Empirical Bayes

→ player quality

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

Model Observe Mookie's mid-season data $\{H, N\}$

$$BA = \frac{H}{N} \sim \frac{1}{N} \cdot \underbrace{\text{Binomial}}_{\#\text{ successes (hits) in } N \text{ trials (at-bats)}}(N, p)$$

→ p is an unobserved measure of his quality

$H = \#\text{hits}$, $N = \#\text{at-bats}$

$$\hat{p}^{(MLE)} = \frac{H}{N} = BA = .300$$

$$\frac{W}{\underset{H}{W}+\underbrace{L}_{N-H}}$$

$$\frac{W+W'}{W+L+W'+L'}$$

* Yesterday we used a PRIOR to stabilize the MLE early in the season. But here we assumed use no other info, so we can't do that.

Q Suppose we know each player's batting average midway through the season. Using no info from previous seasons, predict each player's end-of-season Batting average.

Data $\{\underset{\#\text{hits}}{H_i}, \underset{\#\text{at bats}}{N_i}\}$ player i, $BA_i = \frac{H_i}{N_i}$

Model $$\frac{H_i}{N_i} \sim \frac{1}{N_i} \text{Binomial}(N_i, \underbrace{p_i}_{\text{unobservable measure of batter i's quality}})$$

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

Can we do better?

Previous idea

- Mon: Ridge → shrinkage → make your coeff. estimates smaller so as not to overfit
- Tues: predict end of szn WP → add prior / fake data to enhance our estimates

Q What are we going to shrink to?

towards the overall mean batting avg.

→ Mookie is a baseball player

How to accomplish this?

→ PRIOR

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

Binomial math can be difficult.

Central Limit theorem → the sum or avg. of iid random variables is approx. Normal

$$X_i = \frac{H_i}{N_i} = \frac{\text{Binomial}(N_i, P_i)}{N_i} = \frac{1}{N_i} \sum_{j=1}^{N_i} \begin{cases} 1 & \text{if hit} \\ 0 & \text{if not hit} \end{cases}$$

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

Math with Normal is easy!

$$\mathbb{E}X_i = \frac{1}{N_i} \mathbb{E}H_i = \frac{1}{N_i} \cdot \underbrace{N_i P_i}_{\mathbb{E}\,\text{Binomial}(N_i, P_i)} = P_i$$

$$\text{Var}(X_i) = \text{var}\left(\frac{H_i}{N_i}\right) = \frac{1}{N_i^2} \text{Var}(H_i)$$

$$= \frac{1}{N_i^2} \text{Var}(\text{Binom}(N_i, P_i)) = \frac{N_i P_i(1-P_i)}{N_i^2}$$

$$\text{Var}(cX_i) = c^2 \text{Var}(X_i) =$$

$$\mathrm{Var}(cX_i) = \mathbb{E}(cX_i)^2 - [\mathbb{E}(cX_i)]^2 =$$
$$= c^2 [\mathbb{E} X_i^2 - (\mathbb{E} X_i)^2) = c^2 \mathrm{Var}(X_i)$$

i^{th} batting avg. $X_i \approx N\left(p_i, \frac{p_i(1-p_i)}{N_i}\right)$

$$\longrightarrow \boxed{\begin{cases} X_i \sim N(p_i, \sigma_i^2) \\ p_i \sim N(p, \tau^2) \end{cases}} \longrightarrow$$

PRIOR: batter i is a baseball player whose quality is drawn from some dist across baseball players

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

Bayesian: MAP, $\underline{\text{posterior mean}}$

↓

posterior → compute $\mathbb{P}(p_i \mid X_i)$

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

Posterior:

$$\mathbb{P}(p_i \mid X_i) = \frac{\mathbb{P}(X_i \mid p_i)\,\mathbb{P}(p_i)}{\mathbb{P}(X_i)} \rightarrow \text{Bayes Rule}$$

proportional to $\leftarrow$

$$\propto \mathbb{P}(X_i \mid p_i)\,\mathbb{P}(p_i)$$

$$= \mathbb{P}\left(\mathcal{N}(p_i, \sigma_i^2) = X_i\right) \cdot \mathbb{P}\left(\mathcal{N}(p, \tau^2) = p_i\right.$$

$$= \frac{1}{\sqrt{\cancel{2\pi}\,\cancel{\sigma_i^2}}} \exp\left(-\frac{1}{2\sigma_i^2}(X_i - p_i)^2\right) \cdot$$

$$\frac{1}{\sqrt{\cancel{2\pi}\,\cancel{\tau^2}}} \exp\left(-\frac{1}{2\tau^2}(p - p_i)^2\right)$$

$$\propto \exp\left(-\frac{1}{2}\frac{1}{\sigma_i^2}(X_i - p_i)^2 - \frac{1}{2}\frac{1}{\tau^2}(p - p_i)^2\right)$$

$$\exp(a) \cdot \exp(b) = e^a \cdot e^b = e^{a+b} = \exp(a+b)$$

$$= \exp\left(-\frac{1}{2}\left[\frac{\underleftarrow{x_i^2 - 2x_i\rho_i + \rho_i^2}}{\sigma_i^2} + \frac{\rho^2 - 2\rho\rho_i + \rho_i^2}{\tau^2} \right] \right)$$

$$\propto \exp\left(-\frac{1}{2}\left[\rho_i^2\left(\frac{1}{\sigma_i^2} + \frac{1}{\tau^2}\right) - 2\rho_i\left(\frac{x_i}{\sigma_i^2} + \frac{\rho}{\tau^2}\right) + \cancel{\left(\frac{x^2}{\sigma_i^2} + \frac{\rho^2}{\tau^2}\right)} \right] \right)$$

$$= \exp\left(-\frac{1}{2}\left(\frac{1}{\sigma_i^2} + \frac{1}{\tau^2}\right)\left[\rho_i^2 - 2\rho_i\left(\frac{\frac{x_i}{\sigma_i^2} + \frac{\rho}{\tau^2}}{\frac{1}{\sigma_i^2} + \frac{1}{\tau^2}} \right) + \right] \right)$$

$$\propto \exp\left(-\frac{1}{2}\left(\frac{1}{\sigma_i^2} + \frac{1}{\tau^2}\right) \cdot \left(\rho_i - \left(\frac{\frac{x_i}{\sigma_i^2} + \frac{\rho}{\tau^2}}{\frac{1}{\sigma_i^2} + \frac{1}{\tau^2}} \right) \right)^2 \right)$$

$$(\rho_i - a)^2 = \rho_i^2 - 2\rho_i a + \cancel{a^2}$$

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

$$\propto \exp\left(-\frac{1}{2}\left(\frac{1}{\sigma_i^2}+\frac{1}{\tau^2}\right)\cdot\left(p_i - \left(\frac{\frac{X_i}{\sigma_i^2}+\frac{P}{\tau^2}}{\frac{1}{\sigma_i^2}+\frac{1}{\tau^2}}\right)\right)^2\right)$$

$$\mathbb{P}(p_i | \mathcal{N}(\mu, v)) = \frac{1}{\sqrt{2\pi v^2}} \exp\left(-\frac{1}{2}\frac{1}{v}(p_i-\mu)^2\right)$$

$$\Rightarrow \boxed{p_i | X_i \sim \mathcal{N}\left(\left(\frac{\frac{X_i}{\sigma_i^2}+\frac{P}{\tau^2}}{\frac{1}{\sigma_i^2}+\frac{1}{\tau^2}}\right), \frac{1}{\left(\frac{1}{\sigma_i^2}+\frac{1}{\tau^2}\right)}\right)}$$

We now have the post. dist p_i / X_i

Prior $\quad p_i \sim N(P, \tau^2)$

Posterior: updated our beliefs about p_i after having seen the data X_i.

Posterior mean:

X_i = player i's obs. batting avg

p = overall mean player quality

τ^2 = variability across batters
(variance of p_i over all batters)

σ_i^2 = variability within a batter

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

$$\boxed{\hat{p}_i^{(Bayes)} = \mathbb{E}(p_i \mid X_i) = \left(\frac{\frac{X_i}{\sigma_i^2} + \frac{p}{\tau^2}}{\frac{1}{\sigma_i^2} + \frac{1}{\tau^2}}\right) = p + \frac{\tau^2}{\tau^2+\sigma_i^2}(X_i - p)}$$

$$\left(\frac{\frac{X_i}{\sigma_i^2} + \frac{p}{\tau^2}}{\frac{1}{\sigma_i^2} + \frac{1}{\tau^2}}\right) \frac{\sigma_i^2\tau^2}{\sigma_i^2\tau^2} = \frac{X_i\tau^2 + p\sigma_i^2}{\sigma_i^2 + \tau^2}$$

$$= p\left(\frac{\sigma_i^2 + \tau^2 - \tau^2}{\sigma_i^2 + \tau^2}\right) + X_i\left(\frac{\tau^2}{\sigma_i^2 + \tau^2}\right)$$

$$= p + p\left(\frac{-\tau^2}{\sigma_i^2 + \tau^2}\right) + X_i\left(\frac{\tau^2}{\sigma_i^2 + \tau^2}\right)$$

If $\frac{\tau^2}{\tau^2+\sigma_i^2}=1$, then $\hat{p}_i^{Bayes} = \hat{p}_i^{MLE} = X_i$

but its <1, so $\hat{p}_i^{Bayes}$ is closer to p than $\hat{p}_i^{MLE}$

$$\begin{cases} X_i \sim N(p_i, \sigma_i^2) \\ p_i \sim N(p, \tau^2) \end{cases}$$

Model

$$\hat{P}_i^{(Bayes)} = \frac{\frac{X_i}{\hat{\sigma}_i^2} + \frac{\hat{p}}{\hat{\tau}^2}}{\frac{1}{\hat{\sigma}_i^{-2}} + \frac{1}{\hat{\tau}^2}}$$

observed X_i
estimate p_i

Problem: we never observed p, τ^2, σ_i^2

Empirical Bayes: estimate these other hyperparameters $\hat{p}, \hat{\tau}^2, \hat{\sigma}_i^2$ and plug them in!

$$\sigma_i^2 = \frac{p_i(1-p_i)}{N_i} = \frac{1}{N_i} \operatorname{var}(\operatorname{Binom}(N_i, P_i))$$

Simplifaction: $\sigma_i^2 = \frac{C}{N_i}$ C: some constant

$X_i = \frac{H_i}{N_i}$ $p_i(1-p_i) = \frac{1}{4} \cdot \frac{3}{4} = \frac{3}{16}$

$$\hat{p}_i^{\text{Bayes}} = \frac{\frac{H_i/N_i}{C/N_i} + \frac{\hat{p}}{\tau^2}}{\frac{1}{C/N_i} + \frac{1}{\tau^2}}$$

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

$$\boxed{\hat{p}_i^{(\text{Bayes})} = \frac{H_i + \hat{p}\frac{\hat{C}}{\hat{\tau}^2}}{N_i + 1 \cdot \frac{\hat{C}}{\hat{\tau}^2}}}$$

$\hat{\tau}^2 = 0 \longrightarrow \hat{p}$

$$\hat{\tau}^2 = \infty \longrightarrow \frac{H_i}{N_i}$$

$$\widehat{WP}^{MLE} = \frac{W}{N}, \quad \widehat{WP}^{Bayes} = \frac{W + W'}{N + N'}$$

as $N_i \uparrow$, $\hat{p}_i^{(Bayes)} \longrightarrow \hat{p}_i^{(MLE)}$

$$\boxed{\hat{p}_i^{(Bayes)} = \frac{H_i + \hat{p}\frac{\hat{c}}{\hat{\tau}^2}}{N_i + 1 \cdot \frac{\hat{c}}{\hat{\tau}^2}}}$$

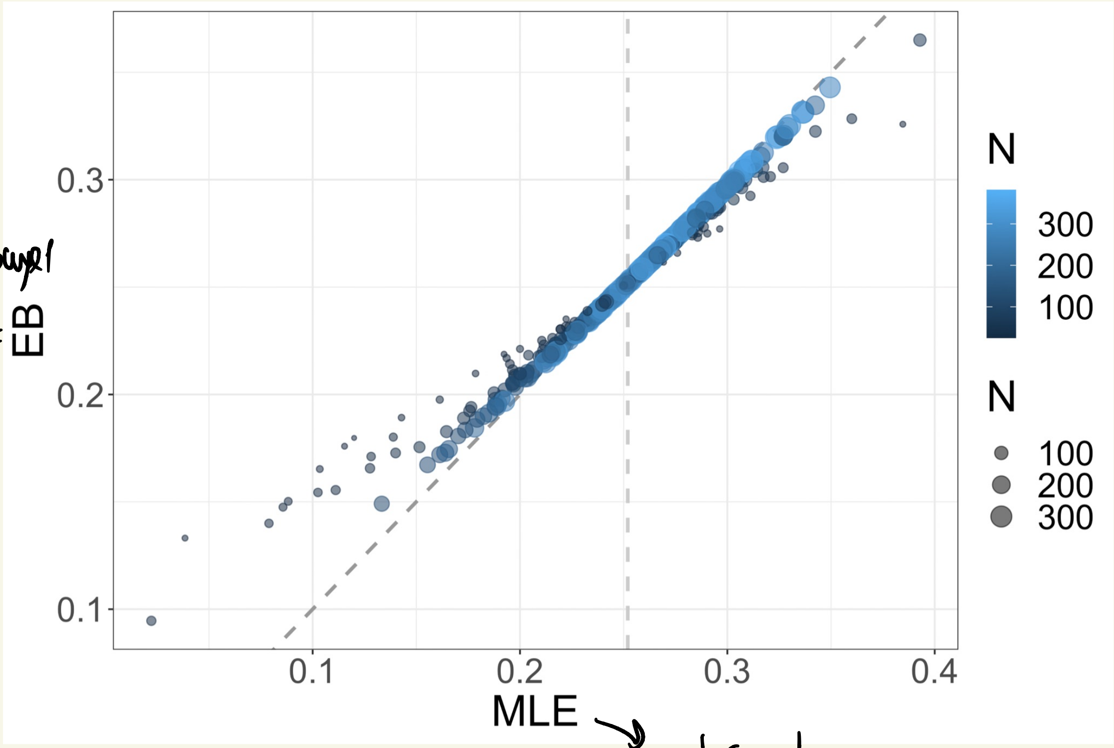
- estimate $\hat{p}, \hat{\tau}^2$ from data $\{X_i\}$ in a smart way

  e.g. $\hat{p} = \text{mean}(X_i)$

- in practice, there is no C b.c. that was a simplification

- treated c as a tuning parameter
  → chose the $\hat{c}$ which had best pred. perf. (lowest RMSE of end-of-szn p_i to $\hat{p}_i$)

$\hat{p}_i^{\text{Bayes}} =$



→ obs. batting average

rmse_MLE	rmse_EB
0.02629828	0.02383808

Takeaway

- Shrinkage — shrank obs. MLE batting avg. to overall mean, shrinking more if have fewer observations, to stabilize estimates and improve predictive performance

- Bayesian modeling — used prior to encode additional information

Consultant's Dilemma (Stein's Paradox)

James Stein's Theorem

Client index i, (unobserved) μ_i each client,

data $X_i \overset{iid}{\sim} N(\mu_i, 1)$ (known variance) the JS estimator $\hat{\mu}_i^{(JS)}$

dominates the MLE,

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

$$\downarrow$$

$$\hat{\mu}_i^{MLE} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Leaderboard:

predict $\hat{\mu}_i$

observe μ_i

score $= \sum_{i=1}^{n}(\hat{\mu}_i-\mu_i)^2$