\frac{E(x)}{1 - E(x)}\right) = f_0 + f_1(X_1) + f_2(X_2) + \cdots + f_k(X_k) \tag{7}$$

where $f_i$'s are smooth functions. The GALM is a specific example of a GAM. Hence, the expected value of the response variable $E(x)$ is related to an additive function of the explanatory variables by a link function of $g$ in Eq. (5).[8]

# APPLICATION

In the simulation study, four different scenarios are designed. Four logistic regression models were fitted using different combinations of smooth functions such as TPS, TPSS and CS. In each case, AIC is calculated for the fitted models. Thus, simulation models are used to obtain information on the performance of the models. First, a random sample of 50, 100, and 1000 observations are generated. Within each sample, three correlated covariates and the response are simulated. Thus, the simulated data sets consist of three correlated continuous variables and one binary response variable. The response variable is generated from binomial distribution while correlated continuous variables are generated from the standard normal distribution. For each data set, the additive logistic regression method is replicated for 1000 times. To illustrate how different correlations could affect the estimation, three different correlation coefficients are considered as negligible, moderate and small in each scenario. Indeed, the user is set to free to define a constant correlation between two variables in the algorithm for forward studies. However in this study, the correlations are limited to ±0.90, ±0.60, and ±0.10. The performance of the fitted models is compared in terms of AIC values. R Software is used for analysis.

## RESULTS

Akaike Information Criteria (AIC) was computed for every different model. The results of additive logistic model are tabulated in terms of AIC values and sample sizes, as well as the correlations among variables in Table 1. As seen in Table 1, the last six columns refer to the smooth functions used in the fitting of the binary response variable. The choice of the basis functions for smooth terms given in columns 4, 5, and 6 are TPS, TPSS and CS, respectively. This means that the smooth terms in Model 1 represent the TPS, TPSS and CS for each predictor respectively. The last three columns in Table 1 shows the modified smooth terms in Model 1 where TPS, TPSS and CS are used at the same time.

The additive logistic model is first applied to the model, $y=0.5+3x_1-5x_2+6x_3$. Here, $x_i$ is generated from standard normal distribution for i =1, 2. $x_1$ and $x_2$ are correlated with correlation r = 0.10, 0.60, 0.90. Additionally, $x_3$ is generated as $x_3 = x_1 + N(0,1)$ for all four models. In the model $x_2$ has a negative effect on y whereas $x_1$ and $x_3$ have positive effect on y. According to the simulation results, AIC value increases when the correlation parameter increases for different combinations of the smooth functions in Model 1 for all sample sizes. The smallest AIC value is obtained as 10.4969 from a cubic splines fit at r = 0.10 whereas the largest AIC is obtained as 15.6184 from a thin plate shrinkage splines fit at r = 0.90 for sample size of 50. The similar results are hold for sample sizes of 100 and 500 (Table 1).

In the second model, $x_i$ is also generated from standard normal distribution for i =1, 2. $x_1$ and $x_2$ are correlated with correlation r = 0.10, 0.60, 0.90 as in the 1st scenario. Moreover, the model is set that $x_1$, $x_{2,}$ and $x_3$ have all positive effect on y in the model. Simulation results show that the smallest AIC value is obtained as 19.025 from a cubic splines fit at r = 0.90 whereas the largest AIC is obtained as 27.0261 from a thin plate splines fit at r = 0.10 for sample size of 50 for the second model, $y=0.75+2x_1+0.5x_2+1.5x_3$. The similar results are hold for sample size of 100. On the other hand, the smallest AIC value is obtained as 249.9221 from a combination of (CS, TPS, CS) of smooth terms fit at r = 0.90 where the largest AIC is obtained as 294.6806 from a combination of (TP, TPSS, CS) of smooth terms at r = 0.10 for sample size of 500. Notice that AIC value decreases when the correlation parameter increases for all sample sizes (Table 1).

In the third model, $x_i$ is generated from standard normal distribution for i =1, 2 as in the 1st and 2nd scenarios. However, $x_1$ and $x_2$ are correlated with correlation r = -0.10, -0.60, -0.90. Here, although $x_1$, $x_{2,}$ are correlated with negative r, the model is set that $x_{1,}$ $x_2$, and $x_3$ have all positive effect on y. Hence, the smallest AIC value is obtained as 22.5958 for a cubic splines fit at r = -0.10 where the largest AIC value is obtained as 33.8512 from a thin plate splines fit at r= -0.90 for sample size of 50 for the third model, $y=0.75+2x_1+0.5x_2+1.5x_3$. The similar results are hold for sample size of 100. On the other hand, the smallest AIC value is obtained as 299.6301 from a combination of (CS, TPS, CS) of smooth terms at r = -0.10 where the largest AIC is obtained as 360.0575 from thin plate shrinkage splines fit at r = -0.90 for sample size of 500. Thus, AIC value increases when the correlation parameter increases in absolute value for all sample sizes (Table 1).

In the fourth model, $x_i$ is generated from standard normal distribution for i =1, 2 as in the previous scenarios. However $x_1$ and $x_2$ are correlated

**TABLE 1:** Akaike information criteria results of the simulation study.

| Model<br>(1) | Sample Size<br>(n)<br>(2) | Correlation Parameter<br>(r)<br>(3) | Thin Plate Spline<br>(TPS)<br>(4) | Thin Plate Shrinkage Spline<br>(TPSS)<br>(5) | Cubic Spline<br>(CS)<br>(6) | TPS<br>CS<br>TPSS<br>(7) | CS<br>TPSS<br>TPS<br>(8) | CS<br>TPS<br>CS<br>(9) |
|---|---|---|---|---|---|---|---|---|
| MODEL 1<br>$y = 0.5 + 3x_1 - 5x_2 + 6x_3$<br>with positive correlations | 50 | 0.10 | 11.2335 | 11.0030 | 10.4969 | 10.8522 | 10.9521 | 10.5641 |
| | | 0.60 | 12.8581 | 13.0200 | 11.7904 | 12.4516 | 12.6395 | 12.1108 |
| | | 0.90 | 15.1172 | 15.6184 | 13.8236 | 14.6542 | 14.9314 | 14.4707 |
| | 100 | 0.10 | 21.4117 | 21.8416 | 19.0117 | 20.3974 | 20.9324 | 20.1984 |
| | | 0.60 | 26.0498 | 26.5989 | 22.8273 | 24.9518 | 25.3715 | 24.3468 |
| | | 0.90 | 30.5384 | 31.0657 | 26.9748 | 29.6248 | 29.6891 | 28.4794 |
| | 500 | 0.10 | 113.9904 | 114.6857 | 113.6364 | 114.0834 | 114.1791 | 114.0392 |
| | | 0.60 | 138.2460 | 139.2595 | 138.0487 | 138.4316 | 138.7142 | 138.5280 |
| | | 0.90 | 162.4402 | 164.0667 | 162.3754 | 162.9664 | 163.2637 | 163.1849 |
| MODEL 2<br>$y = 0.75 + 2x_1 + 0.5x_2 + 1.5x_3$<br>with positive correlations | 50 | 0.10 | 27.0261 | 26.6287 | 22.1714 | 25.4393 | 24.3779 | 22.7335 |
| | | 0.60 | 25.5721 | 25.0575 | 20.9916 | 24.1200 | 23.2024 | 21.6981 |
| | | 0.90 | 23.3861 | 22.7095 | 19.0250 | 21.7891 | 21.2184 | 19.5354 |
| | 100 | 0.10 | 59.4765 | 59.0896 | 55.3316 | 58.7004 | 58.1503 | 56.6435 |
| | | 0.60 | 55.8456 | 55.5017 | 51.7551 | 54.8000 | 54.4489 | 53.1756 |
| | | 0.90 | 49.7524 | 49.2998 | 45.8214 | 48.8786 | 48.5386 | 46.9888 |
| | 500 | 0.10 | 294.5731 | 294.6765 | 294.4920 | 294.6806 | 294.4890 | 294.4065 |
| | | 0.60 | 278.5673 | 278.6108 | 278.4344 | 278.6620 | 278.4174 | 278.3432 |
| | | 0.90 | 250.2094 | 250.1426 | 249.9449 | 250.1770 | 250.0325 | 249.9221 |
| MODEL 3<br>$y = 0.75 + 2x_1 + 0.5x_2 + 1.5x_3$<br>with negative correlations | 50 | -0.10 | 27.4741 | 27.2363 | 22.5958 | 25.8970 | 25.0455 | 23.2643 |
| | | -0.60 | 29.1396 | 28.8466 | 23.4083 | 27.6549 | 26.5452 | 24.4317 |
| | | -0.90 | 33.8512 | 33.4914 | 27.3512 | 31.9290 | 31.4558 | 28.9870 |
| | 100 | -0.10 | 60.2703 | 59.9601 | 56.3572 | 59.6121 | 58.9083 | 57.4580 |
| | | -0.60 | 63.7203 | 63.3633 | 60.3737 | 63.2379 | 62.6634 | 61.5378 |
| | | -0.90 | 73.2449 | 72.9077 | 70.0161 | 72.6721 | 72.4211 | 71.3603 |
| | 500 | -0.10 | 299.7593 | 299.8944 | 299.7432 | 299.9221 | 299.7305 | 299.6301 |
| | | -0.60 | 318.6037 | 318.7693 | 318.5133 | 318.6938 | 318.4705 | 318.4499 |
| | | -0.90 | 359.7732 | 360.0575 | 359.7934 | 359.9033 | 359.8187 | 359.7591 |
| MODEL 4<br>$y = 0.75 - 2x_1 + 0.5x_2 + 1.5x_3$<br>with negative correlations | 50 | -0.10 | 45.1992 | 45.2119 | 36.8272 | 42.8986 | 42.4227 | 39.2779 |
| | | -0.60 | 42.9330 | 42.9295 | 35.1892 | 41.0284 | 40.3572 | 37.3828 |
| | | -0.90 | 38.6097 | 38.6206 | 31.9531 | 36.8773 | 36.5866 | 33.7497 |
| | 100 | -0.10 | 96.8138 | 97.0306 | 96.2252 | 96.9587 | 96.5721 | 96.3271 |
| | | -0.60 | 94.5108 | 94.5462 | 93.0483 | 94.0511 | 93.9939 | 93.6507 |
| | | -0.90 | 85.2246 | 84.9196 | 82.5565 | 84.7648 | 84.3686 | 83.5335 |
| | 500 | -0.10 | 483.6293 | 484.1915 | 483.9036 | 483.9441 | 483.7461 | 483.8251 |
| | | -0.60 | 467.8540 | 468.3894 | 468.0983 | 468.1639 | 467.9549 | 468.0820 |
| | | -0.90 | 426.9172 | 427.3187 | 427.0826 | 427.1735 | 426.9890 | 427.0860 |

with correlation r = -0.10, -0.60, -0.90 as in the 3rd scenario. In the model, $x_1$ has a negative effect on y whereas $x_2$ and $x_3$ have both positive effect on y. For the last model $y=0.75-2x_1+0.5x_2+1.5x_3$, the smallest AIC value is obtained as 31.9531 for a cubic splines fit at r = -0.90 where the largest AIC value is obtained as 45.2119 from a thin plate shrinkage splines fit at r = -0.10 for sample size of 50. The similar results are hold for sample size of 100. On the other hand, the largest AIC value is obtained as 484.1915 from a thin plate shrinkage splines at r = -0.10 where the smallest AIC is obtained as 426.9172 from thin plate splines fit at r = -0.90 for sample size of 500. Notice that AIC value decreases when the correlation parameter in absolute value increases for all sample sizes (Table 1).

## CONCLUSION

Correlated data occur very frequently biostatistics. In this study, generalized additive logistic regression is used to examine the relationship between the correlated explanatory variables and a binary response. This method uses the smooth functions that provide flexibility to the relationship between the variables in the model. For this aim, a binary response variable and some explanatory variables are generated for different sample sizes and correlation parameters. Simulation study is conducted to examine the effect of the correlations between variables in additive logistic regression. The basis used to represent the smooth terms is varied from thin plate regression splines, thin plate shrinkage splines, cubic regression splines, and some combinations of them.

In general, the change in the basis makes little difference to the AIC values at the same correlation parameter. However, the models where smooth terms are represented by cubic regression splines are more often tend to give smaller AIC values than the models where other splines are used. Additionally, when correlation parameter r increases AIC values also increase because of the negative structure of the explanatory variables with the response in the first model. However when correlation parameter r increases AIC values decrease due to the positive structure of the explanatory variables with the response in the second model. In the third model, when the absolute of correlation parameter increases AIC values also increase because of the positive structure of the explanatory variables with the response. In the last model, when the absolute correlation parameter r increases AIC values decrease due to the negative structure of the explanatory variables with the response. As a conclusion of this study, additive logistic regression with high correlated explanatory variables produces better results in terms of AIC for all sample sizes when the consistency is provided between the structure of the explanatory variables with the response variable and correlation parameter. Finally, additive logistic regression when specifically the highly correlated variables are in the data could serve a useful tool for statistical modeling.

***Acknowledgement***

*This study was supported by Anadolu University Scientific Research Projects Commission under the grant no: 1501F018.*

## REFERENCES

1. Hastie T, Tibshirani R. Generalized additive models. Statistical Science 1986;1(3):297-318.
2. Casetti E. Generating models by the expansion method: applications to geographical research. Geographical Analysis 1972;4(1): 81-91.
3. Wrigley N, Dunn R. Graphical diagnostics for logistic oil exploration models. Mathematical Geology 1986;18(4):355-74.
4. Jones K, Wrigley N. Generalized additive models, graphical diagnostics and logistic regression. Geographical Analysis 1995;27(1):1-18.
5. Wood SN. Introducing GAMs. Generalized Additive Models. 1st ed. London: Chapman & Hall/CRC; 2006. p.133-9.
6. Green PJ, Silverman BW. Introduction. Nonparametric Regression and Generalized Linear Models. 1st ed. London: Chapman & Hall; 1995. p.5-6.
7. Duchon J. Splines minimizing rotation-invariant semi-norms in solobev spaces. In: Schemp W, Zeller K, eds. Construction Theory of Functions of Several Variables: Lecture Notes in Mathematics. 1st ed. Vol. 571. Berlin: Springer; 1977. p.85-100.
8. Hastie T, Tibshirani R, Friedman J. Additive models, trees and related methods. The Elements of Statistical Learning. 2nd ed. California, Standford: Springer; 1986. p.295-304.