Lecture 10

Automatic Variable Selection

- Instead of choosing what goes into the model manually, we can find a program/algorithm to do so

- Suppose we have y and q predictors. There are 2^q possible regression models

- A computer can potentially picks a "best" model, but it has to be taught what is meant by "best"

- Keep in mind: bias-variance tradeoff for prediction based on learned model

    - Bias: the prediction error resulting from miss-specifying the model
    - Variance: the prediction error resulting from variations in the data used for fitting

    - Suppose we have a model f to describe response y from feature z. Given a new z_{n+1} we may write

    $$y_{n+1} = \underbrace{f(z_{n+1})}_{z_{n+1}^T \beta} + \varepsilon_{n+1}$$

    - Suppose that we have past data $(y_i, z_i)_{i=1}^n$ we estimate $\hat{f}$ from this data

        In the linear model:
        $$\hat{f}(z) = z^T \hat{\beta} \quad, \quad \hat{\beta} = (Z^T Z)^{-1} Z y$$
        $$z \in \mathbb{R}^p \qquad\qquad Z \in \mathbb{R}^{n \times p}$$

    - we have the decomposition: