

ASSIGNMENT 12. VALIDATING THE REGRESSION MODEL

1. Run geosal1.m, selecting part 2 (assignment 12) from the opening menu. This selection specifies multiple linear regression **with** cross-validation.

The difference from assignment 11 is that this time (1) you will use cross-validation as a “stopping rule” to decide how many predictors are included in the model, and (2) you will generate a long-term reconstruction and related statistics. Besides the figure windows produced in assignment 11, we have two additional windows. One summarizes the change in validation and calibration statistics with each step in the model. The other is a time series plot of the reconstructed flow, with two different measures of reconstruction uncertainty and with flagged “extrapolations.”

Several of the options menus are the same as for the Part 1 run, and are not described again here. In your choice of predictors, again choose some combination of predictor time series and lags that will give you at a predictor pool of at least 6 potential predictors.

At the menu “Maximum number of steps to run model”, accept the default the first time through. This will let all potential predictors enter. Accept the default calibration period and let the model run. This may take a while. At the end of this run, print out and save the “validation” window (Figure 5). You might also want to print out the “equation” window (Figure 2) because you will want to specify the same pool of potential predictors in your next model run.

Use the printout of the change of validation statistics with model size (number of steps) to decide on a cutoff level of steps to include for your final reconstruction model.

2. (Caption to Fig 5) Explain your decision on how many steps to allow the regression to proceed.

Now re-run assign11.m, again in Part-2 mode. Use the same chronologies, same prewhitening option, and same lagging option you used before, but this time answer the “maximum number of steps to run mode” question with the number of steps identified in #1 above. You will now of course have a fresh set of six figure windows, some of which are referred to below.

3. (Caption to Fig 1). Do the calibration statistics indicate that the regression equation has statistically significant explanatory power? Does the validation exercise indicate that the model has any skill when applied to independent data? What is the reconstruction uncertainty in terms of root-mean-square error of validation ($RMSE_v$), and what are the units of that statistic for your particular data?
4. (Caption to Fig 6). From the confidence bands and the range of the predictand variable, does the reconstruction uncertainty seem large or small in a practical sense? The plotted confidence bands is at $\pm 1 RMSE_v$ from the calibration-period mean of the predictand. Given the interpretation of $RMSE_v$ in terms of probability (notes, p 7), do any of your reconstructed values appear to differ from the calibration period mean with 95% certainty? If yes, point one out on the plot. If not, speculate why.

Running geosa11.m for assignment 12

1. >geosa11
2. Message box: message introducing geosa11.m; click OK to remove message and move on
3. Menu: select “Part 2 – Assignment 12
4. Question box: Wide-line graphics? This has to do with widening some of the lines and enlarging some of the markers to make the display more visible in class when projected with powerpoint. For your on-screen and printed versions, answer NO to this question.
5. Respond to input dialog with the name of your data file; click OK
6. Menu: select data set the predictand is to be selected from
7. Menu: select the predictand series (a “Y” appears beside it), then click “satisfied”

You can undo this choice by clicking again on the variable. You must select one and only one predictand series before being “satisfied.”

8. Menu: select whether or not to log10 transform the predictand series

Log10 is the only transform built into the program. For other transformations, you would need to transform the data before building your initial storage files. Log transform is invalid if any data values are zero or negative.

9. Menu: select the pool of potential predictors – up to 8 series allowed; then click “satisfied”
10. Menu: select whether or not to prewhiten the predictors
If “yes”, it will take a few seconds for the prewhitening to be completed
11. Menu: select whether to include lagged predictors

If you decide to include lagged predictors, two additional menus will ask how many positive lags and how many negative lags. Including lags will increase the size of the predictor pool. For example, if you select 1 negative lag and 1 positive lag, and have 3 predictor series, the total number of potential predictor series is 3 series times 3 lags (lags -1, 0, 1), or 9.

Lagged predictors increase the number of observations “left out” in each step of the cross validation. With no lags, the cross validation is “leave-1-out.” Assuming the maximum of the positive and negative lags is m , the cross-validation is “leave- M -out”, where $M=2m+1$. For example, if you allow 2 positive lags, the cross validation is “leave-5-out.” The increased number of observations omitted from calibration is necessary to retain the independence of the validation data from the data used to calibrate the model.

12. Edit dialog: enter maximum number of steps to run model. The first time through, accept the default, which lets all the variables in the pool of potential predictors enter the model.
13. Edit dialog: specify calibration period, or accept the default by clicking OK

The default is the period over which the set of selected series (predictors and predictand) overlap. That’s usually the best choice.

Now lots of windows flash by unreadable as the model is repeatedly fit and cross-validated. Depending on the speed of your computer, you may need to sit awhile till the process is completed.

Six figure windows have now been produced.

Fig 1. Time series plots of the predicted and observed predictand series for the calibration period. Below the plot are summary statistics, including the R^2 for regression. Refer to the notes for descriptions of these statistics. Note that at the first step the number of predictors in the final equation is 1. Note that this figure now lists an RE statistic and a root-mean-square error of cross-validation. Those two statistics are validation statistics, and so did not appear in Fig 1 for assignment 11.

Fig 2. Text window summarizing the regression equation. No different information here than in the same window for assignment 11. You see which variables are in the equation, as well as the regression coefficients and their 95% confidence bands. A key to the predictor time series X1, X2, ... is at right. In the equation, the order of entry is the number in parentheses. The lag on the predictor is coded as follows:

X2L0 – variable X2, lag 0
X2N1 – variable X2, lag -1 years (“N” stands for negative)
X2P1 – variable X2, lag +1 years (“P” stands for positive)
etc

Fig 3. Residuals analysis 1: distribution of residuals, correlation of residuals with predicted values and with the first two predictors. The scatterplots ideally are random in appearance and the histogram looks like that of a normal distribution (see notes). Note that the scatterplots of residuals on predictors other than first two predictors are not shown. Same content as for assignment 11.

Fig 4. Residuals analysis 2: autocorrelation of residuals. The top plot is a time series plot of residuals. This plot is useful in pointing out possible trend in residuals over time, as well as tendency of large residuals to cluster. At lower left is a scatterplot of residuals at time t against residuals at time $t-1$. Ideally this scatterplot shows no dependence. A linear pattern might indicate first-order autocorrelation of residuals. At lower right is the acf of the residuals. Ideally, the acf is close to zero at lags. Annotated below the plots are the Portmanteau statistic and the Durbin-Watson test results. Same content as for assignment 11.

Fig 5. These plots summarize the change in calibration and validation accuracy with each step of the model. At top is the change in reconstruction error as measured by RMSE of calibration and validation. Note that RMSE of calibration is equivalent to the standard error of the estimate for regression. Typically, the reconstruction error drops with the first few steps of the regression. The calibration error may continue dropping, although that may be reversed if the loss of degrees of freedom from additional predictors becomes more important than the incremental increase in the explained variance. The validation error typically begins rising after a few predictors have been entered. The turning point is the step at which additional predictors fail to improve the accuracy of prediction on independent data.

The bottom plots summarize the change in accuracy rather than the change in reconstruction error. In that sense the bottom plots resemble mirror images of the top plots about the horizontal axis (i.e., as error increases, accuracy decreases). R-squared summarizes the calibration period accuracy, and is therefore the counterpart of the root mean square error of calibration. R-squared ALWAYS increases with additional predictors, because an additional predictor will use up degrees of freedom and R-squared as plotted has not been adjusted for loss of degrees of freedom. More predictors automatically yield improved prediction on the training or calibration data.

The RE (reduction of error statistic), in contrast, is not constrained to increase as the model complexity increases. RE is computed on cross-validation data, and increases only as long as the additional predictor improves the ability of the model to predict that independent data. RE is the counterpart of the root mean square error of validation. So you will notice that the maximum in RE (highest validation accuracy) coincides with the minimum RMSEv (lowest validation error).

You can use these plots to decide on a cutoff point for entering predictors in the next run through the model. For example, you may see a well defined maximum in RE at step 3, indicating that the model should probably not be run out beyond step 3.

Fig 6. Time series plot of reconstruction and error bars, with extrapolation flagged. The error bars are plotted at 1 standard error of prediction (variable year to year) and at 1 RMSEv above and below the calibration period mean of the predictand. Note that 2 RMSEv is an approximate 95% confidence band. The terms are explained in the notes.

14. Print out windows from this first run, or otherwise write down the choices you made (e.g., which predictand, predictors, lagging choice, etc., so that you can run steps 1-13 with the identical setting a second time through.

The second time, though, there is a critical difference at step 12. Now you specify only the number of steps you identified as optimal for maximizing validation accuracy. Other steps are identical.

Left over in the workspace is a structure called "Results" that stores data and statistics from the analysis. Included in this storage is the time series of reconstructed values. Please type Results.what at the command line for a definition of data stored in Results.

PROGRAMMING NOTES

geosal1.m relies heavily on user-written functions, including:

armawht1 -- prewhitens time series with AR model
crospul2 – builds pointer to rows of time series matrix for cross-validation
lagyr3 – builds a matrix of lagged predictors
menudm1 – miscellaneous menu function
dwstat –Durbin-Watson statistic
acf – autocorrelation function
portmant – Portmanteau statistic
rederr – reduction-or-error statistic
stepvbl1—stepwise entry of variables based on ability to reduce residual variance

durbinwt.mat – lookup table for significance of D-W statistic
sepred2 – standard error of prediction
hatmtx – “hat matrix”
mce1 – minimum coverage ellipsoid