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 is

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

H = # hits, N = # at-bats

As discussed previously, the MLE of a binomial is

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

and this is our prediction.

Concretely, we predict Mookie's end-of-season batting average to be .300.

Yesterday we used a prior, but we can't do that here since we have no other information.

Q2 Suppose we know each player's batting average midway through the 2023 season. Using no information from any previous season, i.e. only using these 2023 mid-season batting averages, predict each player's end-of-season batting average.

One approach:

Notation player index i
player i's # hits H_i
player i's # at-bats N_i
player i's batting average $\frac{H_i}{N_i}$

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

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

Can we do better??

Previous idea: shrinkage

But, we only have access to batting averages from 2023 and nothing else.

So, we can't shrink to prior information from previous seasons.

What can we shrink to??

Well, we have the midseason batting average of each baseball player, so perhaps we can pool information and shrink to the overall mean batting average.

Idea: to predict the batting average of Mookie Betts, use the batting averages of every other player!

Insight: Mookie Betts is a baseball player
use that information!

Notation player index i
player i's # hits H_i
player i's # at-bats N_i
player i's batting average $X_i = \frac{H_i}{N_i}$

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

Remove players from the dataset with small N_i (say, $N_i < 25$).
Since N_i large,

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

by Central Limit Theorem.

Simplification

p_i unknown, so variance $\sigma_i^2 = \frac{p_i(1-p_i)}{N_i}$ unknown
Much easier to work with known variance
so let's just assume for simplicity
that $\sigma_i^2 = \frac{C}{N_i}$ for some known
constant C.

We'll use $C = 0.035$, treat this as a tuning hyperparameter... (tune using a previous season).

In practice, we use a variance stabilizing transform $h(X)$ which transforms the batting average so that it has a known variance...

Now, $X_i \sim N(\mu, \sigma_i^2)$ σ_i^2 known...

Parametric Bayesian Model

$$\text{PRIOR:} \begin{cases} X_i \sim N(\mu_i, \sigma_i^2) \\ \mu_i \sim N(\mu, \tau^2) \end{cases}$$

insight: each player i is a baseball player
so, let's pool information!
there is a mean μ across all baseball players μ_i and some s.d. τ

$\mu_i = p_i$ unknown $\quad \sigma_i = \frac{\sqrt{C}}{\sqrt{N_i}}$ known

μ, τ unknown

MLE: ignore the prior, only use the information relevant to player i

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

observed mid-season batting average

Bayesians: use the PRIOR!

The Bayesian estimate is the posterior mean, i.e. the mean of the dist of μ_i after seeing the data $\{X_i\}$,

$$\hat{\mu}_i^{(Bayes)} = \mathbb{E}(\mu_i | X_i)$$

need the posterior dist $\mu_i | X_i$,

$$\mathbb{P}(\mu_i | X_i) = \frac{\mathbb{P}(X_i | \mu_i)\,\mathbb{P}(\mu_i)}{\mathbb{P}(X_i)}$$ Bayes Rule

$$\propto \underbrace{\mathbb{P}(X_i | \mu_i)}_{N(\mu_i, \sigma_i^2)}\,\underbrace{\mathbb{P}(\mu_i)}_{N(\mu, \tau^2)}$$ since $\mathbb{P}(X_i)$ has no μ_i term

$$= P\left(N(\mu_i, \sigma_i^2) = x_i\right) \cdot P\left(N(\mu, \tau^2) = \mu_i\right)$$

$$= \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-1}{2}\frac{(x_i-\mu_i)^2}{\sigma_i^2}\right) \cdot \frac{1}{\sqrt{2\pi\tau^2}} \exp\left(\frac{-1}{2}\frac{(\mu_i-\mu)^2}{\tau^2}\right)$$

$$\propto \exp\left(\frac{-1}{2}\left[\frac{x_i^2}{\sigma_i^2} - \frac{2\mu_i x_i}{\sigma_i^2} + \frac{\mu_i^2}{\sigma_i^2} + \frac{\mu_i^2}{\tau^2} - \frac{2\mu_i\mu}{\tau^2} + \frac{\mu^2}{\tau^2}\right]\right)$$

$$\propto \exp\left(\frac{-1}{2}\left[\mu_i^2\left(\frac{1}{\sigma_i^2} + \frac{1}{\tau^2}\right) - 2\mu_i\left(\frac{x_i}{\sigma_i^2} + \frac{\mu}{\tau^2}\right)\right]\right)$$

$$\propto \exp\left(\frac{-1}{2}\left(\frac{1}{\sigma_i^2} + \frac{1}{\tau^2}\right)\left[\mu_i^2 - 2\mu_i \frac{\left(\frac{x_i}{\sigma_i^2} + \frac{\mu}{\tau^2}\right)}{\left(\frac{1}{\sigma_i^2} + \frac{1}{\tau^2}\right)}\right]\right)$$

$$\propto \exp\left(\frac{-1}{2}\left(\frac{1}{\sigma_i^2} + \frac{1}{\tau^2}\right)\left[\mu_i - \frac{\left(\frac{x_i}{\sigma_i^2} + \frac{\mu}{\tau^2}\right)}{\left(\frac{1}{\sigma_i^2} + \frac{1}{\tau^2}\right)}\right]^2\right)$$

$$\Longrightarrow \mu_i | x_i \sim N\left(\frac{\left(\frac{x_i}{\sigma_i^2} + \frac{\mu}{\tau^2}\right)}{\left(\frac{1}{\sigma_i^2} + \frac{1}{\tau^2}\right)}, \frac{1}{\left(\frac{1}{\sigma_i^2} + \frac{1}{\tau^2}\right)}\right)$$

$\Longrightarrow$ Posterior mean

$$\hat{\mu}_i^{(Bayes)} = \frac{\frac{X_i}{\sigma_i^2} + \frac{\mu}{\tau^2}}{\frac{1}{\sigma_i^2} + \frac{1}{\tau^2}} = \mu + \frac{\tau^2}{\tau^2 + \sigma_i^2}(X_i - \mu)$$

$X_i = \frac{H_i}{N_i}$, $\sigma_i^2 = \frac{C}{N_i}$

$\Longrightarrow$

$$\hat{\mu}_i^{(Bayes)} = \frac{\frac{H_i}{C} + \frac{\mu}{\tau^2}}{\frac{N_i}{C} + \frac{1}{\tau^2}}$$

* Looks a lot like $\frac{W+W'}{W+L+W'+L}$ from last lecture!

Problem: μ and τ are unknown quantities

Empirical Bayes: Estimate μ and τ from observed data!

$$\hat{\mu}_i^{(EB)} = \frac{\frac{H_i}{C} + \frac{\hat{\mu}}{\hat{\tau}^2}}{\frac{N_i}{C} + \frac{1}{\hat{\tau}^2}}$$

If $\tau^2 = 0$, $\mu_i \sim N(\mu, \tau^2) \approx N(\mu, 0) = \mu$

$\Rightarrow \hat{\mu}_i = \hat{\mu}$

If $\tau^2 = \infty$, $\mu_i \sim N(\mu, \infty) \approx \text{Uniform}(-\infty, \infty)$

and $\hat{\mu}_i = \frac{H_i}{N_i} = X_i = \hat{\mu}_i^{(MLE)}$

Else, $\hat{\mu}_i = \hat{\mu} + \frac{\hat{\tau}^2}{\hat{\tau}^2 + \sigma_i^2}(X_i - \hat{\mu})$

large σ_i^2 small N_i

is closer to $\hat{\mu}$ if $\frac{\hat{\tau}^2}{\hat{\tau}^2+\sigma_i^2}$ is small

is closer to X_i if it is large → small σ_i^2 large N_i

✳ use $\hat{\mu}^{(MLE)}$, $\hat{\tau}^{2(MLE)}$ to estimate μ, τ^2

Model $\begin{cases} X_i \sim N(\mu_i, \sigma_i^2) \\ \mu_i \sim N(\mu, \tau^2) \end{cases}$

Get Rid of μ_i

Marginal distribution $X_i \sim N(\mu, \tau^2 + \sigma_i^2)$

By Bayes Rule.

Log-Likelihood

$$L(\mu,\tau^2) = \log \mathbb{P}(X \mid \mu,\tau^2)$$

$$= \log \prod_{i=1}^{n} \mathbb{P}(X_i \mid \mu,\tau^2) \quad \text{by independence}$$

$$= \sum_{i=1}^{n} \log \mathbb{P}\left(\mathcal{N}(\mu,\tau^2+\sigma_i^2) = X_i\right)$$

$$= \sum_{i=1}^{n} \log \left[\frac{1}{\sqrt{2\pi(\tau^2+\sigma_i^2)}} e^{-\frac{(X_i-\mu)^2}{2(\tau^2+\sigma_i^2)}} \right]$$

$$= \sum_{i=1}^{n} \left\{ \log\left(\frac{1}{\sqrt{2\pi(\tau^2+\sigma_i^2)}} \right) - \frac{(X_i-\mu)^2}{2(\tau^2+\sigma_i^2)} \right\}$$

$$\propto -\frac{1}{2}\sum_i \log(\tau^2+\sigma_i^2) - \frac{1}{2}\sum_i \left(\frac{(X_i-\mu)^2}{\tau^2+\sigma_i^2}\right)$$

MLE $(\mu,\tau^2) = \underset{\mu,\tau^2}{\operatorname{argmin}} L(\mu,\tau^2) \Longrightarrow$ solve $\frac{dL}{d\mu}=0, \frac{dL}{d\tau^2}=0.$

$$\frac{\partial L}{\partial \mu} = \frac{1}{2}\sum_i \frac{2(X_i-\mu)}{\tau^2+\sigma_i^2} = 0$$

$$\Longrightarrow \mu = \frac{\sum_i X_i/(\tau^2+\sigma_i^2)}{\sum_i 1/(\tau^2+\sigma_i^2)}$$

$$\frac{\partial L}{\partial \tau^2} = -\frac{1}{2}\sum_i \frac{1}{\tau^2+\sigma_i^2} + \frac{1}{2}\sum_i \frac{(x_i-\mu)^2}{(\tau^2+\sigma_i^2)^2} = 0$$

$$\Longrightarrow \sum_i \frac{(x_i-\mu)^2}{(\tau^2+\sigma_i^2)^2} = \sum_i \frac{1}{\tau^2+\sigma_i^2}$$

Problem μ is in terms of τ^2 and τ^2 is in terms of μ, but we need both...

Iteratively Solve for $\hat{\mu}, \hat{\tau}^2$:

$$\hat{\mu} = \frac{\sum x_i/(\hat{\tau}^2+\sigma_i^2)}{\sum 1/(\hat{\tau}^2+\sigma_i^2)}$$

$$\sum \frac{(x_i-\hat{\mu})^2}{(\hat{\tau}^2+\sigma_i^2)^2} = \sum \frac{1}{\hat{\tau}^2+\sigma_i^2}$$

$\mu^{(0)}$ = educated guess

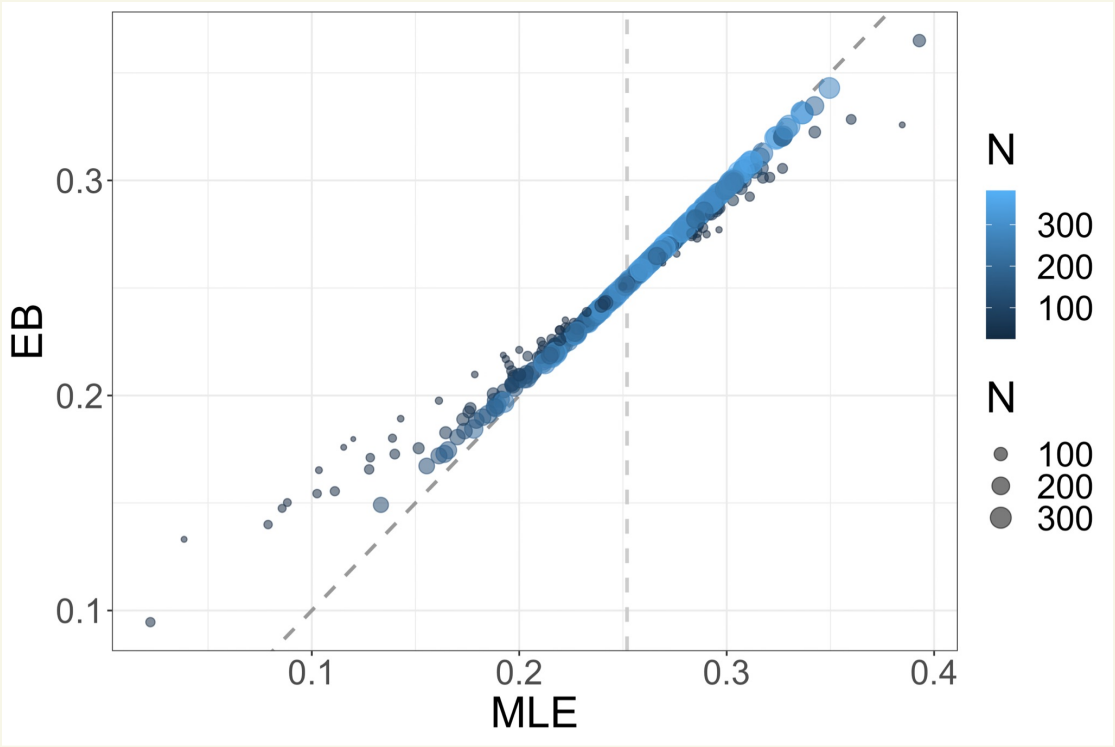
$\tau^{2(0)}$ = educated guess

while not $|\mu^{(t)} - \mu^{(t-1)}| \le \delta$ and $|\tau^{2(t)} - \tau^{2(t-1)}| < \delta$:

$$\mu^{(t)} \leftarrow \frac{\sum_i x_i/(\tau^2+\sigma_i^2)}{\sum_i 1/(\tau^2+\sigma_i^2)}$$

$\tau^{2(t)} \leftarrow$ the solution of $\sum_i \frac{(x_i-\mu)^2}{(\tau^2+\sigma_i^2)^2} = \sum_i \frac{1}{\tau^2+\sigma_i^2}$

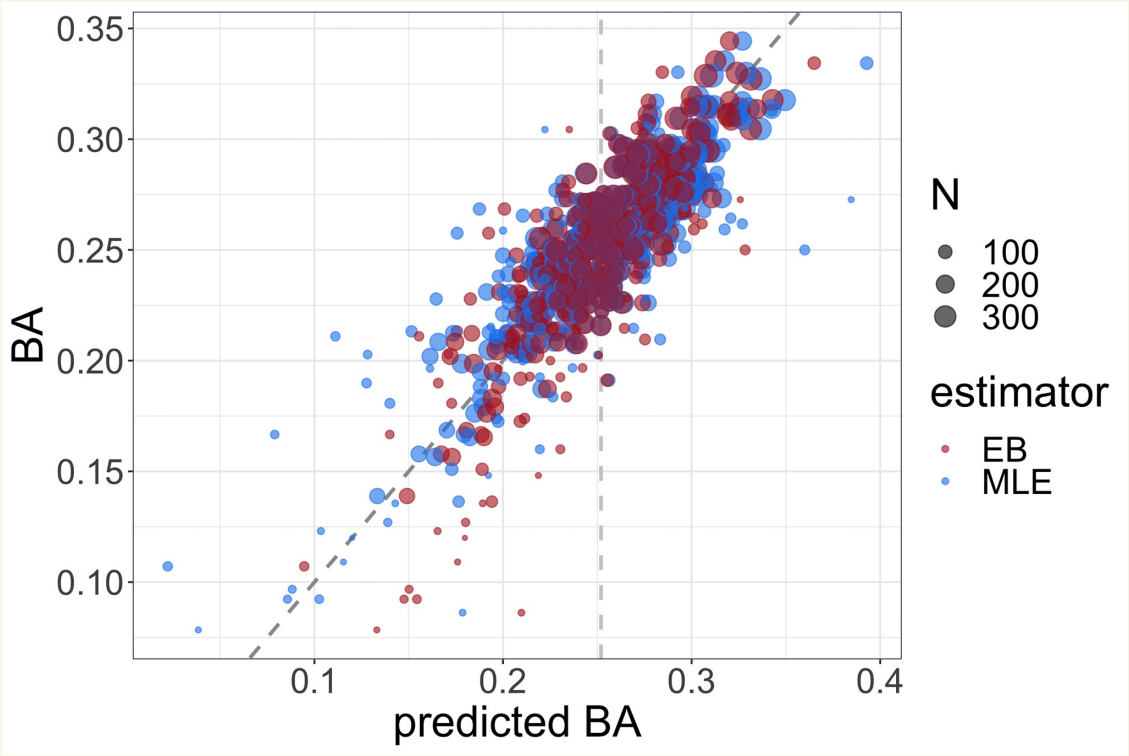
(use uniroot in R)



* Players with smaller N have $\hat{\mu}^{EB}$ shrunk towards the overall mean, while players with larger N have $\hat{\mu}^{(EB)} = \hat{\mu}^{(MLE)}$ their mid-szn BA

rmse_MLE	rmse_EB
0.02629828	0.02383808

* $\hat{\mu}^{(EB)}$ predict better the actual end-of-szn BA!



* There really isn't a huge diference b/t the predictions though...

Takeaway

- shrinkage towards the overall mean helps prediction when have smaller sample sites
- sharing information helps!

Consultants Dilemma (Stein's Paradox)

Billy Beane asks you to project the end-of-season performance of Elvis Andrus, using just Andrus' at bats alone, the MLE is best.

Using all MLB at bats, a shrinkage estimator is better than the MLE on average across all batters.

But, we only need a good prediction for one player, Andrus, not a good prediction on average over all players.

Which estimator do we use ???

An intuitive explanation is that optimizing for the mean-squared error of a *combined* estimator is not the same as optimizing for the errors of separate estimators of the individual parameters. In practical terms, if the combined error is in fact of interest, then a combined estimator should be used, even if the underlying parameters are independent. If one is instead interested in estimating an individual parameter, then using a combined estimator does not help and is in fact worse.

James Stein's Theorem A similar Empirical Bayes shrinkage estimator, the James Stein estimator $\hat{\mu}_i^{(JS)}$, everywhere dominates the MLE,

$$\mathbb{E}_\mu\left[\sum_{i=1}^{n}\left(\hat{\mu}_i^{(JS)} - \mu_i\right)^2\right] < \mathbb{E}_\mu\left[\sum_{i=1}^{n}\left(\hat{\mu}^{(MLE)} - \mu_i\right)^2\right],$$

i.e. it has smaller expected mean squared error. Expected because $\hat{\mu}_i$ is Random because X_i is Random.

Shrinkage, with Access to Previous Seasons' Data

Parametric Empirical Bayes Estimator:

$$\hat{\mu}_i^{(PEB)} = \hat{\mu} + \frac{\hat{\tau}^2}{\hat{\tau}^2 + \sigma_i^2}(X_i - \hat{\mu})$$

General Shrinkage Estimator:

$$\hat{\mu}_i = \hat{\mu} + \beta(X_i - \hat{\mu})$$

Procedure $\{X_i\}$ data

$\hat{\mu}$ = overall mean (sample mean of data from the season of datapoint i)

$\hat{\mu}_i$ = response column
= known end-of-szn batting avg of datapoint i (player i) from a previous szn

Estimate β using Regression, call it $\hat{\beta}$

Then, for future prediction,

$$\hat{\mu}_i = \hat{\mu} + \hat{\beta}(X_i - \hat{\mu}).$$