EC 709: Discussion 2

Liang Zhong[1]

Boston University

samzl@bu.edu

September 2023

[1]Thanks to Rubaiyat Alam for kindly sharing his lecture notes

Overview

Table of Contents

When to adjust for MHT?

1. Multiple outcomes of interest
   - For a treatment on 7 different outcome variables, #tests = 7

2. Multiple measures of the same treatment (e.g., 5 measures of weather: temperature, wind speed, etc)
   - Regress each measure separately on the outcome ⇒ correct for MHT
   - Regress all 5 measures in the same regression: if care about the coefficient on each ⇒ correct for MHT
   - Regress all 5 measures in the same regression: If interested in whether at least one of these is significant ⇒ use an F-test

When to adjust for MHT? (cont.)

3. Multiple subgroups to identify mechanism or heterogeneous treatment effect
   - If 5 groups and 3 coefficients of interest, $\#\text{tests} = 5x3 = 15$

4. Multiple treatments are of interest and desired to determine which treatments have an effect relative to either the control or each of the other treatments
   - Run 1 regression, 10 covariates, 10(5) coefficients of interest: adjust for MHT with $\#\text{tests} = 10(5)$

How to adjust for MHT?

- Controlling for some criterion so that the more tests are carried out, the more difficult it gets to reject a null
  1. Controlling Family-wise Error Rate (FWER)
  2. Controlling False Discovery Rate (FDR)

- References and Good open resources:
  - STATA illustration by Damian Clarke
  - Blog for STATA implementation by David Mckenzie

Notation

We want to test n individual null hypothesis $H_{0,1}, H_{0,2}, ..., H_{0,n}$

	Accepted	Rejected	total
True	U	V	n_0
False	T	S	$n - n_0$
total	$n - R$	R	n

- $FWER = Pr(V \geq 1) = 1 - Pr(V = 0)$
  - Limit the probability of making at least one false discovery
  - Stringent control over false discoveries (Type I error)
  - Reasonable to aspire to no false discoveries when $n = 10$, but less reasonable when $n = 100$ (or more)
- $FDR = E[V/R]$ if $R > 0$; 0 if $R = 0$
  - Limit the proportion of false discoveries
  - $FDR \leq FWER \Rightarrow$ less control over false discoveries, but often at greater power
  - For historical reason size of FDR refer to q rather than α

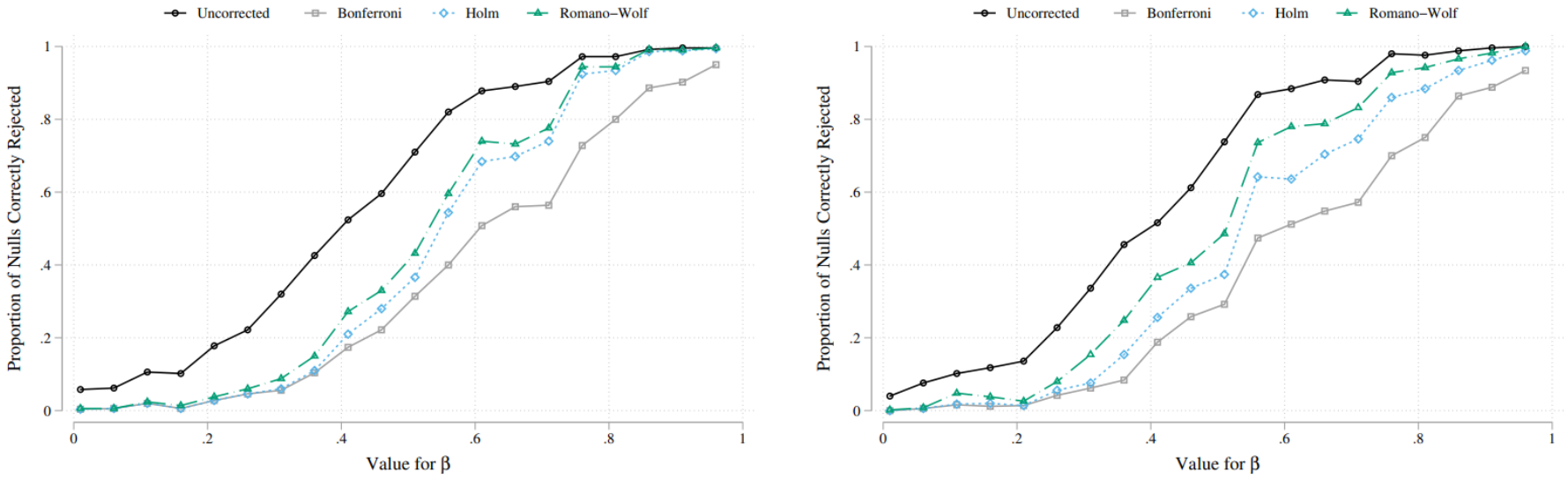
FWER adjustment methods

- Bonferroni (1935) or Sidak (1967)
  - Mixing true and false nulls: Bonferroni's $FWER \leq \frac{n_0}{n}\alpha$
  - $\Rightarrow$ Too conservative, power for false nulls are often affected
  - Assumes that tests are independent
- Holm (1979) (step down)
  - Including both true and false null in the setting so less conservative
  - Still too conservative when tests are not independent
- Romano-Wolf (2005b)
  - Uses a bootstrapping approach to incorporate information about the joint dependence
  - $\Rightarrow$ Section 2.2 of Clarke, Romano, and Wolf (2020) provides good details
  - List, Shaikh, and Vayalinkal (2023) extend the framework by allowing covariate adjustment to increase power

Power Comparison by Damian Clarke

ρ is the correlation between the tests

Figure: Simulated Power to Reject False Null Hypotheses



(a) $\rho = 0.25$ (b) $\rho = 0.75$

Refer to section (4) of the accompanying Stata code `multHyp.do`.

Notes on implementing Romano-Wolf

- **rwolf** is the original STATA program
  - Only allows one treatment variable of interest: Putting in two variable names causes **rwolf** to run the algorithm for two separate single-treatment regressions
- Clarke, Romano, and Wolf (2020) proposed a new version **rwolf2**
  - Now allows for multiple treatments, different commands, different controls in different regressions, and also allows for clustered standard errors
  - David Mckenzie: "seems the theoretically best option for FWER correction at the moment"

```
rwolf2 (areg Y1 treat1 treat2 treat3 treat4, r a(strata)) ///
(areg Y2 treat1 treat2 treat3 treat4, r a(strata)) ///
(areg  Y3 treat1 treat2 treat3 treat4 b_Y3, r a(strata)) ///
(areg Y4 treat1 treat2 treat3 treat4 b_Y4, r a(strata)) ///
(areg Y5 treat1 treat2 treat3 treat4 b_Y5,  r a(strata)), ///
indepvars(treat1 treat2 treat3 treat4, treat1 treat2 treat3 treat4, ///
treat1 treat2 treat3 treat4, treat1 treat2 treat3 treat4, ////
treat1 treat2 treat3 treat4) usevalid seed(123) reps(3000)
```

Figure 1: Example of the syntax

FDR adjustment methods

Order the p-values $p_{(1)}, p_{(2)}, ..., p_{(n)}$ with $H_{(1)}, H_{(2)}, ..., H_{(n)}$ be the corresponding hypotheses, and n_0 is the number of true null hypotheses

- Benjamini-Hochberg (1995) (size is refer to q which is similar to α)
  - For a given q, find the largest k such that $P(k) \leq \frac{k}{n} q$
  - Reject all $H_{(i)}$ for $i = 1, ..., k$
  - $FDR \leq \frac{n_0}{n} q \Rightarrow$ still conservative (Benjamini and Yekutieli, 2001)

- Benjamini, Krieger and Yekutieli (2006): Sharpened q-values
  1. Apply the BH procedure at level $q' = q/(1+q)$. Let c be the number of hypotheses rejected. If $c = 0$, stop; otherwise, continue to step 2.

     $\Leftarrow$ Estimates the number of true hypotheses
  2. Let $\hat{n}_0 = n - c$
  3. Apply the BH procedure at level $q^\star = q' n / \hat{n}_0$

About Benjamini, Krieger and Yekutieli (2006)

- Does not work well if p values are negatively correlated

⇐ Need a more conservative modification (Benjamini and Yekutieli, 2001)

- Michael Anderson has a good program for implementation in practice

- All the programs introduced above also report the adjusted p-values −− the natural analog to the standard p-value
  - the smallest level α at which the hypothesis would be rejected

  ⇐ Performing the procedures for all possible α levels (e.g., 1.000, .999, .998) and recording when each hypothesis ceases to be rejected

  - No need to rerun the previous programs for different α, just compare adjusted p-values with any level α

Comparison of different methods by David Mckenzie

Comparing impacts on just treatment 1 outcomes using rwolf to other methods

	Y1	Y2	Y3	Y4	Y5
Treat 1	0.022	0.043	0.083**	0.079***	0.032
p-value	(0.516)	(0.258)	(0.031)	(0.001)	(0.178)
sharpened q-value	[0.422]	[0.240]	[0.067]	[0.006]	[0.217]
mhtexp FWER p-value	{0.518}	{0.470}	{0.078}	{0.002}	{0.391}
rwolf FWER p-value	<0.527>	<0.417>	<0.060>	<0.003>	<0.327>

- Sharpened q-values can actually be LESS than unadjusted p-values in some cases when many hypotheses are rejected

$\Leftarrow$ If there are many true rejections, you can tolerate several false rejections too, and still maintain the false discovery rate low

- For only 5 tests, the power advantage from FDR is not obvious

Which one should we choose FWER vs. FDR?

The choice between FWER and FDR adjustments may be dominated by the **cost of a false rejection**:

- FWER control limits the probability of making **any** type I error

  $\Rightarrow$ All rejections will be correct with high probability

  - Well-suited to cases in which the cost of a false rejection is **high**

  e.g., Incorrectly concluding some interventions are effective could result in a large-scale misallocation of resources

Which one should we choose FWER vs. FDR?(cont.)

- FDR control allows a small number of type I errors in exchange for greater power than FWER control

  $\Rightarrow$ A high probability that some false positives will occur

  - If the cost of a false rejection is **low to moderate**, then the increased power of FDR control will be appealing, particularly if the family of hypotheses being tested is large

  e.g., In exploratory analysis, we may be willing to tolerate some type I errors in exchange for greater power

- However, if the number of tests is not too large, we should stick with FWER

MHT and two-step procedures

- In practice, we might use MHT to determine our main specification and involve a two-step procedure:

  1. Run "long" model (including main and interaction effects). If coefficients on interactions are significant, stop; Otherwise, continue to step 2
  2. Run "short" model (that ignores interactions) for higher power

- Muralidharan, Romero and Wuthrich (2020): Naive use of inference procedures can be highly misleading
- Generally, we need to adjust the inference method for all "post-model-selection estimators"

Problem of Post-model selection

- What are "post-model-selection estimators"? Two steps:
  1. Select the model you want to estimate, based on:
     - MHT
     - Optimization of a penalized goodness-of-fit criterion (e.g., AIC, BIC): In time series, choose the k for the AR(k) model
     - Cross-validation methods: Choose the parameters for Machine learning models
  2. Estimate the selected model for the parameter of interest
- The sampling properties of post-model-selection estimators are typically significantly different from the nominal distributions that arise if a fixed model is supposed (Leeb and Potscher, 2005)
- Details would be covered in EC711

Table of Contents

Randomization Inference: Sharp null of no effect

- Causal inference is a missing data problem! (Rubin, 1975)
- Sharp Null Hypothesis of No Effect: $H_0 : Y_i(\text{Treated}) = Y_i(\text{Control})\ \forall i$
  - "Sharp": The treatment effect is zero for **all** subjects

  $\Rightarrow$ Implies $ATE = 0$ and much stronger

  E.g., If the treatment effect is 5 for half the subjects and -5 for the other half, ATE is 0, but sharp null is false

- Under the Sharp Null, we solved our missing data problem by assuming $Y_i(\text{Treated}) = Y_i(\text{Control})\ \forall i$!

? All potential outcomes are fixed. How do we conduct the testing?

Reference: Imbens and Rubin (2015), Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction

Sources of Uncertainty for Estimation

1. Sampling variability: when choosing units from the population to sample
   - Estimation would vary due to heterogeneity from sample to sample
   - $\Leftarrow$ The same thing applies when you construct the sample mean for the random variable
   - The larger the sample size, the smaller the sampling variation
   - Bootstrapping inference plays with this sampling uncertainty

- However, there are many cases where there is no sampling
  - If you have county-level data in the U.S. and observe all counties, that's the relevant population
  - You still get a standard error when you run a regression using those datasets
- How do we understand the standard error we have from the regression? Where is the variation?

Sources of Uncertainty for Estimation (cont.)

2. Assignment variability: When choosing the units being treated
   - Estimation would be different due to heterogeneity between treated and control group

   $\Rightarrow$ Estimation vary with assignment vectors

- Accounting assignment uncertainty is a new philosophy for causal inference

  $\Leftarrow$ MHE (2009): the assignment vector is assumed to be fixed over the whole population
  - Called **Design-based inference**: Abadie et al (2020); Card (2022)
  - However, it is in the same spirit as Randomization Inference (RI)

- RI: condition on the potential outcomes and simulate over the random treatment assignments

  $\Leftarrow$ Only care about the estimation of the sample of subjects we have (only care about finite sample properties)
  - Only uncertainty comes from assignment variability

Intuition of Fisher Randomization Testing (1935)

- Consider Test Statistics $T = T(Y(z), z)$:
  - Z: Binary stochastic treatment assignment vector; $z_i = 1$ if treated, $z_i = 0$ if control; Observed Assignment, $Z^{obs} = (Z_1, ..., Z_N)$
  - $Y(z)$: Potential outcome vector under assignment z; Observed Outcome, $Y^{obs} = (Y(Z_1), ..., Y(Z_N))$
  - $Z \sim P(Z)$, the treatment design is random and **known**

- Z is random $\Rightarrow$ T is also random and has a distribution under the null

Step 1. Simulate all possible random assignments $\rightarrow$ exact sampling distribution of T

Step 2. Compare the actually observed value of the test statistic $T^{obs} = T(Y^{obs}, Z^{obs})$ against this distribution

$\Rightarrow$ An observed value that is "very unlikely" will be taken as evidence against the null

- a stochastic version of "proof by contradiction"

Review of FRT Procedure

- SUTVA: no interference, $Y_i(z)$ depends only on z_i
  - Only two potential outcomes, $Y_i(0)$, $Y_i(1)$, for every i.
- H_0 : $Y_i(0) = Y_i(1)$, for every i

FRT procedure

Inputs: $T = T(Y(z), z)$, Z^{obs}, Y^{obs}, P.

1. Calculate: $T_{obs} = T(Y^{obs}, Z^{obs})$.
2. Randomly sample: $Z' \sim P(Z')$, store T_r

$$T_r = T(Y^{obs}, Z') \overset{H_0}{=} T(Y', Z') \overset{\text{d}}{=} T(Y^{obs}, Z^{obs}) = T_{obs}$$

.

3. Obtain p-value: $pval = E[1\{\| T_r \| \geq \| T_{obs} \|\}]$.

Output: Reject if p-value$< \alpha$.

Randomization Inference: Example

- Total of 7 units, and assign treatment to 2 of them
- $\Leftarrow$ The number of all possible assignment vectors: $\frac{7!}{2!5!} = 21$
- Observe the following values after randomization under the null:

$$-7.5, -7.5, -7.5, -4.0, -4.0, -4.0, -4.0, -0.5, -0.5, -0.5,$$

$$-0.5, -0.5, -0.5, 3.0, 3.0, 6.5, 6.5^{obs}, 6.5, 10.0, 10.0$$

- Calculating p-values:
  - Take the absolute value for all the outcomes
  - 8 estimates ≥ 6.5 or ≤ -6.5, hence $pval = 8/21 \approx 0.38$
  - cannot reject sharp null under the typical choice of α
- In practice, I recommend using absolute value and doing the one-side test. Since the Two-side test would have a power issue in some cases

Some technical Notes in practice

- Obtain exact sampling distribution is impossible for large N
  - For N = 50 and 25 treatment assignments: over 126 trillion assignment vectors
  - Looking over every possible randomization becomes impractical
  - $\Leftarrow$ Approximate the sampling distribution by sampling at random from the **set of** all possible assignment vectors
- The statistic should be sensitive to the difference between the null and alternative (have statistical power)
  - See section 5.5 of Imbens and Rubin (2015) for a nice discussion on the choice of the test statistic

Some technical Notes in practice (cont.)

- Need to know the experimental design, $P(Z)$
  - Easy to obtain in Experimental settings
  - Might assume it is completely random when you are using randomization inference in observational studies
- Any "sharp null" can be used: a null hypothesis that allows us to infer all the missing potential outcomes from the observed outcomes. e.g., $Y_i(0) = Y_i(1) + c, \forall i$

$\Rightarrow$ In practice, researchers often use RI when:

  - Sample size is too small (Typically the case in the experimental samples)
  - One of the treatment and control group sizes is too small
  - Underlying data are distributed nonnormally in some case
- Why is RI still useful after nearly 100 years?

Randomization Inference: Pros

1. Non-parametric, no functional form, or homogeneity assumption

⋆2. Good finite sample properties: do not rely on asymptotic theory or distributional assumptions

- In small sample, conventional tests based on asymptotic theory may be misleading

⇐ t-tests based on Robust standard errors over-reject when the null hypothesis is true, and the sample is not large

  - FRT vs. Robust Standard Error is a good blog comparing their performance in practice

  ⇐ It seems some new methods for robust standard error can be as good as randomization inference

Randomization Inference: Cons

⋆1. Sharp Null is too strong

⇐ Zhao and Ding (2021) proposed an extension for testing the weak null hypothesis of zero average treatment effect

Step 1. Run OLS fit of the observed outcome on the treatment, centered covariates, and their interactions for covariate adjustment

Step 2. Treat the robust t-value of the treatment as the test statistics, conduct randomization testing

- Asymptotically valid for the weak null and finite sample valid for the strong null

⇐ Irrespective of whether the linear model is correctly specified or not

2. Sample inference rather than population inference

Thank You!