Beyond Linearity by Default: Generalized Additive Models

Author(s): Nathaniel Beck and Simon Jackman

Source: *American Journal of Political Science*, Apr., 1998, Vol. 42, No. 2 (Apr., 1998), pp. 596-627

Published by: Midwest Political Science Association

Stable URL: https://www.jstor.org/stable/2991772

# *Beyond Linearity by Default: Generalized Additive Models**

Nathaniel Beck *University of California, San Diego***
Simon Jackman, *Stanford University****

Social scientists almost always use statistical models positing the dependent variable as a global, linear function of *X,* despite suspicions that the social and political world is not so simple, or that our theories are so strong. Generalized additive models (GAMs) let researchers fit each independent variable with arbitrary nonparametric functions, but subject to the constraint that the nonparametric effects combine additively. In this way GAMs strike a sensible balance between the flexibility of nonparametric techniques and the ease of interpretation and familiarity of linear regression. GAMs thus offer social scientists a practical methodology for improving on the extant practice of global linearity by default.

We reanalyze published work from several subfields of political science, highlighting the strengths (and limitations) of GAMs. We estimate non-linear marginal effects in a regression analysis of incumbent reelection, nonparametric duration dependence in an analysis of cabinet duration, and within-dyad interaction effects in a reconsideration of the democratic peace hypothesis. We conclude with a more general consideration of the circumstances in which GAMs are likely to be of use to political scientists, as well as some apparent limitations of the technique.

## 1. Introduction

Social scientists make many assumptions in the course of employing statistical models. Linearity must rank as one of the more ubiquitous assumptions, yet one that typically gets fairly short shrift. For instance, while econometrics texts invariably make the linear functional form their first

*We thank Michael Dimock, Gary Jacobson, Gary King and John Oneal for their data, Ari Zentner for research assistance, and Larry Bartels, Richard Carson, William Cleveland, Peter Hall, Trevor Hastie, Clive Loader, Gary King, Robert Kohn, Jonathan Nagler, Adrian Pagan and Matt Wand for helpful discussions. The analysis of the democratic peace draws on material co-authored with Richard Tucker. We are grateful to Jonathan Nagler, the director of the Social Science Computing Facility at the University of California, Riverside, and Kolin Hand, its systems programmer, for allowing us to use their computing facility. Jackman's contribution was facilitated by a Research Fellowship at the Research School of the Social Sciences, Australian National University. Errors and omissions remain our own responsibility. All data analyses and statistical graphs were generated using S-PLUS, Version 3.3 (Statsci 1995). The routines and data are available at ftp://weber.ucsd.edu/pub/nbeck/gam.

**Department of Political Science, University of California, San Diego, La Jolla, CA 92093. e-mail: beck@ucsd.edu

***Department of Political Science, Stanford University, Stanford, CA 94305-2044.

*American Journal of Political Science,* Vol. 42, No. 2, April 1998, Pp. 596–627 

assumption (e.g., Greene 1997, 143), the consequences of violating latter assumptions garner far more space. This strikes us as odd, given that the optimality properties of least squares and maximum likelihood estimators depend upon the researcher "getting the mean right," that is, having specified the correct functional form for the mean of the dependent variable, conditional on the independent variables.

To be sure, introductory texts go on to show how a variety of specific nonlinear functional forms can be estimated in the linear regression framework. But simple nonlinear forms, such as polynomials or logarithmic transformations, often appear as clumsy attempts to capture subtlety in the regression function specific to a region of $X$; as we argue below, these tricks of the trade impose *global* solutions on *local* features of the data.

Why this disparity in emphasis? For one thing, social and political theories rarely suggest a specific functional form. It is true that we occasionally encounter good reasons for using higher order terms or nonlinear transformations, but these situations are relatively rare. And rarer still is guidance about the specific form of any *a priori* nonlinearities. For instance, we examined the 44 articles and research notes appearing in Volume 89 (1995) of the *American Political Science Review* in search of substantive arguments in support of specific functional forms. In 25 articles or research notes using some form of quantitative data analysis, 19 use a multivariate technique (typically regression, or some variant such as logistic regression). Of these 19 articles, only four employ a nonlinear transformation of one or more right-hand side variables. In three of these cases the logarithmic transformation is applied to variables exhibiting positive skew, though only twice is the use of the transformation accompanied by an explicit substantive argument, i.e., diminishing marginal effects in Hall and Houweling (1995, 127) and Bueno De Mesquita and Siverson (1995, 844–6). And Lodge, Stenbergen, and Brau (1995, 323) employ a power transformation of a time counter to estimate the shape of a memory-decay function. But in the vast majority of cases researchers using regression-like techniques rely exclusively on linear functional forms, and almost always without any substantive justification.

The key point here is not a dearth of theories implying nonlinearity or non-monotonicity. Rather, *few social scientific theories offer any guidance as to functional form whatsoever.* Statements like "$y$ increases with $X$" (monotonicity) are as specific as most social or political theories get. As a result, the possibility that the functional form of the regression relationship might vary *locally* over the range of $X$ is implicitly ruled out in many political science applications. Rather than testing whether the relationship between $y$ and $X$ is specific to local regions of $X$, standard practice is to impose a global linear relationship between $y$ and $X$ (i.e., assuming that the relationship between $y$ and $X$ is exactly the same for all possible values of $X$). To the

extent that empirically-oriented political scientists confront this issue at all, the generality of global linear regression is trumpeted over the specificity, complexity, or data-driven properties of alternatives such as piecewise linear regression, transformations of variables, or the nonparametric local regression models we employ below. But we do not find this defense of global linear regression particularly compelling. The use of linear regression in political science is typically by default, and rarely follows from theories of social and political processes implying that a linear functional form between $y$ and $X$ holds globally.

If the sole goal of data analysis was to "get the mean right," then a simple prescription follows: run regressions with dummy variables for every unique value of $X$. But such a strategy represents an absurdly strong commitment to modeling locally and an abandonment of science altogether. Most of us believe that parsimony makes for good social science, and the clear implication for data analysis is to purchase as much explanatory power with as few assumptions and/or explanatory variables as necessary. Of course, one person's parsimony could well be another's gross reduction, over-simplification, or even "ahistoricism" (Isaac and Griffin 1989). Our view is that while the social and political world admits generalization—we do not need dummy variables for every unique realization of $X$—it is probably more subtle than the linear, additive structure characterized by most regression models. On the other hand, our theories typically are not so developed as to give us strong reasons to prefer one functional form over another, or to be confident that (linear) relationships among variables hold globally. Clearly, there is a tradeoff here and hence our advocacy of a methodology that allows the data to speak on the question of functional form, yet still retain much of what we like about linear regression models. In particular, we shall see that the methods we employ encapsulate linear regression as a special case.

To summarize, we are not arguing that nonlinearities necessarily abound; we are saying that we do not know one way or the other. Our theories are typically silent on the issue, and empirical social scientists are neither trained nor encouraged to consider nonlinearity or local fluctuations in the regression function as compelling issues. Nonetheless, we do not see the work-a-day regression-runner as a dupe. Most of us understand that the linear regression model prevails in spite of suspicions that the world is not linear and additive. As we tell our students, linear regression prevails because it is simple—parsimonious, easily estimated, and easily interpreted—and, with perhaps a hint of embarrassment, it seems plausible in the absence of any theoretical story suggesting a specific functional form. But the truth of the matter is that there are seldom *any* prior expectations as to the functional form of $E(y) = m(X)$. Moreover, it hardly seems persuasive to use the absence of theoretical arguments about nonlinearity to buttress use of the linear

regression model. Global linearity by default seems an unduly narrow methodological practice in the face of weak or vague theoretical expectations about the political or social relationships we study, and especially since alternative models can be readily implemented.

## 2. Generalized Additive Models

Here we survey a regression-like model that directly confronts the possibility of nonlinearity: generalized additive models (GAMs). These models and the ideas that underlie them have received considerable attention in the statistics literature, but have yet to percolate into the social sciences.

Additive models recast the linear regression model

$$y_i = \alpha + \sum_{j=1}^{k} \beta_j X_{i,j} + \varepsilon_i. \qquad [1]$$

by modeling $y$ as an additive combination of arbitrary univariate functions of the independent variables, and (exactly as in the linear regression model) a zero mean, independent and identically distributed stochastic disturbance:

$$y_i = \alpha + \sum_{j=1}^{k} m_j\left(X_{i,j}\right) + \varepsilon_i. \qquad [2]$$

where $E(\varepsilon_i) = 0$, and $\text{var}(\varepsilon_i) = \sigma^2, i = 1, \ldots, n$. Here we focus on *loess,* a locally weighted regression smoother, and a widely-used method of estimating $m(\cdot)$. We describe loess more fully in section 3.1. Several other smoothers have been considered in the literature, but our focus on this one smoother does not unduly restrict the generality of our discussion.

*Generalized* additive models extend the framework in Equation [2] in precisely the same way that generalized linear models (GLMs) (McCullagh and Nelder 1989) extend the linear regression model so as to accommodate binary and count dependent variables. For instance, when using a GAM for a binary (0,1) dependent variable we will use the GAM version of a logit or probit model. Unlike the standard logit/probit model, the GAM logit/probit model allows the independent variables to have nonlinear effects on the continuous latent variable.[1]

### *2.1 Interpreting GAMs*

The absence of the regression parameters $\beta_j$ in Equation [2] reflects an important characteristic of GAMs. One does not obtain a set of regression

[1]In Beck and Jackman (1997) we provide a guide to software for GAMs, with an emphasis on the implementation in S-PLUS, including the GLM extensions for qualitative dependent variables.

parameters from a GAM, but rather, estimates of $m_j(X_{i,j})$ for every value of $X_{i,j}$, written as $\hat{m}_j(X_{i,j})$. The coefficient on each $m_j(X_j)$ is set to one by construction, and so it is the $\hat{m}_j(X_j)$ themselves that tells us about the relationship between $X_j$ and the dependent variable.[2] In many applications we will assume that some of the predictors have linear effects, as in the standard regression setup. Thus we will typically estimate a *semi-parametric* additive regression model

$$y_i = \alpha + \sum_{l=1}^{m} \beta_l Z_{i,l} + \sum_{j=1}^{k} m_j\left(X_{i,j}\right) + \varepsilon_i. \qquad [3]$$

Graphical methods are used to interpret the nonparametric component of a GAM. A plot of $X_j$ versus $\hat{m}_j(X_j)$ reveals the nature of any estimated nonlinearities in the relationship between $X_j$ and the dependent variable, holding constant the other components in the model. Standard errors and confidence regions can be calculated and plotted about $\hat{m}_j(X_j)$, providing a guide as to whether the fitted function is distinguishable from a linear fit, or increasing or decreasing in $X_j$. While it may seem easier to examine tables of regression coefficients rather than plots of the $\hat{m}_j$, this ease is only obtained at the cost of (possibly) unwarranted, restrictive, and unnecessary *assumptions* of linearity. On the other hand, additivity ensures that the effects of each of a GAM's predictors can be interpreted net of the effects of other predictors, just as in linear regression.

### *2.2 Other Approaches to Nonparametric Multiple Regression*

Nonparametric statistics is an extremely large and active area of statistical research (e.g., Pagan and Ullah forthcoming), providing many alternatives to the linear, additive regression model. Given this panoply of methods, our focus on GAMs warrants some justification.

GAMs seem to strike a sensible compromise between ease of interpretation and flexibility. Other alternatives to linear regression like neural nets (Ripley 1993; White 1992), projection pursuit (Friedman and Stuetzle 1981), or alternating conditional expectations (Breiman and Friedman 1985; DeVeaux 1990) strike us as sacrificing ease of interpretation in order to better fit the data. For instance, projection pursuit finds orthogonal projections through the space of the predictors that maximize model fit, but these com-

[2]The fitted $m_j(X_j)$ each have zero mean by construction. Without this constraint the intercept in a GAM is unidentified; i.e., any of the fitted $m_j(X_j)$ could be shifted by some arbitrary constant, accompanied by an offsetting shift in the intercept, and the resulting set of estimates would fit the data equally well.

binations of the predictors often lack substantive interpretation, resembling combinations of apples and oranges. Also, it is often difficult to decide when these techniques are overfitting the data, a problem that is especially pressing for projection pursuit and neural networks.

GAMs should also be distinguished from semi-parametric methods which keep the assumption of linearity but allow the error term to be modeled nonparametrically (e.g., Western 1995). While this approach is valuable, we believe that "getting the mean function right" is more important than correct modeling of the error process.[3] On the subject of robustness against outliers, we also note here GAMs have a built-in resistance against outlying data points on $X$, since they estimate the mean of $y$ specific to a series of local regions of $X$, rather than exploiting all the data simultaneously. However, the benefits of local fitting come with the cost of higher variance, a point we elaborate in section 3.2.

We also distinguish GAMs from parametric approaches which attempt to handle nonlinearities. These range from the familiar log and polynomial to the more exotic Box-Cox transformation.[4] While these approaches obviously add flexibility to the linear model, they are still global: one relationship is assumed to hold everywhere. Conversely, the local, nonparametric properties of the GAM make it much more flexible than any parametric transformation. Complicated parametric transformations often work poorly in practice, inducing multicollinearity or floundering on problems caused by odd behavior at the extremes of the data; the more natural approach of the GAM is not subject to such problems. But, as we shall see in our examples, even those committed to parametric transformations will find the GAM useful in selecting and/or justifying those transformations.

As Equation [2] shows, GAMs retain the additive character of the familiar regression model, and so can be interpreted variable-by-variable, unlike projection pursuit or most neural nets. Transformations of variables are restricted to GAMs' right-hand side variables, again making substantive interpretation relatively easy. The results of a GAM, the $\hat{m}_j(X_j)$, can be easily interpreted in the natural units of the research problem and plots of the $\hat{m}_j(X_j)$ against the untransformed independent variables are the nonparametric counterpart of standard regression lines.

[3]But it is possible to combine these two approaches. In particular, GAMs can be estimated using Huber's (1981) "M-estimates" which allow for platykurtic error distribution (Hastie and Tibshirani 1990), or using mixtures of differently scaled Normal distributions (e.g., Smith and Kohn 1997). Our focus here is on the mean function, and so we do not pursue M-estimators for GAMs further.

[4]We exclude splines here, since the most flexible of them, the smoothing spline, is simply another smoother that is incorporated in the GAM approach (see Beck and Jackman 1997).

Furthermore, the amount of smoothing of each predictor in a GAM is governed by calibration parameters that have straightforward substantive interpretations; for instance, the *equivalent degrees of freedom,* which we describe below. And because GAMs are additive, the methodology for comparing GAMs and assessing goodness-of-fit is for all practical purposes identical to that used to compare different linear regression models.

With linear regression at one extreme and neural networks at the other, GAMs lie closer to the former than the latter. We concede that there are research settings where neural nets or ACE or projection pursuit are to be preferred; but our belief at this point is that such situations are rare in political science. In short, GAMs provide much of the flexibility of nonparametric methods while allowing the analyst to keep a close eye on what the statistical procedures are actually doing to the data.

## 3. Scatterplot Smoothing

Like most nonparametric techniques, the statistical theory and accompanying mathematics for GAMs is reasonably complex, and we refer readers to the discussion in Hastie and Tibshirani (1990, chap. 5). However, most of the key intuitions about GAMs follow from ideas having to do with *scatterplot smoothing.* Indeed, GAMs are just additive combinations of bivariate scatterplot smoothers.

Smoothing is an important tool for nonparametric regression, addressing one of the simplest yet most fundamental questions in data analysis: "what is our best guess of $y$, given $X$?" In the most general setting, a *smoother* yields a series of data reductions or summaries of $y$, each specific to (possibly overlapping) regions of the space of the predictors. Bivariate scatterplot smoothers perform this data summary with respect to a single predictor variable, $x$. Computing localized summaries of $y$ at numerous points over the range of $x$ yields an approximation to the mean function for $y$ conditional on $x$, $E(y|x) = m(x)$.

The moving average or moving median is a frequently encountered smoother, summarizing a time series with means or medians specific to overlapping time windows. The resulting summary series exhibits less variability than the raw series, again highlighting that smoothers are tools for data reduction. The amount of smoothing performed by a moving-average is determined by the number of time periods that are averaged, or, more generally, by the size of the *neighborhood* or *bandwidth* over which the averaging is done. The larger this neighborhood, the more smoothing, and the more data reduction. For instance, consider taking the entire sample as the neighborhood. A moving average in this case is simply the sample mean of the observations, reducing the entire series to a constant.

Bivariate linear regression is also a special case of a scatterplot smoother, summarizing a scatterplot with just two parameters (a slope and

an intercept), providing an infinite amount of smoothing.[5] Another way to think about the infinite smoothness of linear regression is to consider that the relationship between $y$ and $x$ is taken to be global, not locally-specific, and hence the smoothing neighborhood is the entire dataset; in other words, the relationship between $x$ and $y$ is deemed fixed over all plausible values of $x$. As we show below, scatterplot smoothers yield fitted functions that tend towards a regression fit as the amount of smoothing increases. At the other extreme, as bandwidth gets smaller, the amount of smoothing decreases, and the fitted function degrades to a function that simply connects the data points in the scatterplot, providing no data reduction at all.

*3.1 Smoothing by Local Regression (Loess)*

In our uses of GAMs we rely on a local regression scatterplot smoother known as *loess* (Cleveland and Devlin 1988), which we describe briefly.[6] Local regression fits a line to a scatterplot by estimating the relationship between $x$ and $y$ at a number of *target points* over the range of the observed $x$ values. Given a target point $x_0$, loess yields $\hat{y}|x_0 = \hat{m}(x_0)$ by fitting a locally weighted regression to the data, according to the following procedure (Cleveland 1993, 100–1):

1. Identify the $q$ *nearest neighbors* of $x_0$, i.e., the $q$ values of $x$ closest to $x_0$. This set is denoted $N(x_0)$. The analyst controls $q$ via a "span" argument, which defines the size of the neighborhood in terms of a proportion of the sample size: i.e., $q$ is span times the sample size, truncated to an integer. Like many smoothers, choosing the span parameter (which in turn controls the bandwidth of the smoother) is the most critical part of smoothing by loess.
2. Calculate $\Delta(x_0) = \max_{N(x_0)}|x_0 - x_i|$, the distance of the near-neighbor most distant from $x_0$.
3. Calculate weights $w_i$ for each point in $N(x_0)$, using the following *tricube weight function:*

$$W\left(\frac{|x_0 - x_i|}{\Delta(x_0)}\right) \qquad [4]$$

[5]Smoothness here is defined as the inverse of the second derivative of the fitted function. Since linear regression fits straight lines, and straight lines have second derivatives that are zero, linear regression fits are considered to be infinitely smooth; similarly for the horizontal line that results from the implicit intercept-only regression when a variable is reduced to its mean.

[6]In Beck and Jackman (1997) we use another widely used scatterplot smoother (cubic smoothing splines), obtaining virtually identical results to those reported below. Constraints of space underlie our focus on smoothing by local regression. Suffice to say that decisions about how much to smooth using a particular type of smoother usually overwhelm choices about what type of smoother to use. Put differently, within-smoother variation dominates across-smoother variation, and hence our focus on smoothing on local regression via loess is not particularly restrictive.

where

$$W(z) = \begin{cases} (1 - z^3)^3 \text{ for } 0 \leq z < 1; \\ 0 \qquad \text{otherwise} \end{cases} \quad [5]$$

Note that $W(x_0) = 1$, and that the weights decline smoothly to zero over the chosen set of nearest-neighbors, such that $W(x_i) = 0$ for all non near-neighbors. The use of the tri-cube weight function here is somewhat arbitrary. Any weight function that has smooth contact with 0 on the boundary of $N(x_0)$ will produce a smooth fitted function (Cleveland, Devlin, and Grosse 1988, 112).

4. Regress $y$ on $x$ and a constant (for local linear fitting), using weighted least squares (WLS) with weights $w$ as defined in the previous step. Quadratic or cubic polynomial local regressions can also be used, or even mixtures of low-order polynomials (Cleveland and Loader 1996), depending on the analyst's beliefs about the order of the underlying function.
5. The smoothed value $\hat{y}|x_0 = \hat{m}(x_0)$ is the predicted value from the WLS fit, evaluated at $x_0$.

Repeating this procedure for every target point traces out a function, the smoothed fit of $y$ against $x$.

Local regression can also be applied beyond the two-dimensional setting encountered in scatterplot smoothing. Consider fitting the following model by local regression

$$y_i = m(x_i, z_i) + \varepsilon_i \quad [6]$$

A target point in the two-dimensional space of the predictors can be denoted as $(x_0, z_0)$, about which we can define a set of $q$ nearest-neighbors. The Euclidean distances of near-neighbors from the target point are passed to the tri-cube weight function in Equation [5], producing a set of weights for the (local linear) regression of $y$ on $x$, $z$, and a constant, or higher-order terms (squares and cross-products). This is a very useful feature of local regression, which we exploit in one of the applications below. In particular, the term $m(x,z)$ in a GAM provides a way of capturing interaction effects nonparametrically, rather than by the familiar (though less flexible) parametric form

$$y_i = \beta_i x_i + \beta_2 z_i + \beta_3 x_i z_i + \varepsilon_i \quad [7]$$

Local regression by loess is an excellent paradigm for smoothing in social-science settings. The use of nearest-neighborhoods lets the smoother adapt

to local fluctuations of the density of $x$.[7] This combination of three features—nearest neighbors, a smooth weight function, and forming $\hat{y}|x_0$ via locally weighted regressions—helps local regression (and loess in particular) outperform many other scatterplot smoothers (e.g, fixed bandwidth kernel smoothers, or more primitive smoothers such as moving averages, overlapping regressions, and so on) in a wide number of settings. Useful summaries of these properties and other strengths of local regression by loess appear in Hastie and Loader (1993), Cleveland and Loader (1996), and Loader (1996).

*Inference for scatterplot smoothers.* All scatterplot smoothers proceed by averaging values of $y$ about some target point $x_0$, and we have seen that for local regression this averaging takes place using locally weighted regressions. Local regression is a special case of a linear smoother, all of which estimate $m(x_0)$ as

$$\hat{m}(x_0) = \sum_{j=1}^{N} S(x_0, x_j; \lambda) y_i, \qquad [8]$$

where $x_j$ and $y_j$ are elements of the data vectors $x$ and $y$, respectively, $x_0$ is the target point, and $S$ is a weight function parameterized by a smoothing parameter $\lambda$. Letting the scatterplot smoother produce $\hat{m}(x_i)$ for each $i = 1, \ldots, n$ yields **S**, an $n$ by $n$ *smoother matrix.* Each row of **S** contains the weights used in generating the smoothed fit at each data point; or, in other words, the $j$th row vector of **S** is a $n$ by 1 vector of the weights each observation picks up in contributing to the fitted smooth at $x_j$. Using matrix notation a linear smoother can be written as

$$\hat{\mathbf{m}}(\mathbf{x}) = \mathbf{S}\mathbf{y}, \qquad [9]$$

where $\hat{\mathbf{m}}(\mathbf{x})$ is the $n$ by 1 vector of estimates of the mean function $E(\mathbf{y}|\mathbf{x}) = \mathbf{m}(\mathbf{x})$, conditional on the smoothing parameter $\lambda$ (Hastie and Tibshirani 1990, 44).

Linear smoothers are thus linear in precisely the same way that the predicted values from a least squares regression are a linear function of the dependent variable. Recall that for least squares linear regression, $\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{H}\mathbf{y}$, where $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ is the well-known "hat matrix" which transforms the observed responses **y** into the fitted responses $\hat{\mathbf{y}}$ (Weisberg 1985, 109–11). The smoother matrix **S** has many of the same

[7]When the data are sparse in the vicinity of $x_0$, the set of $q$ nearest neighbors is wider than that for target points in more dense regions of $x$. This ensures that the local regressions are based on the same number of observations for all target points, which in turn helps keep the overall variance of a local regression fit smaller than that of fixed-bandwidth scatterplot smoothers.

properties as the hat matrix. In fact, **H** is a special case of a smoother matrix, recalling that least squares regression can be thought of as an infinitely smooth scatterplot smoother. The representation of linear smoothers in Equation [9] is extremely convenient. As the bandwidth of a smoother gets wider, off-diagonal elements of **S** get larger, as data points further away from a target point $x_0$ have greater influence in forming $\hat{m}(x_0)$. On the other hand, in the limiting case of a scatterplot smoother that simply interpolates the data, **S** is a diagonal matrix (specifically, an identity matrix), equivalent to running a regression with a dummy variable for every observation.

*Equivalent degrees of freedom.* The inverse relationship between the amount of smoothing and the extent to which **S** is diagonal gives rise to a very useful approximation for linear smoothers. The degrees of freedom consumed by a linear smoother **S** is well approximated by the trace (sum of the diagonal elements) of **S** (Hastie and Tibshirani 1990, 52–5 and Appendix 2).[8] This quantity, $tr(\mathbf{S})$, is referred to as the *equivalent degrees of freedom;* more precisely, $tr(\mathbf{S}) - 1$ is an estimate of the degrees of freedom consumed by the non-parametric component of the smooth fit, since all scatterplot smoothers fit an intercept term.

*Standard errors.* Given an estimate of the error variance $\sigma_2$, and using Equation [9], the variance-covariance matrix of the fitted smoothed values can be written

$$\text{cov}\left[\hat{\mathbf{m}}(\mathbf{x})\right] = \mathbf{SS}'\sigma^2 \qquad [10]$$

with the pointwise standard errors of $\hat{\mathbf{m}}$ given by the square-root of the diagonal elements of $\mathbf{SS}'\sigma^2$. The error variance itself is usually estimated as

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}\left[y_i - \hat{m}(x_i)\right]^2}{n - tr(\mathbf{S})}. \qquad [11]$$

With either a large $n$ or normal errors, pointwise confidence intervals can be constructed about the fitted function using the pointwise standard errors. Examples are in the applications below.

*Specification tests.* Specification tests for scatterplot smoothers/GAMs can be performed under the same conditions for linear regression or GLMs (i.e., large samples or normal errors). In particular, for a GAM with a continuous dependent variable, the sum of squared residuals can be used in a standard $F$ test. The statistic for this test is computed identically to the usual linear regression $F$ statistic. While the distribution of this statistic is only approximately $F$, simulation results indicate that the approximation is not

[8]Recall that for least squares linear regression, the trace of the hat matrix is equal to the degrees of freedom consumed by the regressor's predictors, (including an intercept, if present).

unreasonable (Hastie and Tibshirani 1990, 67).[9] For models with a binary or count dependent variable we can compute the analogue of a log-likelihood (which can be tested with the usual $\chi^2$ test).

For both tests, statistics are computed using the equivalent degrees of freedom consumed by the smoother. A common use of either test is to examine whether a specification with a smooth term is superior to one with a corresponding linear term; since the latter model is nested within the former, the appropriate $F$ or $\chi^2$ test can be used here.

*Estimation.* Finally, we note that estimating GAMs usually poses no great challenge, given modern computing power and purpose-written software. An iterative algorithm known as *backfitting* is usually employed to estimate GAMs (and other nonparametric regression models). We refer interested readers to some of the descriptions of the algorithm in the statistics literature (Breiman and Friedman 1985; Hastie and Tibshirani 1990).

### *3.2 Bandwidth Selection*

Because linear regression is a global fitting technique, all observations contribute to the least squares estimate of a specific $y|x_0$, no matter how distant from the target point. The result of exploiting all the sample information is familiar: least squares has minimum variance among the class of linear, unbiased estimators. But it is precisely this question of bias that is at issue. If least squares linear regression is not "getting the mean right" then we might question the value of its optimality properties. Under an easily-imagined set of circumstances, we might prefer an estimated mean function which tracks local features of the data, to a low variance, global, linear fit, that misses important features of the mean function.

On the other hand, since local fitting does not use all the available sample data, it results in estimated mean functions that are inherently less precise than that produced by a regression fit to the entire sample. In general, a wider bandwidth risks bias while reducing variability, while a narrower bandwidth purchases less bias at the cost of precision.

All scatterplot smoothers wrestle with this tradeoff, via the smoothing parameter that in turn governs bandwidth. Bandwidths are sometimes estimated or derived from the data using cross-validation (e.g., Hastie and Tibshirani 1990, 42–52) or "plug-in" methods (e.g., Ruppert, Sheather, and Wand 1995), but may also be chosen in advance and then fine-tuned by the researcher on a case-by-case basis. For local regression by loess, bandwidth is controlled by the span parameter, governing the size of the set of nearest

[9]These tests are only approximate for a number of fairly technical reasons, that are beyond our scope here (see Cleveland and Devlin 1988; Cleveland, Devlin, and Grosse 1988). In Beck and Jackman (1997) we report new simulation results which indicate that the $F$ approximation is generally good, as well as factors which appear to make the approximation better or worse.

neighbors that have non-zero weights in contributing to the estimated fitted function at a particular target point.

Figure 1 depicts the bias-variance tradeoff, using a scatterplot of social expenditures and per capita GNP on 64 countries in 1966, given in Wilensky (1975, 122–4). The dependent variable is social security spending as a percent of GNP, while the predictor is 1966 per capita GNP, converted into U.S. dollars. We stress that since our interest in these data is purely expository (highlighting the bias-variance tradeoff in scatterplot smoothing), we ignore a number of substantive issues raised by this example (see Castles and McKinlay 1979, 184), as well as some of the more standard tricks-of-the-trade for better fitting these data with global parametric regression (e.g., logging the income variable, or fitting a quadratic model).

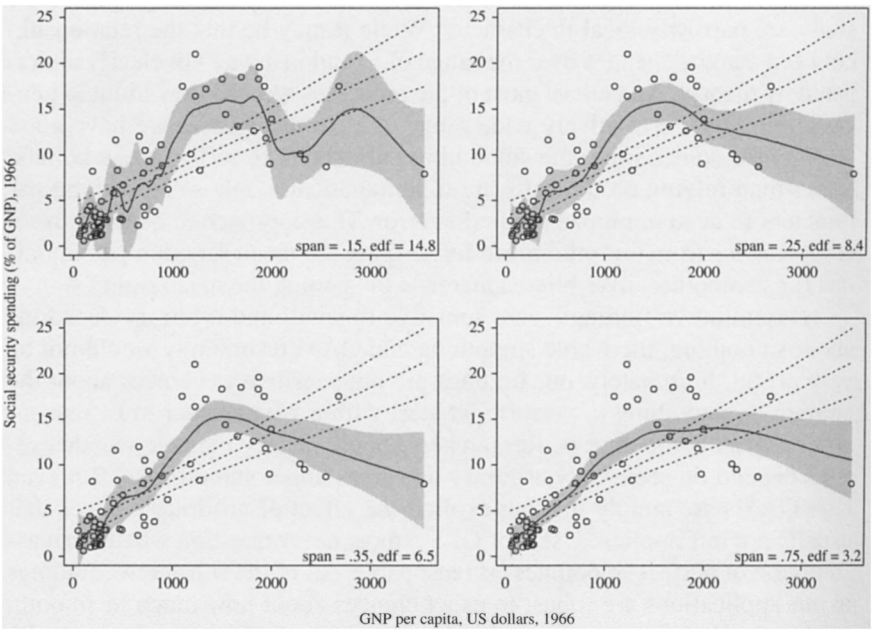
The top left panel of Figure 1 uses a relatively small span, making the fitted function quite sensitive to local changes in the mean function, but with relatively low precision (wider confidence intervals). As the bandwidth of the smoother increases, the resulting fit is less sensitive to local changes in the mean function (risking bias), though more precise (smaller confidence intervals). In the bottom right-hand panel, a local regression fit obtained with a span = .75 is reasonably smooth, and consumes the equivalent of just 1.2 degrees freedom more than if we were to fit a global linear regression model.

Global, linear regression suggests that social expenditures are strongly and positively associated with the wealth of a country; the estimated coefficient here is .0051 ($t = 7.5$), suggesting that every extra \$100 of per capita GNP translates into about half a percentage point of increased social expenditures (as a proportion of GNP). But the local regression models show that such a conclusion is an oversimplification. At low levels of wealth, the positive association between wealth and social expenditures is quite pronounced, and probably greater than the effect estimated with the global linear model. But at around \$1500 GDP per capita, the relationship appears to reverse, or at least become indistinguishable from zero.

The other point to be made here is that the linear regression model is not immune from the bias-variance tradeoff. In all cases, the confidence intervals on the global linear regression fit (dotted lines) are more narrow than those obtained from the nonparametric local regression models. But this precision comes at the expense of possibly high bias, which is clearly the case in this example. Global linear regression simply fails to pick up critical features of the data.

Figure 1 also highlights that the choice of smoothing parameter is a critical feature of scatterplot smoothing. Based on our experience in applying smoothers in social science contexts, this choice is best dealt with on a case-by-case basis, informed by prior beliefs about the smoothness of the process being modeled and diagnostic plots of residuals for evidence of over- and

**Figure 1. Bias-Variance Tradeoff, Local Linear Regression (loess) and Global Least Squares Regression, Fit to Cross-National Data on Social Security Expenditures**




The solid lines are the fitted smooth functions, and the shaded regions indicate 95% pointwise confidence regions, assuming iid normal errors. For comparison, the dotted lines mark a global linear regression fit and its 95% confidence interval. The top panels show how smaller bandwidths increase sensitivity to local changes in the mean function, at the expense of precision. Note the wider confidence intervals in the top panels, compared to those around the smoother fitted functions in the bottom panels.

under-fitting (e.g., Cleveland, Grosse, and Shyu 1992).[10] We prefer this type of approach for scatterplot smoothing and fitting GAMs; graphical inspection lets us keep track of what the smoothing procedures are up to more than if we were to let the software "black-box" the model fitting, say via cross-validation or other automated procedures for choosing smoothing parameters. Left to their own devices, smoothing algorithms that minimize cross-validation suggest fits to the data that generally appear to over-fit the data, producing quite jagged fits.

[10]According to one of the pioneers of loess, the choice of smoothing parameters ". . . must be based on a combination of judgment and of trial and error. There is no substitute for the latter" (Cleveland 1993, 96).

Typically we have prior beliefs that the social and political processes under study are reasonably smooth; that is, while the social world may contain nonlinearities, we doubt it to be abruptly or flamboyantly nonlinear. Or, put differently, we do not believe that the social and political processes we study are narrowly local in character. While it may be that the relationship between $x$ and $y$ changes over the range of $x$ (and in a way not clearly anticipated by theory), we believe most of the processes we study do admit to generalization over a relatively wide range of $x$. In these cases we have good reasons for constraining the smoothing parameters to reflect these beliefs, rather than relying on the software to automatically choose smoothing parameters so as to minimize prediction error. This approach to bandwidth selection stems from our preference for keeping the mean function parsimonious (i.e., smoother) over blind adherence to "getting the mean right."

If qualitative findings were sensitive to small and arbitrary decisions about smoothing, the whole smoothing and GAM technology would not be very useful. Fortunately, our findings are not sensitive to choices about the amount of smoothing to perform, at least within what appear to be reasonable ranges of smoothness. Researchers should clearly indicate whether results depend on seemingly arbitrary decisions about smoothness. But even here GAMs are simply making explicit the effect of arbitrary choices that usually are left implicit. Users of OLS almost never question whether an assumption of perfect smoothness is reasonable. All of the substantive findings in our applications are robust to exact choices about how much to smooth; on the other hand, since our results differ markedly from those obtained with simple linear models, it follows that the assumption of extreme smoothness underlying OLS is substantively consequential.

## 4. GAMs in Practice

GAMs seem an appealing way to analyze political and social data. It is now time to see how helpful they are in actual political science research. The examples that follow come from a variety of subfields of political science, and were chosen to show some of the advantages of the GAM approach. In the conclusion we assess the usefulness of GAMs in political research more generally, suggesting research settings where GAMs are likely to be particularly useful.

### *4.1 The House Bank Scandal and the Congressional Vote: Assessing Transformation*

Our first example is a reanalysis of Jacobson and Dimock's (1994) assessment of the impact of the House Bank Scandal of 1991 on the 1992 House elections. Analyzing 309 races where incumbents sought reelection, they regressed the challenger's vote in the 1992 election (CV) on the log of

the number of overdrafts (*OVERDRAFTS*) (plus one) written by the incumbent, and a series of other "control" variables, comprising the logs of both the challenger's and incumbent's expenditures, the presidential vote for the challenger's party in the district, the log of the Perot vote in the district[11] and dummy variables for whether the incumbent was in a close race in 1990 (Marginal), whether the incumbent was subject to a partisan redistricting between 1990 and 1992 (Part. Redist.), and whether the challenger had ever held prior office. In Table 1 we first present a reestimation of Jacobson and Dimock's regression (their Table 9).[12]

Jacobson and Dimock chose to model the effect of overdrafts on the challengers vote via a logarithmic transformation. Since many incumbents had no overdrafts, they used the natural log of the number of overdrafts plus one. While we might expect that the impact of overdrafts on the vote is monotonically increasing, it is also reasonable to expect decreasing returns to scale. But Jacobson and Dimock neither justify nor examine their choice of a specific transformation. And while zeros in the data require some transformation before logging, adding one to the data is, of course, arbitrary. GAMs provide an ideal way of examining the consequences of Jacobson and Dimock's choices.

The overdrafts variable presents difficulties in that it is highly skewed. While its range is from zero to 851, the median number of overdrafts is two, with 40% of incumbents who sought reelection having had no overdrafts and almost 90% having had fewer than 50 overdrafts. The few incumbents with a large number of overdrafts brings the mean number of overdrafts to over 30.

While highly skewed variables are always a problem, the large number of incumbents with no overdrafts causes severe problems for the logarithmic transformation. While it is conventional to add one to all observations to avoid the log of zero, we could equally well add .01 or 10. Substantively, this decision has serious consequences for estimating the impact of OVERDRAFTS at low levels of that variable.[13] The GAM completely eliminates

[11]Jacobson and Dimock use the actual Perot vote, but a GAM analysis (Beck and Jackman 1997) shows that the logarithmic form is clearly superior.

[12]The data as supplied by Jacobson and Dimock were updated after publication of their article; measurement details are reported there. Analyses reported here use the updated data. Our results differ by at most a few percent from the published results. We report results only with the log of the Perot vote. We investigated, one variable at a time, whether we could improve the specification by replacing linear control variables with a smooth fit; the linear controls were never statistically rejected. Since interest here is on the effect of *OVERDRAFTS,* we did not pursue alternative specifications of the control variables further.

[13]These consequences stem from the logarithmic function having large first and second derivatives in the vicinity of zero. Thus we will get very different estimated impacts if we add .01, 1, or 10 before transforming. Theory is completely silent as to which of these values is appropriate.

**Table 1. Estimates of Challenger's Vote: 1992 Congressional Election.**

| | *OLS* | | *GAM: OVERDRAFTS* | |
|---|---|---|---|---|
| | *b* | *se* | *b* | *se* |
| *Constant* | –1.04 | 3.77 | –0.53 | 3.76 |
| *Prior Office* | 1.04 | .86 | .89 | .86 |
| *ln(Chal. Exp.)* | 2.65 | .26 | 2.63 | .26 |
| *ln(Inc. Exp.)* | –.15 | .60 | –.07 | .59 |
| *Pres. Vote* | .35 | .04 | .36 | .04 |
| *ln(OVERDRAFTS+1)* | .77 | .19 | *see figure 2* | |
| *Marginal* | .82 | .84 | .81 | .83 |
| *Part. Redist.* | 3.56 | 1.23 | 3.63 | 1.23 |
| *ln(Perot Vote)* | 3.46 | .95 | 3.47 | .95 |
| *Sum Sq. Resid.* | 9759.3 | | 9625.3 | |
| *Degrees of freedom* | 300 | | 297.3 | |

the need for the common arbitrary decision about how to avoid the log of zero.

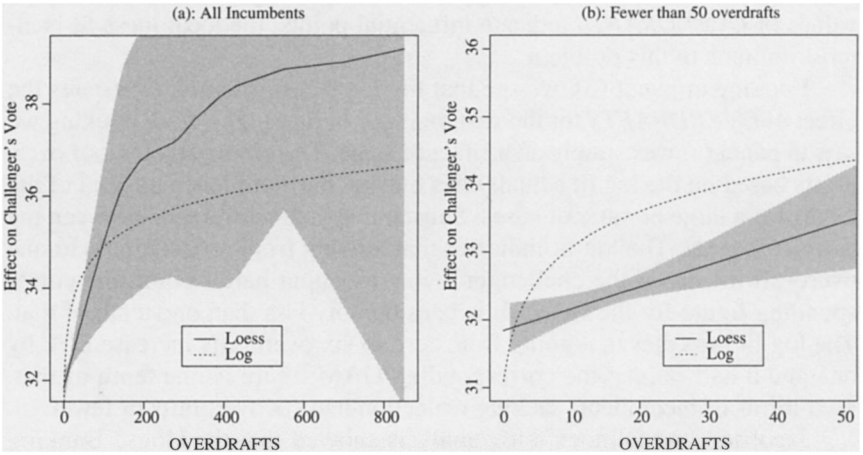
The logarithmic transformation is also problematic in that it is fit globally, not locally. Decreasing returns to scale for large values of *OVERDRAFTS* clearly requires the use of a downward bending transformation. But a log transformation struggles to fit both those incumbents with many overdrafts, as well as the large number of incumbents with few, if any, overdrafts. In particular, the few incumbents with hundreds of overdrafts are influential points; the logarithmic fit is forced upwards so as to better fit those few influential points, where, of course, the challenger did relatively well.

To better assess the relationship between overdrafts and challenger vote, we estimated a GAM relating *CV* to a nonparametric estimate of *OVERDRAFTS*. We chose a smooth loess fit, with a span of 90% of the data. As we note in the conclusion, we generally prefer relatively smooth fits (though not so smooth as the linear!), so that we do not have substantively meaningless small bumps in the smooth fit.[14]

The log and GAM models are not nested. If we nest both inside a bigger model, we find that we cannot reject either specification in favor of the

[14]We also tested the proposed smooth fit against less smooth fits with spans of 25%, 50%, and 75% of the data. The smoother fit was preferred in all cases. The less smooth fits also showed a nonmonotonicity in the neighborhood of 100 overdrafts; the fit we use is monotonic. We have a strong belief that more overdrafts must hurt an incumbent, providing another basis for our choice of the very smooth fit. We believe that specification tests combined with substantive knowledge is the appropriate way to make decisions about how much to smooth.

**Figure 2. Comparison of Logarithmic (Dashed Line) and loess Fit (Span = 90%), Effect of Overdrafts on Challenger's Vote**




An approximate 95% confidence interval for the loess smooth is shaded. Panel (a) is for all incumbents while panel (b) focuses on the region containing almost 90% of the data, allowing the eye to more clearly see the difference between the two models. Both panels are identical except for scale.

other. We could conclude that the GAM analysis is saying that the log transformation is not doing much injustice to the data.[15]

But we can use the estimated loess fit to improve our understanding of the effect of the House Bank scandal on the election. The estimated effect (the fitted values from the GAM, with all other independent variables set to their sample medians) are shown in Figure 2. Panel (a) covers the entire range of *OVERDRAFTS* while panel (b) enables us to better examine the relationship between the log and GAM fits for the region which contains almost all the data.

Panel (a) seems to indicate that the log and GAM fits track relatively closely until the number of overdrafts approaches 100. At that point the log model seems to understate the effect of *OVERDRAFTS*, although the confidence interval around the loess curve is enormous (because there are so few incumbents with huge numbers of overdrafts). The global log fit must balance these two portions of the data; the behavior of the GAM fit at the low end of *OVERDRAFTS* is much less affected by the high end of that scale. This is almost certainly a good thing; when the number of overdrafts is

[15]Another approach we tried was to use a smooth of the log of *OVERDRAFTS*. This analysis showed that we could not reject the null hypothesis of a logarithmic effect in favor of a smooth of that effect. However, the plot of that smooth is consistent with the thrust of the next paragraph.

large, the precise value of that score is clearly less interesting; do we really think it matters if an incumbent had 150 versus 200 overdrafts? The global log is sensitive to errors of this type, and particularly so here, where large values of *OVERDRAFTS* indicate influential points; the local loess fit is almost immune to this problem.

Looking at panel (b), we see that the log transformation overstates the effect of *OVERDRAFTS* for the vast majority of races. The close tracking we saw in panel (a) was simply an artifact of scale. The predicted effect of overdrafts based on the log fit actually lies outside the confidence interval of the GAM for a large number of cases. Substantively, the difference between the two fits is great. The log fit indicates that moving from no overdrafts to one overdraft increased the challenger's vote by about half a point; the corresponding figure for the GAM fit is considerably less than one tenth of that. The log fit indicates that going from zero to six overdrafts increased *CV* by one and a half points; the corresponding GAM figure is one tenth of that. Two-thirds of incumbents seeking reelection had six overdrafts or fewer.

Jacobson and Dimock's log analysis showed that the House Banking scandal had a nontrivial effect on the 1992 election; the GAM analysis shows that, while it is clear the scandal had an impact, it was much smaller than Jacobson and Dimock claim. So while the log transformation may do no great injustice to the *OVERDRAFTS* variable, the GAM appears superior here.[16]

### *4.2 Do Governments Face Rising Hazards?—an Analysis in Discrete, but Smooth, Time*

An enduring controversy in comparative politics is whether the probability of a cabinet failing increases with the age of the cabinet. Warwick (1992) argued that the hazard rate of cabinet failure increased over the life of the cabinet, while Alt and King (1994) argued for a constant hazard rate (both controlling for various institutional rules).[17] This controversy has both statistical and substantive implications. The incorrect assumption of a constant hazard rate leads to wrong standard errors, and wrong, or at best, inefficient, parameter estimates. The substantive issue is whether, controlling for

[16]Perhaps the problem is that the log is not a sufficiently flexible parametric transformation. We also estimated a model with a Box-Cox (Greene 1997, 479–85) transformation, which nests the log transformation. The estimated Box-Cox parameter is, however, not significantly different from 0($\lambda$ = .3, se = .2), suggesting that the log transformation is adequate. The Box-Cox transformation, like the log transformation, is a gobal, parametric transformation and is subject to similar problems. In addition, like the log, it requires a positive variable, which requires an arbitrary additive transformation.

[17]In survival analysis, the hazard rate is, loosely speaking, the probability of a cabinet failing in some small time interval ($t + \Delta t$) given that it had survived up to time $t$.

institutional rules, there is a tendency for cabinets to become more fragile as they age. If so, we need to investigate what happens over the life of a cabinet to make it more fragile; if not, then fragility is purely determined by institutional rules.

Parametric estimation of a continuous time Weibull model shows that the hazard rate is rising, at least for the model used by Alt and King (Beck N.d.). But this finding is based on a strong assumption, that the hazard rate increases (or decreases) monotonically, following the specific Weibull shape. One of us has used discrete time hazard models to investigate this issue (Beck N.d.). This analysis can be dramatically improved by the use of GAMs.

The discrete time approach models the probability of a government failing at the end of some month, given that it survived up to that month (the discrete hazard) as a logit on a series of covariates and a time marker. As is standard in this literature, the covariates we use are a measure of party polarization, whether the government is a majority government, how many prior formation attempts had been made, whether the coalition was formed right after an election, whether the government is a caretaker government, and whether there is an investiture requirement.[18]

Letting *FAILURE* be a dummy variable which marks whether a government fails in a given month, given that it survived up to that month, and removing governments which fail from further analysis, the discrete hazard model is a logit with

$$E\log\left[\frac{P\left(FAILURE_{it}\middle|ALIVE_{i,t-1}\right)}{1-P\left(FAILURE_{it}\middle|ALIVE_{i,t-1}\right)}\right]=\beta_0+\beta_1 x1_i+\cdots+\beta_k xk_i+h(t). \quad [12]$$

We can model $h(t)$ as a series of dummy variables, one for each time point (so $h(k)=h_k$, for $k=1,\ldots,48$). With monthly data, and cabinets surviving up to 48 months, this leads to imprecise estimates of the coefficients on the individual $h_k$'s; these are, not surprisingly, highly collinear. While it appears that the coefficients on the $h_k$'s increase over time, the large standard error on these estimates made this finding uncertain.

One solution to this problem would be to use a parametric model in time. But the shape of the time dependence is unknown and is exactly the point at issue. A low order polynomial approximation is overly restrictive, and, in any event, will likely have odd behavior as $t$ gets large. GAMs provide a better solution. Instead of a series of dummy variables, or an arbitrary polynomial, we simply model $h(t)$ as a relatively smooth function of time.

[18] These covariates are described in King et al. (1990). Our interest here is on the modeling of the hazard rate, not the covariates. Our results on the impact of the covariates are very close to those originally reported in King et al. (1990).

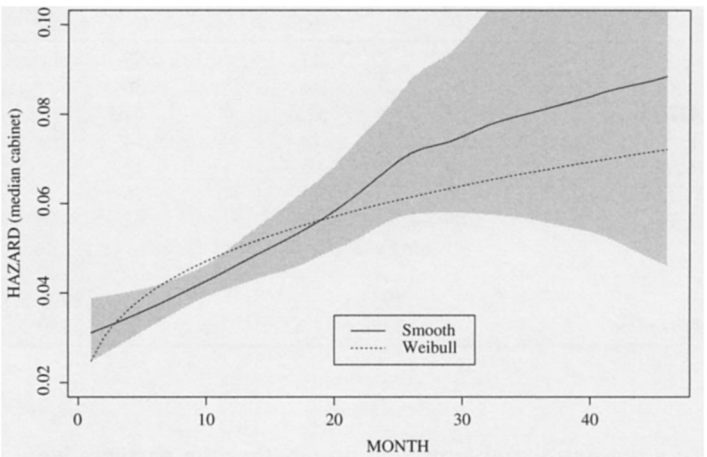
While hazard rates may be time dependent, we would not expect them to make large jumps from month to month. In modeling $h(t)$ with loess we chose a relatively smooth fit with a span of 75% of the data.[19] Since duration analysis is silent on the preferred functional form of $h(t)$, the loess fit is a particularly appropriate form. This fit can then be compared, as we do here, to standard duration analysis functional forms, giving us confidence in the chosen form; alternatively, we can simply use the loess fit as the estimate of $h(t)$.

The discrete hazard rate implied by the loess estimate of $h$ is shown in Figure 3 with coefficient estimates in Table 2. These hazard rates are estimated for a cabinet with all covariates set to their median value. The picture is clear. The hazard of a government falling increases over time. Examination of the 95% confidence interval shows that this decline is not simply random error. A $\chi^2$ test clearly shows that Equation [12] with a smooth function of time is superior to one without a time variable; we can reject the null hypothesis that time does not belong in the specification at $p = .001$. Hazard rates are rising, not constant. This rise is not trivial; the hazard rate doubles over the first two years of a cabinet's life, and then continues to increase, although at a slower rate.

Fully parametric duration models typically make an assumption about the shape of the hazard function. The most standard model, the Weibull, assumes monotonic hazards. Thus it is hard to know if the finding of rising hazards in a Weibull model is real or artifactual. The GAM discrete hazard model estimates the hazard function nonparametrically. Thus our finding of rising hazards gives credence to the Weibull findings. (It is not easy to test the adequacy of the Weibull assumptions using parametric methods.) A comparison of the Weibull hazard with the nonparametrically estimated hazard is also shown in Figure 3. Both hazard rates are similar for the first two years that a cabinet is in power; at that point the nonparametric discrete hazard rises faster than does the Weibull hazard. While this difference between the two hazard rates is not statistically significant, it does show that, if anything, the Weibull underestimates the degree of rising hazards. Thus the Weibull model, with rising hazards, does little injustice to the data. The GAM analysis clearly shows that the Alt and King model is subject to rising hazards. Even holding institutional rules constant, cabinets do appear to become more fragile as they age.

[19]It should be noted that we cannot reject the null hypothesis that the $h(t)$ are linear in $t$. The line clearly has positive slope, and so the hazard rate is monotonically increasing. Unlike many other applications, there is no particular reason to presume that a linear $h(t)$ should be preferred (and almost all events history models assume a nonlinear hazard function). Thus we would not consider an initial estimation of Equation [12] using a linear $h(t)$.

**Figure 3. Comparison of Estimated Smoothed (Discrete) and Weibull Hazard Rate for Median Cabinet**




The shaded area indicates an approximate 95% confidence interval for the smooth hazard.

*4.3 The Democratic Peace: Monadic or Dyadic*

The democratic peace hypothesis asserts that democracies, while no less likely to be involved in wars in general, are less likely to fight other democracies. This hypothesis has spawned much quantitative research (see e.g., Ray 1995). The typical research design used to assess this hypothesis analyzes data on some subset of all pairs of nations, with annual observations on dyadic variables, including the binary dependent variable indicating whether the dyad is involved in a militarized dispute. This data is analyzed by a straightforward logit analysis. Since theory is not specific on exactly what it takes for a nation to be a democracy, researchers have either used arbitrary dichotomizations (or trichotomizations) of relatively continuous democracy measures or have created *ad hoc* continuous measures of dyadic democracy. The two-dimensional loess smooth enables us to do better.

Here we reanalyze conflict in the post-World War II period, using the model and data reported in Oneal and Russett (N.d.). Following Oneal and Russett, we estimate a logit of *DISPUTE* (whether or not there was a militarized interstate dispute) on: a measure of dyadic democracy (see below); the lesser of the average annual growth rates in real GDP of the dyadic partners over the previous three years (*GROWTH*); economic interdependence measured by the lesser of the ratio of dyadic trade to GDP of the partners

**Table 2. Estimation of the Discrete Hazard of Cabinet Failure**

| | *With Smoothed Time* | | *Without Time* | |
|---|---|---|---|---|
| *Variable* | *b* | *se* | *b* | *se* |
| *Constant* | –2.87 | .17 | –2.88 | .17 |
| *INVESTITURE* | .41 | .14 | .36 | .14 |
| *POLARIZATION* | .025 | .006 | .020 | .006 |
| *MAJORITY* | –.63 | .14 | –.54 | .14 |
| *FORMATION* | .14 | .05 | .12 | .05 |
| *POST-ELECTION* | –.90 | .15 | –.76 | .14 |
| *CARETAKER* | 1.89 | .32 | 1.62 | .31 |
| *MONTH* | *see figure 3* | | | |
| *–2LogLike* | 1952.1 | | 1971.5 | |
| *Degrees of Freedom* | 5386.6 | | 5389 | |

(*TRADE*); a dummy variable indicating whether the partners were allied (*ALLIES*) and/or contiguous (*CONTIG*); and the relative balance of forces between the partners, with each partner's capability measured by economic, military, and demographic factors (*CAPRATIO*).[20]

In the earlier work Oneal et al. (1996) used a simple dummy variable to measure dyadic democracy. Oneal and Russett (N.d.) argued that the appropriate measure of dyadic democracy is the democracy score of the less democratic nation in the pair. This newer measure, *MINDEM,* is constructed by creating democracy scores for each member of the dyad (*DEMA* and *DEMB*) and taking the dyadic score as the lesser of the two. The democracy score for a nation is the difference between its measure on a democracy scale (running from zero to ten) and an autocracy scale (also running from zero to 10), so *DEMA, DEMB,* and *MINDEM* run from –10 to 10. While *MINDEM* is not an unreasonable measure of dyadic democracy, the democratic peace hypothesis is not stated so clearly as to make it obviously the correct measure.

The GAM/interaction approach allows us to investigate the interrelationship of *DEMA, DEMB,* and the likelihood of a dyadic dispute in any given year. We simply enter a smooth term in the interaction of *DEMA* and *DEMB* (a two-dimensional loess smooth) in a GAM of disputes.[21] The

[20]This data was provided by Oneal and Russett. We exactly replicated their Equation 1. Details on measurement are in Oneal and Russett (N.d.). The dataset contains 20,990 dyad-years from the 1950–85 period. The dependent variable is drawn from Bremer's (1996) update of the Correlates of War dataset. The democracy scores are drawn from Gurr and Jaggers' (1996) Polity III dataset.

[21]Since the model is symmetric in the dyadic partners, we randomly assign one nation as "A" for each observation.

smooth term used has a span of three quarters of the data; as usual, findings were not sensitive to smoothing choices.[22]

The simple logit setup of Oneal et al. assumes the dyadic observations are temporally independent. This assumption, as Oneal et al. recognize, is implausible. We can view the dispute time-series-cross-section data as duration data on how long dyads remain at peace (Beck N.d.; Beck and Tucker 1996). We therefore included a smooth term for how long a dyad has been at peace (*PEACEYRS*) in the logit specification to correct for temporal dependence.[23]

Results for the simple logit and GAM analyses are reported in Table 3. For reasons of space we focus only on the effect of democracy on peace. The conventional logit analysis shows that *MINDEM* affects the probability of dyadic dispute; as the less democratic partner becomes more democratic, the probability of a militarized dispute decreases. The two-dimensional loess fit gives more information about the role of each partner's level of democracy in inhibiting war. The estimated loess fit is shown in Figure 4. Panel (a) is a perspective plot, while panel (b) is a contour plot. Both have the partners' democracy scores on the $x$ and $y$ axes, with each point associated with a probability of conflict. This probability (in percent) is computed at the median of all the other independent variables, and is given by the inverse logit transformation of the predicted value of the GAM for differing combinations of the democracy scores. The isocurves of the contour plot join points in the $x, y$ space with the same estimated probability of conflict.[24]

A likelihood ratio test clearly indicates that the smooth fit of the democracy interaction belongs in the specification, and is superior to a specification with only linear and multiplicative terms in *DEMA* and *DEMB* ($p < .001$). Of perhaps more interest, estimating a combined model with both *MINDEM* and the smooth interaction shows that even with *MINDEM* already in the specification, the smooth interaction term also belongs ($p < .001$) while the *MINDEM* term does not. The GAM with a bivariate loess

[22]The issue of how much to smooth is not simple. With such a large $n$, $F$-tests indicate that a smaller span of half the data is superior. But this less smooth fit simply has a few more uninterpretable small bumps in the central portion of the democracy space. Any classical hypothesis testing framework will pick less parsimonious specifications when $n$ is large (since tests have enormous power under such a circumstance). Here we choose to go with the more parsimonious fit. In any event, the features of the bivariate smooth discussed here are the same for both smooth fits. But it should be stressed that choices about how much to smooth can never be purely mechanistic.

[23]Since our interest here is in the effect of democracy on disputes, we do not report the smoothed fit for *PEACEYRS* here; it is similar to Beck and Jackman (1997, fig. 8).

[24]The contours should be symmetric around the 45 degree line, but they are not quite so. Disputes are rare events. Thus it is not unlikely, that, by chance, we observe a few more disputes in the northwest region as compared to the southeast. The asymmetry is not great.

**Table 3. Estimates of Logit Model of Post-World War II Disputes[a]**

| | *Logit* | | *GAM* | |
|---|---|---|---|---|
| | *b* | *se* | *b* | *se* |
| *MINDEM* | –.050 | .007 | | |
| *DEMA, DEMB* | | | *see figure 4* | |
| *GROWTH GDP* | –.022 | .009 | –.019 | .009 |
| *ALLIES* | –.82 | .08 | –.41 | .09 |
| *CAPRATIO* | –.0031 | .0004 | –.0030 | .0004 |
| *TRADE* | –66.1 | 13.4 | –13.5 | 10.4 |
| *CONTIG* | 1.31 | .08 | .73 | .09 |
| *PEACEYRS* | | | *see text* | |
| *Constant* | –3.29 | .08 | –1.52 | .07 |
| *–2LogLike* | 6955.1 | | 5226.7 | |
| *Degrees of Freedom* | 20983 | | 20976 | |

[a]Data are in dyad-year form.

Dependent variable is whether dyad engaged in a militarized dispute in a given year.

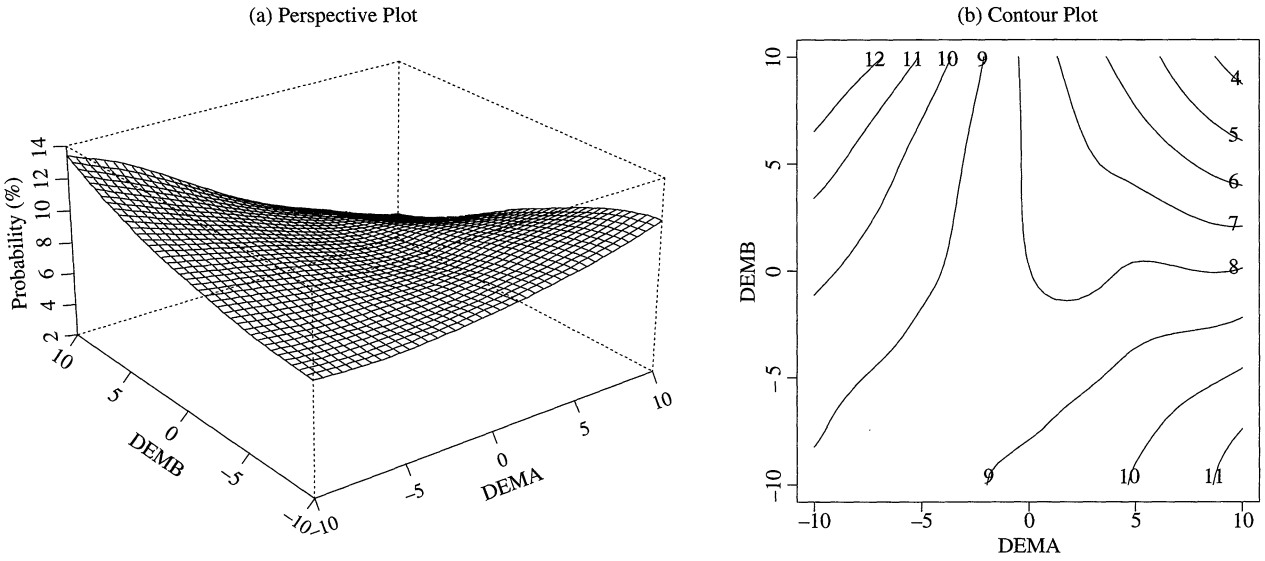
smooth of *DEMA* and *DEMB* is clearly statistically superior to the logit model with *MINDEM*.

The value of the GAM is not, however, in obtaining improved fits, but rather in allowing us to delineate the impact of democracy on war, and in particular to assess whether each nation's democracy matters, or whether only dyadic democracy matters. It is easiest to examine the GAM plots in several different regions separately.

The perspective plot shows that the probability of conflict declines markedly when both partners are democratic; this plot is consistent with the democratic peace hypothesis. The contour plot allows for a more complete assessment of this hypothesis. In the northeast democratic region (where both *DEMA* and *DEMB* are positive), we see the democratic peace. As dyads become more democratic, the probability of a militarized dispute declines. Moving from a dyad where both partners have a democracy score of zero to a dyad where both score 10 decreases the probability of a dispute by about four points. The isocurves are almost downward sloping 45 degree lines (this is even clearer in the plots without the inverse logit transformation). This means that *DEMA* and *DEMB* are good substitutes for each other; what matters in this region is the sum of *DEMA* and *DEMB*. Thus, contrary to the democratic peace hypothesis, it is the democracy scores of each partner separately that matter, rather than some nonadditive dyadic measure.

We note that in this region Oneal and Russett's *MINDEM* measure is also misleading. If *MINDEM* were a good measure, the isocurves should consist of a series of horizontal and vertical lines all radiating from the 45

**Figure 4. Perspective and Contour Plots of Estimated Joint Effects of Each Partner's Democracy Score on the Probability of a Militarized Dispute. They show the "topography" of the fitted surface.**



degree line. This is clearly not the case. We also see that a simple dichotomization of dyadic democracy would not be appropriate; the plots clearly indicate that the effect of democracy on disputes is not simply two plateaus.

Turning to the southwest (nondemocratic) regions of the plots, we see relative flatness; in this region the probability of a dispute hovers between 8% and 10% (and in the majority of the region it is between 8.5% and 9.5%). Thus there is a relatively constant probability of conflict between autocracies, regardless of exactly how autocratic either nation is.[25]

Finally, in the northwest or southeast regions, where democracies face autocracies, we get a puzzling result. In these regions the probability of a dispute *increases* as the democratic partner becomes more democratic. Depending on whether we wish to look at the northwest (or southeast) regions, dyads consisting of the most and least democratic nations are four (or two) points more likely to engage in a dispute than are dyads consisting of two maximally autocratic nations. This does not conflict with the democratic peace hypothesis, which states only that democracies do not fight democracies, but it is a puzzling result which demands explanation.

The bivariate loess transformation of *DEMA* and *DEMB* provides a flexible statistical model of the relationship of dyadic democracy and disputes. It clearly shows that current measures of dyadic democracy are flawed, and the loess transformation itself provides a better understanding of how dyadic democracy relates to international conflict. It also shows that a monadic theory, where each nation's level of democracy matters, may be superior to a dyadic theory, at least in understanding conflict among the more democratic nations. We know of no other method for doing this type of analysis.

## 5. Discussion: the Strengths and Weaknesses of GAMs

Political methodology has made great strides in the last decade. Political scientists now routinely use models driven by a wide variety of statistical distributions, and have started to use nonparametric methods to allow for more realistic error processes. Nonetheless, statistical models in political science frequently impose a linear relationship between covariates and the

[25]It is *graphically* hard to add confidence regions about the surface plotted in Figure 4. A simple method of estimating uncertainty about the plotted surface is to note that three quarters of the point-wise standard errors for the loess smooth are less than .1 (with all being less than .14). We can compute contours adding or subtracting .2 (approximately two standard errors) to the war effect. When converted back to probability units, the probability of war associated with democracy scores of (0,0) is between 7% and 9.5%, showing that the undulations in this region may simply be artifacts of statistical uncertainty. In the extreme democratic region (10,10), a similar computation shows that we can estimate a rough confidence interval for the probability of war at between 3.5% and 4.5%. Movements in the northwest region of the plot reflects much more than statistical noise. The asymmetry between the southeast and northwest corners, on the other hand, may be the result of such noise.

dependent variable. We find this especially puzzling since there is almost always *no* theory that mean functions should follow this simple linear form.

GAMs possess several strengths for confronting this shortcoming.

- GAMs are very flexible in modeling the effects of specific variables, but not so pliable that they find nonlinear patterns where none exist.

GAMs may well be the ideal compromise between simple linear models and much more complex models, such as neural nets and projection pursuit.

- The additivity constraint makes GAMs easy to interpret, a property not shared by its more complicated cousins.

These strengths aside, analysts will choose among the numerous nonparametric regression techniques (perhaps choosing to ignore them altogether), drawing on their experiences with statistical modeling, and their intuitions about the particular substantive problem at hand. But the combination of flexibility and simplicity captured by the GAM means it can lay a strong claim to be the "workhorse" nonparametric regression method.

- GAMs are likely to be useful whenever the analyst has independent variables that take on values on more than just a few discrete categories.

GAMs work equally well on interval and ordinal data, and do not even assume that the effect of a variable is monotonic; for these types of independent variables, GAMs essentially avoid the entire level-of-measurement issue. And as we have seen, GAMs work for a wide variety of dependent variables: continuous, binary, and count.[26] However,

- GAMs can also be used by researchers committed to parametric methods.

GAMs can be used to either test the adequacy of a chosen parametric transformation (including, perhaps most importantly, the implicit assumption of linearity) or to suggest a superior transformation. While we feel that parametric transformations of independent variables are generally inferior to local fitting, a carefully chosen global transformation is better than one

[26] The typical GAM user will be limited to what types of models can be estimated with currently available software. While any GLM can be estimated using the GAM routines in SPLUS, we are unaware of software for utilizing GAMs in contexts such as systems of simultaneous equations, or models of censoring and sample selection. Other than programming difficulties, there is no reason why GAMs can not be used in these situations. Researchers are continually extending the application of GAMs. For example, Kustra (1995) has recently written programs to allow duration modelers to estimate GAMs for the Cox proportional hazards model.

chosen by convention or prayer. In this regard, perhaps the value of a GAMs will stem from the simple fact that they demand analysts graphically inspect their data.

Several potential shortcomings of GAMs should also be pointed out.

- GAMs can be computationally intensive.

Given modern computing power, the added computational demands of GAMs are negligible when working with small to moderately-sized data sets. But our experience is that when confronted with data sets considered large by social-science standards, GAMs can place enormous demands on the computing resources typically available to social-scientists.[27] All the same, we are confident that advances in computing power and software design will continue to transform today's computational nightmares into everyday tasks for the next decade's social scientists.

- A fitted GAM might be seen as harder to communicate than a vector of parameter estimates (and their standard errors).

If the analyst believes that the effect of $x$ on $y$ is a nonparametric function of $x$, varying over the range of $x$, then obviously no single number can summarize the relationship between $x$ and $y$. In these circumstances we see no alternative to plotting the estimated nonparametric effect of $x$ on $y$, as we have here. While graphs of (nonparametric) marginal effects can quickly chew up valuable journal space, we would dispute the claim that they are necessarily "harder to understand" than parameter estimates. If anything, a graph of a nonlinear marginal effect often conveys what is going on more clearly and more efficiently than can prose alone; that is, "a picture tells a thousand words," and in fewer column-inches (e.g., Franklin 1989).

- GAMs appear to make it all too easy for analysts to invent "theories" to account for small bumps in the plots.

While GAMs can be used purely inductively, they can also be used to test theories (e.g., is the shape of the plot consistent with theoretical predictions?). Undersmoothing can lead to many local idiosyncrasies which may cry out for explanation. Since we do not believe that these narrowly local features are usually of interest in the social sciences, we prefer to avoid this problem by estimating rather smooth fits. Careful attention to confidence intervals and specification tests are crucial for distinguishing statistical noise

[27] The data set assembled for the democratic peace example has over 25,000 observations, and estimating a logit model with a two-dimensional loess fit for the levels-of-democracy independent variables took over 12 hours on a large workstation that had to be reconfigured especially for this task. In part this reflects our choice of software, SPLUS (which does not handle data sets of this size particularly well). This example is atypical. The other two analyses were done on a standard desktop machine running Windows.

from interesting local features. But as we saw in the democratic peace example, the GAM can be useful for finding empirical puzzles, which must be then explained by further theoretical work. The GAM is a powerful tool for finding patterns in data; like all such tools, it can be used well or used poorly.

## 6. Conclusion

We presented a variety of applications of the GAM in Section 4. The GAM helped us to better understand the substantive effect of the number of overdrafts drawn on 1992 House reelection contests; this effect is substantively smaller than had been previously claimed. And at the very least, our reanalysis also showed that those committed to parametric forms could better inform their choice of parametric form by using the GAM. These reanalyses show it is straightforward for analysts familiar with regression and logit analysis to recast those models as GAMs.

Our reanalyses of cabinet durations shows that GAMs are an ideal tool for the analysis of discrete time duration data, and can settle important substantive and statistical questions that are almost impossible to deal with any other way. Currently, standard practice is to impose specific forms of age effects, such as the Weibull. The GAM can either be used to justify such forms or to go beyond them.

The reanalysis of the democratic peace data also shows how useful the GAM is for modeling bivariate interactions. While many theories suggest interactions in social and political data, they are hard to estimate precisely in a linear framework. The GAM makes such estimation easy. We know of no other tool that could have shown that the dyadic democratic peace hypothesis holds, but also showing that the hypothesis must be refined: for the more democratic nations, an increase in either nation's level of democracy decreases the probability of conflict.

Analyzing data involves making compromises. The linear regression model is one such compromise (though typically not presented as such), providing a simple statistical answer to the substantive question "does $y$ increase/decrease with $x$?" When its underlying assumptions are valid, the linear regression model is rightly hailed as a powerful tool: for the cost of just one degree of freedom an analyst learns as much as can be learned from the sample data about $\beta$. But as our applications demonstrate, assumptions of global, linear effects do not always hold true, and linear regression may well be providing optimal estimates of a relationship of limited substantive relevance. In such circumstances linear regression stops being a useful compromise, and may even blind researchers to what is actually going on in the data.

Instead, GAMs puts a *range* of compromises between parsimony and fidelity to the data within reach of empirical social scientists, with linear regression at one extreme. Throughout this paper, we have claimed that getting

the mean right is a good thing. If we include statistical power and substantive interpretability as desiderata, then surely GAMs are good things too.

*Manuscript submitted 13 November 1996.*
*Final manuscript received 27 May 1997.*

## REFERENCES

Alt, James, and Gary King. 1994. "Transfers of Government Power: The Meaning of Time Dependence." *Comparative Political Studies* 27:190–210.

Beck, Nathaniel. N.d. 'Modeling Space and Time: The Event History Approach." In *Research Strategies in Social Science,* ed. Elinor Scarbrough and Eric Tanenbaum. Oxford: Oxford University Press.

Beck, Nathaniel, and Simon Jackman. 1997. "Getting the Mean Right is a Good Thing: Generalized Additive Models." Working paper. Political Methodology WWW Site. http://wizard.ucr.edu/polmeth/working_papers 97.

Beck, Nathaniel, and Richard Tucker. 1996. "Conflict in Space and Time: Time-Series—Cross-Section Analysis with a Binary Dependent Variable." Presented at the annual meeting of the American Political Science Association, San Francisco.

Breiman, L., and J. H. Friedman. 1985. "Estimating Optimal Transformations for Multiple Regression and Correlation (with discussion)." *Journal of the American Statistical Association* 80:580–619.

Bremer, Stuart A. 1996. "Militarized Interstate Disputes." Technical report.

Bueno De Mesquita, Bruce, and Randolph M. Siverson. 1995. "War and the Study of Political Leaders: A Comparative Study of Regime Types and Political Accountability." *American Political Science Review* 89:841–855.

Castles, Frank, and Robert D. McKinlay. 1979. "Does Politics Matter: An Analysis of the Public Welfare Commitment in Advanced Democratic States." *European Journal of Political Research* 7:169–186.

Cleveland, William S. 1993. *Visualizing Data.* Summit, NJ: Hobart Press.

Cleveland, William S., and Clive Loader. 1996. "Smoothing by Local Regression: Principles and Methods." In *Statistical Theory and Computational Aspects of Smoothing,* ed. W. Härdle and M. G. Schimek. Heidelberg: Physica Verlag.

Cleveland, William S., E. Grosse, and W. M. Shyu. 1992. "Local Regression Models." In *Statistical Models in S,* ed. John M. Chambers and Trevor J. Hastie. Pacific Grove, CA: Wadsworth and Brooks/Cole.

Cleveland, William S., and S. J. Devlin. 1988. "Locally Weighted Regression: An Approach to Regression Analysis by Local Fitting." *Journal of the American Statistical Association* 83:596–610.

Cleveland, William S., Susan J. Devlin, and Eric Grosse. 1988. "Regression by Local Fitting: Methods, Properties and Computational Algorithms." *Journal of Econometrics* 37:87–114.

DeVeaux, R. 1990. "Finding Transformations for Regression Using the ACE Algorithm." In *Modern Methods of Data Analysis,* ed. John Fox, and J. Scott Long. Newbury Park, CA: Sage.

Franklin, Charles. 1989. "Graphic Displays in Political Science." *The Political Methodologist* 2:7–9.

Friedman, Jerome H., and W. Stuetzle. 1981. "Projection Pursuit Regression." *Journal of the American Statistical Association* 76:817–23.

Greene, William. 1997. Econometric Analysis 3rd ed. Upper Saddle River, NJ: Prentice Hall.

Gurr, Ted Robert, and Keith Jaggers. 1996. "Polity III: Regime Change and Political Authority, 1800–1994." Computer file, Inter-university Consortium for Political and Social Research, Ann Arbor, MI.

Hall, Richard L., and Robert P. Van Houweling. 1995. "Avarice and Ambition in Congress: Representatives' Decisions to Run or Retire from the U.S. House." *American Political Science Review* 89:121–36.

Hastie, Trevor J., and R. J. Tibshirani. 1990. *Generalized Additive Models* London: Chapman and Hall.

Hastie, Trevor J., and Clive Loader. 1993. "Local Regression: Automatic Kernel Carpentry (with Discussion)." *Statistical Science* 8:120–43.

Huber, Peter J. 1981. *Robust Statistics.* New York: Wiley.

Isaac, Larry W., and Larry J. Griffin. 1989. "Ahistoricism in Time-series Analyses of Historical Process: Critique, Redirection, and Illustrations from U.S. Labor History." *American Sociological Review* 54:873–890.

Jacobson, Gary, and Michael Dimock. 1994. "Checking Out: The Effects of Bank Overdrafts on the 1992 House Elections." *American Journal of Political Science* 38:601–24.

King, Gary, James Alt, Michael Laver, and Nancy Burns. 1990. "A Unified Model of Cabinet Dissolution in Parliamentary Democracies." *American Journal of Political Science* 34:846–871.

Kustra, Rafal. 1995. "A GAM Cox Proportional Hazards Model." Software archive. Department of Statistics, University of Toronto. http://lib.stat.cmu.edu/S/cox-ph.

Loader, Clive R. 1996. "Local Regression and Likelihood: A Guide to the LOCFIT software." Lucent Technologies. Typescript. http://cm.bell-labs.com/stat/project/locfit.

Lodge, Milton, Marco R. Stenbergen, and Shawn Brau. 1995. "The Responsive Voter: Campaign Information and the Dynamics of Candidate Evaluation." *American Political Science Review* 89:309–326.

McCullagh, P., and J. A. Nelder. 1989. *Generalized Linear Models.* 2nd ed. London: Chapman and Hall.

Oneal, John R., Francis H. Oneal, Zeev Maoz, and Bruce Russett. 1996. "The Liberal Peace: Interdependence, Democracy and International Conflict." *Journal of Peace Research* 33:11–29.

Oneal, John R., and Bruce Russett. N.d. "The Classical Liberals Were Right: Democracy, Interdependence, and Conflict, 1950–1985." *International Studies Quarterly.* Forthcoming.

Pagan, Adrian, and Aman Ullah. Forthcoming. *Non-Parametric Econometrics*. New York: Cambridge University Press.

Ray, James Lee. 1995. *Democracy and International Conflict: An Evaluation of the Democratic Peace Proposition.* Columbia, SC: University of South Carolina Press.

Ripley, Brian D. 1993. "Statistical Aaspects of Neural Networks." *In Networks and Chaos - Statistical and Probabilistic Aspects,* ed. O. E. Barndorff-Nielsen, J. L. Jensen, and W. S. Kendall. London: Chapman and Hall.

Ruppert, D., S. J. Sheather, and M. P. Wand. 1995. "An Effective Bandwidth Selector for Local Least Squares Regression." *Journal of the American Statistical Association* 90:1257–70.

Smith, Michael, and Robert Kohn. 1997. "A Bayesian Approach to Nonparametric Bivariate Regression." *Journal of the American Statistical Association* 92:1522–35.

Warwick, Paul V. 1992. "Rising Hazards: An Underlying Dynamic of Parliamentary Government." *American Journal of Political Science* 36:857–76.

Weisberg, Sanford. 1985. *Applied Linear Regression.* New York: Wiley.

Western, Bruce. 1995. "Concepts and Suggestions for Robust Regression Analysis." *American Journal of Political Science* 39:786–817.

White, Halbert. 1992. *Artificial Neural Networks: Approximation and Learning Theory.* Oxford: Basil Blackwell.

Wilensky, Harold L. 1975. *The Welfare State and Equality.* Berkeley: University of California Press.