WSABI Summer Research Lab **Summer 2025**

Lab 13: Shrinkage Estimation

Instructor: Jonathan Pipping *Author: Ryan Brill*

13.1 Predicting Second-Half Putting Performance

13.1.1 Problem Setup

Our goal is to predict golfers' putting performance in the second half of the season from their performance in the first half. Unfortunately, we only have access to the data from the first half of this season, so **regression is not an option**. We will use the estimators we've learned in class to make these predictions.

The putters in our dataset are indexed by $i = 1, \ldots, k$. For each golfer, we have access to the following training data:

- X_i: the percentage of putts golfer i made in the first half of the season
- N_i: The number of putts golfer i took in the first half of the season

We will use each golfer's second-half putting performance (X_i') to evaluate our predictions, **not** to fit them. Each prediction will be a single number $\widehat{p}_i$ for each golfer i, and we will use the mean squared error (MSE) on our test data to evaluate our predictions.

$$\text{MSE} = \frac{1}{k}\sum_{i=1}^{k}(\widehat{p}_i - X_i')^2$$

13.1.2 Naïve Estimators

The first estimator we will consider is the overall mean of all golfers' first-half putting performance.

$$\widehat{p}_i^{(\text{mean})} = \frac{1}{k}\sum_{j=1}^{k} X_j \tag{13.1}$$

The second estimator we will consider is the maximum likelihood estimator (MLE) for each golfer

$$\widehat{p}_i^{(\text{MLE})} = X_i \tag{13.2}$$

Based on what you've learned in class, which estimator would you expect to perform better?

However, we also learned that the MLE can be improved by shrinking to the overall mean when estimating a large number of parameters. Two of these shrinkage estimators are the Efron-Morris and Empirical Bayes, and we will fit two versions of each. Recall from Lecture 12 that our original model takes the form

$$X_i \sim \mathcal{N}(p_i, \sigma_i^2) \text{ where } \sigma_i^2 = \frac{p_i(1-p_i)}{N_i}$$

where p_i is the latent "true" putter quality of golfer i. Our estimators work when the variance is **known**, but that is not the case here, and we will need a bit of work to estimate it.

13.1.3 Efron-Morris Version 1

We will begin by simplifying the problem, assuming that $\sigma_i^2 = \frac{C}{N_i}$ where C is a known constant. Specifically, we will use

$$C = \bar{X}(1-\bar{X}) \text{ where } \bar{X} = \frac{1}{k}\sum_{i=1}^{k} X_i$$

Now that we have a known variance, we can transform X_i to have unit variance. Define $\widetilde{X}_i$ as follows:

$$\widetilde{X}_i = \frac{X_i}{\sigma_i} = \frac{X_i}{\sqrt{C/N_i}}$$

Then we know that

$$\widetilde{X}_i \sim \mathcal{N}(\theta_i, 1) \text{ where } \theta_i = \frac{p_i}{\sigma_i} = \frac{p_i}{\sqrt{C/N_i}}$$

Now we have matched the assumptions for the Efron-Morris estimator, and we are ready to define our third estimator,

$$\widehat{p}_i^{(\text{EM 1})} = \sqrt{\frac{C}{N_i}}\widehat{\theta}_i^{(\text{EM 1})} \tag{13.3}$$

where $\widehat{\theta}_i^{(\text{EM 1})}$ is the Efron-Morris estimator for θ_i,

$$\widehat{\theta}_i^{(\text{EM 1})} = \bar{\widetilde{X}} + \left(1 - \frac{k-1}{S^2_{\widetilde{X}}}\right)\left(\widetilde{X}_i - \bar{\widetilde{X}}\right)$$

and $S^2_{\widetilde{X}}$ is the sum of squared deviations from the mean of $\widetilde{X}_i$, defined as

$$S^2_{\widetilde{X}} = \sum_{i=1}^{k}(\widetilde{X}_i - \bar{\widetilde{X}})^2$$

13.1.4 Efron-Morris Version 2

Now, instead of using the previous assumption of known variance, we'll use a **variance-stabilizing transformation** to make the variance known. Recalling the model from Lecture 12,

$$X_i \sim \mathcal{N}(p_i, \sigma_i^2) \text{ where } \sigma_i^2 = \frac{p_i(1-p_i)}{N_i}$$

our goal is to find a transformation T such that $T(X)$ has constant variance. To do this we write out a 1st-order Taylor approximation of X centered at p,

$$T(X) \approx T(p) + T'(p)(X-p)$$

Then we can write the variance of $T(X)$ as

$$\text{Var}\left(T\left(X\right)\right) = \text{Var}\left(T(p) + T'(p)(X-p)\right)$$

Since $T(p)$ is a constant, we know that

$$\text{Var}\left(T\left(X\right)\right) = \text{Var}\left(T'\left(p\right)\left(X-p\right)\right)$$

And by the rules of variance, we have

$$\text{Var}\left(T\left(X\right)\right) = \left[T'\left(p\right)\right]^2 \text{Var}\left(X\right)$$

Then, since we know the distribution of X, we can write

$$\text{Var}\left(T\left(X\right)\right) = \left[T'\left(p\right)\right]^2 \frac{p(1-p)}{N}$$

Our goal is to find a transformation T such that this variance is constant, and we can do this by setting

$$\begin{aligned} [T'(p)]^2 \frac{p(1-p)}{N} &= C \\ \implies T'(p) &= \sqrt{\frac{CN}{p(1-p)}} \end{aligned}$$

We proceed by solving this differential equation.

$$\begin{aligned} T(p) &= \int_0^p \sqrt{\frac{CN}{x(1-x)}}\, dx \\ &= \sqrt{CN} \int_0^p \frac{1}{\sqrt{x(1-x)}}\, dx \end{aligned}$$

We substitute $x = \sin^2(t)$ to get

$$\begin{aligned} T(p) &= \sqrt{CN} \int_0^{\arcsin(\sqrt{p})} \frac{2\sin(t)\cos(t)}{\sqrt{\sin^2(t)(1-\sin^2(t))}}\, dt \\ &= \sqrt{CN} \int_0^{\arcsin(\sqrt{p})} \frac{2\sin(t)\cos(t)}{\sqrt{\sin^2(t)\cos^2(t)}}\, dt \\ &= \sqrt{CN} \int_0^{\arcsin(\sqrt{p})} \frac{2\sin(t)\cos(t)}{\sin(t)\cos(t)}\, dt \\ &= \sqrt{CN} \int_0^{\arcsin(\sqrt{p})} 2\, dt \\ &= \underbrace{2\sqrt{CN}}_{\text{constant}} \arcsin\sqrt{p} \end{aligned}$$

Then $T(p) = \arcsin\sqrt{p}$ is a variance-stabilizing transformation, and we use it to transform X_i to have constant variance. We define $\widetilde{X}_i$ as follows

$$\widetilde{X}_i = T(X_i) = \arcsin\sqrt{X_i} = \arcsin\sqrt{\frac{H_i}{N_i}}$$

where H_i is the number of putts golfer i made in the first half of the season. In practice, we will use a slightly-different transformation that also yields a constant variance,

$$\widetilde{\widetilde{X}}_i = \arcsin\sqrt{\frac{H_i + 3/8}{N_i + 3/4}}$$

Then it follows that

$$\widetilde{\widetilde{X}}_i \sim \mathcal{N}(\widetilde{\theta}_i, \nu_i^2) \text{ where } \widetilde{\theta}_i = \arcsin\sqrt{p_i} \text{ and } \nu_i^2 = \frac{1}{4N_i}$$

And finally, we transform $\widetilde{\widetilde{X}}_i$ to have unit variance,

$$\widetilde{X}_i = \frac{\widetilde{\widetilde{X}}_i}{\nu_i} \sim \mathcal{N}(\widetilde{\theta}_i, 1) \text{ where } \widetilde{\theta}_i = \frac{\widetilde{\widetilde{\theta}}_i}{\nu_i}$$

We are now ready to define our fourth estimator,

$$\widehat{p}_i^{(\text{EM 2})} = \sin^2\!\big(\nu_i\, \widetilde{\theta}_i^{(\text{EM 2})}\big) \tag{13.4}$$

where $\widetilde{\theta}_i^{(\text{EM 2})}$ is the Efron-Morris estimator for $\widetilde{\theta}_i$,

$$\widetilde{\theta}_i^{(\text{EM 2})} = \bar{\widetilde{X}} + \left(1 - \frac{k-1}{S^2_{\widetilde{X}}}\right)\left(\widetilde{X}_i - \bar{\widetilde{X}}\right)$$

and $S^2_{\widetilde{X}}$ is the sum of squared deviations from the mean of $\widetilde{X}_i$, defined as

$$S^2_{\widetilde{X}} = \sum_{i=1}^{k} (\widetilde{X}_i - \bar{\widetilde{X}})^2$$

13.1.5 Empirical Bayes Version 1

Suppose the following distribution for X_i:

$$X_i \mid p_i \sim \mathcal{N}(p_i, \frac{C}{N_i}) \text{ where } C = \bar{X}(1-\bar{X}) \text{ is a known constant}$$

To complete this setup, we add a prior distribution for p_i,

$$p_i \sim \mathcal{N}(\mu, \tau^2)$$

Then the Bayesian estimate for p_i is the posterior mean, defined as

$$\widehat{p}_i^{(\text{Bayes})} = \mathbb{E}\left[p_i \mid X_i\right] = \mu + \frac{\tau^2}{\tau^2 + \frac{C}{N_i}}\left(X_i - \mu\right)$$

However, we know that μ and τ^2 are unknown, so we need to estimate them to get the **Empirical Bayes** estimator we saw in Lecture 12.

$$\widehat{p}_i^{(\text{EB})} = \widehat{\mu} + \frac{\widehat{\tau}^2}{\widehat{\tau}^2 + \frac{C}{N_i}}\left(X_i - \widehat{\mu}\right)$$

One way to do this is to find the MLE of μ and τ^2 using the iterative convergence method in Lecture 12, but this is a bit of overkill. Instead, we will consider the marginal distribution of $\boldsymbol{X}$. From our model setup, we know that

$$X_i \mid p_i \sim \mathcal{N}(p_i, \frac{C}{N_i}) \text{ and } p_i \sim \mathcal{N}(\mu, \tau^2)$$

Then the marginal distribution of X_i is

$$X_i \sim \mathcal{N}(\mu, \tau^2 + \frac{C}{N_i})$$

Then the estimator $\widehat{\mu} = \bar{X}$ is an unbiased estimator of μ. To estimate τ^2, we first rearrange the variance of X_i to get

$$\begin{aligned} \mathrm{Var}(X_i) &= \tau^2 + \frac{C}{N_i} \\ \implies \tau^2 &= \mathrm{Var}(X_i) - \frac{C}{N_i} \end{aligned}$$

So an unbiased estimator of τ^2 is

$$\widehat{\tau}^2 = S_X^2 - C \cdot \mathrm{Mean}(\frac{1}{N_i})$$

where S_X^2 is the sample variance of X_i, defined as

$$S_X^2 = \frac{1}{k-1} \sum_{i=1}^{k} (X_i - \bar{X})^2$$

Then, we define our fifth estimator,

$$\widehat{p}_i^{(\text{EB 1})} = \widehat{\mu} + \frac{\widehat{\tau}^2}{\widehat{\tau}^2 + \frac{C}{N_i}} (X_i - \widehat{\mu}) \tag{13.5}$$

13.1.6 2nd Empirical Bayes Estimator

Once again, we will consider a second estimator that uses a variance-stabilizing transformation to make the variance known. We will use the same transformation as in the Efron-Morris Version 2,

$$\widetilde{X}_i = \arcsin \sqrt{\frac{H_i + 3/8}{N_i + 3/4}}$$

Then it follows that

$$\widetilde{X}_i \sim \mathcal{N}(\widetilde{\theta}_i, \nu_i^2) \text{ where } \widetilde{\theta}_i = \arcsin\sqrt{p_i} \text{ and } \nu_i^2 = \frac{1}{4N_i}$$

We define our prior distribution for $\widetilde{\theta}_i$ as

$$\widetilde{\theta}_i \sim \mathcal{N}(\mu, \tau^2)$$

Then our posterior mean for $\widetilde{\theta}_i$ is

$$\widehat{\widetilde{\theta}}_i^{(\text{EB 2})} = \mu + \frac{\tau^2}{\tau^2 + \nu_i^2} \left(\widetilde{X}_i - \mu \right)$$

And we can estimate the unknown parameters μ and τ^2 using the same method as in the Empirical Bayes Version 1:

$$\begin{aligned} \mu &= \bar{\widetilde{X}} \\ \tau^2 &= S_{\widetilde{X}}^2 - \overline{\nu_i^2} \end{aligned}$$

where $S_{\widetilde{X}}^2$ is the sample variance of $\widetilde{X}_i$, defined as

$$S_{\widetilde{X}}^2 = \frac{1}{k-1} \sum_{i=1}^{k} (\widetilde{X}_i - \bar{\widetilde{X}})^2$$

and $\overline{\nu_i^2}$ is the mean of the ν_i^2s. Then we can estimate $\widetilde{\theta}_i$ and our final estimator is

$$\widehat{p}_i^{(\text{EB 2})} = \sin^2(\widehat{\widetilde{\theta}}_i^{(\text{EB 2})}) \tag{13.6}$$

13.1.7 Your Task

1. Fit the 6 estimators defined in Equations (13.1), (13.2), (13.3), (13.4), (13.5), and (13.6) to the data.
2. Using the second-half putting performance data, compute the out-of-sample loss for each estimator, which we defined as the mean squared error (MSE).
3. Rank the estimators by their MSE. Which was the best? Which was the worst?
4. Make a scatterplot of the true second-half putting performances (X_i') against your predicted estimates ($\widehat{p}_i$) for each of the 6 estimators on the same plot, using different colors for each estimator. Make sure to include an identity line $y = x$ to show what perfect predictions would look like.