Predictive Error and Prediction Variance

Predictive Error for Least Squares

The predictive error (or Predictive Residual Error Sum of Squares, "PRESS") used in this report is the one defined by Allen (1974) and is a variation on the "Jacknife" statistic discussed in Tuckey (1967).

The Least Squares (LS) situation for fitting transformation models by LS can be written by defining the error at GCP "i" as:

$$e_i = \underset{\sim}{\varphi}_i^T \underset{\sim}{\alpha} - z_i$$

where $\underset{\sim}{\varphi}_i$ is the vector of "monomials" evaluated at the *i*'th GCP (x_i,y_i) in the "from" coordinate system, $\underset{\sim}{\alpha}$ is the vector of coefficients and z_i is the coordinate (x or y) being predicted in the "to" coordinate system.

If there are M GCPs then the complete expression for the residuals at the GCP points is:

$$\underset{\sim}{e} = \underset{\sim}{z} - A\underset{\sim}{\alpha}$$

The LS solution vector $\underset{\sim}{\alpha}^*$ minimises the sum of squared residuals or vector "norm":

$$\underset{\alpha}{Min} \left\| \underset{\sim}{e} \right\|^2 = \underset{\sim}{e}^T \underset{\sim}{e} = \sum_{i=1}^{M} e_i^2$$

The solution is well known and can be written in this notation as:

$$\begin{aligned} \underset{\sim}{\alpha}^* &= \left(A^T A \right)^{-1} A^T \underset{\sim}{z} \\ \underset{\sim}{e}^* &= \underset{\sim}{z} - A\underset{\sim}{\alpha}^* \\ &= \left[I - A\left(A^T A \right)^{-1} A^T \right] \underset{\sim}{z} \end{aligned}$$

The consequent estimate for the error called the Residual Mean Square error or "RMS" error is:

$$\begin{aligned} RMS &= \frac{1}{M} \sum_{i=1}^{M} \left| e_i^* \right|^2 \\ &= \frac{1}{M} RSS \end{aligned}$$

Where “RSS” is termed the Residual Sum of Squares which is not normalised by the number of observations M.

For reasons that will become clearer in the following we will denote the matrix:

$$\begin{aligned} A\left(A^T A\right)^{-1} A^T &= H \\ &= \left[h_{ij}\right] \\ h_{ij} &= \underset{\sim}{\varphi}_i^T \left(A^T A\right)^{-1} \underset{\sim}{\varphi}_j \end{aligned}$$

The (symmetric) matrix *H* is a special matrix in that powers of *H* simply result in *H* again as is the case with (*I-H*). It follows that the RSS can be written:

$$\begin{aligned} RSS &= \underset{\sim}{z}^T (I - H)\underset{\sim}{z} \\ &= \underset{\sim}{z}^T \underset{\sim}{z} - \underset{\sim}{z}^T H\underset{\sim}{z} \\ &= TSS - MSS \end{aligned}$$

which simply expresses the “analysis of variance” associated with LS of the total Sum of Squares (TSS) into the sum of Model Sum of Squares (MSS) and Residual sum of Squares (RSS).

The “Predictive” error is developed by noting that the errors at the GCPs after the model is fitted are too small as estimates for the actual statistical error. This comes about as they are used to fit the model. The consequent bias in the RMS is more than the normal allowance for the model degrees of freedom to obtain the estimated variance (EVAR) as in:

$$EVAR = \frac{1}{M - p} \sum_{i=1}^{M} \left|e_i^*\right|^2$$

(where p is the order of the model) will allow.

As an alternative and unbiased estimate of the error, if each GCP is removed from the model fit in turn and the error between the point removed and its prediction by the fitted model without its influence is used in place of the LS error at that point then the Predictive Error (“PRESS”) is obtained.

The method is really a variation of retaining some points as “test points” and some as “training points”. However, the existence of the test points is always a problem in such a scheme in that people wish to use all of the information they have to train and to test. The Predictive Error is a good compromise.

Let us write for the predictive error at the i’th GCP:

$$e_i^{(i)} = z_i - \underset{\sim}{\varphi}_i^T \underset{\sim}{\alpha}^{(i)}$$

where $\underset{\sim}{\alpha}^{(i)}$ is the set of model parameters obtained if the point z_i is not used to fit the model. Then:

$$\text{PRESS} = \frac{1}{M}\sum_{i=1}^{M}\left|e_i^{(i)}\right|^2$$

It may seem that this quantity involves M LS solutions and is quite messy. However, it turns out that the result is immediately available following the LS solution since:

$$\text{PRESS} = \frac{1}{M}\sum_{i=1}^{M} w_i \left|e_i^{*}\right|^2$$

$$w_i = \frac{1}{\left(1-h_{ii}\right)^2}$$

where e_i^* is the full LS residual obtained above and h_{ii} is the i'th diagonal element of the matrix *H* above.

The proof of this result can be found in Golub *et al.* (1979) but is easily and usefully derived as follows:

Noting that we can write the components of the LS solution as:

$$A^T A = \sum_{i=1}^{M} \underset{\sim}{\varphi}_i \underset{\sim}{\varphi}_i^T$$

$$A^T \underset{\sim}{z} = \sum_{i=1}^{M} z_i \underset{\sim}{\varphi}_i$$

it follows that we can express the i'th component of the PRESS as:

$$e_i^{(i)} = z_i - \underset{\sim}{\varphi}_i^T \underset{\sim}{\alpha}^{(i)}$$

where:

$$\underset{\sim}{\alpha}^{(i)} = \left(A^T A - \underset{\sim}{\varphi}_i \underset{\sim}{\varphi}_i^T\right)^{-1}\left(A^T \underset{\sim}{z} - z_i \underset{\sim}{\varphi}_i\right)$$

with the help of the well-known Sherman-Morrison formula (Sherman and Morrison, 1949) for (in this case) a non-singular and symmetric matrix, *H*:

$$(H + \lambda \underset{\sim}{a}\underset{\sim}{a}^T)^{-1} = H^{-1} - \frac{\lambda (H^{-1}\underset{\sim}{a})(H^{-1}\underset{\sim}{a})^T}{1+\lambda \underset{\sim}{a}^T H^{-1}\underset{\sim}{a}}$$

it follows with a little manipulation that:

$$\underset{\sim}{\varphi}_i^T \underset{\sim}{\alpha}^{(i)} = \left(1+\frac{h_{ii}}{1-h_{ii}}\right)\left(\underset{\sim}{\varphi}_i^T \underset{\sim}{\alpha}^* - h_{ii} z_i\right)$$
$$= \frac{\underset{\sim}{\varphi}_i^T \underset{\sim}{\alpha}^* - h_{ii} z_i}{1-h_{ii}}$$

so that

$$e_i^{(i)} = z_i - \underset{\sim}{\varphi}_i^T \underset{\sim}{\alpha}^{(i)}$$
$$= z_i - \frac{\underset{\sim}{\varphi}_i^T \underset{\sim}{\alpha}^* - h_{ii} z_i}{1-h_{ii}}$$
$$= \frac{z_i - \underset{\sim}{\varphi}_i^T \underset{\sim}{\alpha}^*}{1-h_{ii}}$$
$$= \frac{1}{1-h_{ii}} e_i^*$$

QED

Generalised Cross-Validation (GCV)

Golub *et al.* (1979) favoured the Generalised Cross-Validation or GCV statistic over PRESS because they say that PRESS is not “rotation invariant”. Perhaps this is not a vital characteristic of a statistic to have as I believe PRESS is an excellent tool that is sensitive to the distribution of the GCPs. This is the main reason for its use.

However, for completeness and because GCV is used to resolve “ties” in the modelling described here, the GCV needs to be derived as well.

We have seen above that PRESS can be written as:

$$\text{PRESS} = \frac{1}{M}\sum_{i=1}^{M} w_i \left|e_i^*\right|^2$$
$$w_i = \frac{1}{\left(1-h_{ii}\right)^2}$$

The GCV obtained by writing:

$$GCV = \frac{\frac{1}{M}\sum_{i=1}^{M}\left|e_i^*\right|^2}{\left(1-\frac{1}{M}\sum_{i=1}^{M} h_{ii}\right)^2}$$

It has a similar nature to PRESS but does not separately weight the points. This can be a disadvantage when the location of sensitive points and their stability to outlier data is being assessed as well as simply fitting data.

Both PRESS and GCV have the property that as the model order increases the fit becomes less stable and the statistics (as expressed by the how close $(1-h_{ii})$ is to zero) increase the errors. As the model order increases, RMS and RSS will both continue to decrease. PRESS and GCV are therefore commonly used to choose the "best" model order as described in this report.

[NOTE: microBRIAN program Model uses PRESS to help select the "best" order of model to use. Siever also has a calculation for the value of w_i to help identify points where errors will have greatest effect on the LS solution. These points can be the best to use if GCPs are accurate but also the worst to use if GCPs are inaccurate.]

Singular Value Decomposition

The Singular Value Decomposition (Lanczos, 1958; "SVD") provides a convenient method to solve the LS problem and also derive PRESS or GCV.

The SVD of the M row by N column matrix A is the (unique) decomposition of the matrix into factors such that:

$$A = USV^T$$

where U and V are ortho-normal and S is a diagonal matrix such that:

$$\begin{aligned} U^T U &= I_M \\ V^T V &= VV^T = I_p \\ S &= diag[s_1, s_2, ..., s_p] \\ s_1 &\geq s_2 \geq ... \geq s_p \geq 0 \end{aligned}$$

The pxp matrix V is the eigenvector matrix for $A^T A$ and the squares of the singular values (s_j) are the eigenvalues. The Mxp matrix U consists of the first p columns of the eigenvector matrix for AA^T.

If the matrix has Rank "q" (<p) then:

$$\begin{aligned} s_q &> 0 \\ s_{q+1} &= ... = s_p = 0 \end{aligned}$$

Also, if $\underset{\sim}{u}_j$ is the j'th column of U and $\underset{\sim}{v}_j$ is the j'th column of V then the SVD can also be written as:

$$A = USV^T = \sum_{j=1}^{p} s_j \underset{\sim}{u}_j \underset{\sim}{v}_j^T$$

The famous theorem of Eckert and Young (1936) concerning the approximation of a matrix A by a matrix B of lower rank is provided with a solution by the SVD in that the best approximation to A by a matrix B of rank q<p in the sense of minimum square norm is simply:

$$B = \sum_{j=1}^{q} s_j \underset{\sim}{u}_j \underset{\sim}{v}_j^T$$

$$\|A - B\|^2 = \sum_{i=1}^{M} \sum_{j=1}^{p} |a_{ij} - b_{ij}|^2$$

$$= \sum_{j=q+1}^{p} s_j^2$$

The SVD can be used to define a “generalised inverse” for the matrix A such that:

$$A^+ = VS^+U^T$$

$$s_j^+ = \begin{cases} \dfrac{1}{s_j} & if\ s_j > 0 \\ 0 & if\ s_j = 0 \end{cases}$$

The relationship with LS is that when the matrix A is the LS matrix above:

$$\underset{\sim}{\alpha}^* = A^+ \underset{\sim}{z}$$

$$\underset{\sim}{e}^* = (I - AA^+)\underset{\sim}{z}$$

$$H = AA^+ = UU^T$$

where in this case the columns of the matrix U are reduced to those corresponding to non-zero singular values. It follows that h_{ii} may be easily derived if the SVD is used to solve the LS problem since:

$$h_{ii} = \sum_{j=1}^{p} u_{ij}^2$$

$$w_i = \frac{1}{\left(1 - \sum_{j=1}^{p} u_{ij}^2\right)^2}$$

The SVD is also used to check that LS solutions are “well posed” in that singular values are either zero or not “close” to zero. Near-zero singular values correspond to unstable parameters. Various methods exist to modify the LS equations (often called “ridge” regression) to provide more stable (but biased) solutions. These will not be pursued here.

[NOTE: microBRIAN program Model uses SVD to solve the LS problem and to provide stabilisation when the singular values become small. This can happen with higher order models and few and poorly distributed GCPs.]

Predictive Variance

Predictive Variance (PV) is a related but different idea of the variation away from control. In this case, the resulting equation expresses how the model may vary at places away from the control due to the variation we know is in our estimate of the model parameters.

We can consider that the data vector values $\underset{\sim}{z}$ may have a variance, which we can estimate from the residual errors or (better) the PRESS. We will assume (although mainly because it is rare to have enough information to do otherwise – the expressions can be derived under different knowledge) that the error model is:

$$\begin{aligned}
&\underset{\sim}{z} = \underset{\sim}{z}' + \delta\underset{\sim}{z} \\
&E(\delta\underset{\sim}{z}) = \underset{\sim}{0} \\
&Var(\delta\underset{\sim}{z}) = E(\delta\underset{\sim}{z}\delta\underset{\sim}{z}^T) = \sigma^2 I
\end{aligned}$$

Even accepting that the LS model predicted data values are an estimate for $\underset{\sim}{z}'$ there will be an expected variation around the model parameters ($\delta\underset{\sim}{\alpha}$) with mean zero and variance:

$$\begin{aligned}
&\delta\underset{\sim}{\alpha} = \left(A^T A\right)^{-1} A^T \delta\underset{\sim}{z} \\
&E(\delta\underset{\sim}{\alpha}) = \underset{\sim}{0} \\
&Var(\delta\underset{\sim}{\alpha}) = E(\delta\underset{\sim}{\alpha}\delta\underset{\sim}{\alpha}^T) = \sigma^2\left(A^T A\right)^{-1}
\end{aligned}$$

The variation of the model in the “to” coordinate system at a point predicted from the point (x,y) in the “from” system is therefore:

$$\begin{aligned}
\delta z(x,y) &= \underset{\sim}{\varphi}(x,y)^T \delta\underset{\sim}{\alpha} \\
E(\delta z) &= 0 \\
Var(\delta z) &= \sigma^2 \underset{\sim}{\varphi}(x,y)^T \left(A^T A\right)^{-1} \underset{\sim}{\varphi}(x,y) \\
&= \sigma^2\, h(x,y)
\end{aligned}$$

The function $h(x,y)$ is related to the Predictive Error above since at the GCP points:

$$h(x_i, y_i) = h_{ii}$$

The function $h(x,y)$ may be plotted over the “from” space coordinate system to indicate how variations may increase away from control. To allow for the possibility

of higher order models being present this is often done using PRESS for the data variance and a higher order of model than the one used for the fitting. Areas of high predictive variation are ones where more control should be obtained if possible.

[NOTE: In microBRIAN the function $h(x, y)$ was called the Predictive Error function but it would be better to call it the "predictive variation" due to the possibility of confusion with the PRESS defined above. It, along with Model, Siever and Mosmod provided a useful analysis of the effectiveness of GCP modelling.]

References

Allen, D.M. (1974). The relationship between variable selection and data augmentation and a method of prediction. *Technometrics*, **16**, 125-127.

Eckert, C. and Young, G. (1936). The approximation of one matrix by another of lower rank. *Psychometrika*, **1**, 211-218.

Golub, G.H., Heath, M. and Wahba, G. (1979). Generalised cross validation as a method for choosing a good ridge parameter. *Technometrics*, **21**, 215-223.

Lanczos, C. (1958). Linear systems in self-adjoint form. *Am. Math. Month.*, **65**, 665-679.

Sherman, J. and Morrison, W.J. (1949). Adjustment of an inverse matrix corresponding to a change in one element of a given matrix. *Ann. Math. Statist.*, **21**, 124-127.

Tuckey, J.W. (1967). Discussion of Anscome (1967). *J.R. Statist. Soc. B,* **29**, 47-48.