

Estimation of GLMs with unspecified link function: a P-spline approach

*Stima dei GLM con funzione legame non specificata:
un approccio con P-splines*

Giancarlo Ferrara, Vito M. R. Muggeo, Anna Maria Taormina

Dipartimento di Scienze Statistiche e Matematiche “S. Vianelli”, Università di Palermo
e-mail: ferrara@dssm.unipa.it

Riassunto: I modelli lineari generalizzati (GLM) costituiscono un’ampia classe di modelli statistici nei quali l’effetto delle variabili esplicative sulla risposta viene modellato attraverso la specificazione del predittore lineare, della funzione *link* e della funzione di varianza. Solitamente, nelle applicazioni poca importanza è data alla selezione della funzione *link* e quasi sempre la scelta ricade sui *link* canonici. In questo lavoro viene proposta la stima dei GLM con funzione link non specificata e modellata mediante B-splines con penalizzazione sui relativi coefficienti (P-splines): viene proposto un algoritmo di stima ed i risultati illustrati attraverso l’analisi di un dataset reale.

Keywords: generalized linear models, link function, B-splines, penalized likelihood.

1. An introduction to GLMs and notation

Generalized Linear Models (GLMs) constitute a unified representation of a wide class of models aimed to investigate the effect of explanatory variables on a ‘response’ variable of main interest. Formally, given a set of p explanatory variables $X_1, X_2, \dots, X_p$ and a response variable Y , the specification of GLMs requires the choice of a distribution (within the exponential family) for the random variable Y and of a *link* function $g(\cdot)$ relating $\mu_i = E(Y_i|x_i)$ to the linear predictor $\eta_i = x_i'\beta$ (McCullagh and Nelder, 1989), namely: $g(\mu_i) = \eta_i$ for $i = 1, \dots, n$ sample units. In practice the choice of the link function is usually overlooked, and the canonical link is generally selected, the reason being the simple interpretation of the model parameters. However, linear relations imposed in linear predictor η and/or linearity between $g(\mu_i)$ and η_i are not guaranteed to provide always the best fit to the data. Generalized Additive Models (GAMs) generalize GLMs replacing the linear predictor η by an additive predictor given by a sum of smooth functions of the explanatory variables in the form: $\eta = \beta_0 + \sum_{j=1}^p f_j(X_j)$, in which the $f_j(\cdot)$ s are nonparametric smooth functions estimated from the data. However in the GAM framework the relationship ‘fitted values - additive predictor’ is still linear and the smoothed variables X_j have to be *continuous*.

A further generalization of GLMs is to assume a smooth unspecified relation between $g(\mu)$ and η , through:

$$g(\mu_i) = \psi(\eta_i) = \psi(x_i'\beta) \quad (1)$$

for some smooth function $\psi(\cdot)$. Thus, model (1) leaves the parametric component $\eta_i = x_i'\beta$, and introduces the function $\psi(\cdot)$ to relax the linearity assumption; when

$\psi(\cdot)$ is the identity function, model (1) reduces to standard GLM with link function $g(\cdot)$. Therefore, model (1) is a generalized linear model with an unknown or unspecified link function. The problem of the “missing link” in GLMs has already been considered by Weisberg and Welsh (1994); such problem is also known in econometrics as ‘Single Index Model’ (SIM) where the linear predictor η_i is referred as the single index (Ichimura, 1993; Carroll *et al.*, 1997). The gain by using this model is that it avoids the so-called “curse of dimensionality” by focusing the attention on a univariate quantity, the single-index, rather than the single covariates X_j ; indeed, the appeal of this model is that it represents a compromise between the fully parametric models provided by GLMs and the fully non parametric models given by GAMs. An alternative approach to the problem of the “missing link” is supplied by assuming a family of link functions which depends on one or more parameters. For instance, the power link family is specified as: $\eta = \mu^\kappa$ if $\kappa > 0$, and $\eta = \log(\mu)$ if $\kappa = 0$. Possible drawbacks associated with the power link functions include the interpretation of model parameters.

In this work we will consider the estimate of model (1) with unspecified or unknown $\psi(\cdot)$ using parametric B-splines with penalty. In Section 2 we give a brief introduction to B-splines and penalties and present an estimating algorithm for model (1); Section 3 includes an application and conclusions are reported in the last section.

2. P-splines: a summary

For sake of simplicity, consider the simple regression of Y on X and let (x_i, y_i) the n observed data points: we are interested in finding a smooth function $\psi(\cdot)$ such that $E[Y_i|x_i] = \mu_i = \psi(x_i)$. There are several smoothing techniques in literature, like nonparametric smoothers (including loess, kernel and smoothing splines) or parametric B-splines. A B-spline basis of order q on X is a matrix $\mathbf{B} = [B_1(x), \dots, B_q(x)]$ constructed from polynomial pieces of a specified degree, joined at certain values within the range of X , the knots, and such that the fitted values are given by $\mu_i = \sum_j^q B_j(x_i)b_j$, or in matrix notation $\boldsymbol{\mu} = \mathbf{B}\mathbf{b}$. While the degree of the B-spline is a minor issue, the choice of knots is a difficult task since too many (too few) knots lead to undersmoothing (oversmoothing); besides the number of knots, also their location is a major point. To overcome this problem and to limit the influence of knots on fitted values, Eilers and Marx (1996) suggest to select a large number of knots and to set some constraints on the relevant coefficients: this approach is known as penalized splines, or P-splines. They propose to introduce a penalty based on (higher-order) finite differences of the coefficients of adjacent B-splines, in the form: $\lambda\{(b_1 - b_2)^2 + (b_2 - b_3)^2 + \dots\}$ or $\lambda\{(b_1 - 2b_2 + b_3)^2 + (b_2 - 2b_3 + b_4)^2 + \dots\}$ for first- or second-order differences, respectively. Here $\lambda > 0$ is the penalty (or smoothing) parameter known or to be selected by some criterion, such as AIC, BIC or GCV equal to $\|\mathbf{y} - \hat{\boldsymbol{\mu}}\|^2 \{1 - \text{tr}(\mathbf{S}_\lambda)\}^{-2}$ where $\mathbf{S}_\lambda$ is the hat matrix. In matrix notation the (penalized) objective function becomes $\|\mathbf{y} - \mathbf{B}\mathbf{b}\|^2 + \lambda \mathbf{b}'\mathbf{D}'\mathbf{D}\mathbf{b}$, where the second term is the roughness penalty which penalizes fits that are too rough. $\mathbf{D}$ is the matrix representation of the difference operator.

In model (1) we replace $\psi(\eta)$ by the B-spline smoother, therefore the model becomes

$$g(\mu_i) = \psi(\eta_i) = \psi(\mathbf{x}_i'\boldsymbol{\beta}) = \mathbf{B}(\mathbf{x}_i'\boldsymbol{\beta})\boldsymbol{\delta} \quad (2)$$

where $\boldsymbol{\delta}$ is the q -dimensional vector of the coefficient of the B -spline. Therefore the

model parameters are β and δ as well, and the (penalized) log-likelihood may be written as $l^*(\beta, \delta) = l(\beta, \delta) - \lambda \delta' D' D \delta$ for some proper known matrix D . To maximize the log-likelihood note that when δ or β are known the model reduces to a specific GLM, hence an estimating algorithm is:

1. fit a standard GLM $g(\mu_i) = \mathbf{x}_i' \beta$;
2. known β and $\eta_i = \mathbf{x}_i' \beta$, fit a GLM $g(\mu_i) = \delta' B(\eta_i)$;
3. compute the working response $z_i = \eta_i + (y_i - \mu_i) / \frac{\partial \mu_i}{\partial \eta_i}$, the weights $w_i = (\frac{\partial \mu_i}{\partial \eta_i})^2 / e^{\eta_i}$ and fit the weighted linear model $z_i = \mathbf{x}_i' \beta$ with weights w_i ;
4. repeat step 2. and 3. up to convergence.

Note that the aforementioned algorithm is a modified version of the IWLS for GLMs with link function $g(\cdot)$; the ‘only’ difference is that $\frac{\partial \mu}{\partial \eta}$ depends on the derivative of the B-spline.

3. Application

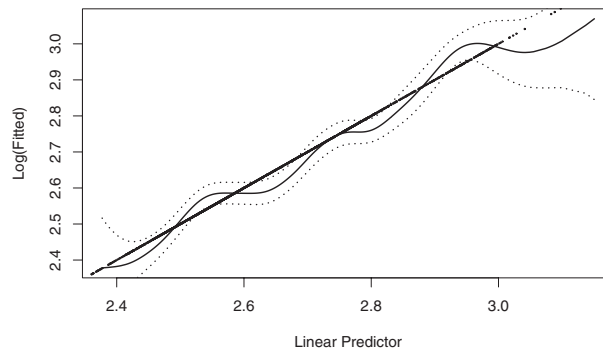
We have applied the model to the daily time series data of mortality and meteorological variables in Palermo, recorded 1997 to 2001. The goal of the analysis is to quantify the mortality excesses due to air pollution (particulate matter, PM_{10}) and temperature, especially the heat defined as temperature values greater then $25^\circ C$. There is an extensive epidemiological literature investigating the health effects of environmental factors, and standard analysis have made use of GLM with log-link to obtain risk estimates related to PM_{10} and temperature.

Table 1: Point estimates ($\times 10$) and relevant t -values (in parenthesis) for the effect of heat and PM_{10} on natural mortality in Palermo, 1997-2001.

Term	Power link*	Log link	Nonparam. link
PM_{10}	-3.26e-04 (2.93)	0.011 (2.86)	0.012 (5.10)
Heat	-1.45e-02 (6.29)	0.388 (5.53)	0.410 (8.11)

* $\hat{\kappa} = -1.388$

Figure 1: $\log(\text{Fitted values})$ versus linear predictor in two considered regression models. Straight line: Standard Poisson model; Curve line with relevant pointwise confidence intervals: unspecified link model.



Here we investigate to which extend the ‘linear’ assumption between $\log(\mu)$ and η is questionable. For comparisons, we have fitted three regression models: i) standard Poisson GLM with log link; ii) GLM with power link function and iii) GLM with nonparametric link (model (2)). In each case the linear predictor includes ‘seasonality’ (via natural spline with 5 degree of freedom per year), influenza epidemics, day-of-week, PM_{10} , and temperature by means of two complementary linear terms to account for cold and heat. Both air pollution and temperature have been evaluated as lag 0 series.

Table 1 displays point estimates (and relevant t values) for the parameter of main interest, PM_{10} and heat and Figure 1 shows the plot of $\log(\text{fitted})$ versus linear predictor according to the log-link and nonparametric-link models; the confidence intervals for model (2) reflect the uncertainty in the $\hat{\delta}$ estimate. Although departure from linearity is not so impressive, the linear assumption does not appear fully reliable and somewhat important differences may be recognized in Table 1; also, results from the power link model emphasize that the canonical log-link is not supported by data ($\hat{\kappa}$ is far from zero). Here note the opposite sign of coefficients due to the negative value of the exponent κ .

4. Conclusions

In this paper we have discussed GLMs with unspecified link function: unlike previous works published in literature we have used B-spline with penalty selected by GCV; parametric splines offer several computational advantages with respect to the nonparametric smoothers and penalty is aimed to avoid undersmoothing in data. We have suggested a ‘see-saw’ estimating algorithm, alternating estimation of two regression models to update the estimates of β and δ . From a practical viewpoint nonparametric link models may turn out quite useful to check the linear assumption while gaining new insights in phenomenon under study. Although not addressed in this paper, possible generalizations of such model include partial and/or multiple index model such as $g(\mu_i) = \psi_1(x'_{1i}\beta_1) + \psi_2(x'_{2i}\beta_2) + z'_i\gamma$.

References

- Carroll R. J., Fan J., Gijbels I., Wand M.P. (1997) Generalized Partially Linear Single-Index Models, *Journal of the American Statistical Association*, 92, 477–489.
- de Boor C. (2001) *A Practical Guide to Splines*, Springer.
- Eilers P. H. C., Marx B. D. (1996) Flexible Smoothing with B-splines and Penalties, *Statistical Science*, 11, 89–121.
- Kaiser M. S. (1997) Maximum Likelihood Estimation of Link Function Parameters, *Computational Statistics & Data Analysis*, 24, 79–87.
- Ichimura H. (1993) Semiparametric Least Squares (SLS) and Weighted SLS Estimation of Single-Index Models, *Journal of Econometrics*, 58, 71–120.
- McCullagh P., Nelder J. A. (1989) *Generalized Linear Models*, Chapman & Hall.
- Simonoff J.S. (1998) *Smoothing Methods in Statistics*, Springer.
- Weisberg S., Welsh A. H. (1994) Adapting for the Missing Link, *The Annals of Statistics*, 22, 1674–1700.