Chapter 8

Generalized Models

What happens if the response is not continuous? This the same problem that motivated the extension of linear models to generalized linear models (GLM).

In this Chapter we will discusss two method based on a likelihood approach simlar to GLMs.

8.1 Generalized Additive Models

We extend additive models to generalized additive models in a similar way to the extension of linear models to generalized linear models.

Say Y has conditional distribution from an exponential family and the conditional mean of the response $E(Y|X_1,\dots,X_p) = \mu(X_1,\dots,X_p)$ is related to an additive model through some link functions

$$g\{\mu_i\} = \eta_i = \alpha + \sum_{j=1}^{p} f_j(x_{ij})$$

with μ_i the conditional expectation of Y_i given $x_{i1}, \dots, x_{ip}$. This motivates the use of the IRLS procedure used for GLMs but incorporating the backfitting algorithms used for estimation in Additive Models.

As seen for GLM the estimation technique is again motivated by the approximation:

$$g(y_i) \approx g(\mu_i) + (y_i - \mu_i)\frac{\partial \eta_i}{\partial \mu_i}$$

This motivates a weighted regression setting of the form

$$z_i = \alpha + \sum_{j=1}^{p} f_j(x_{ij}) + \varepsilon_i, \, i = 1, \dots, n$$

with the εs, the working residuals, independent with $\mathrm{E}(\varepsilon_i) = 0$ and

$$\mathrm{var}(\varepsilon_i) = w_i^{-1} = \left(\frac{\partial \eta_i}{\partial \mu_i}\right)^2 V_i$$

where V_i is the variance of Y_i.

The procedure for estimating the function f_js is called the *local scoring procedure*:

1. Initialize: Find initial values for our estimate:

$$\alpha^{(0)} = g\left(\sum_{i=1}^{n} y_i/n\right); f_1^{(0)} = \dots, f_p^{(0)} = 0$$

2. Update:

- Construct an adjusted dependent variable

$$z_i = \eta_i^{(0)} + (y_i - \mu_i^{(0)})\left(\frac{\partial \eta_i}{\partial \mu_i}\right)_0$$

with $\eta_i^{(0)} = \alpha^{(0)} + \sum_{j=1}^{p} f_j^{(0)}(x_{ij})$ and $\mu_i^{(0)} = g^{-1}(\eta_i^{(0)})$

- Construct weights:

$$w_i = \left(\frac{\partial \mu_i}{\partial \eta_i}\right)_0^2 (V_i^{(0)})^{-1}$$

- Fit a weighted additive model to z_i, to obtain estimated functions $f_j^{(1)}$, additive predictor $\eta^{(1)}$ and fitted values $\mu_i^{(1)}$.
  Keep in mind what a fit is.... $\hat{\mathbf{f}}$.
- Compute the convergence criteria

$$\Delta(\eta^{(1)}, \eta^{(0)}) = \frac{\sum_{j=1}^p ||f_j^{(1)} - f_j^{(0)}||}{\sum_{j=1}^p ||f_j^{(0)}||}$$

- A natural candidate for $||f||$ is $||\mathbf{f}||$, the length of the vector of evaluations of f at the n sample points.

3. Repeat previous step replacing $\eta^{(0)}$ by $\eta^{(1)}$ until $\Delta(\eta^{(1)}, \eta^{(0)})$ is below some small threshold.

8.1.1 Penalized Likelihood

How do we justify the local scoring algorithm? One way is to minimize a penalized likelihood criterion.

Given a generalized additive model let

$$\eta_i = \alpha + \sum_{j=1}^p f_j(x_{ij})$$

and consider the likelihood $l(f_1, \dots, f_p)$ as a function $\boldsymbol{\eta} = (\eta_1, \dots, \eta_p)'$.

Consider the following optimization problem: Over p-tuples of functions $f_1, \dots, f_p$ with continuous first and second derivatives and integrable second derivatives find one that minimizes

$$pl(f_1, \dots, f_p) = l(\boldsymbol{\eta}; \mathbf{y}) - \frac{1}{2} \sum_{j=1}^p \lambda_j \int \{f_j''(x)\}^2 \, dx$$

where $\lambda_j \geq 0, j = 1, \dots, p$ are smoothing parameters.

Again we can show that the solution is an additive cubic spline with knots at the unique values of the covariates.

In order to find the $\mathbf{f}$s that maximize this penalized likelihood we need some optimization algorithm. We will show that the Newton-Raphson algorithm is equivalent to the local-scoring procedure.

As before we can write the criterion as:

$$pl(\mathbf{f}_1, \dots, \mathbf{f}_p) = l(\boldsymbol{\eta}, \mathbf{y}) - \frac{1}{2} \sum_{j=1}^{p} \lambda_j \mathbf{f}_j' \mathbf{K}_j \mathbf{f}_j.$$

In order to use Newton-Raphson we let $\mathbf{u} = \partial l / \partial \boldsymbol{\eta}$ and $\mathbf{A} = -\partial^2 l / \partial \boldsymbol{\eta}^2$. The first step is then taking derivatives and solving the score equations:

$$\begin{pmatrix} \mathbf{A} + \lambda_1 \mathbf{K}_1 & \mathbf{A} & \dots & \mathbf{A} \\ \mathbf{A} & \mathbf{A} + \lambda_2 \mathbf{K}_2 & \dots & \mathbf{A} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{A} & \mathbf{A} & \dots & \mathbf{A} + \lambda_p \mathbf{K}_p \end{pmatrix} \begin{pmatrix} \mathbf{f}_1^1 - \mathbf{f}_1^0 \\ \mathbf{f}_2^1 - \mathbf{f}_2^0 \\ \vdots \\ \mathbf{f}_p^1 - \mathbf{f}_p^0 \end{pmatrix} = \begin{pmatrix} \mathbf{u} - \lambda_1 \mathbf{K}_1 \mathbf{f}_1^0 \\ \mathbf{u} - \lambda_1 \mathbf{K}_1 \mathbf{f}_2^0 \\ \vdots \\ \mathbf{u} - \lambda_1 \mathbf{K}_1 \mathbf{f}_p^0 \end{pmatrix}$$

where both $\mathbf{A}$ and $\mathbf{u}$ are evaluated at $\boldsymbol{\eta}^0$. In the exponential family with canonical family, the entries in the above matrices are of simple form, for example the matrix $\mathbf{A}$ is diagonal with diagonal elements $a_{ii} = (\partial \mu_i / \partial \eta_i)^2 V_i^{-1}$.

To simplify this further, we let $\mathbf{z} = \boldsymbol{\eta}^0 + \mathbf{A}^{-1}\mathbf{u}$, and $\mathbf{S}_j = (\mathbf{A} + \lambda_j \mathbf{K}_j)^{-1}\mathbf{A}$, a weighted cubic smoothing-spline operator. Then we can write

$$\begin{pmatrix} \mathbf{I} & \mathbf{S}_1 & \dots & \mathbf{S}_1 \\ \mathbf{S}_2 & \mathbf{I} & \dots & \mathbf{S}_2 \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{S}_p & \mathbf{S}_p & \dots & \mathbf{I} \end{pmatrix} \begin{pmatrix} \mathbf{f}_1^1 \\ \mathbf{f}_2^1 \\ \vdots \\ \mathbf{f}_p^1 \end{pmatrix} = \begin{pmatrix} \mathbf{S}_1 \mathbf{z} \\ \mathbf{S}_2 \mathbf{z} \\ \vdots \\ \mathbf{S}_p \mathbf{z} \end{pmatrix}$$

Finally we may write this as

$$\begin{pmatrix} \mathbf{f}_1^1 \\ \mathbf{f}_2^1 \\ \vdots \\ \mathbf{f}_p^1 \end{pmatrix} = \begin{pmatrix} \mathbf{S}_1(\mathbf{z} - \sum_{j\neq 1} \mathbf{f}_j^1) \\ \mathbf{S}_2(\mathbf{z} - \sum_{j\neq 2} \mathbf{f}_j^1) \\ \vdots \\ \mathbf{S}_p(\mathbf{z} - \sum_{j\neq p} \mathbf{f}_j^1) \end{pmatrix}$$

Thus the Newton-Raphson updates are an additive model fit; in fact they solve a weighted and penalized quadratic criterion which is the local approximation to the penalized log-likelihood.

Note: any linear smoother can be viewed as the solution to some penalized likelihood. So we can set-up to penalized likelihood criterion so that the solution is what we want it to be.

This algorithm converges with any linear smoother.

8.1.2 Inference

Deviance

The deviance or likelihood-ratio statistic, for a fitted model $\hat{\boldsymbol{\mu}}$ is defined by

$$D(\mathbf{y}; \hat{\boldsymbol{\mu}}) = 2\{l(\boldsymbol{\mu}_{max}; \mathbf{y}) - l(\hat{\boldsymbol{\mu}})\}$$

where $\boldsymbol{\mu}_{max}$ is the parameter value that maximizes $l(\hat{\boldsymbol{\mu}})$ over all $\boldsymbol{\mu}$ (the saturated model). We sometimes unambiguously use $\hat{\boldsymbol{\eta}}$ as the argument of the deviance rather than $\hat{\boldsymbol{\mu}}$.

Remember for GLM if we have two linear models defined by η_1 nested within η_2, then under appropriate regularity conditions, and assuming η_1 is correct, $D(\hat{\eta}_2; \hat{\eta}_1) = D(y; \hat{\eta}_1) - D(y; \hat{\eta}_2)$ has asymptotic χ^2 distribution with degrees of freedom equal to the difference in degrees of freedom of the two models. This result is used extensively in the analysis of deviance tables etc...

For non-parametric we can still compute deviance and it still makes sense to compare the deviance obtained for different models. However, the asymptotic approximations are undeveloped.

H&T present heuristic arguments for the non-parametric case.

Standard errors

Each step of the local scoring algorithm consists of a backfitting loop applied to the adjusted dependent variables $\mathbf{z}$ with weights $\mathbf{A}$ given by the estimated information matrix. If $\mathbf{R}$ is the weighted additive fit operator, then at convergence

$$\begin{aligned}\hat{\boldsymbol{\eta}} &= \mathbf{R}(\hat{\boldsymbol{\eta}} + \mathbf{A}^{-1}\hat{\boldsymbol{\mu}}) \\ &= \mathbf{R}\mathbf{z},\end{aligned}$$

where $\hat{\mathbf{u}} = \partial l / \partial \hat{\boldsymbol{\eta}}$. The idea is to approximate $\mathbf{z}$ by an asymptotically equivalent quantity $\mathbf{z}_0$. We will not be precise and write $\approx$ meaning asymptotically equivalent.

Expanding $\hat{\mathbf{u}}$ to first order about the true $\boldsymbol{\eta}_0$, we get $\mathbf{z} \approx \mathbf{z}_0 + \mathbf{A}_0^{-1}\mathbf{u}_0$, which has mean $\boldsymbol{\eta}_0$ and variance $\mathbf{A}_0^{-1}\phi \approx \mathbf{A}\phi$.

Remember for additive models we had the fitted predictor $\hat{\boldsymbol{\eta}} = \mathbf{R}\mathbf{y}$ where $\mathbf{y}$ has covariance $\sigma^2\mathbf{I}$. Here $\hat{\boldsymbol{\eta}} = \mathbf{R}\mathbf{z}$, and $\mathbf{z}$ has asymptotic covariance $\mathbf{A}_0^{-1}$. $\mathbf{R}$ is not a linear operator due to its dependence on $\hat{\mu}$ and thus $\mathbf{y}$ through the weights, so we need to use its asymptotic version $\mathbf{R}_0$ as well. We therefore have

$$\mathbf{cov}(\hat{\boldsymbol{\eta}}) \approx \mathbf{R}_0\mathbf{A}_0^{-1}\mathbf{R}_0'\phi \approx \mathbf{R}\mathbf{A}^{-1}\mathbf{R}'\phi$$

Similarly

$$\mathbf{cov}(\hat{\mathbf{f}}_j) \approx \mathbf{R}_j\mathbf{A}^{-1}\mathbf{R}_j'\phi$$

where $\mathbf{R}_j$ is the matrix that produces $\hat{\mathbf{f}}_j$ from z.

Under some regularity conditions we can further show that $\hat{\boldsymbol{\nu}}$ is asymptotically normal, and this permits us to construct confidence intervals.

8.1.3 Degrees of freedom

Previously we described how we defined the degrees of freedom of the residuals as the expected value of the residual sum of squares. The analogous quantity in generalized models is the deviance. We therefore use the expected value of the deviance to define the *relative degrees of freedom*.

We don't know the exact or asymptotic distribution of the deviance so we need some approximation that will permit us to get an approximate expected value.

Using a second order Taylor approximation we have that

$$\mathrm{E}[D(\mathbf{y};\hat{\boldsymbol{\mu}})] \approx \mathrm{E}[(\mathbf{y}-\hat{\boldsymbol{\mu}})'\mathbf{A}^{-1}(\mathbf{y}-\hat{\boldsymbol{\mu}})]$$

with $\mathbf{A}$ the Hessian matrix defined above. We now write this in terms of the "linear terms".

$$\mathrm{E}[(\mathbf{y}-\hat{\boldsymbol{\mu}})'\mathbf{A}(\mathbf{y}-\hat{\boldsymbol{\mu}})] \approx (\mathbf{z}-\hat{\boldsymbol{\eta}})'\mathbf{A}(\mathbf{z}-\hat{\boldsymbol{\eta}})$$

and we can show that this implies that if the model is unbiased

$$\mathrm{E}(D) = df\phi$$

with

$$df = n - \mathrm{tr}(2\mathbf{R} - \mathbf{R}'\mathbf{A}\mathbf{R}\mathbf{A}^{-1})$$

This gives the degrees of freedom for the whole model not for each smoother. We can obtain the dfs for each smoother by adding them one at a time and obtaining

$$\mathrm{E}[D(\hat{\boldsymbol{\eta}}_2;\hat{\boldsymbol{\eta}}_1)] \approx \mathrm{tr}(2\mathbf{R}_1 - \mathbf{R}_1'\mathbf{A}_1\mathbf{R}_1\mathbf{A}_1^{-1}) - \mathrm{tr}(2\mathbf{R}_2 - \mathbf{R}_2'\mathbf{A}_2\mathbf{R}_2\mathbf{A}_2^{-1})$$

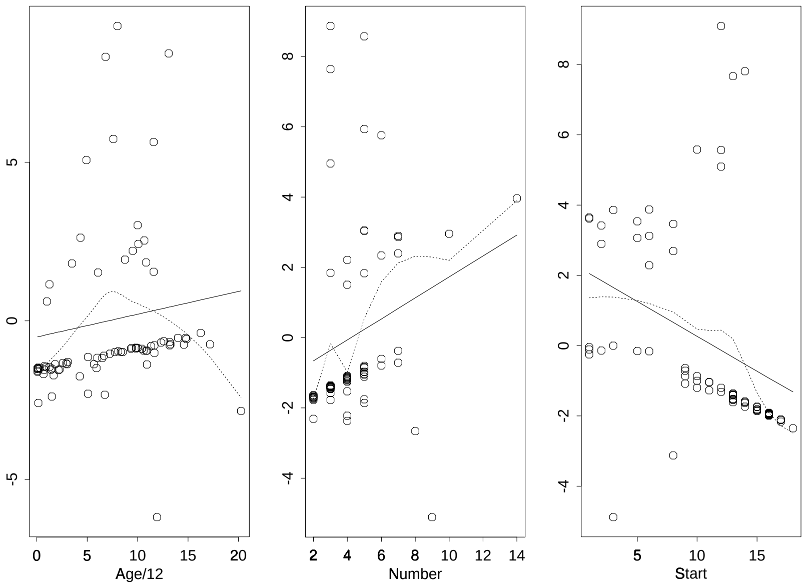
In general, the crude approximation $df_j = \mathrm{tr}(\mathbf{S}_j)$ is used.

8.1.4 An Example

The kyphosis data frame has 81 rows representing data on 81 children who have had corrective spinal surgery. The binary outcome Kyphosis indicates the presence or absence of a postoperative deformity (called Kyphosis). The other three variables are `Age` in months, `Number` of vertebra involved in the operation, and the beginning of the range of vertebrae involved (`Start`).

Using GLM these are the results we obtain

```
                   Value  Std. Error   t value
(Intercept) -1.213433077 1.230078549 -0.986468
        Age  0.005978783 0.005491152  1.088803
     Number  0.298127803 0.176948601  1.684827
      Start -0.198160722 0.065463582 -3.027038

Null Deviance: 86.80381 on 82 degrees of freedom

Residual Deviance: 65.01627 on 79 degrees of freedom
```

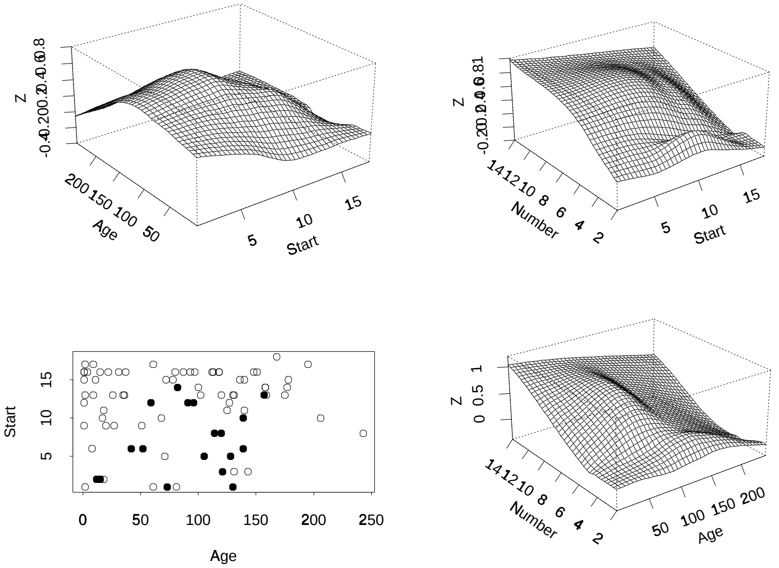


The dotted lines are smooths of the residuals. This does not appear to be a very good fit.

We may be able to modify it a bit, by choosing a better model than a sum of lines.

We’ll use smoothing and GAM to see what “the data says”.

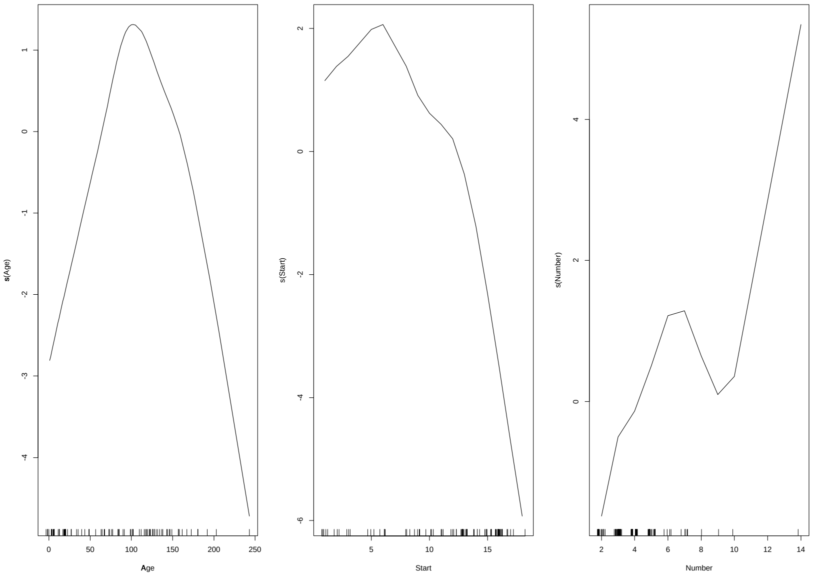
Here are some smooth versions of the data:



And here are the gam results:

```
Null Deviance: 86.80381 on 82 degrees of freedom

Residual Deviance: 42.74212 on 70.20851 degrees of freedom

Number of Local Scoring Iterations: 7

DF for Terms and Chi-squares for Nonparametric Effects

            Df Npar Df Npar Chisq    P(Chi)
(Intercept)  1
     s(Age)  1     2.9   6.382833 0.0874180
   s(Start)  1     2.9   5.758407 0.1168511
  s(Number)  1     3.0   4.398065 0.2200849
```

Notice that it is a much better fit and not many more degrees of freedom. Also notice that the tests for linearity are close to “rejection at the 0.05 level”.



We can either be happy considering these plots as descriptions of the data, or we can use it to inspire a parametric model:

Before doing so, we decide not to include Number becuase it seems to be associated with “Start” and not adding much to the fit. This and other considerations suggest we not include `Number`. The gam plots suggest the following “parametric” model.

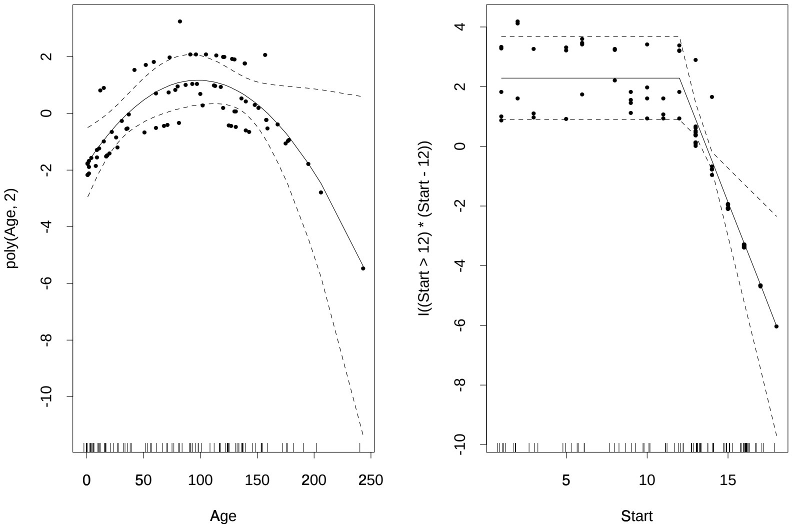
```
glm2 <- glm(Kyphosis˜poly(Age,2) + I((Start > 12) * (Start - 12)),
            family=binomial)
```

Here are the results of this fit... much better than the original GLM fit.

```
Coefficients:
                                     Value Std. Error    t value
```

```
                      (Intercept)  -0.5421608  0.4172229 -1.2994512
                 poly(Age, 2)1   2.3659699  4.1164283  0.5747628
                 poly(Age, 2)2 -10.5250479  5.2840926 -1.9918364
I((Start > 12) * (Start - 12))  -1.3840765  0.5145248 -2.6900094

(Dispersion Parameter for Binomial family taken to be 1 )

    Null Deviance: 86.80381 on 82 degrees of freedom

Residual Deviance: 56.07235 on 79 degrees of freedom

Number of Fisher Scoring Iterations: 6
```

Here are the residual plots:



8.1.5 Prediction using GAM

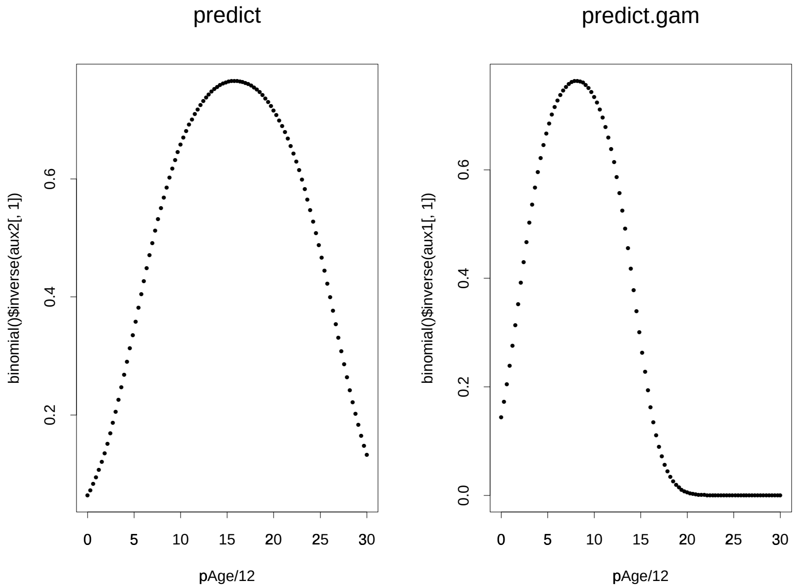
Often we wish to evaluate the fitted model at some new values.

With parametric models this is simple because all we do is form a new design matrix and multiply by the estimated parameters.

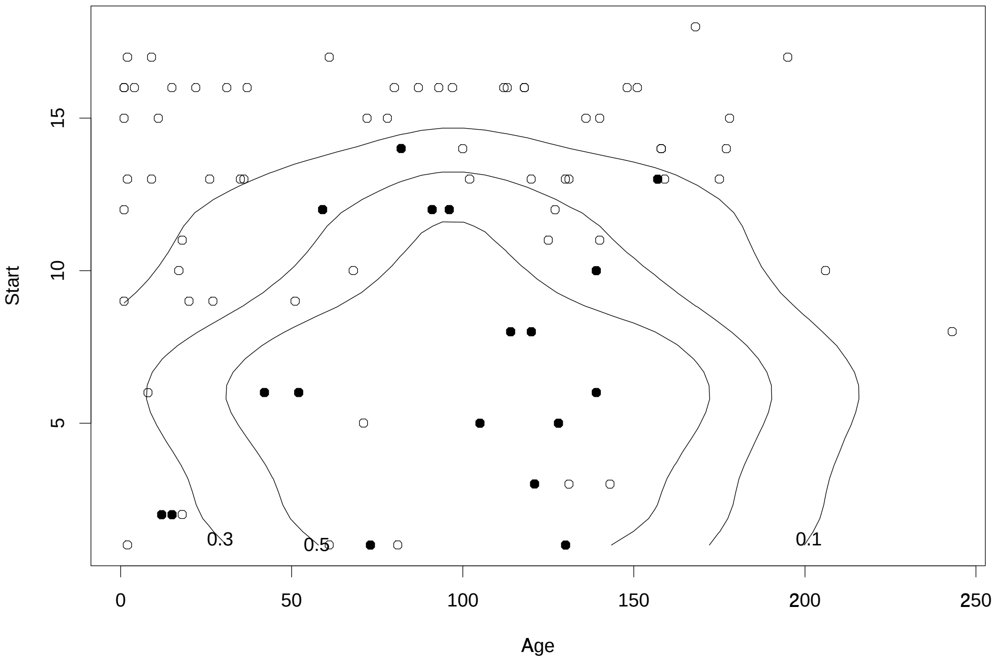
Some of the functions used to create design matrices in lm, glm a and gam are data dependent. For example `bs()`, `poly()`, make some standardization of the covariate before fitting and therefore new covariates would change the meaning of the parameters.

As an example look at what happens when we predict fitted values for new values of AGE in the Kyphosis example using `predict()`.

The solution is to use `predict.gam()` that takes this into account



`predict.gam` is especially useful when we want to make surface plots. For example:


Start
Age
0
50
100
150
200
250
5
10
15
0.3
0.5
0.1

8.1.6 Over-interpreting additive fits

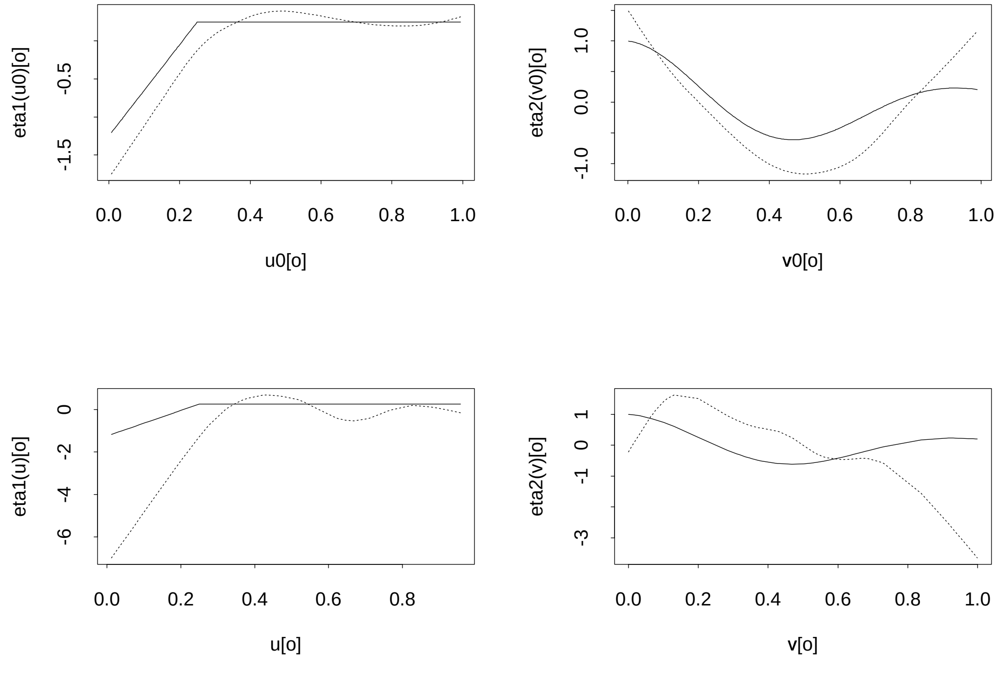
One of the advantages of GAM is their flexibility. However, because of this flexibility we have to be careful not to "over-fit" and interpret the results incorrectly.

Binary data is especially sensitive. We construct a simulated example to see this.

The following figure shows the functional components f_1 and f_2 of a GAM

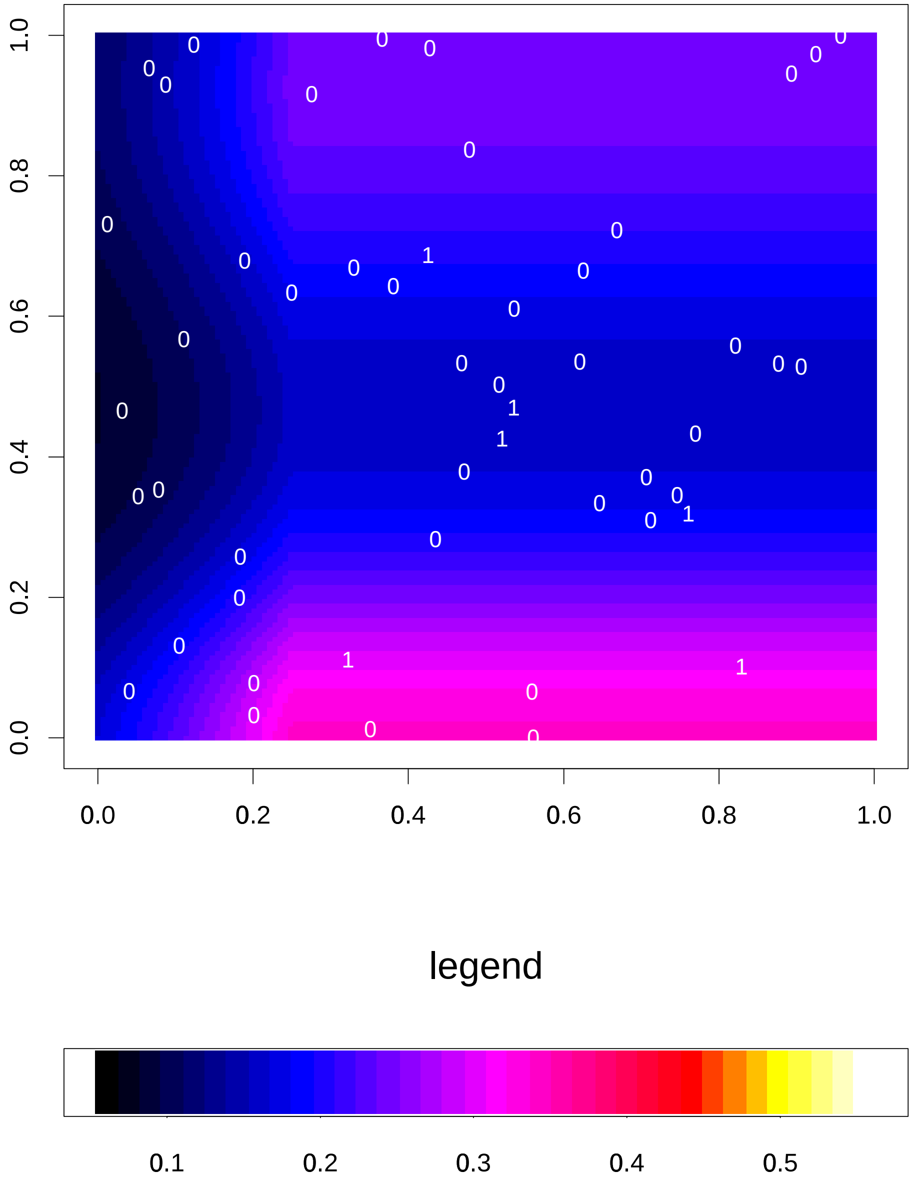
$$\text{logit}\{\Pr(Y = 1|U, V)\} = -1 + f_1(U) + f_2(V)$$

with U and V independent uniform(0,1).



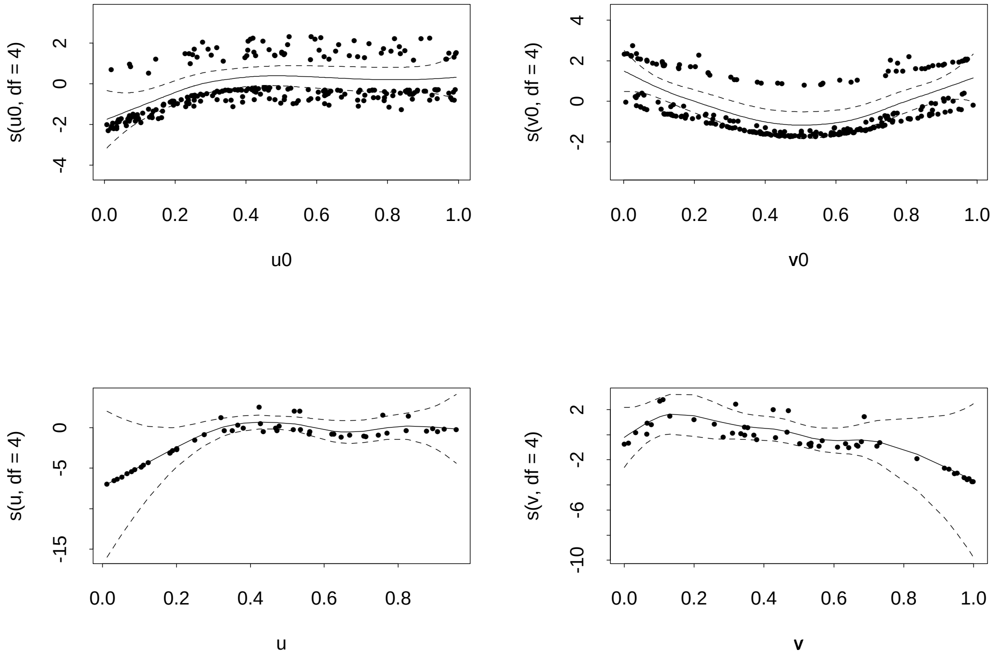
We also show the "smooths" obtained for a data set of 250 observations and a data set of 50 observations. Notice how "bad" the second fit is.

If we make a plot of the mean $\mu(u, v)$ and of it's estimate we see why this happens.



We have relatively large neighborhoods of $[0,1] \times [0,1]$ that contain only 1s or only 0s. The estimates in these regions will have linear part close to infinity and minus infinity!

One way to detect this when we don't know "the truth" is to look at the estimates with standard errors and partial residuals. If the partial residuals follow the fit to closely and the standard errors "explode" we know something is wrong.



8.2 Local Likelihood

Suppose we have independent observation s $\{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)\}$ that are the realization of a response random variable Y given a $P \times 1$ covariate vector $\mathbf{x}$ which we consider to be known. Given the covariate $\mathbf{x}$, the response variable Y follows a parametric distribution $Y \sim g(y; \theta)$ where θ is a function of $\mathbf{x}$. We are interested in estimating θ using the observed data.

The log-likelihood function can be written as

$$l(\theta_1, \dots, \theta_n) = \sum_{i=1}^{n} \log g(y_i; \theta_i) \tag{8.1}$$

where $\theta_i = s(\mathbf{x}_i)$. A standard modeling procedure would assume a parsimonious form for the θ_is, say $\theta_i = \mathbf{x}_i' \boldsymbol{\beta}$, $\boldsymbol{\beta}$ a $P \times 1$ parameter vector. In this case the log-likelihood $l(\theta_1, \dots, \theta_n)$ would be a function of the parameter $\boldsymbol{\beta}$ that could be estimated by maximum likelihood, that is by finding the $\hat{\boldsymbol{\beta}}$ that maximizes $l(\theta_1, \dots, \theta_n)$.

The local likelihood approach is based on a more general assumption, namely that $s(\mathbf{x})$ is a "smooth" function of the covariate $\mathbf{x}$. Without more restrictive assumptions, the maximum likelihood estimate of $\boldsymbol{\theta} = \{s(\mathbf{x}_1), \dots, s(\mathbf{x}_n)\}$ is no longer useful because of over-fitting. Notice for example that for the case of regression with all the $\mathbf{x}_i$s distinct the maximum likelihood estimate would simply reproduce the data.

Suppose we are interested in estimating only $\theta_0 = \theta(\mathbf{x}_0)$ for a fixed covariate value $\mathbf{x}_0$. The local likelihood estimation approach is to assume that there is some neighborhood N_0 of covariates that are "close" enough to $\mathbf{x}_0$ such that the data $\{(\mathbf{x}_i, y_i); \mathbf{x}_i \in N_0\}$ contain information about θ_0 through some *link function* η of the form

$$\begin{aligned} \theta_0 &= s(\mathbf{x}_0) \equiv \eta(\mathbf{x}_0, \boldsymbol{\beta}) \text{ and} & (8.2) \\ \theta_i &= s(\mathbf{x}_i) \approx \eta(\mathbf{x}_i, \boldsymbol{\beta}), \text{ for } \mathbf{x}_i \in N_0. & (8.3) \end{aligned}$$

Notice that we are abusing notation here since we are considering a different $\boldsymbol{\beta}$ for every $\mathbf{x}_0$. Throughout the work we will be acting as if θ_0 is the only parameter of interest and therefore not indexing variables that depend on the choice of $\mathbf{x}_0$.

The local likelihood estimate of θ_0 is obtained by assuming that, for data in N_0, the true distribution of the data, $g(y_i;\theta_i)$ is approximated by

$$f(y_i;\mathbf{x}_i,\boldsymbol{\beta}) \equiv g(y_i;\eta(\mathbf{x}_i,\boldsymbol{\beta})), \tag{8.4}$$

finding the $\hat{\boldsymbol{\beta}}$ maximizes the local log-likelihood equation

$$l_0(\boldsymbol{\beta}) = \sum_{\mathbf{x}_i \in N_0} w_i \log f(y_i;\boldsymbol{\beta}), \tag{8.5}$$

and then using Equation (8.2) to obtain the local likelihood estimate $\hat{\theta}_0$. Here w_i is a weight coefficient related to the "distance" between $\mathbf{x}_0$ and $\mathbf{x}_i$. In order to obtain a useful estimate of θ_0, we need $\boldsymbol{\beta}$ to be of "small" enough dimension so that we fit a parsimonious model within N_0.

Hastie and Tibshirani (1987) discuss the case where the covariate $\mathbf{x}$ is a real valued scalar and the link function is linear

$$\eta(x_i,\boldsymbol{\beta}) = \beta_0 + x_i\beta_1$$

Notice that in this case, the assumption being made is that the parameter function $s(x_i)$ is approximately linear within "small" neighborhoods of x_0, i.e. locally linear.

Staniswalis (1989) presents a similar approach. In this case the covariate $\mathbf{x}$ is allowed to be a vector, and the link function is a constant

$$\eta(\mathbf{x}_i,\beta) = \beta$$

The assumption being made here is that the parameter function $s(x_i)$ is locally constant.

If we assumes a density function of the form

$$\log g(y_i;\theta_i) = C + (y_i - \theta_i)^2/\phi \tag{8.6}$$

where K and ϕ are constants that do not depend on the θ_is, local regression may be considered a special case of local likelihood estimation.

Notice that in this case the local likelihood estimate is going to be equivalent to the estimate obtained by minimizing a sum of squares equation. The approach in Cleveland (1979) and Cleveland and Devlin (1988) is to consider a real valued covariate and the polynomial link function

$$\eta(\mathbf{x}_i, \boldsymbol{\beta}) = \sum_{j=0}^{d} x_i^j \beta_j.$$

In general, the approach of local likelihood estimation, including the three above-mentioned examples, is to assume that for "small" neighborhoods around $\mathbf{x}_0$, the distribution of the data is approximated by a distribution that depends on a constant parameter $\boldsymbol{\beta}(\mathbf{x}_0)$, i.e. we have locally parsimonious models. This allows us to use the usual estimation technique of maximum likelihood. However, in the local version of maximum likelihood we often have an a priori belief that points "closer" to $\mathbf{x}_0$ contain more information about θ_0, which suggest a weighted approach.

The asymptotic theory presented in, for example, Staniswalis (1989) and Loader (1986) is developed under the assumption that as the size (or radius) of some neighborhood of the covariate of interest $\mathbf{x}_0$ tends to 0, the difference between the true and approximating distributions within such neighborhood becomes negligible. Furthermore, we assume that despite the fact that the neighborhoods become arbitrarily small, the number of data points in the neighborhood somehow tends to ∞. The idea is that, asymptotically, the behavior of the data within a given neighborhood, is like the one assumed in classical asymptotic theory for non-IID data: The small window size assure that the difference between the true and approximating models is negligible and the large number of independent observations is available to estimate a parameter of fixed dimension that completely specifies the joint distribution. This concept motivates the approach taken in the following sections to derive a model selection criteria.

Bibliography

[1] Cleveland, W. S. and Devlin, S. J. (1988), “Locally weighted regression: An approach to regression analysis by local fitting,” *Journal of the American Statistical Association*, 83, 596–610.

[2] Hastie, T. and Tibshirani, R. (1987), “Generalized additive models: Some applications,” *Journal of the American Statistical Association*, 82, 371–386.

[3] Loader, C. R. (1996), “Local likelihood density estimation,” *The Annals of Statistics*, 24, 1602–1618.

[4] Loader, C. R. (1999), *Local Regression and Likelihood*, New York: Springer.

[5] Staniswalis, J. G. (1989), “The kernel estimate of a regression function in likelihood-based models,” *Journal of the American Statistical Association*, 84, 276–283.

[6] Tibshirani, R. and Hastie, T. (1987), “Local likelihood estimation,” *Journal of the American Statistical Association*, 82, 559–567.