

ASSIGNMENT 11. MULTIPLE LINEAR REGRESSION

1. Run `geosal1.m`, selecting part I (assignment 11) from the first menu. This selection specifies multiple linear regression without cross-validation.

Menus appear allowing you 1) select time series to be used as predictand and predictors, 2) log-transform the predictand, 3) prewhiten the predictors, 4) include lagged predictors (lags up to $t-2$ or $t+2$ relative to the year of the predictand, and 5) specify a calibration period. Make choices that seem reasonable from your knowledge of the data and results of previous analyses. For this exercise, make sure your pool of potential predictors includes at least 6 variables. This can be achieved by various combinations of number of predictor series and lags. For example, two predictor series and lags -1 through $+1$ gives six potential predictors. Or six predictor time series and no lags also six potential predictors.

You will then begin an interactive modeling process that consists of stepwise entry of predictors one-by-one, with examination of model statistics and residuals analysis at each step. Note that at each step you must press “being modeling or add another predictor”, followed by “review results.” For this exercise, run the model out to 4 steps. Then press “Quit.” You should have 4 predictors are in your final model.

You will have four Figure windows. Answer the following questions and turn in your answers with printouts of the figure windows.

2. (Caption to Fig 1) What is the explanatory power of the relationship? Refer to the R^2 , overall F of the regression, the p -value for F in explaining your answer. Does the F -value support a statistically significant relationship? Why is adjusted R^2 smaller than R^2 ?
3. (Caption to Fig. 2) For which predictors are the regression coefficients significantly different than zero? What is multicollinearity? Name one undesirable effect of multicollinearity. Is multicollinearity a problem with your set of potential predictors? (Provide evidence in the form of results provided by running function `vif.m` on the matrix of your predictors – this matrix is `Xc(:,Lin)` in the workspace after running `geosal1`; you

can call the function as $R = \text{vif}(Xc(:,Lin),1)$, and can read the structure R.what for a definition of output)

4. (Caption to Fig 3) What is the assumption on the form of the distribution of the regression residuals? Does your analysis of residuals suggest that the assumption is satisfied? What can you say about the distribution of residuals for the extreme case of zero explanatory power of regression? (A “trick” question)

5. (Caption to Fig 4) What is the assumption on autocorrelation of residuals? Do the residuals violate this assumption. Refer the Portmanteau statistic, Durbin-Watson statistic, and visual appearance of acf of residuals in your answer.

Running goesa11.m

1. >geosa11
2. Respond to input dialog with the name of your data file; click OK
3. Menu: select “Part 1 – Assignment 11
4. Menu: select data set the predictand is to be selected from
5. Menu: select the predictand series (a “Y” appears beside it), then click “satisfied”
6. Menu: select whether no transform or log10 transform of predictand series
7. Menu: select the pool of potential predictors – up to 8 series allowed; then click “satisfied”
8. Menu: select whether or not to prewhiten the predictors
9. If “yes” to previous question, it will take a few seconds for the prewhitening to be completed
10. 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.

11. Edit dialog: enter an 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.

12. Menu: You are instructed to choose one: either 1) begin modeling or add another predictor, or 2) review results

The first time you encounter this menu you have obviously not yet begun modeling, so the choice should be to begin modeling. Click that and 4 figure windows appear, summarizing the regression modeling and residuals analysis for the first step of the stepwise modeling.

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.

Fig 2. Text window summarizing the regression equation. Here you see which variables are in the equation, as well as the regression coefficients and their 95% confidence band. 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.

Fig 4. Residuals analysis 2: autocorrelation of residuals. The top plot is a time series plot of residual. 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.

13. After reviewing the figures summarizing the regression for step 1, click on “Begin modeling or add another predictor” again, which enters a second predictor into the equation. Again browse through the four figures. Note how the explanatory power of the regression equation increases with the additional predictor.
14. Repeat step 13 through as many steps as desired. Then press “Quit” to stop the stepwise entry of predictors

PROGRAMMING NOTES

geosal1.m relies heavily 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”
mcel – minimum coverage ellipsoid

Many of the above functions are not used until assignment 12, which brings cross-validation into the regression model.