

Homework 1 - ST 790b - Model Selection

1. (**Background**) Breiman (1992) defined Model Error for a data set in the following way for a fixed design \mathbf{X} . Let \mathbf{Y}^{new} be an independent vector with the exact same distribution as the original \mathbf{Y} , i.e., $\mathbf{Y}^{\text{new}} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}^{\text{new}}$, where \mathbf{Y}^{new} is independent of $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$. Then prediction error (PE) for a predicted response $\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$ is defined by

$$\text{PE} = \text{E}(\mathbf{Y}^{\text{new}} - \hat{\mathbf{Y}})^T (\mathbf{Y}^{\text{new}} - \hat{\mathbf{Y}}) = (\hat{\mathbf{Y}} - \boldsymbol{\mu})^T (\hat{\mathbf{Y}} - \boldsymbol{\mu}) + n\sigma^2,$$

where the expectation is only with respect to \mathbf{Y}^{new} , and the simplification is obtained by adding and subtracting $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$ within the parentheses and then expanding. Model Error (ME) is then defined to be the first term above,

$$\text{ME} = (\hat{\mathbf{Y}} - \boldsymbol{\mu})^T (\hat{\mathbf{Y}} - \boldsymbol{\mu}) = (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T \mathbf{X}^T \mathbf{X} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}). \quad (1)$$

Note that ME is a random quantity and equal to squared error $(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ for orthogonal designs where $\mathbf{X}^T \mathbf{X}$ is the identity matrix. In a Monte Carlo study it is typical to estimate

$$\text{E}(\text{ME}) = \text{tr}\{\text{Cov}(\hat{\mathbf{Y}})\} + \{\text{E}(\hat{\mathbf{Y}}) - \boldsymbol{\mu}\}^T \{\text{E}(\hat{\mathbf{Y}}) - \boldsymbol{\mu}\}$$

by averaging ME over replications. Now there are three parts to answer.

a) In the random design case, we assume $(\mathbf{X}_1, Y_1), \dots, (\mathbf{X}_n, Y_n)$ are iid from some distribution, $Y_i = \mathbf{X}_i^T \boldsymbol{\beta} + e_i$, $i = 1, \dots, n$, and $\text{E}(e_i | \mathbf{X}_i) = 0$ so that $\text{E}(Y_i | \mathbf{X}_i) = \mathbf{X}_i^T \boldsymbol{\beta}$. Let $(\mathbf{X}_{n+1}, Y_{n+1})$ be an independent new pair from the distribution (plays the role of the “new” data above, but note that it is a single pair in contrast to \mathbf{Y}^{new} plus \mathbf{X} that is a full data set). Proceeding as above, show that Model Error should be defined as

$$\text{ME} = (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T \Gamma (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}). \quad (2)$$

where $\Gamma = \text{E}\{\mathbf{X}_{n+1} \mathbf{X}_{n+1}^T\}$.

b) In a Monte Carlo study, the fixed case $\text{E}(\text{ME})$ is easy to estimate using

$$\frac{1}{N} \sum_{j=1}^N \text{ME}_j = \frac{1}{N} \sum_{j=1}^N (\hat{\mathbf{Y}}_j - \boldsymbol{\mu}_j)^T (\hat{\mathbf{Y}}_j - \boldsymbol{\mu}_j), \quad (3)$$

where N refers to the number of Monte Carlo replications and ME_j is Model Error for the j th data set. How would you estimate $E(ME)$ for the random case when you do not have a simple form for Γ (thus you have to estimate it from the Monte Carlo samples)? In the Monte Carlo study you will have N independent data sets, each composed of n iid pairs $(\mathbf{X}_1, Y_1), \dots, (\mathbf{X}_n, Y_n)$.

c) In a recent JASA submission, for the random design case, the author defined

$$\text{Out-of-Sample } R^2 = 1 - \frac{\sum_{i=1}^n (Y_i^* - \mathbf{X}_i^* \widehat{\boldsymbol{\beta}})^2}{\sum_{i=1}^n (Y_i^* - \bar{Y}^*)^2},$$

where $(\mathbf{X}_1^*, Y_1^*), \dots, (\mathbf{X}_n^*, Y_n^*)$ is a replicate data set with exactly the same distribution as the original $(\mathbf{X}_1, Y_1), \dots, (\mathbf{X}_n, Y_n)$. Then, for a Monte Carlo study, they generated N of these duplicate data sets in addition to the N original data sets. For each original data set, they obtain $\widehat{\boldsymbol{\beta}}$ and compute Out-of-Sample R^2 from one of the duplicate data sets. They then average these N Out-of-Sample R^2 . Explain how this Monte Carlo average relates to $E(ME)$. (Hint: Approximate the denominator by an expectation and then take iterated expectations.)

2. The baseball data set is on my Variable Selection website,

<http://www4.stat.ncsu.edu/~boos/var.select/baseball.html>

There it explains that the data set contains salary information for 337 Major League Baseball (MLB) players who were not pitchers and played at least one game during both the 1991 and 1992 seasons. The purpose of the study was to determine whether a baseball player's salary is a reflection of his offensive performance. For each player, the salary from the 1992 season along with 12 offensive statistics from the 1991 season were collected. In addition to these variables, there are 4 indicator variables that identify free agency and eligibility for arbitration. You are to find a good but parsimonious model relating salary to the predictor variables. At minimum I would like you to download SAS GLMSELECT Procedure from

<http://support.sas.com/rnd/app/da/glmselect.html>

and use it along with measures like C_p , BIC (SBC), AIC and the forward and backward sequences. You might look at plots and think about adding quadratic terms and/or interactions.