Homework 7

Due Thursday, 10/31/19

1. (Leave one out cross validation) Let $\boldsymbol{X} = \begin{bmatrix} \boldsymbol{x}_1^T \\ \vdots \\ \boldsymbol{x}_n^T \end{bmatrix} \in \mathbb{R}^{n \times p}$ be a full rank design matrix, $\boldsymbol{H} = \boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T$, h_{ii} be the ith diagonal element of $\boldsymbol{H}$ and $\boldsymbol{Y} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} \in \mathbb{R}^n$ be the response. For each $i = 1, \ldots, n$, let $\boldsymbol{X}_{(-i)} \in \mathbb{R}^{(n-1)\times p}$ be the design matrix with the ith row removed and $\boldsymbol{Y}_{(-i)} \in \mathbb{R}^{n-1}$ be the response with the ith element removed. Define

$$\hat{\boldsymbol{\beta}}_{(-i)} = \left(\boldsymbol{X}_{(-i)}^T\boldsymbol{X}_{(-i)}\right)^{-1}\boldsymbol{X}_{(-i)}^T\boldsymbol{Y}_{(-i)}$$

to be the ordinary least squares estimator that ignores the ith sample.

   (a) Show that $\left(\boldsymbol{X}_{(-i)}^T\boldsymbol{X}_{(-i)}\right)^{-1} = \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1} + (1-h_{ii})^{-1}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{x}_i\boldsymbol{x}_i^T\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}$.

   (b) Show that the leave one out cross validation error,

$$PRESS = \sum_{i=1}^{n}\left(Y_i - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}_{(-i)}\right)^2,$$

   can be written as

$$PRESS = \sum_{i=1}^{n}\left(\frac{Y_i - \hat{Y}_i}{1-h_{ii}}\right)^2,$$

   where $\hat{Y}_i$ is the ith element of $\boldsymbol{HY} \in \mathbb{R}^n$.

2. (AIC and BIC) Assume $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ is a full rank design matrix and $\boldsymbol{Y} \sim N\left(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I}_n\right)$. Let $L(\boldsymbol{b}, v) = (2\pi v)^{-n/2}\exp\left\{-\frac{1}{2v}(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{b})^T(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{b})\right\}$ be the likelihood for the normal distribution and $\hat{L} = L\left(\hat{\boldsymbol{\beta}}_{(MLE)}, \hat{\sigma}^2_{(MLE)}\right)$, where $\hat{\boldsymbol{\beta}}_{(MLE)}$ and $\hat{\sigma}^2_{(MLE)}$ are the maximum likelihood estimates for $\boldsymbol{\beta}$ and σ^2. Show that up to constants that do not depend on $\boldsymbol{Y}$, $\boldsymbol{X}$ or p,

$$AIC = n\log(SSE) + 2p, \quad BIC = n\log(SSE) + \log(n)p,$$

   where AIC and BIC are defined on slide 15 of lecture 15. Use this to show that for Gaussian data and n suitably large, the model chosen by AIC is always at least as large as that chosen by BIC.

3. Consider the data Boston.txt that we looked at in lecture on 10/24 and regress mvalue onto all predictors. Using residual × covariate plots, suggest at least one way to improve the mean model. Use AIC, BIC or adjusted R2 (or all 3) to show that your new mean model is superior to the one we considered in class.

4. Consider the data "Steam.txt" where you are trying to predict steam usage using fat, glycerine, wind, freezday and temp. Perform forward and backward selection with $\alpha_1 = 0.1$ and $\alpha_2 = 0.2$, respectively. Which covariates did you include in the full model? How does this compare when you perform best subset regression using AIC to rank models? Does the model chosen with best subset regression change when you use BIC to rank models?