Opening

Merci d'etre venu pour ma soutenance. Thank you for coming to my presentation. The subject I am going to present today is an application of the Empirical Bayes method in selecting the most labor efficient hospitals in France. Thus, giving the name "Hopital" selection rule. PUN INTENDED.

Introduction

My presentation connects two fields of interests. The first one is on productivity analysis. I want to draw a comparison between the public and private hospitals. Due to the constraint of data, I only focus on the use of labor, more specifically nurses. And abstract from quality and other measure etc. Instead of estimating a production function, such as in the stochastic frontier model or use a non parametric approach data envelopment analysis, I follow \cite{croiset2024} and use a simple specification that can have the interpretation of **conditional demand function**. So the reduced form relationship is that the dependent variable is the number of labor input, and the right-hand side is a list of medical outputs. Under this specification, the less input needed the more efficient the hospital is obviously.

The second strand of literature is actually the main focus of today's exposition. As pointed out by \cite{Gu2023}. It is of human nature to rank and select. This tendency is given a name called league table mentality. However, unfortunately, most of the time, what we have is the noisy estimate of the things we want to rank or select. The estimated fixed effect from a panel data estimation. Secondly, I care about the collective performance of my decision, not just one single individual decision, which is why we call it a **compound** decision. The compound decision framework and the corresponding compound risk is closely related to the Bayesian risk. When I say Bayes, that is to say, I want to assume a prior distribution of the true parameters the parameters that I want to rank or select. Then I can estimate the prior empirically Using parametric assumption, or as applied in this presentation, NPMLE.

Having surveyed/gone through the two strands of literature, the simple/question and the easy exercise that I want to do is making comparison between the public and private hospitals from a granular perspective. Out of the around 1600 hospitals in the sample, I want to select the top 20% percent, the best performing units. I classify them by legal status and see how many are public and how many are private. I want to control for the expected number of mistakes. What would be the selection outcome if I control for the mistakes? Besides, there are different ranking statistics that I can use. What would be the outcome if I use different ranking statistics and hold different assumptions?

A quick recap. I need to first have an estimate of the hospital efficiency. This is done by the usual panel data estimator. In my case, since it is an input demand specification, the fixed effect is the inefficiency measure. The smaller the theta, the more efficient the hospital i is. Secondly I need to construct a prior distribution. Finally, I can perform selection incorporating the information from G. Let's assume that we have already accomplished the first steps. Holding the inefficiency measure, I want to go directly to step 2 and 3.

Data

A quick look at the data. I follow the paper by croiset and gary-bobo. The data I have is from statistique annuelle d'etablissement. It is publically available. Covering all french hospitals from 2013 to 2022. However, year 2020 is completely missing. The number of hospitals is quite stable over the years. The data itself only distinguishes public, private for profit and private non profit. However, it is important to single out the teaching hospitals from the public. Teaching hospitals are innately very different. They are humongous hospitals that not only treat patient but also train medical students. They need to allocate a significant amount of resources to teaching and research, unlike the other types. For this reason, I will exclude it from the estimation.

Turn to the output of each type of hospitals. We can see that each type of hospitals differ in the mix of services that they provide. Emergency care is mostly taken care of by public hospitals and private hospitals are strong in medical sessions and outpatient stays. As can be reconfirmed, the teaching is quite different (large). I will exclude them from estimation so that it is more reasonable to assume that all hospitals have similar input demand function.

Compound decision framework

Now Let's jump to the main sections. Selection as compound decision. Let's say we observe a vector of estimate of the true parameter θ. Each estimate conditional on the true parameter follows a certain distribution of P. The delta is my decision based on the observed estimate. For example, in selection problem, delta is 1 or 0, indicating whether i is selected or not. included in the selection set or not.

The reason why this is a compound decision is due to the way we define the loss function. The collective/compound loss function is the simple sum of individual loss. Therefore, the objective of the selection problem is to minimize the expected compound loss -- compound risk. The compound decision framework is first coined by Robbins in 1950s. How is it related to bayesian risk? The compound risk is the bayesian risk when we replace the sum by integration over the prior G. The compound risk is related to the frequentist view, by treating θ_i as fixed unknown parameter. The bayesian risk assumes that the θ_i does follow a distribution G. Since neither Gn or G is known, we need to estimate it empirically.

How do we do that? Kiefer and Wolowitz xxx. The g is the likelihood of observing y. Kiefer and Wolowitz have established the consistency of such a NP estimation. As can be easily shown that this is an infinite dimension convex optimization with linear constraints. The recent development in computation method helps solving the NPMLE by discretizing the support.

Selection task

Recall that the selection problem is to select the bottom 20% of θ_i, note that we want to select the true theta i not theta i hat. Also, we want to control the number of falsely selected. The expected number of wrongly selected/all selected. Previously, I denote the loss by L. And this L depends on the specific problem we are trying to tackle. For a selection problem, the loss function is defined as the following. First, We want to minimize the number of i that belongs to the set but unselected. h corresponds to whether the true value belongs to the set. δ corresponds to whether unit i is selected. Second, it is the constraint that control the expected false discovery. Third, it is the constraint that limits our interests to the bottom 20%.

The loss function is explicitly defined, but since we don't know the true value of each θ. We will take expectation of the loss. By taking expectation, we retrieve the most important component of the compound risk, posterior tail probability. This is the probability of being in the tail given the estimate hat theta.

To put the question into context, let's say the true inefficiency parameter is theta_i, we observe a sequence of Y_it that follows a normal distribution centered at the true value. The worst case scenario is that we don't know theta nor the variance sigma. But we have two sufficient statistics for them. Y_i and S_i with respective distribution.

For example, y_it is the loginput-logouput times beta, but subject to assumptions that the error term in the input demand specification is normal.

Given the two sufficient statistics, The posterior tail p is the p that theta belongs to the set given our estimate of theta_i and sigma_i. This is the bayes rule without any issue.

Posterior tail probability

Now we have defined the ptp, the two constraints can be expressed in terms of ptp. Back to our risk function. minimize the first term subject to two constraint. The first term is saying that we select one by one from the highest tail prob to the lowest. The constraints are saying we stop selecting once a constraint is hit. Therefore, there exists a cutoff λ such that we select all i whose tp is higher than lambda. And one constraint binds.

A quick recap.

1. We have a longitudinal panel. For each i, we got a T observation. Each observation Y_it is normal around the true theta_i.
2. Given the yit, we perform NPMLE to get a G. and smoothed it.
3. Given the estimated G, we calculate ptp for each i. as well as the two constraints.
4. Solve the problem and find the optimal cutoff lamda such that all i whose ptp higher than lambda is selected.

Results

What would it be like in the hospital application?

Previously, I have stated the most general case where either theta_i nor sigma_i is known. An estimate gives me the LHS. A density function over theta and sigma. In many of the literature reference, due to the difference in data and model, they assume that sigma is known. I also apply it by setting the estimated s_i as the true value, which gives me the RHS. a density function over theta. Though I believe the LHS is a more realistic assumption.

I can also smooth it by kernel smoothing. (biweight kernel with kernal width)

Let us look at the case we have been talking about, bottom 20% and FDR 20%. The LHS corresponds to the case where FDR is imposed. Imposing FDR cst shrinks the selected set by a dozen. The private are 6-7 times more than public.

What if we take sigma as known? Then surprisingly, the FDR is not binding. (Intuition) is that when sigma is assumed to be known, the estimate theta hat is less noisy, giving rise to a lower chance of making mistakes, thus FDR is less likely to bind. This approach is not the same as taking the face value of sigma hat and select. As we can see from the figure. For MLE, we are selecting based on the theta hat. while in the compound decision approach, we are selecting based on the posterior tail probability.

Estimation

Now I may talk about the first step which is estimating the fixed effect. The standard method is WG FD estimation. Yet, the medical output is exogenous is a questionable assumption not to mention the strict exogeneity. Therefore, in order to relax it, I use the first difference GMM. The system GMM is also feasible if I impose further assumptions. But the overidentification hypothesis is rejected. However, when I assume that some regressors are exogenous but I still overidentified moment condition, the sargan test is not rejecting the null hypothesis. Statistically I can do that, but it doesn't make economics sense to me why some regressors are exogenous while some are not. Due to the fact that I spent most time learning the empirical bayes, I am obliged to save it for future exploration.

Conclusion

A few points in conclusion. is clear that the FDR constraint shrinks the selection set by some amount. It is also worth mentioning that It is also intuitive that the larger the capacity (the larger the α), the less binding the FDR constraint is. The idea is that when the decision-maker can select more units, the probability of making mistakes decreases.

Another observation comes from the assumption we make in the estimating prior G. In \citet{gu2023invidious}, the authors have pointed out that while the known variance assumption in $Y_i|\theta_i,\sigma_i$ may be plausible in some applications, it is more common to be faced with only an estimate of the variance.

The two assumptions give rise to a different level of \textit{stringency} in response to the constraints, especially when the decision-maker wants to control for the expected false discovery rate.

With respect to the private-public comparison, among the top 20% performers. A preliminary conclusion is that in terms of labor employment efficiency, there are more efficient private hospitals among the top performers.

However, one cautionary note is the interpretation of the fixed effect estimate. Since the fixed effect captures all time-invariant components of the unit, whether it is only the unobserved heterogeneity of individual hospitals or an actual measure of inefficiency is questionable. This issue is discussed in \citet{greene2005fixed}.

Lastly, despite the fact that it is human nature to construct rankings and make selections, every step of the procedure requires attention to specification, identification, and justifiable assumptions. Incorporating constraints such as FDR in defining the problem may be helpful, but the decision is still subject to great uncertainty and should be made with great care and caution.

With that, I conclude. Thank you for your attention:)