Appendix C

Fitting Quantiles by Combining Nonlinear and Linear Regression

The approach outlined below was motivated by a July 3, 1997, memorandum from Timothy Barry, Senior Analyst, Office of Policy and Re-Invention, to Jackie Moya, Environmental Engineer, Office of Research and Development.

Let F(x) be the cumulative distribution function (CDF) of a nonnegative continuous random variable X, that is, $F(x)=P[X \leq x]$=the probability of a value $\leq x$. Since X is continuous, F is continuous and strictly increasing, and its inverse FINV exists, so that F[FINV(*p*)]=*p* and FINV[F(x)]=x. Let $Y=a\,X^r$ be a power transform of X with both a and r strictly positive (a>0, r>0), and let $G(y)=P[Y \leq y]$ be the CDF of Y. Recall that y_p is the *p* th quantile of Y iff $G(y_p)=p$. Here iff denotes logical equivalence (“if and only if”).

Using basic algebra, set theory, and probability, it can be shown that

$$\log(y_p)=r \cdot \log[\mathrm{FINV}(p)]+\log(a). \qquad \text{(C.1)}$$

Hence, if F and its inverse FINV are known, and there are empirical quantiles y_p for several different values of *p*, then the power transform parameters a and r by linear regression of log (y_p) on log [FINV(p)] can be estimated. This is easily extended to cover distributions that are nonnegative and continuous except for a point mass M at zero. To see this, let H(y)=0 for y<0, H(y)=M+(1-M) G(y) for $y \geq 0$, and note that $H(y_p)=p$ iff $G(y_p)=(p-M)/(1-M)$. Hence for *p*>M, the *p*th quantile y_p for H is obtained by solving $G(y_p)=p_1$, where $p_1=(p-M)/(1-M)$. This leads to

$$\log(y_p)=r \cdot \log[\mathrm{FINV}(p_1)]+\log(a). \qquad \text{(C.2)}$$

These arguments suggest the following combined nonlinear/linear regression approach to fitting the five-parameter generalized F distribution with a point mass M at zero.

Let pmin be the smallest p for which a positive empirical quantile y_p exceeds zero. Then M should not exceed $p$min.

1. Perform an outer search on M, or simply use a grid of M values, such as

   M = 0, 0.1 $p$min, 0.2 $p$min, . . ., 0.9 $p$min.

2. For a given value of M, perform a two-dimensional search on the degrees-of-freedom parameters df_1, df_2 of the generalized F distribution.

3. Given M, df_1, and df_2, estimate a and r by solving the linear regression problem defined by Equation C.2.