14.771 Development Economics: Microeconomic Issues and Policy Models
Fall 2008

# 14:771: Recitation Handout #8
Normal Learning and Inference with Multiple Outcomes

## Bayesian Learning

Lots of the technology adoption and learning literature has used the notion of Baysian learning, particularly the normal learning model, which is tractable and relatively straightforward to work with. We will here review a few things in Baysian learning so as to make sure that everybody is on the same page.

### Bayes' Rule

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

### Normal learning

Foster and Rosenzweig (1995) uses a very common tool for learning models based on the fact that one receives a signal that is equal to the "real" parameter plus some normal noise

$$\text{signal: } \widetilde{\theta}_{ijt} = \theta^* + u_{ijt}$$

where $u_{ijt}$ is a iid variable with distribution:

$$u_{ijt} \sim N\left(0, \sigma_u^2\right)$$

The $t = 0$ prior of the farmer is given by:

$$\theta^* \sim N\left(\hat{\theta}_{j0}, \sigma_{\theta j0}^2\right)$$

At the end of the first period, the farmer learns what $\widetilde{\theta}_{ij1}$ on each plot he planted was and can then update his prior on $\theta^*$. The normal learning model is generally easier to think about in terms of precision: $h = \frac{1}{\sigma^2}$. We can rewrite everything as

$$\begin{aligned} \text{signal} &: \quad \widetilde{\theta}_{ijt} \sim N\left(\theta^*, \frac{1}{h_u}\right) \\ \text{prior} &: \quad \theta^* \sim N\left(\hat{\theta}_{j0}, \frac{1}{h_{\theta j}}\right) \end{aligned}$$

Normal learning involves some nasty math - however, the final product is very easy to write using precision (you've all probably done the nasty math at some point on a 381 problem set, so I'll omit it here - in the future all you'll want to remember is the updating formula anyway):

$$\text{posterior}_{t=1}\text{: } \theta^* \mid \widetilde{\theta}_{ij1} \sim N\left(\frac{h_u \widetilde{\theta}_{ij1} + h_{\theta j}\hat{\theta}_{j0}}{h_{\theta j} + h_u}, \frac{1}{h_{\theta j} + h_u}\right)$$

So, what happens when we have multiple periods of learning, as we have in Foster and Rosenzweig? At $t = 2$ we just treat our posterior from $t = 1$ as our new prior and incorporate the next signal:

$$\text{posterior}_{t=2} : \theta^* \mid \widetilde{\theta}_{ij1}, \widetilde{\theta}_{ij2} \sim N\left(\frac{(h_{\theta j} + h_u)\frac{h_u\widetilde{\theta}_{ij1} + h_{\theta j}\hat{\theta}_{j0}}{h_{\theta j} + h_u} + h_u\widetilde{\theta}_{ij2}}{h_u + (h_{\theta j} + h_u)}, \frac{1}{(h_{\theta j} + h_u) + h_u}\right)$$

simplifying everything out we get:

$$\text{posterior}_{t=2} : \theta^* \mid \widetilde{\theta}_{ij1}, \widetilde{\theta}_{ij2} \sim N\left(\frac{h_u\left(\widetilde{\theta}_{ij1} + \widetilde{\theta}_{ij2}\right) + h_{\theta j}\hat{\theta}_{j0}}{2h_u + h_{\theta j}}, \frac{1}{2h_u + h_{\theta j}}\right)$$

then we can extrapolate this out to time $T$, when we get

$$\text{posterior}_{t=T} : \theta^* \mid \widetilde{\theta}_{ij1}, \widetilde{\theta}_{ij2}, ...\widetilde{\theta}_{ijT} \sim N\left(\frac{h_u\sum_{t=1}^{T}\widetilde{\theta}_{ijt} + h_{\theta j}\hat{\theta}_{j0}}{Th_u + h_{\theta j}}, \frac{1}{Th_u + h_{\theta j}}\right)$$

The other wrinkle in FR is that we can also learn from our neighbors - so in each period we get some signal from our own plots and some signal from their plots. This is fine, as you've probably already noticed that the normal learning model is additive. Let's now denote the signal you get from you neighbor according to:

$$\text{signal (neighbor): } \widetilde{\theta}_{ij't} \sim N\left(\theta^*, \frac{1}{h_n}\right)$$

Let's also assume for now that the own signal and neighbor signal are independent. We can then incorporate neighbor learning into the model:

$$\text{posterior}_{t=T} : \theta^* \mid \widetilde{\theta}_{ij1}, .., \widetilde{\theta}_{ijT}, \widetilde{\theta}_{ij'1}, .., \widetilde{\theta}_{ij'T} \sim N\left(\frac{h_u\sum_{t=1}^{T}\widetilde{\theta}_{ijt} + h_n\sum_{t=1}^{T}\widetilde{\theta}_{ij't} + h_{\theta j}\hat{\theta}_{j0}}{T(h_u + h_n) + h_{\theta j}}, \frac{1}{T(h_u + h_n) + h_{\theta j}}\right)$$

- What happens regarding the importance of a farmer's prior as time progresses?
- What happens to precision as time progresses?
- Will learning progress faster or slower when signals are more precise?
- Will learning progress faster or slower when priors are more precise?

# Multiple outcomes in analysis

Multiple inference corrections are important in studies with multiple outcomes - if I run regressions of my treatment on 100 outcomes, how many significant results will I find when my treatment has absolutely no impact? This is a big issue because surveys in randomized trials generally ask about lots of outcomes, and economists like publishing papers with

significant results. You should always have this in mind when reading papers, particularly those on randomized trials (the WISE experiment is one example)...

There are two main multiple inference techniques - reducing the number of tests (more common) and correcting p-values (less common). Here are some handy references:

- Anderson (JASA Forthcoming) - "Multiple Inference and Gender Differences in the Effects of Early Intervention: A Reevaluation of the Abecedarian, Perry Preschool,a nd Early Training Projects"
- Katz, Kling, and Liebman (2007, EMA) - "Experimental Analysis of Neighborhood Effects"
- Kling and Liebman (2004 Working Paper) - "Experimental Analysis of Neighborhood Effects on Youth"

I will focus on Anderson (2008), which considers several early childhood education trials, employs both these techniques if you'd like an applied example.

## Reducing the number of tests

### Anderson (2008)'s Method

We just follow these steps:

1. Choose a specific set of outcomes based on a-priori notions of importance. Anderson (2008) chooses IQ scores, grade retention, special education, high school graduation, college attendance, employment, earnings, government transfers, arrests, convictions or incarcerations, drug use, teen pregnancy, and marriage. Overall, there are 47 outcomes. In practice, he uses 9 groups (groups are the $J$ referred to below) for each gender defined by three categories in each domain (academic, economic,and social), experiment (he looks at 3 experiments) and age.

2. For all outcomes, switch signs where necessary so that the positive direction always indicates a "better" outcome

3. Demean all outcomes and convert them to effect sizes by dividing each outcomes by its control group standard deviation. Call the transformed outcomes $\tilde{y}$. (This conversion normalizes outcomes to be of comparable scale).

4. Define $J$ groupings of outcomes (also referred to as areas, domains, or families). Every outcome $y_{jk}$ is assigned to one of these $J$ areas, giving $K_j$outcomes in each area $j$.

5. Create a new variable $\bar{s}_{ij,}$ that is a weighted average of $\tilde{y}_{ijk}$ for each individual $i$ in area $j$. When constructing $\bar{s}_{ij}$, weight its inputs by the inverse of the covariance matrix of the transformed outcomes in area j - note that this is an efficient GLS estimator. (KKL (2007) weight outcomes equally - that is, they just sum stuff up - so you don't have to do the GLS weighting if you don't want to).

6. Regress the new variable $\bar{s}_{ij}$ on treatment status to estimate the effect of treatment in area $j$. A standard t-test assesses the significant of each coefficient.

**The advantages of this method include:**

- The probability of a false rejection does snot increase as additional outcomes are added to the index
- They are potentially more powerful - several outcomes that are marginally significant when examined separately may produce a significant change in the index (job market papers have been made this way!)

**The disadvantages include:**

- Magnitudes are difficult to interpret. We refer to "effect sizes", but this is mushy.
- Impacts on specific outcomes of interest are obscured
- This method relies on the ability to classify some outcomes as "better" than others. It is not always clear cut what is better (i.e. child intraversion vs. extraversion....)
- The choice of grouping can affect results

**Kling and Liebman (2004)'s Method**

The first part is just as above: First, one needs to change the outcomes so that they are all going in the "same" direction. For example, if one has three measures of schooling: test score, attendance and drop-out rates, one would convert the later in a "non-drop-out" rate so that if the program has a positive effect on schooling, all these outcomes will be increased. Define the effect size one finds for each outcome i as $\pi_i$ and then compute the variation in that outcome in the control group

$$\sigma_i^2 = Var\left(Y_i | Z = 0\right)$$

and then find the overall treatment effect by

$$\tau = \frac{1}{I} \sum \frac{\pi_i}{\sigma_i}$$

To compute the standard error on that measure, one needs to take into account the fact that the outcomes may be correlated. To do so, the easiest thing to do is to run the same regression using all outcomes in a Seemingly Unrelated Regression (SUR). When you use exactly the same covariates on the left-hand side of the equations, the point estimates produced by SUR are algebraically identical to those of an OLS. Since SUR does not assume that the error terms of each regression are independent, it will estimate the covariance between each outcome. Then, one can test the hypothesis whether $\tau > 0$ as a test of a

linear combination of parameters of the regressions. To do this in STATA, one uses the function SUEST followed by LINCOM.

Also note that for whatever reason, this technique got changed to the "sum up the index" technique discussed above for the published Econometrica paper, so you may want to go ahead and do that instead of the SUR technique that they used in the working paper. You will see this used in other papers that deal with multiple outcomes though, so it's good to be aware of it.

## Familywise Error Rate (FWER)

- The FWER is the probability that at least one of the J true hypotheses from a family of M hypotheses is rejected.

As the number of hypotheses in a family increases, the probability of rejecting at least one of them (and therefore concluding that some effect is statistically significant) at a given $\alpha$ level increases. There are two popular methods to adjust the FWER:

1. *Bonferroni correction* - for each test, multiply the p-value by M, the total number of tests performed. The advantage of this correction is that it's simple. The disadvantage is that it doesn't have much power, and the calculated p-values can often be greater than 1.

2. *Free step-down resampling method (Westfall and Young, 1993)* - one advantage of this method is that it yields an exact probability rather than an upper bound. Another advantage is that when a hypothesis is rejected, it is removed from the family being tested, increasing the power of the remaining tests. A third advantage is that it incorporates dependence between outcomes. We implement this procedure as follows (Anderson, p. 14):

   (a) Sort outcomes $y_1, ...y_m$ in order of decreasing significance (increasing p value): $p_1 < p_2 < ... < p_M$

   (b) Simulate the data set under the null hypothesis of no treatment effect using the resampling procedure described in section 3.1 of Anderson (2008).

   (c) Calculate a set of simulated p-values for outcomes $y_1, ...y_m$ ($p_1^* < p_2^* < ... < p^*$) using the simulated treatment status variable. Note that they will not display the same monotonicity in terms of p values as your original list.

   (d) Enforce the original monotonicity: Compute $p^{**} = \min\left\{p_r^*, p_{r+1}^*, ...p_M^*\right\}$ ($r$ denotes the original significant rank of the outcome, with $r = 1$ being the most significant and $r = M$ being the least significant).

   (e) Perform $L \geq 100,000$ replications of steps b-d. For each outcome $y_r$, tabulate $S_r$, the number of times that $p^{**} < p_r$.

   (f) Compute $p_r^{FWER^*} = \frac{S_r}{L}$

(g) Enforce monotonicity a final time:

$$p_r^{FWER} = \min\left\{p_r^{FWER^*}, p_{r+1}^{FWER^*}, ..., p_M^{FWER^*}\right\}$$

This final monotonicity adjustment ensures that larger unadjusted p-values always correpond to larger adjusted p-values.

## False Discovery Rate

While the FWER controls for the number of any false rejects, the FDR controls for the proportion of any false rejections. The researcher can decide which method to use based on the cost of a false rejection. See Anderson (2008) for implementation details.