A Simple, Graphical Procedure for Comparing Multiple Treatments

Brennan S. Thompson, Ryerson University
Matthew D. Webb, Carleton University

2017 Canadian Stata Users Group Meeting

For Stata code please email matt.webb@carleton.ca
The full paper can be found on Repec: link
When comparing multiple treatments, we want to know:

(A) Whether or not each treatment effect is different from zero
(B) Whether or not each treatment effect is different from all others

With $k$ treatments, this involves making a total of

\[ k \binom{k}{2} + \binom{k+1}{2} \]

unique comparisons (e.g., with 4 treatments, there are a total of 10 comparisons)
We consider the following regression model:

\[ Y_t = \beta_0 \text{CONTROL}_t + \sum_{i=1}^{k} \beta_i \text{TREAT}_{i,t} + z_t' \delta + U_t \]

The (average) treatment effect of the \( i \)th treatment is

\[ \alpha_i \equiv \beta_i - \beta_0, \quad i = 1, \ldots, k, \]

so we want to test

(A) \( \alpha_i = 0 \quad (\iff \beta_i = \beta_0) \), for each \( i \in \{1, \ldots, k\} \)

(B) \( \alpha_i = \alpha_j \quad (\iff \beta_i = \beta_j) \), for each unique pair \((i,j) \in \{1, \ldots, k\}^2\)

or, more simply,

\[ \beta_i = \beta_j, \quad \text{for each unique pair } (i,j) \in \{0, 1, \ldots, k\}^2 \]
NOTE: This is very different from a single joint test:

$$\beta_0 = \ldots = \beta_k$$

(the alternative here is uninformative)
Simple Example: Teacher Incentives

- Field experiment from Muralidharan & Sundararaman (2011)
- Considers the effects of \( k = 2 \) teacher incentive pay treatments:
  - Incentives based on test scores of the teacher’s own students
  - Incentives based on test scores of all students in a teacher’s school
- The effects of these interventions are compared to test scores of students in similar schools (the control group)
- \( Z_t \) includes 49 county dummies and the pre-treatment test score
- Standard errors are clustered by school (we use wild cluster bootstrap when applying our procedure below)
- We focus on combined (math and language) test scores; there are a total of 29,760 obs.
1. Any effect of individual incentive treatment?
   Test $\alpha_1 = 0 \Leftrightarrow \beta_1 = \beta_0$
   $T$-stat: 4.84 \quad (p_{asy} = 1.298 \times 10^{-6})

2. Any effect of group incentive treatment?
   Test $\alpha_2 = 0 \Leftrightarrow \beta_2 = \beta_0$
   $T$-stat: 2.70 \quad (p_{asy} = 0.007)

3. Any difference between individual incentive and group incentive?
   Test $\alpha_1 = \alpha_2 \Leftrightarrow \beta_1 = \beta_2$
   $T$-stat: 1.91 \quad (p_{asy} = 0.056)
Our approach to this multiple testing problem is to seek to control the familywise error rate (FWER): the probability of finding at least one spurious difference (Type I error) between the parameters. It is straightforward to modify our procedure to target control of a less stringent error rate such as the false discovery rate (Benjamini & Hochberg, 1995).
FWER Error Rates

(A) $k$ independent $T$-tests at 5% level
(B) $\binom{k}{2}$ independent $T$-tests at 5% level
Utilize procedure of Bennett & Thompson (2017, JASA), which can be seen as a resampling-based generalization of Tukey’s (1953) procedure.

The approach is to plot each parameter estimate $\hat{\beta}_{n,i}$ together with its corresponding uncertainty interval,

$$[L_{n,i}(\gamma), U_{n,i}(\gamma)] = \left[\hat{\beta}_{n,i} \pm \gamma \times \text{se} \left(\hat{\beta}_{n,i}\right)\right],$$

where $\gamma$ is chosen to control the FWER.

We infer that $\beta_i > \beta_j$ if $L_{n,i} > U_{n,j}$.
Why not use confidence intervals

- Comparisons based on the non-overlap of confidence intervals are not reliable:
- With a single comparison ($k = 1$), non-overlap of CI’s lead to serve under-rejection
- When the number of comparisons grows, non-overlap of CI’s lead to over-rejection
Ideal choice of $\gamma$

- The “ideal” choice of $\gamma$ is the smallest value satisfying

$$
\text{Prob}_P \left\{ \max_{n,i} L_{n,i}(\gamma) > \min_{n,i} U_{n,i}(\gamma) \right\} \leq \alpha
$$

Probability of at least one non-overlap

when all $k$ parameters are equal

- This choice is infeasible since $P$ is unknown
Data-driven choice of $\gamma$

- We choose $\gamma$ to satisfy the bootstrap analogue of the above condition:

$$\operatorname{Prob}_{\hat{\rho}_n} \left\{ \max L_{n,i}^*(\gamma) > \min U_{n,i}^*(\gamma) \right\} \leq \alpha,$$

where

$$[L_{n,i}^*(\gamma), U_{n,i}^*(\gamma)] = \left[ (\hat{\beta}_{n,i}^* - \hat{\beta}_{n,i}) \pm \gamma \times \text{se} \left( \hat{\beta}_{n,i}^* \right) \right],$$
Data-driven choice of $\gamma$: 0.497
Empirical researchers are typically interested only in the $\alpha$ coefficients (the marginal treatment effects).

Accordingly, we can plot $\hat{\alpha}_{n,i}$ along with the re-centered uncertainty interval for $\beta_i$:

$$\left[ \hat{\beta}_{n,i} - \hat{\beta}_{n,0} \pm \gamma \times \text{se} \left( \hat{\beta}_{n,i} \right) \right]_{\hat{\alpha}_{n,i}}$$

We also include the re-centered uncertainty interval for $\beta_0$:

$$\left[ \hat{\beta}_{n,0} - \hat{\beta}_{n,0} \pm \gamma \times \text{se} \left( \hat{\beta}_{n,0} \right) \right]_{0}$$
Teacher Incentives Example: Marginal Treatment Effects

Year 2 Score – Marginal
Gamma – Uncertainty Intervals

Dotted line corresponds to upper endpoint of re-centered uncertainty interval for $\beta_0$
Bennett & Thompson show that, under fairly general conditions, the procedure:

1. Bounds the FWER by $\alpha$ asymptotically
2. Is consistent in the sense that the ordering of all parameter pairs are correctly inferred asymptotically

Simulation evidence in both Bennett & Thompson and Thompson & Webb suggests that the finite sample properties of the procedure are satisfactory
If the procedure fails to resolve all pairwise comparisons, it may be possible to do so via a global refinement which is analogous to the stepdown procedures of Romano & Wolf (2005) and others.
A Modified Procedure

- The above procedure controls the FWER error rate across all pairwise comparisons.
- This approach allows for a (potentially complete) ranking of all the treatments:
  - Assuming larger values of outcome variable are “better”, one could infer that treatment $i$ is the “best” if
    $$L_{n,i} > U_{n,j}, \quad \text{for all } j \neq i$$
  - Similarly, one may be able to identify a “second best” treatment, a “third best” treatment, etc.
While such a complete ranking may occasionally be of value, interest often centers on identifying only the (first) best treatment.

Specifically, we may only want to know whether or not the treatment effect which is estimated to be the largest is actually statistically distinguishable from the other treatments effects (and zero).

Such a problem is the focus of multiple comparisons with the best procedures.

Here, we follow BT in developing a modification of the basic overlap procedure to focus on this problem.
Let \([1], [2], \ldots, [k + 1]\), be the random indices such that

\[
\hat{\beta}_n,[1] > \hat{\beta}_n,[2] > \cdots > \hat{\beta}_n,[k+1]
\]

Note that \(\beta_{[1]}\) is the true value of the parameter which is estimated to be largest, and not necessarily the largest parameter value.

Similarly, \(L_{n,[1]}\) is the lower endpoint of the uncertainty interval associated with the largest point estimate, which is not necessarily the largest lower endpoint (the standard error of \(\hat{\beta}_n,[1]\) might be relatively large).
Similar to before, we infer that $\beta_{[1]}$ is the largest parameter value in the collection if $L_{n,[1]} > U_{n,[j]}$ for all $j > 1$

Our “ideal” choice of $\gamma$ is the smallest value satisfying

$$\text{Prob}_P \left\{ L_{n,[1]}(\gamma) > \max_{j \neq 1} U_{n,[j]}(\gamma) \right\} \leq \alpha$$

when all $k$ parameters are equal

A feasible choice of $\gamma$ is the smallest value satisfying

$$\text{Prob}_{\hat{P}_n} \left\{ L_{n,[1]}^*(\gamma) > \max_{j \neq 1} U_{n,[j]}^*(\gamma) \right\} \leq \alpha$$

This choice of $\gamma$ will be (weakly) smaller than the choice resulting from the basic procedure, leading to greater power
Data-driven choice of $\gamma$: 0.316 (compare with 0.497)
Charitable Giving Example

- Data comes from field experiment by Karlan & List (2007)
- Experiment was designed to examine the effect of matching grants on charitable giving
- Letters sent out to \( n = 50,083 \) previous donors
- 1/3 of letter recipients belonged to control group
- Remaining 2/3 of letter recipients got one of the \( k = 36 \) treatments that varied by
  1. Matching ratio: 1:1, 2:1, or 3:1
  2. Maximum size of matching grant: $25,000, $50,000, $100,000, or none
  3. Amount used as illustration: 1, 1.25, or 1.50 \( \times \) donor’s prev. max.
Charitable Giving Example