International Journal of Advanced Statistics and Probability, 2 (2) (2014) 84-101
©Science Publishing Corporation
www.sciencepubco.com/index.php/IJASP
doi: 10.14419/ijasp.v2i2.3624
Research Paper

Semi-parametric mixed effects models for longitudinal data with applications in business and economics

Sunil K. Sapra

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

Abstract

Longitudinal data is becoming increasingly common in business, social sciences, and biological sciences due to the advantages it offers over cross-section data in modeling and incorporating heterogeneity among subjects and in being able to make causal inferences from observational data. Parametric models and methods are widely used for analyzing longitudinal data for continuous, discrete, and count data occurring in these disciplines. Some popular models are Gaussian, Logit, and Poisson fixed and random effects models. These models are unreliable in situations in which the link function is nonlinear and the form of nonlinearity is not known with certainty. This paper employs a semi-parametric extension of fixed and random effects models called generalized additive mixed models (GAMMs) to analyze several longitudinal data sets. These semi-parametric models are flexible and robust extensions of generalized linear models. Following Wood [19], the GAMMs are represented using penalized regression splines and estimated by penalized regression methods treating the penalized component of each smooth as a random effect term and the unpenalized component as a fixed effect term. The degree of smoothness for the unknown functions in the linear predictor part of the GAMM is estimated as the variance parameter of the term. Applications of GAMMs studied include analysis of anti-social behavior, decision to use a professional tax preparer, and analysis of patent data on manufacturing firms. For each application, several GAMMs are compared with their parametric counterparts.

Keywords: *Generalized Additive Mixed Models (GAMMS), Generalized Linear Mixed Models (GLMMS), Logit Models, Poisson Regression Models, Penalized Regression Splines.*

1. Introduction

Linear regression model is the workhorse of empirical research across many disciplines. Generalized linear models (GLMs) extend the linear regression model by allowing for response variables, which are bounded or discrete. These models are used for modeling continuous, categorical, count, and ordinal data on the response variable. These models 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. Gaussian, Logit, and Poisson regression models are among the most widely used GLMs. Common applications of Logit models include analysis of brand choice data in marketing (Baltas [2] and Guadagni and Little [7]) and transportation choice data in economics (Greene [6] and Manski and McFadden [12]). Applications of Poisson regression models include analysis of data on patents, number of trips to a doctor's office, and number of shipping accidents. McCullagh and Nelder [11] provide an authoritative account of GLMs and Cameron and Trivedi [4] and Greene [6] provide econometric applications. These models are appropriate for cross-section data and do not account for heterogeneity among subjects. In order to account for heterogeneity among subjects, longitudinal data is used for which the generalized linear mixed model (GLMM) extension of GLMs is needed. In GLMMs, some of the unknown coefficients in the model linear predictor are treated as random variables. These random effects are viewed as having a covariance structure that itself depends on some unknown fixed parameters. This allows the use of more complex models for the random component of data, which leads to improvements in modeling over-dispersed and correlated data. Generalized additive models (GAM) developed by Hastie and Tibshirani [8], [9] and extended by Wood [19] among others, are a powerful semi-parametric generalization of GLMs in which part of the linear predictor is a sum of unknown smooth functions of explanatory variables. GAMs are very flexible and are very useful in the nonparametric exploration of continuous, discrete, and count data. Sapra [14] presented several applications of these models to cross-section data in business and economics and demonstrated that GAMs generally provided a better fit to data than GLMs.

This paper presents econometric applications of the generalized additive mixed models (GAMM) extensions of the generalized linear mixed models (GLMMs) for longitudinal data, which includes the conventional random effects models and demonstrates that the GAMMs can overcome a serious weakness of the GLMMs: failing to identify the nonlinearities in the link function. The paper is organized as follows. To begin with, we introduce the generalized additive mixed model (GAMM) in section 2 and present the penalized regression method for the estimation of GAMMs in section 3. In the following sections, several econometric applications of GAMM are presented. These applications include a Gaussian GAMM for analysis of anti-social behavior among children in section 4, a GAMM Logit model for analysis of data on choice of a paid tax-preparer in section 5, and a GAMM Poisson regression model for analysis of patent data for manufacturing firms in section 6.

2. Generalized additive mixed models

GAMMs extend generalized additive models (GAMs) by including random effects to allow for heterogeneity and correlation among subjects. Generalized additive models (GAMs) are nonparametric generalized linear models. GAMs extend traditional linear models in another way, namely by allowing for a link between the nonlinear predictor f(x_1... 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, these models may not be adequate for modeling discrete responses such as counts, or bounded responses such as proportions.

Generalized linear models (GLMs) 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. The GLM models (1) can be extended to generalized linear mixed models (GLMMs) by incorporating random effects into the GLMs. Suppose that observations of the ith of n units consist of a response variable y_i and p covariates $x_i = (1, x_{1i},, . . ., x_{pi})^T$ associated with fixed effects and a q x 1 vector of covariates z_i associated with random effects. Let y_{it} denote the response of the ith subject at time t and let x_{1it}, x_{2it}… x_{pit} be the associated covariates. Given a q x 1 vector u of random effects, the observations y_{it} on the ith unit at time t are assumed to be conditionally independent with means E $(y_{it}|u_i) = \mu_{it}$ and variances $\mathrm{Var}(y_{it}|u_i) = \phi m_{it}^{-1} v(\mu_{it})$, where v(.) is a specified variance function, m_{it} is a prior weight (e.g. a binomial denominator) and ϕ is a scale parameter, and follow a generalized additive model. Under the GLMMs, the mean μ_{it} = E (y_{it} | x_{1it}, x_{2it}, …, x_{pit}, u_i) is linked to the linear predictor $x_{it}^T\beta + z_{it}^T u_i$, through the link function

$$\eta_{it} = g(\mu_{it}) = x_{it}^T \beta + z_{it}^T u_i \,. \tag{2}$$

The generalized additive mixed models (GAMMs) extend the GLMMs by linking the mean μ = E(y | x_1, x_2, …, x_p) to the nonlinear nonparametric predictor through the link function

$$\eta_{it} = g(\mu_{it}) = \sum_{j=1}^{p} s_j\,(x_{jit}) + z_{it}^T u_i\,, \tag{3}$$

where s_1 (·)... s_p (·) are smooth nonparametric functions. The most commonly used link function is the canonical link, for which η = θ. The random effects u are assumed to be distributed iid as N (0, D(ψ)), where ψ is a cx1 vector of variance components.

3. Estimation of GAMMs

Estimation of GAMMs consists in representing the GAMM as a GLMM with a variance component controlling the amount of smoothing for each additive component. The Bayesian model of spline smoothing introduced by Wahba [16] and Silverman [15] has led to the possibility of estimating the degree of smoothness of terms in a generalized additive model as variances of the wiggly components of the smooth terms treated as random effects. Several algorithms for GAMM estimation have exploited this connection (see Wang [17] and Ruppert et al. [13]). In the normal errors identity link case, estimation can be performed using general linear mixed effects modeling software such as the lme package in R. In the generalized case, only approximate inference is possible using the Penalized Quasi-Likelihood approach of Breslow and Clayton [3]. An advantage of this approach is that it allows correlated errors to be treated via random effects or the correlation structures available. However, using correlation structures beyond the strictly additive form requires using a GEE approach to fitting.

Some details of how GAMs are represented as mixed models and estimated using maximum likelihood or penalized quasi-likelihood methods can be found in Wood [18], [19]. In addition, these methods obtain a posterior covariance matrix for the parameters of all the fixed effects and the smooth terms. A similar approach due to Lin & Zhang [10] obtains the covariance matrix of the data in the additive case (or pseudo-data in the generalized case) implied by the weights, correlation and random effects structure based on the estimates of the parameters of these terms, which is used to obtain the posterior covariance matrix of the fixed and smooth effects. The bases used to represent smooth terms in GAMMs are the same as those used in GAMs. The normal GAMMs can be described by the conditional density of responses given the random effects

$$f\ (y|u) = \exp\{y^T(X\beta + Zu) - 1^Tb(X\beta + Zu) + 1^Tc(y)\}, \quad (4)$$

And the probability density function of the random effects

$$f\ (u) = (2\pi)^{-q/2}|D\ (\psi)|^{-1/2}\exp\ (-1/2\ u^T\ D\ (\psi)^{-1}u^T). \quad (5)$$

Following Ruppert et al [13], we assume that (4) represents a generalized semi-parametric additive model with D1+D2 predictors of which the first D1 predictors form the columns of X and enter the model linearly and the last D2 predictors, which form the columns of Z enter the model non-parametrically as p-th degree splines. For each of the last D2 predictors, the powers of degree 1 through p are columns of X, while the truncated power functions form columns of Z. Under these conditions, the GAMM in (4) and (5) is represented as a GLMM and the methods for estimation of GLMMs become available for GAMMs.

Lin and Zhang [10] propose constructing nonparametric smoothing spline estimators of the s functions and then estimating the smoothing parameter λ and variance components ψ using marginal quasi-likelihood as follows.

The parameters in the model are (β, ψ) and the likelihood function is

$$L\ (\beta, \psi) = \int_{R^q} f\ (y|u)\ f\ (u)\ du$$
$$= (2\pi)^{-q/2}|D\ (\psi)|^{-1/2}\exp\ (1^Tc(y))\ J\ (\beta, \psi), \quad (6)$$

where

$$J\ (\beta, \psi) = \int_{R^q} \exp\ \{y^T\ (X\beta + Zu) - 1^Tb\ (X\beta + Zu) - 1/2\ u^T\ D\ (\psi)^{-1}u\}. \quad (7)$$

Maximization of L (β, ψ) is intractable due to the presence of the q-dimensional integral J (β, ψ) in equation (7). Several methods have been proposed for circumventing this problem. The penalized quasi-likelihood method maximizes the penalized log-likelihood

$$\log f\ (y|u) - 1/2\ u^T\ D\ (\psi)^{-1}u \quad (8)$$

to obtain estimates of (β,u) for given ψ. Fixing (β,u) at their current values, Breslow and Clayton [3] suggest updating ψ at each stage of the iteration using maximum likelihood or restricted maximum likelihood applied to the pseudo-data. Alternatively, the variance components ψ can be estimated via cross-validation.

4. The generalized additive Gaussian model

The generalized additive Gaussian model assumes that the link functions are

$$g\,(\mu_{it}) = \mu_{it} = \sum_{j=1}^{p} s_j\,(x_{jit}) + z_{it}^T u_i \quad (9)$$

for the identity link and

$$g\,(\mu_{it}) = \ln \mu_{it} = \sum_{j=1}^{p} s_j\,(x_{jit}) + z_{it}^T u_i \quad (10)$$

for the log link, where $u_i \sim N\,(0, D(\psi))$.

4.1. An empirical application of GAM Gaussian model to data on anti-social behavior

4.1.1. Variable definitions and data description

The data are taken from Allison [1]. The sample is drawn from the National Longitudinal Survey of Youth (NLSY; Center for Human Resource Research, 2002). We use Allison's smaller sample of 581 children, which was drawn by the author from a much larger sample. These children were interviewed in 1990, 1992, and 1994, but we use the data in 1990 and 1994 only. The dependent variable is ANTI and all of the remaining variables are explanatory variables. The data are summarized in Table 1.

ANTI = Anti-social behavior (scale ranges from 0 to 6)
SELF = Self-esteem (scale ranges from 6 to 24)
POV = 1 if family is in poverty, 0 otherwise.
BLACK = 1 if child is BLACK, 0 otherwise
HISPANIC = 1 if child is HISPANIC, 0 otherwise
CHILDAGE = child's age in 1990
MARRIED = 1 if mother was currently married in 1990, otherwise 0

GENDER = 1 if female, 0 otherwise
MOMAGE = Mother’s age at birth of child
MOMWORK = 1 if mother was employed in 1990, 0 otherwise
TIME_2 = 1 if the year is 1992, 0 othewise
TIME_3 = 1 if the year is 1994, 0 othewise
MSELF and MPOV are the person-specific means for the variables SELF and POV respectively.
DSELF and DPOV are deviations around the person-specific means for the variables SELF and POV respectively.

Table 1: Summary statistics

Variable	Obs	Mean	Std. Dev.	Min	Max
MOMAGE	581	20.65577	2.188982	16	25
ANTI90	581	1.567986	1.470728	0	6
ANTI94	581	1.595525	1.559149	0	6
GENDER	581	0.5043029	0.5004123	0	1
CHILDAGE	581	8.943632	0.6013551	8	10
HISPANIC	581	0,2444062	0.4301049	0	1
BLACK	581	0.363167	0.4813268	0	1
MOMWORK	581	0.3356282	0.4726165	0	1
MARRIED	581	0.2358003	0.4248638	0	1
SELF90	581	20.07057	3,191613	9	24
SELF92	581	20.36213	3.528692	6	24
SELF94	581	20,6179	3.26176	9	24
POV90	581	0.3356282	0.4726165	0	1
POV92	581	0.3287435	0.4701613	0	1
POV94	581	0.3218589	0.4675918	0	1

Data source: Fixed Effects Regression Models by Allison [1].

4.1.2. Models

The following models were fit to the data. Model 1 is a fixed effects linear regression model with ANTI as the dependent variable and SELF, POV, and TIME_2 and TIME_3 as the independent variables, which is a GLM with identity link. Given the nonlinearity of the link function in SELF displayed in the partial residual plots of SELF in Fig.1, Model 2 is a fixed effects GAM with the identity link, which introduces a nonparametric smooth term s (SELF). Model 3 is a random effects GLM with the identity link with ANTI as the dependent variable and SELF, POV, TIME_2, TIME_3, BLACK, HISPANIC, CHILDAGE, MARRIED, GENDER, MOMAGE, and MOMWORK as the independent variables. Model 4 is a hybrid GLM with the identity link, ANTI as the dependent variable and DSELF, DPOV, MSELF. MPOV, TIME_2, TIME_3, BLACK, HISPANIC, CHILDAGE, MARRIED, GENDER, MOMAGE, and MOMWORK as the independent variables. Due to the nonlinearity of the link function in DSELF, CHILDAGE, and MOMAGE displayed in the partial residual plots of these variables in Fig. 2, Model 5 is a hybrid GAMM with ANTI as the dependent variable, which includes parametric terms for DPOV, MSELF, MPOV, TIME_2, TIME_3, BLACK, HISPANIC, MARRIED, GENDER, and MOMWORK, and nonparametric terms for DSELF, CHILDAGE, and MOMAGE. Estimation and inference results are presented in tables 2 through 6.

4.1.3. Nonparametric exploration of nonlinearity in the link function

The following partial residual plots help us identify the nature of nonlinearity in the link functions.

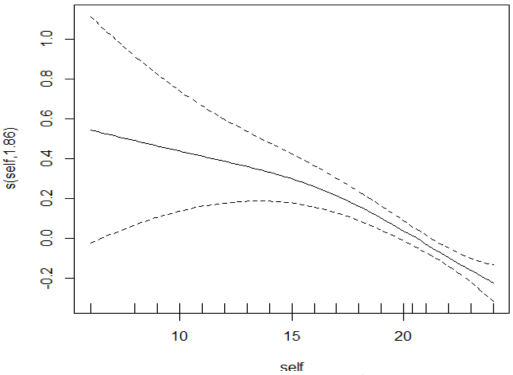

Fig. 1: Partial Residuals Plot of SELF

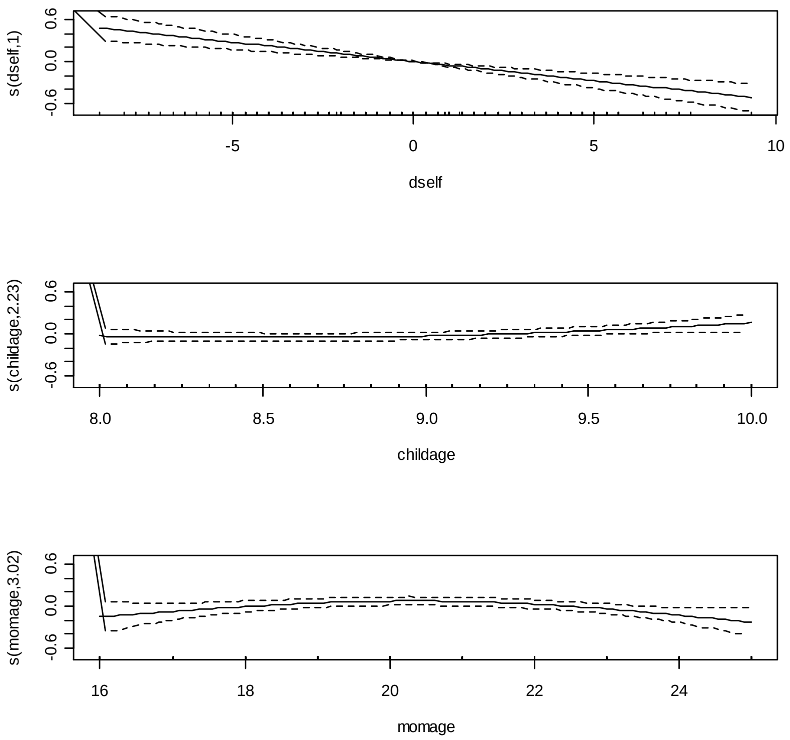

Fig. 2: Partial Residuals Plot of DSELF, CHILDAGE, and MOMAGE

Model 1: *Fixed Effects GLM Normal with Identity Link for ANTI*

$$\eta = g(\mu) = \beta_1 + \beta_2 SELF + \beta_3 POV + \beta_4 TIME\_2 + \beta_5 TIME\_3$$

Model 2: *Fixed Effects GAM Normal with Identity Link for ANTI*

$$\eta = g(\mu) = \beta_1 + \beta_2 s(SELF) + \beta_3 POV + \beta_4 TIME\_2 + \beta_5 TIME\_3$$

Model 3:*Random Effects GLM Normal with Identity Link for ANTI*

$$\begin{aligned}\eta = g(\mu) = {} & \beta_1 + \beta_2 SELF + \beta_3 POV + \beta_4 TIME\_2 + \beta_5 TIME\_3 \\ & + \beta_6 BLACK + \beta_7 HISPANIC + \beta_8 CHILDAGE + \beta_9 MARRIED \\ & + \beta_{10} GENDER + \beta_{11} MOMAGE + \beta_{12} MOMWORK + u\end{aligned}$$

Model 4: *Mixed Effects Hybrid GLM Normal with Identity Link for ANTI*

$$\begin{aligned}\eta = g(\mu) = {} & \beta_1 + \beta_2 DSELF + \beta_3 DPOV + \beta_4 MSELF + \beta_5 MPOV \\ & + \beta_6 TIME\_2 + \beta_7 TIME\_3 + \beta_8 BLACK + \beta_9 HISPANIC \\ & + \beta_{10} CHILDAGE + \beta_{11} MARRIED + \beta_{12} GENDER + \beta_{13} MOMAGE \\ & + \beta_{14} MOMWORK + Zu\end{aligned}$$

Model 5: *Hybrid GAMM Normal with Identity Link for ANTI and nonparametric smooth terms for DSELF*, *CHILDAGE*, *and MOMAGE*

$$\begin{aligned}\eta = g(\mu) = {} & \beta_1 + \beta_2 s(DSELF) + \beta_3 DPOV + \beta_4 MSELF + \beta_5 MPOV \\ & + \beta_6 TIME\_2 + \beta_7 TIME\_3 + \beta_8 BLACK + \beta_9 HISPANIC \\ & + \beta_{10} s(CHILDAGE) + \beta_{11} MARRIED + \beta_{12} GENDER + \beta_{13} s(MOMAGE) \\ & + \beta_{14} MOMWORK + Zu,\end{aligned}$$

where $u \sim N(0, D(\theta))$ in models 3, 4, and 5.

Table 2: Model (1) Fixed Effects GLM Normal with Identity Link

Variable	Estimate	Std. Error	t-ratio	p-value
SELF	-0.05515	0.01053	-5.240	1.91x10^{-7***}
POV	0.11247	0.09341	1.204	0.228797
TIME_2	0.04439	0.05858	0.758	0.448741
TIME_3	0.21074	0.05880	3.584	0.000352***

Signif. Codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9971 on 1158 degrees of freedom
Multiple R-squared: 0.8721, Adjusted R-squared: 0.8075
F-statistic: 13.5 on 585 and 1158 DF, p-value: < 2.2e-16
AIC: 5395.467

Table 3: Model (2) Fixed Effects GAM Normal with Identity Link Fit by Penalized Quasi Maximum Likelihhod

Variable	Estimate	Std. Error	t-ratio	p-value
POV	0.114297	0.093352	1.224	0.221059
TIME_2	0.050225	0.058766	0.855	0.392913
TIME_3	0.214835	0.058878	3.649	0.000275***

Approximate significance of smooth terms:

Variable	Estimated df	Refined df	F	p-value
s(SELF)	1.856	2.343	12.54	8.67x10^{-7}***

Signif. Codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
R-sq. (adj) = 0.601 Deviance explained = 87.2%
GCV score = 1.4952 Scale est. = 0.99266 n = 1743
AIC: 5393.267

Table 4: Model (3) Random Effects GLM Normal with Identity Link Estimate as Linear Mixed Effects Model Fit by Maximum Likelihood

Variable	Estimate	Std. Error	t-ratio	p-value
INTERCEPT	2.5314313	1.0935304	2.314916	0.0208
SELF	-0.0620764	0.0095203	-6.520446	0.0000***
POV	0.2471376	0.0804133	3.073343	0.0022***
TIME_2	0.0473396	0.0587324	0.806022	0.4204
TIME_3	0.2163811	0.0589054	3.673364	0.0003***
BLACK	0.2267537	0.1254321	1.80778	0.0712.
HISPANIC	-0.2182088	0.1379317	-1.582006	0.1142
CHILDAGE	0.0884559	0.0908965	0.973150	0.3309
MARRIED	-0.0495647	0.1261522	-0.392896	0.6945
GENDER	-0.4834488	0.1062911	-4.548348	0.0000***
MOMAGE	-0.0219197	0.0252337	-0.868668	0.3854
MOMWORK	0.2611318	0.1144528	2.281568	0.0229.

Signif Codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Linear mixed-effects model fit by maximum likelihood

AIC	BIC	logLik
5882.398	5958.885	-2927.199

Random effects:
Formula: ~1 | id

	(Intercept)	Residual
StdDev	1.132552	0.9964282

4.1.4. Correlation matrix of parameter estimates

```
         (Intr) self   pov    time_2 time_3 black hispnc childg marrid gender momage
Self       0.165
pov      -0.033  0.001
time_2   -0.019 -0.047  0.009
time_3   -0.013 -0.088  0.019  0.502
black    -0.116  0.005 -0.198 -0.002 -0.004
hispanic -0.140  0.040 -0.071 -0.003 -0.005  0.438
childage -0.859 -0.001 -0.007  0.000  0.000  0.013  0.057
married  -0.053  0.030 -0.117 -0.003 -0.005 -0.008 -0.028 -0.001
gender   -0.102  0.042 -0.024 -0.002 -0.004 -0.001  0.037  0.071  0.001
momage   -0.654 -0.029  0.073  0.002  0.004  0.111  0.092  0.232  0.046 -0.015
```

```
momwork  -0.028  0.016 -0.174 -0.002 -0.005  0.060 -0.048 -0.008  0.089 -0.011 -0.004
```

Standardized Within-Group Residuals:

Min	Q1	Med	Q3	Max
-3.5961743	-0.5637367	-0.1141457	0.5087177	3.3422232

Number of Observations: 1743
Number of Groups: 581

Table 5: Model (4) Hybrid Estimates Combining Fixed and Random Effects Model: Model Fit by Restricted Maximum Likelihood (Reml): Random Intercept Model and Random Slope

Variable	Estimate	Std. Error	t-ratio	p-value
INTERCEPT	2.9210139	1.1525255	2.534446	0.0114
DSELF	-0.0530277	0.0112317	-4.721249	0.0000***
DPOV	0.1142690	0.0935692	1.221224	0.2222
MSELF	-0.0924467	0.0218917	-4.222912	0.0000***
MPOV	0.6134043	0.1556909	3.939885	0.0001***
TIME_2	0.0380959	0.0585363	0.650808	0.5153
TIME_3	0.2033742	0.0588074	3.458310	0.0006***
BLACK	0.1038250	0.1311180	0.791844	0.4288
HISPANIC	-0.2816312	0.1382723	-2.036787	0.0421
CHILDAGE	0.0880806	0.0902103	0.976392	0.3293
MARRIED	-0.1357811	0.1277337	-1.063001	0.2882
GENDER	-0.5168850	0.1059485	-4.878644	0.0000***
MOMAGE	-0.0096507	0.0252677	-0.381937	0.7027
MOMWORK	0.1500985	0.1183270	1.268507	0.2051

Signif Codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Linear mixed-effects model fit by maximum likelihood

AIC	BIC	logLik
5875.479	5973.82	-2919.74

```
Random effects:
Formula: ~1 + dself | id
Structure: General positive-definite, Log-Cholesky parametrization
StdDev     Corr
(Intercept) 1.12955207 (Intr)
dself       0.07076595 -0.259
Residual    0.97564551

Correlation:
         (Intr) dself  dpov   mself  mpov   black  hispnc childg marrid gender momage momwrk time_2
dself    -0.014
dpov     -0.001 -0.012
mself    -0.367  0.031 -0.001
mpov     -0.072  0.003  0.000  0.033
black    -0.086 -0.003  0.000 -0.001 -0.366
hispanic -0.150 -0.002  0.000  0.088 -0.135  0.449
childage -0.808  0.002  0.000 -0.003 -0.014  0.016  0.058
married  -0.057  0.000 -0.001  0.061 -0.221  0.053  0.000  0.001
gender   -0.122  0.000  0.001  0.095 -0.043  0.011  0.048  0.071  0.013
momage   -0.598  0.000  0.002 -0.062  0.138  0.067  0.071  0.229  0.018 -0.024
momwork  -0.018 -0.001  0.001  0.025 -0.325  0.143 -0.011 -0.005  0.139  0.002 -0.039
time_2   -0.025 -0.055  0.013 -0.008 -0.002 -0.002  0.000  0.002 -0.001  0.002  0.002  0.003
time_3   -0.026 -0.093  0.027 -0.005  0.002 -0.005 -0.004  0.003 -0.003  0.004  0.002 -0.001  0.504

Standardized Within-Group Residuals:
Min        Q1        Med        Q3        Max
-3.4509485 -0.5378020 -0.1204521  0.5052164  3.3143416
Number of Observations: 1743
```

Number of Groups: 581

Table 6: Model (5) Random Effects GAMM Normal with Identity Link Estimate as Linear Mixed Effects Model Fit by Maximum Likelihood

Variable	Estimate	Std. Error	t-ratio	p-value
INTERCEPT	1.60003	0.06076	26.335	<2e-16***
POV	0.24714	0.05673	4.357	1.40e-05***
TIME_2	0.04935	0.05866	0.841	0.400272
TIME_3	0.21771	0.05872	3.708	0.000216***
BLACK	0.22099	0.05841	3.784	0.000160***
HISPANIC	-0.21622	0.06278	-3.444	0.000587***
MARRIED	-0.04492	0.05774	-0.778	0.436667
GENDER	-0.47690	0.04835	-9.863	<2e-16***
MOMWORK	0.26880	0.05308	5.064	4.55e-07***

Approximate significance of smooth terms:

Variable	Estimated df	Refined df	F	p-value
s(DSELF)	1	1	51.960	<22e-16***
s(CHILDAGE)	2.155	2.155	2.905	0.0510
s(MOMAGE)	3.069	3.069	4.231	0.0051**

Family: gaussian
Link function: identity

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
R-sq.(adj) = 0.0825 lmer.REML score = 5893.4 Scale est. = 0.99525 n = 1743
Linear mixed model fit by REML

```
AIC   BIC   logLik  Deviance REMLdev
5930  6033  -2946    5846        5892
```

4.2. Comparing the models

A comparison of models using the AICs presented in Table 7 suggests that Models 1 and 2, which employ ANTI as the response variable and SELF, TIME_2, and TIME_3 as explanatory variables have the lowest AICs among the models considered and are therefore the best models. Both models are fixed effects models of which Model 2 is a generalized additive fixed effects model with a nonparametric smooth term for the variable SELF. At the other extreme, Models 3 and 5 have the highest AIC, BIC, and deviance. This may suggest that a pure random effects model does not perform as well as a fixed effects model and that including nonparametric nonlinear terms in DSELF, CHILDAGE, and MOMAGE may lead to over-fitting. Following Allison [1], Models 4 and 5 employ a hybrid approach, which combines some of the merits of fixed effects and random effects models. Under this approach, the time-varying variables are transformed into deviations from their individual-specific means, but the response variable is left unchanged. Unlike fixed effects models, the time-invariant variables are included in the regression model. Additionally, variables, which are the individual-specific means for each of the time-varying variables, are also included. Instead of OLS, a random effects model is estimated. The correlation matrices display the estimated sampling correlations among the fixed-effects coefficient estimates, which are not usually of direct interest. Very large correlations, however, are indicative of an ill-conditioned model (Frees [5]). In all of the models, correlation matrices of parameter estimates display weak correlations between parameter estimates after conditioning on random effects confirming that random effects specification is desirable in all of the cases considered. Results presented in tables 2 through 6 may be summarized as follows. The variables SELF and TIME_3 are highly significant across all of the fixed and random effects models as are the variables DSELF and MSELF across all of the hybrid models. The variable GENDER is highly significant in models 3, 4, and 5 in which it is included. Surprisingly, POV is statistically insignificant in all of the models except Models 3 and 5, a random effect GLM and a hybrid GAMM respectively. Most importantly, MOMAGE and MOMWORK are highly significant in Model 5 only, which is the only semi-parametric model that captures the nonlinearity of the link function in these variables.

Table 7: Models and the Aics,

MODEL	AIC
1	5395.467
2	5393.267
3	5882.398
4	5875.479
5	5930

5. The generalized additive mixed Logit model

The generalized additive Logit model assumes that the link function is

$$g(\mu_{it}) = \text{logit}(\mu_{it}) = \ln\left(\frac{\mu_{it}}{1-\mu_{it}}\right) = \sum_{j=1}^{p} s_j(x_{jit}) + z_{it}^{T} u_i\,, \tag{11}$$

where $\mu_{it} = p_{ti} = E(y_{it} = 1 \,|\, x_{jit}, z_{it})$

$$= \exp\{\sum_{j=1}^{p} s_j(x_{jit}) + z_{it}^{T} u_i\} / \exp\{1 + \sum_{j=1}^{p} s_j(x_{jit}) + z_{it}^{T} u_i\}. \tag{12}$$

$u_i \sim N(0, D(\psi))$, where ψ is a vector of variance components.

5.1. An empirical application of GAM Logit model to choice of using a professional tax preparer

The dataset is from Frees [5]. Following Frees [5], we model choice of using a professional tax preparer (PREP) using demographic and economic characteristics of taxpayers. The data are from Statistics of Income (SOI) panel of Individual Returns. The SOI panel represents a simple random sample of unaudited individual income tax returns filed for tax years 1979-1990. The data were compiled from a stratified probability sample of unaudited individual income tax returns filed by US taxpayers. The estimates obtained from these data are intended to represent all returns filed for the income tax years under review. All returns presented are subjected to sampling except tentative and amended returns.

Following Frees [5], we use a balanced panel from 1983-1984 and 1986-1987 taxpayers included in the SOI panel, a 4% sample of this comprises our sample of 258 taxpayers. These years are chosen because they contain the interesting information on paid-preparer usage. Specifically, these data include line-item tax return data and a binary variable noting the presence of a paid tax preparer for years 1982-1984 and 1986-1987. The variable definitions are presented in Table 8 and summary statistics for the data are displayed in Table 9.

Table 8: Tax Preparer Data

Dependent Variable	
PREP is a variable indicating the presence of a paid preparer.	
Independent Variables - Demographic Characteristics	
MS	is an indicator variable of the taxpayer's marital status. It is coded one if the taxpayer is married and zero otherwise.
HH	is an indicator variable, one if the taxpayer is a head of household and zero otherwise.
DEPEND	is the number of dependents claimed by the taxpayer.
AGE	is the presence of an indicator for age 65 or over.
Independent Variables - Economic Characteristics	
F1040A	is an indicator variable of the taxpayer’s filing type. It is coded one if the taxpayer uses Form 1040A and zero otherwise.
F1040EZ	is an indicator variable of the taxpayer’s filing type. It is coded one if the taxpayer uses Form 1040EZ and zero otherwise.
TPI	is the sum of all positive income line items on the return
TXRT	is a marginal tax rate. It is computed on TPI less exemptions and the standard deduction.
MR	is an exogenous marginal tax rate. It is computed on TPI less exemptions and the standard deduction.
EMP	is an indicator variable, one if Schedule C or F is present and zero otherwise. Self-employed taxpayers have greater need for professional assistance to reduce the reporting risks of doing business.
PREP	is a variable indicating the presence of a paid preparer.
Additional Variables	
TAX	is the tax liability on the return.
SUBJECT	Subject identifier, 1- 258.
TIME	Time identifier, 1-5.
LNTAX	is the natural logarithm of the tax liability on the return.
LN^{TPI}	is the natural logarithm of the sum of all positive income line items on the return.

Source: Longitudinal and Panel Data Models by Frees [5]).

Table 9: Some Data Characteristics of the Tax preparer Data

Variable	Obs	Mean	Std. Dev.	Min	Max
MS	1290	0.62	0.48	0	1
HH	1290	0.09	0.28	0	1
DEPEND	1290	2.42	1.34	0	6
AGE	1290	0.12	0.32	0	1
F1040A	1290	0.18	0.38	0	1
F1040EZ	1290	0.11	0.31	0	1
TPI	1290	30279.65	36634.47	0.88	552403.19
TXRT	1290	21	10.55	0	50
MR	1290	23.52	11.45	0	50
EMP	1290	0.15	0.36	0	1
PREP	1290	0.48	0.50	0	1
TAX	1290	4095.57	8612.29	0	141461.27
SUBJECT	1290	129.50	74.51	1	258
TIME	1290	3.00	1.41	1.00	5.00

5.1.1. Models

The following models were fit to the data. Models 1 and 2 are GLMMs and models 3 and 4 are GAMMs. The response variable in all of the models is PREP, choice of a professional tax-preparer. Model 1 is a generalized linear mixed Poisson regression model, which employs LNTPI, TAX, AGE, DEPEND and EMP as the explanatory variables. Model 2 is also a generalized linear mixed Poisson regression model, which uses LNTPI, MR, and EMP as the explanatory variables. The partial residual smoothing plot of LNTPI in Fig. 3 displays a high degree of nonlinearity in LNTPI. The dotted curves around the solid curve represent +-2 standard errors around the solid curve. Given the nonlinearity of the Logit link function in LNTPI, Model 3 is a generalized additive mixed Logit model, which introduces a nonparametric smooth term s (LNTPI) and replaces MR with TAX. Model 4 is also a generalized additive mixed Logit model, which ads DEPEND to the list of explanatory variables.

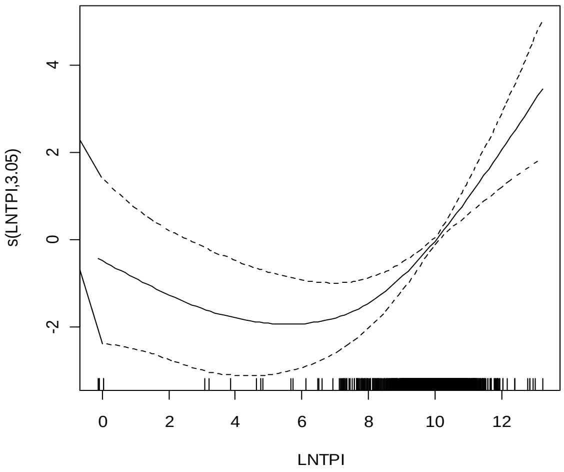

Fig. 3: Partial Residuals Plot of LNTPI in GAM Regression

$$Model\ 1: \mathrm{GLMM}\ Logit\ Re\ gression\ 1$$

$$\eta = g(\mu) = \beta_1 + \beta_2 LNTPI + \beta_3 TAX + \beta_4 AGE + \beta_5 DEPEND + \beta_6 EMP + zu$$

$$Model\ 2: GLMM\ \ Logit\ Re\ gression\ 2$$

$$\eta = g(\mu) = \beta_1 + \beta_2 LNTPI + \beta_3 MR + \beta_4 EMP + zu$$

$$Model\ 3: \ GAMM\ \ Logit\ Re\ gression\ Model\ 1\ with\ Nonparametric\ smooth$$

$$term\ \ for\ LNTPI$$

$$\eta = g(\mu) = \beta_1 + \beta_2 s(LNTPI) + \beta_3 TAX + \beta_4 AGE + \beta_5 EMP + zu$$

$$Model\ 4: GAMM\ \ Logit\ Re\ gression\ Model\ 2\ with\ Nonparametric\ \mathrm{smooth}$$

$$term\ \ for\ LNTPI$$

$$\eta = g(\mu) = \beta_1 + s(LNTPI) + \beta_3 TAX + \beta_4 AGE + \beta_5 DEPEND + \beta_6 EMP + zu,$$

where $u \sim N(0, D(\psi))$ and $g(\mu)$ is defined in eqn. (11) above.

Table 10: Model (1) Random Effects Logit with Taxpayer Characteristics as Predictors and TAX as the Tax Liability Variable

Variable	Coefficient Estimate	Std. Error	t-ratio	p-value
Intercept	-3.841	1.314	-2.924	0.00346**
LNTPI	0.1902	0.1305	1.458	0.14497
TAX	8.198x10^{-5}	4.504x10^{-5}	1.820	0.06875
AGE	1.946	0.06485	3.000	0.00270**
DEPEND	0.03824	0.01611	2.373	0.01764*
EMP	1.733	0.05298	3.270	0.00107**

```
Signif. Codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
R-sq.(adj) =  0.0577glmer.ML score = 1047.9  Scale est. = 1        n = 1290
Generalized linear mixed model fit by maximum likelihood
Family: binomial ( Logit )
     AIC       BIC            logLik     deviance
1061.9109  1098.0477 -523.9555 1047.9109
Random effects:
Groups           Name        Std.Dev.
SUBJECT (Intercept)       4.3
Number of obs: 1290, groups: SUBJECT, 258
```

Table 11: Model (2) Random Effects Logit with Taxpayer Characteristics as Predictors and MR as the Tax Liability Variable

Variable	Coefficient Estimate	Std. Error	t-ratio	p-value
Intercept	-3.11529	1.34872	-2.310	0.020899*
LNTPI	0.22803	0.15613	1.461	0.144144
MR	0.01395	0.02035	0.685	0.493258
EMP	1.79349	0.54102	3.315	0.000916***

```
Generalized linear mixed model fit by the Laplace approximation
AIC  BIC   logLik deviance
1074 1099 -531.8     1064
Random effects:
Groups          Name        Variance   Std.Dev.
SUBJECT (Intercept)     19.838      4.4539
Number of obs: 1290, groups: SUBJECT, 258

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

5.1.2. Correlation matrix of parameter estimates

```
(Intr) LNTPI  MR   EMP
LNTPI   -0.931
MR        0.370 -0.630
EMP     -0.257  0.194  -0.064
```

Table 12: Model (3) GAM Mixed Logit with Taxpayer Characteristics as Predictors and A Smooth Nonparametric Term for LNTPI Parametric Coefficients

Variable	Coefficient Estimate	Std. Error	t-ratio	p-value
Intercept	-0.0776	0.03596	-2.162	0.03060*
TAX	1.373x10^{-5}	4.310x10^{-5}	0.319	0.74999
AGE	2.026	0.06548	3.094	0.00198**
EMP	1.656	0.05301	3.124	0.00179**

Approximate significance of smooth terms:

Variable	Estimated df	Refined df	Chi-sq	p-value
s(LNTPI)	2.202	2.202	7.757	0.026*

```
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
R-sq.(adj) =  0.0692 glmer.ML score = 1049.6  Scale est. = 1        n = 1290

Generalized linear mixed model fit by maximum likelihood
```

```
Family: binomial ( Logit )
     AIC          BIC           logLik     Deviance
1063.5630    1099.6998  -524.7815  1049.5630
Random effects:
 Groups         Name        Std.Dev.
 SUBJECT (Intercept)    4.337
 Xr              s(LNTPI)    11.397
Number of obs: 1290, groups: SUBJECT, 258; Xr, 8

Family: binomial
Link function: Logit
Formula:
PREP ~ s(LNTPI) + TAX + AGE + EMP
Estimated degrees of freedom:
2.2  total = 6.2
Glmer.ML score: 1049.563
```

Table 13: Model (4) Gam Logit with Taxpayer Characteristics as Predictors and a Smooth Nonparametric Term for Lntpi Parametric Coefficients

Variable	Coefficient Estimate	Std. Error	t-ratio	p-value
Intercept	-1.590	0.05537	-2.871	0.00409**
TAX	3.272x10^{-5}	4.419x10^{-5}	0.740	0.45901
AGE	2.087	0.06533	3.195	0.00140**
DEPEND	0.03025	0.01632	1.853	0.06390
EMP	1.638	0.05282	3.102	0.00192**

Approximate significance of smooth terms:

Variable	Estimated df	Refined df	Chi-sq	p-value
s(LNTPI)	1.926	1.926	3.83	0.137

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

```
Generalized linear mixed model fit by maximum likelihood
 Family: binomial ( Logit )
AIC       BIC             logLik    Deviance
1062.5350 1103.8342  -523.2675 1046.5350
Random effects:
Groups         Name        Std.Dev.
SUBJECT (Intercept)   4.307

R-sq. (adj) = 0.0675   glmer.ML score = 1046.5 Scale est. = 1         n = 1290
```

5.2. Comparing the models

Estimation results are presented in tables 10 through 13. A comparison of models using the AIC presented in Table 14 suggests that Models 1 and 4, which employ PREP as the response variable and TAX, AGE, DEPEND, EMP and LNTPI as explanatory variables have the lowest AICs among the models considered and are therefore the best models. Model 1 is a generalized linear random effects Logit model while Model 4 is a generalized additive mixed effects Logit model with a nonparametric smooth term for the variable LNTPI. At the other extreme, Models 2, a random effects Logit model, which includes LNTPI, MR and EMP has the highest AIC, BIC, and deviance. This may suggest that a pure random effects model, which omits AGE and replaces TAX with MR as the tax liability variable, does not perform as well as the other random effects GLMMs and GAMMs.

Table 14: Models and the Aics

MODEL	AIC
1	1061.9109
2	1074
3	1063.5630
4	1062.5350

Correlation matrices of parameter estimates display weak correlations between parameter estimates after conditioning on random effects confirming that random effects specification is desirable in all of the cases considered.

Examination of tables 10-13 suggests that the variable EMP is statistically significant at 1% significance level in all of the models. However, the tax liability variables TAX and MR are statistically insignificant across all models. The variable LNTPI is statistically insignificant in all of the models except Model 3. The variable AGE is statistically significant at 1% significance level in all of the models. The positive signs of the coefficient estimates are all expected in all of the models. For instance, the signs of AGE, DEPEND, EMP and LNTPI are positive in all of the models indicating that the odds of choosing a professional tax preparer are higher for a taxpayer who is 65 years or older or has dependents, is self-employed or whose total income increases than for a taxpayer who does not have these traits. The surprising statistical insignificance of tax liability variables TAX and MR in the GAMM models could be attributed to possible over-fitting in these models.

6. The generalized additive mixed effects poisson models

Poisson and Negative Binomial regression models are the most widely used count data models. In these models, the outcome, y_{it} is a count variable. Generalized additive mixed extensions of these models replace $\sum x_{jit}\beta_j$, the linear component of the model with an additive component $\sum f_j(x_{jit})$ in the link function and introduce a random effects term. We wish to model p $(y_{it}|x_{1it}, x_{2it}\ldots x_{pit})$, the probability of an event given variables $x_{1it}, x_{2it}\ldots x_{pit}$. The Poisson regression model assumes that the link function is linear:

$$\ln\mu_{it} = \beta_0 + x_{1it}\beta_1 + \ldots + x_{pit}\beta p \quad (13)$$

The generalized additive mixed Poisson model assumes instead that

$$\ln\mu_{it} = s_1(x_{1it}) + \ldots + s_p(x_{pit}) + z_{it}^T u_i\text{, where } u_i \sim N(0, D(\psi)), \quad (14)$$

where $s_1, s_2, \ldots, s_p$ are smooth nonparametric functions, which are estimated by maximizing a penalized quasi-log likelihood approach described in Section 3 above.

6.1. An empirical application of GAMM poisson models to patents and R & D data

6.1.1. Data and variable definitions

The data are from Bronwyn Hall, Zvi Griliches, and Jerry Hausman (1986), "Patents and R&D: Is There a Lag?", International Economic Review, 27, 265-283. The following variables were used in econometric analysis.
CUSIP = Compustat's identifying number for the firm (Committee on Uniform Security Identification Procedures number).
ARDSIC = A two-digit code for the applied R&D industrial classification
(Roughly that in Bound, Cummins, Griliches, Hall, and Jaffe, in the Griliches R&D, Patents, and Productivity volume).
SCISECT = Dummy equal to one for firms in the scientific sector.
LOGK = the logarithm of the book value of capital in 1972.
SUMPAT = the sum of patents applied for between 1972-1979.
LOGR70- = the logarithm of R&D spending during the year (in 1972 dollars).
LOGR79
LOGR =LOGR79-LOGR75, LOGR1 =LOGR78-LOGR74, LOGR2 =LOGR77-LOGR73,
LOGR3 =LOGR76-LOGR72, LOGR4 =LOGR75-LOGR71, LOGR5 =LOGR74-LOGR70
PAT70- = the number of patents applied for during the year that were
PAT79 eventually granted.
TIME DUMMIES: DYEAR2 = 1 if year =2, 0 otherwise; DYEAR3 = 1 if year =3, 0 otherwise; DYEAR4 = 1 if year =4, 0 otherwise; DYEAR5 = 1 if year =5, 0 otherwise.
The dependent variable is PAT and all of the remaining variables are independent variables. The data are summarized in Table 15.

6.1.2. Models

The following models were fit to the data using PAT as the response variable. Model 1 is a generalized linear mixed effects Poisson regression model, which employs LOGR, LOGR1, LOGR2, LOGR3, LOGR4, and LOGR5 as the explanatory variables. Model 2 extends model 1 by including time dummies for four of the five years using Year1 as the reference year. Next, we explore nonlinearities in the link function non-parametrically and present partial residual plots. Smooth nonparametric terms are included in the link functions if nonlinearities are confirmed through these plots. Given the nonlinearity of the Poisson link function in LOGR1 as displayed in Fig. 4, Model 3 was chosen to be a semi-parametric Poisson regression model in which the link function is of the additive form and includes a nonparametric smooth term for LOGR1 and parametric linear terms for all other variables.

Table 15: Summary Statistics for the Patent Data

Variable	Obs.	Mean	Std. Dev.	Min	Max
OBSNO	1730	173.50	99.91	1.00	346.0
YEAR	1730	3.0	1.41	1.00	5.00
CUSIP	1730	531201.21	281748.41	800	989399
ARDSSIC	1730	-19.18	169.17	-999	21
SCISECT	1730	0.42	0.49	0	1
LOGK	1730	3.92	2.09	-1.77	9.67
SUMPAT	1730	284.73	570.45	0	3806
PAT	1730	34.77	70.88	0	515.0
PAT1	1730	35.87	72.76	0	528
PAT2	1730	36.70	75.12	0	595
PAT3	1730	36.72	75.53	0	595
PAT4	1730	37.17	76.54	0	595
LOGR	1730	1.26	2.01	-3.85	7.03
LOGR1	1730	1.23	1.98	-3.85	7.07
LOGR2	1730	1.22	1.97	-3.85	7.07
LOGR3	1730	1.21	1.95	-3.85	7.07
LOGR4	1730	1.20	1.94	-3.67	7.07
LOGR5	1730	1.20	1.93	-3.67	7.07

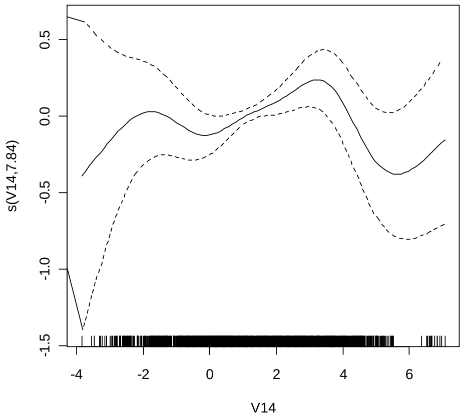

Fig. 4: Partial Residual Plot of LOGR1 (V14)

Model 1 : Mixed Poisson Regression Model with Log Link

$$\eta = g(\mu) = \beta_1 + \beta_2 LOGR + \beta_3 LOGR1 + \beta_4 LOGR2 + \beta_5 LOGR3 + \beta_5 LOGR4 + LOGR5 + Zu$$

Model 2 : Mixed Poisson Regression Model with Time Dummies and Log Link

$$\eta = g(\mu) = \beta_1 + \beta_2 LOGR + \beta_3 LOGR1 + \beta_4 LOGR2 + \beta_5 LOGR3 + \beta_6 LOGR4 + \beta_7 LOGR5 + \beta_8 DYEAR2 + \beta_8 DYEAR3 + \beta_9 DYEAR4 + \beta_{10} DYEAR5 + Zu$$

Model 3 : Generalized Additive Mixed Poisson Regression Model with smooth nonparametric term for LOGR1

$$\eta = g(\mu) = \beta_1 + \beta_2 LOGR + \beta_3 s(LOGR1) + \beta_4 LOGR2 + \beta_5 LOGR3 + \beta_6 LOGR4 + \beta_7 LOGR5 + \beta_8 DYEAR2 + \beta_8 DYEAR3 + \beta_9 DYEAR4 + \beta_{10} DYEAR5 + Zu$$

Table 16: Model (1) Mixed Effects Poisson Regression Model

Variable	Coefficient Estimate	Std. Error	t-ratio	p-value
INTERCEPT	1.05527	0.07368	14.322	<2e-16***
LOGR	0,17241	0.03902	4.419	$9.92\text{x}10^{-6***}$
LOGR1	0.04326	0.04652	0.930	0.352322
LOGR2	0.17403	0.04379	3.974	$7.05\text{x}10^{-05***}$
LOGR3	0.13602	0.04046	3.362	0.000774***
LOGR4	0.05657	0.03675	1.540	0.123674
LOGR5	0.04383	0.03100	1.414	0.157446

Signif Codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

```
(Dispersion parameter for poisson family taken to be 1)

Generalized linear mixed model fit by the Laplace approximation
AIC  BIC logLik deviance
4963 5007  -2474     4947
Random effects:
Groups Name        Variance Std.Dev.
id     (Intercept) 1.2799   1.1313
Number of obs: 1730, groups: id, 346
```

6.1.3. Correlation matrix of parameter estimates

```
(Intr)  LOGR    LOGR1    LOGR2    LOGR3    LOGR4   LOGR5
LOGR    -0.092
LOGR1  -0.054    -0.554
LOGR2  -0.055     0.069     -0.497
LOGR3  -0.055    -0.021     0.064       -0.569
LOGR4 -0.028    -0.145     0.022        0.095        -0.496
LOG R5 -0.068   -0.202   -0.115       -0.030        0.114        -0.460
```

Table 17: Model (2) Mixed Effects Poisson Regression Model

Variable	Coefficient Estimate	Std. Error	t-ratio	p-value
INTERCEPT	0.928353	0.066869	13.883	<2e-16***
LOGR	0.487805	0.042367	11.514	<2e-16***
LOGR1	-0.002582	0.047993	-0.054	0.957100
LOGR2	0.139574	0.044801	3.115	0.001837**
LOGR3	0.063122	0.041322	1.528	0.126622
LOGR4	0.028317	0.037616	0.753	0.451582
LOGR5	0.086404	0.030996	2.788	0.005311**
DYEAR2	-0.047085	0.013134	-3.585	0.000337***
DYEAR3	-0.057034	0.013366	-4.267	1.98e-05***
DYEAR4	-0.192322	0.013798	--13.938	<2e-16***
DYEAR5	-0.256210	0.014225	-18.011	<2e-16***

```
Generalized linear mixed model fit by the Laplace approximation
AIC   BIC   logLik  Deviance
4529 4594 -2252    4505
Random effects:
Groups Name        Variance Std.Dev.
id     (Intercept) 0.96982 0.9848
Number of obs: 1730, groups: id, 346

Signif Codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

6.1.4. Correlation matrix of parameter estimates

```
(Intr)        LOGR    LOGR1    LOGR2    LOGR3    LOGR4    LOGR5   DYEAR2 DYEAR3 DYEAR4 DYEAR5
LOGR      -0.100
LOGR1    -0.073    -0.546
LOGR2    -0.045     0.030       -0.481
LOGR3    -0.035    -0.033       0.048     -0.567
LOGR4    -0.010    -0.144       0.003      0.088       -0.480
LOGR5    -0.075    -0.177      -0.100     -0.032       0.107       -0.472
DYEAR2 -0.089    -0.167       0.260     -0.094      -0.044      -0.020       0.032
DYEAR3 -0.064    -0.208       0.162      0.144      -0.156       -0.047      0.037        0.510
DYEAR4 -0.022    -0.279       0.162      0.041       0.059       -0.118       0.028        0.485         0.515
DYEAR5 0.031    -0.361        0.126      0.047      -0.009       0.088      -0.044       0.461          0.491         0.517
Signif Codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Table 18: Model (3) Generalized Additive Mixed Effects Poisson Regression Model

Variable	Coefficient Estimate	Std. Error	t-ratio	p-value
INTERCEPT	0.91893	0.05981	15.364	<2e-16***
LOGR	0.47337	0.03342	14.164	<2e-16***
LOGR2	0.14348	0.04430	3.239	0.001201**
LOGR3	0.08240	0.04019	2.050	0.040327*
LOGR4	0.01403	0.03641	0.385	0.700040
LOGR5	0.08952	0.02494	3.589	0.000332***
DYEAR2	-0.04913	0.01311	-3.748	0.000178***
DYEAR3	-0.05758	0.01329	-4.331	1.48e-05***
DYEAR4	-0.18614	0.01360	-13.683	<2e-16***
DYEAR5	-0.24023	0.01367	-17.569	<2e-16***

Approximate significance of smooth terms:

Variable	Estimated df	Refined df	Ch-squared	p-value
s(LOGR1)	7.836	7.836	1400	$<2x10^{-16***}$

Family: poisson
Link function: log

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

```
R-sq.(adj) =  0.485 glmer.ML score = 4473.8  Scale est. = 1         n = 1730
Generalized linear mixed model fit by the Laplace approximation
AIC    BIC   logLik Deviance
4500   4571    -2237     4474
Random effects:
Groups Name        Variance     Std.Dev.
id     (Intercept)   0.92382        0.96115
Xr.1   s (V14)     148.03991     12.16717
Number of obs: 1730, groups: id, 346; Xr.1, 8
```

6.2. Comparing the models

As for the Gaussian and Logit GLMMs, correlation matrices of parameter estimates for Poisson GLMMs also display weak correlations between parameter estimates after conditioning on random effects confirming that random effects specification is desirable in all of the cases considered.

A comparison of models using the AICs is presented in Table 19. Based on the AIC, BIC, and Deviance criteria in Table 19, Model 3, a GAMM model with the nonparametric term s(LOGR1) appears to be the best since it has the lowest AIC index among all models. At the other extreme, Model 1, a GLMM Poisson regression model with no time dummies, has the highest AIC and is consequently the poorest model. The estimation and significance testing results are presented in tables 16 through 18. The variables LOGR, LOGR2, and LOGR3 are statistically significant in Model 1, which does not include time dummies. However, in models 2 and 3, which include time dummies, LOGR, LOGR2, and LOGR5 are highly statistically significant, but LOGR3 is not. All of the time dummies are also highly significant across all models. The remaining variables are statistically insignificant across all of the models.

Table 19: GLM and GAM Poisson Models and Their Aics

MODEL	AIC
1	4963
2	4529
3	4500

7. Conclusion

The paper has studied applications of generalized additive mixed models (GAMMs), including GAMM Gaussian, Logit, and Poisson regression models in business and economics. Unlike GLMs and GLMMs, these models allow us to explore the relationship between the response and multiple predictor variables non-parametrically in the presence of nonlinearities in the link function using longitudinal data. The applications studied ranged from the analysis of anti-social behavior to the choice of a professional tax-preparer to the number of patents issued to a manufacturing firm. In all of the empirical applications, the semi-parametric GAMMs generally performed better than the parametric generalized linear mixed models (GLMMs).

The GAMMs employed in the paper are widely applicable to the analysis of continuous, count, and binary response with longitudinal data occurring frequently in business and social sciences. GAMMs offer important advantages over GLMMs, including extension of nonparametric regression to more than one regressor circumventing the curse of dimensionality, non-parametric exploration of nonlinearities and interactions among explanatory variables as well as accounting for correlation and over-dispersion among responses. Nevertheless, the GAMMs are not without drawbacks. The computational algorithms are complex due to the presence of multiple integrals in the likelihood function and difficulties in interpretations. Poor prediction performance due to over-fitting in some situations is also a drawback of GAMMs. These models are useful mainly when parametric GLMs and GLMMs provide an inadequate fit for the data.

References

[1] Allison, P. Fixed Effects Regression Models, Sage, (2009).

[2] Baltas, G., Determinants of store brand choice: a behavioral analysis. Journal of Product & Brand Management, Vol. 6, No.5, (1997), pp. 315 – 324. http://dx.doi.org/10.1108/10610429710179480.

[3] Breslow, N. and Clayton, D. Approximate inference in generalized linear mixed models. Journal of the American Statistical Association, Vol. 88, No.1, (1993), pp. 9-25.

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

[5] Frees, E. Longitudinal and Panel Data. Cambridge University Press, New York, (2004). http://dx.doi.org/10.1017/CBO9780511790928.

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

[7] 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. http://dx.doi.org/10.1287/mksc.2.3.203.

[8] Hastie, T. & Tibshirani, R, Generalized Additive Models, Statistical Science Vol.1, No.3, (1986), pp. 297-318. http://dx.doi.org/10.1214/ss/1177013604.

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

[10] Lin, X.and Zhang D. Inference in Generalized Additive Mixed Model by Using Smoothing Splines. Journal of the Royal Statistical Society, Vol. 61, No. 2, (1999), 381-400. http://dx.doi.org/10.1111/1467-9868.00183.

[11] McCullagh, P. and J. Nelder, Generalized Linear Models, Chapman and Hall, London (1989). http://dx.doi.org/10.1007/978-1-4899-3242-6.

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

[13] Ruppert, D., M. Wand, and R. Carrol, Semiparametric Regression, Cambridge University Press, Cambridge, (2003). http://dx.doi.org/10.1017/CBO9780511755453.

[14] Sapra, S., Generalized Additive Models in Business and Economics, International Journal of Advanced Statistics and Probability, Vol.1, No. 3, (2013), pp. 64-81.

[15] Silverman, B. W. Spline smoothing: The equivalent variable kernel method. Annals of Statistics, Vol 12, No. 4, (1984), pp. 898-916. http://dx.doi.org/10.1214/aos/1176346710.

[16] Wabha, G. Bayesian confidence intervals for the :cross validated" smoothing spline, Journal of the Royal Statistical Society Series B, Vol. 45, No. 1, (1983), pp. 133-150.

[17] Wang, Y. Mixed affects smoothing analysis of variance, Journal of the Royal Statistical Society Series B, Vol. 60, No. 1, (1998), pp. 159-174. http://dx.doi.org/10.1111/1467-9868.00115.

[18] Wood, S. N., R-package gamm4, (2012).

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