Empirical Bayes

Q1 Suppose Mookie Betts' batting average midway through the season is .300. Using no other information, predict his end-of-season batting average.

Model mid-season batting average

$H = \#$ hits

$N = \#$ at-bats

$$\frac{H}{N} \sim \frac{1}{N} \text{Binomial}(N, p)$$

the MLE of p is $\hat{p}_{MLE} = \frac{H}{N}$

Last week we talked about using a Prior but we said "no other information" so we can't quite do that here.

Q2 Suppose we know each players' batting average midway through the 2023 season. Using no ~~other~~ info from any previous season, i.e. only using these 2023 mid-season BA predict each players' end-of-season batting average

One approach

i index player

H_i = # hits for player i midway thru season

N_i = # at bats

p_i = player i's "true" hitting quality

Model $BA_i = \frac{H_i}{N_i} \sim \frac{1}{N_i}\,\text{Binomial}(N_i, p_i)$

MLE $\hat{p}_i^{(MLE)} = \frac{H_i}{N_i}$

Can we do better?

Previous idea use a prior / shrink to something

But all we have access to is this season's 2023 midway Batting Average data, no data from previous seasons,

What can we shrink to?

Use data from all the players to create a prior and shrink to the overall mean baseball player.

Mookie Betts himself is a baseball player

Model

i index player

H_i = # hits for player i midway thru season

N_i = # at bats

p_i = player i's "true" hitting quality

Model $X_i = \frac{H_i}{N_i} \sim \frac{1}{N_i}\,\text{Binomial}(N_i, p_i)$

data: batting average midway thru the season

Remove all players with low N_i (say $N_i < 25$)

CLT $X_i \approx \mathcal{N}\left(p_i, \frac{p_i(1-p_i)}{N_i}\right)$

Simplification

p_i unknown

we are interested in estimating it

it will be a lot easier if the variance were known

Let $\sigma_i^2 = \frac{C}{N_i}$ $\rightarrow C = (.24)(1-.24) = .18$

Bayesian Model

likelihood $\quad X_i \sim \mathcal{N}\left(p_i, \frac{c}{\underbrace{N_i}_{\text{known}}}\right) \qquad c = .18$

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

there is a distribution on the talent of baseball players and we suppose that distribution is Normal with some mean μ and some variance σ^2

μ = average batting talent
τ^2 = variance of batting talent

MLE: ignore the prior, only using the info relevant to player i,

$$\boxed{\hat{p}_i^{(MLE)} = X_i}$$

Bayesians: use the prior!
The Bayesian estimate is the posterior mean

$$\begin{cases} X_i \sim \mathcal{N}(p_i, \frac{c}{N_i}) & \text{likelihood} \\ p_i \sim \mathcal{N}(\mu, \tau^2) & \text{prior} \end{cases}$$

$$\hat{p}_i^{(Bayes)} = \overbrace{\mathbb{E}[p_i | X_i]}^{\mathbb{E}(\text{param} | \text{data})}$$

to calculate the posterior mean we'll need to know the posterior dist:

$$\mathbb{P}(p_i | X_i) = \frac{\mathbb{P}(X_i | p) \cdot \mathbb{P}(p_i)}{\mathbb{P}(X_i)} \quad \text{Bayes Rule}$$

proportional to

$$\propto \mathbb{P}(X_i | p_i) \cdot \mathbb{P}(p_i)$$

$$= \mathbb{P}\left(\mathcal{N}(p_i, \tfrac{c}{N_i}) = X_i\right) \cdot \mathbb{P}\left(\mathcal{N}(\mu, \tau^2) = p_i\right)$$

$$\propto \exp\left(-\frac{1}{2} \frac{(X_i - p_i)^2}{c/N_i}\right) \cdot \exp\left(-\frac{1}{2} \frac{(\mu - p_i)^2}{\tau^2}\right)$$

$$\propto \exp\left(-\frac{1}{2}\left[\cancel{\frac{x_i^2}{c/N_i}} - \frac{2p_i x_i}{c/N_i} + \frac{p_i^2}{c/N_i} + \cancel{\frac{\mu^2}{\tau^2}} - \frac{2p_i\mu}{\tau^2} + \frac{p_i^2}{\tau^2}\right]\right)$$

$$= \exp\left(-\frac{1}{2}\left[p_i^2\left(\frac{1}{c/N_i}+\frac{1}{\tau^2}\right) - 2p_i\left(\frac{x_i}{c/N_i}+\frac{\mu}{\tau^2}\right)\right]\right)$$

$$= \exp\left(-\frac{1}{2}\left(\frac{1}{c/N_i}+\frac{1}{\tau^2}\right)\left[p_i^2 - 2p_i\frac{\left(\frac{x_i}{c/N_i}+\frac{\mu}{\tau^2}\right)}{\left(\frac{1}{c/N_i}+\frac{1}{\tau^2}\right)}\right]\right)$$

$$\propto \exp\left(-\frac{1}{2}\left(\frac{1}{c/N_i}+\frac{1}{\tau^2}\right)\cdot\left[p_i - \frac{\left(\frac{x_i}{c/N_i}+\frac{\mu}{\tau^2}\right)}{\left(\frac{1}{c/N_i}+\frac{1}{\tau^2}\right)}\right]^2\right)$$

$$\Longrightarrow \mathbb{P}(p_i|x_i) \propto \exp\left(-\frac{1}{2}\frac{\left[p_i - \frac{\left(\frac{x_i}{c/N_i}+\frac{\mu}{\tau^2}\right)}{\left(\frac{1}{c/N_i}+\frac{1}{\tau^2}\right)}\right]^2}{\frac{1}{\left(\frac{1}{c/N_i}+\frac{1}{\tau^2}\right)}}\right)$$

$Y \sim N(\mu, \sigma^2) \longrightarrow \mathbb{P}(Y=y) \sim \exp\left(-\frac{1}{2}\frac{(Y-y)^2}{\sigma^2}\right)$

$$\Longrightarrow \boxed{p_i|x_i \sim N\left(\frac{\left(\frac{x_i}{c/N_i}+\frac{\mu}{\tau^2}\right)}{\left(\frac{1}{c/N_i}+\frac{1}{\tau^2}\right)}, \frac{1}{\left(\frac{1}{c/N_i}+\frac{1}{\tau^2}\right)}\right)}$$

Posterior distribution

Posterior mean

$$\hat{p}_i^{(Bayes)} = \mathbb{E}(p_i \mid X_i)$$

$$= \frac{\left(\frac{X_i}{C/N_i} + \frac{\mu}{\tau^2}\right)\left(\tau^2 \cdot \frac{c}{N_i}\right)}{\left(\frac{1}{C/N_i} + \frac{1}{\tau^2}\right)\left(\tau^2 \cdot \frac{c}{N_i}\right)}$$

$$= \frac{X_i \tau^2 + \mu \frac{c}{N_i}}{\tau^2 + \frac{c}{N_i}}$$

$$= \left(\frac{\tau^2}{\tau^2 + \frac{c}{N_i}}\right) X_i + \mu \left(\frac{C/N_i}{\tau^2 + C/N_i}\right)$$

$$= \left(\frac{\tau^2}{\tau^2 + \frac{c}{N_i}}\right) X_i + \mu \left(1 - \frac{\tau^2}{\tau^2 + \frac{c}{N_i}}\right)$$

$$= \mu + \left(\frac{\tau^2}{\tau^2 + \frac{c}{N_i}}\right)(X_i - \mu).$$

$$\hat{p}_i^{(\text{Bayes})} = \frac{\dfrac{X_i}{C/N_i} + \dfrac{\mu}{\tau^2}}{\dfrac{1}{C/N_i} + \dfrac{1}{\tau^2}}$$

$$\begin{cases} X_i \sim N\left(p_i, \frac{c}{N_i}\right) & \text{likelihood} \\ p_i \sim N(\mu, \tau^2) & \text{prior} \end{cases}$$

$$\hat{p}_i^{(\text{Bayes})} = \mu + \underbrace{\left(\frac{\tau^2}{\tau^2 + \frac{c}{N_i}}\right)}_{\text{some fraction } < 1}(X_i - \mu)$$

a form of regression to the mean

the Bayes estimate is a weighted average of the data and the PRIOR.

the $\frac{1}{C/N_i}$ and $\frac{1}{\tau^2}$ terms dictate the relative weighting of the prior to the data in our weighted average

$$\tau^2 = 0, \quad p_i \sim N(\mu, 0), \quad \hat{p}_i = \mu$$

$$\tau^2 = \infty, \quad p_i \sim N(\mu, \infty), \quad \hat{p}_i = X_i$$

The weight of μ in $\hat{p}_i$ is $\frac{1}{\tau^2}$

$$\begin{cases} X_i \sim N(p_i, \frac{c}{N_i}) & \text{likelihood} \\ p_i \sim N(\mu, \tau^2) & \text{prior} \end{cases}$$

Problem : μ, τ^2 are unknown hyperparameters

Empirical Bayes : estimate μ and τ^2 from data
plug in $\hat{\mu}$ and $\hat{\tau}^2$ into $\hat{p}$

$$\boxed{\hat{p}_i^{(Bayes)} = \mu + \left(\frac{\tau^2}{\tau^2 + \frac{c}{N_i}}\right)(X_i - \mu)}$$

$$\boxed{\hat{p}_i^{(Empirical\ Bayes)} = \hat{\mu} + \left(\frac{\hat{\tau}^2}{\hat{\tau}^2 + \frac{c}{N_i}}\right)(X_i - \hat{\mu})}$$

Estimate μ, τ^2 from the data $\{X_i, N_i\}$

1. MLE: Find $\hat{\mu}^{(MLE)}$, $\hat{\tau}^{2(MLE)}$

   $\hookrightarrow$ a bit complicated and unnecessarily difficult

2. Tricks

$$\begin{cases} X_i \sim N\left(p_i, \frac{c}{N_i}\right) & \text{likelihood} \\ p_i \sim N(\mu, \tau^2) & \text{prior} \end{cases}$$

Marginal dist $X_i \sim N\left(\mu, \tau^2 + \frac{c}{N_i}\right)$

↳ integrate out p_i

$$P(X_i \mid \mu, \tau^2) = \int_0^1 P(X_i \mid p_i) \cdot P(p_i \mid \mu, \tau^2)\, dp_i$$

$$\begin{cases} \hat{\mu} = \text{Mean}(\{X_i\}) \\ \hat{\tau}^2 = \cancel{\text{Var}(\{X_i\}) - \frac{c}{N}} \end{cases}$$

$$\cancel{\text{var}(\{X_i\}) - \frac{c}{\text{Mean}(\{N_i\})}}$$

$$\text{var}(\{X_i\}) - \text{Mean}\left(\left\{\frac{c}{N_i}\right\}\right)$$

$\text{var}(\{X_i\}) = \tau^2 + \frac{c}{N_i}$

$\tau^2 = \text{var}(\{X_i\}) - \frac{c}{N_i}$

$\tau^2 = \text{mean}(\tau^2) = \text{mean}\left(\left\{\text{var}(X_i) - \frac{c}{N_i}\right\}\right)$

$= \text{var}(\{X_i\}) - \text{mean}\left(\left\{\frac{c}{N_i}\right\}\right)$

$$\boxed{\hat{p}_i^{(\text{Empirical Bayes})} = \hat{\mu} + \left(\frac{\hat{\tau}^2}{\hat{\tau}^2 + \frac{c}{N_i}}\right)(X_i - \hat{\mu})}$$

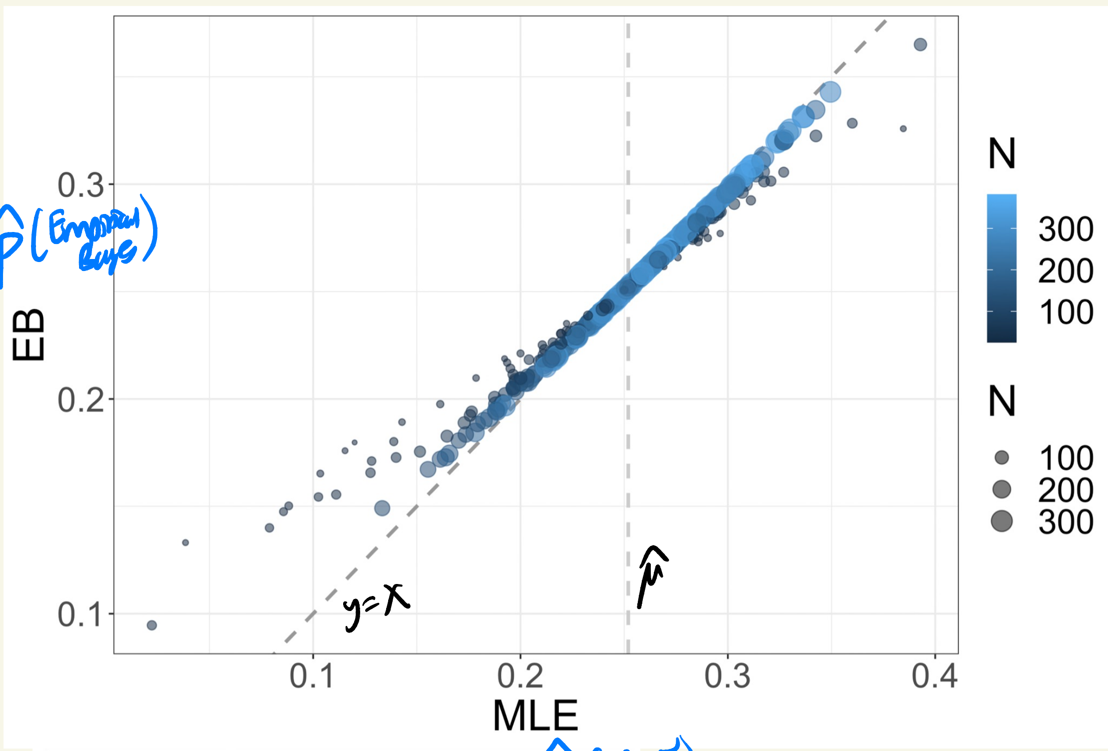
$$\hat{\mu} = \text{mean}(\{X_i\}) = \frac{1}{N}\sum_{i=1}^{N} X_i$$

$$\boxed{\hat{p}_i^{(MLE)} = X_i}$$

$N = \#\text{players}$

$$\hat{\tau}^2 = \text{var}(\{X_i\}) - \text{mean}\left(\left\{\frac{c}{N_i}\right\}\right)$$

$$= \frac{1}{N-1}\sum_{i=1}^{N}\left(X_i - \frac{1}{N}\sum_{i=1}^{N} X_i\right)^2 - \frac{1}{N}\sum_{i=1}^{N}\frac{c}{N_i}$$



$\hat{p}^{(\text{Empirical Bayes})}$

$\hat{p}^{(MLE)}$

We have 2 estimates of p_i : $\hat{p}_i^{(MLE)}$, $\hat{p}_i^{(EB)}$

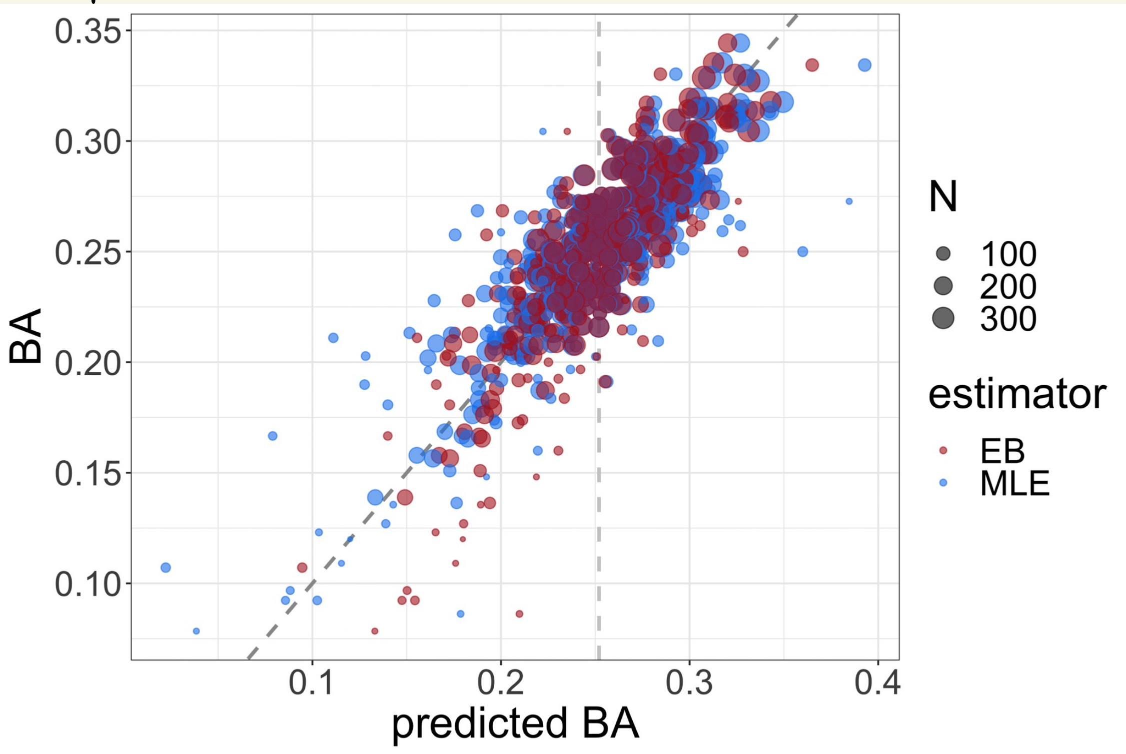
Evaluate which is better using out-of-sample prediction

Predict BA in 2^{nd} half of the season, X_i'

$$RMSE(\hat{p}, X') = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{p}_i - X_i')^2}$$

RMSE $\hat{p}^{(MLE)}$ → .026

RMSE $\hat{p}^{(EB)}$ → .023



<u>Takeaway</u>

- Shrinkage (to the overall mean) helps prediction especially when sample sizes are smaller
- Sharing information across individuals helps