Something interesting

Fu Zixuan*

July 4, 2024

Abstract

someting interesting

*Last compiled on June 17, 2024

Contents

1 Introduction

It is almost of human nature to compare, rank and select. And competition, be it good or bad, emerges in the wake. As invidious as ranking and selection can be, in many cases it is one of the driving forces behind any improvement (e.g. natural selection in the history of evolution). The society itself is constantly constructing league table. It rewards the meritous and question or even punishes the unsatisfactory. The measure based on which rank is constructed ranges from teacher's evaluation to communities' mobility index as pioneered by [**?**].

The present article extends the practice to the health sectors. To be more specific, it studies the labor efficiency across all hospitals in France. By exploring a comprehensive database (SAE) of French hospitals, we first construct a measure of labor efficiency. Then based on the estimates, we compare the public and private hospitals by selecting the top-performing units. We borrow from the recent developments in Empirical Bayes method to achieve the comparison.

Conclusion here

The article bridges two fields of interests. One is on productivity analysis. The most popular methods are Data Envelopment Analysis [**?**] and Stochastic Frontier Analysis [**?**]. Yet we abstract from both of them for the convenience of statistical inference. We use the simple method applied in [**?**] by estimating a conditional input demand function. To put it simply, we estimate a linear function of how much labor input is needed to produce a give list of outputs [1]. We only focus on the employment level of nurses for (reasons) [2] in the specification.

The second area of interests is the Empirical Bayes Methods. The name "Empirical Bayes" is self-telling in the sense that Bayes implies a prior distribution while Empirical means empirically estimate the prior from the data. Details on how it can be of use in ranking and selection will be presented in the rest of the section.

In [**?**], the authors argue that public hospital is less efficient than private counterpart in the sense that it would need a smaller size of personnel if it were to use the input demand function of the private hospital, which is the main result of their counterfactuals.

Having roughly replicated the results after doubling the length of the panel, the paper differentiates itself by including/adding the standard/classical panel data methods in demand function estimation, specifically the fixed-effect within-group estimation and GMM [**?**].

A tiny bit about the literature of GMM

Though the original focus of the panel data estimator is on the β parameters. It also provides us with a noisy estimate of the underlying unobserved heterogeneity term denoted as θ_i. (It's important to note that this heterogeneity is not necessarily indicative of inefficiency). Now that we have set the stage for empirical bayes, it is time to bring about the prior distribution of θ_i, denoted as G_θ. If the prior distribution G is known, having observed an estimate $\hat{\theta}_i$ of θ_i, we can update our noisy estimate $\hat{\theta}_i$ using or incorporating our knowledge

[1] We refer the audience to xxx for detailed reasons of adopting this approach.

[2]

of G.

The usefulness of a prior G is further exemplified/highlighted in the ranking and selection problem mentioned above, when the object of interests is the noisy estimate of θ_i. For example, in [**?**], we are given the task of selecting the top 20% out of the population of α_i, that is to say selecting those i whose $alpha_i > G^{-1}(0.8)$, while controlling for the overall false discovery rate ($\frac{x}{y}$) at 5%. In [**?**], the authors try to develop an optimal decision rule for the given task. To put it in the language of optimization, they want to have a decision rule that optimizes the performance of selection, equivalently, minimize the loss of selection

$$\delta^* = \min_{\delta} \text{Loss} \quad \text{subject to contstraints}$$

where the loss function can take different forms, for example the expected number of total type 1 and 2 mistakes.

The task at hand falls naturally under the compound decision framework pioneered by [**?**] if we define the loss function in such a way that takes into account the results of all the individual decisions $\delta(Y_i)$. For instance, summing all mistakes would be one way to aggregate/compound individual decisions,

It is obvious that in order to impose the two stated constraints (capacity and FDR) in formulating the optimization problem, we need to know the prior distribution G. Despite the importance of the G, it does not fall from heaven. Therefore, Empirical Bayes methods come to the rescue, as its name suggests, we will have to empirically estimating the unknown prior G.

Often times "empirically estimating G" is done with parametric assumption that G is normal. Notable use cases are found in [**?**] which aroused the ensuing/lasting interests on teacher evaluation and social mobility in communities. By assuming a Gaussian G, they shrink the estimated fixed effect linearly, thus giving the name "linear shrinkage". However, departure from normality may render the linear shrinkage rule as unhelpful. Thanks to the foundational work of [**?**] who has shown that non-parametric estimation is also feasible and consistent, it is preferable to relax the normality assumption and estimate the prior G non-parametrically. In terms of computation method, [**?**] has formulated the non-parmetric estimation as a convex optimization problem. Compared to other popular methods such as EM algorithm [**?**], recent advancements in convex optimization computation methods [**?**] has made the novel approach of [**?**] computationally more attractive.

It is worth mentioning here that though a discrete G with at most x atoms can be estimated using the REBayes package [**?**], we are not free of imposing any assumptions, that is the distribution of estimate $\hat{\alpha}_i | \alpha_i, \sigma_i^2$. To illustrate, in the case of the estimate of fixed effect, we have

$$\begin{aligned}\hat{\alpha}_i &= \frac{1}{T}\sum_t (y_{it} - x_{it}\hat{\beta}) \\ &= \frac{1}{T}\sum (\alpha_i + \varepsilon_{it} + x_{it}(\beta - \hat{\beta})) \\ &\rightarrow^d \alpha_i + \frac{1}{T}\sum_t \varepsilon_{it}\end{aligned}$$

The asymptotic distribution follows from the consistency of $\hat{\beta}$ when $N \to \infty$, a reasonable assumption in wide panels.

If we may boldly assume that the errors are *i.i.d.* normally distributed for each i

$$\varepsilon_{it} \sim N(0, \sigma_i^2)$$

Then a fixed/small T won't do jeopardize/imperil xxx our argument too much since we do not need to invoke central limit theorem to have

$$\hat{\alpha}_i \rightarrow^d N(\alpha_i, \sigma_i^2/T)$$

However without the normality assumption on the error terms, we have to resort to the central limit theorem from the claim that $T \to \infty$, which may seem unrealistic for a wide panel ($N >> T$).

The rest of the paper is organized as follows. Section 2 briefly describes the data and lays out the reduced form estimation of the input demand function, treating the number of nurses as the dependent variable and a list of 9 output measures as the regressors. Section 3 applies the classical panel data estimators to the same specification, distinguishing between whether strict exogeneity is assumed. In section 4, we introduce the compound decision framework and specifically/xxx define the selection problem. Section 5 follows with a comparison of the different selection outcome as a result of imposing varying constraints and assumptions. In section 6, we try to draw preliminary conclusion on the comparative performance of public and private hospitals. Section 7 discusses potential issues and concludes.