Lecture 19: Multiple Endpoints

POL-GA 1251
Quantitative Political Analysis II
Prof. Cyrus Samii
NYU Politics

April 18, 2022

Motivating Example

ESTIMATING THE IMPACT OF THE HAJJ: RELIGION AND TOLERANCE IN ISLAM'S GLOBAL GATHERING*

DAVID CLINGINGSMITH
ASIM IJAZ KHWAJA
MICHAEL KREMER

(Clingingsmith et al. 2009)

Motivating Example

TABLE IX
SELECTED SURVEY QUESTIONS

Question	Coding	Coef.	p-value	Comp. mean	Obs.	R^2
(1) Do you believe others regard you as religious?	1 = Religious, 0 = Not religious	0.100	.000	0.772	1,541	.033
(2) Do you pray "Tahajjud Namaz"?	1 = Yes (regularly, occasionally), 0 = No (rarely, never)	0.184	.000	0.281	1,605	.047
(3) How often did you fast outside of Ramadan during the past year?	1 = Several times per month or more, 0 = Once per month or less	0.041	.006	0.049	1,605	.030
(4) Is your general view of Indonesian people:	2 = Very positive, −2 = Very negative	0.217	.000	0.362	1,583	.055
(5) Is your general view of Saudi people:	2 = Very positive, −2 = Very negative	0.110	.026	1.034	1,593	.026
(6) In your opinion, overall how are people of a different religion compared to your people?	0 = Better or worse, 1 = Same	0.084	.004	0.389	1,604	.025
(7) Do you believe that people of different religions can live in unity & agreement (harmony) in a given society by making agreements over their differences?	1 = Yes, 0 = No	0.063	.074	0.589	1,270	.036
(8) Do you ever pray in the mosque of a different maslak than your own?	Binary: 1 = Frequently, 0 = Less often/never	0.034	.021	0.049	1,463	.027
(9) Do you believe the goals for which Osama is fighting are correct?	1 = Not correct at all/slightly incorrect, 0 = Correct/absolutely correct	0.063	.014	0.068	761	.054
(10) Do you believe the methods Osama uses in fighting are correct?	1 = Absolutely never/almost never correct, 0 = To small extent/some extent/strongly correct	0.051	.112	0.159	761	.063
(11) How important do you believe peace with India is for Pakistan's future?	1 = Important, 0 = Not important	0.044	.016	0.913	1,155	.020
(12) Please tell me what you think about the correctness of the following: family members physically punishing someone who has dishonored the family	0 = Correct, 1 = Never correct	0.044	.112	0.261	1,459	.033
(13) In your opinion, how do men and women compare to each other with respect to the following traits: spiritually	0 = Men are better/equal, 1 = Women are better	0.057	.006	0.111	1,497	.034
(14) What is your opinion about the quality of women's lives in each of the following countries/regions? Indonesia/ Malaysia	1 = Greater than in Pakistan, 0 = Lower than or equal that in Pakistan; Base variables 5 = Very high, 1 = Very low	0.094	.088	0.262	551	.058
(15) What is your opinion about the quality of women's lives in each of the following countries/regions? Saudi Arabia	1 = Greater than in Pakistan, 0 = Lower than or equal that in Pakistan; Base variables 5 = Very high, 1 = Very low	0.051	.145	0.322	1,180	.048
(16) What is your opinion about the quality of women's lives in each of the following countries/regions? West	1 = Greater than in Pakistan, 0 = Lower than or equal that in Pakistan; Base variables 5 = Very high, 1 = Very low	0.087	.051	0.186	646	.091
(17) Do you think there are too many crimes against women in Pakistan? Overall	Binary: 0 = No, 1 = Yes	0.052	.075	0.597	1,605	.045
(18) Do you think there are too many crimes against women in Pakistan? Relative to men	1 = Against women score < against men score, 0 = Against women score ≥ against men score; Base scores 1 = Yes, a lot; 4 = No, not at all	0.053	.052	0.171	1,135	.026
(19) In your opinion, girls should attend school	Binary: 0 = Disagree, 1 = Agree	0.028	.039	0.933	1,604	.027
(20) Until what level would you prefer allow/permit girls in your family to attend coeducational schools (boys and girls in the same school)?	0 = Never, 1 = Primary, secondary, or all levels	0.055	.036	0.722	1,550	.035
(21) Until what level would you prefer allow/permit boys in your family to attend coeducational schools (boys and girls in the same school)?	0 = Never, 1 = Primary, secondary, or all levels	0.059	.024	0.729	1,550	.036
(22) Would you like for your daughters or female grand-children to have a career other than caring for the household?	0 = No, 1 = Yes	0.045	.156	0.540	1,605	.029
(23) How important are the following characteristics in your son's, grandson's wife?: Good employment or business	0 = Not important, 1 = Important	0.054	.073	0.457	1,562	.028

How to interpret without being statistically reckless?

Motivating Example

TABLE V
TOLERANCE

	AES coefficients		
	Base	Controls	Restricted subsample
(1) Views of other countries	0.150***	0.147***	0.151***
	(0.04)	(0.04)	(0.04)
(2) Views of other groups	0.131***	0.108**	0.122**
	(0.05)	(0.05)	(0.06)
(3) Harmony	0.128***	0.117***	0.126***
	(0.04)	(0.04)	(0.05)
(4) Peaceful inclination	0.111***	0.121***	0.128***
	(0.03)	(0.03)	(0.04)
(5) Political Islam index	−0.050	−0.044	−0.043
	(0.04)	(0.03)	(0.04)
(6) Views of West	0.029	0.039	0.011
	(0.04)	(0.04)	(0.04)

Notes. See notes to Table IV. Index component questions with number of components indicated in parentheses: Index 1 (6): General view of people from other countries, positive to negative: Saudis, Indonesians, Turks, African, Europeans, Chinese. Index 2 (3): How do members of the following groups compare to your group: different sect? different religion? different ethnicity? Index 3 (4): Do you believe the following groups can live in unity and harmony through compromise over disagreements: sects of Islam? religions? Pakistani ethnic groups? Do you ever pray in a mosque of a different school of thought? Index 4 (8): Belief in incorrectness of: Osama's goals? Osama's methods? How important is peace with India for Pakistan? Should the current India/Pakistan boundary be the permanent border if this leads to peace? Should Pakistan not support/only partly support those fighting the Indian government in Kashmir? How incorrect are: suicide attacks? attacks on civilians in war? physical punishment of someone who dishonors family? Index 5 (5): Agree that: government should enforce Islamic injunctions? religious leaders have right to dispense justice? religious leaders should have direct influence on government? better for politicians/officials to have strong religious beliefs? religious beliefs important in voting for candidate? Index 6 (4): Is it bad for Pakistanis to adopt: Western social values? Western technology? Believe there was Western/Jewish role in 9/11 and 2005 London bombing? Believe West does not take into account interests of countries such as Pakistan?

Testing and Error Rates

- We've focused on *estimation* of coefficients and standard errors.
- Methods we have learned are valid no matter how many outcomes you have, even when you have a lot of related outcomes:
  - A consistent estimate remains consistent even when you add other outcomes to the analysis.
  - A consistent standard error estimate remains consistent even when you add other outcomes to the analysis.
  - Possible to "borrow strength" across outcomes although put that aside for the moment (we will get back to this later today).
- New issues arise when it comes to testing statistical significance.

Testing and Error Rates

- In testing, we fix a hypothesis and alternative.
- We determine what is the probability of obtaining our estimate under the null, given the alternative. This is the p-value.
- Typically, we then establish a rejection criterion based on a confidence level $(1-\alpha)$: if $p < \alpha$, we reject the null.

Testing and Error Rates

- Suppose $D_i = 0, 1$ and potential outcomes Y_{1i} and Y_{0i}, we have random assignment, and we wish to estimate $\rho = \mathrm{E}[Y_{1i} - Y_{0i}]$.
- There are two types of null hypotheses and (two-sided) alternatives about the causal effect of D_i:
- Sharp null:

$$H_{SN} : Y_{1i} = Y_{0i} \text{ for all } i \text{ vs. } H_{SN}^{A} : Y_{1i} \neq Y_{0i} \text{ for some } i.$$

- Average null:

$$H_{AN} : \rho = 0 \text{ vs. } H_{AN}^{A} : \rho \neq 0.$$

- $H_S \Rightarrow H_A$ but $H_A \nRightarrow H_S$.
- Standard regression t-test output provides p-value for H_{AN}.

Testing and Error Rates

- Consider two-sided t-test for H_{AN}.
- Suppose simple random sampling of N units from large population, OLS with K regressors (including constant), and normal population residuals. Under H_{AN},

$$\frac{\hat{\beta}}{s.e.(\hat{\beta})} \overset{H_{AN}}{\sim} t_{N-K}.$$

  (For non-normal residuals, this is a finite-sample-adjusted approx. based on asymp. normal distribution of $\hat{\beta}$.)
- By this fact, it is apparent that for any α

$$\Pr\left[\left|\frac{\hat{\beta}}{s.e.(\hat{\beta})}\right| > t_{\alpha/2} \middle| H_{AN}\right] = \alpha > 0.$$

Testing and Error Rates

- There is always *some* chance of rejecting *even if null is true*.
- More tests means higher chance this occurs at least once.

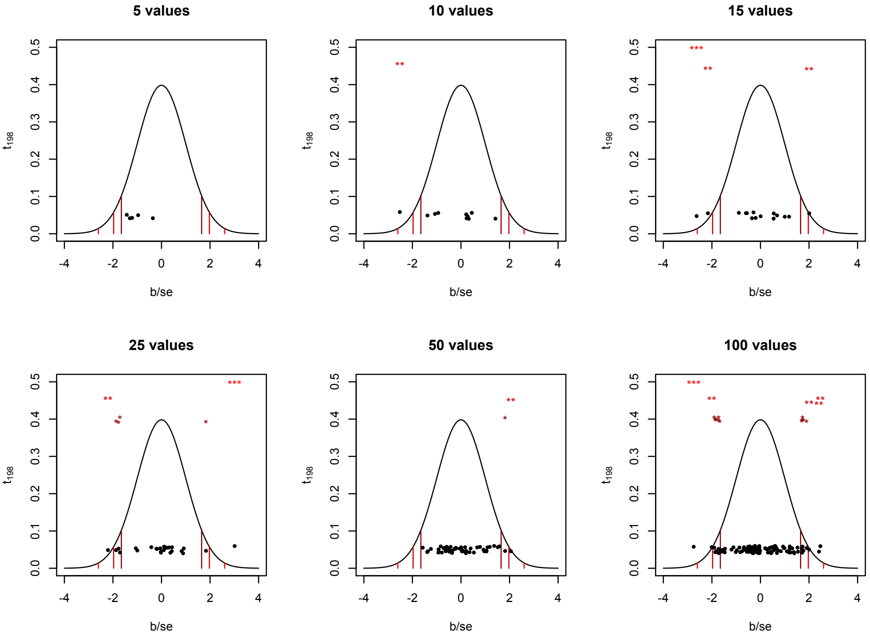
Testing and Error Rates

- A simulation demonstrates:

```
N <- 200
i <- 1:N
draw.vec <- c(5,10,15,25,50,100)
for(j in 1:length(draw.vec)){
  b.se <- rep(NA, draw.vec[j])
  for(s in 1:draw.vec[j]){
    Y0 <- rnorm(N)
    Y1 <- rnorm(N)
    D <- rep(0,N)
    D[sample(i, floor(N/2))] <- 1
    Y <- D*Y1 + (1-D)*Y0
    fit <- lm(Y~D)
    b.se[s] <- summary(fit)[[4]][2,3]
  }
}
```

- `draw.vec` is number of outcomes analyzed.

Testing and Error Rates



Testing and Error Rates

- The "stars" are all false positives.
- We *rejected the null when we should not have*.
- These are "type I" errors.
- This is an instance of the "multiple comparisons" problem.
- When we look at lots of outcomes, the concern is that what we take be a "significant" effect is really just noise.

Testing and Error Rates

- Think about what this means for studies that look at lots of outcomes. How do you know what you've found is "real"?

Testing and Error Rates

- We could just as well have set up a simulation where H_{AN}^{A} was true in all cases and we still sought to test H_{AN}.
- Then, there is positive probability that we would *fail to reject the null* in some tests.
- These would be false negatives or "type II" errors.

Testing and Error Rates

- If we test M hypotheses in one analysis, the M hypotheses are called a family.
- Suppose that for $J \leq M$ of tests the null hypothesis is true.
- The "familywise error rate" (FWER) is the probability that at least one of the J null hypotheses is rejected.
- The "false discovery rate" (FDR) is the expected proportion of rejections that will be false rejections.
- Adjusting the testing procedure to reduce either FWER or FDR will reduce the number of false rejections.
- Reducing FWER is more stringent than FDR.

Classical Approaches to Dealing with Multiple Endpoints

Two ways to reduce testing error:

1. *Reduce the number of tests* applied to a family by aggregating information and then applying multivariate “omnibus” test.
2. *Adjust p-values or critical values* in individual tests.

There are goals other than reducing testing error when doing inference with multiple outcomes:

- Making summary statements about “overall effectiveness” of a treatment by incorporating information from many outcomes.
- Examining what are the “key drivers” of overall effectiveness.

Classical Approaches to Dealing with Multiple Endpoints

- Classical method for multivariate "omnibus" testing is Hotelling's T^2.
- It generalizes the t-test to multiple outcomes.
- It amounts to testing whether the mean values of the different outcomes are the same under treatment and control.
- Problems:
  - No distinction between positive and negative mean deviations.
  - Does not allow us to test for "overall efficacy" of a treatment.
  - Just tests whether patterns of outcome means are different. (O'Brien, 1984).
- These are problems with many multivariate tests.

Classical Approaches to Dealing with Multiple Endpoints

- Classical p-value adjustment uses Bonferroni union bound.
- Suppose 2 tests of H_{01} and H_{02} with confidence α for each.
- If H_{0m} true, probability of rejection is α, non-rejection $1-\alpha$.
- Define A_m as event that H_{0m} *not* rejected. If both nulls are true, probability of at least one rejection is,

$$\begin{aligned}
\text{FWER} = 1 - \Pr[A_1 \cap A_2] &= 1 - (\Pr[A_1] + \Pr[A_2] - \Pr[A_i \cup A_2]) \\
&= 1 - \Pr[A_1] - \Pr[A_2] + \underbrace{\Pr[A_i \cup A_2]}_{\leq 1} \\
&\leq 2 - \Pr[A_1] - \Pr[A_2] = 2 - 2(1-\alpha) = 2\alpha.
\end{aligned}$$

- So, adjusting confidence for each test to use $\alpha_B = \alpha/2$ means:

$$\text{FWER} \leq 2\frac{\alpha_B}{2} = \alpha.$$

- We are at least "$1-\alpha$ confident there are no false discoveries among the rejected hypotheses" (Romano et al. nd.)

Classical Approaches to Dealing with Multiple Endpoints

- Generally, for M tests, we use α/M to ensure FWER $\leq \alpha$.
- Equivalently, you can multiply p-values by M and test against α.
- Problem: Bonferroni correction can be way over-conservative:
  - Does not account for outcome correlations and thus dependence between p values.
    - Consider the case were all M outcomes are perfectly correlated. Then, FWER equals α and no need for adjustment. Bonferroni ignores that.
    - This makes Bonferroni and other tests assuming independent p values "suboptimal in terms of power" (Romano et al.)
  - Mechanically rises in M, possibly yielding adjusted $p > 1$.

Modern Approaches to Dealing with Multiple Endpoints

Modern Approaches to Dealing with Multiple Endpoints

Multiple Inference and Gender Differences in the Effects of Early Intervention: A Reevaluation of the Abecedarian, Perry Preschool, and Early Training Projects

Michael L. ANDERSON

The view that the returns to educational investments are highest for early childhood interventions is widely held and stems primarily from several influential randomized trials—Abecedarian, Perry, and the Early Training Project—that point to super-normal returns to early interventions. This article presents a de novo analysis of these experiments, focusing on two core issues that have received limited attention in previous analyses: treatment effect heterogeneity by gender and overrejection of the null hypothesis due to multiple inference. To address the latter issue, a statistical framework that combines summary index tests with familywise error rate and false discovery rate corrections is implemented. The first technique reduces the number of tests conducted; the latter two techniques adjust the p values for multiple inference. The primary finding of the reanalysis is that girls garnered substantial short- and long-term benefits from the interventions, but there were no significant long-term benefits for boys. These conclusions, which have appeared ambiguous when using "naive" estimators that fail to adjust for multiple testing, contribute to a growing literature on the emerging female–male academic achievement gap. They also demonstrate that in complex studies where multiple questions are asked of the same data set, it can be important to declare the family of tests under consideration and to either consolidate measures or report adjusted and unadjusted p values.

KEY WORDS: False discovery rate; Familywise error rate; Multiple comparisons; Preschool; Program evaluation.

- Anderson (2008) presents a modern approach to handling multiple inference.
- His methods overcome problems of the classical approaches.

Modern Approaches to Dealing with Multiple Endpoints

The substantive problem that he analyzes:

- Debate over early intervention (pre-school) programs on long-term developmental outcomes.
- Three major randomized field experiments: ABC, PPP, and ETP in North Carolina, Michigan, and Tennessee.
- These studies each assess short-, medium-, and long-term outcomes on a number of dimensions.
- Looking at outcomes one-by-one, you get a mixed bag of positive, negative, and null effects.
- This has led to conflicting interpretations.

Modern Approaches to Dealing with Multiple Endpoints

Anderson proposes a unified analysis to make *summary judgments* about effectiveness while also *exploring drivers* of any positive effects:

- Define groups of outcomes that should be assessed jointly and perform inference with summary indices of these outcomes.
- To make summary judgments, correct FWER in testing a collection of indices, but in a manner that does not go overboard like Bonferroni.
- To explore drivers, use a more permissible p-value correction based on FDR minimization to explore effects on raw outcomes.
- Combination of confirmatory and exploratory analysis.

Summary Index Methods

- Anderson examines 47 outcome variables ranging from IQ test scores at different ages, to grades at different ages, college attendance, employment, and criminal record.
- These are aggregated and analyzed in a set of thematic indices.
- The summary index method:
  - Automatically reduces error rate,
  - Provides measure of "overall effect" of program, and
  - Potentially increases power: marginally significant, noisy results have the potential to aggregate into a cleaner statement of actual significance.

Summary Index Methods

- Summary index constructed using inverse covariance weighting (ICW).
- ICW provides optimal linear aggregation of information for a set of noisy measures of a common latent factor (O'Brien, 1984).
- Distinct from factor scoring via factor analysis or principal component analysis.
- Factor scoring methods hunt out different dimensions of variability. ICW optimally collapses into *one* dimension.
- ICW ensures "outcomes that are highly correlated with each other receive less weight when added into the index [given their redundancy], while outcomes that are uncorrelated and thus represent new information receive more weight" (Anderson 2008, 1485).

Summary Index Methods

- *Inverse covariance weighting* optimizes information content for index constructed from *items determined to be related a priori*. Equiv. to a single factor latent variable model:

$$\begin{pmatrix} Y_{1i} \\ \vdots \\ Y_{Ki} \end{pmatrix} = \begin{pmatrix} z_i + \varepsilon_{1i} \\ \vdots \\ z_i + \varepsilon_{Ki} \end{pmatrix}$$

- ICW is then equivalent to fitting this as a varying intercept regression using FGLS.

Summary Index Methods

- *Contrast with factors scores* or *principal component scores* isolate and extract shared variation in *different* latent dimensions. Equivalent to a multifactor linear latent variable model with orthogonal factors:

$$\begin{pmatrix} Y_{1i} \\ \vdots \\ Y_{Ki} \end{pmatrix} = \begin{pmatrix} \beta_1 z_{1i} + \ldots + \beta_K z_{Ki} + \nu_{1i} \\ \vdots \\ \beta_1 z_{1i} + \ldots + \beta_K z_{Ki} + \nu_{Ki} \end{pmatrix}$$

  where $\mathbf{z}_k' \mathbf{z}_l = 0$ for all $k \neq l$.
- (IRT models are analogous approaches that use GLMs for binary, categorical, etc. variables.)

Summary Index Methods

- Choice of ICW vs factor scores/principal component scores depends on the ways that the indicators correlate with each other.
- Anderson uses ICW but this may not always be the best choice...

Summary Index Methods

ICW summary index method steps:

- Scale all outcomes so that larger values always mean "better."
- Standardize outcomes (e.g., subtract pooled mean and divide by control group standard deviation). Label the standardized outcome vector $\tilde{y}$.
- Assign each outcome to one of J thematic groupings. Label outcome vectors as $\tilde{y}_{jk}$, giving K_j outcome vectors in grouping j indexed by k.

Summary Index Methods

Outcome data are thus,

$$\tilde{\mathbf{Y}} = \begin{pmatrix} & & & & \vdots & & & \\ \cdots & \tilde{y}_{ijk} & \tilde{y}_{ij,k+1} & \tilde{y}_{ij,k+2} & \cdots & \tilde{y}_{ij,K_j} & \tilde{y}_{i,j+1,1} & \cdots \\ \cdots & \tilde{y}_{i+1,jk} & \tilde{y}_{i+1,j,k+1} & \tilde{y}_{i+1,j,k+2} & \cdots & \tilde{y}_{i+1,j,K_j} & \tilde{y}_{i+1,j+1,1} & \cdots \\ & & & & \vdots & & & \end{pmatrix}$$

Summary Index Methods

- Create the index, $\bar{s}_{ij}$, by taking ICW weighted average of the K_j standardized outcomes for individual i in grouping j:

$$\bar{s}_{ij} = (\iota'_{K_j}\hat{\Sigma}_j^{-1}\iota_{K_j})^{-1}(\iota'_{K_j}\hat{\Sigma}_j^{-1}\tilde{y}_{ij}),$$

  where $\hat{\Sigma}_j$ for the K_j outcomes in grouping j.
- Matrix implementation to create vector of indices for all N units:

$$\bar{s}'_j = (\iota'_{K_j}\hat{\Sigma}_j^{-1}\iota_{K_j})^{-1}(\iota'_{K_j}\hat{\Sigma}_j^{-1}\tilde{Y}'_j),$$

- NB: cannot have any missing data.
- For interpretability, scale $\bar{s}_{ij}$ again, centering on the control group mean and standard deviation of the new index.
- Yields a standardized index on the scale of control group standard deviations.

Summary Index Methods

Approaches to analysis:

- Anderson uses $\bar{s}_{ij}$ as the outcome in regressions to estimate an "overall effect" for grouping j, measured in terms of control group standard deviations.
  - For inference, just treats the index as a regular outcome to obtain a p-value, although this may overstate precision since it does not account for estimating $\hat{\Sigma}_j$.
  - Improvement might come from bootstrapping the whole process.

Summary Index Methods

Approaches to analysis:

- Anderson uses $\bar{s}_{ij}$ as the outcome in regressions to estimate an "overall effect" for grouping j, measured in terms of control group standard deviations.
  - For inference, just treats the index as a regular outcome to obtain a p-value, although this may overstate precision since it does not account for estimating $\hat{\Sigma}_j$.
  - Improvement might come from bootstrapping the whole process.
- Clingingsmith et al. (2009) do this slightly differently:
  - Index is a simple average of standardized effects.
  - Standard error estimate and associated p-value account for correlation between effect estimates by using the covariance matrix from a "seemingly unrelated regression" model (cf. Davidson and MacKinnon, 2004, Ch. 12).
  - Could also use bootstrap.

Summary Index Methods

Approaches to analysis:

- Anderson uses $\bar{s}_{ij}$ as the outcome in regressions to estimate an "overall effect" for grouping j, measured in terms of control group standard deviations.
  - For inference, just treats the index as a regular outcome to obtain a p-value, although this may overstate precision since it does not account for estimating $\hat{\Sigma}_j$.
  - Improvement might come from bootstrapping the whole process.
- Clingingsmith et al. (2009) do this slightly differently:
  - Index is a simple average of standardized effects.
  - Standard error estimate and associated p-value account for correlation between effect estimates by using the covariance matrix from a "seemingly unrelated regression" model (cf. Davidson and MacKinnon, 2004, Ch. 12).
  - Could also use bootstrap.
- If you use a PCA factor score, also just estimate effects on the score, but again would want to account for estimation of factor loadings (e.g., with bootstrap).

Summary Index Methods

Table 2. Summary index components

Project	Stage	Summary index components
ABC	Preteen	IQ (5, 6.5, 12), Retained in Grade (12), Special Education (12)
Perry	Preteen	IQ (5, 6, 10), Repeat Grade (17), Special Education (17)
ETP	Preteen	IQ (5, 7, 10), Retained in Grade (17), Special Help (17)
ABC	Teen	IQ (15), HS Grad (18), Teen Parent (19)
Perry	Teen	IQ (14), HS Grad (18), Unemployed (19), Transfers (19), Teen Parent (19), Arrested (19)
ETP	Teen	IQ (17), HS Dropout (18), Worked (18)
ABC	Adult	College (21), Employed (21), Convicted (21), Felon (21), Jailed (21), Marijuana (21)
Perry	Adult	College (27), Employed (27, 40), Income (27, 40), Criminal Record (27), Arrests (27), Drugs (27), Married (27)
ETP	Adult	College (21), Receive Income (21), On Welfare (21)

NOTE: Age of measurement in parentheses. For Perry and Early Training grade repetition and special education variables, it was not possible to isolate pre-9th grade outcomes in the data.

18 groupings defined by program location, timing, and gender.

Summary Index Methods

Table 3. Summary index effects

		Female				Male				Gender
Project	Age	Effect	Naive p value	FWER p value	n	Effect	Naive p value	FWER p value	n	difference t statistic
ABC	Preteen	.445 (.194)	.026	.125	54	.417 (.181)	.026	.184	51	.11
Perry	Preteen	.537 (.177)	.004	.028	51	.150 (.172)	.387	.943	72	1.53
ETP	Preteen	.362 (.251)	.160	.349	30	.148 (.245)	.552	.958	34	.61
ABC	Teen	.422 (.202)	.042	.156	53	.162 (.194)	.407	.943	51	.93
Perry	Teen	.613 (.156)	0	.003	51	.035 (.096)	.716	.977	72	3.32
ETP	Teen	.456 (.299)	.138	.349	29	.123 (.377)	.747	.977	32	.68
ABC	Adult	.452 (.144)	.003	.024	53	.312 (.166)	.066	.372	51	.64
Perry	Adult	.353 (.150)	.022	.125	51	−.012 (.130)	.927	.977	72	1.83
ETP	Adult	−.069 (.186)	.714	.701	29	−.710 (.260)	.011	.090	31	1.98

NOTE: Parentheses contain OLS standard errors. Naive p values are unadjusted p values based on the t distribution. FWER p values adjust for multiple testing at the summary index level and are computed as described in Section 3.2.2. The t statistics test the difference between female and male treatment effects. See Table 2 for the components of each summary index.

FWER Control Methods

- For Anderson, summary indices allowed for separate inference on 18 groupings.
- A general statement of effectiveness required aggregation over the indices.
- We could create a meta-summary index that combined all the groupings, but then the assumption of a common underlying latent factor becomes tenuous.
- So, Anderson turns to FWER control p-value adjustments to ascertain whether the program exhibited effects beyond what we would expect by pure chance.

FWER Control Methods

- Anderson uses a p-value adjustment algorithm from the data-mining literature: "free step-down resampling method" (Westfall & Young, 1993) that controls FWER.
- Basic step-down (Holm):
  - Suppose our 2 tests example where 1 is a true null.
  - Rank the p-values from smaller to larger.
  - Apply sequential tests to these ranked p-values: reject $H_{(1)}$ if $p_{(1)} \leq \alpha/2$ (Bonferroni); if rejected, go to next and reject if $p_{(2)} \leq \alpha$.
  - The true null could rank either first or second.
  - So at worst, we would reject a true null at level α.
  - Thus, FWER $\leq \alpha$.

FWER Control Methods

"Free step-down resampling" accounts for dependence, boosting power:

FWER Control Methods

"Free step-down resampling" accounts for dependence, boosting power:

- Rank M outcomes wrt p-values (M is largest): $y_{p(1)},...,y_{p(M)}$.
- Permute treatment under sharp null and compute "p^*_m" values for each outcome: $p^*_{p(1)},...,p^*_{p(M)}$.
- Enforce monotonicity by constructing $p^{**}_{p(1)},...,p^{**}_{p(M)}$ such that $p^{**}_r = \min\{p^*_r, p^*_{r+1},...,p^*_M\}$.
- Repeat 100,000 times, generating vectors of p^{**}_r values.
- Calculate $p^{fwer*}_r = |\{p^{**}_r : p^{**}_r < p_r\}|/100,000$.
- Enforce monotonicity: $p^{fwer**}_r = min\{p^{fwer*}_r, p^{fwer*}_{r+1},...,p^{fwer*}_M\}$.

	$p(1)$	...	$p(M)$
$\mathbf{Z}$	$\mathbf{Y}_{p(1)}$	...	$\mathbf{Y}_{p(M)}$
$\mathbf{Z}^{*(1)}$	$p^{*(1)}_{p(1)}$	...	$p^{*(1)}_{p(M)}$
	$p^{**(1)}_{p(1)}$	...	$p^{**(1)}_{p(M)}$
$\mathbf{Z}^{*(2)}$	$p^{*(2)}_{p(1)}$	...	$p^{*(2)}_{p(M)}$
	$p^{**(2)}_{p(1)}$	...	$p^{**(2)}_{p(M)}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
	$p^{fwer*}_{p(1)}$	...	$p^{fwer*}_{p(M)}$
	$p^{fwer**}_{p(1)}$	...	$p^{fwer**}_{p(M)}$

FWER Control Methods

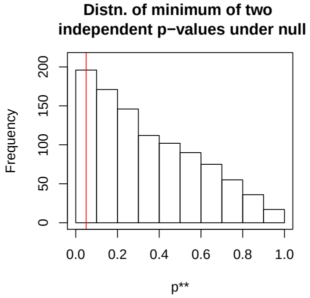
"Free step-down resampling" accounts for dependence, boosting power:

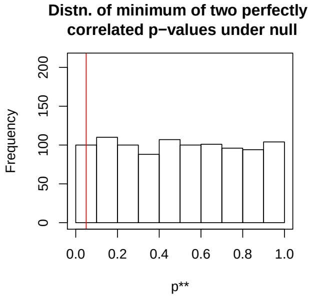
- Rank M outcomes wrt p-values (M is largest): $y_{p(1)}, \ldots, y_{p(M)}$.
- Permute treatment under sharp null and compute "p^*_m" values for each outcome: $p^*_{p(1)}, \ldots, p^*_{p(M)}$.
- Enforce monotonicity by constructing $p^{**}_{p(1)}, \ldots, p^{**}_{p(M)}$ such that $p^{**}_r = \min\{p^*_r, p^*_{r+1}, \ldots, p^*_M\}$.
- Repeat 100,000 times, generating vectors of p^{**}_r values.
- Calculate $p^{fwer*}_r = |\{p^{**}_r : p^{**}_r < p_r\}|/100,000$.
- Enforce monotonicity: $p^{fwer**}_r = min\{p^{fwer*}_r, p^{fwer*}_{r+1}, \ldots, p^{fwer*}_M\}$.

	$p(1)$	$\ldots$	$p(M)$
$\mathbf{Z}$	$\mathbf{Y}_{p(1)}$	$\ldots$	$\mathbf{Y}_{p(M)}$
$\mathbf{Z}^{*(1)}$	$p^{*(1)}_{p(1)}$	$\ldots$	$p^{*(1)}_{p(M)}$
	$p^{**(1)}_{p(1)}$	$\ldots$	$p^{**(1)}_{p(M)}$
$\mathbf{Z}^{*(2)}$	$p^{*(2)}_{p(1)}$	$\ldots$	$p^{*(2)}_{p(M)}$
	$p^{**(2)}_{p(1)}$	$\ldots$	$p^{**(2)}_{p(M)}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
	$p^{fwer*}_{p(1)}$	$\ldots$	$p^{fwer*}_{p(M)}$
	$p^{fwer**}_{p(1)}$	$\ldots$	$p^{fwer**}_{p(M)}$

- Results are FWER adjusted p-values.
- Stata `.ado` on Anderson's website; in R, the `multtest` and `coin` packages and `p.adjust` function have FWER methods.

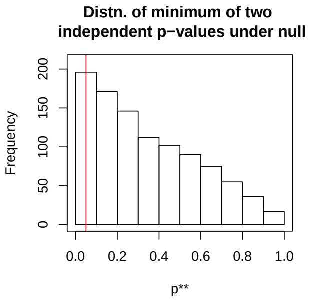
FWER Control Methods

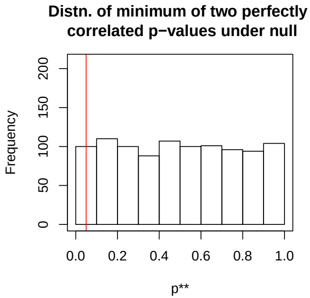




- Permutation preserves outcome dependence and thus *p*-value dependence.
- Distn of min. for independent *p* values is more skewed than distn for positively correlated p-values.

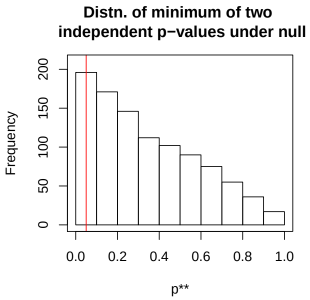
FWER Control Methods

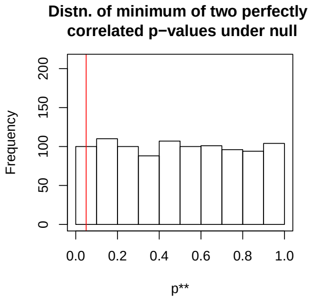




- Permutation preserves outcome dependence and thus *p*-value dependence.
- Distn of min. for independent *p* values is more skewed than distn for positively correlated p-values.
- Bonferroni assumes independent *p* values. Sets adjusted *p* as mass to left of cutpoint in top graph. "Over-corrects" if pos. correlation.

FWER Control Methods





- Permutation preserves outcome dependence and thus p-value dependence.
- Distn of min. for independent p values is more skewed than distn for positively correlated p-values.
- Bonferroni assumes independent p values. Sets adjusted p as mass to left of cutpoint in top graph. "Over-corrects" if pos. correlation.
- Permutation based methods find an adjusted p that has mass to the left of cutpoint on the distribution that accounts for correlation.
- This will be less stringent and thus yields more power.

FWER Control Methods

Table 3. Summary index effects

		Female				Male				
Project	Age	Effect	Naive p value	FWER p value	n	Effect	Naive p value	FWER p value	n	Gender difference t statistic
ABC	Preteen	.445 (.194)	.026	.125	54	.417 (.181)	.026	.184	51	.11
Perry	Preteen	.537 (.177)	.004	.028	51	.150 (.172)	.387	.943	72	1.53
ETP	Preteen	.362 (.251)	.160	.349	30	.148 (.245)	.552	.958	34	.61
ABC	Teen	.422 (.202)	.042	.156	53	.162 (.194)	.407	.943	51	.93
Perry	Teen	.613 (.156)	0	.003	51	.035 (.096)	.716	.977	72	3.32
ETP	Teen	.456 (.299)	.138	.349	29	.123 (.377)	.747	.977	32	.68
ABC	Adult	.452 (.144)	.003	.024	53	.312 (.166)	.066	.372	51	.64
Perry	Adult	.353 (.150)	.022	.125	51	−.012 (.130)	.927	.977	72	1.83
ETP	Adult	−.069 (.186)	.714	.701	29	−.710 (.260)	.011	.090	31	1.98

NOTE: Parentheses contain OLS standard errors. Naive p values are unadjusted p values based on the t distribution. FWER p values adjust for multiple testing at the summary index level and are computed as described in Section 3.2.2. The t statistics test the difference between female and male treatment effects. See Table 2 for the components of each summary index.

FDR Control Methods

- The summary index plus the FWER control p-value adjustment allows one to conclude that the programs tend to be effective in some domains for girls, though not for boys. What might be driving these results?
- To explore this question, Anderson proposes to allow some more leeway in testing, choosing to control FDR rather than FWER in order to “allow” possible effects to reveal themselves.
- This more exploratory analysis is done on the raw outcomes.

FDR Control Methods

- Basic FDR control (Benjamini & Hochberg 1995):

FDR Control Methods

- Basic FDR control (Benjamini & Hochberg 1995):

r	p_r	$.05r/M$	$p_r < .05r/M$?	$q^* = p_r M/r$
1	.01	.0125	yes	.04
2	.02	.025	yes	.04
3	.05	.0375	no	.07
4	.10	.05	no	.10

  - Rank outcomes wrt p-values: $y_{p(1)}, \ldots, y_{p(M)}$.
  - Choose a FDR level, q (analogous to α, e.g., .05).
  - Reject null when $p_r < qr/M$.
  - Ensures FDR is no greater than $q(m_0/M)$, where m_0 is number of true nulls. (Benjamini & Yekutieli, 2001, Thm. 1.2).
    - E.g., if $M = 2$, $m_0 = 1$, then

$$FDR = 1 \cdot \Pr[\text{all rejections false}] + \frac{1}{2}\Pr[\text{half rejections false}]$$

$$\leq 1 \cdot \frac{q}{2}\frac{2q}{2} + \frac{1}{2} \cdot \left[\frac{1}{2}\frac{q}{2}\frac{2q}{2} + \frac{1}{2}\frac{q}{2}\frac{2q}{2}\right] = q\frac{3q}{4}$$

    - B & H 2001 Thm. 1.2 tightens this bound.
  - Obtain "q-values": find min q resulting in rejection under above.

FDR Control Methods

- Basic FDR control (Benjamini & Hochberg 1995):

r	p_r	$.05r/M$	$p_r < .05r/M$?	$q^* = p_r M/r$
1	.01	.0125	yes	.04
2	.02	.025	yes	.04
3	.05	.0375	no	.07
4	.10	.05	no	.10

  - Rank outcomes wrt p-values: $y_{p(1)}, \ldots, y_{p(M)}$.
  - Choose a FDR level, q (analogous to α, e.g., .05).
  - Reject null when $p_r < qr/M$.
  - Ensures FDR is no greater than $q(m_0/M)$, where m_0 is number of true nulls. (Benjamini & Yekutieli, 2001, Thm. 1.2).
    - E.g., if $M = 2$, $m_0 = 1$, then
    $$\begin{aligned} FDR &= 1 \cdot \Pr[\text{all rejections false}] + \frac{1}{2}\Pr[\text{half rejections false}] \\ &\leq 1 \cdot \frac{q}{2}\frac{2q}{2} + \frac{1}{2} \cdot \left[\frac{1}{2}\frac{q}{2}\frac{2q}{2} + \frac{1}{2}\frac{q}{2}\frac{2q}{2}\right] = q\frac{3q}{4} \end{aligned}$$
    - B & H 2001 Thm. 1.2 tightens this bound.
  - Obtain "q-values": find min q resulting in rejection under above.
  - Benjamini et al. (2006) brings FDR control closer to q.
  - Anderson provides Stata `.ado`. In R, see `p.adjust` function.

FDR Control Methods

Table 8. Effects on adult academic outcomes

Outcome	Age	Project	Female Effect	Female CM	Female Naive p value	Female FDR q value	Female n	Male Effect	Male CM	Male Naive p value	Male FDR q value	Male n	Gender difference t statistic
In college	21	ABC	.293 (.116)	.107	.016	.077	53	.148 (.121)	.174	.267	1.000	51	.87
Any college	27	Perry	.160 (.137)	.280	.260	.336	50	−.005 (.110)	.308	.971	1.000	72	.94
In post–high school education	21	ETP	.121 (.191)	.300	.524	.453	29	−.486 (.171)	.636	.004	.082	31	2.37

NOTE: Parentheses contain robust standard errors. CM refers to control mean. Sample size varies within experiments due to attrition for some variables. The p and q values are computed as described in Section 3; t statistics test the difference between female and male treatment effects.

FDR Control Methods

Table 9. Effects on adult economic outcomes

			Female					Male					Gender difference
Outcome	Age	Project	Effect	CM	Naive p value	FDR q value	n	Effect	CM	Naive p value	FDR q value	n	t statistics
Employed	21	ABC	.104 (.137)	.536	.427	.405	53	.188 (.142)	.455	.199	1.000	50	−.43
Employed	27	Perry	.255 (.136)	.545	.078	.216	47	.036 (.121)	.564	.773	1.000	69	1.20
Annual income	27	Perry	2,567 (2,686)	8,986	.347	.390	47	2,363 (2,708)	12,495	.391	1.000	66	.05
Monthly income	27	Perry	396 (236)	651	.101	.245	47	537 (247)	830	.026	.388	68	−.41
Employed	40	Perry	.015 (.115)	.818	.931	.574	46	.200 (.120)	.500	.112	.741	66	−1.12
Annual income	40	Perry	3,492 (5,491)	17,374	.538	.453	46	6,228 (5,958)	21,119	.299	1.000	66	−.34
Monthly income	40	Perry	162 (431)	1,615	.704	.505	46	436 (562)	1,839	.459	1.000	66	−.39
Receive income	21	ETP	−.074 (.200)	.600	.697	.505	29	−.159 (.134)	.909	.304	1.000	31	.36
Receive welfare	21	ETP	−.042 (.157)	.200	.826	.537	30	NA					

NOTE: Parentheses contain robust standard errors. CM refers to control mean. Sample size varies within experiments due to attrition for some variables. The p and q values are computed as described in Section 3; t statistics test the difference between female and male treatment effects. Males are ineligible for welfare.

Multiple outcomes: from vice to virtue

Index measures can give you more power by borrowing strength:

Multiple outcomes: from vice to virtue

Index measures can give you more power by borrowing strength:

- Suppose N random draws of Y_1 and Y_2, each with variance σ^2 and correlation between them ρ.
- Variance of the sample mean for either one would be σ^2/N.
- Now consider their average: $\tilde{Y}_i = (Y_{i1} + Y_{i2})/2$.

Multiple outcomes: from vice to virtue

Index measures can give you more power by borrowing strength:

- Suppose N random draws of Y_1 and Y_2, each with variance σ^2 and correlation between them ρ.
- Variance of the sample mean for either one would be σ^2/N.
- Now consider their average: $\tilde{Y}_i = (Y_{i1} + Y_{i2})/2$.
- Sample mean of the average is,

$$\bar{\tilde{Y}} = \frac{1}{N}\sum_{i=1}^{N}\tilde{Y}_i = \frac{1}{N}\sum_{i=1}^{N}\frac{Y_{1i}+Y_{2i}}{2} = \frac{1}{2N}\sum_{i=1}^{N}(Y_{1i}+Y_{2i}).$$

- Sampling variance is

$$\mathrm{Var}\left[\bar{\tilde{Y}}\right] = \frac{1}{4N^2}\sum_{i=1}^{N}(2\sigma^2 + 2\sigma^2\rho) = \frac{\sigma^2}{N}\frac{1+\rho}{2}$$

- Note that $1/2 < \frac{1+\rho}{2} < 1$, meaning a reduction in variance.

Multiple outcomes: from vice to virtue

Index measures can give you more power by borrowing strength:

- Suppose N random draws of Y_1 and Y_2, each with variance σ^2 and correlation between them ρ.
- Variance of the sample mean for either one would be σ^2/N.
- Now consider their average: $\tilde{Y}_i = (Y_{i1} + Y_{i2})/2$.
- Sample mean of the average is,

$$\bar{\tilde{Y}} = \frac{1}{N}\sum_{i=1}^{N}\tilde{Y}_i = \frac{1}{N}\sum_{i=1}^{N}\frac{Y_{1i}+Y_{2i}}{2} = \frac{1}{2N}\sum_{i=1}^{N}(Y_{1i}+Y_{2i}).$$

- Sampling variance is

$$\text{Var}[\bar{\tilde{Y}}] = \frac{1}{4N^2}\sum_{i=1}^{N}(2\sigma^2 + 2\sigma^2\rho) = \frac{\sigma^2}{N}\frac{1+\rho}{2}$$

- Note that $1/2 < \frac{1+\rho}{2} < 1$, meaning a reduction in variance.
- Implies composite measures can yield more power *if* one uses an aggregation procedure that doesn't introduce too much noise.

Multiple outcomes: from vice to virtue

Composite measures allow one to test richer theoretical implications, often with more power:

- Caughy et al. (2015) consider theories that give rise to a set of hypotheses and associated statistical tests.
- Cf. “pattern matching.”
- Intuition: refine one’s inference to account for when results are “generally consistent” with a large number of predictions, even if some are not statistically significant.
- Method: combine the various tests to get a “global p-value” for the set of propositions.
- See also Young (2015a) “Channeling Fisher” paper for “global tests” of experimental effects.

Multiple outcomes: from vice to virtue

Athey et al. (2016) study the "surrogate outcomes" problem:

Multiple outcomes: from vice to virtue

Athey et al. (2016) study the “surrogate outcomes” problem:

- Sometimes we want to study a long term outcome—e.g., outbreak of war—but our period of analysis is limited, and so we have few cases (wars) to learn from.
- What would be nice would be to find a “surrogate”— that is, something that is highly predictive of war, but for which we get more variation in the short term.

Multiple outcomes: from vice to virtue

Athey et al. (2016) study the "surrogate outcomes" problem:

- Sometimes we want to study a long term outcome—e.g., outbreak of war—but our period of analysis is limited, and so we have few cases (wars) to learn from.
- What would be nice would be to find a "surrogate"— that is, something that is highly predictive of war, but for which we get more variation in the short term.
- Athey et al. propose the following strategy:
  - Find an auxiliary dataset that covers a longer period and has lots of wars and then also has *tons* of variables that might be predictive of wars.
  - Find variables that are indeed predictive and that also vary in the short term. Use them to construct a "surrogate index"—that is, the expected value of the long term outcome conditional on the short term measures. Machine learning would be useful here.
  - Use these short term measures in your shorter term study, combining them into the surrogate index.
- Validity depends on criteria similar to Pearl's "front door."

Remarks

- Deep consideration of multiple endpoints is pretty new for social scientists.
- Multiple endpoints and multiple comparisons is especially important for political science research: we have to work with complex or vague concepts that require multiple measures.
- The methods described here show you how to *be careful* with multiple measures.
- They also suggest how to *make good use* of multiple measures.