Lab: Bias-Variance Tradeoff

1. Park effects simulation study

It would be great to compute bias and variance via

$\text{Bias}(\hat{f})^2 = (f - \mathbb{E}_{\mathcal{D}}\hat{f})^2$ and $\text{Var}(\hat{f}) = \mathbb{E}_{\mathcal{D}}(\hat{f} - \mathbb{E}_{\mathcal{D}}\hat{f})^2$

for real-world estimators $\hat{f}$ of real-world sports datasets $\mathcal{D}$, but the "true" f is unknown and $\mathbb{E}_{\mathcal{D}}$ is an expectation over the randomness of drawing the dataset $\mathcal{D}$, which is incalculable.

To understand the nature of the bias-variance tradeoff, we turn to a simulation study, using park effects as our example.

In this simulation study, we assume that the park, team offensive quality, and team defensive quality coefficients are known.

Specifically, generate a "true" parameter vector β according to

$$\begin{cases} \beta_0 = 0.4, \\ \beta_j^{(\text{park})} \overset{\text{iid}}{\sim} \mathcal{N}(0.04, 0.065), \\ \beta_k^{(\text{off})} \overset{\text{iid}}{\sim} \mathcal{N}(0.02, 0.045), \\ \beta_k^{(\text{def})} \overset{\text{iid}}{\sim} \mathcal{N}(0.03, 0.07). \end{cases}$$

Then, we assemble our park effects data matrix X associated with the model

$$y_i = \beta_{\text{park}(i)} + \beta_{\underset{\text{team } j \in n}{\text{off}(i)}} - \beta_{\underset{\text{team } s \in n}{\text{def}(i)}} + \varepsilon_i$$

consisting of each half inning in the dataset.

Then, for each $m \in \{1, .., M=100\}$, simulate a "true" outcome vector $y^{(m)}$ according to $y_i^{(m)} \sim \text{Round}(\mathcal{N}_+(x_i \cdot \beta, 1))$, where $\mathcal{N}_+$ is the normal distribution truncated to be positive.

Our goal is to recover the park effects $\beta^{(\text{park})}$ from the simulated data $(X, y^{(m)})$.

For each m, use OLS and Ridge Regression to estimate $\beta^{(\text{park}, m)}$, yielding the vectors $\hat{\beta}^{(m, \text{park}, \text{OLS})}$ and $\hat{\beta}^{(m, \text{park}, \text{Ridge})}$.

Then, estimate $\mathbb{E}_{\mathcal{D}}[\hat{\beta}^{(\text{park}, \text{OLS})}]$ by $\frac{1}{M} \sum_{m=1}^{M} \hat{\beta}^{(\text{park}, \text{OLS}, m)}$,

and then estimate the avg. bias by $\| \beta^{(\text{park})} - \hat{\mathbb{E}}_{\mathcal{D}}[\hat{\beta}^{(\text{park}, \text{OLS})}] \|_2$ where $\|\cdot\|_2$ is the 2-norm.

Also estimate the bias for Ridge.

Do something similar to estimate variance.

Compare OLS to Ridge via estimated bias & variance.