

### 3.1 Generalized additive models (GAM's)

**Model description** A very useful generalization of the ordinary multiple regression

$$y_i = \mu + \beta_1 x_{1,i} + \cdots + \beta_p x_{p,i} + \varepsilon_i,$$

is the class of additive models,

$$y_i = \mu + f_1(x_{1,i}) + \cdots + f_p(x_{p,i}) + \varepsilon_i. \quad (3.1)$$

Here, the  $f_j$  are ‘nonparametric’ components which can be modelled by penalized splines. When this generalization is carried over to generalized linear models, and we arrive at the class of GAM's (Hastie & Tibshirani 1990). From a computational perspective penalized splines are equivalent to random effects, and thus GAM's fall naturally into the domain of ADMB-RE.

For each component  $f_j$  in (3.1) we construct a design matrix  $\mathbf{X}$  such that  $f_j(x_{i,j}) = \mathbf{X}^{(i)}\mathbf{u}$ , where  $\mathbf{X}^{(i)}$  is the  $i$ th row of  $\mathbf{X}$  and  $\mathbf{u}$  is a coefficient vector. We use the R-function `splineDesign` (from the `splines` library) to construct a design matrix  $\mathbf{X}$ . To avoid overfitting we add a first order difference penalty (Eilers & Marx 1996) :

$$- \lambda^2 \sum_{k=2} (u_k - u_{k-1})^2, \quad (3.2)$$

to the ordinary GLM loglikelihood, where  $\lambda$  is a smoothing parameter to be estimated. By viewing  $\mathbf{u}$  as a random effects vector with the above Gaussian prior, and by taking  $\lambda$  as a hyper-parameter, it becomes clear that GAM's are naturally handled in ADMB-RE.

#### Implementation details

- A computationally more efficient implementation is obtained by moving  $\lambda$  from the penalty term to the design matrix, i.e.  $f_j(x_{i,j}) = \lambda^{-1}\mathbf{X}^{(i)}\mathbf{u}$ .
- Since (3.2) does not penalize the mean of  $\mathbf{u}$ , we impose the restriction that  $\sum_{k=1} u_k = 0$  (see the `union.tpl` for details). Without this restriction the model would be over-parameterized since we already have an overall mean  $\mu$  in (3.1).
- To speed up computations the parameter  $\mu$  (and other regression parameters) should be given ‘phase 1’ in ADMB, while the  $\lambda$ 's and the  $\mathbf{u}$ 's should be given ‘phase 2’.

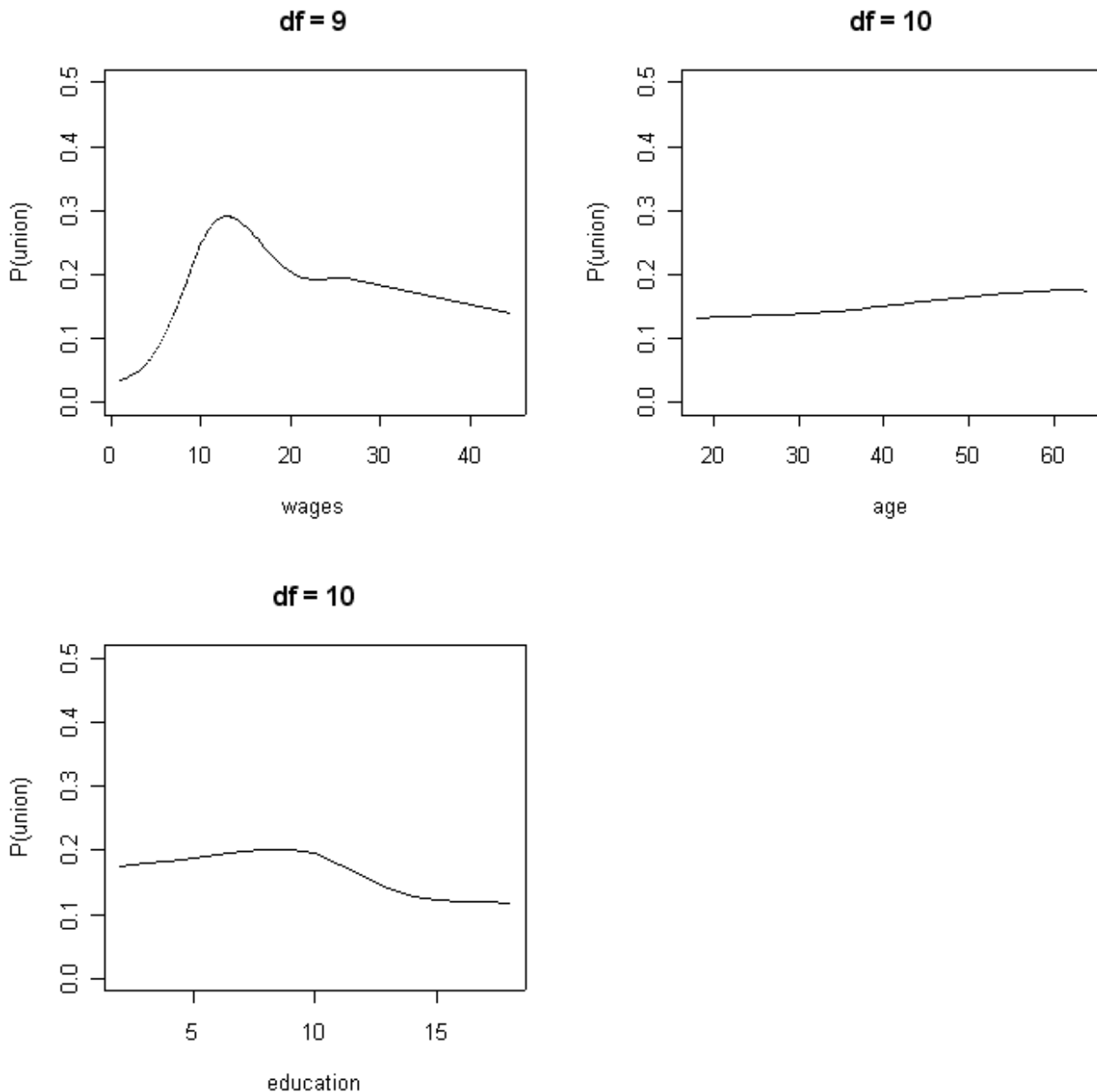


Figure 3.1: Union data: Probability of membership as a function of covariates. In each plot, the remaining covariates are fixed at their sample means. The effective degrees of freedom (df) are also given (Hastie & Tibshirani 1990).

**The Wage-union data** The data, which are available from Statlib ([lib.stat.cmu.edu/](http://lib.stat.cmu.edu/)), contain information for each of 534 workers about whether they are members ( $y_i = 1$ ) of a workers union or not ( $y_i = 0$ ). We study the probability of membership as a function of six covariates. Expressed in the notation used by the R (S-Plus) function `gam` the model is:

```
union ~race + sex + south + s(wage) + s(age) + s(ed), family=binomial
```

Here,  $\mathbf{s}()$  denotes a spline functions with 20 knots each. For **wage** a cubic spline is used, while for **age** and **ed** quadratic splines are used. The total number of random effects that arise from the three corresponding  $\mathbf{u}$  vectors is 64. Figure 3.1 shows the estimated nonparametric components of the model. The time taken to fit the model was 165 seconds.

### Extensions

- The linear predictor may be a mix of ordinary regression terms ( $f_j(x) = \beta_j x$ ) and nonparametric terms. ADMB-RE offers a unified approach to fitting such models, in which the smoothing parameters  $\lambda_j$  and the regression parameters  $\beta_j$  are estimated simultaneously.
- It is straightforward in ADMB-RE to add ‘ordinary’ random effects to the model, for instance to accommodate for correlation within groups of observations, as in Lin & Zhang (1999).