School of Economics Working Paper
2016-03

University of East Anglia

SCHOOL OF ECONOMICS

Generalized additive modelling of the repayment performance of Korean borrowers

by Young Ah Kim*
Peter G. Moffatt**

*University of Essex
**School of Economics, University of East Anglia

Abstract

Data from a sample of around 32,000 customers taking out personal loans from a Korean bank, are analysed. The focus of analysis is a binary variable indicating default, defined as any sort of failure to meet the obligations of the loan during a time period up to a fixed reference date. Around 1.5% of the sample defaulted. The Generalized Additive Modelling (GAM) framework is used to investigate the combined effect of a number of factors on the likelihood of default. The GAM framework allows flexibility in the effects of continuously-distributed predictors. The B-spline smoothing approach is used for each of these effects. An extensive model-selection process is implemented. It is found that the statistical fit improves if some, but not all, of the predictors are modelled with the B-spline. The two variables found to have non-linear effects are amount borrowed and age of borrower. The predicted probability curve obtained for the former shows that borrowers least likely to default are those that have borrowed at the 85th percentile of the amount-borrowed distribution. The prediction curve for the latter shows that the default probability declines in a stepwise fashion with age, falling abruptly at certain ages, but appearing to level off for significant periods within the life-cycle.

School of Economics
University of East Anglia
Norwich Research Park
Norwich NR4 7TJ
United Kingdom
www.uea.ac.uk/economics

GENERALIZED ADDITIVE MODELLING OF THE REPAYMENT PERFORMANCE OF KOREAN BORROWERS

Young Ah KIM
University of Essex

and

Peter G MOFFATT
University of East Anglia

12 MAY 2016

ABSTRACT

Data from a sample of around 32,000 customers taking out personal loans from a Korean bank, are analysed. The focus of analysis is a binary variable indicating default, defined as any sort of failure to meet the obligations of the loan during a time period up to a fixed reference date. Around 1.5% of the sample defaulted. The Generalized Additive Modelling (GAM) framework is used to investigate the combined effect of a number of factors on the likelihood of default. The GAM framework allows flexibility in the effects of continuously-distributed predictors. The B-spline smoothing approach is used for each of these effects. An extensive model-selection process is implemented. It is found that the statistical fit improves if some, but not all, of the predictors are modelled with the B-spline. The two variables found to have non-linear effects are amount borrowed and age of borrower. The predicted probability curve obtained for the former shows that borrowers least likely to default are those that have borrowed at the 85^{th} percentile of the amount-borrowed distribution. The prediction curve for the latter shows that the default probability declines in a stepwise fashion with age, falling abruptly at certain ages, but appearing to level off for significant periods within the life-cycle.

1. Introduction

Binary data models such as logit and probit are popular tools in the modelling of loan defaults (see for example Greene, 1998). In this paper, we extend this framework to allow predictor variables to have fully flexible effects on the outcome. The extended framework is known as the Generalized Additive Modelling (GAM) framework (Hastie and Tibshirani, 1990). The essence of the approach is that each individual effect is modelled using a scatter-plot smoothing procedure.

The smoothing procedure adopted here is the B-spline smoother (de Boor, 1978). The B-spline procedure offers an attractive compromise between polynomial regression and kernel regression. The former provides global fit, in the sense that the position of the smoother at any point is determined by all observations even those furthest from the point; the latter provides local fit, in the sense that only local observations determine the position of any point of the smoother. The B-spline procedure lies somewhere in between.

The B-spline approach has another major advantage that is not widely discussed. It is a non-parametric technique that can be performed as a (generalised) linear regression, since it simply amounts to a regression of the dependent variable on a set of basis functions. This clearly makes implementation relatively straightforward. A further advantage is that by-products of regression analysis such as statistical significance tests may be exploited to the full. Statistical testing is often an awkward problem in the context of nonparametric regression. Using the B-spline, it becomes possible, using standard regression-based tests, to adjudicate between models, and in particular to make valid judgements on whether a predictor should be represented flexibly at all, in preference to assuming a linear effect.

Application of the GAM approach to the modelling of loan defaults has already been suggested by Taylan et al. (2007).

In Section 2, we motivate and outline the GAM framework. In Section 3, we describe the data, which is binary default data from a sample of customers of a Korean bank. In Section 4, we present and interpret the results from applying the GAM framework to this data set. Section 5 concludes.

2. Generalized Additive Models (GAMs)

Traditional regression models frequently fail for the simple reason that the effects of interest are often non-linear. To characterise such effects, flexible statistical methods such as nonparametric regression are a useful first step (Fox, 2002).

However, if the number of independent variables is large, many forms of nonparametric regression do not perform well. To overcome this difficulty, Stone (1985) proposed additive models. These models estimate an additive approximation of the multivariate regression function.

For non-continuous (e.g. binary) outcomes, further generality is required. Hastie and Tibshirani (1990) introduced the framework of Generalized Additive Models (GAMs). These include a link function that allows for the discrete nature of the dependent variable.

For the case of a binary dependent variable y_i and a total of m available predictors, and if we assume a probit link function, the model takes the following form:

$$P(y_i = 1|x_{1i}, \cdots, x_{mi}) = \Phi\left(\beta_0 + \sum_{j=1}^{m} f_j\left(x_{ji}\right)\right) \qquad (1)$$

where $\Phi(.)$ is the standard normal c.d.f. In general the functions f_j are piecewise polynomials (or "splines") although they do not have to be so. Some predictors are better modelled linearly (so $f_j(x_{ji}) = \beta_j x_{ji}$ in the context of (1)) and, of course, some (e.g. binary dummy variables) can only be modelled in this way.

Splines form a useful compromise between the global fit of polynomial regression, and the local fit of kernel smoothers. The "pieces" of the piecewise polynomials are separated by a sequence of K "knots", $\xi_1, \ldots, \xi_K$, and are forced to join smoothly at these knots.

Cubic splines are usually chosen, and the smoothness requirement is that the piecewise cubic functions are continuous and have continuous first and second derivatives at the knots[1].

The more knots are used, the more flexible the smoother. However more knots also means more parameters to estimate, and therefore fewer degrees of freedom. Clearly the choice of the number of knots must depend on the sample size: the larger the sample, the more knots can be used. Another choice that needs to be made is the positioning of the knots. An obvious strategy is to spread the knots uniformly over the range of the predictor variable, giving rise to what is known as a "cardinal spline". A more adaptive approach is one that places knots at appropriate quantiles of the predictor variable; for example, three knots, one at each of the three quartiles of the predictor. A more adventurous scheme would be one that places a higher concentration of knots in parts of the range of the predictor in which the nonlinearities are seen to be most marked. Such a scheme must of course be conducted by trial and error, since a smoother must be seen in order for nonlinearities to be identified.

The necessity of using ad hoc procedures for choosing the number and position of the knots is often seen as one of the (few) shortcomings of the spline approach.

[1] According to Hastie and Tibshirani (1990, p.22) "our eyes are skilled at picking up 2nd and lower order discontinuities, but not higher".

The most popular approach for obtaining a piecewise cubic smoother with the required properties is the B-spline approach (de Boor, 1978). This amounts to a linear regression of the dependent variable on a set of *basis functions*, with no intercept. If there are *K* knots, there are *K+4* basis functions in total, although for practical reasons, only *K+2* of them are used in the regression. For illustration, Figure 1 shows the basis functions obtained from the variable "age", as used in the model of later sections. There are five knots, and therefore seven basis functions used in the regression[2].

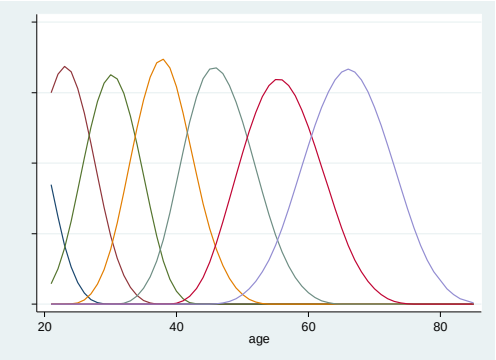

Figure 1: basis functions for variable age; knots at 23,30,38,44,55.

If the basis functions to be used in the B-spline regression are $B_1(.) \cdots B_{K+2}(.)$, then the piecewise cubic functions $f_j(.)$ appearing in (1) may be expressed as:

$$f_j(x_{ji}) = \sum_{k=1}^{K+2} \gamma_{jk} B_k(x_{ji}) \quad j = 1, \cdots, m \qquad (2)$$

It is important that (2) does not contain an intercept. This is necessary for the model intercept (β_0 in (1)) to be identified.

Note also that (2) would lead to a fully general GAM in the sense that all *m* of the predictors are being assumed to have flexible effects. As noted in the discussion following (1), there are strong reasons for *not* modelling the effect of every predictor in accordance with (2).

As mentioned in Section 1, an understated advantage of the "GAM with B-spline" approach is that it can be estimated with a linear regression. Although the regression coefficients on the basis functions (i.e. estimates of the γ_{jk}'s in (2)) are themselves hard to interpret, it is a straightforward matter to perform regression-based tests (e.g. for predictor *j*, a joint test of H_0: $\gamma_{j1}=0;\ldots \gamma_{j,K+2}=0$) in order firstly to test for the presence of non-linear effects of predictors, and secondly to adjudicate between models. It is well known that hypothesis testing can be a very awkward problem in the context of non-parametric regression; such problems are clearly avoided under the approach adopted here.

[2] These basis functions are obtained using the following STATA (add-on) command (Newson, 2000): `bspline, x(age) kn(23,30,38,44,55) power(3) gen(bsage)`

3. **Data**

The data is from a Korea-based commercial bank founded in 1969, and headquartered in Jeonju. In 2011, the bank had 87 branches distributed throughout South Korea. The bank is engaged in the provision of a wide range of commercial and consumer banking services, including deposits, loans, foreign currency exchange, trust business, trade finance, exchange rate and interest rate information, and credit cards.

This research focuses on individual loans, and each unit of observation is an individual borrower. The original dataset contains data on 65,534 borrowers. However, information on individual income is only available on 32,310 borrowers, and since income is one of the key predictors in the models we estimate, we use data on this sub-sample in estimation.

Descriptive statistics for all variables used in the analysis are presented in Table A2 of the Appendix, and variable definitions are provided in Table A3. All loans commenced between May 1992 and June 2012, with dates of redemption between June 2001 and February 2051. The reference date for the default information is 31 December 2012: "default" is defined as any sort of failure to meet the obligations of the loan during the time period between commencement of the loan and this date.

Banks in Korea aim to achieve that overall default rate of individual loans is between 1.5% and 2.0%. As seen from the first row of Table A2, the overall default rate of this bank is marginally above 1.5%.

4. **Results**

We estimate six binary probit models, and the results are reported in Table A1 of the Appendix. The dependent variable is 1 if the borrower defaulted and zero otherwise. The predictors are: log(amount borrowed); log(income); age in years; male (1 if male, 0 if female); mar (1 if married, 0 otherwise); dep (number of dependents); pur1 (purpose of loan is accommodation); pur2 (purpose of loan is living expenses). The base case purpose-of-loan dummy is "other purposes". Data on 32,310 borrowers is used in all estimations.

Model 1 is a simple probit model with linear effects only. Model 2 is a simple probit model with quadratic effects assumed for the three continuous predictors (log(amount borrowed), log(income), and age). On the basis of the AIC[3], Model 2 is superior, and we see in particular that amount borrowed (definitely) and income (possibly) have non-linear effects on the default propensity.

Models 3-6 are GAMs. Model 3 assumes a flexible effect of log(amount borrowed). On the basis of the AIC this model is hugely superior to models 1 and 2, suggesting that a flexible specification for this variable is essential.

[3] AIC is Akaike's Information Criterion, defined as 2(k-LogL) where k is the number of parameters being estimated and LogL is the maximized log-likelihood. The preferred model is the one with the *lowest* AIC.

Model 4 assumes a flexible effect of log(income). We see that this model is inferior to both of the simple probit models, suggesting that the effect of log(income) may be linear after all. Model 5 assumes a flexible effect of age. This model is superior to Model 1, but inferior to Model 2. It is not completely clear whether the effect of age should be modelled flexibly.

Finally, Model 6 assumes flexible effects for both log(amount borrowed) and age. Once again using AIC, this model is the best performer of the six models, and its superiority over Model 3 vindicates the assumption of a flexible effect for age as well as for log(amount borrowed).

It is well known that the coefficients on the basis functions are hard to interpret. However, it is a relatively straightforward matter to use the estimated coefficients to generate predicted probability curves against each of the variables for which flexible effects are assumed. For this purpose, all other predictors are set to representative values. The resulting plots are shown in Figure 2. These plots clearly reveal the highly nonlinear nature of the effects. In the case of log(loan amount), the most striking feature is the pronounced dip in the interior, with a minimum at a value of around 17.5, which is roughly the 85^{th} percentile of the log(loan amount) distribution. The implication is that borrowers whose amount borrowed is around this percentile may be seen to be the least likely to default.

In the case of age, we see overall a downward trend in default probability with age; it appears that older Korean borrowers are less likely to default. However, we also see hints that the effect of age takes the form of a step function: the probability of default falls very steeply (from a very high starting point) in the early 20s, but appears to level off in the late 20s. It falls again during the 30s, appearing to level off again in the 40s. It falls once again in the 50s, finally bottoming-out at ages beyond 60. Note that it has only been possible to uncover this complicated pattern by virtue of assuming a flexible effect of age, as the GAM framework has allowed us to do.

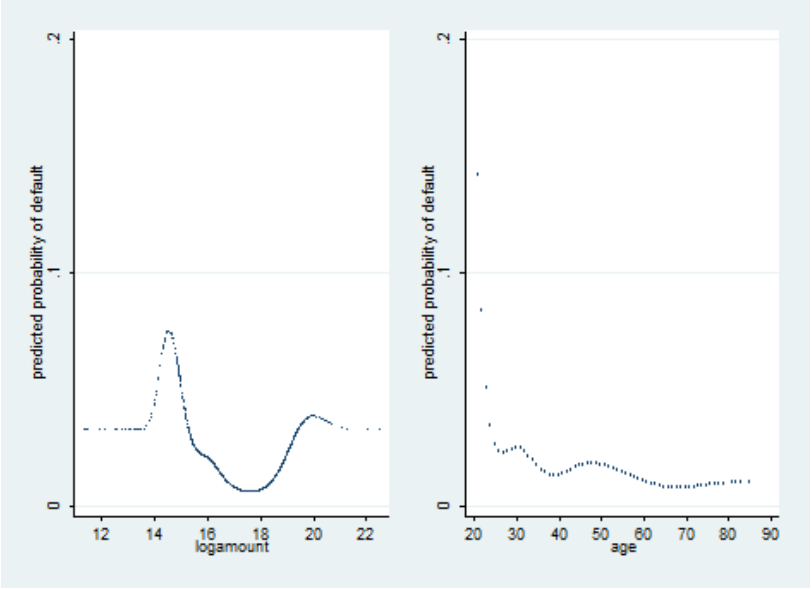

Figure 2: Predicted probability of default against logamount (left panel) and against age (right panel). All other variables set to representative values (male=1; mar=0; dep=0; pur2=1; continuous variables set to sample means).

Turning to the effects of the other predictors, we see that (ceteris paribus) male borrowers are significantly more likely to default, while marital status and number of dependants are apparently unimportant. Purpose of loan is important, with those borrowing for "accommodation" purposes being more likely to default than those borrowing for "other" purposes, and those borrowing for "living expenses" purposes being significantly more likely to default.

5. **Conclusion**

We have proposed the GAM approach for modelling the probability of loan default. The B-spline smoother is the chosen method for modelling the flexible effects. A major advantage of this approach that has been stressed is the straightforwardness of estimation and testing, leading to a reliable and unambiguous strategy for model selection. Because models estimated within the framework differ widely in terms of the number of parameters estimated, it is important to make adjustments for this in the model selection process. For this reason, we have adopted Akaike's Information Criterion (AIC) as the principal model selection criterion.

The estimation framework has been applied to loan default data on a sample of Korean borrowers. One important finding is that flexible effects should not be routinely assumed for all continuously-distributed predictors. Flexible effects should only be assumed if they are seen to give a better fit (even adjusting for the number of parameters estimated). There is an example in the results presented in Section 4: we found that the assumption of a flexible effect of income was counter-productive since an assumption of a linear income effect led to a better fit.

The two variables that did appear to have non-linear effects were amount borrowed and age of borrower. The nature of these non-linear effects became very clear when plots of predicted default probability were obtained from the estimation results. These plots conveyed interesting findings regarding the types of borrower most likely or least likely to default, and these sorts of findings may be useful in future scorecard construction.

APPENDIX

	(1) default	(2) default	(3) default	(4) default	(5) default	(6) default
logamount	-0.193***	-1.163***		-0.191***	-0.193***	
	(-9.13)	(-3.92)		(-8.97)	(-9.10)	
loginc	-0.311***	1.267	-0.275***		-0.311***	-0.275***
	(-11.42)	(1.63)	(-9.72)		(-11.36)	(-9.64)
age	-0.00774**	-0.0209	-0.00676**	-0.00745**		
	(-3.10)	(-1.30)	(-2.66)	(-2.94)		
male	0.213***	0.219***	0.226***	0.229***	0.213***	0.227***
	(4.84)	(4.96)	(5.07)	(5.15)	(4.83)	(5.05)
mar	-0.0451	-0.0282	-0.0269	-0.0384	-0.0147	0.00448
	(-0.86)	(-0.51)	(-0.50)	(-0.73)	(-0.26)	(0.08)
dep	-0.0102	-0.00895	-0.00920	-0.00936	-0.0120	-0.0108
	(-0.90)	(-0.78)	(-0.80)	(-0.82)	(-1.05)	(-0.94)
Pur1	0.345*	0.253	0.320*	0.351*	0.342*	0.318*
	(2.33)	(1.67)	(2.00)	(2.36)	(2.29)	(1.98)
Pur2	0.599***	0.610***	0.628***	0.605***	0.590***	0.618***
	(10.18)	(10.19)	(10.37)	(10.22)	(10.00)	(10.20)
logamount2		0.0308***				
		(3.30)				
loginc2		-0.0477*				
		(-2.04)				
age2		0.000146				
		(0.82)				
bsam1			0.500*			0.487*
			(2.12)			(2.06)
bsam2			0.331			0.331
			(1.93)			(1.92)
bsam3			-0.188			-0.193
			(-1.07)			(-1.10)
bsam4			-0.145			-0.144
			(-0.88)			(-0.87)
bsam5			-0.611**			-0.618**
			(-3.05)			(-3.07)
bsam6			-0.766**			-0.768**
			(-2.90)			(-2.91)
bsam7			0.198			0.188
			(0.53)			(0.51)
bsinc1				0.136		
				(0.37)		

bsinc2				0.0975 (0.39)		
bsinc3				-0.0762 (-0.30)		
bsinc4				-0.515* (-2.19)		
bsinc5				-0.509 (-1.91)		
bsinc6				-0.795** (-2.79)		
bsinc7				-0.676 (-1.44)		
bsage1					3.761** (2.71)	3.635** (2.59)
bsage2					0.0703 (0.12)	-0.0239 (-0.04)
bsage3					0.664 (1.18)	0.570 (1.02)
bsage4					0.0753 (0.14)	-0.0164 (-0.03)
bsage5					0.358 (0.64)	0.283 (0.51)
bsage6					0.275 (0.53)	0.191 (0.37)
bsage7					-0.124 (-0.18)	-0.167 (-0.24)
_cons	5.971*** (12.07)	0.835 (0.13)	2.355*** (4.70)	0.931* (2.19)	5.361*** (7.34)	1.857* (2.56)
LogL	-2257.0	-2249.8	-2211.6	-2253.1	-2248.0	-2203.0
n	32310	32310	32310	32310	32310	32310
k	9	12	15	15	15	21
AIC	4531.9	4523.7	4453.2	4536.1	4526.1	4448.0

t statistics in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table A1: Results of six binary probit models of loan default. Model 1: linear effects only. Model 2: quadratic effects. Model 3: GAM with flexible effect of log(amount borrowed) (knots: 14.9,15.4,16.1,16.8,18.1). Model 4: GAM with flexible effect of log(income) (knots: 15.9,16.8,17.3,17.8,18.2). Model 5: GAM with flexible effect of age (knots: 23,30,38,44,55). Model 6: GAM with flexible effects of both log(amount borrowed) and age. AIC = 2(k-LogL) (where k is the number of parameters) is a measure of model fit. The best-fitting model is the one with the lowest AIC.

Variable	Obs	Mean	Std. Dev.	Min	Max
default	32,310	.0151037	.1219673	0	1
logamount	32,310	16.33415	1.025021	11.35041	22.51503
loginc	32,310	17.24473	.7178109	12.67608	21.05973
male	32,310	.7125967	.4525583	0	1
age	32,310	44.23278	9.107085	21	85
mar	32,310	.8011452	.3991447	0	1
dep	32,310	3.115816	1.759105	0	10
pur1	32,310	.0502631	.2184907	0	1
pur2	32,310	.6560508	.4750317	0	1

Table A2: Descriptive statistics of all variables used.

Variable name	Definition
default	1 if borrower defaulted; 0 otherwise
logamount	natural log of loan amount
loginc	natural log of annual income
male	1 if male; 0 if female
mar	1 if married; 0 otherwise
dep	number of dependants
pur1	1 if purpose of loan is accommodation; 0 otherwise
pur2	1 if purpose of loan is living expenses; 0 otherwise
pur3	1 if purpose of loan is “other” (used as base case)

Table A3: Definitions of all variables used

References

De Boor C., 2001, *Practical Guide to Splines*, Berlin: Springer Verlag.

Fox J., 2002, *Nonparametric Regression, Appendix to an R and S-Plus Companion to Applied Regression*. London: Sage Publications.

Greene, W. (1998). Sample selection in credit-scoring models. *Japan and the world economy*, *10*(3), 299-316.

Hastie T.J. and Tibshirani R.J., 1990, *Generalized Additive Models*, New York: Chapman and Hall.

Newson R., 2000, sg151: B-splines and splines parameterized by their values at reference points on the X-axis. *Stata Technical Bulletin* 57: 20-27.

Stone, C.J., 1985, Additive regression and other nonparametric models. *The Annals of Statistics*, 13(2), 689–705.

Taylan P., Weber G.W. and Beck A., 2007, New Approaches to Regression by Generalized Additive Models and Continuous Optimization for Modern Applications in Finance, Science and Technology. *Optimization* 56, 675-698.