CROSS-VALIDATION ERROR STATISTIC AS A TOOL FOR MODEL EVALUATION

Randal T. Wortman, Hydraulic Engineer, U. S. Army Corps of Engineers, Portland, OR, randal.t.wortman@usace.army.mil

Abstract: Mathematical model development typically proceeds through a model calibration phase followed by a model validation phase. Cross-validation techniques, in the form of split-sample analysis, have long been employed as a tool in evaluating hydrologic simulation models, and with the advent of modern computer systems, have been extended in a variety of ways and applied to both model calibration and model validation. Over the past thirty years statisticians have extensively investigated advanced validation techniques and explored their use in the development and evaluation of both linear and non-linear models. The cross-validation standard error (CVSE) statistic is presented herein as a statistic particularly applicable to linear regression models. The CVSE provides a robust statistic superior to the traditional standard error statistic for evaluating the capability of a model to provide a prediction for a sample set not utilized in the model calibration. A case study deriving new water supply forecast equations for Libby, Montana demonstrates the use of the CVSE in model calibration and selection.

INTRODUCTION

Mathematical Modeling Process Modeling of water resource systems typically involve either physical models or mathematical models, with the mathematical models falling into either the category of conceptual simulation models (see Singh and Woolhiser 2002 for a recent survey of watershed models) or statistical models. Optimization techniques may utilize components of either simulation or statistical models, or both. The mathematical modeling process itself may be represented by four phases: Conceptualization, formulation, calibration, and validation (McCuen 2003; McCuen 2005), with a potential feedback-control process of the validation procedure to return one to the formulation step to reconsider an alternate model type or form with subsequent recalibration of the model parameters (Salas et al 1980). The model conceptualization phase considers the general and broad considerations for the purpose of the model, the complexity of the model, the available data, along with the experience and biases of the investigator (often not recognized!). The model formulation phase considers various alternatives to the model composition, model type, and model parameters. Model calibration seeks to determine values for the model parameters, while model validation evaluates the performance of the model through some type of error analysis. Most water resource models, outside of the statistical models, perform inadequate error analysis (Singh and Woolhiser 2002; McCuen 2005).

STATISTICS FOR MODEL SELECTION AND EVALUATION

Cross-validation Validation in hydrologic simulation models has traditionally utilized the split-sample methodology as a tool to assist in evaluating the model calibration and to judge the "fit" of the simulation model (WMO 1975; WMO 1986; McCuen 2003; McCuen 2005). The split-sample approach divides the data set into two subsets, typically sequential halves, using one data set for calibration (or "training") and the second data set for prediction and error analysis using the model calibrated to the first data set. This approach was all that was reasonable when considering a complex simulation model and extensive subjective calibration of the model

parameters. Invariably the model would perform more poorly on the data set that was withheld for calibration than with the data set used for calibration. It should be readily apparent that the split-sample technique has significant limitations, including (1) that model calibration is performed on a significantly reduced subset of the data, and is thus guaranteed to be suboptimal over the full data set, i.e. the calibration statistics will always be overly optimistic, and (2) that there is little or no flexibility in the choice of the split-sample subsets, especially if the data are serially correlated and must be grouped in a time-ordered sequence. Prior to the 1970's split-sample validation was similarly employed in fitting and evaluating regression models.

The availability of virtually unlimited computer resources in the 1970's provided the catalyst for extending the concept of split-sample validation and the development of a wide variety of cross-validation techniques (Stone 1974; Efron 1982). The original split-sample validation methodology was constructed under the premise that both the training and validation sets were of equal size, N/2, where N is the sample size. There being no inherent restriction on the sizes of the training and validation sets, researchers expanded cross-validation to include a training set of any size k, with $1 < k < N/2$, with the corresponding validation set then being of size N-k. The computation of a validation (goodness-of-fit) statistic from the prediction errors results from using the calibrated model to provide predictions on the validation data set. The ultimate extension of this approach leads to the jackknife or leave-one-out cross-validation (LOO-CV) procedures, both based on systematically removing one observation from the dataset, calibrating the model to the remaining N-1 data points, then using this model calibrated on the reduced data set to calculate a statistic of interest (Jackknife) or predict the withheld observation (LOO-CV). Efron (1982) and Stone (ibid) discuss the details of the differences between Jackknife and LOO-CV. Other cross-validation techniques explore the option of random sample splits. The bootstrap method invented by Efron in 1977, is another resampling approach, in that it draws subsamples of a set size from the original data set *with replacement*, and the statistics of the resulting models studied. The jackknife, bootstrap, and various cross-validation strategies have been well discussed in the statistical and technical literature (Stone ibid; Efron 1982; Diaconis and Efron 1983; Wilmott et al 1985). The statistical and technical literature on cross-validation and resampling techniques generally, with few exceptions, restricts the discussion to objective statistical models, (mostly linear models, but including non-linear and artificial neural network models), and not subjectively calibrated hydrologic simulation models. The subsequent discussion herein will likewise limit its scope to the statistical model, with a particular focus on the linear regression model.

Regression Model Development Cross-validation has been particularly attractive as a tool for use in regression analysis, particularly in application to the model development problem of selecting the subset of "best" regression variables (Stone 1974; Hocking 1976; Snee 1977; Wu 1986). Although many criteria functions for subset selection are discussed in the literature (e.g. Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), Mallow's C_p, MSEP) , Miller (1984), in a thorough discussion of the regression subset selection problem, states that the theory behind their derivations is not valid when model selection and estimation are from the same data. Hocking issues a similar warning that "the basic idea behind partitioning the data is that the data used for analysis should not be used for assessment", a sentiment echoed by Meko (2005). Miller also reminds us of Hocking's caution that standard least-squares theory is not applicable when the model (variable selection) has not been determined *a priori*, and

emphasizes the problem that large biases in the regression coefficients may result when the same data have been used for both model selection (choosing the predictor variables) and parameter estimation (calibrating the regression coefficients). Stone and Hocking advocate cross-validation techniques as the appropriate integration of the parameter estimation phase and model selection phase.

Although there are a variety of purposes for which one could be developing a linear regression model, is must be recognized that the appropriate criterion for model selection and evaluation should be related to the intended use of the regression model. A modeler focused on the estimation of the model parameters should consider a different criterion than one focused on prediction or extrapolation. In the traditional approach to comparing two regression models, used for prediction, and calibrated over the full data set, i.e. without cross-validation, it is nearly universal to employ a criterion that penalizes against "over-fitting", the use of too many variables in proportion to the number of available observations. Traditional textbook approaches would consider the use of either the adjusted R-squared or the standard error criterion, as they both contain a term to adjust for the "degrees of freedom" lost due to fitting additional parameters. Intermediate techniques would consider such criterion as the AIC, BIC, and Mallow's C_p, although we have already established that use of these criterion should be limited to circumstances that do not involve use of the same data for parameter fitting and model selection.

<u>Cross-validation statistics</u> Allen (1971a; 1971b) introduced the Predicted Residual Sum of Squares (PRESS) statistic to aid in the variable selection process. PRESS is computed from the Leave-One-Out residuals in a fashion analogous to the commonplace sum-of-squares-residuals:

$$PRESS = \sum_{i=1}^{n} \left(y_i - \hat{y}_{(i)} \right)^2 \tag{1}$$

where i is the index to the observations in data set of size n, y_i is the ith observation of the dependent variable, and $\hat{y}_{(i)}$ is the prediction derived from the regression equation fit to the data set with the ith observation removed. Hocking (1976) confirms that PRESS has in intuitive appeal if the objective is prediction.

In practice it would be extremely laborious to recalibrate a mathematical model for each training set with the ith observation removed (thus the use of split-sample validation for hydrologic simulation models). Fortunately, the matrix algebra that provides the least-squares solution to the regression model also provides an efficient solution to the computation of the PRESS statistic, since

$$PRESS = \sum_{i=1}^{n} \left(y_i - \hat{y}_{(i)} \right)^2 = \sum_{i=1}^{n} \frac{\left(y_i - \hat{y}_i \right)^2}{\left(1 - h_i \right)^2} \tag{2}$$

where $\hat{y}_i$ is the full model prediction for observation i, h_i is the leverage for the ith observation, and the h_i are the diagonal terms of the Hat (projection) matrix $X^T(X^TX)^{-1}X$.

A close derivative to PRESS is the cross-validation mean-square-error (CV MSE)

$$CV\ \ MSE = \frac{PRESS}{n} \tag{3}$$

which is also frequently used as a criterion for evaluating a regression model (Allen 1971a; Efron and Tibshirani 1995; Tibshirani and Knight 1999).

Recall that the Standard Error statistic (SE), traditionally used to evaluate regression models, is defined as the root-mean-squared-error (RMSE) from the full model prediction errors:

$$SE = \sqrt{\frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{n-p}} \tag{4}$$

where p is the number of independent variables fit in the regression equation, including a y-intercept term. The analogous cross-validation standard error is computed as

$$CVSE = \sqrt{\frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_{(i)}\right)^2}{n-p}} = \sqrt{\frac{PRESS}{n-p}} \tag{5}$$

The criterion introduced by Stone (1974, page 124), MSE_q, is equivalent to the term under the radical sign in the CVSE criterion. Although the PRESS statistic by itself is frequently employed as the criterion to evaluate the fit of a regression model (Hocking 1976; Stone 1974; Snee 1977; Carpio and Hermosilla, 2002; Prairie el at 2005), the PRESS statistic does not utilize a penalty term to make a degrees-of-freedom adjustment to prevent over-fitting the regression model with too many variables. Plutowski (1996) notes that the "LOO-CV can be too conservative, in the sense that it may select a model of excessive size".

The cross-validation standard error statistic (CVSE) is the ultimate extension of the split-sample methodology and with today's computer resources the CVSE is easily calculated using the PRESS statistic (available through many statistical software packages) using (5), or deriving the Hat matrix using matrix algebra and using (2). The CVSE is then calculated for each candidate regression model and used as a parsimonious criterion to evaluate the goodness-of-fit of each model.

Hocking (1976) and Miller (1984), among others, have noted that the variables selection process is confounded when there is a high degree of multicollinearity (intercorrelation) in the predictor variables. If the predictor variables are orthogonal (i.e. uncorrelated), "the effects of the individual variables are clear and the problems of estimation and subset selection are elementary" (Hocking ibid). Principal components regression provides a tool that will remove

the effects of intercorrelation among the predictor variables (Haan 1972; Diaconis and Efron 1983, Carpio and Hermosilla 2002).

APPLICATION

Water Supply Forecasting Model Combining principal components regression with the cross-validation standard error criterion for model selection provides a valuable tool for the development and evaluation of linear regression models in general and specifically for the development of statistical water supply forecasting models (Wortman 1989; Garen 1992; Wortman 2005). Individual monthly water supply forecast equations were developed for predicting the April-July inflow into Lake Koocanusa, above Libby Dam, Montana. The NRCS REG software program developed by Garen at the National Water and Climatic Center, National Resources Conservation Service, was used to develop the equations from the observed data available on the first of each month during the winter and spring forecast season. For each first-of-month equation NRCS REG selects the predictor variables from a large pool of variables, including a climatic variable (SOI index), monthly precipitation station values, and when available, first-of-month snow water equivalent (SWE) station values. In most cases the number of variables (up to 80) far exceeded the number of available observations (less than 50). In developing a list of "best" models, the NRCS REG procedure first evaluates a base list of all the one variable regression models and then examines each of the two-variable principal components regression models. Only the significant principal components are retained as predictor variables, with the non-significant components being discarded. The pool of all candidate principal component regression models are then compared and evaluated based on the CVSE statistic. Up to 30 "best" models are retained from this step and the process continues for another round of adding variables, continuing until no further principal components regression models can be developed that enter into the pool of best models. After REG has developed the pool of 30 best models for each forecast date, candidate models are compared and a model for each forecast date is subjectively chosen by the modeler to provide the greatest level of month-to-month consistency in predictor variables (i.e. find the top models that contain a consistent set of stations).

The original forecasting equations for the inflow to Lake Koocanusa were developed in the 1970s during project construction, and updated in the 1980s to reflect a longer period of record and to adjust for precipitation and snow stations whose operation had been discontinued. These original equations were developed using standard regression model techniques based on the standard error criterion. Figure 1 displays the results of new monthly forecasting equations developed using the NRCS REG model, based on principal components regression and the CVSE criterion, as previously described. The error statistics for both model sets are based on the full set of data through 2004; the standard error statistic is presented for the standard regression equations and the CVSE statistic is presented for the new principal component regression equations. Use of a climatic variable, the Southern Oscillation Index (SOI), in the new equations allowed for the development of forecast equations two months earlier than the original models, prior to the availability of winter snow and precipitation observations. A 20% to 30% improvement (reduction in prediction errors) is noted for all but the last late-spring forecast date.

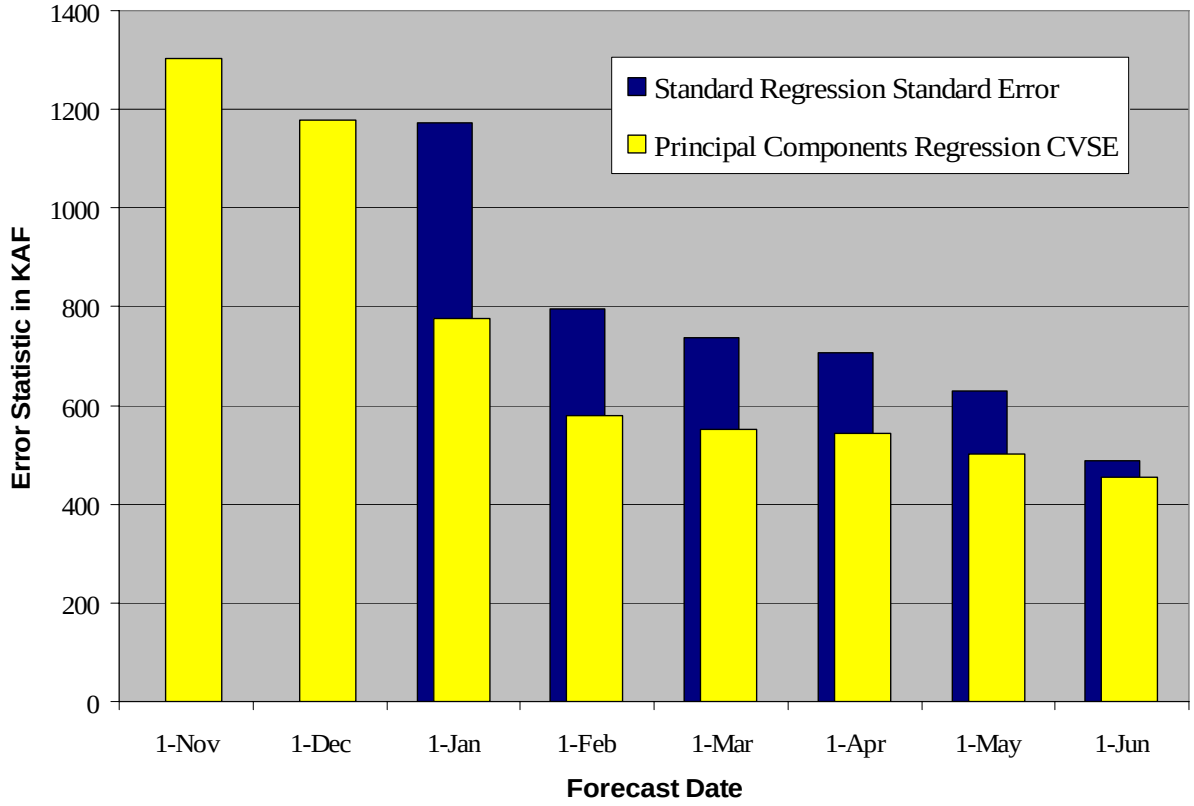

Figure 1 Libby water supply forecasting model statistics

The use of principal components regression models, based on equations selected using the CVSE criterion are seen to offer significant improvement over standard regression procedures. These procedures have been widely adopted for use in developing water supply forecasting equations by agencies across the Pacific Northwest, including the following agencies: Natural Resources Conservation Service; Northwest Division, U.S. Army Corps of Engineers; Northwest River Forecast Center, NOAA; Bonneville Power Administration; and British Columbia Hydropower.

REFERENCES

Allen, D.M. (1971a). "Mean square error of prediction as a criterion for selecting variables", Technometrics 13, pp 469-475.

Allen, D.M. (1971b). "The prediction sum of squares as a criterion for selecting prediction variables", Technical Report No. 23, Dept. of Statistics, University of Kentucky.

Carpio, K.J.E., and Hermosilla, A.Y. (2002). "On multicollinearity and artificial neural networks", Compexity International, Monash University, AU. (draft manuscript: http://www.complexity.org.au/ci/draft/draft/hermos01/hermos01s.pdf)

Diaconis, P., and Efron, B. (1983). "Computer intensive methods in statistics", Scientific American, Vol. 248, pp 116-130.

Efron, B. (1982). The Jackknife, the Bootstrap, and Other Resampling Plans. Society for Industrial and Applied Mathematics, Philadelphia, PA.

Efron, B., and Tibshirani, R. (1995). “Cross-validation and the bootstrap: Estimating the error rate of a prediction rule”, Technical Report (TR-477), Dept. of Statistics, Stanford University, Palo Alto, CA.

Garen, D. (1992). “Improved techniques in regression-based streamflow volume forecasting”, Journal of Water Resources Planning and Management, ASCE, Vol. 118, No. 6, pp 654-670.

Haan, C.T. (1977). Statistical Methods in Hydrology. Iowa State University Press, Ames, IA.

Hocking, R.R. (1976). The analysis and selection of variables in linear regression”, Biometrics 32, pp 1-49.

McCuen, R. H. (2003). Modeling Hydrologic Change. Lewis Publishers, Boca Raton, FL

McCuen, R.H. (2005). “Accuracy Assessment of Peak Discharge Models”, Journal of Hydrologic Engineering, ASCE, Vol. 10, No. 1, pp 16-22.

Meko, D. M. (2005). “Validating the regression model”, Applied Time Series Analysis (GEOS 585A), Lesson 12 notes, University of Arizona, AZ (http://www.ltrr.arizona.edu/~dmeo/geos585e.html)

Miller, A.J. (1984). “Selection of subsets of regression variables”, Journal of the Royal Statistical Society, Series A, 147 pp 389-425.

Plutowski, M.E. (1996). “Survey: Cross-validation in theory and in practice”, unpublished manuscript (http://www.emotivate.com/CvSurvey.doc). David Sarnoff Research Center, Princeton, N.J.

Prairie, J.R., Rajagopalan, B., Fulp, T.J., and Zagona, E.A. (2005). “Statistical nonparametric model for natural salt estimation”, Journal of Environmental Engineering, ASCE, Vol. 131, No. 1, pp 130-138.

Salas, J.D., Delleur, J.W., Yevjevich, V., Lane, W.L. (1980). Applied Modeling of Hydrologic Time Series. Water Resources Publications, Littleton, CO.

Singh, V.P., and Woolhiser, D.A. (2002). “Mathematical modeling of watershed hydrology”, Journal of Hydrologic Engineering, Vol. 7, No. 4, pp 270-292.

Snee, R.D. (1977). “Validation of Regression Models: Methods and Examples”, Technometrics, Vol. 19, No. 4, pp 415-428.

Stone, M. (1974). “Cross-validatory choice and assessment of statistical predictions (with discussion)”, Journal of the Royal Statistical Society, Series B, 36 pp 111-147.

Tibshirani, R., and Knight, K. (1999). “The covariance inflation criterion for adaptive model selection”, Journal of the Royal Statistical Society, Series B, 61 pp 529-546.

Willmott, C.J., Ackleson, S.G., Davis, R.E., Feddema, J.J., Klink, K.M., Legates, D.R., O’Donnell, J., and Rowe, C.M. (1985). Journal of Geophysical Research, Vol. 90, No. C5, pp 8995-9005.

WMO (1975). “Intercomparison of conceptual models used in operational hydrological forecasting”, Operational Hydrology Paper No. 429, World Meteorological Organization, Geneva.

WMO (1986). “Intercomparison of models of snowmelt runoff”, Operational Hydrology Paper 646, World Meteorological Organization, Geneva.

Wortman, R.T. (1989). “Statistical forecast model for Libby Basin, Montana” Western Snow Conference, Fort Collins, CO.

Wortman, R.T. (2005) “Improved water supply forecasts for the Kootenay Basin”, Tri-Service Infrastructure Systems Conference, NDIA, St. Louis, MO.

Wu, C.F.J. (1986). “Jackknife, bootstrap and other resampling methods in regression analysis (with discussion)”, The Annals of Statistics, Vol. 14, No. 4, pp 1261-1350.