Appendix C

Generalized least-squares model description and assumptions

Consider a region with n gaging stations as follows.
At each gaged site, a streamflow characteristic is estimated, such as the logarithm of the 50-year peak flow,

$$y_i = \Psi_i + \eta_i \ , \quad (1)$$

where ψ_i is the true (but unknown) log of the 50-year peak and η_i is a random error. If y_i is an unbiased estimate of ψ_i, then η_i (sometimes called time sampling error) has a mean of zero and a variance that is a function of how many years of data are available for the site and the standard deviation of water-year peaks. In addition, there are k basin characteristics, such as log of drainage area, that are measured with negligible error.

Assuming that (within the region defined by the basin characteristics at the n stations) ψ is approximately linearly related to the basin characteristics (x's), then the model formulation can be written as:

$$\Psi_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} \ldots + \beta_k x_{ki} + \varepsilon_i \quad (i=1,2,...,n;\ n>k) \ , \quad (2)$$

where ε_i is a model error assumed uncorrelated from observation to observation, with mean zero and constant variance, γ^2. Substituting into equation 1,

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_k x_{ki} + \eta_i + \varepsilon_i \ . \quad (3)$$

In matrix notation:

$$\mathbf{Y} = \mathbf{X}\beta + \upsilon \ , \quad (4)$$

where

$$\mathbf{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \ldots \\ y_n \end{bmatrix} \quad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{21} & \ldots & x_{k1} \\ 1 & x_{12} & x_{22} & \ldots & x_{k2} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 1 & x_{1n} & x_{2n} & \ldots & x_{kn} \end{bmatrix} \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \ldots \\ \beta_k \end{bmatrix} \quad \upsilon = \begin{bmatrix} \varepsilon_1 + \eta_1 \\ \varepsilon_2 + \eta_2 \\ \ldots \\ \varepsilon_n + \eta_n \end{bmatrix} , \quad (5)$$

where $E[\upsilon]=\mathbf{0}$, and $E[\upsilon\upsilon^T]=\Lambda$. Now the GLS estimator of β is:

$$\mathbf{b} = (\mathbf{X}^T\Lambda^{-1}\mathbf{X})^{-1}\mathbf{X}^T\Lambda^{-1}\mathbf{Y} \ . \quad (6)$$

The problem with this estimator is that Λ is unknown and must be estimated from the data. In OLS, Λ is estimated as $\sigma^2\mathbf{I}$, which would be a good estimate if all stations in that region had approximately the same lengths of record, or if the variance of η_i is small relative to the variance of ε_i at every station in the region.

Because this assumption may be hard to justify, a better estimate of Λ is attempted. Denote this estimated covariance matrix $\hat{\Lambda}$, and the GLS estimator, b, will be referred to as an Estimated Generalized Least Squares (EGLS) estimator.

EGLS Regression

An example illustrates how $\hat{\Lambda}$ is estimated. Suppose that y_i is the log of the 50-year peak estimated from m_i years of record and that the water-year peaks follow a log-Pearson Type III (LPIII) distribution at all sites. Further, to minimize notation, assume that the skew coefficient at all sites is zero. The elements of $\hat{\Lambda}$ would be given by:

$$\lambda_{ij} = \begin{cases} \gamma^2 + \dfrac{\sigma^2{}_i(1+0.5K^2)}{m_i} for(i=j) \\ or \\ \dfrac{\rho_{ij}\sigma_i\sigma_j m_{ij}(1+0.5K^2)}{m_i m_j} for(i \neq j) \end{cases} . \qquad (7)$$

In this equation, K (LPIII standard deviate for zero skewness and 50-year recurrence interval), m_i (record length at station i), m_j (record length at station j), and m_{ij} (concurrent record length for stations i and j) are known, but σ_i (standard deviation of water-year peaks at station i), ρ_{ij} (cross correlation of water-year peaks at stations i and j), and γ^2 (variance of model error) must be estimated from the data. Furthermore, we cannot use s_i (the sample estimate of σ_i) as an estimate of σ_i without introducing bias, and the use of r_{ij} (sample cross correlations) for ρ_{ij} often causes numerical problems. Therefore, we estimate σ_i and ρ_{ij} as follows.
The standard deviation of water-year peaks, σ_i, is estimated from a regional regression of the form:

$$ln(s_i) = b_0 + b_1 x_{1i} + b_2 x_{2i} + \ldots + b_k x_{ki} \qquad (8)$$

By estimating the standard deviations, s_i, that enter into equation 7 with equation 8, we are assured that the rows of the Λ matrix are not correlated with the observed dependent variable **Y**. This quality is necessary for the estimates of β to be unbiased.

The cross correlation coefficient, ρ_{ij}, is estimated by developing an empirical relation between sample cross correlations, r_{ij}, and distance between stations of the form:

$$r_{ij} = \Theta^{\left[\frac{d_{ij}}{\alpha d_{ij}+1}\right]} . \qquad (9)$$

Estimating the cross correlations in this manner assures us that the matrix Λ will be positive definite. Figure 1 below shows a smooth curve with Θ=.9812 and α=.00412 based on data from Illinois. This curve was developed by running the GLSNET program that will be described later.

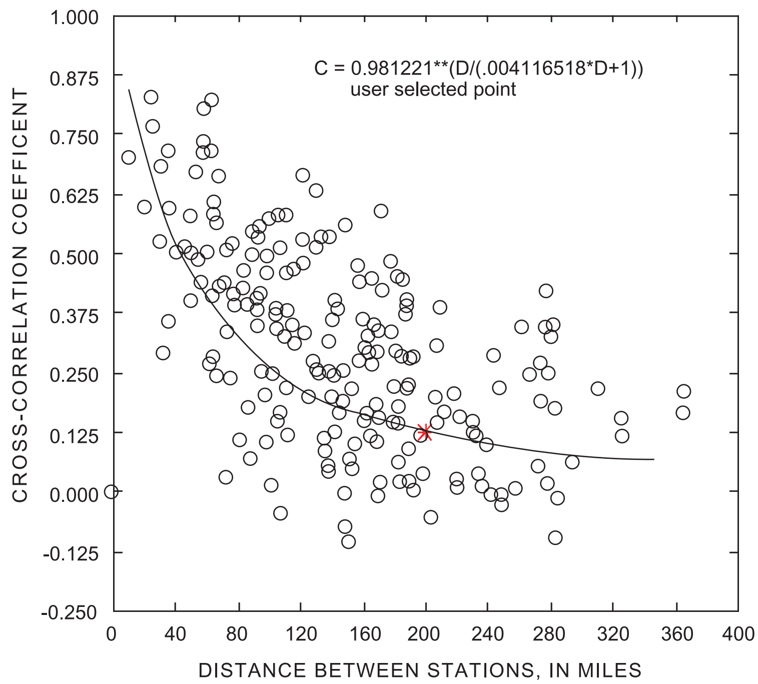

Figure 1. Relation between cross correlation and distance.

Now the only parameters left to find in the EGLS model are the regression coefficients, **b**, and variance of the model error, γ^2. The model error variance, γ^2, and regression coefficients, **b**, are found by iteratively searching for the best non-negative solution to the equation:

$$E\{(\mathbf{Y} - \mathbf{X}\beta)^T \Lambda^{-1} (\mathbf{Y} - \mathbf{X}\beta)\} = n - k - 1 \quad . \tag{10}$$

The GLSNET/AIDE package leads one through the development of equations 8 and 9 in preparation for the estimation of the GLS regression coefficients.

Reporting results and errors

The predicted response at ungaged site k with basin characteristics $\mathbf{x}_k = (1, x_{k,1}, x_{k,2}, ..., x_{k,p})$ is:

$$\hat{y}_k = \mathbf{x}_k \mathbf{b}. \tag{11}$$

The standard error of the prediction in OLS regression is:

$$S(\hat{y}_k) = \{\sigma^2[1 + \mathbf{x}_k(\mathbf{X'X})^{-1}\mathbf{x'}_k]\}^{0.5} \quad (12)$$

In GLS regression, the standard error of prediction is:

$$S(\hat{y}_k) = \sqrt{\hat{\gamma}^2 + \mathbf{x}_k \mathbf{X}' \hat{\Lambda}^{-1} \mathbf{X}^{-1} \mathbf{x}'_k} \ . \quad (13)$$

The $S(\hat{y}_k)$ is a function of **x** and the computed standard error of a prediction in percent will also be a function of **x.**

Standard Errors in Percent

When a standard error or average prediction error in log units follows a normal distribution, the error may be expressed in percent of the predicted value in cubic feet per second (ft^3/s). Denote σ as the standard error in log (base 10) units, S_{cfs} as the standard error in ft^3/s, and $E(q|\mathbf{x}_k)$ as the predicted value of q, in ft^3/s, given $\mathbf{x}_k$, and $\mathbf{x}_k = (1, x_{k,1}, x_{k,2}, ..., x_{k,p})$ is a vector of basin characteristics. The standard error in percent, $S_{percent}$ is given by:

$$S_{percent} = 100\frac{S_{cfs}}{E(q|x_k)} = 100\sqrt{(e^{5.302\sigma^2} - 1)} \quad (14)$$

(Aitcheson and Brown, 1957).

Sometimes it is said in OLS that two-thirds of the points lie within one standard error of estimate of the regression function. This is true for the log unit standard error of estimate, σ, but it generally is not correct for $S_{percent}$. This is true because the errors in log space are symmetrically distributed under the assumption of normality of the log errors, but the errors in ft^3/s are skewed. You can, however, calculate +percent and -percent errors with the following formulas:

$$S_{plus} = 100(10^{\sigma} - 1) \text{ ; and} \quad (15)$$

$$S_{minus} = 100(10^{-\sigma} - 1) \ . \quad (16)$$

The three formulas above apply not only to the standard error of estimate for an regression, but they also apply to the standard error of the model, $\hat{\gamma}$, in GLS regression, the average prediction error, and standard error of a prediction in both OLS and GLS.

Average prediction error (APE)

One overall measure of how good the regression model is for prediction is the average prediction error (Hardison, 1971), where the average is taken over prediction sites with X variables identical to the observed data. This measure assumes the observed data have been collected at a representative set of sites in the region. It is computed as:

$$APE = \left(\sum_{i=1}^{n} \frac{\hat{\gamma}^2_i}{n} + \sum_{i=1}^{n} \frac{x_i (X' \hat{\Lambda}^{-1} X)^{-1} x'_i}{n} \right)^{1/2} . \qquad (17)$$

The first term in the brackets on the right side of equation 17 represents an estimate of the average squared model error for the n sites and the second term inside the brackets is an estimate of the average squared error due to estimating true model parameters from a sample of data.

Prediction interval

Users of the regression model are probably more interested in a measure of error in a particular prediction rather than an average prediction. A good measure of the error of a particular prediction is the confidence interval of a prediction, or prediction interval. Let x_0 represent the usual row vector of basin characteristics at a prediction site. As usual x_0 is augmented by a 1 as the first element. The predicted value is $\hat{y}_0 = x_0 b$. A 100(1-α) prediction interval would be:

$$\hat{y}_0 - T \le y_0 \le \hat{y}_0 + T , \qquad (18)$$

where

$$T = t_{\frac{\alpha}{2}, n-p'} \sqrt{(\hat{\gamma}^2_0 + x_0 (X' \hat{\Lambda}^{-1} X)^{-1} x'_0)} , \qquad (19)$$

where $t_{\alpha/2, n-p'}$ is the critical value from a t-distribution for n-p′ degrees of freedom.
If a log transform had been made so that $y_0 = \log_{10}(q_0)$, then the prediction interval would be:

$$10^{\hat{y}_0 - T} \le q_0 \le 10^{\hat{y}_0 + T} . \qquad (20)$$

SELECTED REFERENCES

Acreman, M.C., and Wiltshire, S.E., 1987, Identification of regions for regional flood frequency analysis [abs,], EOS, v. 68, no. 44.

Aitchison, J., and Brown, J.A.C., 1957, *The Log-Normal Distribution*: London, Cambridge University Press.

Hardison, C.H., 1971, Prediction error of regression estimates of streamflow characteristics at ungaged sites: U.S. Geological Survey Professional Paper 750-C, p. 228-236.