Dose-Response and Trend Analysis in Epidemiology: Alternatives to Categorical Analysis

Sander Greenland

Standard categorical analysis is based on an unrealistic model for dose-response and trends and does not make efficient use of within-category information. This paper describes two classes of simple alternatives that can be implemented with any regression software: fractional polynomial regression and spline regression. These methods are illustrated in a problem of estimating historical trends in human immunodeficiency virus incidence. Fractional polynomial and spline regression are especially valuable when important nonlinearities are anticipated and software for more general nonparametric regression approaches is not available. (Epidemiology 1995;6:356–365)

Keywords: biostatistics, epidemiologic methods, logistic regression, relative risk, risk assessment.

Dose-response and trend analyses in epidemiology are commonly conducted in a very simple and often naive fashion. At worst, authors only conduct a trend test such as the Mantel test, or fit a regression model with a single exposure term and test the significance of the slope (coefficient) for the exposure. Such an approach can be very misleading because, in essence, it *assumes* that the dose-response or trend curve follows a specific model form (usually logistic).[1]

More desirably, authors may break the range of the study exposure into categories and look for trends in the category-specific coefficients or relative risks.[2] Such an approach can be adequate if numbers allow the use of categories that reflect biologically homogeneous response groups or are very narrow. Too often, however, categories are chosen via a mechanical algorithm such as the percentile method, in which equal-sized categories (tertiles, quartiles, or quintiles) are chosen in the belief that such an approach will maximize accuracy and minimize subjectivity in the analysis. The potential pitfalls of percentiles are most dramatic when most subjects are exposed in a very narrow range or when exposure effects are limited to extreme ends of the exposure scale, such as very low nutrient levels or very high occupational exposure levels. In such situations, individuals placed at elevated risk by exposure will be submerged among lower-risk members of their percentile category. This hazard can sometimes be mitigated by basing percentiles on the case distribution, rather than the distribution of all subjects, but would be desirable to avoid altogether.

Many authors have recommended nonparametric regression as a means of avoiding the categorization problem altogether.[1,3,4] This is a preferable approach, especially when one can safely assume nothing about the form of the trend or the exposure-disease (dose-response) relation. It is mildly hindered by lack of widely available software, although this obstacle is gradually disappearing. Another occasional drawback is that the computing limits (maximum numbers of covariates and subjects) for nonparametric regression tend to be much lower than those for conventional regression. Because of these limits, and because several books on the topic are available,[3–5] I will not discuss nonparametric regression here. Instead, I will describe two alternative curve-fitting methods that seem under-used in epidemiologic research. The two methods, fractional polynomial regression and spline regression, can be performed with any regression program simply by adding some transformed exposure variables to the regression. Both methods are intermediate between simple regression and nonparametric regression in behavior, with fractional polynomials closer to simple regression (but still a vast improvement) and spline regression falling closer to nonparametric regression (so close that it may be considered an approximation to nonparametric regression). As will be discussed below, both categorical analysis and splines can be viewed as special types of category-specific regression, but splines are based on more realistic category-specific models.

In what follows, I will denote the exposure of interest by x. All points apply even when x is only a time variable for which trends are to be plotted, or a confounder for which close control is desired. The following analysis of secular trends in human immunodeficiency virus (HIV) infection incidence will serve to illustrate all of the

From the Department of Epidemiology, UCLA School of Public Health, Los Angeles, CA 90095-1772.

This work was partially supported by the HIV Epidemiology Program of the Los Angeles County Department of Health Services.

Submitted October 14, 1994; final version accepted February 10, 1995.

TABLE 1. New AIDS Diagnoses in Los Angeles County through 1992 Reported by 1994 among White Non-Hispanic Men Who Have Sex with Men Reporting No Injection Drug Use, 1951–1960 Birth Cohort, and Naive HIV Incidence-Rate Estimates

Year	Years since 1976 (j)	New AIDS Cases (y)	White Non-Hispanic Person-Years, in Thousands (n_x)	Naive HIV Rate Estimate*
1979	3	0	362	0
1980	4	0	360	100
1981	5	4	355	519
1982	6	10	353	76
1983	7	46	351	728
1984	8	92	349	1,744
1985	9	201	348	433
1986	10	329	348	65
1987	11	456	348	1,820
1988	12	476	347	111
1989	13	606	347	2,013
1990	14	642	348	192
1991	15	791	344	57
1992	16	645	339	119
Total		4,298		

* Number of infections per 100,000 person-years (see Appendix).

methods discussed in this paper. Throughout, the focus will be on estimation of the shape of dose-response or trend; a companion article[6] describes the advantages of splines in testing for dose-response and trends.

General Description of Example

A major task in the study of acquired immunodeficiency syndrome (AIDS) is estimation of historical trends in HIV infection incidence.[7] Table 1 presents the 4,298 AIDS cases diagnosed in Los Angeles County through 1992 and reported by 1994 among white non-Hispanic men who have sex with men (MSM) born 1951–1960 who reported no injection drug use (IDU). Because there are no reliable data on cohort-specific prevalences of behaviors that define HIV transmission groups (such as sexual behavior), the HIV rates refer to the number of HIV MSM cases that reported no injection drug use among white non-Hispanic men born 1951–1960, rather than the number of HIV cases among non-IDU white non-Hispanic MSM born 1951–1960.

Because HIV incidence has not been directly observed, historical HIV incidence is computed from observed AIDS incidence using estimates of the distribution of incubation time from HIV infection to AIDS diagnosis.[7,8] The final column of Table 1 presents HIV rate estimates derived from a backcalculation equation, given in the Appendix, that relates AIDS to HIV incidence. These naive estimates involve no model or grouping of years. As a result, they present a noisy pattern and would fluctuate wildly in response to minor changes in the data or the estimation method.

More stable estimates require use of a model for the HIV rates. In the examples below, a series of models for these rates will be fitted via a Poisson regression method described in detail elsewhere[8,9] and summarized in the Appendix. The important elements for the present discussion are the structural forms of the models. To describe them, let x be years since 1976 (1976, then, is year 0, which is commonly taken as the start of the epidemic), n_x the person-years at risk in year x, and r_x the HIV incidence rate in year x. The simple log-linear model

$$r_x = \exp(\alpha + \beta x) \qquad (1)$$

is out of the question, because it implies that HIV incidence rates continued to increase exponentially through the 1980s and beyond, contrary to extensive evidence of leveling and decline in the 1980s.[5] Hence βx must be replaced by a more flexible set of trend terms. Figure 1 presents the fitted HIV incidence rates derived from Table 1 using five different choices for these terms, each with four coefficients (beyond the intercept): (1) four category indicators for five categories (*dotted line*); (2) fractional polynomial with four powers of untransformed time (*short dashes*); (3) fractional polynomial with four powers of log time (*solid curve*); (4) linear spline with four categories of log time (*long dashes*); (5) quadratic spline with three categories of log time (*solid curve* again—it almost perfectly agrees with the fractional polynomial with log time). The remainder of the paper will describe each choice in detail.

As a special caution in interpreting Figure 1, note that the very long incubation time between HIV infection and AIDS incidence (median time on the order of 10 or more years[10]) implies that the data in Table 1 contain almost no information on HIV incidence after 1989.

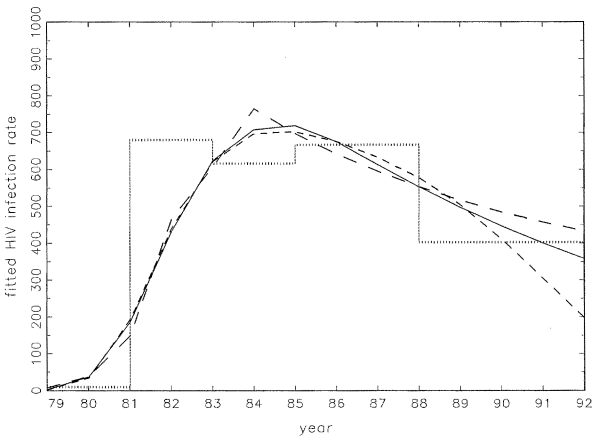

FIGURE 1. Fitted HIV infection incidence in Los Angeles County, 1979–1992: non-IDU MSM cases per 100,000 person-years among white non-Hispanic men born 1951–1960. *Short dashes:* fractional polynomial curve in untransformed time; *solid curve:* fractional polynomial curve in log time and (coinciding) quadratic spline curve; *dotted line:* step function from category indicators; *long dashes:* linear spline.

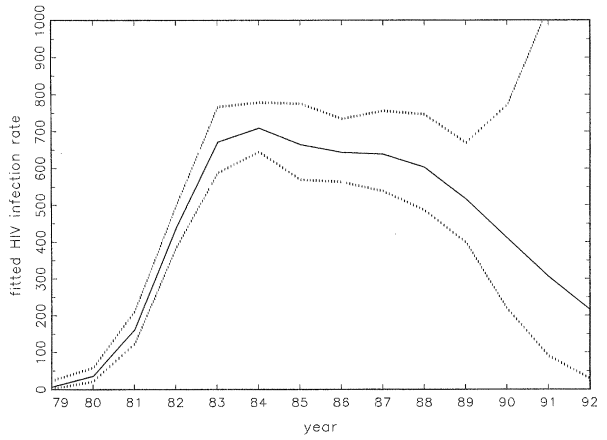

FIGURE 2. Fitted penalized spline HIV incidence curve (*solid curve*) with pointwise 95% confidence limits (*dotted curves*).

Thus, after 1989 the curves are little more than extrapolations from previous years (hence the increasing divergence beyond 1989). On the other hand, the data do provide a reasonable amount of information on HIV incidence before 1989. This point is illustrated in Figure 2, which shows the fitted curve and pointwise confidence limits obtained by using a penalized spline smoother as part of a multivariate model for HIV incidence[9] (the kinks in these three curves are artifacts of the graphing program). As a secondary caution, note that the curves in Figure 1 cannot be obtained by fitting models to the naive HIV rates in Table 1; all fitting must instead be done via backcalculation from observed AIDS incidence (Appendix Eq A1).

Fractional Polynomial Regression

Many authors recommend that one try to examine polynomial terms (at least a quadratic term x^2) in addition to the basic linear term x in the dose-response model.[1] There are problems with polynomial regression, however. Although in theory with enough polynomial terms one can approximate any smooth curve, in reality the number of terms required may be so large as to result in numerically unstable estimates. Polynomials greater than quadratic tend to produce artifactual turns in the fitted curve, whereas quadratics have extremely limited flexibility.

Recently, Royston and Altman[11] have emphasized that a great deal more flexibility and stability can be obtained by examining fractional and inverse powers of x, such as x^{-2}, x^{-1}, $x^{-1/2}$, and $x^{1/2}$ in addition to x and x^2. (Terms of the form $x^p[\ln(x)]^j$ are also included in the family of curves considered by Royston and Altman, but these cannot be used if x can be zero or negative.) Royston and Altman point out that models containing as few as three different powers of x between x^{-2} and x^2 encompass a dramatic range of shapes.

Fractional polynomials do have important limitations.[12,13] For example, x cannot be negative if fractional powers are used, and the results will be sensitive to the position of the zero-level of exposure x. Thus, fractional polynomials may be problematic if x is not ratio scaled; that is, it is advisable that x have an absolute zero level (unexposed level) and be coded so that this level is zero.[12] Nonetheless, many, if not most, epidemiologic exposures have an absolute zero, so that this limitation may be of infrequent practical importance. (If x can be negative, Royston and Altman recommend adding a positive number to force it to be positive, but this approach essentially introduces a new nonlinear parameter into the model, because the optimal number to add is unknown.)

In the HIV example, x does have an absolute zero: it is the time at which the epidemic started, which has not been precisely determined but is customarily taken to be 1976. A similar problem (of an absolute but imprecisely known zero) arises when using age in studies of adult noninfectious diseases: Age is often a surrogate for time since start of an unmeasured background exposure (for example, hormones) or etiologic process. In such situations, it may be advisable to replace age by a more biologically relevant time scale in which risk becomes nonzero only after time zero. For example, time since puberty could serve as such a scale in certain studies of cancers of the reproductive system.

Another problem, one which also afflicts polynomial regression, is how to decide which terms to include. Royston and Altman propose a special stepwise procedure, which (like all stepwise procedures) is questionable in concept and requires special programming. Ideally, one should specify in advance the shape of curves one would want the fitted model to encompass. To do so, however, requires a sense of what shapes are encompassed by each power of x. For most epidemiologic purposes, it suffices to recall that, as x increases above 0, x^2 starts more slowly but soon increases more rapidly than x, and that $x^{1/2}$ starts more rapidly but soon increases more slowly than x. From this, a simple qualitative dose-response analysis might always include x (the linear term) and then:

1. Include x^2 if one expects the slope of the trend or dose-response curve (that is, the steepness, or effect per unit exposure) to increase in absolute value as exposure increases (as with cigarettes per day and lung cancer[14]), or if one expects the curve to change direction.
2. Include $x^{1/2}$ if one expects the slope to decrease in absolute value as exposure increases.
3. Include both $x^{1/2}$ and x^2 if one wants to allow for either possibility.

One may, of course, use a higher power of x in place of x^2 and a lower power of x in place of $x^{1/2}$ if one expects more rapid changes in slope over the range of exposure, and one may include more terms if greater flexibility is desired.

If x can only be positive (as with typical cardiovascular and anthropometric measurements), $\ln(x)$ can be

used in place of $x^{1/2}$ to yield a curve with a very gradually declining slope. In fact, if one uses a logistic or exponential (log-linear) model for risks or rates and x can only be positive, it can be argued that $\ln(x)$ should be included in all dose-response and trend analyses. This is because the use of x alone in such models implies that the rate, risk, or odds ratio for exposure level x *vs* zero is $e^{\beta x}$, which increases exponentially with x if $\beta > 0$. Use of $\ln(x)$ instead yields a rate, risk, or odds ratio of $\exp[\beta \ln(x)] = x^{\beta}$, which can increase much less rapidly than exponentially and can even increase less than linearly if $0 < \beta < 1$.

Example

The *short dashes* in Figure 1 trace the fitted curve obtained from Table 1 using:

$$r_x = \exp(\alpha + \beta_1 x^{1/2} + \beta_2 x + \beta_3 x^{3/2} + \beta_4 x^2).$$

A virtually identical curve was obtained using x^3 instead of $x^{3/2}$. The *solid line* traces the fitted curve obtained using powers of $\ln(x)$ in place of powers of x as the time covariate. Both curves exhibit essentially exponential growth until 1982, followed by rapid slowing with a peak in 1984, and gradual decline thereafter.

Although fractional polynomials with only two or three exposure terms can produce quite a variety of curves, one should be aware when examining such curves that their exact shape and location can be strongly influenced by one or a few data points. In particular, the fitted values for a point can be strongly influenced by data at points far away on the graph.[13] This is also a problem with curves fit by quadratic or cubic regression, and with the single slope produced by a simple regression with only x (the linear term) included. Thus, it is especially important to evaluate regressions with few exposure terms using influence analysis, which involves seeing how much results change when the most influential data points are deleted from the analysis (in the present HIV example, the basic conclusions are unchanged by single deletions). Inclusion of confidence limits in the curve graph can also help indicate what portions of the curve are poorly estimated. (Methods for constructing confidence limits are described in the Discussion.)

Spline Regression

Category Indicators Revisited

Consider ordinary categorical dose-response analysis[2] from the following perspective: One divides the observed range of exposure x into K categories indexed by $k = 1, \ldots, K$ with $K - 1$ internal boundaries $c_1, \ldots, c_{K-1}$. Then, within each category, one fits a completely horizontal line as the dose-response "curve" relating exposure to the outcome within the category. For example, in categorical logistic regression, one simultaneously fits K category-specific models for the logit (log odds) of risk R:

$$\text{logit}(R|x \text{ in category } k) = \alpha_k^*, k = 1, \ldots, K, \quad (2)$$

which says that x has *no effect whatsoever within categories*, no matter how large its effect between categories!

To illustrate, suppose x is daily intake of ascorbic acid, R is mortality risk, and the internal boundaries for x are at 20, 50, and 100 mg per day, with the boundaries included in the lower category. The categorical dose-response model then says that there is no difference in risk between 0 and 20 mg per day but allows there to be an arbitrarily large jump in risk between 20 and 21 mg per day. This is biologically absurd, given that 0 mg per day represents a relatively rapidly fatal deficiency state, 20 mg per day does not, and the difference between 20 and 21 mg per day is biologically trivial. Although a categorical model can be viewed as providing estimates of average risk within categories, one should question the value of averaging risks that are known to be as disparate as those for 0 and 20 mg per day of ascorbic acid. Furthermore, under nonlinear models, the estimates of average risk provided by category-indicator regression can produce a biased impression of the exposure-specific dose-response curve.[15]

The preceding type of model, called a step function, is precisely what one is fitting when one breaks exposure into categories and then fits a model with $K - 1$ indicator variables $i_2, \ldots, i_K$, where $i_k = 1$ if x is in category k, 0 otherwise:

$$\text{logit}(R|x) = \alpha_1 + \alpha_2 i_2 + \cdots + \alpha_K i_K. \quad (3)$$

Here, $\alpha_1 = \alpha_1^*$ and $\alpha_k = \alpha_k^* - \alpha_1^*$ for $k > 1$. The results from such a model will not be misleading if risk changes little within categories. Unfortunately, selection of category boundaries based on percentiles in no way guarantees that this criterion will be met. In fact, use of percentiles virtually guarantees that the criterion will be violated if most subjects are concentrated within a narrow subrange of exposure and the exposure does have a large effect beyond that subrange.

The only way to ensure constancy of risk within categories is to use very narrow categories. This will often yield many more categories than the standard four or five—perhaps as many as 10, or even 20. If so, numbers may become so small within categories that the category-specific estimates are uselessly unstable, as in Table 1. Conventional recommendations (of four or five categories) seek to minimize variance by using few categories, but they unrealistically assume that boundaries will be set in an ideal fashion. If, however, the boundaries are not well chosen, bias will result. The variance-bias tension is especially severe in categorical dose-response modeling because of the unrealistic model that underlies the analysis.

Example

The *small dots* in Figure 1 trace the step function obtained by fitting the categorical model:

$$r_x = \exp(\alpha_1 + \alpha_2 i_2 + \alpha_3 i_3 + \alpha_4 i_4 + \alpha_5 i_5),$$

where i_2, i_3, i_4, i_5 are indicators for the categories 1981–1982, 1983–1984, 1985–1987, and ≥1988. The later

categories are broader because of the declining stability of the estimates over time. The visual impression provided by these estimates is less accurate than that from the other curves, failing to locate well the early climb and later decline in rates. The curve produced by connecting the midpoints instead of the ends of the categories (not shown) is a little better but is still not as plausible as the smooth curves. Narrower categories (not shown) were tried but failed to help, and instead produced erratically fluctuating steps. (Note that if one used categories based on the AIDS case percentiles from Table 1, the results would be disastrous: the first quartile extends through 1987, so that the fitted step function would represent the dramatic 1980–1987 trend by one constant rate!)

Linear Spline Regression

How can one avoid the absurdity, pitfalls, and tensions of category-indicator analysis for continuous variables? One simple solution is to allow the within-category lines to have nonzero slopes, so that the model will allow risk to vary *within* as well as *between* categories. Furthermore, we can fit these lines in such a way that there is no sudden jump in risk across category boundaries, so that fitted risk changes in a continuous manner within and across categories. The simplest method for doing so is called linear spline regression, which can be performed with conventional regression programs.

For logistic regression, the objective might be to simultaneously fit the K category-specific linear models:

$$\text{logit}(R|x \text{ in category } k) = \alpha_k^* + \beta_k^* x. \quad (4)$$

We should want these K models to fit together in a biologically sensible way, meaning that we want continuity (no sudden jumps) in risk across the category boundaries. This in turn requires any adjacent pair of category-specific models to predict the same risk at their common boundary j. For a logistic model, this means we must have:

$$\text{logit}(R|x = c_k) = \alpha_k^* + \beta_k^* c_k = \alpha_{k+1}^* + \beta_{k+1}^* c_k \quad (5)$$

for all k less than K. One way to force Eqs 4 and 5 to hold for all k less than K is to fit the following *linear spline model* to all of the data:

$$\text{logit}(R|x) = \alpha + \beta_1 x + \beta_2 s_2 + \cdots + \beta_K s_K, \quad (6)$$

where $s_k = 0$ if $x \leq c_k$, $x - c_k$ if $x > c_k$. s_k is sometimes called the positive part of $x - c_k$ and can also be defined as $s_k = \max(0, x - c_k)$. The parameters in Eq 6 are simple functions of the parameters in the K models in Eq 4: $\alpha = \alpha_1^*$ and $\beta_1 = \beta_1^*$, whereas for $k > 1$, $\beta_k = \beta_k^* - \beta_{k-1}^*$ is the change in the slope of dose-response in going from category $k - 1$ to category k. The graph of Eq 6 will look like a series of connected line segments.

Example

The *long dashes* in Figure 1 trace the linear spline obtained by fitting the model:

$$r_x = \exp(\alpha + \beta_1 \ln(x) + \beta_2 s_2 + \beta_3 s_3 + \beta_4 s_4)$$

where

$$s_2 = 0 \text{ if } x \leq 6,\ \ln(x) - \ln(6) \quad \text{if } x > 6,$$

$$s_3 = 0 \text{ if } x \leq 8,\ \ln(x) - \ln(8) \quad \text{if } x > 8,$$

and

$$s_4 = 0 \text{ if } x \leq 11,\ \ln(x) - \ln(11) \quad \text{if } x > 11$$

(x = 6, 8, 11 correspond to 1982, 1984, 1987). Apart from the artificially sharp peak in 1984, this model conveys essentially the same pattern as the fractional polynomial curves.

The general idea exemplified by Eqs 4–6 is to fit regression models simultaneously within each category, subject to constraints that maintain reasonable relations across the strata. These constraints also keep the analysis parsimonious. With K separate category-specific linear regressions, the total number of coefficients fit would have been $2K$ (K intercepts and K slopes). Nevertheless, the intercepts $\alpha_2, \ldots, \alpha_K$ are eliminated because of the continuity (no-jump) constraint, leaving only one intercept. The total number of parameters in Eq 6 is thus $K + 1$, only one more than the step function model for the same categories (Eq 3). Furthermore, unlike the step function (Eq 3), the linear spline (Eq 6) does not depend on risk being constant within categories for validity and thus can be used with fewer categories than required for valid use of the step function (Eq 3).

More General Spline Regression

Although a linear spline function is a dramatic improvement over a step function, it still does not have full biological plausibility because of the sharp bends (kinks) at the boundaries where the slope of the function abruptly changes. Also, linear-spline regression can suffer from instabilities and sensitivities to choice of category boundaries, although usually not as severely as category-indicator regression. To address these problems, we can create a curve with no sharp bends and a more smooth, plausible appearance simply by adding a quadratic term to each category-specific model, for example:

$$\text{logit}(R|x \text{ in category } k) = \alpha_k^* + \beta_k^* x + \gamma_k^* x^2. \quad (7)$$

As before, we want no jumps, which means adjacent category-specific models must agree at their common boundary:

$$\begin{aligned} \text{logit}(R|x = c_k) &= \alpha_k^* + \beta_k^* c_k + \gamma_k^* c_k^2 \\ &= \alpha_{k+1}^* + \beta_{k+1}^* c_k + \gamma_{k+1}^* c_k^2. \end{aligned} \quad (8)$$

To obtain a smooth appearance, we also want adjacent models to have the same slope (derivative) at their common boundary, which corresponds to requiring that:

$$\beta_k^* + 2\gamma_k^* c_k = \beta_{k+1}^* + 2\gamma_{k+1}^* c_k. \quad (9)$$

Simultaneously fitting the K category-specific quadratic models (Eq 7) subject to the continuity constraint (Eq 8) and the smoothness (slope) constraint (Eq 9) is equivalent to fitting the single *quadratic spline model*:

$$\text{logit}(R|x) = \alpha + \beta x + \gamma_1 x^2 + \gamma_2 s_2^2 + \cdots + \gamma_K s_K^2, \quad (10)$$

where $\alpha = \alpha_1^*$, $\beta = \beta_1^*$, $\gamma_1 = \gamma_1^*$, and, for $k > 1$, $\gamma_k = \gamma_k^* - \gamma_{k-1}^*$ is the change in the quadratic term (departure from linearity) of the dose-response function in going from category $k - 1$ to category k. This quadratic spline model (Eq 10) has only one more parameter than the linear spline model for the same categories (Eq 6). Furthermore, it would ordinarily require fewer categories for accuracy than the linear spline model, so that in practice no more parameters are needed than for the latter model. As with the linear spline, it can easily be fit with conventional regression programs.

Example

In addition to tracing the log-time fractional polynomial curve, the *solid curve* in Figure 1 coincides with the quadratic spline obtained by fitting the model:

$$r_x = \exp(\alpha + \beta_1 \ln(x) + \gamma_1 \ln(x)^2 + \gamma_2 s_2^2 + \gamma_3 s_3^2),$$

where

$$s_2 = 0 \text{ if } x \le 7,\ \ln(x) - \ln(7) \quad \text{if } x > 7,$$

and

$$s_3 = 0 \text{ if } x \le 10,\ \ln(x) - \ln(10) \quad \text{if } x > 10$$

(x = 7,10 correspond to 1983,1986).

Like higher-order polynomials, quadratic splines can suffer from odd behavior in open-ended tails of the exposure distribution. When this happens, we can further reduce the number of parameters and improve tail behavior by restricting the fitted curve to be linear in open-ended categories. To restrict the lower tail, one need only drop x^2 from the model. To restrict the upper tail, one drops s_K^2 from Model 10 and replaces the remaining s_k^2 by $s_k^2 - s_K^2$; one also replaces x^2 by $x^2 - s_K^2$ if the lower tail is not restricted. The quadratic spline with both tails restricted to be linear is:

$$\text{logit}(R|x) = \alpha + \beta x + \gamma_2(s_2^2 - s_K^2) + \cdots + \gamma_{K-1}(s_{K-1}^2 - s_K^2). \quad (11)$$

This model has only K coefficients including the intercept. In other words, it has exactly the same number of parameters as the crude step-function model (Eq 3), given that the same number of categories are used. Yet, unlike the step function, it can reproduce a wide variety of smooth curves.

Example

Because of its closeness to the other curves, Figure 1 omits the curve obtained by fitting the restricted quadratic spline:

$$r_x = \exp[\alpha + \beta_1 \ln(x) + \gamma_2(s_2^2 - s_5^2) + \gamma_3(s_3^2 - s_5^2) + \gamma_4(s_4^2 - s_5^2)]$$

where

$$s_2 = 0 \text{ if } x \le 5,\ \ln(x) - \ln(5) \quad \text{if } x > 5,$$

$$s_3 = 0 \text{ if } x \le 7,\ \ln(x) - \ln(7) \quad \text{if } x > 7,$$

$$s_4 = 0 \text{ if } x \le 9,\ \ln(x) - \ln(9) \quad \text{if } x > 9,$$

$$s_5 = 0 \text{ if } x \le 12,\ \ln(x) - \ln(12) \quad \text{if } x > 12$$

(these spline terms are based on the same categories as the earlier category-indicator model). This curve is very similar to the linear-spline curve, but rounded at the peak and at other category boundaries.

The type of restricted spline just described should not be confused with so-called natural splines,[5] in which the fitted curve is restricted to be linear below the smallest and above the largest observed value of x. These natural splines have the same number of parameters as unrestricted splines. They are obtained by treating $\min(x)$ and $\max(x)$ as additional category boundaries and then fitting a restricted spline to the expanded set of $K + 2$ categories. Within the range of the data, the resulting curve is identical to that produced by the unrestricted spline.

As the reader may have surmised, one may further extend the category-specific models and constraints. The form preferred by most statisticians is the *cubic spline model*,[16] which in its unrestricted form may be written:

$$\text{logit}(R|x) = \alpha + \beta x + \gamma x^2 + \delta_1 x^3 + \delta_2 s_2^3 + \cdots + \delta_K s_K^3, \quad (12)$$

This model may be derived by adding a cubic term $\delta_k^* x^3$ to the category-specific quadratic models (Eq 7) and then constraining the curves to be continuous and have equal slopes and second derivatives at the boundaries. The linear, quadratic, and cubic splines (Models 6 and 10–12) are all examples of *spline functions*, which are extensively used in the physical sciences and engineering but surprisingly rare in epidemiology. In the spline literature, the category boundaries $c_1, \ldots, c_{K-1}$ are called *knots* or join points, because they are the points at which the category-specific curves are tied together.[3–5,16] Natural cubic splines can be extended to produce a nonparametric smoother (called a cubic spline smoother) by placing a knot at each distinct exposure value and constraining the resulting saturated model with a penalty function.[3–5] It is also possible to constrain splines to produce only monotonic curves (that is, curves with no trend reversals).[17]

Discussion

Some external evidence regarding the true epidemic curve in the example is available, all of it indicating that the smooth curves are better estimates than the category-indicator step function. Backcalculations based

on much more extensive national data[8] indicate that a single sharp peak occurred around 1984–1985. More generally, both theoretical[5] and simulation[6] evidence indicates that smooth splines have better statistical properties than comparably parameterized step functions. Of course, one may conduct both a traditional step-function analysis and a spline analysis. The primary point of this paper is simply that some sort of smooth curve fitting is advisable when the study covariate is continuous and numbers do not permit the use of narrow categories.

All of the above methods can be applied to multiple covariates in a model. When applied to confounders, however, fractional-polynomial and spline regressions can produce more complete confounder control than step functions; this is because only the former control for confounder effects *within* strata as well as across strata. Generalized additive models[3,5] offer the same advantage, but within a given computing capacity, fractional polynomials and splines can be fit to larger datasets with more subjects and covariates, and can be fit with any regression software.

Unexposed Subjects

An issue that often arises when x is a ratio-scaled exposure (such as alcohol consumption) is whether to delete the unexposed during dose-response analysis. As explained elsewhere,[18] deletion of the unexposed (zero-exposed) is not always the best approach and is, in fact, an inadvisable waste of information if the unexposed and exposed are comparable with respect to factors that affect validity (such as uncontrolled confounders and selection factors). An advantage of highly flexible models (with more than a few exposure terms) over simpler models is that the overall curve will usually be less influenced by the unexposed than in simpler models, and hence the decision to retain or delete the unexposed will be less momentous. In nonparametric regression with ample data, smoothing neighborhoods can be made small, in which case the unexposed will exert little or no influence on the curve beyond their immediate low-exposure neighborhood. For situations in which the validity of retaining the unexposed is in question, a separate indicator variable for the unexposed category can be entered in the regression, which will eliminate direct influence of the unexposed on the curve. If this is done, the resulting fitted curve will not necessarily pass through the fitted rate at $x = 0$, reflecting the fact that the unexposed have been effectively eliminated from the curve-fitting process. See Greenland and Poole[18] for further discussion of this approach.

Choice of Splines

The improved smoothness of quadratic splines over linear splines leads me to prefer the former. In contrast, for epidemiologic purposes, there seem to be practical disadvantages and little if any advantage to using cubic splines instead of the quadratic splines. The primary disadvantage of cubic splines is that the cubic form of the category-specific models can produce very strange shapes in broad categories and in open-ended categories. With any spline, category boundaries can be adjusted to remove anomalies, whereas end-category anomalies can be prevented or removed by further constraining the end-category models to be linear.[16] Unfortunately, for cubic splines, the latter constraint requires that more complicated covariates than the s_k defined above be used in the regression. A more minor disadvantage of cubic splines is the poor interpretability of the coefficients, especially when end constraints are needed.

With enough well-chosen categories, cubic splines can closely approximate virtually any smooth curve.[4] This advantage seems of doubtful utility for epidemiologic analysis, however, because plausible trends and dose-response curves are usually very simple in form compared with many of the response functions found in engineering and the physical sciences. The primary gain from using cubic splines is that they yield very smooth curves. Nonetheless, I have not yet found epidemiologic data for which a gain from using cubic instead of quadratic splines is graphically noticeable. In the HIV example used here, a 5-parameter cubic spline model with one knot in the mid-1980s yields nearly the same curve as the fractional polynomial and quadratic spline curves in Figure 1.

There are certain advantages to using unrestricted splines (such as Models 10 and 12) over splines with end-category restrictions (such as Model 11). An unrestricted quadratic spline contains the ordinary quadratic regression model (the model with x and x^2 only) as a special case. Hence, the ordinary quadratic model can be checked against the more general unrestricted spline model (Eq 10) by testing the hypothesis that the spline coefficients are zero ($\gamma_2 = \cdots = \gamma_K = 0$ in Model 10). The restricted spline model (Eq 11) does not contain the quadratic model as a special case and so cannot be used in this way. Another drawback of restricted splines is that, perhaps counter to intuition, an end-category restriction can strongly affect the entire shape of the curve and enhance sensitivity of the overall shape to outliers. Nonetheless, restricted splines can be useful when linear end-category behavior is considered preferable to the nonmonotone end-category behavior that unrestricted splines can exhibit.

Choice of Categories and Terms

There are various schools of thought regarding choice of categories for splines. One school seeks automatic methods that optimize some statistical criterion, such as minimizing a goodness-of-fit statistic or the cross-validation sum of squared residuals.[3–5] Others prefer simple visual assessment of smoothness: Start with many categories, then reduce their number and adjust boundaries so that implausible blips, dips, and irregularities are eliminated. Another visual approach (suggested by a referee) is to use the curve from a smoother to suggest where cutpoints should be. All of these approaches have limitations. Automated methods (such as stepwise selection of

knots) can invalidate conventional tests and confidence intervals for trends,[16,19,20] whereas visual choice runs the risk of introducing subjective biases. Visual choice does allow one to use vague prior information about curve shape. Absent such information, some authors prefer to use percentile categories[16]; the latter can perform adequately with splines even when they perform poorly with category indicators.[6]

The problems just discussed are even more acute for ordinary category-indicator regression, because the latter is so sensitive to category choice. In particular, use of percentile categories can severely harm power and precision in category-indicator regression if the exposure effect is concentrated in a tail of the exposure distribution.[6] Unlike category indicators, splines make use of within-category risk variation and so can be less sensitive to category choice,[6] although, like category indicators, they can be sensitive to choice of tail categories when those categories are open ended.

Fractional polynomial regression avoids the problem of category choice but instead faces an analogous problem in choice of terms. As with category choice, mechanical algorithms for choice of terms invalidate conventional tests and can perform badly in small to modest samples, whereas visual choice runs the risk of introducing subjective biases of the analyst. These choice issues also arise in nonparametric regression, in which the analyst must visually select a value for the smoothing parameter, or else have it chosen by an algorithm.[3,5] In sum, every dose-response or trend analysis (from conventional categorical to advanced nonparametric) must choose the degree of smoothness or complexity in the fitted curve via choice of categories, model terms, or smoothing parameter. Regardless of the approach one uses, graphical inspection of the final fitted curve will greatly aid in determining whether the choices made yielded credible or surprising results.

Cutpoint Analysis and Thresholds

An issue of prominence in recent literature is that of choosing the proper cutpoint for dichotomous analysis of continuous exposures. Special concerns have been raised about "cutpoint bias," in which cutpoints are chosen to maximize significance or size of estimates.[21,22] Nonparametric curves and quadratic or cubic splines can largely finesse such issues by providing a single curve that simultaneously conveys rates or relative risks across the full range of exposure, without collapsing together disparate exposure levels. If there is a threshold for the exposure effect, it will be reflected by a steep portion of the smooth curve following a near-level portion. One should not, however, expect to see a single sharp (vertical) threshold point, because both exposure measurement error and individual variation in threshold will stretch out the threshold portion of the curve over some range of exposure.

Diagnostics

As with all regression, the methods discussed here (including conventional category-indicator regression, as well as the alternatives) need to be coupled with regression diagnostics (model checking) such as tests of fit, residual analysis, and influence analysis. In nonparametric regression, the effects of influential data points tend to be visually more dramatic but more localized than in conventional parametric regression[3]; similar comments apply to the flexible alternatives discussed here. Marked influences often show up in tails of the fitted curve, which can be strongly pulled toward outlying points. Diagnostics such as influence analysis help distinguish observed patterns that are resistant to modest changes in the data from those that are "driven" by just one or two unusual data points. Sensitivity of patterns to conventional model assumptions can also be explored by comparing conventional results to the results from flexible models.

Sample-Size Considerations

Fractional-polynomial and spline regression are *not* inherently large-sample techniques and can be applied with exact regression programs such as LogXact.[23] When applied in conjunction with large-sample (asymptotic) methods such as maximum-likelihood logistic regression, however, checks on sample size adequacy are advisable. Perhaps the easiest way of checking adequacy for maximum-likelihood logistic spline regression is to examine tabular cross-classifications based on the categories used to define the spline. By one rough criterion, if there are no product terms between exposure and other covariates, one should have at least five cases and five noncases in each category when applying maximum-likelihood methods. I am not aware of an equally simple sample-size criterion for maximum-likelihood estimation of fractional polynomials.

Confidence Limits

For clarity, confidence limits were omitted from Figure 1, but in practice, it can be helpful to include them, as in Figure 2. Confidence limits for points on the regression curve are an option in many software packages, and these options can be invoked when fitting fractional polynomials and splines.

When such options are not available, one may compute limits directly using large-sample analogues of standard formulas.[24] As an illustration, suppose we want 95% limits at the point x under the quadratic logistic spline model (Eq 10). Define the full parameter vector $\underline{\Theta}$ as:

$$\underline{\Theta} = (\Theta_1, \ldots, \Theta_{K+2})' = (\alpha, \beta, \gamma_1, \gamma_2, \ldots, \gamma_K)'$$

and the full covariate vector $\underline{z}$ as:

$$\underline{z} = (z_1, \ldots, z_{K+2})' = (1, x, x^2, s_2^2, \ldots, s_K^2)'.$$

Also, let $\hat{c}_{ij}$ be the estimated covariance of the parameter estimates $\hat{\Theta}_i$ and $\hat{\Theta}_j$ (the $\hat{c}_{ij}$ are available by requesting the covariance matrix output option from the regression

software). The fitted logit of risk $\hat{l}_x$ at x is then the dot product of $\underline{\hat{\Theta}}$ and $\underline{z}$,

$$\underline{\hat{\Theta}}'\underline{z} = \sum_i \hat{\Theta}_i z_i. \quad (13)$$

Approximate pointwise 95% limits for the risk at x are then given by:

$$\text{expit}(\underline{\hat{\Theta}}'\underline{z} \pm 1.96\hat{v}_x^{1/2}) \quad (14)$$

where $\text{expit}(u) = e^u/(1 + e^u)$ is the logistic transform, $\hat{v}_x$ is the estimated logit variance:

$$\hat{v}_x = \sum_i \sum_j \hat{c}_{ij} z_i z_j = \underline{z}'\hat{C}\underline{z}, \quad (15)$$

and $\hat{C}$ is the estimated covariance matrix for $\underline{\hat{\Theta}}$.

To estimate the ratio of odds at two different exposure levels with full covariate vectors $\underline{z}_1$ and $\underline{z}_0$, let $\underline{d} = \underline{z}_1 - \underline{z}_0$ be the vector of differences of the $\underline{z}_1$ and $\underline{z}_0$ components. The fitted log odds ratio is then

$$\underline{\hat{\Theta}}'\underline{d} = \sum_i \hat{\Theta}_i d_i \quad (16)$$

and approximate 95% limits for the odds ratio are given by:

$$\exp(\underline{\hat{\Theta}}'\underline{d} \pm 1.96\hat{v}_d^{1/2}) \quad (17)$$

where

$$\hat{v}_d = \sum_i \sum_j \hat{c}_{ij} d_i d_j = \underline{d}'\hat{C}\underline{d}. \quad (18)$$

For cohort data, approximate limits for the risk ratio can be obtained using the conditional method of Flanders and Rhodes,[25] whereas rate ratio limits can be obtained from an exponential-multiplicative rate model via Formulas 16–18.

The above formulas can be used when multiple covariates (exposure, confounders, and products among them) are present in the full covariate vector $\underline{z}$. The chief caution in their use is that they are large-sample approximations and can become inaccurate if the data are too limited. Computations are most easily performed using a matrix language such as GAUSS, MATLAB, SAS Proc Matrix, or S-Plus.

Approximate simultaneous 95% confidence limits can be constructed by replacing the normal 97.5th percentile of 1.96 by the square-root of the 97.5th percentile of a χ^2 distribution with degrees of freedom equal to the number of parameters ($K + 2$ for the unrestricted quadratic spline). One should note, however, that these simultaneous limits do not provide an accurate 95% confidence band for the true regression curve; see section 3.82 of Hastie and Tibshirani[3] for a discussion of this point and of bootstrap options for construction of confidence bands for the entire curve.

Conclusion

The present paper has argued that epidemiologic analyses of dose-response and trend, as well as methods for control of continuous confounders, should be expanded beyond simple categorical and linear (single-coefficient) approaches to include flexible curves that make use of intracategory information. Such expansion can be accomplished with little difficulty via fractional polynomial regression and spline regression. These methods can be especially valuable when important nonlinearities are anticipated, as in studies of health effects of alcohol, nutrients, and other life-style factors.

Acknowledgments

I would like to thank Philip Kass, Jennifer Kelsey, Stephan Lanes, Malcolm Maclure, Wendy McKelvey, and the referees for their helpful comments on the initial draft of this paper.

References

1. Maclure M, Greenland S. Tests for trend and dose response: misinterpretations and alternatives. Am J Epidemiol 1992;135:96–104.
2. Rothman KJ. Modern Epidemiology. Boston: Little, Brown, 1986.
3. Hastie T, Tibshirani R. Generalized Additive Models. New York: Chapman and Hall, 1990.
4. Härdle W. Applied Nonparametric Regression. New York: Cambridge, 1990.
5. Green PJ, Silverman BW. Generalized Linear Models and Nonparametric Regression. New York: Chapman and Hall, 1994.
6. Greenland S. Avoiding power loss associated with categorization and ordinal scores in dose-response and trend analysis. Epidemiology 1995;6:450–454.
7. Brookmeyer R, Gail MH. AIDS Epidemiology: A Quantitative Approach. New York: Oxford, 1994.
8. Bacchetti P, Segal MR, Jewell NP. Backcalculation of HIV infection rates. Stat Sci 1993;8:82–119.
9. Greenland S. Historical HIV incidence in regional subgroups: use of flexible discrete models with penalized splines based on prior curves. Stat Med 1996;15 (in press).
10. Alcabes P, Muñoz A, Vlahov D, Friedland GH. Incubation period of human immunodeficiency virus. Epidemiol Rev 1993;15:304–318.
11. Royston P, Altman DG. Regression using fractional polynomials of continuous covariates: parsimonious parametric modelling (with discussion). Appl Stat 1994;43:425–467.
12. Ross GJS. Discussion of Royston and Altman. Appl Stat 1994;43:456.
13. Hastie T, Tibshirani R. Discussion of Royston and Altman. Appl Stat 1994;43:460.
14. Doll R, Peto R. Cigarette smoking and lung cancer: reanalysis of the British doctor's data. J Epidemiol Community Health 1978;32:303–313.
15. Greenland S. Problems in the average-risk interpretation of categorical dose-response analyses. Epidemiology (in press).
16. Durrleman S, Simon R. Flexible regression models with cubic splines. Stat Med 1989;8:551–561.
17. Ramsay JG. Monotone regression splines in action (with discussion). Stat Sci 1988;3:425–461.
18. Greenland S, Poole C. Interpretation and analysis of differential exposure variability and zero-exposure categories for continuous exposures. Epidemiology 1995;6:326–328.
19. Freedman LS, Pee D. Return to a note on screening regression equations. Am Stat 1989;43:279–282.
20. Hurvich CM, Tsai C-L. The impact of model selection on inference in linear regression. Am Stat 1990;44:214–217.
21. Wartenberg D, Savitz DA. Evaluating exposure cutpoint bias in epidemiologic studies of electric and magnetic fields. Bioelectromagnetics 1993;14:237–245.
22. Schulgen G, Lausen B, Olsen JH, Schumacher M. Outcome-oriented cutpoints in analysis of quantitative exposures. Am J Epidemiol 1994;140:172–184.
23. LogXact. Release 1.0. Cambridge, MA: Cytel, 1994.

24. Graybill FA. Theory and application of the linear model. North Scituate, MA: Duxbury, 1976.
25. Flanders WD, Rhodes PH. Large sample confidence limits for regression standardized risks, risk ratios, and risk differences. J Chron Dis 1987;40:697–704.

Appendix

Assuming that migration and AIDS case reporting are nondifferential, the expectation μ_j for the observed AIDS case count y_j in epidemic year j is:

$$\mu_j = q_j \sum_{x=1}^{j} p_{jx} n_x r_x \qquad \text{(A1)}$$

where q_j is the probability that someone diagnosed with AIDS in year j is reported by the study time (1994), p_{jx} is the probability that someone contracting HIV in year x is diagnosed with AIDS in year j, n_x is the person-years at risk in year x, and r_x is the rate of non-IDU MSM HIV infections for the population in year x. In the examples, p_{jx} is taken from the stationary 3-parameter Weibull curve fit by Bacchetti *et al*[8] to the San Francisco hepatitis B cohort, with leveling of the hazard at its maximum. The denominators n_x are estimated from census data, whereas the q_j are estimated directly from the Los Angeles County AIDS surveillance data, which supplies both diagnosis and reporting dates.

Given the p_{jx}, n_x, q_j, and a model for r_x, the r_x are estimated by maximizing the Poisson loglikelihood $\Sigma_j[y_j \ln(\mu_j) - \mu_j]$ over the unknown model parameters.[8] The naive estimates in Table 1 were obtained by treating the log HIV rates $\alpha_x = \ln(r_x)$ as independent parameters. This corresponds to using a saturated log-linear model with an indicator for each year. The backcalculation equation (Eq A1) has no unique solution under this model, but a solution can be obtained by adding a penalty function to the loglikelihood.[8] The penalty function used for Table 1 is $\Sigma_x(\hat{\alpha}_x - \bar{\alpha})^2/t^2$, where $t^2 = 1.499 \times 10^7$ is the largest value that yielded a solution for Eq A1, and $\bar{\alpha}$ is the information-weighted average of the current log HIV rate estimates $\hat{\alpha}_x$. This penalty produces very mild shrinkage of the year-specific rates toward the weighted mean rate. Note that, counter to intuition, the naive estimates do not average to produce the categorical-model results in Figure 1. This is because the HIV rate estimates for each year are highly nonlinear functions of the AIDS incidence observed in all later years, and these functions differ across models as well as across years.