Journal of Analysis and Applications
Vol. 21 (2023), No.2, pp.113-122
ISSN: 0972-5954

URL : www.sasip.net

Estimability of semiparametric switching meta-regression model*

I.A. Mangaya-ay

Abstract. The primary objective of this study is to assess the estimability of the proposed model, referred to as the mixed semiparametric switching meta-regression model. The model is expressed as a Generalized Additive Model (GAM), which accommodates a high-dimensional set of covariates not usually considered in metadata. Furthermore, the applicability of the mixed semiparametric switching meta-regression model in meta-analysis settings is established in this study.

AMS Subject Classification (2020): 03C30, 03C65

Keywords: Meta-regression, mixed-effect model, switching regression, Semi-parametric

1. Introduction

The two popular statistical models for meta-regression are the fixed effect model, which is the most commonly used [8], and the random effects model [1]. This study postulates a mixed-effect model that contains both fixed and random effects, which is not commonly done in meta-analysis.

The model also incorporates switching regression, which is not usually dealt with in many meta-regression studies but may provide higher modelling flexibility. Switching regression accounts for the possibility that the study sample may be grouped into two (possibly more groups) and based on a classifier, regression equation may switch from one group to another.

*This research is supported by Bohol Island State University for the resources and financial assistance provided, which are necessary for the conduct and completion of this research work.

The main objective of this study is to evaluate the estimability of the proposed model mixed semiparametric switching meta-regression model expressed as a Generalized Additive Model (GAM) for metadata with a high dimensional set of covariates.

2. Preliminaries

2.1. Generalized Additive Model

The classical linear regression model is of the form

$$\mathbf{y} = \boldsymbol{\alpha} + \boldsymbol{X}^T\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{1}$$

For n observations $y = (y_1, y_2, ..., y_n)^T$ is the target variable (or dependent) and $\boldsymbol{X} = \boldsymbol{x_1}, \boldsymbol{x_2}, .., \boldsymbol{x_p}$ is an nxp matrix containing vector of p predictor variables. Moreover, $\boldsymbol{\beta} = \boldsymbol{\beta_1}, ..., \boldsymbol{\beta_p}$ consists a p dimensional vector of coefficients that corresponds to each of the predictor variables, $\boldsymbol{\alpha}$ is the intercept and $\boldsymbol{\epsilon}$ is the error term.

The error term ϵ is assumed to be identically and independently normally distributed with zero mean and constant variance. Aside from these assumptions, the limitations of the basic linear model also include linearity in the regression equation [4]. The classical model is usually estimated using ordinary least square method.

A larger class of model that is popularized by McCullagh and Nelder [10] is the Generalized Linear Model (GLM). It is still of the form (1) but the assumption does not limit the distribution of $\boldsymbol{y}$ to a normal distribution but can be any member of the exponential family of distributions (e.g. Gaussian, Bernoulli, Poisson, Gamma).

The use of GLM is reasonable since a normal distribution, which is a continuous distribution, is often inadequate when modelling count data (e.g. proportions, presence or absence of characteristic and frequencies).

Furthermore, it is superficial to always assume that the variance of the data is constant in all observations.

In a GLM, the predictor variables $\boldsymbol{x_j}, j = 1, 2, \cdots, p$ are linearly combined to have a predictor that is related to the expected value $\boldsymbol{\mu} = E(\boldsymbol{y})$ of the response variable y through a link function $g()$ [4]. The link function for a linear regression model is

$$g(E(\boldsymbol{y})) = \boldsymbol{\alpha} + \boldsymbol{X}^T\boldsymbol{\beta} \tag{2}$$

Aside from the link function, GLM also depends on a variance function that describes how the variance of y depends on the mean, e.g.

$$var(\boldsymbol{y}) = \phi V(\boldsymbol{\mu}) \tag{3}$$

where ϕ is a constant dispersion parameter. Though GLM does not assume a direct linear relationship between the response and the predictor variable, it does assume linear relationship of the transformed responses in terms of the link function and the linear predictor variables.

GLM is usually estimated using an iterative reweighted least squares [5], which reduces to MLE with additional assumptions. See [10] for further details.

The class of GLM is further generalized into Generalized Additive Model (GAM). It replaces the linear predictors $\eta = \sum_{i=1}^{p} \boldsymbol{\beta_j x_j}$ with additive smooth functions $\eta = \sum_{i=1}^{p} \boldsymbol{s_j x_j}$. The response variable y has an exponential distribution with mean

$$\boldsymbol{\mu} = E(\boldsymbol{y}|\boldsymbol{x_1}, ..., \boldsymbol{x_p}) \tag{4}$$

and the link function

$$g(\boldsymbol{u}) = \boldsymbol{\alpha} + \sum_{j=1}^{p} \boldsymbol{f_j x_j} \tag{5}$$

subject to

$$E[\boldsymbol{f_j}(\boldsymbol{x_j})] = \mathbf{0} \tag{6}$$

Hastie and Tibshirani [5] used GAM as an alternative model different from the usual parametric function to mitigate nonlinearities. In addition, GAM does not assume rigid dependence of the response variable and the predictor variables. It is a nonparametric model that let the data decide on functional form.

In estimating the GAM, Hastie and Tibshirani [5] used local scoring algorithm, a general form of the iterative reweighted least squares algorithm for solving likelihood and nonlinear regression equations. This algorithm estimates the functions $\boldsymbol{f_j}(\boldsymbol{x_j})$ nonparametrically using a scatterplot smoother. Since GAM is flexible, it has the tendency to overfitting.

A conservative degrees freedom of the fitted smooth can be used so that overfitting can be prevented [2]. Low degrees of freedom of the smooth can also reduce the computation cost [7].

A smoother is a means of summarizing the relationship between the dependent variable $\boldsymbol{y}$ and one or more independent variables using the local average of the observations (x_i, y_i) [5].

A scatterplot smooth of the data $(\boldsymbol{x_j}, \boldsymbol{y})$ at the point $\boldsymbol{x_i}$ can be viewed as an estimate of $E(\boldsymbol{y}|x_i)$ denoted by $\boldsymbol{S}(\boldsymbol{y}|x_i)$. Buja et al. [2] have proven convergence in the estimation algorithm if the smoothers are linear, symmetric and shrinking, that is, $\hat{\boldsymbol{y}} = \boldsymbol{S}\boldsymbol{y}$ for some nxn matrix and the eigenvalues of $\boldsymbol{S}$ have absolute values not larger than 1. The cubic smoothing splines possess these characteristics and it is the solution to the minimization problem that follows:

$$min_f \sum_{n}^{i} (y_i - f(x_i))^2 + \lambda_a^b [f^{''}(x)]^2 d_x \quad a \leq x_1 \leq ...x_n \leq b \tag{7}$$

Here $f^{'}$ is absolutely continuous and $f^{"} \in L_2$, and λ is a fixed tuning constant. The first term represents the least square criterion and it measures departure of the model from the data.

On the other hand, the second term measures the amount of smoothness in the model. The smoothing parameter λ manages the tradeoff between the two terms, i.e., it controls the goodness of fit and the curvature of the function.

Cubic splines used in the backfitting algorithm converge to a solution of the penalized least squares problem:

$$min_f \sum_{i=1}^{n}(y_i - \sum_{j=1}^{p} f_j((x_{ji})^2) + \sum_{j=1}^{p} \lambda_j [f^{''}(x)]^2 d_x \tag{8}$$

among functions defined in a Sobolev space.

2.2. The Backfitting algorithm

The backfiiting algorithm or Gauss-Seidel algorithm is a process of estimating (5) using the following steps [5]:

(i) Initialize: $\hat{\boldsymbol{f}}_{\boldsymbol{j}} = \boldsymbol{f}_{\boldsymbol{j}}^{\boldsymbol{0}}, j = 1, 2, ..., p$

(ii) Cycle: $j = 1, 2, ..., p$
$\hat{f}_j = S_j(\boldsymbol{y} - \boldsymbol{a}) - \sum_{k \neq j} \hat{\boldsymbol{f}}_k | x_j)$

(iii) Continue (ii) until individual functions don't change.

where S_j is the smoothing operator.

The algorithm estimates each smooth function while other functions fixed. Buja et al. [2] showed that there is no need to worry of until in the last step (iii) in the backfitting algorithm because the right choice of smoother guarantees convergence.

2.3. The proposed semiparametric switching meta-regression model

$$\boldsymbol{y}^1 = \sum_{j=1}^{p} \beta_j^1 \boldsymbol{x_j} + \boldsymbol{\mu} \quad if \quad h(\boldsymbol{u}) \geq b \tag{9}$$

and

$$\boldsymbol{y}^2 = \sum_{j=1}^{p} \beta_j^2 \boldsymbol{x_j} + \boldsymbol{\mu} \quad if \quad h(\boldsymbol{u}) < \boldsymbol{b} \tag{10}$$

here $h(\boldsymbol{u}) = \boldsymbol{\delta} + \boldsymbol{\epsilon}$ and $\boldsymbol{\delta} = \boldsymbol{g(w)}$.

Here $\boldsymbol{y}^d$ are the effect estimates of $\boldsymbol{d}^{th}$ regime, $\boldsymbol{x}_j^d$ are study covariates (discrete or continuous), $\boldsymbol{\beta}_j^d$ are regression coefficients, and $\boldsymbol{u}$ is the random component with two sources of variation-the within study error ($\boldsymbol{\epsilon}$), which is a random imprecision of estimates ; and $\delta = g(w)$, which are explanatory variables related to study selection.

3. Results

Theorem 3.1 (Estimability of the proposed model). *Under a known regime, the proposed model is estimable via a cubic smoothing spline smoother for the fixed component and the REML for the random component.*

Proof. The penalized least square that leads to cubic smoothing spline minimizes the following:

$$\sum_{i=1}^{n} (y_i - g((x_i))^2 + \lambda \int_{-\infty}^{\infty} [g(x)]^2 d_z \tag{11}$$

with $g^{'}$ absolutely continuous and $g^{''} \in L_2$. □

According to Green and Yandell [3], (11) can be represented by basis functions, e.g., quadratic function. The objective function in (11) then reduces to

$$argmin_g Q(g) = ||\boldsymbol{y} - g(\boldsymbol{x})||^2 + \lambda g^T(\boldsymbol{x}) \boldsymbol{K} g(\boldsymbol{x}) \tag{12}$$

where $K = \Delta^t C^{-1} \Delta$. Here, Δ is a tridiagonal matrix with

$$\Delta_{ii} = \frac{1}{h_i}, \Delta_{i,i} = -\left(\frac{1}{h_i} + \frac{1}{h_{i+1}}\right), \Delta_{i,i+2} = \frac{1}{h_{i+1}} \tag{13}$$

Also, C is a symmetric tridiagonal matrix with order $n-2$ with

$$C_{i-1,i} = C_{i,i-1} = \frac{h_i}{6}, C_{ii} = (h_i + h_{i+1}) \tag{14}$$

and h_i is the binwidth.

Assuming the inverses exist, Green and Yandell [3] derived the solution to (12) as:

$$\hat{g}(\boldsymbol{x}) = (1 + \lambda K)^{-1}.\boldsymbol{y} \tag{15}$$

For a linear smoother $\boldsymbol{S}$, $\hat{\boldsymbol{g}}$ can be written in the form $\hat{\boldsymbol{g}} = \boldsymbol{S}\boldsymbol{y}$ [2].

Hence, (15) can be expressed as:

$$\hat{g}(\boldsymbol{x}) = S\boldsymbol{y} \tag{16}$$

where

$$\boldsymbol{S} = (\boldsymbol{I} + \lambda \boldsymbol{K})^{-}1 \tag{17}$$

Provided that symmetric invertible $\boldsymbol{S}$ exist, $\hat{\boldsymbol{g}}$ can be characterized as stationarity solution of

$$argmin_g Q(g) = ||\boldsymbol{y} - g(\boldsymbol{x})||^2 + \lambda g^T(\boldsymbol{x})[\boldsymbol{S}^{-1} - \boldsymbol{I}]g(\boldsymbol{x}) \tag{18}$$

For a smoother with positive real eigenvalues, the solution exists since $Q(g) \geq 0$ [2].

According to Buja et al. [2], the penalized constrained least square approach can be extended to additive regression by penalizing the residual sum of square separately for each component function,

$$Q(g) = ||y - \sum_{j=1}^{p} g_j(\boldsymbol{x})||^2 + \sum_{j=1}^{p} g_j^T(\boldsymbol{x})(\boldsymbol{S}_j^-) - I)g_j(\boldsymbol{x}) \tag{19}$$

and the solution still exists and is computed via the backfitting algorithm.

Furthermore, Buja et al. [2] were able to show convergence of a backfitting algorithm.

Now, suppose $\hat{\delta}$ is estimated as $\hat{\delta}$ in (10) for a given a single regime $d = 1, 2$, then

$$\boldsymbol{y}^* = \boldsymbol{y} - \hat{\boldsymbol{\delta}} = \sum_{l=1}^{m} f_l(v_l) \tag{20}$$

Since (10) is additive and is computed for separate regimes, $\boldsymbol{y}^*$ can be $\boldsymbol{y}$ in (19), i.e., the solution of

$$argmin_f Q(f) = ||y - \sum_{l=1}^{m} f_l^{(d)}(\boldsymbol{v}_l - \delta)||^2 + \sum_{l=1}^{m} f_l^{(d)T}(\boldsymbol{S}_l^- - I) f_(^d l) \tag{21}$$

exists.

The solution $\hat{\delta}$, is determined by REML through an iterative backfitting algorithm. Existence of $\hat{f}$ is also guaranteed since the modified backfitting algorithm uses cubic smoothing spline which is linear, symmetric and has positive real eigenvalues.

4. Conclusion

Using cubic splines to smooth the fixed nonparametric functions and REML to estimate random component, the proposed model is theoretically proven to be estimable.

Acknowledgement. The author would like to express her sincere gratitude to the referees for their valuable suggestions and comments which improved the paper.

References

[1] M. Borenstein, L.V. Hedges, J.P. Higgins and H.R. Rothstein, *A ba-*

sic introduction to fixed-effect and random-effects models for meta-analysis. Research synthesis methods, 1 (2010), 97-111.

[2] A. Buja, T. Hastie and R. Tibshirani, *Linear smoother and additive models*, The Annals of Statistics, 17 (1989), 453-510.

[3] P. Green and B. Yandell, *Semi-parametric generalized linear models*, Generalized Linear Models, Lecture Notes in Statist.. 32 (1985), 44-55, Springer Verlag, Berlin.

[4] A. Guisan, T.C. Edwards, Jr, T. Hastie, *Generalized linear and generalized additive models in studies of species distributions: setting the scene*, Ecological modelling, 157 (2002), 89-100.

[5] T. Hastie and R. Tibshirani, *Generalized additive models : some applications*, J. Amer. Statis. Assoc. 82 (1987), 371-386. doi: 10.2307/2289439

[6] T. J. Hastie and R.J. Tibshirani, Generalized Additive Model. New York : Chapman and Hall (1990).

[7] W. Liu, C. Vu and J. Cela, *Generalizations of generalised additive models (GAM): a case of credit risk modelling*, Conference Paper. Washington: SAS Program (2009).

[8] R. Overton, *A comparison of fixed-effects and mixed (random-effects) Models for Meta-Analysis Tests of Moderator Variable Effects*, Psychological Methods, 3 (1989), 354-379. doi: 101037

[9] R. Tibshirani and T. Hastie, *Local likelihood estimation*, J. Amer. Assoc., 82 (1987), 559-568. doi:10.2307/2289465

[10] J.A. Nelder and R.W. Wedderburn, *Generlized Linear Models.*, J. Royal Statistical Society, Series A (General), 135 (1972), 370-384. Retrieved from http://www.jstor.org/stable/2344614

Department of Mathematics and Natural Sciences
College of Arts and Sciences
Bohol Island State University
CPG Avenue, 6300 Tagbilaran City
Bohol, Philippines
E-mail: ivycorazon.mangayaay@bisu.edu.ph

(*Received: May, 2023; Revised: June, 2023*)