Chapter 2 A Review of Parametric and Nonparametric Regression

We will spend a short time reviewing the basics of parametric and nonparametric regression. We will first set forth a conventional form for the model. This model is not exhaustive, but it does provide us with a basic form from which we shall work. For n paired observations the model is given by

$$Y_i = \boldsymbol{q}(\mathbf{x}_i) + \boldsymbol{e}_i \text{ for } i = 1 \text{ to } n.$$

Here, $\mathbf{x}_i$ can be a scalar or a vector depending on how many "independent" variables are involved, and $\mathbf{x}_i \in$ C = (a,b), where a and b (a < b) are real numbers. $\boldsymbol{q}(\mathbf{x}_i)$ is the mean response, Y_i is the observed scalar response with some unit measure, and the scalar $\boldsymbol{e}_i$ (having the same unit measure) is the stochastic driver in the equation, since it locates Y_i, and $\boldsymbol{e}_i \sim$ F($0\ unit, \boldsymbol{s}_i^2\ unit^2$), with $E(\boldsymbol{e}_i^4) = \boldsymbol{s}_i^4\ unit^4 < S\ unit^4,\ for\ S \in \Re, and\ i = 1,\ldots,n$. In subsequent chapters we will place various restrictions on this model.

We shall discuss parametric regression first, since this topic is necessary for the discussion of nonparametric regression.

Part 2a A Review of Parametric Regression

A predetermined form for $\boldsymbol{q}(\mathbf{x}_i;\boldsymbol{b})$ indicates the parameters that are to be estimated by the user. Here we add the vector of parameters $\boldsymbol{b}$ to indicate that $\boldsymbol{q}$ is dependent upon this vector as well. Here are some examples of $\boldsymbol{q}(\mathbf{x}_i;\boldsymbol{b})$.

a) $\boldsymbol{q}(\mathbf{x}_i;\boldsymbol{b}_0,\boldsymbol{b}_1) = \boldsymbol{b}_0 + \boldsymbol{b}_1 x_i$. A basic linear model that usually has Gaussian residuals.[].

b) $\boldsymbol{q}(\mathbf{x}_i;\boldsymbol{b}_0,\boldsymbol{b}_1) = e^{\boldsymbol{b}_0 + \boldsymbol{b}_1 x_i}$. A model that is often used for modeling a Poisson response variable.[].

c) $\boldsymbol{q}(\mathbf{x}_i, \boldsymbol{b}_0, \boldsymbol{b}_1) = \frac{1}{1+e^{b_0+b_1 x_i}}$. A model that is often used for modeling a Bernoulli response variable.[].

Once the parameters are targeted, there are generally two criteria by which users ascertain the solutions: generalized least squares and maximum likelihood estimation. Whether or not corresponding solutions are identical depends on the joint distribution of $\mathbf{e}$, the residual vector.

When solving by generalized least squares (GLS) the user is faced with minimizing $\sum \frac{(Y_i - \hat{\boldsymbol{q}}_i)^2}{\boldsymbol{s}_i^2}$ with respect to $\boldsymbol{b}$, which leads to solving a score equation of the form

$$\mathbf{s} = D^T V^{-1}(\boldsymbol{U} - \hat{\boldsymbol{q}}) = \mathbf{0}.$$

Here D is a matrix of derivatives of the mean response function with respect to $\boldsymbol{b}$, and V is a variance matrix (often diagonal). If the residuals are independently and identically distributed (i.i.d.), and we have a model similar to example a), then the score equation is easily solved in closed form with a solution like $(X^T X)^{-1} X^T \boldsymbol{U} = \hat{\boldsymbol{b}}$ where X is the matrix of stacked $\mathbf{x}_i$ vectors. Solutions of this type are called ordinary least squares (OLS) solutions. In other cases the score equation may be solved iteratively by methods such as iterated reweighted least squares (IRLS). IRLS is a computational method in which the user

1) Estimates $\hat{\boldsymbol{b}}_0$ (often by using OLS)
2) Obtains weights for V using $\hat{\boldsymbol{b}}_0$
3) Forms $\hat{\boldsymbol{b}}_1$ using V as in $\hat{\boldsymbol{b}}_1 = \hat{\boldsymbol{b}}_0 + (D^T V D)^{-1} D^T V^{-1}(\mathbf{Y} - \hat{\boldsymbol{q}}(\mathbf{x}_i; \hat{\boldsymbol{b}}_0))$
4) Replaces $\hat{\boldsymbol{b}}_0$ with $\hat{\boldsymbol{b}}_1$
5) Repeats steps 1-4 for a number of cycles C (or until step 4 becomes unnecessary).

This is a sound method and, used with today's 400-1000 MHz processors, makes solving most GLS problems quite easy (for more information on this see Carroll and Ruppert (1988)). The GLS concept has been extended to cases where (often by repeated measures) V is not a diagonal matrix because of correlation between observations.

A particular (to models having exponential error distributions) extension of the GLS concept is referred to as generalized estimating equations (GEE). GEE solves these cases by utilizing IRLS (or similar algorithms) at various substages of the solution process. GEE produces consistent (if not always efficient) estimates of $\hat{\mathbf{b}}$ and robust estimates of V. See Liang and Zeger (1986) for more details on this. Presently modifications (possible improvements) are being made to the GEE concept. Again, current computational capabilities allow problems of this nature to be handled nicely.

The second criterion for solutions, maximum likelihood estimation, is also quite popular. Nelder and Wedderburn (1972) set forth a general form for the cases in which the distribution of the response variable is a member of the exponential family of distributions.
They called this particular form the generalized linear model (GLM) and demonstrated that in these cases obtaining the maximum likelihood estimator (MLE) is equivalent to obtaining the GLS solution. In cases outside the GLM realm the user is forced to deal with the problem theoretically (to obtain the MLE) before proceeding to iterative schemes to obtain the solution. In the GLM cases, however, because we are really getting the GLS solution, IRLS can be used immediately after substituting a readily available form for $D^T V^{-1}$ in the score equation listed earlier.

We wish to point out that many functional forms can be modeled via parametric techniques. And even if a functional form is not available, a close approximation can usually be found. But in the latter instances one is compelled to query as to the reasonableness of the facsimile. For example, what if two or more very different functional forms are combined over a compact interval…can one parametric estimate approximate the mean response? This type of question led to the development of nonparametric regression which we shall briefly investigate next.