5.7 Matrix formulation

Warning: this is a more advanced, optional section and assumes knowledge of matrix algebra.

Recall that multiple regression model can be written as

$$y_t = \beta_0 + \beta_1 x_{1,t} + \beta_2 x_{2,t} + \cdots + \beta_k x_{k,t} + \varepsilon_t$$

where ε_t has mean zero and variance σ^2. This expresses the relationship between a single value of the forecast variable and the predictors.

It can be convenient to write this in matrix form where all the values of the forecast variable are given in a single equation. Let $\boldsymbol{y} = (y_1, \dots, y_T)'$, $\boldsymbol{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_T)'$, $\boldsymbol{\beta} = (\beta_0, \dots, \beta_k)'$ and

$$\boldsymbol{X} = \begin{bmatrix} 1 & x_{1,1} & x_{2,1} & \dots & x_{k,1} \\ 1 & x_{1,2} & x_{2,2} & \dots & x_{k,2} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{1,T} & x_{2,T} & \dots & x_{k,T} \end{bmatrix}.$$

Then

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}.$$

where $\boldsymbol{\varepsilon}$ has mean $\mathbf{0}$ and variance $\sigma^2\boldsymbol{I}$. Note that the $\boldsymbol{X}$ matrix has T rows reflecting the number of observations and $k+1$ columns reflecting the intercept which is represented by the column of ones plus the number of predictors.

Least squares estimation

Least squares estimation is performed by minimising the expression $\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon} = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$. It can be shown that this is minimised when $\boldsymbol{\beta}$ takes the value

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$$

This is sometimes known as the "normal equation." The estimated coefficients require the inversion of the matrix $\boldsymbol{X}'\boldsymbol{X}$. If $\boldsymbol{X}$ is not of full column rank then matrix $\boldsymbol{X}'\boldsymbol{X}$ is singular and the model cannot be estimated. This will occur, for example, if you fall for the "dummy variable trap," i.e., having the same number of dummy variables as there are categories of a categorical predictor, as discussed in Section 5.4.

The residual variance is estimated using

$$\hat{\sigma}_e^2 = \frac{1}{T-k-1}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})'(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}).$$

Fitted values and cross-validation

The normal equation shows that the fitted values can be calculated using

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} = \boldsymbol{H}\boldsymbol{y},$$

where $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ is known as the "hat-matrix" because it is used to compute $\hat{\boldsymbol{y}}$ ("y-hat").

If the diagonal values of $\boldsymbol{H}$ are denoted by $h_1, \dots, h_T$, then the cross-validation statistic can be computed using

$$\text{CV} = \frac{1}{T}\sum_{t=1}^{T}[e_t/(1-h_t)]^2,$$

where e_t is the residual obtained from fitting the model to all T observations. Thus, it is not necessary to actually fit T separate models when computing the CV statistic.

Forecasts and prediction intervals

Let $\boldsymbol{x}^*$ be a row vector containing the values of the predictors (in the same format as $\boldsymbol{X}$) for which we want to generate a forecast . Then the forecast is given by

$$\hat{y} = \boldsymbol{x}^*\hat{\boldsymbol{\beta}} = \boldsymbol{x}^*(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$$

and its estimated variance is given by

$$\hat{\sigma}_e^2\left[1 + \boldsymbol{x}^*(\boldsymbol{X}'\boldsymbol{X})^{-1}(\boldsymbol{x}^*)'\right].$$

A 95% prediction interval can be calculated (assuming normally distributed errors) as

$$\hat{y} \pm 1.96\hat{\sigma}_e\sqrt{1 + \boldsymbol{x}^*(\boldsymbol{X}'\boldsymbol{X})^{-1}(\boldsymbol{x}^*)'}.$$

This takes into account the uncertainty due to the error term ε and the uncertainty in the coefficient estimates. However, it ignores any errors in $\boldsymbol{x}^*$. Thus, if the future values of the predictors are uncertain, then the prediction interval calculated using this expression will be too narrow.