Chapter 1

Updating Techniques and Cross Validation

1.1 Cross Validation

Cross validation techniques for linear regression employ the use 'leave one out' re-calculations. In such procedures the regression coefficients are estimated for $n-1$ covariates, with the Q^{th} observation omitted.

Let $\hat{\beta}$ denote the least square estimate of β based upon the full set of observations, and let $\hat{\beta}^{-Q}$ denoted the estimate with the Q^{th} case excluded.

In leave-one-out cross validation, each observation is omitted in turn, and a regression model is fitted on the rest of the data. Cross validation is used to estimate the generalization error of a given model. alternatively it can be used for model selection by determining the candidate model that has the smallest generalization error.

Evidently leave-one-out cross validation has similarities with 'jackknifing', a well known statistical technique. However cross validation is used to estimate generalization error, whereas the jackknife technique is used to estimate bias.

1.1.1 The Hat Matrix

The projection matrix H (also known as the hat matrix), is a well known identity that maps the fitted values $\hat{Y}$ to the observed values Y, i.e. $\hat{Y} = HY$.

$$H = \quad X(X^T X)^{-1} X^T \tag{1.1}$$

The hat matrix, also known as the projection matrix, is well known in classical linear models. The diagonal elements h_{ii} are known as 'leverages'. The properties of $\boldsymbol{H}$,such as symmetry and idempotency, are well known.

$$\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X'X})^{-1}\boldsymbol{X'}$$

$$\boldsymbol{H} = \begin{bmatrix} h_{ii} & \boldsymbol{h}_i' \\ \boldsymbol{h}_i & \boldsymbol{H}_{(i)} \end{bmatrix}$$

$\boldsymbol{H}_{(i)}$ is an $(n-1) \times (n-1)$ matrix. It's inversion for each i is computationally expensive.

$$\boldsymbol{C} = \boldsymbol{H}^{-1} = \begin{bmatrix} c_{ii} & \boldsymbol{h}_c' \\ \boldsymbol{c}_i & \boldsymbol{C}_{(i)} \end{bmatrix}$$

H describes the influence each observed value has on each fitted value. The diagonal elements of the H are the 'leverages', which describe the influence each observed value has on the fitted value for that same observation. The residuals (R) are related to the observed values by the following formula:

$$R = (I - H)Y \tag{1.2}$$

The variances of Y and R can be expressed as:

$$\text{var}(Y) = H\sigma^2$$
$$\text{var}(R) = (I - H)\sigma^2 \tag{1.3}$$

Updating techniques allow an economic approach to recalculating the projection matrix, H, by removing the necessity to refit the model each time it is updated. However this approach is known for numerical instability in the case of down-dating.

1.2 Efficient updating theorem

It is convenient to write partitioned matrices in which the i-th case is isolated. The partitioned matrix is written as $i = 1$, but the results apply in general.

If $\boldsymbol{C'}_i = [c_{ii}, \boldsymbol{c'}_i]$, such that $\boldsymbol{C}_i$ is the i-th column of $\boldsymbol{H}^{-1}$ then

- $m_i = \frac{1}{c_{ii}}$
- $\breve{x}_i = \frac{1}{c_{ii}} \boldsymbol{X'C}_i$
- $\breve{\boldsymbol{z}}_{ji} = \frac{1}{c_{ii}} \boldsymbol{Z'}_j \boldsymbol{C}_i$
- $\breve{y}_i = \frac{1}{c_{ii}} \boldsymbol{y'C}_i$

Once $\boldsymbol{H}^{-1}$ is determined, an efficient updating formula can be applied.

$$\boldsymbol{H}^{-1} = \boldsymbol{I} - \boldsymbol{Z}(\boldsymbol{D}^{-1} + \boldsymbol{ZZ})^{-1}\boldsymbol{Z'} \tag{1.4}$$

1.3 Cross Validation: Updating standard deviation

The variance of a data set can be calculated using the following formula.

$$S^2 = \frac{\sum_{i=1}^{n}(x_i^2) - \frac{(\sum_{i=1}^{n} x_i)^2}{n}}{n-1} \tag{1.5}$$

While using bivariate data, the notation Sxx and Syy shall apply to the variance of x and of y respectively. The covariance term Sxy is given by

$$Sxy = \frac{\sum_{i=1}^{n}(x_i y_i) - \frac{(\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n}}{n-1} \tag{1.6}$$

Let the observation j be omitted from the data set. The estimates for the variance identities can be updating using minor adjustments to the full sample estimates. Where (j) denotes that the jth has been omitted, these identities are

$$Sxx^{(j)} = \frac{\sum_{i=1}^{n}(x_i^2) - (x_j)^2 - \frac{((\sum_{i=1}^{n} x_i) - x_j)^2}{n-1}}{n-2} \tag{1.7}$$

$$Syy^{(j)} = \frac{\sum_{i=1}^{n}(y_i^2) - (y_j)^2 - \frac{((\sum_{i=1}^{n} y_i) - y_j)^2}{n-1}}{n-2} \tag{1.8}$$

$$Sxy^{(j)} = \frac{\sum_{i=1}^{n}(x_i y_i) - (y_j x_j) - \frac{((\sum_{i=1}^{n} x_i) - x_j)(\sum_{i=1}^{n} y_i) - y_k)}{n-1}}{n-2} \tag{1.9}$$

The updated estimate for the slope is therefore

$$\hat{\beta}_1^{(j)} = \frac{Sxy^{(j)}}{Sxx^{(j)}} \tag{1.10}$$

It is necessary to determine the mean for x and y of the remaining $n-1$ terms

$$\bar{x}^{(j)} = \frac{(\sum_{i=1}^{n} x_i) - (x_j)}{n-1}, \tag{1.11}$$

$$\bar{y}^{(j)} = \frac{(\sum_{i=1}^{n} y_i) - (y_j)}{n-1}. \tag{1.12}$$

The updated intercept estimate is therefore

$$\hat{\beta}_0^{(j)} = \bar{y}^{(j)} - \hat{\beta}_1^{(j)} \bar{x}^{(j)}. \tag{1.13}$$

1.4 Updating Estimates

1.4.1 Updating of Regression Estimates

Updating techniques are used in regression analysis to add or delete rows from a model, allowing the analyst the effect of the observation associated with that row. In time series problems, there will be scientific interest in the changing relationship between variables. In cases where there a single row is to be added or deleted, the procedure used is equivalent to a geometric rotation of a plane.

Updating techniques are used in regression analysis to add or delete rows from a model, allowing the analyst the effect of the observation associated with that row.

Consider a $p \times p$ matrix X, from which a row x_i^T is to be added or deleted. **?** sets $A = X^T X$, $a = -x_i^T$ and $b = x_i^T$, and writes the above equation as

$$(X^T X \pm x_i x_i^T)^{-1} = \quad (X^T X)^{-1} \mp \quad \frac{(X^T X)^{-1}(x_i x_i^T (X^T X)^{-1}}{1 - x_i^T (X^T X)^{-1} x_i} \tag{1.14}$$

This approach allows an economic approach to recalculating the projection matrix, V, by removing the necessity to refit the model each time it is updated.

This approach is known for numerical instability in the case of downdating.

1.4.2 Updating Regression Estimates

Let the observation j be omitted from the data set. The estimates for the variance identities can be updating using minor adjustments to the full sample estimates. Where (j) denotes that the jth has been omitted, these identities are

$$Sxx^{(j)} = \frac{\sum_{i=1}^{n}(x_i^2) - (x_j)^2 - \frac{((\sum_{i=1}^{n} x_i) - x_j)^2}{n-1}}{n-2} \tag{1.15}$$

$$Syy^{(j)} = \frac{\sum_{i=1}^{n}(y_i^2) - (y_j)^2 - \frac{((\sum_{i=1}^{n} y_i) - y_j)^2}{n-1}}{n-2} \tag{1.16}$$

$$Sxy^{(j)} = \frac{\sum_{i=1}^{n}(x_i y_i) - (y_j x_j) - \frac{((\sum_{i=1}^{n} x_i) - x_j)(\sum_{i=1}^{n} y_i) - y_k)}{n-1}}{n-2} \tag{1.17}$$

The updated estimate for the slope is therefore

$$\hat{\beta}_1^{(j)} = \frac{Sxy^{(j)}}{Sxx^{(j)}} \tag{1.18}$$

It is necessary to determine the mean for x and y of the remaining $n-1$ terms

$$\bar{x}^{(j)} = \frac{(\sum_{i=1}^{n} x_i) - (x_j)}{n-1}, \tag{1.19}$$

$$\bar{y}^{(j)} = \frac{(\sum_{i=1}^{n} y_i) - (y_j)}{n-1}. \tag{1.20}$$

The updated intercept estimate is therefore

$$\hat{\beta}_0^{(j)} = \bar{y}^{(j)} - \hat{\beta}_1^{(j)} \bar{x}^{(j)}. \tag{1.21}$$

1.5 Sherman Morrison Woodbury Formula

The 'Sherman Morrison Woodbury' Formula is a well known result in linear algebra;

$$(A + a^T B)^{-1} \quad = \quad A^{-1} - A^{-1} a^T (I - b A^{-1} a^T)^{-1} b A^{-1} \tag{1.22}$$

This result is highly useful for analyzing regression diagnostics, and for matrices inverses in general. Consider a $p \times p$ matrix X, from which a row x_i^T is to be added or deleted. **?** sets $A = X^T X$, $a = -x_i^T$ and $b = x_i^T$, and writes the above equation as

$$(X^T X \pm x_i x_i^T)^{-1} = \quad (X^T X)^{-1} \mp \quad \frac{(X^T X)^{-1}(x_i x_i^T (X^T X)^{-1}}{1 - x_i^T (X^T X)^{-1} x_i} \tag{1.23}$$

The projection matrix H (also known as the hat matrix), is a well known identity that maps the fitted values $\hat{Y}$ to the observed values Y, i.e. $\hat{Y} = HY$.

$$H = \quad X(X^T X)^{-1} X^T \tag{1.24}$$

H describes the influence each observed value has on each fitted value. The diagonal elements of the H are the 'leverages', which describe the influence each observed value has on the fitted value for that same observation. The residuals (R) are related to the observed values by the following formula:

$$R = (I - H)Y \tag{1.25}$$

The variances of Y and R can be expressed as:

$$\begin{aligned} \text{var}(Y) &= H\sigma^2 \\ \text{var}(R) &= (I - H)\sigma^2 \end{aligned} \tag{1.26}$$

Updating techniques allow an economic approach to recalculating the projection matrix, H, by removing the necessity to refit the model each time it is updated. However this approach is known for numerical instability in the case of down-dating.

1.5.1 Hat Values for MCS regression

With A as the averages and D as the casewise differences.

```
fit = lm(D~A)
```

$$H = A\left(A^\top A\right)^{-1} A^\top,$$

1.6 Updating Estimates

1.6.1 Updating Standard deviation

A simple, but useful, example of updating is the updating of the standard deviation when an observation is omitted, as practised in statistical process control analyzes. From first principles, the variance of a data set can be calculated using the following formula.

$$S^2 = \frac{\sum_{i=1}^{n}(x_i^2) - \frac{(\sum_{i=1}^{n} x_i)^2}{n}}{n-1} \tag{1.27}$$

While using bivariate data, the notation Sxx and Syy shall apply hither to the variance of x and of y respectively. The covariance term Sxy is given by

$$Sxy = \frac{\sum_{i=1}^{n}(x_i y_i) - \frac{(\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n}}{n-1}. \tag{1.28}$$

1.6.2 Inference on intercept and slope

$$\hat{\beta}_1 \pm t_{(\alpha,n-2)}\sqrt{\frac{S^2}{(n-1)S_x^2}} \tag{1.29}$$

$$\frac{\hat{\beta}_0 - \beta_0}{SE(\hat{\beta}_0)} \tag{1.30}$$

$$\frac{\hat{\beta}_1 - \beta_1}{SE(\hat{\beta}_0)} \tag{1.31}$$

1.6.3 Inference on correlation coefficient

This test of the slope is coincidentally the equivalent of a test of the correlation of the n observations of X and Y.

$$H_0 : \rho_{XY} = 0$$

$$H_A : \rho_{XY} \neq 0 \tag{1.32}$$