Empirical Bayes

David Chan, MD, PhD

April 19, 2023

Introduction

- Researchers increasingly have access to detailed, large data with unit identifiers
- The units (e.g., physicians, hospitals) may matter for outcomes of interest
- Developments in computing power allow analysis accounting for units
- A recent body of work on empirical Bayes (EB) methods provides tools for analyzing *unit-specific parameters* when *observations per unit are finite*

Note: this cyberseminar borrows heavily from Gu (2022) and Walters (2022)

Poll

How familiar are you with empirical Bayes methods?

1. I have used empirical Bayes methods in my work
2. I have some understanding of empirical Bayes methods but have not used them in my work
3. I have only heard of the term empirical Bayes
4. I have not heard of the term empirical Bayes

Outline

Basic Setup

- Patients i, assigned to one of J physicians
  - Assignment function $j(i)$; $N_j = \sum_i \mathbf{1}(j(i) = j)$ is count of patients assigned to j

- $Y_i(j)$ is the outcome (e.g., spending) for patient i when assigned to physician $j \in \{1, \ldots, J\}$

- Simple additive model of potential outcomes:

$$Y_i(j) = \beta_j + \varepsilon_i$$

  - β_j is the **value-added** of physician j
  - $\beta_j - \beta_{j'}$ represents a treatment effect of being assigned to physician j instead of j'
  - ε_i represents other patient characteristics. Normalize $E[\varepsilon_i] = 0$.

Object of Interest

- The object of interest in this model is a *unit-specific parameter* β_j or the set $\{\beta_j\}_{j \in \{1,\ldots,J\}}$
- Contrast this with models for other empirical questions, e.g.,

$$Y_i = \beta D_i + \alpha_{j(i)} + \varepsilon_i$$

  - Object of interest may be the effect β of a treatment D_i
  - α_j would be a nuisance parameter
- When interested in $\{\beta_j\}_{j \in \{1,\ldots,J\}}$, important to ask how large is J and how many observations for each j
  - Fundamental issue in empirical Bayes: finite sample of observations for each j

Relevant Research and Policy Questions

- Questions
  - What is the value-added for a particular physician, e.g., β_1?
  - What does the distribution of $\{\beta_j\}_{j \in \{1,\ldots,J\}}$ look like? Are there outliers?
  - Which physicians can we classify as being high performers, e.g., the top 10%?

- With infinite observations for each physician, these questions would be trivial
  - In practice, we have finite observations, sometimes very few, for each physician
  - Empirical Bayes methods can provide tools to answering these questions

Outline

Value-Added Model Regression

- Recall the value-added model:

$$Y_i(j) = \beta_j + \varepsilon_i$$

- Assume that patients are randomly assigned to physicians
- Then can estimate value-added in an OLS regression:

$$Y_i = \beta_{j(i)} + \varepsilon_i$$

  - By random assignment, $\varepsilon_i \perp\!\!\!\perp \beta_{j(i)}$ (can relax this with more complicated causal models)

Fixed Effects and Random Effects

- Two statistical assumptions about $\beta_1, \ldots, \beta_J$:
  - Fixed effects: $\beta_1, \ldots, \beta_J$ are treated as unknown parameters
  - Random effects: $\beta_1, \ldots, \beta_J$ are treated as random variables with distribution G (i.e., $\beta_j \sim_{\text{i.i.d.}} G$)

- Two corresponding estimators:
  - Fixed effect estimator:
    $$\hat{\beta}_j^{FE}(\mathbf{Y}) = \frac{1}{N_j} \sum_i \mathbf{1}(j(i) = j)\, Y_i$$
    Uses only data corresponding to $j(i) = j$.
  - Empirical Bayes estimator with linear shrinkage (assume $E[\beta_j] = 0$):
    $$\hat{\beta}_j^{EB}(\mathbf{Y}) = \lambda(\mathbf{Y})\, \hat{\beta}_j^{FE}(\mathbf{Y})$$
    $\lambda(\mathbf{Y})$ uses all data.

Objective of Effect Estimation

- If we are only interested in one parameter (e.g., β_1), then we may have a loss function as

$$L\left(\beta_1, \hat{\beta}_1(\mathbf{Y})\right) = \left(\beta_1 - \hat{\beta}_1(\mathbf{Y})\right)^2$$

  I.e., our objective is to minimize the expected difference between β_1 and $\hat{\beta}_1$. In this case, $\hat{\beta}_1^* = \hat{\beta}_1^{FE}(\mathbf{Y})$, as $E\left[\hat{\beta}_1^{FE}(\mathbf{Y})\right] = \beta_1$.

- If we are interested in multiple parameters (e.g., $\beta_1, \ldots, \beta_J$), then we may have a loss function as

$$\frac{1}{J}\sum_{j=1}^{J} L\left(\beta_j, \hat{\beta}_j(\mathbf{Y})\right)$$

  The expectation of this is known as **compound risk**, and minimizing it is a **compound decision problem**

- Note: both of these are frequentist objectives

Theory

- Linear shrinkage estimator $\hat{\beta}_j^{EB}(\mathbf{Y}) = \lambda(\mathbf{Y})\,\hat{\beta}_j^{FE}(\mathbf{Y})$ (under $E[\beta_j] = 0$, can be relaxed) produces lower compound risk than the fixed-effect estimator for any $J \geq 3$ (James and Stein 1961)

- Shrinkage estimator is biased for any individual β_j (i.e., $E\left[\hat{\beta}_j^{EB}(\mathbf{Y})\right] \neq \beta_j$) in return for better average performance over $j \in \{1, \ldots, J\}$.

- Therefore, use empirical Bayes shrinkage when interested in performance of estimator over many units

- Empirical Bayes estimator uses all data:
  - "Borrowing strength from the ensemble" (Efron and Morris 1973; Morris 1983)
  - "Learning from the experience of others" (Efron 2012)

Philosophy of Random Effects

- Recall random effects definition: random variables with distribution G
- How do we think about the distribution G?
- Literal view: observed units are random draws from a larger population of units; may be unsatisfying depending on the context (e.g., units are VA hospitals)

Philosophy of Random Effects

- Pragmatic view: even with a fixed number of units (e.g., VA hospitals), empirical Bayes allows provides useful insights
  - $\{\beta_j\}_{j=1}^J$ imply a (discrete) distribution G even with a fixed set of j; continuous modeling of G can be viewed as a useful approximation
  - What is the best set of predictions of $\{\beta_j\}_{j=1}^J$ given the data? Other important policy-relevant questions can be illuminated by G
  - Distinction between random effects vs. fixed effects here is not about correlation with covariates but about focus on $\{\beta_j\}_{j=1}^J$

Outline

Implementing Empirical Bayes

- Shrinkage depends on the distribution G, which is unknown
- Empirical Bayes involves plugging in estimates of G from the data $\mathbf{Y}$ (recall the simple linear shrinkage $\lambda(\mathbf{Y})$)
- Parametric and non-parametric methods for approximating G

Parametric Normal-Normal Model

- Consider set of estimates of $\{\beta_j\}_{j=1}^J$ and corresponding standard errors: $\left\{\left(\hat{\beta}_j, s_j\right)\right\}_{j=1}^J$
- Assume normal-normal **hierarchical model**:

$$\begin{aligned} \beta_j &\sim N\left(\mu, \sigma^2\right) \\ \hat{\beta}_j \Big| \beta_j, s_j^2 &\sim N\left(\beta_j, s_j^2\right) \end{aligned}$$

- $G = N\left(\mu, \sigma^2\right)$ is a **mixing distribution**; $\hat{\beta}_j \Big| s_j \sim F_j = N\left(\mu, \sigma^2 + s_j^2\right)$ (F_j is a **mixture distribution**)
- **Deconvolution**: estimating G from $\left\{\left(\hat{\beta}_j, s_j\right)\right\}_{j=1}^J$
  - In parametric normal-normal model, this reduces to estimating **hyperparameters** μ and σ^2

Estimating Normal Hyperparameters

- Common approach:

$$\begin{aligned}\hat{\mu} &= \frac{1}{J}\sum_{j=1}^{J}\hat{\beta}_j \\ \hat{\sigma}^2 &= \frac{1}{J}\sum_{j=1}^{J}\left[\left(\hat{\beta}_j - \hat{\mu}\right)^2 - s_j^2\right]\end{aligned}$$

- Subtract out s_j^2 to account for sampling error in $\hat{\beta}_j$ relative to β_j

Shrinkage and Posterior Means

- Posterior mean of β_j, given known hyperparameters and $\left(\hat{\beta}_j, s_j\right)$:

$$\beta_j^* = E\left[\beta_j | \hat{\beta}_j, s_j\right] = \left(\frac{\sigma^2}{\sigma^2 + s_j^2}\right)\hat{\beta}_j + \left(\frac{s_j^2}{\sigma^2 + s_j^2}\right)\mu$$

- Shrinkage factor $\lambda = \frac{\sigma^2}{\sigma^2 + s_j} \in [0, 1]$ reflects signal-to-noise ratio

- Linear regression interpretation: λ is coefficient in linear regression of β_j on $\hat{\beta}_j \Rightarrow$ in class of linear functions, β_j^* minimizes MSE for even non-normal G

- Empirical Bayes posterior mean plugs in estimated hyperparameters $(\hat{\mu}, \hat{\sigma}^2)$:

$$\hat{\beta}_j^* = \left(\frac{\hat{\sigma}^2}{\hat{\sigma}^2 + s_j}\right)\hat{\beta}_j + \left(\frac{s_j^2}{\hat{\sigma}^2 + s_j}\right)\mu$$

Outline

Incorporating Unit Covariates

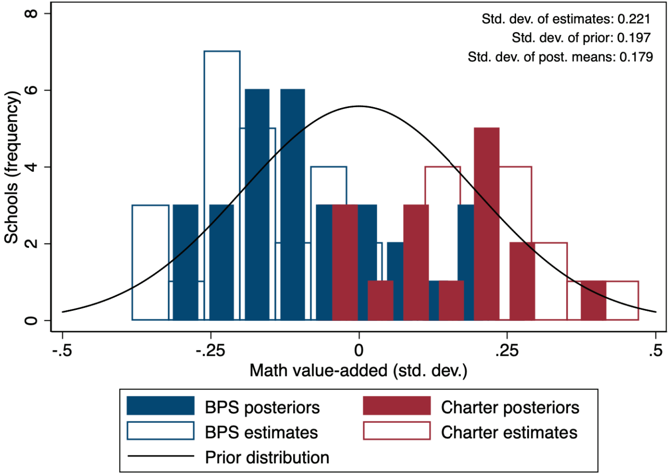
- May observe different groups of units (e.g., nurse practitioners and physicians) or characteristics of units (e.g., provider gender, age); can use these for shrinkage
  - Akin to asking *which other units* should we learn from (c.f., Efron 2012)
- Can model G with mean (and variance) accounting for covariates $\mathbf{X}_j$:

$$\beta_j | \mathbf{X}_j \sim N\left(\mathbf{X}_j'\gamma, \sigma_r^2\right)$$
$$\hat{\beta}_j \Big| \beta_j, s_j \sim N\left(\hat{\beta}_j, s_j^2\right)$$

- Estimate γ from regressing $\hat{\beta}_j$ on $\mathbf{X}_j$; deconvolve residuals $\hat{r}_j = \hat{\beta}_j - \mathbf{X}_j'\gamma$ (with knowledge of s_j) to estimate σ_r^2. For groups (e.g., nurse practitioners and physicians), could model group-specific $\sigma_{r(j)}^2$.
- Empirical Bayes posterior mean:

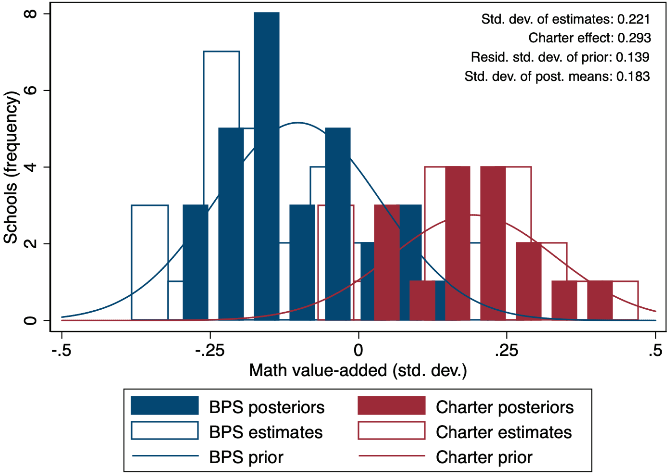
$$\hat{\beta}_j^* = \left(\frac{\hat{\sigma}_r^2}{\hat{\sigma}_r^2 + s_j}\right)\hat{\beta}_j + \left(\frac{s_j^2}{\hat{\sigma}_r^2 + s_j}\right)\mathbf{X}_j'\hat{\gamma}$$

School Value-Added (Angrist et al. 2017)



Source: Walters (2022)

School Value-Added (Angrist et al. 2017)



Source: Walters (2022)

EB for Bias Correction

- Can use EB framework to improve predictions of parameters when we have multiple estimates, some possibly biased (Angrist et al. 2017)
  - Suppose we have a precise but biased OLS estimate for β_j, with bias b_j:

$$\hat{\beta}_j^{OLS} \Big| \beta_j, b_j, s_{j,OLS}^2 \sim N\left(\beta_j + b_j, s_{j,OLS}^2\right)$$

  - Suppose we also have a noisy but (asymptotically) unbiased IV estimate:

$$\hat{\beta}_j^{IV} \Big| \beta_j, s_{j,IV}^2 \sim N\left(\beta_j, s_{j,IV}^2\right)$$

- Suppose Hausman test rejects equality of $\hat{\beta}_j^{OLS}$ and $\hat{\beta}_j^{IV}$. Should we throw away $\hat{\beta}_j^{OLS}$?

EB for Bias Correction

$$
\begin{aligned}
\hat{\beta}_j^{OLS} \Big| \beta_j, b_j, s_{j,OLS}^2 &\sim N\left(\beta_j + b_j, s_{j,OLS}^2\right) \\
\hat{\beta}_j^{IV} \Big| \beta_j, s_{j,IV}^2 &\sim N\left(\beta_j, s_{j,IV}^2\right) \\
\beta_j &\sim N\left(\mu, \sigma^2\right)
\end{aligned}
$$

- Use $\left\{\left(\hat{\beta}_j^{OLS}, \hat{\beta}_j^{IV}, s_{j,OLS}^2, s_{j,IV}^2\right)\right\}$ to estimate $G(\beta, b)$
- MSE-minimizing posterior $\hat{\beta}_j^* = E_{\hat{G}(\beta,b)}\left[\beta_j | \hat{\beta}_j^{OLS}, \hat{\beta}_j^{IV}\right]$

$$
\hat{\beta}_j^* = \hat{\lambda}_{IV}\hat{\beta}_{j,IV} + \hat{\lambda}_{OLS}\left(\hat{\beta}_{j,OLS} - \hat{E}\left[b_j\right]\right) + \left(1 - \hat{\lambda}_{IV} - \hat{\lambda}_{OLS}\right)\hat{\mu}
$$

  - Performs better than EB using only unbiased estimates, or $\hat{\beta}_{j,IV}^* = E_{\hat{G}(\beta)}\left[\beta_j | \hat{\beta}_j^{IV}\right]$

Normal Transformations

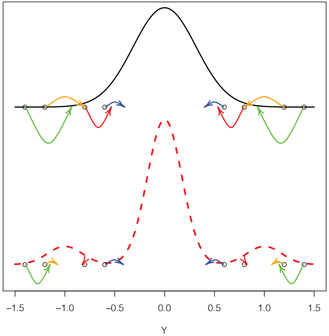
- Efron and Morris (1975): predict hits H_j out of N_j bats of baseball players for the remainder of the season
$$H_j \sim Binom\left(N_j, p_j\right)$$
- Transform H_i to an approximately normal distribution:
$$\begin{aligned} \tilde{H}_j &= \sqrt{N_j}\arcsin\left(2H_j/N_j - 1\right) \approx N\left(\beta_j, 1\right), \\ \beta_j &= \sqrt{N_j}\arcsin\left(2p_j - 1\right) \end{aligned}$$
- Use realized $\hat{p}_j = H_j/N_j$ to calculate $\hat{\beta}_j$. $\beta_j \sim G = N\left(\mu, \sigma^2\right)$. Deconvolve to find G.
- EB prediction $p_j^*\left(\mathbf{H}, \mathbf{N}\right)$ performs much better than using $\hat{p}_j$ alone to predict H_j

Non-Parametric G

- Recent advances estimate more flexible non-parametric G

$$
\begin{aligned}
(\beta_j, s_j^2) &\sim G \\
\hat{\beta}_j \,\Big|\, \beta_j, s_j^2 &\sim N(\beta_j, s_j^2)
\end{aligned}
$$

- Simulated example: mixture of three normal distributions (Gu 2022)

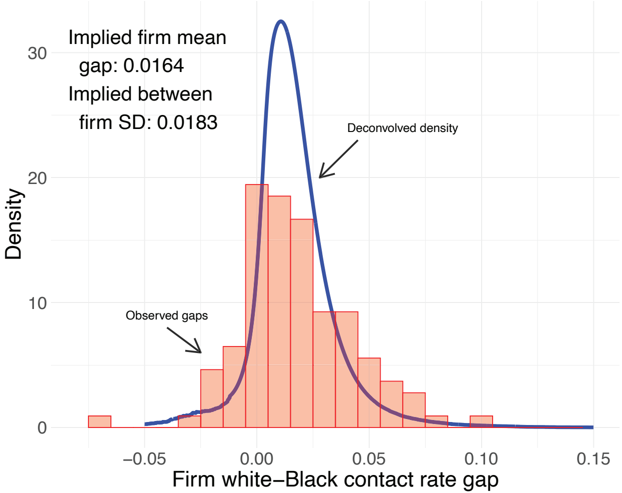


Non-Parametric G Techniques

$$
\begin{aligned}
z_j = \beta_j / s_j &\sim G \\
\hat{\beta}_j \Big| \beta_j, s_j^2 &\sim N\left(\beta_j, s_j^2\right)
\end{aligned}
$$

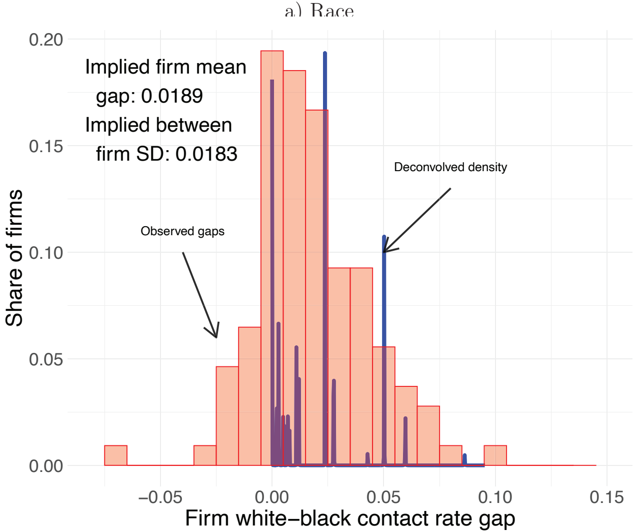
- Efron (2016): approximate G with flexible splines
  - Implement with `deconvolveR` R package (Narashimhan and Efron 2020)
- Non-parametric maximum likelihood estimator (NPMLE) (Robbins 1950): approximate G as discrete distribution with at most K mass points
  - Implement with `REBayes` R package (Koenker and Gu 2017)

Non-Parametric *G* Illustration: Efron (2016)



Source: Kline et al. (2022)

Non-Parametric *G* Illustration: NPMLE



Source: Kline et al. (2022)

Outline

Machine Learning

- EB methods are closely related to **machine learning** (ML) with **regularized regressions**
  - Both have many (J) parameters; goal of both to improve predictions with finite observations per J
- Consider normal-normal model with J physicians and N patients per physician:

$$\begin{aligned} Y_i &= \beta_{j(i)} + \varepsilon_i \\ \varepsilon_i &\sim N\left(0, \sigma_\varepsilon^2\right) \\ \beta_j &\sim N\left(0, \sigma_\beta^2\right) \end{aligned}$$

- Recall unbiased fixed effect estimator: $\hat{\beta}_j^{FE} = \frac{1}{N}\sum_i \mathbf{1}\left(j\left(i\right) = j\right) Y_i$; using it is akin to **overfitting**, a problem ML seeks to solve
- EB posterior distribution for β_j is normal $\Rightarrow$ posterior mean β_j^* = posterior mode, also known as **maximum a posteriori** (MAP)

Machine Learning

- Plugging in normal densities for ε_{ij} and β_j (i.e., G), we can solve for the MAP $\left(\beta_1^*, \ldots, \beta_J^*\right)$:

$$
\begin{aligned}
\left(\beta_1^*, \ldots, \beta_J^*\right) &= \arg \min_{(\beta_1, \ldots, \beta_J)} \sum_{j=1}^{J} \sum_{i=1}^{N} \mathbf{1}\left(j\left(i\right) = j\right)\left(Y_i - \beta_j\right)^2 + \frac{\sigma_\varepsilon^2}{\sigma_\beta^2} \sum_{j=1}^{J} \beta_j^2 \\
&= \arg \min_{(\beta_1, \ldots, \beta_J)} \sum_{j=1}^{J} \sum_{i=1}^{N} \mathbf{1}\left(j\left(i\right) = j\right)\left(Y_i - \beta_j\right)^2 + \lambda p\left(\beta_1, \ldots, \beta_J\right)
\end{aligned}
$$

  This is the solution to a regularized regression, with penalty $p\left(\cdot\right)$ and tuning parameter $\lambda = \sigma_\varepsilon^2 / \sigma_\beta^2$

- The particular regularized regression is known as **ridge regression**

- In spirit of EB, data are used to choose λ

Machine Learning

- Thus, ML regularization often has EB interpretation
  - Ridge regression estimates (L2 penalization): posterior means/modes from a model with normal priors
  - LASSO regression estimates (L2 penalization): posterior modes from double exponential (Laplace) priors
- May be useful to think about implicit EB prior distribution of parameters (i.e., G)
  - Some ML approaches will perform better under certain implicit parameter prior distributions (Abadie and Kasy 2019)

Multiple Hypothesis Testing

- With $\hat{G}$ and $\left(\hat{\beta}_j, s_j^2\right)$, can use EB to make relevant policy assessments that are compound decision problems, e.g.,
  - Which doctors are in the top quintile of performance ($\beta_j > G^{-1}(0.8)$)?
  - Which VA hospitals are discriminating against Black patients ($\beta_j > 0$)?
- Such problems are “large-scale inference” problems, closely related to multiple-testing problems (Efron 2012)

Multiple Hypothesis Testing

- Example: Kline et al. (2022) interested in classifying firms as discriminating against Black applicants (see also Gu and Koenker (2022))
  - Send random applications with Black-sounding vs. white-sounding names to J firms
  - Estimate $\left\{\left(\hat{\beta}_j, s_j^2\right)\right\}_{j=1}^{J}$ for J firms
  - Can perform one-tailed t-test: $\beta_j = 0$ vs. $\beta_j > 0$. Implies test statistic $z_f = \hat{\beta}_j / s_j$ and p-value $p_j = 1 - \Phi(z_j)$.
- Decision rule: classify firm as discriminatory if $p_j \leq \overline{p}$
  - How many mistakes do we expect to make (i.e., **false discovery rate** or FDR) for a given $\overline{p}$? What should we pick for $\overline{p}$?

Multiple Hypothesis Testing

- From definition of p-value, $\overline{p} = \Pr\left(p_j \le \overline{p} \middle| \beta_j = 0\right)$; interested in $FDR = \Pr\left(\beta_j = 0 \middle| p_j \le \overline{p}\right)$

- By Bayes rule,

$$
\begin{aligned}
FDR\,(\overline{p}) &= \Pr\left(\beta_j = 0 \middle| p_j \le \overline{p}\right) \\
&= \frac{\Pr\left(p_j \le \overline{p} \middle| \beta_j = 0\right) \Pr\left(\beta_j = 0\right)}{\Pr\left(p_j \le \overline{p}\right)} \\
&= \frac{\overline{p}\,\Pr\left(\beta_j = 0\right)}{\Pr\left(p_j \le \overline{p}\right)}
\end{aligned}
$$

  - $\Pr\left(p_j \ge \overline{p}\right)$ is a function of the data; $\Pr\left(\beta_j = 0\right)$ depends on G

- Set $FDR\,(\overline{p})$ based on cost of type I (false positives) vs. type II errors (false negatives)

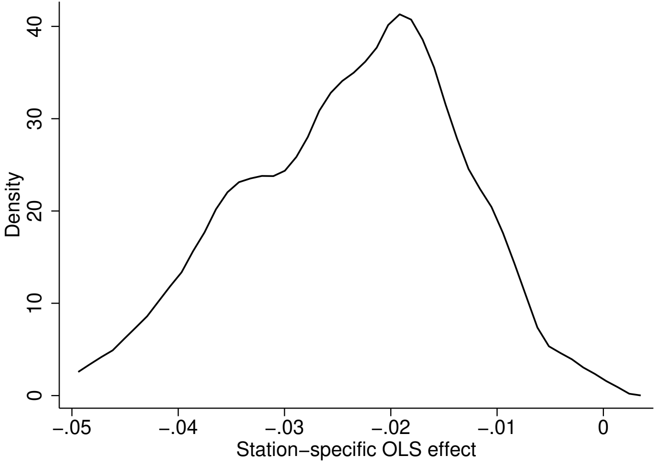
Outline

VA vs. Non-VA Care

- Chan et al. (2022): quasi-experimental assignment of ambulances to dually eligible veterans above age 65
  - Veterans may receive care at VA or non-VA emergency departments (EDs)
  - IV approach: ambulances have different propensities to transport veterans to the VA
  - Condition on zip code: each location is part of a different quasi-experiment (each with its own VA and non-VA hospitals of interest)
- Main finding: VA hospitals reduce mortality by 45% in IV design; robust 21% reduction in mortality by OLS
- How might this result vary across VA stations?

VA vs. Non-VA Care

Station-specific VA mortality effect (OLS design): all stations reduce mortality

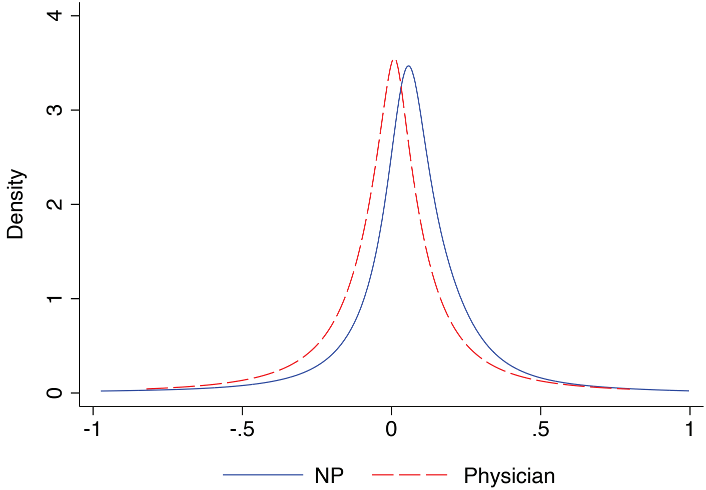


NP vs. Physician ED Productivity

- Chan and Chen (2022): quasi-experimental assignment of ED patients to NPs vs. physicians
  - IV approach: availability of NPs and physicians at the time of patient arrival
- Main finding: on average, NPs have lower productivity, using more resources but achieving worse outcomes
- What about the productivity distribution within professions?

NP vs. Physician ED Productivity

Deconvolved distributions of productivity: 38% overlap in productivity



Conclusion

- Empirical Bayes methods provide tools to jointly assess effects across important units of interests (e.g., physicians, hospitals)
- The increasing granularity of data (including at the VA), combined with computational tools, has led to a rise in recent methods and applications
- Compound decision problems and large-scale inference made possible by these methods are extremely policy-relevant

References

Abadie, Alberto, and Kasy, Maximilian. 2019. "Choosing among regularized estimators in empirical economics: the risk of machine learning." *Review of Economics and Statistics:* 101(5).

Angrist, Joshua D., Peter D. Hull, Parag A. Pathak, and Christopher R. Walters. 2017. "Leveraging Lotteries for School Value-Added: Testing and Estimation." *Quarterly Journal of Economics* 132 (2): 871-919.

Chan, Jr, David C., David Card, and Lowell Taylor. 2022. "Is There a VA Advantage? Evidence from Dually Eligible Veterans." Working Paper 29765. National Bureau of Economic Research.

Chan, Jr, David C., and Yiqun Chen. 2022. "The Productivity of Professions: Evidence from the Emergency Department." Working Paper 30608. National Bureau of Economic Research.

Efron, Bradley. 2016. "Empirical Bayes deconvolution estimates." *Biometrika* 103(1).

Efron, Bradley. 2012. *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction.* Cambridge University Press.

Efron, Bradley, and Carl Morris. 1973. "Stein's Estimation Rule and Its Competitors—An Empirical Bayes Approach." *Journal of the American Statistical Association* 68 (341): 117-30. Gu, Jiaying. Empirical Bayes Methods. 2022. National Bureau of Economic Research.

References

Gu, Jiaying, and Roger Koenker. 2023. “Invidious Comparisons: Ranking and Selection as Compound Decisions.” *Econometrica* 91(1): 1–41.

James, W., and Stein, Charles. 1961. “Estimation with quadratic loss." *Berkeley Symposium on Mathematical Statistics and Probability* 1.

Kline, Patrick, Evan K Rose, and Christopher R Walters. 2022. “Systemic Discrimination Among Large U.S. Employers.” *Quarterly Journal of Economics* 137 (4): 1963-2036.

Koenker, Roger and Jiaying Gu. 2017. “Rebayes: an R package for empirical Bayes mixture methods.” *Journal of Statistical Software* 82(8).

Narasimhan, Balasubramanian, and Bradley Efron. 2020. “deconvolveR: a G-modeling program for deconvolution and empirical Bayes estimation.” *Journal of Statistical Software* 94(11).

Robbins, Herbert. (1950). “A generalization of the method of maximum likelihood: estimating a mixing distribution.” *Annals of Mathematical Statistics* 21(2).

Walters, Christopher. Empirical Bayes Methods. 2022. National Bureau of Economic Research.