Variable Selection

- Subset selection with AIC/BIC
- Regularization methods: Ridge and Lasso
- Case study: Boston Housing Data.

Introduction to Variable Selection

In modern statistical applications nowadays, we have many potential predictors, i.e., p is large and we could even have $p \gg n$.

In some applications, the key question we need to answer is to identify a subset of the predictors that are most relevant to Y.

If our goal is simply to do well on prediction/estimation (i.e., we don't care whether the predictors employed by our linear model are really relevant to Y or not), then should we care about variable selection? To understand this, let's examine the training and the test errors.

Test vs Training Error

- Training data $(\mathbf{x}_i, y_i)_{i=1}^n$. Fit a linear model on the training data and define

$$\text{Train Err} = \|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2,$$

where $\hat{\boldsymbol{\beta}} \in \mathbb{R}^p$ is the LS estimate of the regression parameter.

- Test data $(\mathbf{x}_i, y_i^*)_{i=1}^n$ is an independent data set collected at the same location $\mathbf{x}_i$'s. Define

$$\text{Test Err} = \|\mathbf{y}^* - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2.$$

- Note that the two errors are random; In the two equations above, terms are colored differently representing difference sources of randomness. Next we decompose the expectation of the two errors into three components.

We can show that

$$\begin{aligned}\mathbb{E}[\text{Train Err}] &= (\text{Unavoidable Err}) - p\sigma^2 + \text{Bias} \\ \mathbb{E}[\text{Test Err}] &= (\text{Unavoidable Err}) + p\sigma^2 + \text{Bias}\end{aligned}$$

where

- Unavoidable Err: we usually model $Y = f^*(X) + \text{err}$, so even if we know f^*, we still cannot predict Y perfectly.
- Bias: we could encounter this error if the true function f^* is not linear or the current model misses some relevant variables.
- Notice the sign of the term $p\sigma^2$, which increases the "Test Err" (on average) while decreases the "Training Err". So even if our goal is purely prediction, it's not true that the more the predictors the better the prediction. We should benefit from removing some irrelevant variables.