Smoothing in environmental epidemiology

Smoothing is a branch of regression analysis that allows for a predictor to impact the response as an arbitrary smooth function. It is also known as **nonparametric regression**. The purpose is to alleviate the restrictiveness of parametric functional relationships, such as that imposed by linearity. Environmental epidemiology has been a prime area for application of smoothing for two reasons. First, in part because of the traditional ties to respiratory epidemiology, it has dealt with continuous predictors almost from the beginning. In contrast, categorical variables have typified much of the rest of epidemiology until more recently. Continuous variables make the questions of the shape of the dose–response curve explicit. In addition, because of its inherent links to the standard setting, the question of dose–response has been crucial in environmental epidemiology. Searches for potential thresholds have been common for decades and the results of the epidemiology have often fed into risk assessments, where the shape of the dose–response curve has been a crucial question. At the same time, our theoretical and mechanistic understanding of the disease processes involved have rarely been advanced enough for us to specify the form of the dose–response a priori. All of this makes a flexible approach to determining that relationship particularly valuable.

Figure 1 provides a simple illustration of smoothing in environmental epidemiology. The response variable is daily mortality for the city of Milan, Italy over a 10-year period. This variable is plotted against temperature. The curve is the result of smoothing the data and indicates a highly nonlinear relationship between mean mortality and temperature. Such a relationship would be difficult to model parametrically. Such a curve is sometimes called a *smooth* of the data.

There is a large battery of techniques available for obtaining a smooth from a scatterplot. The most popular fall into the categories of kernel smoothing (e.g. [1, 4, 7] and [10]) and spline smoothing (e.g. [3, 5] and [9]). In environmental epidemiology it is common for the response variable to be discrete, particularly a binary or count variable. For such data the standard approach is to use likelihood-based models such as **logistic regression** and Poisson regression. The concepts of *local likelihood* and *penalized likelihood* allow for the extension to nonparametric functional relationships. The Milan mortality data are counts, so the smooth in Figure 1 represents a nonparametric extension of Poisson regression with a logarithmic link function.

Most smoothing techniques require the *amount of smoothing* to be specified. An intuitive measure of the amount of smoothing is the effective number of *degrees of freedom* (e.g. [6]). This extends the classical notion of number of parameters. In Figure 2 the curves correspond to 2, 7 and 35 degrees of freedom. Note that 2 degrees of freedom corresponds to a linear Poisson regression fit; 7 degrees of freedom corresponds to the smooth in Figure 1; while 35 degrees of freedom corresponds to a wiggly overfitting of the data. From Figure 2 it is apparent that the degrees of freedom has a profound effect on the result, and its choice should be treated with caution. Subsequently, there are a variety of data driven rules for choosing the degrees of freedom, many of which are based on classical model selection criteria such as Akaike's information criterion (AIC) and predicted residual sum of squares (PRESS). A computationally simpler approximation of the PRESS criterion, known as generalized cross validation (GCV) [2], is the most common criterion for selecting the degrees of freedom automatically.

Typically, environmental epidemiological studies involve assessment of the impact that a particular environmental exposure has on a health-related

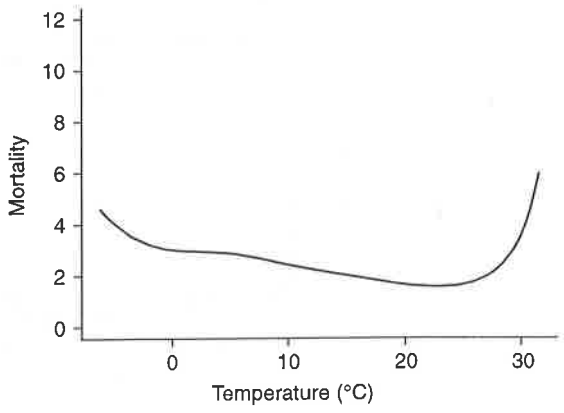

Figure 1 Smooth of Milan daily mortality data counts against temperature (°C)

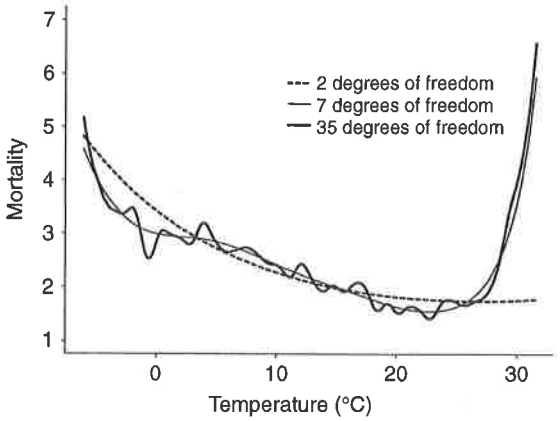

Figure 2 Three different smooths of the Milan daily mortality data counts against temperature (°C)

outcome. However, other factors that impact the outcome need to be taken into consideration, since they are likely to be confounded with the exposure. Consequently, *multiple predictor* regression models are much more common in environmental epidemiology. In the Milan mortality example, scientific interest centers on the impact of total suspended particles (TSP) on mortality, while temperature is merely a confounder for which one would like to control. Other measured variables with possible confounding effects are relative humidity and seasonality (day number). A parametric Poisson regression model with these variables is

$$\text{mortality}_t \sim \text{Poisson}[\exp(\alpha + \beta\ \text{TSP}_t + \gamma_1 \times \text{temperature}_t + \gamma_2\ \text{humidity}_t + \gamma_3 t)] \quad (1)$$

However, as illustrated in Figure 1, temperature has a nonlinear effect. Also, for 10 years of data, one would expect the effect of temperature to be approximately sinusoidal rather than linear. These considerations suggest the extension to

$$\text{mortality}_t \sim \text{Poisson}\{\exp[\alpha + \beta\ \text{TSP}_t + f_1(\text{temperature}_t) + f_2(\text{humidity}_t) + f_3(t)]\} \quad (2)$$

for smooth functions f_1, f_2 and f_3. This is an example of a **generalized additive model** [6]. It has become an important vehicle for analysis of data from environmental epidemiological studies [8], partly due to its availability in the commercial software package S-PLUS™.

Figure 3 shows the result of fitting model (2) to the Milan data, together with plus and minus twice approximate pointwise standard error estimates. The fit was obtained using the function gam( ) in

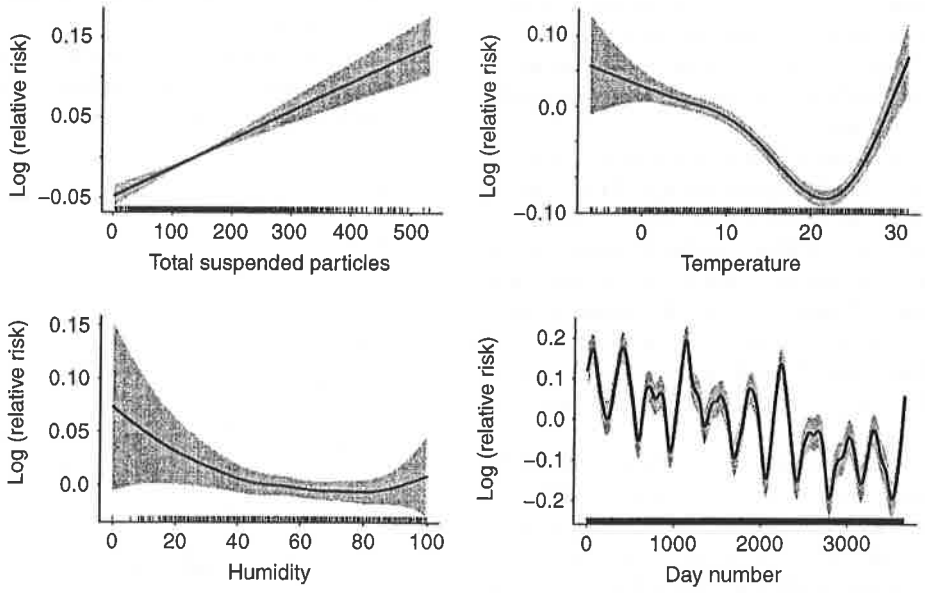

Figure 3 Components of the fit of (2) to the Milan mortality data. Each component is centered about its mean. The shaded regions correspond to approximate ±2× approximate pointwise standard error estimates

S-PLUS™. The smooths for temperature and humidity each involved 4 degrees of freedom, while that for day number involved 4 degrees of freedom per year. Data driven smoothing parameter selection is quite challenging for multiple predictor models such as this, so ad hoc choices like '4 degrees of freedom per year' are common. The function estimates are shown on the log scale and are centered about their average. They can be interpreted as the logarithm of the relative risk due to the value of the corresponding variable. TSP is seen to have a positive impact on mortality. Relative risk estimates and approximate confidence intervals can be computed from the model fit. However, for these data it has been argued [11] that more sophisticated models that account for mortality displacement provide even better relative risk estimates.

Smoothing has a great deal to offer in environmental epidemiology. It allows for the detection of nonlinear relationships and better controlling for covariates. In particular, generalized additive models allow for the handling of discrete responses and multiple predictors that typically arise in environmental epidemiology and, due to their availability in a commercial software package, are accessible to practitioners.

References

[1] Bowman, A.W. & Azzalini, A. (1997). *Applied Smoothing Techniques for Data Analysis*, Clarendon Press, Oxford.

[2] Craven, P. & Wahba, G. (1977). Smoothing noisy data with spline functions, *Numerische Mathematik* **31**, 377–403.

[3] Eubank, R.L. (1988). *Spline Smoothing and Nonparametric Regression*, Marcel Dekker, New York.

[4] Fan, J. & Gijbels, I. (1996). *Local Polynomial Modelling and Its Applications*, Chapman & Hall, London.

[5] Green, P.J. & Silverman, B.W. (1994). *Nonparametric Regression and Generalized Linear Models*, Chapman & Hall, London.

[6] Hastie, T.J. & Tibshirani, R. (1990). *Generalized Additive Models*, Chapman & Hall, London.

[7] Loader, C. (1999). *Local Regression and Likelihood*, Springer-Verlag, New York.

[8] Schwartz, J. (1994). The use of generalized additive models in epidemiology, in Proceedings of the International Biometrics Society Biannual Meeting, Hamilton, Canada, 1994, pp. 55–60.

[9] Wahba, G. (1990). *Spline Models for Observational Data*, SIAM, Philadelphia.

[10] Wand, M.P. & Jones, M.C. (1995). *Kernel Smoothing*, Chapman & Hall, London.

[11] Zanobetti, A., Wand, M.P., Schwartz, J. & Ryan, L.M. (2000). Generalized additive distributed lag models: quantifying mortality displacement, *Biostatistics* **1**, 279–292.

M.P. WAND & J. SCHWARTZ