

12 Validating the Regression Model

Regression R-squared, even if adjusted for loss of degrees of freedom due to the number of predictors in the model, can give a misleading, overly optimistic view of accuracy of prediction when the model is applied outside the calibration period. Application outside the calibration period is the rule rather than the exception in dendroclimatology. The calibration-period statistics are typically biased because the model is “tuned” for maximum agreement in the calibration period. Sometimes too large a pool of potential predictors is used in automated procedures to select final predictors. Another possible problem is that the calibration period itself may be anomalous in terms of the relationships between the variables: modeled relationships may hold up for some periods of time but not for others. It is advisable therefore to “validate” the regression model by testing the model on data not used to fit the model. Several approaches to validation are available. Among these are cross-validation and split-sample validation. In cross-validation, a series of regression models is fit, each time deleting a different observation from the calibration set and using the model to predict the predictand for the deleted observation. The merged series of predictions for deleted observations is then checked for accuracy against the observed data. In split-sample calibration, the model is fit to some portion of the data (say, the second half), and accuracy is measured on the predictions for the other half of the data. The calibration and validation periods are then exchanged and the process repeated. In any regression problem it is also important to keep in mind that modeled relationships may not be valid for periods when the predictors are outside their ranges for the calibration period: the multivariate distribution of the predictors for some observations outside the calibration period may have no analog in the calibration period. The distinction of predictions as extrapolations versus interpolations is useful in flagging such occurrences.

12.1 Validation

Validation strategies. Several alternative strategies for validation are available, and some may be better than others depending on the data and purpose of analysis. Three different ways of validating are:

- 1) *Compare predictions made by the model with records of some proxy for the predictand.*
 - a) Calibrate on the entire available length of the overlap of predictors and predictand
 - b) Apply the model to predict outside the calibration period
 - c) Compare the predictions outside the calibration period with observations of some proxy for the predictand
 - d) Pro: uses all available data for calibration; a long calibration time series generally gives a more stable model
 - e) Con: validation semi-qualitative
- 2) *Validate the model with a time series segment of the predictand withheld from calibration.*
 - a) Calibrate on just a part of the period of overlap of predictors and predictand
 - b) Apply model to generate predictions for the data withheld from calibration
 - c) Compare the predicted and observed predictand for the period withheld from calibration
 - d) Use the model from (a) for final prediction
 - e) Pro: model validated is same model used for final prediction
 - f) Con: requires long time series of predictand if some data are to be sacrificed to validation

3) *Cross-validation*

- a) Divide period of overlap of predictors and predictand into two or more subsets
- b) At each step in cross-validation, omit a subset and calibrate on remaining data
- c) Use sub-period model from (b) to predict for the omitted sub-period
- d) Repeat steps (b) and (c), each time omitting a different subset from calibration
- e) Aggregate the predictions from the various steps (d) into a single “predicted” series
- f) Compare the aggregated predictions with the observed predictand
- g) Re-calibrate using full-length predictand data for final model for prediction model
- h) Pro: optimum use of relatively short predictand series
- i) Con: model validated not exactly same as final model for predictions

These methods might be referred to as “leave n out.” At one extreme n is half the sample length. This type of cross-validation is *split-sample validation*. In split-sample validation the model is calibrated on some fraction (say first half) of the data and validated on the other fraction (Snee 1977). Then the calibration/validation periods are exchanged and the calibration and validation done again. The final prediction model is then calibrated using the full available length of predictand data.

At the other extreme is “leave-one-out” cross-validation, which is equivalent to “cross-validation” as described by Michaelsen (1987) and to the predicted-residual-sum-of-squares procedure, or *PRESS procedure* as described by Weisberg (1985). Say the full available period for calibration is length of n years. Models are repeatedly estimated using data sets of $n - 1$ years, each time omitting a different observation from calibration and using the estimated model to generate a predicted value of the predictand for the deleted observation. At the end of this procedure, a time series of n predictions assembled from the deleted observations is compared with the observed predictand to compute validation statistics of model accuracy and error.

Validation statistics. Validation statistics measure the error or accuracy of the prediction for the validation period. The statistics can generally be expressed as functions of just a few simple terms, or build blocks. We begin by defining the building blocks.

Validation errors. All of the statistics described here are computed as some function of the validation error, which is the difference of the observed and predicted values:

$$\hat{e}_{(i)} = y_i - \hat{y}_{(i)} \quad (1)$$

where y_i and $\hat{y}_{(i)}$ are the observed and predicted values of the predictand in year i , and the notation (i) indicates that data for year i were not used in fitting the model that generated the prediction $\hat{y}_{(i)}$.

Sum of squares of errors, validation (SSE_v). SSE_v is the sum of the squared differences of the observed and predicted values:

$$\text{SSE}_v = \sum_{i=1}^{n_v} (\hat{e}_{(i)})^2 \quad (2)$$

where the summation is over the n_v years making up the validation period.

Mean squared error of validation (MSE_v). MSE_v is the average squared error for the validation data, or the sum-of-squares of errors divided by the length of validation period:

$$\text{MSE}_v = \frac{\text{SSE}_v}{n_v} \quad (3)$$

The closer the predictions to the actual data, the smaller the MSE_v . Recall that the calibration period equivalent of MSE_v is the residual mean square, MSE, which was listed in the ANOVA table in the previous notes.

Root mean squared error of validation (RMSE_v). The RMSE_v is a measure of the average size of the prediction error for the validation period, and is computed as the square root of the mean squared error of validation:

$$\text{RMSE}_v = \sqrt{\text{MSE}_v} = \left[\frac{\sum_{i=1}^{n_v} (\hat{e}_{(i)})^2}{n_v} \right]^{1/2} \quad (4)$$

RMSE_v has the advantage over MSE_v of being in the original units of the predictand. The calibration equivalent of RMSE_v is the standard error of the estimate, s_e . RMSE_v will generally be greater than s_e because s_e reflects the “tuning” of the model to the data in the calibration period. The difference between RMSE_v and s_e is a practical measure of the importance of this tuning of the model. If the difference is small, the model is said to be validated, or to verify well. What is meant by “small” is somewhat subjective. For example, in a reconstruction of annual precipitation for agriculture, a difference of 0.2 inches between RMSE_v and s_e might be judged inconsequential if an error of 0.2 inches makes no appreciable difference to the health of the crop.

Reduction of error (RE). RE measures the *skill* of a regression model, defined as its accuracy relative to a prediction based on no knowledge. In defining RE, it is first necessary to specify the “no-knowledge” prediction. Frequently this prediction is simply the calibration-period mean of the predictand, \bar{y}_c . In other words, with no other knowledge about the predictand other than its calibration-period data, it makes sense simply to substitute the calibration-period mean of the predictand as the predicted value for any year outside the calibration period. Following Fritts et al. (1990), RE is then given by

$$\text{RE} = 1 - \frac{\text{SSE}_v}{\text{SSE}_{null}} \quad (5)$$

where SSE_v is the sum of squares of validation errors as defined previously and

$$\text{SSE}_{null} = \sum_{i=1}^{n_v} (y_i - \bar{y}_c)^2 \quad (6)$$

RE has a possible range of $-\infty$ to 1. An RE of 1 indicates perfect prediction for the validation period, and can be achieved only if all the residuals are zero (i.e., $\text{SSE}_v = 0$). On the other hand, the minimum possible value of RE cannot be specified, as RE can be negative and arbitrarily large if SSE_v is much greater than SSE_{null} . As a rule of thumb, a positive RE is accepted as evidence of some skill of prediction. In contrast, if $\text{RE} \leq 0$, the prediction or reconstruction is deemed to have no skill.

Recall that the equation for computing the regression R^2 is

$$R^2 = 1 - \frac{SSE}{SST} \quad (7)$$

The similarity in form of the equations for R^2 and RE (equations (7) and (5)) suggests that RE be used as a validation equivalent of regression R^2 , and that a value of RE “close to” the value of R^2 be considered as evidence of validation. The rationale for this comparison is easily seen for leave-one-out cross-validation. In both equations, the numerator is a sum of squares of prediction errors, and the denominator is the sum of squares of departures of the observed values of the predictand from a constant. For leave-1-out cross-validation the constant is equal to the calibration-period mean for both (5) and (7). This is so because for leave-1-out cross-validation the aggregate “validation” period is essentially the same as the calibration period: each year of the calibration period is individually and separately used as a validation period in the iterative cross-validation, and the aggregate of these validation years is the “validation period.”

PRESS Statistic. PRESS is an acronym for “predicted residual sum of squares” (Weisberg 1985, p. 217). The PRESS procedure is equivalent to “leave-1-out” cross-validation, as described previously. The PRESS statistic is defined as

$$PRESS = \sum_{i=1}^n \hat{e}_{(i)}^2 \quad (8)$$

where $\hat{e}_{(i)}$ is the residual for observation i computed as the difference between the observed value of the predictand and the prediction from a regression model calibrated on the set of $n - 1$ observations from which observation i was excluded. The PRESS statistic is therefore identical to the sum of squares or residuals for validation, SSE_v , defined in equation (2) which was described previously.

12.2 Cross-validation stopping rule

As described earlier, the automated entry of predictors into the regression equation runs the risk of over-fitting, as R^2 is guaranteed to increase with each predictor entering the model. The adjusted R^2 is one alternative criterion to identify when to halt entry of predictors (e.g., Meko et al. 1980), but the adjusted R^2 has two major drawbacks. First, the theory behind adjusted R^2 assumes the predictors are independent, while in practice the predictors are often inter-correlated. Consequently, entry of an additional predictor does not necessarily mean the loss of one degree of freedom for estimation of the model. Second, the adjusted R^2 does not address the problem of selecting the predictors from a pool – sometimes a large pool – of potential predictors. If the pool of potential predictors is large, R^2 can be seriously biased (high), and the bias will not be accounted for by the adjustment for number of variables in the model used by the algorithm for adjusted R^2 (Rencher and Pun 1980).

An alternative method of guarding against over-fitting the regression model is to use cross-validation as a guide for stopping the entry of additional predictors (Wilks 1995). By evaluating the performance of the model on data withheld from calibration **at every step of the stepwise procedure**, the level of complexity (number of predictors) above which the model is over-fit can be estimated. Graphs of change in calibration and validation accuracy statistics as a function of step in forward stepwise entry of predictors can be used as a guide for cutting off entry of predictors into the model. For example, in a graph of RMSE_v against step in a model run out to many steps (e.g., 10 steps), the step at which the RMSE_v is minimized (or approximately so) can be set as the final step for the model. The same result would be achieved from a plot of RE against step, except that the maximum in RE indicates the “best” model.

Extending the entry of predictors beyond the indicated steps amounts to “over-fitting” the model. Over-fitting refers to the tuning of the model to noise rather than to any real relationship between the variables. In the extreme, over-fitting is illustrated by a model whose number of predictors equals the number of observations for calibration: the model will explain 100% of the variance of the predictand even if the predictor data is merely random noise.

12.3 Prediction (Reconstruction)

Predictions are the values of the predictand obtained when the prediction equation

$$\hat{y}_i = \hat{b}_0 + \hat{b}_1 x_{i,1} + \hat{b}_2 x_{i,2} + \dots + \hat{b}_K x_{i,K} \quad (9)$$

is applied outside the period used to fit the model. For example, in dendroclimatology, the tree-ring indices (x 's) for the long-term record are substituted into (9) to get estimates of past climate. The prediction is called a *reconstruction* in this case because the estimates are extended into the past rather than the future. Once the regression model has been estimated, the generation of the reconstruction is a trivial mathematical step, but important assumptions are made in taking the step.

First, the multivariate relationship between predictand and predictors in the calibration period is assumed to have applied in the past. This assumption might be violated for many possible reasons. For example, in a tree-ring reconstruction, the climate for the calibration period may have been much different than for the earlier period, such that a threshold of response was exceeded in the earlier period. Or the quality of the tree-ring data might have decreased back in time because of a drop-off in sample size (number of trees) in the chronologies. Many other data-dependent scenarios could be envisioned that would invalidate the application of the regression data to reconstruct past climate. For time series in general, regardless of the physical system, it is important to statistically check the ability of the model to predict outside its calibration period or to *validate* the model, as described in the preceding section.

12.4 Error bars for predictions

A reconstruction should always be accompanied by some estimate of its uncertainty. The uncertainty is frequently summarized by error bars on a time series plot of the reconstruction. Error bars can be derived by different methods:

1) **Standard error of the estimate**, s_e . Recall that s_e is computed as the square root of the mean squared residuals, MSE. Following Wilks (1995, p. 176), the Gaussian assumption leads to an expected 95% confidence interval of roughly

$$CI_1 \approx \hat{y}_i \pm 2s_e \quad (10)$$

Confidence bands by this method are the same width for all reconstructed values. The $\pm 2s_e$ rule of thumb is often a good approximation to the 95% confidence interval, especially if the sample size for calibration is large (Wilks 1995, p. 176). But because of uncertainty in the sample mean of the predictand and in the estimates of the regression coefficients, the prediction variance for data not used to fit the model is somewhat larger than indicated by MSE, and is not the same for all predicted values. This consideration gives rise to a slightly more precise estimate of prediction error called the standard error of prediction (see next section). Also note that the “2”

in (10) is a rounded-off value of the 0.975 probability point on the cdf of the normal distribution (1.96 rounded to 2). Strictly speaking, the appropriate multiplier (2 in the example) should come from a “t” distribution with $n-K-1$ degrees of freedom, where n is the sample size for calibration and K is the number of predictors in the model (Weisberg 1985). The distinction will be important only for small sample sizes or models for which the number of predictors is dangerously close to the number of observations for calibration.

2) **Standard error of prediction, $s_{\hat{y}}$** This improved estimate of prediction error is proportional to s_e , but in addition takes into account the uncertainty in the estimated mean of the predictand and the in the estimates of the regression coefficients. Because of these additional factors, the prediction error is larger when the predictor data are far from their calibration-period means, and vice versa. For simple linear regression, the standard error of the estimate and standard error of prediction are related as follows:

$$s_{\hat{y}} = s_e \left[1 + \frac{1}{n} + \frac{(x_* - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right]^{1/2} \quad (11)$$

where s_e is the standard error of the estimate, n is the sample size (number of years) for the calibration period, x_i is the value of the predictor in year i , \bar{x} is the calibration-period mean of the predictor, x_* is the value of the predictor in the reconstruction year in question, $s_{\hat{y}}$ is the standard error of prediction for that year, and the summation in the denominator is over the n years of the calibration period.

Note first that $s_{\hat{y}} > s_e$, and that the difference has contributions from the two right-hand terms inside the square root. The first source of difference is uncertainty due to the fact that the estimate of the predictand will not equal its expectation; this contribution can be made smaller by increasing the sample size. The second source is the uncertainty in the estimates of the regression constant and coefficient. The consequence of this term is that the prediction error is greater when the predictor is farther from its calibration-period mean. This feature is what causes the “flaring out” of the prediction intervals in a plot of the predicted values against the predictor values. More on this topic can be found in Weisberg (1985, p. 22, 229) and Wilks (1995, p. 176).

The equation for the standard error of prediction in MLR is more complicated than given by (11), which applies to simple linear regression, as $s_{\hat{y}}$ depends on the variances and covariances of the estimated regression coefficients. The equation for $s_{\hat{y}}$ in the multivariate case is best expressed in matrix terms. The MLR model, following Weisberg (1985, p. 229) can be written in vector-matrix form as

$$\begin{aligned} \mathbf{Y} &= \mathbf{X}\boldsymbol{\beta} + \mathbf{e} \\ \mathbf{Y} &= \text{column vector of predictand for calibration period} \\ \mathbf{X} &= \text{matrix of predictors for calibration period} \\ \boldsymbol{\beta} &= \text{row vector of regression coefficients, with regression constant first} \\ \mathbf{e} &= \text{column vector of regression residuals} \end{aligned} \quad (12)$$

If the model is used to predict data outside the calibration period, and the predictor data for some year to be predicted is given by the row vector \mathbf{x}_* , the predicted value for that year is given by

$$\tilde{y}_* = \mathbf{x}_*^T \hat{\boldsymbol{\beta}} \quad (13)$$

Assuming the linear model is correct, the estimate is an unbiased point estimate of the predictand for the year in question, the variance of the prediction is

$$\begin{aligned} s_{\tilde{y}}^2 &= \text{varpred}(\tilde{y}_* | \mathbf{x}_*) = \sigma^2 \left(1 + \mathbf{x}_*^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_* \right) \\ &= \sigma^2 (1 + h_*) \end{aligned} \quad (14)$$

where σ^2 is generally estimated as the residual mean square, or s_e^2 . The estimated standard error of prediction is the square root of the above conditional variance

$$s_{\tilde{y}} = \text{sepred}(\tilde{y}_* | \mathbf{x}_*) = \sigma \sqrt{1 + h_*} \quad (15)$$

2) **Root-mean-squared error of validation, RMSE_v.** Another possible way of assigning a confidence interval to predictions is use the validation error as an estimate of the expected error of reconstruction or prediction. For example, with leave-1-out cross-validation, or the PRESS procedure, $\text{RMSE}_v = \sqrt{\text{PRESS}/n_v}$ is the validation equivalent of the standard error of prediction, and if normality is assumed, can be used in the same way as described for s_e or $s_{\tilde{y}}$ to place confidence bands at a desired significance level around the predictions. For example, an approximate 95% confidence interval is $\hat{y}_i \pm 2 \text{RMSE}_v$. Weisberg (1985, p. 230) recommends this approach as a “sensible estimate of average prediction error.”

12.5 Interpolation vs extrapolation

A regression equation is estimated on a data set called the *construction data set*, or calibration data set. For this construction set the predictors have a defined range. For example, in regressing annual precipitation on tree-ring indices, perhaps the tree-ring data for the calibration period are range between 0.4 and 1.8 -- or 40% to 180% of “normal” growth. The relationship between the predictand and predictors expressed by the regression equation applies strictly only when the predictors are “similar” to their values in the calibration period. If the form of the regression equation is not known a priori, then we have no information on the relationship outside the observed range for the predictor in the calibration period. When the model is applied to generate predictions outside the calibration period, an important question is how “different” can the predictor data be from its values in the calibration period before the predictions are considered invalid. When the predictors are acceptably similar to their values in the calibration period, the predictions are called *interpolations*. Otherwise, the predictions are called *extrapolations*. Extrapolations in a dendroclimatic reconstruction model present a dilemma: the most interesting observations are often extrapolations, while the regression model is strictly valid only for interpolations. A compromise to simply tossing out extrapolations is to flag them in the reconstruction.

Algorithm for identifying extrapolations. Extrapolations are identified by locating the predictor data for any given prediction year relative to the multivariate “cloud” of the predictor data for the calibration period. Identification is trivial for the simple linear regression, as any

prediction year for which the predictor is outside its range for the calibration period can be regarded as an extrapolation. For MLR, any prediction for which the predictor data fall outside the predictor “cloud” for the calibration period can be regarded as an extrapolation.

In MLR, extrapolations can be defined more specifically as observations that fall outside an ellipsoid that encloses the predictor values for the calibration period. This is an ellipsoid in p -dimensional space, where p is then number of predictors. For the simple case of one predictor, the “ellipsoid” is one-dimensional, and any values of x outside the range of x for the calibration period would lead to an extrapolation. For MLR with two variables, the ellipsoid is an ellipse in the space defined by the axes for variables x_1 and x_2 .

For the general case of an MLR regression with p predictors and an calibration period of n years, Weisberg (1985, p. 236) suggests an ellipsoid defined by constant values of the diagonal of the “hat” matrix \mathbf{H} , defined in matrix algebra as

$$\mathbf{H} = \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \quad (16)$$

where \mathbf{X} is the n by $(p + 1)$ time series matrix of predictors, with ones in the first column to allow for the constant of regression. For each prediction year with predictor values in the vector \mathbf{x}_* , the scalar quantity

$$h_* = \mathbf{x}_*^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_* \quad (17)$$

is computed, and any prediction for which

$$h_* > h_{\max} \quad (18)$$

where h_{\max} is the largest h_{ii} in the diagonal of the hat matrix \mathbf{H} , is regarded as an extrapolation.

12.6 References

- Fritts, H.C., Guiot, J., and Gordon, G.A., 1990, Verification, in Cook, E.R., and Kairiukstis, L.A., eds., *Methods of Dendrochronology, Applications in the Environmental Sciences*: Kluwer Academic Publishers, p. 178-185.
- Fritts, H.C., 1976, *Tree rings and climate*: London, Academic Press, 567 p.
- Meko, D.M., Stockton, C.W., and Boggess, W.R., 1980, A tree-ring reconstruction of drought in southern California: *Water Resources Bulletin*, v. 16, no. 4, p. 594-600.
- Michaelsen, J., 1987, Cross-validation in statistical climate forecast models, *J. of Climate and Applied Meteorology* 26, 1589-1600.
- Rencher, A.C., and Pun, Fu Ceayong, 1980, Inflation of R^2 in best subset regression, *Technometrics* 22 (1), 49-53.
- Snee, R.D., 1977, Validation of regression models: Methods and examples, *Technometrics* 19, 415-428.
- Weisberg, S., 1985, *Applied Linear Regression*, 2nd ed., John Wiley, New York, 324 pp.
- Wilks, D.S., 1995, *Statistical methods in the atmospheric sciences*: Academic Press, 467 p.