Problem Set 6

K-fold cross-validation

Cross-validation is a re-sampling method that is used for model validation and selection to avoid overfitting in a variety of modeling contexts, especially for machine learning. The basic algorithm for model selection is written below.

Algorithm 1 K**-fold adjusted cross-validation**

- randomly divide the set of observations into k groups of (approx.) equal size.
- first fold is treated as a validation set
- method is fit on the remaining $k-1$ folds
- MSE is then computed on the observations in the held-out fold
- procedure is repeated k times; each time, a different group of observations is treated as a validation set

This process results in k estimates of the test error,

$\text{MSE}_1, \text{MSE}_2, \ldots, \text{MSE}_k$ The k-fold CV estimate is computed by averaging these values,

$$\text{CV}_{(k)} = \frac{1}{k} \sum_{i=1}^{k} \text{MSE}_i \tag{1}$$

LOOCV is a special case of k-fold CV in which k is set to equal n.

Application

We will use this to evaluate whether we should include a squared term in a linear regression equation.

The true data generating process is given by

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

Where $\beta_0 = 1$, $\beta = 0.1$ and $\varepsilon \sim \mathcal{N}(0, 10)$ for $n = 100$. We want to evaluate, given the data generating process above, whether we should include $\beta_2 * x_i^2$ when estimating $\widehat{\boldsymbol{\beta}}$ (subsequently called full model).

- Pick one seed, estimate the full model on this data and plot the data along with the estimated function from the full model.
- Implement the cross-validation procedure as described above for $k = 2$. For each sample estimate the CV-error for the full and for the simple model.
- Repeat the process above 100 times. Record the number of times the full model is selected and repeat this for $k = 5$ and $k = 10$.