Statistical Modeling with Spline Functions
Methodology and Theory

Mark H. Hansen
University of California at Los Angeles

Jianhua Z. Huang
University of Pennsylvania

Charles Kooperberg
Fred Hutchinson Cancer Research Center

Charles J. Stone
University of California at Berkeley

Young K. Truong
University of North Carolina at Chapel Hill

January 5, 2006

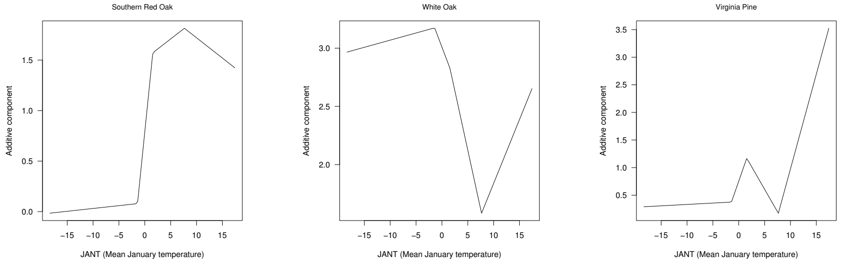

FIGURE 3.27. Dependence of each tree species on mean January temperature.

4
Generalized Linear Models

In the previous chapter, we described the relationship between an input vector $\boldsymbol{X}$ and a response Y through the conditional mean and variance functions

$$E(Y|\boldsymbol{X}=\boldsymbol{x}) = f(\boldsymbol{x}) \qquad \text{and} \qquad \operatorname{var}(Y|\boldsymbol{X}=\boldsymbol{x}) = \sigma^2\,; \tag{4.0.1}$$

and our interest focused on describing the significant features of f. We approached this problem by introducing a *linear model*, or more precisely, a finite-dimensional linear space $\mathbb{G}$. Given a series of independent observations, $(\boldsymbol{X}_1, Y_1), \ldots, (\boldsymbol{X}_n, Y)$, we took as an estimate for f the ordinary least squares (OLS) projection into $\mathbb{G}$. By choosing linear spaces $\mathbb{G}$ based on their flexibility, or rather, their ability to represent a variety of functional forms, we constructed an estimate of f with favorable empirical and theoretical properties under relatively weak assumptions about f (requiring only that f be smooth in some sense). Adaptive methods were introduced to select from among a number of competing linear spaces $\mathbb{G}$, helping to further resolve features like peaks and valleys.

The use of OLS as a fitting criterion was derived originally under the distributional assumption that Y be a normal random variable with the mean and variance given in (4.0.1). While OLS is applied even in cases when strict normality is questionable, there is certainly a limit to its effectiveness. The class of *generalized linear models* (GLM's) encompasses a number of common estimation problems for which OLS is not an appropriate fitting criterion. The term "generalized" refers first to the fact that the distribution of Y can be any one of a large class of so-called *exponential families*. In addition, f may no longer be a simple conditional mean as in (4.0.1), but instead can represent some transformation of the mean of Y. In each case, however, flexible linear spaces $\mathbb{G}$ are introduced to model f, and hence we retain the term "linear." We begin this chapter with two applications: count data, for which Y is taken to be a Poisson random variable; and binary data, corresponding to a Bernoulli distribution for Y. As we will see, computational issues become much more important for GLM's, as estimates are no longer obtained through simple OLS projections, but instead involve iterative optimization techniques. Because our adaptive procedures depend on auditioning a large number of basis funcions, computing a new fit for each candidate is impractical. Issues concerning estimation and adaptation can be treated quite generally in the context of GLM's. In fact, many of the more elaborate estimation problems we consider in later chapters share basic properties with GLM's. The techniques we introduce now will be applied repeatedly throughout the text.

4.1 Applications

4.1.1 Health effects of particulate matter

Background

Largely through regulatory efforts, air quality in the United States has improved considerably over the past three decades. Despite this improvement, many researchers have observed significant health effects resulting from current air pollution levels. Recent epidemiological studies have linked an increase in particulate matter (PM) with increased mortality among the elderly. Of primary concern is the impact of airborne particles that are formed by the combustion of fossil fuels.[1] In this section, we will consider data collected from the city of Philadelphia over a 15 year period from January 1, 1974 through December 31, 1988. Statistics related to mortality, air pollution levels and weather conditions were culled from publicly available sources by a group of researchers at Johns Hopkins University and appear in Kelsall, Samet, Zeger and Xu (1997) as well as Dominici, Samet and Zeger (2000) and Dominici, McDermott and Hastie (2004). These authors conducted a careful analysis of the data and proposed an additive smoothing spline model for mortality. Our analysis will instead explore adaptive schemes for knot placement.

For the 15 year period under study, complete nonaccidental mortality figures are available. These have been stratified by age and cause of death. In Figure 4.1 we present total daily mortality counts for the last three years covered by the data, from December 31, 1985 through December 31, 1988,

[1] Other "secondary particles" are also formed from the sulfur and nitrogen dioxides that are released during combustion.

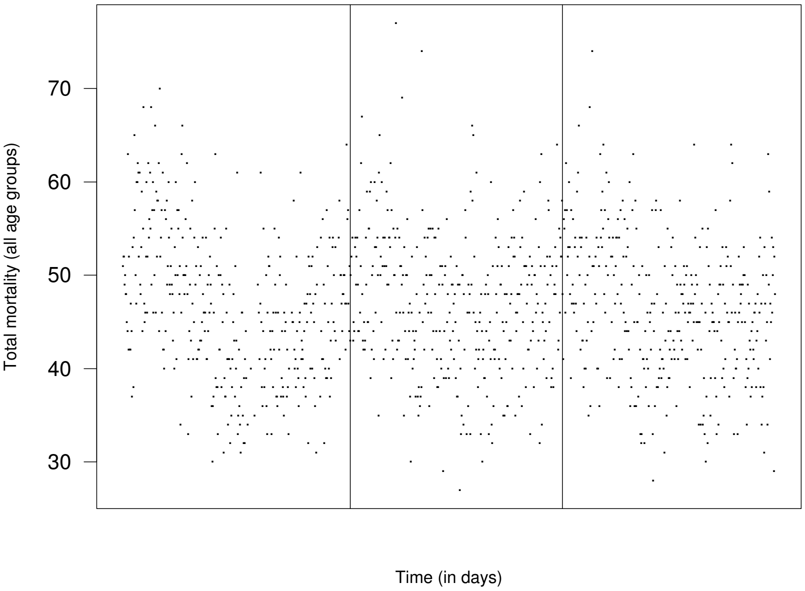

FIGURE 4.1. Total mortality in Philadelphia for 1985–1988. Peaks denote winter months.

marking the dates December 31, 1986 and December 31, 1987 with vertical lines. Notice that the peaks in mortality correspond to winter months. In Figure 4.2 we divide the nonaccidental mortality data into three categories based on the cause of death: cardiovascular disease (CVD), respiratory disease (Resp), and "other." These counts are then further stratified to examine seasonal effects across three different age groups: those less than 65 years old, those between 65 and 75, and those older than 75). In Figure 4.2 we present boxplots of daily mortality counts, where each set is ordered by season (fall, winter, spring and summer). The peak in the winter months among the oldest members of the population is clear for all three causes of death. The "other" category involves mainly the youngest age group, and the seasonal effect is the weakest for this combination of factors. To explore time trends in mortality, we now consider a spline model in which the total deaths from all three causes are modeled as a smooth function of time and separate indicator variables for the three age groups.

Poisson regression

As we are tracking counts, it is common to use a Poisson regression model to capture the effects of the candidate predictors on mortality. Let Y denote

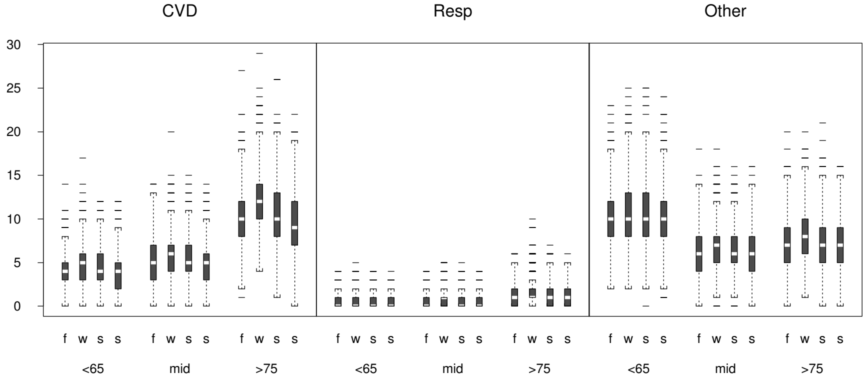

FIGURE 4.2. Daily mortality counts stratified by season and cause of death.

daily mortality, and let $\boldsymbol{X}$ denote a vector of covariates. Set

$$p(Y=k|\boldsymbol{X}=\boldsymbol{x})=\frac{e^{-\mu(\boldsymbol{x})}\mu(\boldsymbol{x})^k}{k!}\,, \tag{4.1.1}$$

so that the conditional mean and variance of Y are given by

$$E(Y|\boldsymbol{X}=x)=\mu(\boldsymbol{x}) \quad \text{and} \quad \operatorname{var}(Y|\boldsymbol{X}=x)=\mu(\boldsymbol{x})\,. \tag{4.1.2}$$

Following the programme from the previous chapter, we attempt to capture the important features in $\mu(\boldsymbol{x})$ via (tensor product) splines. In this case, however, we do not work directly with the mean function, but instead find it more convenient to construct spline estimates of $\log\mu(\boldsymbol{x})$. The major ingredient of this so-called *log-linear model* is a linear space $\mathbb{G}$ with basis functions $B_1,\ldots,B_J$, where each $g\in\mathbb{G}$ can be written as

$$g(\boldsymbol{x};\boldsymbol{\beta})=\beta_1B_1(\boldsymbol{x})+\cdots+\beta_JB_J(\boldsymbol{x})$$

for some vector $\boldsymbol{\beta}=(\beta_1,\ldots,\beta_J)$. Then, given data $(\boldsymbol{X}_i,Y_i)$, $i=1,\ldots,n$, we evaluate members of $\mathbb{G}$ based on the log-likelihood

$$\ell(g)=\sum_{i=1}^n\left[-\exp g(\boldsymbol{X}_i)+Y_i\,g(\boldsymbol{X}_i)\right]\,,\quad g\in\mathbb{G}\,, \tag{4.1.3}$$

where we have substituted $\exp g(\boldsymbol{x})$ for $\mu(\boldsymbol{x})$ in (4.1.1) and have dropped a constant that does not depend on $g(\boldsymbol{x})$. As in the previous chapter, we let $\widehat{g}=\operatorname{argmax}_{g\in\mathbb{G}}\ell(g)$ denote the maximum likelihood estimate. Writing the log-likelihood in terms of $\boldsymbol{\beta}$,

$$\ell(\boldsymbol{\beta})=\sum_{i=1}^n\left[-\exp\left(\sum_{j=1}^J\beta_jB_j(\boldsymbol{X}_i)\right)+Y_i\sum_{j=1}^J\beta_jB_j(\boldsymbol{X}_i)\right]\,, \tag{4.1.4}$$

Type	Variable
Weather[1]	Daily temperature Daily dew point
Air Pollution[2]	Total suspended particles (PM_{10}) Sulfur dioxide (SO_2) Nitrogen dioxide (NO_2) Carbon monoxide (CO) Ozone (O_3)
Mortality[3]	Cardiovascular diseases Respiratory diseases Other

[1] Source: Nation Weather Center
[2] Source: Aerometric Information Retrieval Service, US EPA
[3] Source: National Center for Health Statistics

TABLE 4.1. Summary of variables available in the JHU dataset. Data provided by Francesca Dominici, Department of Biostatistics, Johns Hopkins University.

we can again write $\widehat{g}(\boldsymbol{x}) = g(\boldsymbol{x}; \widehat{\boldsymbol{\beta}})$ where $\widehat{\boldsymbol{\beta}} = \operatorname{argmax}_{\boldsymbol{\beta} \in \mathbb{R}^J} \ell(\boldsymbol{\beta})$. Unfortunately, there generally is no closed-form expression for the value of $\boldsymbol{\beta}$ that maximizes the (4.1.4). By working with a transformation of $\mu(\boldsymbol{x})$, we are able to derive a simple iterative scheme for finding $\boldsymbol{\beta}$ that makes use repeated least squares fits. For the moment, however, this background is sufficient to motivate our discussion of the mortality data.

Modeling mortality

In Table 4.1 we present a set of variables included in the JHU dataset. Meteorological variables like average daily temperature are known to have considerable impact on mortality, especially among the elderly. In Table 4.1 we also find several kinds of measurements that characterize daily pollution levels. The most important is PM_{10}, the concentration per cubic meter of suspended particles with diameter greater than 10μm. Recent studies like Dominici et al. (2000) have built additive models to isolate the relative increase in mortality that can be expected from an increase of PM_{10} by $10\ \mu g\, m^{-3}$. We apply a variant of the adaptive spline methodology discussed in the previous chapter to identify both the important variables and their functional form.

We consider the total nonaccidental mortality for the oldest group in the study, those aged 75 or more. Of the 5479 days between January 1, 1974 and December 31, 1988, we have complete mortality, meteorological, and air pollution data for $n = 5254$. We begin with a simple model using each of the weather and air pollutant variables in Table 4.1 together with a categorical variable for day of the week (7 levels), and season (4 levels).

	Coefficient	SE	z	P-value
Intercept	3.1	3.0×10^{-2}	104.2	.00
Weather				
Dew	2.3×10^{-3}	6.6×10^{-4}	3.5	.00
Dew (lag)	-1.6×10^{-3}	8.1×10^{-4}	-2.0	.05
Temp	-1.2×10^{-3}	9.8×10^{-4}	-1.2	.22
Temp (lag)	-3.4×10^{-3}	1.1×10^{-3}	-3.2	.00
Pollution				
TSP	5.2×10^{-4}	2.0×10^{-4}	2.6	.01
NO2	-6.0×10^{-4}	3.7×10^{-4}	-1.6	.11
SO2	1.1×10^{-3}	4.6×10^{-4}	2.3	.02
CO	7.1×10^{-6}	5.7×10^{-6}	1.3	.21
O3	1.4×10^{-3}	3.6×10^{-4}	3.9	.00
Time				
Day	1.4×10^{-2}	8.3×10^{-4}	16.5	.00
		$\vdots$		

TABLE 4.2. Initial fit to mortality. We have not reported the coefficients from a 7-level factor representing day of the week and a 4-level factor representing season.

We also included two lagged variables, averages of the previous three days values for temperature and dew point. The maximum likelihood estimates of the coefficients together with their standard errors are given in Table 4.2. Notice that many of the effects are significant; the P-values were calculated using approximate distributional results for the standardized coefficients in a Poisson regression. We will have more to say about this in the next section.

Much of the recent literature on mortality and particulate matter centers on correctly capturing weather effects and other slowly-varying but unmeasured seasonal phenomena. For this reason, these studies often include a more elaborate effect of time than the simple linear term listed in Table 4.2. To attempt to assess the possible deficiencies in the fit, we consider the so-called *Pearson residuals* from the fit,

$$r_i = [\,Y_i - \exp \widehat{g}(\boldsymbol{X}_i)\,] \,/ \sqrt{\exp \widehat{g}(\boldsymbol{X}_i)}\,, \qquad i = 1, \ldots, n\,. \tag{4.1.5}$$

The autocorrelation function computed for a time-series of these residuals reveals significant structure at several lags, the implication being that the fitted model is not rich enough, leaving out some component of variability that is related to time. Next, recall that for a Poisson model, the conditional variance of Y is the same as its mean (4.1.2). To examine this assumption, we consider the statistic

$$\hat{\sigma}^2 = \frac{1}{n-J}\sum_{i=1}^{n} r_i^2$$

where in this case the dimension of the selected spline space is $J = 20$. Ideally, for large samples the value of $\hat{\sigma}^2$ should be close to one.[2] In our case, $\hat{\sigma}^2 = 1.14$ indicating again that there is an extra component of variability unexplained by the model. In the literature on GLMs, this effect is known as over-dispersion.

To try to make up for the time component and to illustrate an initial application of splines in this context, we will introduce a natural cubic spline in time. To be precise, let x be an index for the number of days since the start of our data set, January 1, 1974. Then, following the simple recipe given in the previous chapter, we build a space of natural cubic splines by adding knots sequentially. At each step, we add the knot that creates the greatest increase in the log-likelihood (4.1.3). In principle, this can be an expensive operation. Unlike in the regression problems of the previous chapter, introducing a new basis function means having to compute maximum likelihood estimates, for which there are no closed-form solutions. Ignoring the complexity of implementation at the moment, we first perform a very direct search for new knots. As in Section 3.6, we place boundary knots at the first and last day in our study and then add basis functions one at a time. In Figure 4.3, we plot the results of our search. The first line consists of the $-2\ell(\widehat{g}_{t_1})$ where $\widehat{g}_{t_1}$ is the MLE in the natural cubic spline space with knots $(0, t_1, 5479)$. We take $\widehat{t}_1$ to be the point at which this quantity is minimized. The dashed line below this curve tracts $-2\ell(\widehat{g}_{t_2})$ where $\widehat{g}_{t_2}$ is the MLE in the natural cubic spline space with knots $(0, \widehat{t}_1, \widehat{t}_2, 5479)$. Proceeding in this fashion, we create nested spaces of splines $\mathbb{G}_1 \subset \mathbb{G}_2 \subset \cdots$.

We can constrain the addition process using the concept of *allowable spaces* introduced in Chapter 3. Our observations about variance and knot spacing hold in the context of Poisson regression as well as for OLS. At each step in the process we add the candidate basis function that causes the greatest boost in the log-likelihood (or, equivalently, the greatest drop in -2ℓ) subject to these constraints. Starting from the simple model in Table 4.2, we added terms to the natural spline model (again, with an implicit pair of knots at the endpoints of the data) until we had added 15 knots. Given a sequence of spline spaces $\mathbb{G}_\nu$ with dimension J_ν, $\nu = 1, 2, \ldots, K$, we select a model using the AIC-type criterion

$$\mathrm{AIC}_\alpha(\nu) = -2\ell(\widehat{g}_\nu) + \alpha J_\nu\ ,$$

[2] The Pearson goodness of fit statistic is given by $\sum_i r_i^2$ and is known to have an (asymptotically) χ^2 distribution with $n - J$ degrees of freedom.

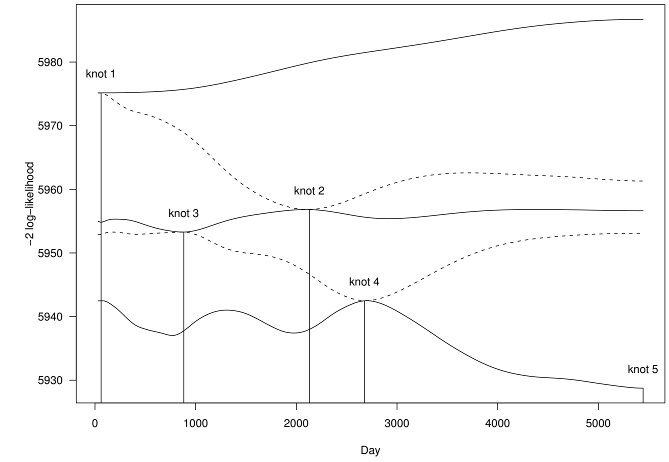

FIGURE 4.3. Stepwise addition of knots for an natural cubic spline in time. The spline space was simply added to the fit in Table 4.2.

where $\widehat{g}_\nu$ is the maximum likelihood estimate in $\mathbb{G}_\nu$. We set $\alpha = \log n$, which corresponds to the Bayesian information criterion, BIC. We will discuss the derivation of this selection rule in the next section. The model in Table 4.2 had an AIC `MHH: Should AIC be BIC here?` value of 6169. The best model found in terms of AIC `MHH: BIC?` during the forward selection process had a value of 6128 and involved 11 knots in time. This is only a modest improvement. The knots were also somewhat evenly spaced over the first half of the data, with three pairs very close together (separated by just over 30 days). The total mortality for the age group we are considering has greater variability before the spring of 1980, an effect which might explain the bulk of our added knots entering prior to this date.

In this setup, we have constrained the effect of TSP to enter the model linearly; that is, our only adaptation has been through adding knots to a cubic natural spline in the time index. This kind of model is often referred to as *semiparametric* in the sense that we fix part of the structure in our model (all of the variables in Table 4.1) and treat the remaining part as an unknown function. In this setup, we can judge the impact that our smooth component in time has on the effect of TSP. To make this precise, suppose $\boldsymbol{x}_0$ and $\boldsymbol{x}_1$ are two values of our predictors that vary only in the value of TSP. Given the log-linear structure of our model, we can easily calculate the relative mortality for these two conditions as `MHH: should TSO be TSP here?`

$$\exp\left[\, \widehat{g}(\boldsymbol{x}_1) - \widehat{g}(\boldsymbol{x}_0) \,\right] = \exp\left[\, \beta_{\mathrm{TSO}}(\mathrm{TSO}_1 - \mathrm{TSO}_0) \,\right] \,.$$

Variable	Type	DF	Coefficient
Intercept	1	1	3.04
TSP	linear	1	0.0007
Dew point	linear	1	0.001
Temp, lag	spline knot at 76	2	–
Time	spline knots at 2021, 2238, 2426, 2509, 2630, 5385	7	–

TABLE 4.3. A simple additive fit to the JHU data set.

In Dominici et al. (200,2004), `MHH: 200 doesn't seem right here` interest focuses on the percent mortality increase associated with `MHH: should TSO be TSP here?` $1000\beta_{\text{TSO}}$. For the model in Figure 4.2, the increase in mortality is 0.5%. As we add knots, this number changes, varying from 0.4% to 0.7%, with the value for the 11 knot model being again 0.5%. `MHH: I don't see that these percentages are correct. In particular, if` $\beta_{\texttt{TSO}} = 0.0007$`, then` $\exp(1000\beta_{\texttt{TSO}}) = \exp(.7) \doteq 2.014$`, which translates into over a 100% increase as I see it.`

Now that the basic mechanism for knot adaptation is clear, we can extend the reach of our procedure to include modeling similar to that followed for the tree example in the previous chapter. That is, we will conduct full stepwise addition and deletion of spline terms, where candidates are drawn not just from the time index, but from any of the available variables in Table 4.1. For simplicity, we again return to spaces of continuous, piecewise linear spines. Since Dominici et al. (200,2004) `MHH: 200 doesn't seem right here` focus on additive models, we will also only consider additive models. To guide the addition process, we made sure that a variable entered linearly before we added spline basis elements; that is, the function x_i was in the model before terms of the form $(x_i - t)_+$ were added. Essentially, we can follow the recipe in the previous chapter.

The final model consists of 12 basis functions and has an AIC `MHH: BIC?` value of 5950. The separate components are displayed in Table 4.3. The effect of the lagged temperature variable changes at 77.7 degrees Fahrenheit. Kelsall et al. (1997) also make use of a linear spline in lagged temperature, but choose a breakpoint at 80 degrees after some exploratory fitting with a cubic smoothing spline. Note that again, the effect of TSP enters linearly, and the percent mortality increase associated with $1000\beta_{\text{TSP}}$ is 0.7%. In this case, the stepwise selection procedure determined that a linear effect of TSP was sufficient, unlike the previous semiparametric modeling exercise in which we forced it to be linear. The knots associated with the time component has a cluster of knots near a run of missing values in the spring of

1980. In part, the problem comes from the fact that we forced only 7 points between candidate knots (following the advice in Section 3.6). Given the regularity of the design in the time variable in all regions except for this sequence of missing points, the forward selection method got "lost" following spurious structure. We can partially repair this problem by increasing the separation between candidate knots. Rather than pursue this in detail, we will consider another solution.

As the time component of our model is meant to stand for seasonal effects on mortality that are not captured by the other variables, we anticipate a more "regular" effect. The greedy scheme has difficulty discovering the effect of time in a bottom-up fashion, leaving unexplored a large fraction of the space of candidate models. To alleviate this problem, we take as our initial model a natural cubic spline having knots spaced equally in time: 15 knots corresponds to one knot per year, while 120 would allow for 8 knots per year. Starting from an initial fit in time, we then perform our full stepwise scheme with the piecewise linear basis functions used in the fit for Table 4.2. During deletion, we first restricted the removal process to just the added covariate effects, leaving the initial spline model unchanged.

When we ran this algorithm, we noted that for the most part TSP appeared at most linearly and in some cases was replaced by a different pollution variable. To make consistent comparisons across the different initial fits with time, we also forced TSP linearly into each of the models. In Figure 4.4 we present the fitted time effects (including the intercept term, setting all other variables to zero) for 3, 5, 15, 60, 90 and 180 equally spaced knots, corresponding to one knot every 5 years, 3 years, 1 year, 3 months, 2 months and 1 month, respectively. The three uppermost curves correspond to the three lowest density knot sets, while the lower curves correspond to the highest density knot sets. The models found by stepwise addition and deletion for the three low-density knot sets all resemble that in Table 4.2. They involve a linear effect of dew point and a spline in lagged temperature around 77 degrees. By contrast, the high-density knot sets involve a linear effect of dew point with a single-knot spline term in lagged dew point, not temperature. The pollution variable O3 also appears in these models, but only linearly.

The BIC values for these six fits is smallest when we place one knot every three years (a value of 5985). Unfortunately, this fit also leaves behind considerable time series structure in the residuals. The largest model (one knot per month, a total of 186 degrees of freedom in the model) has relatively clean residuals. The value of $\hat{\sigma}^2$ for the largest model is 1.02, a big improvement over our initial fit in Table 4.2. The percentage drop in mortality associated with $1000\beta_{\mathrm{TSP}}$ varies for these six models from 0.7% at its largest (least dense knots) to 0.5% at its smallest (most dense knots).

As a final elaboration of the models we have tried so far, we can consider performing stepwise addition from an initial model dense with knots in time, and then perform backward deletion on both the added variables as

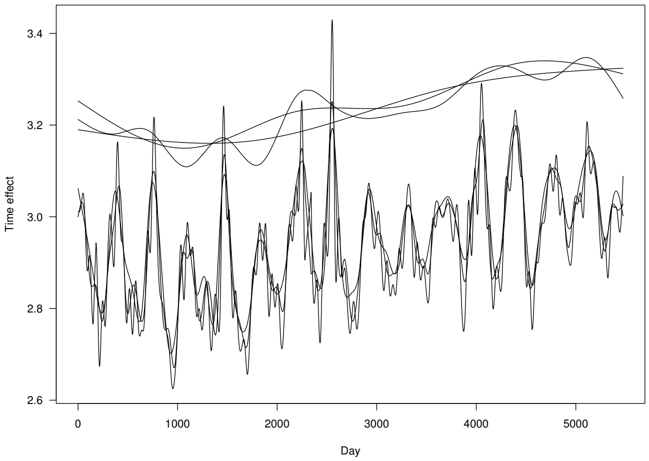

FIGURE 4.4. Comparing six fitted time effects. The upper curves correspond to low-density knot arrangements, while the lower curves all have at least one knot every three months. `MHH: I find the lower curves too hard to read. How about deleting the graph for knots every two months?`

Variable	Type	DF	Coefficient
Intercept	1	1	3.29
TSP	linear	1	0.0008
Dew point	linear	1	0.00002
NO2	linear	1	-0.0028
Dew × NO2	linear × linear	1	0.00005
Temp, lag	spline knot at 77	2	
Time	spline	23	

TABLE 4.4. A simple MARS-like fit to the JHU data set.

well as the spline terms. For the model in Table 4.4, we started with 60 initial knots (4 breakpoints per year, and an initial linear effect of TSP), and based on AIC selected a model with 36 knots. This was taken as our starting point and then standard addition and deletion on the covariates was performed.

The model in Table 4.4 has an BIC score of 5880, well below any fit so far. In terms of covariates, this model introduces the concentration of NO2 and its interaction with dew point. The percent increase in mortality associated with $1000\beta_{\mathrm{TSP}}$ `MHH: Something is missing here.` 0.8%. The autocorrelation function for the residuals from this model is between -0.03

this fit was run using interactions; i need to rerun it for just the additive model. sorry.

and 0.03 for up to 20 lags, indicating little unexplained structure in time. In Section **??** we will return to this example and apply a cross-validation technique to provide some guidance as to which of these models portrays a reasonable picture of the health risks associated with TSP.

4.1.2 Obesity and urban sprawl

Background

The percentage of adults and children in the U.S. who qualify as either overweight or obese has become the subject of great concern among health professionals. Obesity contributes to higher rates of diabetes and cardiovascular disease, and it has been linked to an increased risk of cancer. Technically, obesity is classified by a person's *body mass index* or BMI. This is calculated as the ratio of a person's weight to the square of their height

$$\text{BMI} = \frac{\text{Weight in pounds}}{(\text{Height in inches})^2} \times 703\,.$$

The Center for Disease Control has determined that people with BMI scores below 18.5 are underweight, between 18.5 and 24.9 are normal, between 25 and 30 are overweight, and over 30 are obese. For example, for someone who is $5'10''$, the overweight range is 174–209 pounds.[3]

Health professionals consider obesity to be a problem that rivals smoking in terms of its overall impact on society. This has spurred a number of researchers to investigate the underlying causes of obesity and, in particular, to consider what environmental conditions might be contributing to the problem. Lopez (2004) suggests that Americans living in areas of *urban sprawl* are at greater risk of becoming obese than those who live in more densely developed areas. Does poor urban planning encourage lifestyles that lead to obesity? Previously, Ewing, Schmid, Killingsworth, Zlot and Raudenbush (2003) found that residents in sprawling urban centers tend to rely more on automobiles and are less likely to walk during leisure time. Both Lopez (2004) and Ewing et al. (2003) base their studies on data collected by an annual telephone survey of adults conducted by the Center for Disease Control.

In 2000, the Behavioral Risk Factor Surveillance System (BRFSS) contacted over 180,000 adults in the U.S. and collected information on over 200 different health-related attributes. From each respondent's (self-reported) height and weight, their BMI was computed and a binary variable indicating whether or not this person was obese was derived. Lopez (2004) then considers the seven possible covariates listed in Table 4.1.2. `MHH: My .ps version here printed out as Table 4.1.2 rather than as Table 4.5`

[3] It should be added that BMI does not measure body fat and hence is only one piece of information about a person's health profile.

Age	in years
Sex	0/1, 1=F
Hispanic	0/1, 1=Y
AfAmerican	0/1, 1=Y
Education	1-6, ordinal
Income	1-8, ordinal
Sprawl	Index defined in (4.1.6)

TABLE 4.5. Covariates used from the BRFSS survey to predict obesity status.

`here. Any clues?` Six of these are taken directly from the BRFSS survey and are known to have a relationship with obesity. While education and income are categorical variables, their labels represent increasing years of education and annual income, respectively. Lopez (2004) introduces these factors as linear terms rather than as separate indicator functions representing each level. Because we are going to apply a flexible spline methodology, we have also decided to leave them as single covariates; if certain levels of these factors turn out to be important, knot adaptation can separate out the effect.

To quantify the concept of sprawl, Lopez (2004) considers the population density in census tracts, defining high and low population density regions; those tracts with more than 3,500 people per square mile are said to have high density. The cutoff value is set at the point where people in the tract begin using forms of transportation other than automobiles. With this notion of population density, Lopez (2004) constructed a sprawl index (SI) as follows:

$$\text{SI} = \frac{\text{Percentage of population in high density tracts}}{\text{Percentage of population in low density tracts}} \,. \qquad (4.1.6)$$

`MHH: This discussion seems inverted to me. In particular, it seems to me that the most dense urban areas should have high SI scores, not low SI scores, since their numerators would be high and their denominators low.` The index was then normalized so that it ranged between 1 and 100. For example, in California the most dense urban areas include Los Angeles and Orange Counties (scores less than 15); while the SI was high for Sonoma County (a score over 50). In New York State, the counties surrounding New York City had the lowest values of SI (roughly 6.7), while Dutchess County exhibited significantly more sprawl (scoring over 70). Hudson County in New Jersey was the most densely packed region of the country (with a score of about 4) and Calhoun County in Alabama was maximally sprawling with a score of 100. Using data from the 2000 Census, Lopez (2004) computed the SI for 330 metropolitan regions in the U.S. and combined these scores to the 2000 BRFSS survey responses. The resulting data set will be used to consider the association between obesity and sprawl.

Logistic regression

We now create a model for the simple binary outcome classifying people as either obese or not. In this setting, it is common to employ *logistic regression* to assess the effect of variables like sprawl. Let Y denote a person's weight status and let $\boldsymbol{X}$ be a vector of the covariates listed in Table 4.1.2. Then, we express the probability that a person is obese as a function of the covariates

$$P(Y = y|\boldsymbol{X} = \boldsymbol{x}) = \pi(\boldsymbol{x})^y (1 - \pi(\boldsymbol{x}))^{1-y}\,, \qquad y \in \{0, 1\},$$

so that the conditional mean and variance of Y are given by

$$E(Y|\boldsymbol{X} = \boldsymbol{x}) = \pi(\boldsymbol{x}) \quad \text{and} \quad \mathrm{var}(Y|\boldsymbol{X} = \boldsymbol{x}) = \pi(\boldsymbol{x})(1 - \pi(\boldsymbol{x}))\,.$$

Our goal is to learn the important features of the *risk function* π. As was the case with Poisson regression, we will find it easier to work with a transformation of π rather than with π itself. Here, we will model

$$\mathrm{logit}(\pi) = \log \frac{\pi}{1 - \pi}\,. \tag{4.1.7}$$

We choose a model for logit(π) again using some linear space $\mathbb{G}$ with basis $B_1, \ldots, B_J$, where each $g \in \mathbb{G}$ can be expressed uniquely as a sum

$$g(\boldsymbol{x}; \boldsymbol{\beta}) = \beta_1 B_1(\boldsymbol{x}) + \cdots \beta_J B_J(\boldsymbol{x})$$

for a coefficient vector $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_J)$. Inverting (4.1.7), we have that

$$\pi(\boldsymbol{x}) = \frac{\exp g(\boldsymbol{x})}{1 + \exp g(\boldsymbol{x})}\,.$$

Then given observations $(\boldsymbol{X}_1, Y_n), \ldots, (\boldsymbol{X}_n, Y_n)$, we evaluate members of $\mathbb{G}$ based on the log-likelihood

$$\begin{aligned}\ell(g) = \sum_{i=1}^{n} \Big(Y_i \log \frac{\exp g(\boldsymbol{x}_i)}{1 + \exp g(\boldsymbol{x}_i)} + (1 - Y_i) \log \frac{1}{1 + \exp g(\boldsymbol{x}_i)} \Big) \\ = \sum_{i=1}^{n} \big(Y_i g(\boldsymbol{x}_i) - \log\left[1 + \exp g(\boldsymbol{x}_i)\right] \big). \quad (4.1.8)\end{aligned}$$

As was done for Poisson regression, we let $\widehat{g} = \mathrm{argmax}_{g \in \mathbb{G}}\, \ell(g)$ denote the maximum likelihood estimate of logit(π). Expressing the log-likelihood in terms of $\boldsymbol{\beta}$, we have that

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left(Y_i \sum_j \beta_j B_j(\boldsymbol{x}_i) - \log \left[1 + \exp \left(\sum_j \beta_j B_j(\boldsymbol{x}_i) \right) \right] \right)$$

	Coefficient	Relative risk
(Intercept)	−0.866	
Age	0.006	1.006
Sex	−0.086	0.918
Hispanic	0.128	1.137
AfAmerican	0.665	1.944
Education	−0.147	0.863
Income	−0.048	0.953
Sprawl	0.002	1.002

TABLE 4.6. Coefficients from a simple logistic regression fit to the seven covariates. `MHH: Shouldn't Relative risk be odds ratio here?`

and $\widehat{g} = g(\boldsymbol{x}; \widehat{\boldsymbol{\beta}})$ where $\widehat{\boldsymbol{\beta}} = \mathrm{argmax}_{\boldsymbol{\beta} \in \mathbb{R}^J}\, \ell(\boldsymbol{\beta})$. While there is no closed form expression for this maximizer, it is possible to derive a simple iterative scheme to find $\boldsymbol{\beta}$. We will describe it in more detail in the next section. As with Poisson regression, it will consist of a series of weighted least squares fits.

Modeling obesity

The BRFSS is a complex survey with a stratified sampling plan. The CDC has a weighting scheme to account for differences in coverage and response rates. In this first look at the data, we are going to ignore these two effects and instead highlight the methodology. At the end of this section we will consider the use of weights in our analysis and indicate how it affects our underlying adaptive procedure. In terms of the actual results, the effects change in magnitude but the overall picture, the important variables and their functional form, remains the same.

Do we think introducing the weighted results now is too much?

Following Lopez (2004), we first drop all of the records for people living outside of the 330 metropolitan areas under study. This yields 108,661 records. We then remove all respondents with missing values leaving us with 90,639 observations. (In a thorough treatment of these data, we should consider the impact that so many missing values might have on our fit.) Roughly 20% of the people represented in this sample are classified as obese. A larger proportion of African Americans (32%) and Hispanics (22%) were classified as obese. Next, we fit a simple logistic regression taking g to to be a linear function of the seven covariates. In Table 4.1.2 `MHH: In my .ps version this shows up as Table 4.1.2, not Table 4.6. Any clues?` , we present the coefficients from this fit as well as the relative risk `MHH: shouldn't this be odds ratio instead of relative risk?`. For each covariate, we compute the latter by considering conditions $\boldsymbol{x}_0$ and $\boldsymbol{x}_1$ and evaluating

$$\frac{\pi(\boldsymbol{x}_0)}{\pi(\boldsymbol{x}_1)} = \exp\left[\widehat{g}(\boldsymbol{x}_0) - \widehat{g}(\boldsymbol{x}_1)\right]$$

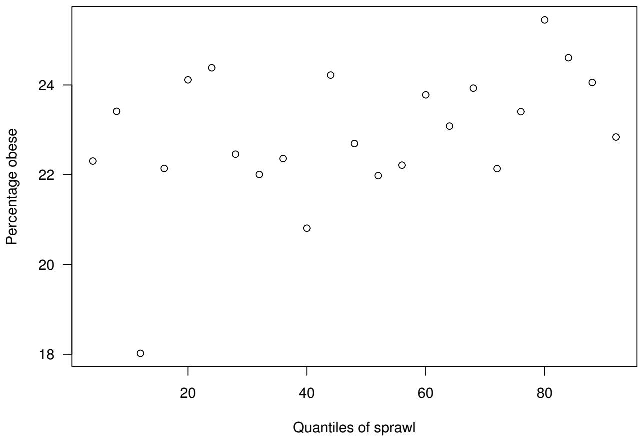

FIGURE 4.5. Percentage of the population that is obese as a function of SI, the index of sprawl.

`MHH: shouldn't the left side of the above equation be the odds ratio instead of the relative risk?` For each row of Table 4.1.2 `MHH: This also showed up as Table 4.1.2.` (excluding the intercept), we take $\boldsymbol{x}_0$ and $\boldsymbol{x}_1$ to differ by 1 in a single covariate, and the relative risk `MHH: odds ratio?` measures the effect of a unit change in that variable.[4] Lopez (2004) demonstrates that each of these coefficients is strongly significant, albeit the relative risk `MHH: odds ratio?` associated with regions of urban sprawl is quite small `MHH: ?` . He argues that even small effects are important when you consider the size of the population at risk. In Figure 4.5 we present a plot of the percentage of the population that is obese as a function of sprawl. Here we have divided into 25 groups based on evenly spaced quantiles of SI `MHH: Please clarify.`.

Starting from this simple model, we first consider an additive fit to the data. Here, we apply the simple additive methodology outlined in Section 3.4.3. As with the Poisson regression example, we again work with linear splines. A greedy algorithm introduces basis functions one at a time, adding simple linear effects before spline elements. After performing stepwise addition to a model of size 20 was found `MHH: please clarify`, we performed stepwise deletion and chose the best model according to AIC `MHH: BIC?`. In Table 4.7 we present the separate effects in the final model.

[4] These results differ somewhat from those in Lopez (2004) because we have not incorporated the survey weights. However, each of our risk figures are within 5% of his weighted estimates, with the single exception of Sex; the coefficient has the right sign, but the risk is 40% higher with our calculation.

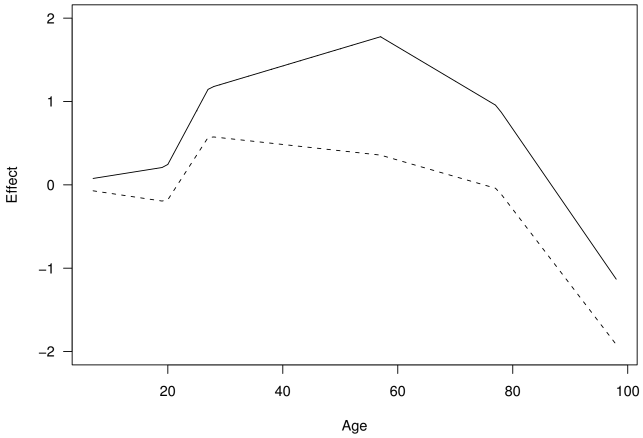

FIGURE 4.6. The effect of age on the probability of being obese. Estimates from the simple additive model (solid line) and the fit including interaction terms (dashed line).

Note that two factors, age and income involve spline terms with 4 knots and 1 knot, respectively. The actual curve for age is plotted in Figure 4.6 (solid line). The dependence on income is very similar to a straight line with negative slope and we omit the figure. The effect of age has a slow increase until the early 20s, a faster rate of increase through the 50s, and then a decrease through the 60s and beyond. Presumably, the increased health risks associated with obesity manifest themselves and explain this drop.

While the additive model is an improvement over the model fit by Lopez (2004) in terms of AIC `MHH: BIC?`, the coefficients of the linear terms are not dramatically different. We have lost some of effect of sprawl, reducing

	Type	DF	Coefficient
(Intercept)	1	1	−1.617
Age	spline	5	–
Sex	linear	1	−0.095
Hispanic	–	–	–
AfAmerican	linear	1	0.608
Education	linear	1	−0.159
Income	spline	2	–
Sprawl	linear	1	0.0014

TABLE 4.7. Fitting an additive model to the BRFSS obesity data.

	Type	DF
Age × Education	composite	2
Age × Sex	simple	1
Income × Education	composite	2
Income × Sex	composite	2
Income × AfAmerican	simple	1
AfAmerican × Sex	simple	1

TABLE 4.8. Modeling with pairwise interactions.

the relative risk associated with SI from 1.002 to 1.00144. Further improvements to the model can be made by considering interactions. We next performed the fitting routine outlined in Section 3.4, allowing for interactions involving at most two factors. Again we fit with linear splines and again we entertained a model consisting of at most 25 terms (experiments with larger maximal model sizes did not change the final model fit). In this case, the model minimizing AIC contained the additive model as a subset; the only spline terms involved age and income and these had the same knot sequences as were found for the additive case. The curve associated with the effect of age is plotted with a dashed line in Figure 4.6. A bivariate plot of the interaction between Age and Education is given in Figure 4.7.

Notice that every variable except SI is involved in some kind of interaction. This makes it easy to compare the relative risks `MHH: odds ratios?` associated with sprawl across all three of the models we have considered so far: the simple logistic fit and the additive and interaction spline models. As we better capture the structure of the other covariates, we have reduced an already small effect. From an initial relative risk of 1.002, we drop it to 1.00144 and finally to 1.00146 for the interaction model. At this point it seems sensible to question whether these results are practically significant.

4.2 GLMs and approximation spaces

4.2.1 Conditional Likelihood for a GLM

We begin by introducing a large class of likelihoods that will be used to capture the conditional distribution of Y given some value of the input vector. In the case of the normal linear model, Y was (conditionally) a Gaussian random variable. Here, for fixed values of the parameters θ and ϕ, we consider univariate densities (or probability functions in the case of discrete data) of the form

$$p(y, \theta, \phi) = \exp\Big(\frac{\theta y - b(\theta)}{a(\phi)} + c(y, \phi)\Big), \tag{4.2.1}$$

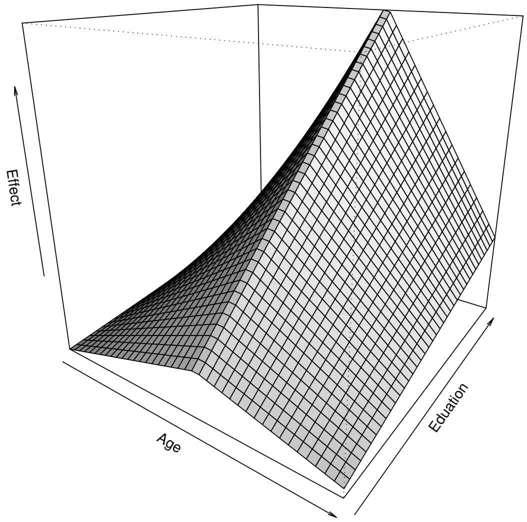

FIGURE 4.7. Surface estimate of the interaction between Age and Income. `MHH: Here education seems to have a positive effect, but in Table 4.7 it has a negative effect, which seems more plausible.`

where y, θ and ϕ and the ranges of a, b and c are real-valued. The function a is taken to be strictly positive over the range of possible values of ϕ. To make this concrete, we revisit the two examples from the previous section. First, the probability function for the Poisson distribution with mean λ can be written as

$$\frac{\lambda^y e^{-\lambda}}{y!} = \exp(y \log \lambda - \lambda - \log y!), \qquad y \in \{0, 1, 2, \ldots\}.$$

After setting $\theta = \log \lambda$ and $\phi = 1$, the individual components of (4.2.1) are easily seen to be given by

$$a(\phi) = 1, \quad b(\theta) = e^\theta \quad \text{and} \quad c(y, \phi) = -\log y!. \tag{4.2.2}$$

The Bernoulli distribution is our second example from the previous section. Let π denote the probability of success on a given trial. Then the corresponding probability function is given by

$$\pi^y (1-\pi)^{1-y} = \exp\Big(y \log \frac{\pi}{1-\pi} + \log(1-\pi)\Big), \qquad y \in \{0, 1\}$$

Eyeing expression (4.2.1), we set

$$\theta = \log \frac{\pi}{1-\pi} \qquad \text{and} \qquad \phi = 1,$$

so that

$$a(\phi) = \phi, \quad b(\theta) = \log(1 + e^{\theta}) \quad \text{and} \quad c(y, \phi) = 1. \tag{4.2.3}$$

We refer $\log[\pi/(1-\pi)]$ as the *logit* of π.

In addition to the Poisson and Bernoulli and general binomial distributions, the family (4.2.1) also includes the gamma, Gaussian, and inverse-Gaussian densities. We can say quite a lot about this class, and we begin with some elementary moment results. For fixed values of the parameters θ and ϕ, let Y have a distribution of the form (4.2.1). Then

$$1 = \int p(y, \theta, \phi)\, dy = \int \exp\Big(\frac{\theta\, y - b(\theta)}{a(\phi)} + c(y, \phi)\Big)\, dy,$$

where we take the integral to be a sum if we are dealing with discrete data. Then, as $b(\theta)$ and $a(\phi)$ do not depend on y, we can multiply both sides by $\exp[b(\theta)/a(\phi)]$. Taking logarithms, we arrive at the equality

$$\frac{b(\theta)}{a(\phi)} = \log \int \exp\Big(\frac{\theta y}{a(\phi)} + c(y, \phi)\Big)\, dy. \tag{4.2.4}$$

Now, assuming we can differentiate both sides with respect to θ, we derive an expression for the expectation of Y:

$$\begin{aligned} b'(\theta) &= \frac{\displaystyle\int y \exp\Big(\frac{\theta y}{a(\phi)} + c(y, \phi)\Big) dy}{\displaystyle\int \exp\Big(\frac{\theta y}{a(\phi)} + c(y, \phi)\Big) dy} \\ &= \int y\, \exp\Big(\frac{\theta y - b(\theta)}{a(\phi)} + c(y, \phi)\Big) dy \\ &= \int y\, p(y, \theta, \phi)\, dy\,, \end{aligned}$$

from which we see that

$$EY \;=\; b'(\theta)\,. \tag{4.2.5}$$

Taking another derivative of (4.2.4) and simplifying much as we did above, we find that the variance of Y is given by

$$\operatorname{var} Y = b''(\theta)\, a(\phi)\,. \tag{4.2.6}$$

Using the definitions of θ, b and a for the Poisson and binomial families, it is straightforward to verify these two moment expressions directly. Because of (4.2.6), we refer to ϕ as a *dispersion parameter.* It is also common to refer to θ as the *natural* or *canonical* parameter.

The densities given by (4.2.1) are said to specify the *random component* of a GLM in the sense that this family is used to describe the conditional distribution of Y given some value of the input vector $\boldsymbol{X}$. In a classical GLM, the dependence of Y on $\boldsymbol{X}$ is captured by letting the parameter θ vary with $\boldsymbol{x}$:

$$\theta = \theta(\boldsymbol{x}), \qquad \boldsymbol{x} \in \mathcal{X}. \tag{4.2.7}$$

Substituting this expression in (4.2.1), we can write the conditional distribution of Y given $\boldsymbol{X} = \boldsymbol{x}$ as

$$p(y, \theta(\boldsymbol{x}), \phi) = \exp\Big(\frac{y\theta(\boldsymbol{x}) - b(\theta(\boldsymbol{x}))}{a(\phi)} + c(y, \phi)\Big). \tag{4.2.8}$$

To see that this is sensible proposition, we consider a somewhat more familiar member of the GLM family, the normal linear model. Rewriting the density of a Gaussian random variable with mean μ and variance σ^2, we arrive at

$$\frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big(-\frac{(y-\mu)^2}{2\sigma^2}\Big) \;=\; \exp\Big(\frac{y\mu - \mu^2/2}{\sigma^2} - \frac{y^2/\sigma^2 + \log(2\pi\sigma^2)}{2}\Big)$$

for $y \in \mathbb{R}$. Here, $\theta = \mu$ and $\phi = \sigma^2$, so that the individual components of (4.2.1) are easily seen to be given by

$$a(\phi) = \phi, \quad b(\theta) = \theta^2/2 \quad \text{and} \quad c(y, \phi) = -\frac{y^2/\sigma^2 + \log(2\pi\sigma^2)}{2}.$$

Therefore, the parameter θ appears as the mean of the normal family and ϕ as its variance. By letting θ vary with $\boldsymbol{x}$, we create the estimation setup studied in the previous chapter, where the covariates $\boldsymbol{X}$ influenced the distribution of the response Y only by shifts in mean. Flexible linear spaces were used to characterize the effect of the covariates, and functional ANOVA models provided us with a formalism for exploring the (mean) relationship between Y and collections of functions of $\boldsymbol{X}$. In the next section, we extend these ideas to a GLM's.

4.2.2 Canonical linear regression and approximation spaces

Using the moment relationship (4.2.5), we can write the conditional means and variances for a GLM as

$$\mu(\boldsymbol{x}) = E(Y|\boldsymbol{X} = \boldsymbol{x}) = b'\big(\theta(\boldsymbol{x})\big) \tag{4.2.9}$$

and

$$\sigma^2(\boldsymbol{x}) = \mathrm{var}(Y|\boldsymbol{X} = \boldsymbol{x}) = a(\phi)b''\big(\theta(\boldsymbol{x})\big). \tag{4.2.10}$$

Given these expressions, we have a choice. In the previous chapter, we modeled the conditional mean of Y directly (in that case μ was denoted by the unknown regression function f), making use of adaptively determined linear spaces to highlight its essential features. In the case of a GLM, it can be difficult numerically to work directly with the conditional mean. For example, the structure of the Poisson model requires that $\mu(\boldsymbol{x})$ be nonnegative for all values of $\boldsymbol{x}$; while for the binomial model $\pi(\boldsymbol{x})$ must map to the open interval $(0, 1)$. These *range constraints* can complicate analysis.

A computationally attractive alternative to the conditional mean is the canonical parameter $\theta(\boldsymbol{x})$. Comparing the relationships above with (4.0.1), we see that unlike the simple linear model, the conditional mean of Y is related to θ via a transformation. Recalling our assumption that a be a positive function, we know from the second expression above that b'' is positive, and hence b' is strictly increasing. This means that an inverse $(b')^{-1}$ exists. For the Poisson and binomial cases, $(b')^{-1}$ is given by

$$\lambda = e^{\theta} \quad \text{and} \quad \pi = \frac{e^{\theta}}{1 + e^{\theta}}, \tag{4.2.11}$$

respectively. Therefore, a model for θ can be transformed into a model for the conditional mean. Notice that in each case, θ is allowed to range over the entire real line. Therefore, from a numerical perspective, we prefer to work with θ over the conditional mean.

As in the previous chapter, we introduce a linear approximation space $\mathbb{G}$ for use as a source of possible descriptions for $\boldsymbol{\theta}(\boldsymbol{x})$. Therefore, by substitution into the (conditional) GLM density, we have that

$$p(y, g(\boldsymbol{x}), \phi)\,, \quad g \in \mathbb{G}\,. \tag{4.2.12}$$

Then, given observations $(\boldsymbol{X}_1, Y_1), \ldots, (\boldsymbol{X}_n, Y_n)$, we construct the likelihood

$$\ell(g) = \sum_{i=1}^{n} \log p\left(Y_i, g(X_i), \phi\right)\,, \qquad g \in \mathbb{G}\,. \tag{4.2.13}$$

Let $\mathbb{G}$ have a basis $B_1, \ldots, B_J$ so that any $g \in \mathbb{G}$ can be written in the form

$$g(\boldsymbol{x}) = g(\boldsymbol{x}; \boldsymbol{\beta}) = \beta_1 B_1(\boldsymbol{x}) + \cdots + \beta_J B_J(\boldsymbol{x})\,,$$

for some choice of the vector $\boldsymbol{\beta} = (\beta, \ldots, \beta)$. Then, maximizing (4.2.13) over g is equivalent to finding $\widehat{\boldsymbol{\beta}} = \text{argmax}_{\boldsymbol{\beta}}\, \ell(\boldsymbol{\beta})$ and taking $g(\boldsymbol{x}; \widehat{\boldsymbol{\beta}})$ as an estimate of $\theta(\boldsymbol{x})$. Finding $\widehat{\boldsymbol{\beta}}$ is more difficult than finding an OLS projection. In fact, we will see in the next section that, under certain conditions, the maximizer can be obtained by a series of OLS projections. To recover an estimate for the conditional mean $\mu(\boldsymbol{x})$, we simply apply the inverse transformation $\hat{\mu}(\boldsymbol{x}) = b'\big(g(\boldsymbol{x}; \widehat{\boldsymbol{\beta}})\big)$, which we know must exist given the discussion above.

The function $(b')^{-1}$ is referred to as a *link function* and is commonly denoted by the symbol η. It is said to transform the conditional mean of Y to a *scale* on which the underlying estimation problem is simpler. To further connect with standard terminology, we have introduced the linear space $\mathbb{G}$ as a tool for capturing the features of the *structural* or *systematic component* of a GLM. Recall that the conditional densities (4.2.1) specify the random component, having a given conditional mean and variance. The function η then provides a link between the systematic and random components in a GLM. So far we have only discussed one link function, the so-called *natural link* $(b')^{-1}$. In general, it is common to either let computational considerations or the specifics of the problem under study dictate the appropriate transformation η. In the next section, we will discuss this approach in more depth, introducing different link functions for the case of binomial data.

4.2.3 Link functions

The exponential family form (4.2.8) describes the distribution of Y given some value of the covariates $\boldsymbol{X}$, with the associated conditional mean μ (4.2.9) and variance σ^2 (4.2.10). We have referred to this as the random component of a GLM. In the previous section, we identified the canonical parameter $\theta(\boldsymbol{x})$ as a sensible candidate for modeling. However, given any link function η, we can define the systematic component of a GLM to be given by $f = \eta(\mu)$. We then introduce linear approximation spaces $\mathbb{G}$ appropriate for capturing the features evident in f. The motivation for favoring one link η over another depends on the underlying modeling problem. For binary data, an alternative to the logit of π is the so-called *probit link*

$$\eta(\pi) \;=\; \Phi^{-1}(\pi)\,, \tag{4.2.14}$$

where $\Phi(\cdot)$ is the Gaussian cumulative distribution function. For this type of data, the inverse of any continuous, cumulative distribution function having support on the entire real line can be used as a link. As we will see, the logit can be recommended on the basis of its favorable mathematical properties, simplifying both theoretical analysis and numerical algorithms. We will return to these issues in Section 4.3.

The function f defined in (4.2.14) is the target of our analysis, and we take as an estimate some member of a linear space $\mathbb{G}$. The fitting criterion will be based on the likelihood of the data rather than OLS. This requires expressing the conditional distribution of Y in terms of f rather than the canonical parameter θ. If we choose some arbitrary η, then the density (4.2.12) is more complicated. The function

$$\eta^{-1}\left((b')^{-1}(g)\right)$$

is now on the scale of θ, and substituting this into (4.2.1) yields the rather clumsy expression

$$p\left(y,\, \eta^{-1}\left((b')^{-1}(g)\right), \phi\right)\,, \quad g \in \mathbb{G}\,.$$

where p is defined in (4.2.1). Then, given observations $(\boldsymbol{X}_1, Y_1), \ldots, (\boldsymbol{X}_n, Y_n)$, we construct the likelihood according to (4.2.13). Let $\mathbb{G}$ have a basis $B_1, \ldots, B_J$ so that any $g \in \mathbb{G}$ can be written in the form

$$g(\boldsymbol{x}) = g(\boldsymbol{x}; \boldsymbol{\beta}) = \beta_1 B_1(\boldsymbol{x}) + \cdots + \beta_J B_J(\boldsymbol{x}) \,,$$

for some choice of the vector $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_J)$. Then, maximizing (4.2.13) over g is equivalent to finding $\widehat{\boldsymbol{\beta}} = \operatorname{argmax}_{\boldsymbol{\beta}} \ell(\boldsymbol{\beta})$ and taking $g(\boldsymbol{x}; \widehat{\boldsymbol{\beta}})$ as an estimate for f.

It should be clear the simplifications that take place when $\eta = (b')^{-1}$, and hence this choice is referred to as the *canonical link* because the target of our estimation schemes f becomes the canonical parameter θ. In fact, such problems are often referred to as *canonical linear regression models.* The functions given in (4.2.11) are the canonical links for the Poisson and binomial families, respectively. These can be obtained by directly computing $(b')^{-1}$ from the expressions in (4.2.2) and (4.2.3).

4.3 Estimation and adaptation

In the previous section, we specified the dependence of a response Y, with conditional distribution of the form (4.2.1), by modeling the canonical parameter for the family as a function $g \in \mathbb{G}$. Except in special cases like the normal distribution (or when $\mathbb{G}$ is saturated; that is, when the dimension of the restriction of $\mathbb{G}$ to the design set $\mathcal{X}' = \{\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n\}$ equals $\#(\mathcal{X}')$), there is not a closed form expression for the function g that maximizes the likelihood (4.2.13). This is true not only for GLMs, but also for every other modeling context introduced in the remaining chapters of this text. Therefore, we have decided to collect several generic tools for computing maximum likelihood estimates and performing stepwise addition and deletion of variables. We will use this framework to derive the adaptive fitting schemes employed in the Poisson and logistic regression examples introduced at the beginning of the chapter.

4.3.1 Quadratic approximations

In both this and the previous chapters, we have introduced linear approximation spaces $\mathbb{G}$, and our attention has focused on first finding a suitable estimate $\widehat{g} \in \mathbb{G}$. Then, through simple stepwise procedures, we construct chains of candidate spaces in which each approximation space $\mathbb{G}_1$ differs from its predecessor $\mathbb{G}_0$ by the addition or deletion of one or more basis functions. To be computationally feasible, we have to conduct these fundamental operations quickly; any single model might include tens of variables, and at each step in the selection process there are typically many candidate basis elements that can be added or deleted. For simple linear models, we were able to derive fast updating schemes that let us entertain a large number of alterations to any single model efficiently.

Maximum likelihood estimation

We now consider working with a generic likelihood function $\ell(g)$ for g in a J-dimensional space $\mathbb{G}$. As we have done several times, this likelihood can also be expressed in terms of the coefficient vector $\boldsymbol{\beta}$, where

$$g = g(\boldsymbol{x};\boldsymbol{\beta}) = \beta_1 B_1(\boldsymbol{x}) + \cdots + \beta_J B_J(\boldsymbol{x})\,.$$

Suppose that our log-likelihood function is reasonably smooth in some neighborhood of $\boldsymbol{\beta}_0$. We can apply a multivariate Taylor expansion to approximate $\ell(\boldsymbol{\beta})$ in a neighborhood around $\boldsymbol{\beta}_0$. The quadratic expansion around $\boldsymbol{\beta}_0$ involves first and second partial derivatives of $\ell(\boldsymbol{\beta})$ evaluated at $\boldsymbol{\beta}_0$. Define the J-vector $\nabla\ell(\beta)$ and the $J \times J$ matrix $\mathbf{H}(\boldsymbol{\beta})$ by

$$[\nabla\ell(\boldsymbol{\beta})]_j = \frac{\partial}{\partial\beta_j}\ell(\boldsymbol{\beta}) \quad \text{and} \quad [\mathbf{H}(\boldsymbol{\beta})]_{jk} = \frac{\partial^2}{\partial\beta_j\partial\beta_k}\ell(\boldsymbol{\beta})\,.$$

The matrix $\mathbf{H}(\boldsymbol{\beta})$ is called the Hessian; its negative $\boldsymbol{I}(\boldsymbol{\beta}) = -\mathbf{H}(\boldsymbol{\beta}$ is referred to as the Fisher information matrix. Recall from Chapter 2 that the quadratic expansion Q_0 around $\boldsymbol{\beta}_0$ is given by

$$Q_0(\boldsymbol{\beta}) = \ell(\boldsymbol{\beta}_0) + [\nabla\ell(\boldsymbol{\beta}_0)]^{\mathrm{T}}(\boldsymbol{\beta}-\boldsymbol{\beta}_0) - \frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)^{\mathrm{T}}\boldsymbol{I}(\boldsymbol{\beta}_0)(\boldsymbol{\beta}-\boldsymbol{\beta}_0). \quad (4.3.1)$$

Now, if $\mathbf{H}(\boldsymbol{\beta}_0)$ is negative definite (or, equivalently, if $\boldsymbol{I}(\boldsymbol{\beta}_0)$ is positive definite), the quadratic function Q_0 has a unique global maximum. This occurs at the point

$$\boldsymbol{\beta}_1 = \boldsymbol{\beta}_0 + [\boldsymbol{I}(\boldsymbol{\beta}_0)]^{-1}\nabla\ell(\boldsymbol{\beta}_0). \quad (4.3.2)$$

The *update* from $\boldsymbol{\beta}_0$ to $\boldsymbol{\beta}_1$ is the essential ingredient in the Newton–Raphson method for finding the maximum of the likelihood function $\ell(\boldsymbol{\beta})$. Given some starting value $\boldsymbol{\beta}_0$, we iteratively apply these updates creating a sequence of values $\boldsymbol{\beta}_0, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \ldots$, where

$$\boldsymbol{\beta}_{k+1} = \boldsymbol{\beta}_k + [\boldsymbol{I}(\boldsymbol{\beta}_k)]^{-1}\nabla\ell(\boldsymbol{\beta}_k). \quad (4.3.3)$$

If the log-likelihood function is well behaved, then the successive $\boldsymbol{\beta}_k$ should get closer and closer to the maximum likelihood estimate $\widehat{\boldsymbol{\beta}}$. For example, if the log-likelihood is strictly concave and we have a reasonably good starting value, Newton–Raphson is guaranteed to converge to $\widehat{\boldsymbol{\beta}}$.[5]

[5] With slight alterations to this scheme, we can guarantee that our sequence will converge to $\widehat{\boldsymbol{\beta}}$ no matter where we start. See the material in Chapter 2 related to *step-halving.*

Addition of a single basis function

In our adaptive methodology, spline basis functions are added sequentially so as to create the greatest drop in the log-likelihood. The number of candidates we could consider at each step would be severely limited if we had to conduct Newton–Raphson iterations for each. Instead, we again turn to the quadratic Taylor expansion to approximate the change in log-likelihood by introducing another basis function. Using (4.3.1) and (4.3.2), we find that

$$Q_0(\boldsymbol{\beta}_1) = \ell(\boldsymbol{\beta}_0) + \frac{1}{2}[\nabla\ell(\boldsymbol{\beta}_0)]^{\mathrm{T}}[\boldsymbol{I}(\boldsymbol{\beta}_0)]^{-1}\nabla\ell(\boldsymbol{\beta}_0)\,.$$

Therefore, the increase in the quadratic approximation Q_0 in moving from $\boldsymbol{\beta}_0$ to $\boldsymbol{\beta}_1$ is given by

$$2[Q_0(\boldsymbol{\beta}_1) - Q_0(\boldsymbol{\beta}_0)] = [\nabla\ell(\boldsymbol{\beta}_0)]^{\mathrm{T}}[\boldsymbol{I}(\boldsymbol{\beta}_0)]^{-1}\nabla\ell(\boldsymbol{\beta}_0). \tag{4.3.4}$$

If $\boldsymbol{\beta}_0$ is the maximum likelihood estimate $\widehat{\boldsymbol{\beta}}_0$ in a subspace $\mathbb{G}_0$, then the right side of this expression is known as the Rao (score) statistic. It is used for testing the hypothesis that the "true" value of $\boldsymbol{\beta}$ lies in $\mathbb{G}_0$.

To see why this is helpful for stepwise addition of basis functions, suppose our candidate model is $\mathbb{G}_1$, which consists of a single basis function addition to $\mathbb{G}_0$, and suppose that the coefficient vector associated with $\mathbb{G}_1$ places this basis function in position $J+1$. Let $\widehat{\boldsymbol{\beta}}_0$ correspond to the MLE in the subspace $\mathbb{G}_0$, so that $[\widehat{\boldsymbol{\beta}}_0]_{J+1} = 0$. With this as our starting value, we can use (4.3.4) to approximate the increase in log-likelihood by considering the larger space $\mathbb{G}_1$. To compute this test statistic, we do not have to perform any more iterations, but instead compute a few extra inner products. We will see this more clearly in the context of GLMs in the next section.

Deletion of a single basis function

In the last chapter, we saw that deleting spline basis functions was equivalent to imposing one or more linear constraints on the coefficients. If we choose to delete a basis function that has the smallest decrease in log-likelihood, then we are again forced into performing a set of Newton–Raphson iterations for each candidate. Again, we turn to the quadratic approximation to help simplify the task.

Let Q be the quadratic approximation to the log-likelihood function about the maximum likelihood estimate $\widehat{\boldsymbol{\beta}}$ of the coefficient vector corresponding to a basis of $\mathbb{G}$, and let $\mathbb{G}_0$ be the subspace of $\mathbb{G}$ corresponding to those coefficient vectors such that $\boldsymbol{A\beta} = \boldsymbol{0}$, where $\boldsymbol{A}$ has full rank. Then the maximum of Q corresponding to $\mathbb{G}_0$ occurs uniquely at

$$\widehat{\boldsymbol{\beta}}_{00} = \widehat{\boldsymbol{\beta}} - \boldsymbol{I}^{-1}(\widehat{\boldsymbol{\beta}})\boldsymbol{A}^{\mathrm{T}}[\boldsymbol{A}\boldsymbol{I}^{-1}(\widehat{\boldsymbol{\beta}})\boldsymbol{A}^{\mathrm{T}}]^{-1}\boldsymbol{A}\widehat{\boldsymbol{\beta}}\,. \tag{4.3.5}$$

To verify (4.3.5), observe that $\boldsymbol{A}\widehat{\boldsymbol{\beta}}_{00} = \boldsymbol{0}$. Moreover, $\boldsymbol{0} = \nabla Q(\widehat{\boldsymbol{\beta}}) = \nabla Q(\widehat{\boldsymbol{\beta}}_{00}) - \boldsymbol{I}(\widehat{\boldsymbol{\beta}})(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{00})$, so $\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{00} = \boldsymbol{I}^{-1}(\widehat{\boldsymbol{\beta}})\nabla Q(\widehat{\boldsymbol{\beta}}_{00})$. By the Lagrange multiplier theorem, there is a vector $\boldsymbol{\lambda}$ such that $\nabla Q(\widehat{\boldsymbol{\beta}}_{00}) = \boldsymbol{A}^{\mathrm{T}}\boldsymbol{\lambda}$. Thus $\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{00} = \boldsymbol{I}^{-1}(\widehat{\boldsymbol{\beta}})\boldsymbol{A}^{\mathrm{T}}\boldsymbol{\lambda}$, so $\boldsymbol{A}\widehat{\boldsymbol{\beta}} = \boldsymbol{A}(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{00}) = \boldsymbol{A}\boldsymbol{I}^{-1}(\widehat{\boldsymbol{\beta}})\boldsymbol{A}^{\mathrm{T}}\boldsymbol{\lambda}$ and hence $\boldsymbol{\lambda} = [\boldsymbol{A}\boldsymbol{I}^{-1}(\widehat{\boldsymbol{\beta}})\boldsymbol{A}^{\mathrm{T}}]^{-1}\boldsymbol{A}\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{00} = \boldsymbol{I}^{-1}(\widehat{\boldsymbol{\beta}})\boldsymbol{A}^{\mathrm{T}}[\boldsymbol{A}\boldsymbol{I}^{-1}(\widehat{\boldsymbol{\beta}})\boldsymbol{A}^{\mathrm{T}}]^{-1}\boldsymbol{A}\widehat{\boldsymbol{\beta}}$, which yields the desired result.

Applying (4.3.5) to (4.3.1) we find that the decrease in the quadratic approximation Q in going from $\widehat{\boldsymbol{\beta}}$ to $\widehat{\boldsymbol{\beta}}_{00}$ is given by

$$2[Q(\widehat{\boldsymbol{\beta}}) - Q(\widehat{\boldsymbol{\beta}}_{00})] = (\boldsymbol{A}\widehat{\boldsymbol{\beta}})^{\mathrm{T}}[\boldsymbol{A}\boldsymbol{I}^{-1}(\widehat{\boldsymbol{\beta}})\boldsymbol{A}^{\mathrm{T}}]^{-1}\boldsymbol{A}\widehat{\boldsymbol{\beta}}. \tag{4.3.6}$$

The right side of (4.3.6) is known as the Wald statistic. It is used for testing the hypothesis that $\boldsymbol{\beta} \in \mathbb{G}_0$ under the assumption that $\boldsymbol{\beta} \in \mathbb{G}$.

From the standpoint of variable deletion, we see that an application of Wald's test for each of the candidate basis functions involves a fixed number of computations involving $\boldsymbol{A}$, but no further Newton–Raphson iterations. Once it is decided which basis function to delete, (4.3.5) can be used to obtain the starting value for maximum likelihood estimation corresponding to $\mathbb{G}_0$.

4.3.2 Application to GLMs

Recall the conditional probability for a response Y under a GLM based on a linear model $\mathbb{G}$

$$p(y, g(\boldsymbol{x}), \phi) \;=\; \exp\Big(\frac{y\, g(\boldsymbol{x}) - b(g(\boldsymbol{x}))}{a(\phi)} + c(y,\phi)\Big), \qquad g \in \mathbb{G}\,. \tag{4.3.7}$$

Given observations $(\boldsymbol{X}_1, Y_1), \ldots, (\boldsymbol{X}_n, Y_n)$, the log-likelihood is given by

$$\ell(\boldsymbol{\beta}) = \frac{1}{a(\phi)} \sum_{i=1}^{n} \Big[Y_i \sum_j \beta_j B_j(\boldsymbol{x}_i) - b\Big(\sum_j \beta_j B_j(\boldsymbol{x}_i)\Big)\Big] + \cdots\,,$$

where we have suppressed terms that do not involve $\boldsymbol{\beta}$. Clearly, maximizing over $\boldsymbol{\beta}$ does not involve ϕ, and we can temporarily ignore the dispersion effect. (This happens in the ordinary linear model as well; we can estimate the coefficients independently of the noise variance. We then have

$$\begin{aligned}[\nabla\ell(\boldsymbol{\beta})]_j &= \sum_{i=1}^{n} \Big[Y_i B_j(\boldsymbol{x}_i) - B_j(\boldsymbol{x}_i) b'\Big(\sum_j \beta_j B_j(\boldsymbol{x}_i)\Big)\Big] \\ &= \sum_i Y_i B_j(\boldsymbol{x}_i) - \sum_i \mu(\boldsymbol{x}_i;\boldsymbol{\beta}) B_j(\boldsymbol{x}_i), \end{aligned} \tag{4.3.8}$$

where we have set $\mu(\boldsymbol{x};\boldsymbol{\beta}) = b'(\sum_j \beta_j B_j(\boldsymbol{x}))$ following (4.2.9). Let $\mathbf{B}$ be the $n \times J$ design matrix having entries $[\mathbf{B}]_{ij} = B_j(\boldsymbol{x}_i)$, $\boldsymbol{Y}$ be the column vector of observations $Y_1, \ldots, Y_n$, and $\boldsymbol{\mu}(\boldsymbol{\beta})$ the column vector of means $\mu(\boldsymbol{x}_1;\boldsymbol{\beta}), \ldots, \mu(\boldsymbol{x}_n;\boldsymbol{\beta})$. Then we can rewrite (4.3.8) as

$$\nabla\ell(\boldsymbol{\beta}) = \mathbf{B}^{\mathrm{T}}(\boldsymbol{Y} - \boldsymbol{\mu}(\boldsymbol{\beta}))\,, \tag{4.3.9}$$

FIGURE 4.8. Comparing the Rao statistic with the actual rise in log-likelihood for stepwise addition of knots in time for the Poisson regression example.

which starts to resemble some of the components from the normal equations for simple linear models.

The information matrix $\boldsymbol{I}(\boldsymbol{\beta}) = -\mathbf{H}(\boldsymbol{\beta})$ is given by

$$\begin{aligned}[\boldsymbol{I}(\boldsymbol{\beta})]_{jk} &= \sum_{i=1}^{n} B_j(\boldsymbol{x}_i)B_k(\boldsymbol{x}_i)b''\Big(\sum_j \beta_j B_j(\boldsymbol{x}_i)\Big)\\ &= \frac{1}{a(\phi)}\sum_i B_j(\boldsymbol{x}_i)B_k(\boldsymbol{x}_i)\sigma^2(\boldsymbol{x}_i;\boldsymbol{\beta}),\end{aligned}$$

where $\sigma^2(\boldsymbol{x};\boldsymbol{\beta}) = a(\phi)b''(\sum_j \beta_j B_j(\boldsymbol{x}))$ according to (4.2.10). Let $\boldsymbol{W}$ denote the $n \times n$ diagonal matrix having diagonal entries $[\boldsymbol{W}(\boldsymbol{\beta})]_{ii} = \sigma^2(\boldsymbol{x}_i;\boldsymbol{\beta})$. Then

$$\boldsymbol{I}(\boldsymbol{\beta}) = \frac{1}{a(\phi)}\mathbf{B}^{\mathrm{T}}\boldsymbol{W}(\boldsymbol{\beta})\mathbf{B}. \tag{4.3.10}$$

Combining (4.3.9) and (4.3.10), we find that the Newton–Raphson iteration is given by

$$\boldsymbol{\beta}_{k+1} = \boldsymbol{\beta}_k + a(\phi)(\mathbf{B}^{\mathrm{T}}\boldsymbol{W}_k\mathbf{B})^{-1}\mathbf{B}^{\mathrm{T}}(\boldsymbol{Y} - \boldsymbol{\mu}_k)$$

where $\boldsymbol{W}_k = \boldsymbol{W}(\boldsymbol{\beta}_k)$ and $\boldsymbol{\mu}_k = \boldsymbol{\mu}(\boldsymbol{\beta}_k)$. The update that carries us from $\boldsymbol{\beta}_k$ to $\boldsymbol{\beta}_{k+1}$ is basically a weighted least squares fit operating on the observations $\boldsymbol{Y}$ adjusted by the mean at the previous step $\boldsymbol{\mu}_k$, where the weights depend on the conditional variances σ_k^2, also from the previous step. Correspondingly, the overall fitting procedure is referred to as *iteratively reweighted least squares.*

We now turn to the Rao and Wald statistics for GLMs. Following (4.3.4) and (4.3.6), we have that

$$2[Q(\boldsymbol{\beta}_1) - Q(\boldsymbol{\beta}_0)] = a(\phi)[\boldsymbol{Y} - \boldsymbol{\mu}(\boldsymbol{\beta}_0)]^{\mathrm{T}}\mathbf{B}[\mathbf{B}^{\mathrm{T}}\boldsymbol{W}(\boldsymbol{\beta}_0)\mathbf{B}]^{-1}\mathbf{B}^{\mathrm{T}}[\boldsymbol{Y} - \boldsymbol{\mu}(\boldsymbol{\beta}_0)]$$

for the Rao statistic and

$$2[Q(\widehat{\boldsymbol{\beta}}) - Q(\widehat{\boldsymbol{\beta}}_{00})] = \frac{1}{a(\phi)}(\boldsymbol{A}\widehat{\boldsymbol{\beta}})^T\left(\boldsymbol{A}[\mathbf{B}^{\mathrm{T}}\boldsymbol{W}(\widehat{\boldsymbol{\beta}})\mathbf{B}]^{-1}\boldsymbol{A}^{\mathrm{T}}\right)^{-1}\boldsymbol{A}\widehat{\boldsymbol{\beta}}$$

for the Wald statistic.

To judge the closeness of these approximations, we consider again stepwise selection of basis elements for the time effect in the particulate matter study. In Figure 4.3, we presented the actual rise in log-likelihood for adding a new cubic spline term. In Figure 4.8, we use the relationship in Chapter 3 explicitly to enter a single basis function and employ the Rao approximation.

4.4 A general methodology