WSABI Summer Research Lab **Summer 2025**

Lab 16: Bias-Variance Tradeoff

Instructor: Jonathan Pipping *Author: Ryan Brill*

16.1 Park Effects Simulation Study

16.1.1 Motivation

It would be great to compute the bias and variance

$$\text{Bias}(\widehat{f}) = (f - \mathbb{E}_{\mathcal{D}}[\widehat{f}])$$
$$\text{Var}(\widehat{f}) = \mathbb{E}_{\mathcal{D}}[\widehat{f}^2] - (\mathbb{E}_{\mathcal{D}}[\widehat{f}])^2$$

for real-world estimators $\widehat{f}$ of real-world sports datasets $\mathcal{D}$, but this is impossible because the "true" function f is unknown and $\mathbb{E}_{\mathcal{D}}$ is an expectation over the randomness of drawing dataset $\mathcal{D}$, which is impossible to calculate. To understand the nature of this tradeoff, we turn to a **simulation study**, using park effects as our example.

16.1.2 Simulation Setup

In this simulation study, we assume that the park effects, team offensive quality, and team defensive quality have known coefficients. You will use these coefficients to simulate y, the runs scored in each half-inning in the dataset. Then you will model your data using OLS and Ridge Regression, using bias and variance to compare the two models.

16.1.3 Your Task

1. Generate a "true" parameter vector $\boldsymbol{\beta}$ in `R` according to the following distributions:

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

2. Assemble your park effects data matrix $\boldsymbol{X}$ associated with the model

$$y_i = \beta_0 + \beta_{\text{park}(i)} + \beta_{\text{off}(i)} + \beta_{\text{def}(i)} + \epsilon_i$$

where each i is a half-inning.

3. Use a for loop to simulate a "true" outcome vector $y^{(m)}$ for $m = 1, \ldots, 100$ datasets according to the following formula:

$$y_i^{(m)} = \text{Round}\left(\mathcal{N}_+\left(x_i^T\boldsymbol{\beta}, 1\right)\right)$$

where $\mathcal{N}_+$ is the normal distribution truncated to be positive.

4. Your goal is to recover the park effects $\boldsymbol{\beta}^{(\text{park})}$ from the simulated data $(X, y^{(m)})$. Do this by fitting OLS and Ridge Regression models, which will give estimates $\widehat{\boldsymbol{\beta}}_{\text{OLS}}^{(m,\text{park})}$ and $\widehat{\boldsymbol{\beta}}_{\text{Ridge}}^{(m,\text{park})}$.

5. Then, estimate $\mathbb{E}_{\mathcal{D}}[\widehat{\boldsymbol{\beta}}_{\text{OLS}}^{(m,\text{park})}]$ and $\mathbb{E}_{\mathcal{D}}[\widehat{\boldsymbol{\beta}}_{\text{Ridge}}^{(m,\text{park})}]$ by averaging your estimates $\boldsymbol{\beta}^{(m)}$ over the 100 datasets. Then, estimate the average bias for each modelusing the following formula:

$$||\beta^{(\text{park})} - \mathbb{E}_{\mathcal{D}}[\widehat{\boldsymbol{\beta}}_{\text{model}}^{(m,\text{park})}]||_2$$

   where $||\cdot||_2$ is the Euclidean norm.

6. Estimate the variance similarly by averaging over the 100 datasets.

7. Compare OLS to Ridge Regression via estimated bias and variance. Visualize your results.