International Journal of Advanced Statistics and Probability, *1 (3) (2013) 64-81*
©Science Publishing Corporation
www.sciencepubco.com/index.php/IJASP

Generalized additive models in business and economics

Sunil K. Sapra

California State University, Los Angeles, USA
E-mail: ssapra@calstatela.edu

Abstract

The paper presents applications of a class of semi-parametric models called generalized additive models (GAMs) to several business and economic datasets. Applications include analysis of wage-education relationship, brand choice, and number of trips to a doctor's office. The dependent variable may be continuous, categorical or count. These semi-parametric models are flexible and robust extensions of Logit, Poisson, Negative Binomial and other generalized linear models. The GAMs are represented using penalized regression splines and are estimated by penalized regression methods. The degree of smoothness for the unknown functions in the linear predictor part of the GAM is estimated using cross validation. The GAMs allow us to build a regression surface as a sum of lower-dimensional nonparametric terms circumventing the curse of dimensionality: the slow convergence of an estimator to the true value in high dimensions. For each application studied in the paper, several GAMs are compared and the best model is selected using AIC, UBRE score, deviances, and R-sq (adjusted). The econometric techniques utilized in the paper are widely applicable to the analysis of count, binary response and duration types of data encountered in business and economics.

Keywords: *Generalized additive models (GAMs), Generalized Linear Models (GLMs), Logit Models, Poisson Regression Models, Penalized Regression Splines.*

1 Introduction

Discrete choice and count data regression models are pervasive in business and economics. Logit and Poisson regression models are the most commonly employed models for non-normal data. Common applications of Logit models include analysis of brand choice data in marketing (Baltas [1] and Guadagni and Little [4]) and transportation choice data in economics (Greene [3] and Manski and McFadden [12]). Poisson regression models have been applied in the analysis of data on patents, number of trips to a doctor's office, and number of shipping accidents. These regression models belong to the class of generalized linear models (GLMs), which relax the assumption that the response is normally distributed by allowing it to follow any distribution from the exponential family, such as normal, Poisson, binomial, gamma etc. Inference for GLMs is based on likelihood theory. McCullagh and Nelder [11] provide an authoritative account of GLMs and Cameron and Trivedi [2] and Greene [3] provide econometric applications.
In recent years, semi-parametric extensions of linear regressions have been employed in business and economics. Nevertheless, there have been relatively few applications of semi-parametric extensions of generalized linear models in business and economics despite applications in other fields (see Hastie and Tibshirani [7]). This paper attempts to fill this gap. We present several econometric applications of the generalized additive model (GAM), a semi-parametric extension of GLMs and demonstrate the usefulness of these semiparametric models for the analysis of continuous, discrete, and count data. A GAM is a semi-parametric GLM in which part of the linear predictor is specified in terms of a sum of unknown smooth functions of explanatory variables. Generalized additive models (GAMs) are a powerful generalization of linear, logistic, and Poisson regression models. GAMs are very flexible, and can provide an elegant fit in the presence of nonlinear relationships. GAMs and GLMs can be applied in similar situations, but they serve different analytic objectives. GLMs emphasize estimation and inference for the parameters of the model, while GAMs focus on exploring data nonparametrically. The GAM approach offers more flexibility in model form than the GLM approach does. These models estimate an additive approximation to the multivariate link function. The main benefit of this approach is that each of the individual additive terms is estimated using a univariate smoother instead of a multivariate smoother for a high-dimensional non-parametric term circumventing the curse of dimensionality: the slow convergence of an estimator to the true value in high dimensions. The GAM formulation of the GLM regression models allows us to build a regression surface as a sum of lower-dimensional nonparametric terms. Two popular approaches to estimation of GAMs are backfitting with local scoring algorithm (Hastie and Tibshirani [6]) and penalized regression

splines (Wood ([14] and [15])). Backfitting has the advantage that it can be used with any scatterplot smoother while the penalized regression splines method has the advantage that estimation of the smoothing parameter using generalized cross validation is integrated into estimation (see Wood [15]). For a discussion of other approaches to GAM estimation, see Wood [15].

This paper presents econometric applications of the GAM extensions of the generalized linear models (GLMs) including the linear regression model and demonstrates that the GAMs can overcome a serious weakness of the GLMs in failing to identify the nonlinearities in the link function. The paper is organized as follows. Section 2 introduces the generalized additive model (GAM). Section 3 presents the penalized regression splines method for the estimation of GAMs. Section 4 presents an econometric application of GAM Gaussian model to wage-education data. Section 5 presents an econometric application of GAM Logit model to Cracker data on choice between four cereal brands and Section 6 presents an application of GAM Poisson regression model to the number of trips to a doctor's office. Section 7 provides some concluding remarks.

2 Generalized additive models

Generalized additive models are nonparametric generalized linear models. These models extend traditional linear models by allowing for a link between the nonlinear predictor $f(x_1, \ldots, x_p)$ and the expected value of y. This amounts to allowing for an alternative distribution for the underlying random variation besides just the normal distribution. While Gaussian models can be used in many statistical applications, for several types of problems they are not appropriate. For example, the normal distribution may not be adequate for modeling discrete responses such as counts, or bounded responses such as proportions.

Generalized additive models consist of a random component, an additive component, and a link function relating these two components. The response y, the random component, is assumed to have a density in the exponential family

$$f_Y(y;\theta,\varphi) = exp\left\{\frac{y\theta - b(\theta)}{a(\varphi)} + c(y.\varphi)\right\}, \tag{1}$$

where θ is called the natural parameter and ϕ is the scale parameter. The normal, binomial, and Poisson distributions are all in this family. Unlike the generalized linear models, the mean $\mu = E(y \mid x_1, x_2, \ldots, x_p)$ is not linked to the linear predictor $\sum x_{ij}\beta_j$, but to the nonlinear nonparametric predictor

$$\eta = g(\mu) = \alpha + \sum_{j=1}^{p} s_j(x_j), \tag{2}$$

where $s_1(\cdot), \ldots, s_p(\cdot)$ are smooth nonparametric functions, which defines the additive component. Finally, the relationship between the mean μ of the response variable and η is defined by a link function $g(\mu) = \eta$. The most commonly used link function is the canonical link, for which $\eta = \theta$.

A combination of backfitting and local scoring algorithms are used in the actual fitting of the model. In order to fit GAMs to the data, following Wood [15], we use basis expansions of smooth functions and penalized likelihood maximization for model estimation in which wiggly models are penalized more heavily than smooth models in a controllable manner, and degree of smoothness is chosen based on cross validation, or AIC or Mallows' criterion.

The generalized additive models are fit to count data or binary response data by maximizing a penalized log likelihood or a penalized log partial-likelihood. To maximize it, the backfitting procedure is used in conjunction with a maximum likelihood or maximum partial likelihood algorithm (Hastie and Tibshirani ([6] and [8]) and Wood [15]). The Newton-Raphson method for maximizing log-likelihoods in these models can be presented in an IRLS (iteratively reweighted least squares) form. It involves a repeated weighted linear regression of a constructed response variable on the covariates: each regression yields a new value of the parameter estimates which give a new constructed variable, and the process is iterated. In the generalized additive model, the weighted linear regression is simply replaced by a weighted backfitting algorithm (Hastie and Tibshirani [6]).

3 Estimation of generalized additive models using Penalized regression method

Algorithm for Penalized Iteratively Re-weighted Least Squares (PIRLS) (Wood [15])

The GAMs are fit to data by maximizing a penalized log likelihood or a penalized log partial-likelihood. The following algorithm is used to implement these methods. The R-package mgcv (Wood [14]) was used for computations.

Step 1: Initialize $\hat{\alpha} = g(1/N\sum_{i=1}^{N} y_i), s_j^0 = 0, j = 1,2,...,p.$

Step 2: Construct an adjusted dependent variable z_{ij} *as*

$z_i = \eta_i^0 - (y_i - \mu_i^0)(\partial\eta_i / \partial\mu_i)_0$

$\eta_i^0 = g(\mu_i^0) = \alpha^0 + \sum_{j=1}^{p} s_j^0(x_{ij})$ *and* $\mu_i^0 = g^{-1}(\eta_i^0).$

Step 3: Compute weights

$w_i = (\partial\mu_i / \partial\eta_i)_0 (V_i^0)^{-1},$

where V_i^0 *is the variance of y at* μ_i^0,

Step 4: Penalized Spline Re gression

Minimize $\left\|\sqrt{W}(z - X\beta)\right\|^2 + \lambda\beta' S\beta$ *with respect to* β, *where X is the matrix of data on basis functions used to represent the regressionfunction, W is a diagonal matrix with i-th diagonal element* w_i, *S is a matrix of known coefficients in the penalty function* $\beta' S\beta$ *and* λ *is a smoothing parameter. Compute* s_j^1, η^1, *and* μ_i^1, *the second stage estimates of* s_j, η, *and* μ_i.

Step 5: Repeat steps 2-4 replacing η^0 *by* η^1 *until the difference between two successive values of* η *is less than a small prespecified number and convergence is obtained.*

4 The generalized additive Gaussian model

The generalized additive Gaussian model assumes that

$g(\mu_i) = \mu_i = \alpha + \sum_{j=1}^{p} s_j(x_{ij})$ *for the identity link and*

$g(\mu_i) = \ln \mu_i = \alpha + \sum_{j=1}^{p} s_j(x_{ij})$ *for the log link.*

The adjusted dependent variable *z* and the weights *w* used in the algorithm above are

$z_i = y_i$ *and* $w_i = 1$ *for all i for the identity link and*

$z_i = \ln y_i$ *and* $w_i = 1$ *for all i for the log link.*

The functions $s_1, s_2, ..., s_p$ are estimated by the algorithm described earlier.

4.1 An Empirical Application of GAM Gaussian Model to data on wages

Variable Definitions and Data Description

A small subset of CPS data on labor force status are taken are from Hill et al [9]. The dependent variable is *WAGE*. The variables are defined as follows.

WAGE = Earnings per hour
EDUC = Years of education
EXPER = Post education years experience
HRSWK = Usual hours worked per week
MARRIED = 1 if married.

Table 1: Summary Statistics

Variable	Obs	Mean	Std Dev	Min	Max
WAGE	1000	20.20122	12.1038	2.03	72.13
EDUC	1000	10.689	2.44013	1	16
EXPER	1000	26.501	12.99041	3	64
HRSWK	1000	39.24	11.44611	0	99

Models

The following models were fitted to the data. Model 1 is a linear regression model with *WAGE* as the dependent variable and *EDUC*, *EXPER*, and *HRSWK* as the independent variables, which is a GLM model with identity link, Model 2 introduces a quadratic term, the square of *HRSWK*. Model 3 is a GAM model with the identity link, which introduces a nonparametric smooth term for *HRSWK* since a high degree of nonlinearity is observed in a partial residual plot of *HRSWK* displayed in Fig. 1. Models 4, 5, 6, and 7 use the identity link with ln(*WAGE)* as the dependent variable. Model 4 is a GAM model with independent variables *EDUC*, *EXPER*, and *HRSWK* , which is a GLM model with identity link for *ln(WAGE)*. Model 5 is a GAM model which employs parametric terms for *EDUC* and *EXPER* but a nonparametric smooth term for *HRSWK*. Model 6 is a GAM model, which includes parametric terms for *EDUC* and *EXPER*, a parametric interaction term for interaction between *EXPERIENCE* and *MARRIED*, but a nonparametric smooth term for *HRSWK*. Model 7 replaces the parametric interaction term for interaction between *EXPERIENCE* and *MARRIED*, with a nonparametric smooth interaction term.

$$Model\ 1: GLM\ \ Normal\ \ with\ \ Identity\ \ Link$$

$$\eta = g(\mu) = \beta_1 + \beta_2 EDUC + \beta_3 EXPER + \beta_4 HRSWK$$

$$Model\ 2: GLM\ \ Normal\ \ with\ a\ Quadratic\ Term\ and\ \ Identity\ \ Link$$

$$\eta = g(\mu) = \beta_1 + \beta_2 EDUC + \beta_3 EXPER + \beta_4 HRSWK$$

$$+\beta_4 HRSWKSQ$$

$$Model\ 3: Generalized\ \ Additive\ \ Re\ gression\ \ Models\ \ with\ \ Identity\ \ Link\ \ for\ \ Wages\ \ and\ \ a$$

$$Nonparametric\ \ Smooth\ \ Term\ \ for\ \ HRSWK$$

$$\eta = g(\mu) = \beta_1 + \beta_2 EDUC + \beta_3 EXPER + \beta_4 HRSWK + s(HRSWK)$$

$$Model\ 4: GLM\ \ Normal\ \ with\ \ Identity\ \ Link\ \ for\ \ ln(Wages)$$

$$\eta = g(\mu) = \beta_1 + \beta_2 EDUC + \beta_3 EXPER + \beta_4 HRSWK$$

$$Model\ 5: GAM\ \ Normal\ \ with\ \ Identity\ \ Link\ \ for\ \ ln(Wages)\ \ and\ \ a\ \ Nonparametric\ \ Smooth$$

$$Term\ \ for\ \ HRSWK$$

$$\eta = g(\mu) = \beta_1 + \beta_2 EDUC + \beta_3 EXPER + s(HRSWK)$$

$$Model\ 6: GAM\ \ Normal\ \ with\ \ Identity\ \ Link\ \ for\ \ ln(Wages),\ a\ \ Nonparametric\ \ Smooth\ \ Term$$

$$for\ \ HRSWK\ \ and\ \ an\ \ Interaction$$

$$\eta = g(\mu) = \beta_1 + \beta_2 EDUC + \beta_3 EXPER + \beta_4 EXPER * MARRIED + s(HRSWK)$$

$$Model\ 7: GAM\ \ Normal\ \ with\ \ Identity\ \ Link\ \ for\ \ ln(Wages),\ a\ \ Nonparametric\ \ Interaction$$

$$Term\ \ and\ \ a\ \ Smooth\ \ Term\ \ for\ \ HRSWK$$

$$\eta = g(\mu) = \beta_1 + \beta_2 EDUC + \beta_3 EXPER + s(EXPER * MARRIED) + s(HRSWK)$$

Table 2: Model 1: GLM Normal with Identity Link

Variable	Estimate	Std. Error	t-ratio	p-value
INTERCEPT	-14.93450	1.95933	-7.622	$5.80 \times 10^{-14***}$
EDUC	2.14545	0.13904	15.430	$<2 \times 10^{-16***}$
EXPER	0.12594	0.02578	4.886	$1.20 \times 10^{-6***}$
HRSWK	0.22593	0.02920	7.737	$2.49 \times 10^{-14***}$

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for Gaussian family taken to be 108.5920)
Null deviance: 146356 on 999 degrees of freedom
Residual deviance: 108158 on 996 degrees of freedom
AIC: 7531.5
Number of Fisher Scoring iterations: 2

Table 3: Model 2: GLM Normal with a Quadratic term and Identity Link

Variable	Estimate	Std. Error	t-ratio	p-value
INTERCEPT	-18.284918	2.357264	-7.757	$2{,}15 \times 10^{-14***}$
EDUC	2.155801	0.138724	15.540	$<2 \times 10^{-16***}$
EXPER	0.123334	0.025728	4.794	$1.89 \times 10^{-6***}$
HRSWK	0.423466	0.083016	5.101	$4.04 \times 10^{-7***}$
HRSWK^2	-0.002659	0.001047	-2.541	0.0112*

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for Gaussian family taken to be 108.0003)
Null deviance: 146356 on 999 degrees of freedom
Residual deviance: 107460 on 995 degrees of freedom
AIC: 7527
Number of Fisher Scoring iterations: 2

Table 4: Model 3: GAM Normal with Identity Link

Variable	Estimate	Std. Error	t-ratio	p-value
INTERCEPT	-5.70533	1.75403	-3.253	0.001118**
EDUC	2.11854	0.13787	15.367	$<2x10^{-16}$
EXPER	0.12307	0.02546	4.833	$1.56x10^{-6}$

Approximate significance of smooth terms:

Variable	Estimated df	Refined df	F	p-value
s(HRSWK)	4.295	5.284	16.51	$< 2x10^{-16}$**

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
R-sq.(adj) = 0.278 Deviance explained = 28.3%
GCV score = 106.52 Scale est. = 105.75 n = 1000
AIC = 7508.195
Null Deviance: 146355.6 on 999 degrees of freedom
Residual Deviance: 108157.6 on 996 degrees of freedom
Number of Local Scoring Iterations: 2

Nonparametric exploration of nonlinearity

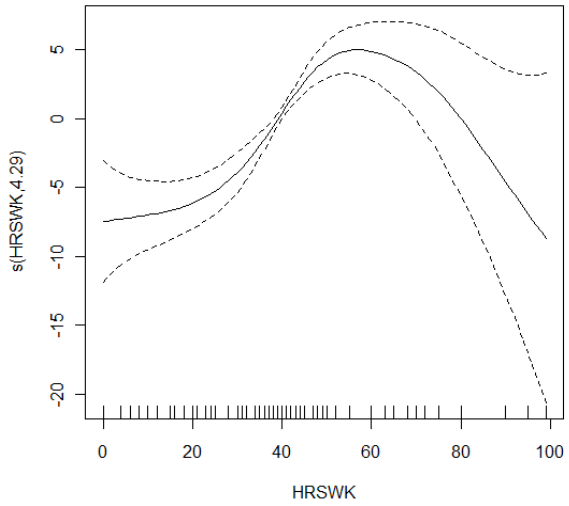


Fig. 1: Partial residuals plot of *HRSWK*

Table 5: Model 4: GLM Normal with Identity Link for ln(Wages)

Variable	Estimate	Std. Error	t-ratio	p-value
INTERCEPT	1.260476	0.104063	12.113	$< 2x10^{-16***}$
EDUC	0.112250	0.007222	15.544	$< 2x10^{-16***}$
EXPER	0.005845	0.001250	4.678	$3.30x10^{-6***}$
HRSWK	0.008831	0.001358	6.503	$1.25x10^{-10***}$

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for Gaussian family taken to be 107.7871)
Null deviance: 146356 on 999 degrees of freedom
Residual deviance: 107356 on 996 degrees of freedom
AIC: 7524.032
Number of Fisher Scoring iterations: 6

Table 6: Model 5: GAM Normal with Identity Link for ln(Wages)

Variable	Estimate	Std. Error	t-ratio	p-value
INTERCEPT	1.624786	0.095485	17.016	$<2x10^{-16***}$
EDUC	0.110449	0.007118	15.518	$<2x10^{-16***}$
EXPER	0.005555	0.001235	4.497	$7.7x10^{-6***}$

Approximate significance of smooth terms:

Variable	Estimated df	Refined df	F	p-value
s(HRSWK)	3.974	4.961	14.20	$2.33x10^{-13}$**

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
R-sq.(adj) = 0.293 Deviance explained = 29.7%
GCV score = 104.37 Scale est. = 103.64 n = 1000
AIC = 7487.733
Null Deviance: 146355.6 on 999 degrees of freedom
Residual Deviance: 102866.4 on 992.9999 degrees of freedom
Number of Local Scoring Iterations: 7

Table 7: Model 6: GAM Normal with an Interaction Identity Link for ln(Wages)

Variable	Estimate	Std. Error	t-ratio	p-value
INTERCEPT	1.649134	0.095606	17.249	$<2x10^{-16***}$
EDUC	0.109249	0.007107	15.373	$<2x10^{-16***}$
EXPER	0.002676	0.001576	1.698	0.08985
*EXPR*MARRIED*	0.003586	0.001172	3.060	0.00228**

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Approximate significance of smooth terms:

Variable	Estimated df	Refined df	F	p-value
s(HRSWK)	3.885	4.859	14.54	$1.83x10^{-13}$**

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
R-sq.(adj) = 0.299 Deviance explained = 30.4%
GCV score = 103.55 Scale est. = 102.74 n = 1000
AIC = 7479.896
Null Deviance: 146355.6 on 999 degrees of freedom
Residual Deviance: 101865.4 on 991.9999 degrees of freedom
Number of Local Scoring Iterations: 7

Table 8: Model 7: GAM Normal with an Interaction term and Identity Link for ln(Wages)

Variable	Estimate	Std. Error	t-ratio	p-value
INTERCEPT	1.638620	0.102779	15.943	$<2x10^{-16***}$
EDUC	0.108653	0.007056	15.399	$<2x10^{-16***}$
EXPER	0.005633	0.001888	2.983	0.00293**

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Approximate significance of smooth terms:

Variable	Estimated df	Refined df	F	p-value
s(HRSWK)	3.928	4.908	13.251	$2.46x10^{-12}$**
*s(EXPER*MARRIED)*	2.571	3.229	6.937	$7.95x10^{-5}$**

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
R-sq.(adj) = 0.308 Deviance explained = 31.4%
GCV score = 102.28 Scale est. = 101.31 n = 1000
AIC = 7467.548
Null Deviance: 146355.6 on 999 degrees of freedom

Residual Deviance: 100082.5 on 989 degrees of freedom
Number of Local Scoring Iterations: 6

4.2 Comparing the Models

Estimation results are presented in tables 2 through 8. A comparison of models using the AIC presented in Table 9 suggests that models 6 and 7, which employ *ln(WAGES)* as the response variable and allow for interaction between *EXPER* and *MARRIED* have the lowest AICs among the models considered and are therefore the best models. Model 7, a generalized additive model for *ln(WAGES)*, which includes a nonparametric interaction term between *EXPER* and *MARRIED* has the lowest AIC and UBRE score among the seven models studied . Model 6, which allows parametric interaction term between *EXPER* and *MARRIED* has the second lowest AIC and UBRE score. At the other extreme, Model 1, a generalized linear Gaussian model for *WAGES*, has the highest AIC suggesting that it is the poorest model among the models considered.

Table 9: Models and the AICs

MODEL	AIC
1	7531.5
2	7527
3	7508.195
4	7524.032
5	7487.733
6	7479.896
7	7467.548

5 The generalized additive Logit model

The generalized additive Logit model assumes that

$$g(\mu_i) = logit(\mu_i) = ln\left(\frac{\mu_i}{1-\mu_i}\right) = \alpha + \sum_{j=1}^{p} s_j(x_{ij}),$$

$$where\ \mu_i = p_i = E(y_i = 1 | x_i) = exp\{\alpha + \sum_{j=1}^{p} s_j(x_{ij})\} / exp\{1 + \alpha + \sum_{j=1}^{p} s_j(x_{ij})\}.$$

The adjusted dependent variable z and the weights w used in the algorithm above are

$$z_i = \eta_i + (y_i - p_i) / p_i(1-p_i),$$
$$w_i = p_i(1-p_i)$$
$$where\ p_i = g^{-1}(\eta_i), \eta_i = \alpha + \sum_{j=1}^{p} s_j(x_{ij})$$

The functions $s_1, s_2, \ldots, s_p$ are estimated by an algorithm like the one described earlier.

5.1 An Empirical Application of GAM Logit Model to data on brand choice of crackers

The dataset is from Jain et al. [10] and Paap and Franses [13]. We use an optical scanner panel data set on purchases of saltine crackers in the Rome (Georgia) market, collected by Information Resources Incorporated. The data set contains information on all 3292 purchases of crackers of 136 households over a period of two years, including brand choice, actual price of the purchased brand and shelf price of other brands, and whether there was a display and/or newspaper feature of the considered brands at the time of purchase. The data file contains 17 variables, arranged in five rows for each observation as follows.
ID: individual identifiers
CHOICE: one of Sunshine, Keebler, Nabisco, Private Label
DISP.z: is there a display for brand z ?
FEAT.z: is there a newspaper feature advertisement for brand z ?
PRICE.z: price of brand z

1: Household ID

2-5: Purchase/no-purchase of Sunshine, Keebler, Nabisco, Private Label
6-9: Display/no-display of Sunshine, Keebler, Nabisco, Private Label
10-13: Feature/no-feature of Sunshine, Keebler, Nabisco, Private Label
14-17: Price in $/unit for Sunshine, Keebler, Nabisco, Private Label

Table 10 shows some data characteristics. There are three major national brands in our database, that is, Sunshine, Keebler and Nabisco with market shares of 7%, 7% and 54%, respectively. The local brands are collected under `Private label', which has a market share of 32%; see the first row of Table 10. `Display' refers to the fraction of purchase occasions that a brand was on display and `feature' refers to the fraction that a brand was featured. The market leader, Nabisco, is relatively often on display (34%) and featured (9%). The `average price' denotes the mean of the price of a brand over the 3292 purchases. The Keebler crackers seem to be the most expensive crackers in our data set. Table 10 provides information of the number of brand switches in the sample. For example, in 39% of the cases households buying Sunshine on the current purchase occasion buy the same brand on the next purchase

Table 10: Some data characteristics of the saltine crackers Sunshine, Keebler, Nabisco, and Private label

BRAND	Sunshine	Keebler	Nabisco	Private Label
MKT. SHARE	0.07	0.07	0.54	0.32
DISPLAY a	0.13	0.11	0.34	0.10
FEATURE b	0.04	0.04	0.09	0.05
AVERAGE PRICE	0.96	1.13	1.08	0.68
ESTIMATED S	0.39	0.07	0.35	0.19
BRAND K	0.09	0.50	0.30	0.11
SWITCHING N	0.04	0.04	0.84	0.08
PROBABILITIES c p	0.04	0.03	0.12	0.81

The dependent variable *CHOICE* is defined as follows.

NABISCO = 1 if Sunshine is chosen,
= 0 if any other brand is chosen
PRICE.SUNSHINE = Price of a box of Sunshine
PRICE.KEEBLER = Price of a box of Sunshine
PRICE.NABISCO = Price of a box of Sunshine
PRICE.PRIVATE = Price of a box of Sunshine
DISP.SUNSHINE = 1 if Sunshine is displayed at time of purchase, otherwise = 0
DISP.KEEBLER = 1 if Keebler is displayed at time of purchase, otherwise = 0
DISP.NABISCO = 1 if Nabisco is displayed at time of purchase, otherwise = 0
DISP.PRIVATE = 1 if Private Label is displayed at time of purchase, otherwise = 0

Models

The following models were fitted to the data. Models 1 and 2 are Logit models. Given the nonlinearity of the Logit link function in *PRATIO* as displayed in the partial residual plots of *PRICE.NABISCO* in Fig.2, Model 3, a Generalized Additive Logit model introduces a nonparametric smooth term *s(PRICE.NABISCO)*. Model 4, another GAM model introduces nonparametric smooth terms *s(PRICE.NABISCO)*, *s(PRICE.SUNSHINE)* and *s(PRICE.KEEBLER)* in addition to parametric terms for *FEAT.NABISCO*, *FEAT.SUNSHINE*, and *FEAT.KEEBLER*, Finally, Model 5 is a GAM Logit model, which contains no parametric terms, but contains nonparametric smooth terms *s(PRICE.NABISCO)*, *s(PRICE.SUNSHINE)*, *s(PRICE.KEEBLER)*, and *s(PRICE.PRIVATE)*.

Model 1: Logit Re gression Model 1

$$\eta = g(\mu) = \beta_1 + \beta_2 PRICE.NABISCO + \beta_3 DISP.NABISCO + \beta_4 FEAT.NABISCO$$

Model 2: Logit Re gression Model 2

$$\eta = g(\mu) = \beta_1 + \beta_2 PRICE.KEEBLER + \beta_3 DISP.KEEBLER + \beta_4 FEAT.KEEBLER$$

Model 3:Generalized Additive Logit Re gression Models with a Nonparametric Interaction Term

$$\eta = g(\mu) = \beta_1 + s(PRICE.NABISCO) + \beta_3 DISP.NABISCO + \beta_4 FEAT.NABISCO$$

Model 4: Generalized Additive Logit Re gression Models with a Nonparametric Interaction Term and a dummy var iable for whether a particular brand is featured at the time of purchase

$$\eta = g(\mu) = \beta_1 + \beta_2 FEAT.NABISCO + \beta_3 FEAT.SUNSHINE + \beta_4 FEAT.KEEBLER + s(PRICE.NABISCO) + s(PRICE.SUNSHINE) + s(PRICE.KEEBLER)$$

Model 5: Logit Re gression Model with a Nonparametric Smooth for the Pr ice of each Brand

$$\eta = g(\mu) = \beta_1 + s(PRICE.NABISCO) + s(PRICE.SUNSHINE) + s(PRICE.KEEBLER) + s(PRICE.PRIVATE)$$

Table 11: Model 1: GLM Logit with Nabisco characteristics as predictors: Cracker Data

Variable	Coefficient Estimate	Std. Error	t-ratio	p-value
Intercept	2.941133	0.290899	10.111	$<2x10^{-16***}$
PRICE.NABISCO	-0.026398	0.002626	-10.051	$<2x10^{-16***}$
DISP.NABISCO	0.138151	0.076595	1.804	0.0713
FEAT.NABISCO	0.588538	0.140506	4.189	$2.81x10^{-5***}$

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for binomial family taken to be 1)
Null deviance: 4537.7 on 3291 degrees of freedom
Residual deviance: 4385.8 on 3288 degrees of freedom
AIC: 4393.8
Number of Fisher Scoring iterations: 4

Table 12: Model 2: GLM Logit with Keebler characteristics as predictors

Variable	Coefficient Estimate	Std. Error	t-ratio	p-value
Intercept	-0.724835	0.383416	-1.890	0.0587
PRICE.KEEBLER	0.007856	0.003369	2.332	0.0197*
DISP.KEEBLER	0.187321	0.122768	1.526	0.1271
FEAT.KEEBLER	-0.031748	0.186839	-0.170	0.8651

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for binomial family taken to be 1)
Null deviance: 4537.7 on 3291 degrees of freedom
Residual deviance: 4530.7 on 3288 degrees of freedom
AIC: 4538.7
Number of Fisher Scoring iterations: 3

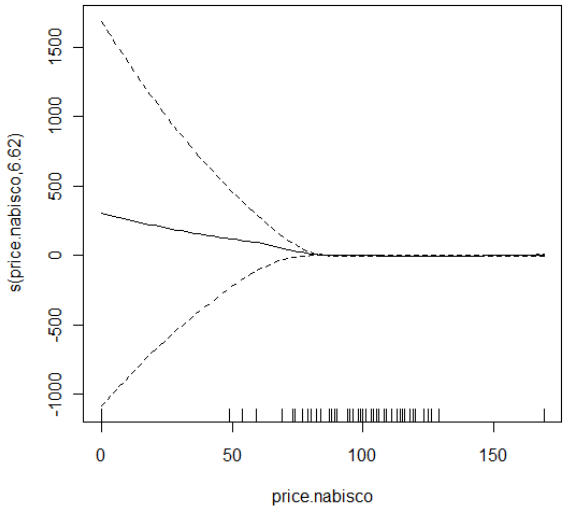


Fig. 2: Partial residuals plot of PRICE.NABISCO

Table 13: Model 3: GAM Logit with a nonparametric smooth term for PRICE.NABISCO

Variable	Coefficient Estimate	Std. Error	t-ratio	p-value
Intercept	1.20856	1.27053	0.951	0.3415
DISP.NABISCO	0.15956	0.07876	2.026	0.0428*
FEAT.NABISCO	0.63135	0.14305	4.414	$1.02x10^{-5***}$

Approximate significance of smooth terms:

Variable	Estimated df	Refined df	Chi-squared	p-value
s(PRICE.NABISCO)	6.616	6.861	48.02	$3.04x10^{-8}$***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
R-sq.(adj) = 0.0499 Deviance explained = 4.53%
UBRE score = 0.32181 Scale est. = 1 n = 3292
AIC = 4351.397
Null Deviance: 4537.747 on 3291 degrees of freedom
Residual Deviance: 4352.563 on 3285 degrees of freedom
Number of Local Scoring Iterations: 9

Table 14: Model 4: A GAM Logit Model with a nonparametric smooth term for each brand price and a dummy variable for whether a brand is featured at the time of purchase

Variable	Coefficient Estimate	Std. Error	t-ratio	p-value
Intercept	67.59466	46.81464	1.444	0.1488
FEAT.SUNSHINE	0.02945	0.21710	0.136	0.8921
FEAT.NABISCO	0.30313	0.14966	2.025	0.0428*
FEAT.KEEBLER	0.34561	0.19080	1.811	0.0701

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:

Variable	Estimated df	Refined df	Chi-squared	p-value
s(PRICE.SUNSHINE)	7.159	8.147	17.98	0.0231*
s(PRICE.NABISCO)	7.575	7.880	64.17	$6.15X10^{-11***}$
s(PRICE.KEEBLER)	8.819	8.982	121.41	$<2x10^{-16}$

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) = 0.0959 Deviance explained = 8.43%
UBRE score = 0.279 Scale est. = 1 n = 3292
AIC = 4210.49
Null Deviance: 4537.747 on 3291 degrees of freedom
Residual Deviance: 4265.343 on 3276 degrees of freedom
Number of Local Scoring Iterations: 9

Table 15: Model 5: A GAM Logit Model with a nonparametric smooth term for each brand price

Variable	Estimated df	Refined df	Chi-squared	p-value
s(PRICE.SUNSHIN)	7.767	8.585	72.91	$2.56x10^{-12}$*
s(PRICE.NABISCO)	7.474	7.763	88.42	$7.18x10^{-16}$***
s(PRICE.KEEBLER)	8.660	8.940	95.01	$<2x10^{-16}$***
s(PRICE.PRIVATE)	8.977	9.000	386.09	$<2x10^{-16}$***

Parametric Coefficients

Variable	Coefficient Estimate	Std. Error	t-ratio	p-value
Intercept	88.91	45.06	1.973	0.0485*

Approximate significance of smooth terms:
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
R-sq.(adj) = 0.222 Deviance explained = 18.7%
UBRE score = 0.14172 Scale est. = 1 n = 3292
AIC: 3758.557
Null Deviance: 4537.747 on 3291 degrees of freedom
Residual Deviance: 3945.672 on 3275 degrees of freedom
Number of Local Scoring Iterations: 9

5.2 Comparing the Models

Estimation results are presented in tables 11 through 15. A comparison of models using the AIC is presented in Table 16.

Table 16: Models and the AICs

MODEL	AIC
1	4393.78
2	4538.65
3	4351.397
4	4210.49
5	3758.557

Model 5, a GAM, which includes a nonparametric smooth term for the price of each brand, has the lowest AIC and UBRE score among the five models studied and is therefore the best model. The Logit regression model (Model 2), which uses the characteristics of the competing brand, Keebler has the highest AIC. This is not surprising since the Logit model misses the nonlinearity in the price of each brand in the link function. Model 5 also has the lowest deviance of 3945.672 on 3275 degrees of freedom, while Model 2 has the highest deviance of 4530.7 on 3288 degrees of freedom. The statistical significance of brand prices differs markedly between the Logit regression model and the various GAM Logit models employed here. The signs of the coefficient estimates are all expected in all of the models but Model 2. For instance, the sign of *DISP.KEEBLER* is positive in model 2 indicating that the odds of choosing *NABISCO* over *KEEBLER* increase if *KEEBLER* is displayed at the time of purchase and decrease if *NABISCO* is displayed at the time of purchase. The analysis of deviance in Tables 13, 14, and 15 indicates significant nonlinear contribution from the variables *PRICE.NABISCO* as well as other brand price variables. The high degree of nonlinearity in *PRICE.NABISCO* is also seen in the partial residual smoothing plot of *PRICE.NABISCO* in Fig. 2. The dotted curves around the solid curve represent +-2 standard errors around the solid curve. The only surprising result is the negative sign of the variable *FEAT.KEEBLER* in Table 12.

6 The generalized additive Poisson and Negative Binomial Models

Generalized additive models can be used in virtually any setting where linear models are used. The basic idea is to replace $\sum x_{ij}\beta_j$, the linear component of the model with an additive component $\sum f_j(x_{ij})$.
In the Poisson regression model the outcome, y_i is a count variable, such as number of visits to a doctor's office as in Gurmu [5]. We wish to model $p(y_i|x_{i1}, x_{i2,\ldots}, x_{ip})$ the probability of an event given factors $x_{i1}, x_{i2,\ldots}, x_{ip}$. The Poisson regression model assumes that the link function is linear:

$$ln\,\mu_i = \beta_0 + x_{i1}\beta_1 + \ldots + x_{ip}\beta p$$

The generalized additive Poisson model assumes instead that

$$ln\,\mu_i = \beta_0 + s_1(x_{i1}) + \ldots + s_p(x_{ip})$$

The functions $s_1, s_{2,\ldots,}s_p$ are estimated by maximizing a penalized log likelihood or a penalized log partial-likelihood using the PIRLS algorithm described above.

6.1 An Empirical Application of GAM Poisson and GAM Negative Binomial Models to data on doctor visits

Data and variable definitions

The data are a 1986 cross section sample from the US consisting of 485 observations and are drawn from Gurmu [5]. These data came from the 1986 Medicaid Consumer Survey sponsored by the Health Care Financing Administration. The following variables were used in econometric analysis.
DOCTOR = the number of doctor visits
CHILDREN = the number of children in the household
ACCESS = is a measure of access to health care
HEALTH = a measure of health status (larger positive numbers are associated with poorer health

Table 17: Summary Statistics for the Doctor data

Variable	Obs.	Mean	Std. Dev.	Min	Max
DOCTOR	485	1.610	3.346809	0	48
CHILDREN	485	2.264	1.319136	1	9
ACCESS	485	0.3812	0.186105	0	0.92
HEALTH	485	-0.00004124	1.433520	1	1

A likelihood ratio test of H_0: Poisson against H_1: NB model yielded a Chi-Square Test Statistic = 599.6065 and p-value = < 2.2e-16 and overdispersion was confirmed. Accordingly, both Poisson and Negative Binomial GLM and GAM were fitted to the data.

Models

The following models were fitted to the data. Models 1P and 1NB are Poisson and Negative Binomial regression models, which use CHILDREN, ACCESS and HEALTH as independent variables. Given the nonlinearity of the Logit link function in *HEALTH* as displayed in the partial residual plots of *HEALTH* in Fig. 3, Models 2P and 2NB are Generalized Additive Poisson and Negative Binomial models respectively, which introduce a nonparametric smooth term *s(HEALTH)*. Models 3P and 3NB are Generalized Additive Poisson and Negative Binomial models respectively, which introduce nonparametric smooth terms *s(ACCESS)* and *s(HEALTH)*. Models 4P and 4NB are Generalized Additive Poisson and Negative Binomial models, which introduce a nonparametric smooth interaction term *s(ACCESS*HEALTH)* in addition to nonparametric smooth terms *s(ACCESS)* and *s(HEALTH)*. Models 5P and 5NB replace the nonparametric interaction *s(ACCESS*HEALTH)* with another nonparametric interaction *s(ACCESS*CHILDREN).* Additional GAM models with more general nonparametric interaction terms *s(ACCESS,HEALTH)* and *s(ACCESS,CHILDREN)* were also estimated but did not result in improvements and are not reported here.

Model 1P: Poisson Re gression Model with Identity Link

$\eta = g(\mu) = \beta_1 + \beta_2 CHILDREN + \beta_3 ACCESS + \beta_4 HEALTH$

Model 1NB: Negative Binomial Re gression with Identity Link

$\eta = g(\mu) = \beta_1 + \beta_2 CHILDREN + \beta_3 ACCESS + \beta_4 HEALTH$

Model 2P:Generalized Additive Poisson Re gression Model

$\eta = g(\mu) = \beta_1 + \beta_2 CHILDREN + \beta_3 ACCESS + s(HEALTH)$

Model 2NB: Generalized Additive Negative Binomial Re gression Model

$\eta = g(\mu) = \beta_1 + \beta_2 CHILDREN + \beta_3 ACCESS + s(HEALTH)$

Model 3P: Generalized Additive Poisson Re gression Model with smooth nonparametric terms for access and health

$\eta = g(\mu) = \beta_1 + \beta_2 CHILDREN + s(ACCESS) + s(HEALTH)$

Model 3NB: Generalized Additive Negative Binomial Re gression Model with smooth nonparametric terms for access and health

$\eta = g(\mu) = \beta_1 + \beta_2 CHILDREN + s(ACCESS) + s(HEALTH)$

Model 4P: Generalized Additive Poisson Re gression Model with smooth nonparametric terms for access and health and nonparametric int eraction between ACCESS and HEALTH

$\eta = g(\mu) = \beta_1 + \beta_2 CHILDREN + s(ACCESS) + s(HEALTH) + s(ACCESS * HEALTH)$

Model 4NB: Generalized Additive Negative Binomial Re gression Model with smooth nonparametric terms for access and health and nonparametric int eraction between ACCESS and HEALTH

$\eta = g(\mu) = \beta_1 + \beta_2 CHILDREN + s(ACCESS) + s(HEALTH) + s(ACCESS * HEALTH)$

Model 5P: Generalized Additive Poisson Re gression Model with smooth nonparametric terms for access and health and nonparametric int eraction between ACCESS and CHILDREN

$\eta = g(\mu) = \beta_1 + \beta_2 CHILDREN + s(ACCESS) + s(HEALTH) + s(ACCESS * CHILDREN)$

Model 5NB: Generalized Additive Negative Binomial Re gression Model with smooth nonparametric terms for access and health and nonparametric int eraction between ACCESS and CHILDREN

$\eta = g(\mu) = \beta_1 + \beta_2 CHILDREN + s(ACCESS) + s(HEALTH) + s(ACCESS * CHILDREN)$

Table 18: Model 1P: Poisson Regression Model

Variable	Coefficient Estimate	Std. Error	t-ratio	p-value
INTERCEPT	0.37509	0.11016	3.405	0.000662***
CHILDREN	-0.17592	0.03164	-5.560	2.70×10^{-8}
ACCESS	0.93694	0.19278	4.860	1.17×10^{-6}
HEALTH	0.28984	0.01825	15.880	$<2 \times 10^{-16}$

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for poisson family taken to be 1)
Null deviance: 1766.2 on 484 degrees of freedom
Residual deviance: 1508.8 on 481 degrees of freedom
AIC: 2179.5
Number of Fisher Scoring iterations: 6

Table 19: Model 1NB: Negative Binomial Regression Model

Variable	Coefficient Estimate	Std. Error	t-ratio	p-value
INTERCEPT	0.48207	0.24005	2.008	0.0452*
CHILDREN	-0.16573	0.06744	-2.457	0.0144*
ACCESS	0.60193	0.43715	1.377	0.1692
HEALTH	0.30344	0.04990	6.081	$2.43 \times 10^{-9***}$

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Negative Binomial(2) family taken to be 2.653272)
Null deviance: 940.44 on 484 degrees of freedom
Residual deviance: 813.49 on 481 degrees of freedom
AIC: 1692.6
Number of Fisher Scoring iterations: 5

Table 20: Model 2P: Generalized Additive Poisson Regression Model

Variable	Coefficient Estimate	Std. Error	t-ratio	p-value
INTERCEPT	0.3182	0.1140	2.792	0.00523**
CHILDREN	-0.1772	0.0319	-5.555	$2.77x10^{-8***}$
ACCESS	0.9933	0.1990	4.991	$6.02x10^{-7***}$

Family: Poisson
Link function: log
Approximate significance of smooth terms:

Variable	Estimated df	Refined df	Ch-squared	p-value
s(HEALTH)	8.62	8.952	300.5	$<2x10^{-16***}$

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
R-sq.(adj) = 0.142 Deviance explained = 17.6%
UBRE score = 2.0474 Scale est. = 1 n = 485
AIC: 2140.685
Null Deviance: 1766.246 on 484 degrees of freedom
Residual Deviance: 1443.896 on 475.0002 degrees of freedom

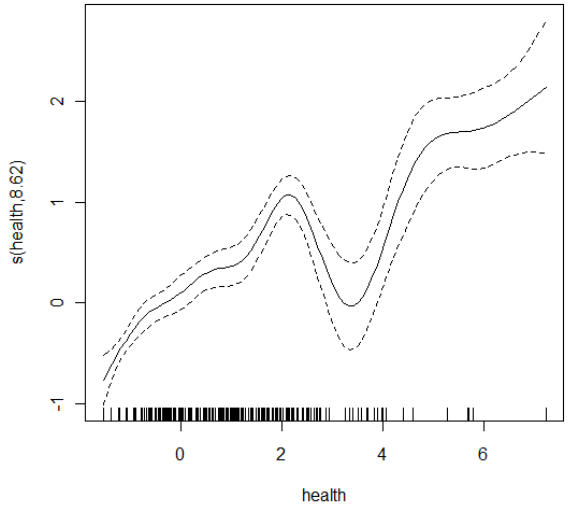


Fig. 3: Partial Residual plot of HEALTH

Table 21: Model 2NP: GAM Negative Binomial (2) Regression Model

Variable	Coefficient Estimate	Std. Error	t-ratio	p-value
INTERCEPT	0.50987	0.14996	3.400	0.000674***
CHILDREN	-0.17116	0.04199	-4.076	$4.58x10^{-5***}$
ACCESS	0.48808	0.27428	1.779	0.075162

Family: Negative Binomial (2)
Link function: log

Approximate significance of smooth terms:

Variable	Estimated df	Refined df	Ch-squared	p-value
s(HEALTH)	8.268	8.836	105.9	$<2x10^{-16***}$

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
R-sq.(adj) = 0.115 Deviance explained = 15.7%
UBRE score = 0.68084 Scale est. = 1 n = 485
AIC: 1686.335
Null Deviance: 1766.246 on 484 degrees of freedom
Residual Deviance: 1478.552 on 477.9999 degrees of freedom

Table 22: Model 3P: GAM Poisson Regression Model with nonparametric smooth terms for ACCESS and HEALTH

Variable	Coefficient Estimate	Std. Error	t-ratio	p-value
INTERCEPT	0.62717	0.07672	8.175	$2.97\ x10^{-16***}$
CHILDREN	-0.16883	0.03165	-5.334	$9.63x10^{-8***}$

Approximate significance of smooth terms:

Variable	Estimated df	Refined df	Ch-squared	p-value
s(ACCESS)	8.62	8.962	300.5	$<2x10^{-16***}$
S(HEALTH)	8.437	8.902	238.4	$<2x10^{-16***}$

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
R-sq.(adj) = 0.175 Deviance explained = 22.7%
UBRE score = 1.8955 Scale est. = 1 n = 485
AIC: 2067.047 Family: Poisson
Link function: log
Null Deviance: 1766.246 on 484 degrees of freedom
Residual Deviance: 1443.896 on 475.0002 degrees of freedom

Table 23: Model 3NB: GAM Negative Binomial Regression Model with nonparametric smooth terms for ACCESS and HEALTH

Variable	Coefficient Estimate	Std. Error	t-ratio	p-value
INTERCEPT	0.62302	0.10382	6.001	$1.96\ x10^{-9***}$
CHILDREN	-0.16414	0.04218	-3.891	$9.97x10^{-5***}$

Approximate significance of smooth terms:

Variable	Estimated df	Refined df	Ch-squared	p-value
s(ACCESS)	8.861	8.991	48.14	$2.39x10^{-7***}$
S(HEALTH)	8.041	8.734	94.46	$<2x10^{-16***}$

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
R-sq.(adj) = 0.137 Deviance explained = 20.8%
UBRE score = 0.61309 Scale est. = 1 n = 485
AIC: 1653.477

Table 24: Model 4P: GAM Poisson Regression Model

Variable	Coefficient Estimate	Std. Error	t-ratio	p-value
INTERCEPT	0.60408	0.07785	7.759	$8.54\ x10^{-15***}$
CHILDREN	-0.16390	0.03206	-5.113	$3{,}17x10^{-7***}$

Approximate significance of smooth terms:

Variable	Estimated df	Refined df	Ch-squared	p-value
s(ACCESS)	8.790	8.978	79.14	$2.34x10^{-13***}$
S(HEALTH)	8.099	8.730	33.51	$8.86x10^{-5***}$
*s(ACCESS*HEALTH)*	5.427	6.630	36.31	$4.47x10^{-6***}$

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
R-sq.(adj) = 0.232 Deviance explained = 25.6%

UBRE score = 1.8088 Scale est. = 1 n = 485
AIC: 2024.952
Null Deviance: 1766.246 on 484 degrees of freedom
Residual Deviance: 1380.621 on 471.0001 degrees of freedom

Table 25: Model 4NB: GAM Negative Binomial Regression Model with Nonparametric Interaction

Variable	Coefficient Estimate	Std. Error	t-ratio	p-value
INTERCEPT	0.60298	0.10415	5.790	7.05×10^{-9}***
CHILDREN	-0.15887	0.04227	-3.759	0.000171***

Approximate significance of smooth terms:

Variable	Estimated df	Refined df	Ch-squared	p-value
s(ACCESS)	8.717	8.966	50.110	9.95×10^{-8}***
S(HEALTH)	1.263	1.467	1.555	0.3289
*s(ACCESS*HEALTH)*	4.465	5.548	18.901	0.0031**

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
R-sq.(adj) = 0.15 Deviance explained = 21.6%
UBRE score = 0.58768 Scale est. = 1 n = 485
AIC: 1641.151

Table 26: Model 5P: GAM Poisson Regression Model with Nonparametric interaction

Variable	Coefficient Estimate	Std. Error	t-ratio	p-value
INTERCEPT	0.10745	0.18992	0.566	0.572
CHILDREN	0.05892	0.08160	0.722	0.470

Approximate significance of smooth terms:

Variable	Estimated df	Refined df	Ch-squared	p-value
s(ACCESS)	8.921	8.997	116.010	$<2 \times 10^{-16}$***
S(HEALTH)	8.568	8.940	243.771	$<2 \times 10^{-16}$***
*s(ACCESS*CHILDREN)*	1.341	1.625	8.819	0.00772**

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
R-sq.(adj) = 0.188 Deviance explained = 23.2%
UBRE score = 1.8812 Scale est. = 1 n = 485
AIC: 2060.085
Null Deviance: 1766.246 on 484 degrees of freedom
Residual Deviance: 1426.358 on 471 degrees of freedom

Table 27: Model 5NB: GAM Negative Binomial Regression Model with Nonparametric Interaction

Variable	Coefficient Estimate	Std. Error	t-ratio	p-value
INTERCEPT	0.33961	0.24609	1.380	0.168
CHILDREN	-0.03982	0.10631	-0.375	0.708

Approximate significance of smooth terms:

Variable	Estimated df	Refined df	Ch-squared	p-value
s(ACCESS)	8.862	8.992	49.488	1.33×10^{-7}***
S(HEALTH)	8.175	8.797	94.602	$<2 \times 10^{-16}$***
*s(ACCESS*CHILDREN)*	1.000	1.001	1.527	0.217

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
R-sq.(adj) = 0.144 Deviance explained = 21%
UBRE score = 0.61398 Scale est. = 1 n = 485
AIC: 1653.908

6.2 Comparing the models

Estimation results are presented in tables 18 through 27. A comparison of models using the AIC is presented in Tables 28 and 29. Table 28 compares GLM and GAM Poisson models and Table 29 compares GLM and GAM Negative Binomial models. Model 4P, a GAM Poisson model, which includes nonparametric smooth terms for *ACCESS*, *HEALTH* and the interaction *ACCESS*HEALTH* has the lowest AIC and UBRE score among the five models studied and is therefore the best model. The GLM Poisson regression model (Model 1P), which uses the predictors *CHILDREN*, *ACCESS* and *HEALTH* has the highest AIC. This is not surprising since the Poisson model misses the nonlinearity in the predictors *CHILDREN*, *ACCESS*, *HEALTH*, and interaction *ACCESS*HEALTH*. Similarly, Model 4NB, a GAM Negative Binomial model, which includes nonparametric smooth terms for *ACCESS*, *HEALTH* and the interaction *ACCESS*HEALTH* has the lowest AIC and UBRE score among the five models studied and is therefore the best model among the Negative Binomial GLM and GAM models. Finally, among the Poisson and Negative Binomial GLM and GAM models studied here, the clear winner is Model 4NB, which has the lowest AIC and UBRE scores. Model 4P also has the lowest deviance of 1380.621 on 471.0001 degrees of freedom among the Poisson GLM and GAM models, while Model 1P has the highest deviance of 1508.8 on 481 degrees of freedom. The statistical significance of the predictors *CHILDREN*, *ACCESS* and *HEALTH* are generally similar among all models since all of the variables are found highly significant in all of the models. The signs of the coefficient estimates are all expected in all of the models but Model 2. For instance, the sign of *ACCESS* is positive as is the sign of *HEALTH*, but the sign of *CHILDREN* is negative indicating that as access to health services increases, individuals make more trips to a doctor's office even for routine checkups. Individuals with poor health with higher numbers on *HEALTH* tend to make more trips to a doctor's office as reflected in the positive sign of *HEALTH*. The negative sign of the variable *CHILDREN* may be surprising, but may indicate that households with fewer children make more trips to a doctor's office perhaps because these households in the sample are more health-conscious than households with too many children in this sample. The analysis of deviance in Tables 21 through 27 indicates significant nonlinear contribution from the variables *ACCESS*, *HEALTH*, and interaction *ACCESS*HEALTH*. The high degree of nonlinearity in *HEALTH* is also seen in the partial residual smoothing plot of *HEALTH* in Figure 3. The dotted curves around the solid curve represent +-2 standard errors around the solid curve.

Table 28: GLM and GAM Poisson Models and the AICs

MODEL	AIC
1P	2179.487
2P	2140,685
3P	2067.047
4P	2024.952
5P	2060.085

Table 29: GLM and GAM Negative Binomial Models and the AICs

MODEL	AIC
1NB	1692.614
2NB	1686.335
3NB	1653.477
4NB	1641.151
5NB	1653.908

7 Conclusion

The paper has presented econometric applications of generalized additive Gaussian, Logit, Poisson, and Negative Binomial regression models as alternatives to the generalized linear Normal, Logit, Poisson, and Negative Binomial regression models respectively. The Gaussian generalized additive models (GAMs) were applied to data on wages, education, and experience. The Logit GAMS were applied to data on brand choice. The Poisson and Negative Binomial GAMs were applied to data on doctor visits. In all of the empirical applications, each of the GAMs provides a much better fit than the corresponding generalized linear model (GLM) as reflected in lower AICs and lower deviances. The econometric techniques used in the paper are widely applicable to the analysis of count, binary response and duration types of data occurring frequently in economics and business. GAMs extend nonparametric regression to more than one predictor helping to circumvent the curse of dimensionality. Another advantage of the GAM approach used in the paper is that we take account of nonlinearities in predictors and interactions among predictors non-parametrically. Nevertheless, the GAMs are not without drawbacks. The computational algorithms are complex and interpretations can

be difficult. These models are useful mainly when simple models for the linear predictor provide an inadequate fit for the data.

References

[1] Baltas, G., Determinants of store brand choice: a behavioral analysis. Journal of Product & Brand Management, Vol. 6, No.5, (1997), pp. 315 – 324.

[2] Cameron, A. C. and P. Trivedi, Microeconometrics. Cambridge University Press, New York, (1998).

[3] Greene, W. H., Econometric Analysis. Pearson/Prentice Hall, New York, (2008).

[4] Guadagni, P. M. and J. D. Little, A Logit Model of Brand Choice Calibrated on Scanner Data, Marketing Science, Vol. 2, No. 3, (1983), pp. 203-238.

[5] Gurmu, S., Semiparametric estimation of hurdle regression models with an application to medicaid utilization. Journal of Applied Econometrics, Vol. 12, No. 3, (1997), pp. 225-242.

[6] Hastie, T. & Tibshirani, R., Generalized Additive Models, Statistical Science, Vol. 1, No. 3, (1986), pp. 297-318.

[7] Hastie, T. & Tibshirani, R., Generalized Additive Models: Some Applications, Journal of the American Statistical Association, Vol. 82, No. 398, (1987), pp. 371-386.

[8] Hastie, T. & Tibshirani, R., Generalized Additive Models. Chapman and Hall, London, (1990).

[9] Hill, R. C., W. E. Griffiths, and G. C. Lim, Principles of Econometrics, Wiley, New Jersey, (2012).

[10] Jain, D. C., N. J. Vilcassim and P. K. Chintagunta, A random–coefficients logit brand–choice model applied to panel data", Journal of Business and Economics Statistics, Vol. 12, No. 3. (1994), pp. 317-328.

[11] McCullagh, P. and J. Nelder, Generalized Linear Models, Chapman and Hall, London (1989).

[12] Manski, C. and D. McFadden, Structural Analysis of Discrete Data and Econometric Applications, MIT Press, Cambridge, (1981).

[13] Paap, R. and P. H Frances, A dynamic multinomial probit model for brand choices with different short–run effects of marketing mix variables", Journal of Applied Econometrics, Vol. 15, No. 6, (2000), pp. 717–744.

[14] Wood, S. N., R-package mgcv, (2012).

[15] Wood, S.N., Generalized Additive Models: an introduction with R, CRC, Boca Raton, (2006).