Beyond the basics

Mauricio Romero

Beyond the basics

Randomization inference

Extensive vs. Intensive Margin Effects

Factorial Designs

Beyond the basics

Randomization inference

Extensive vs. Intensive Margin Effects

Factorial Designs

The Lady Tasting Tea

Chapter II of Fisher's The Design of Experiments begins: A lady declares that by tasting a cup of tea made with milk she can discriminate whether the milk or the tea infusion was first added to the cup.

- The lady was biologist Muriel Bristol, who worked with Fisher at the Rothamsted Experimental Station in Harpenden, UK
- H0: Fisher believes that Dr. Bristol cannot taste the difference
- A test of the hypothesis: Experiment consists in mixing eight cups of tea, four in one way and four in the other, and presenting them to the subject for judgment

Rule 1: do not confound the treatment

- Critical assumption: if Dr. Bristol is unable to detect whether the milk was poured in first, then she will choose 4 cups at random
  - Fisher points out that the experimenter could screw this up: If all those cups made with the milk first had sugar added, while those made with the tea first had none, this might well ensure that all those made with sugar should be classed alike
  - Gerber and Green refer to this as excludability (or the treatment no correlated with the potential outcomes)

Rule 2: do not accidentally confound your own treatment

Fisher:It is not sufficient remedy to insist that 'all the cups must be exactly alike' in every respect except that to be tested. For this is a totally impossible requirement

- To minimize the likelihood of accidentally confounding your treatment, the best approach is to constrain yourself by randomizing
- Randomization minimizes the likelihood of unfortunate coincidences
- Highly controversial position at the time, and it is still debated in some circles
- The alternative is to force balance on observables, and hope that unobservables don't matter

The Lady Tasting Tea: a Hypothesis Test

How should we interpret data from this experiment?

- Suppose Dr. Bristol correctly identified all 4 "treated" cups
  - How likely is it that this outcome could have occurred by chance?
  - There are $\binom{8}{4} = \frac{8!}{4!4!} = 70$ possible ways to choose 4 cups
  - Only one is correct; a subject with no ability to discriminate between treated and untreated cups would have a $1/70$ chance of success
  - The p-value associated with this outcome is $1/70 \approx 0.014$

The Lady Tasting Tea: a Hypothesis Test

How should we interpret data from this experiment?

- Suppose Dr. Bristol correctly identified 3 "treated" cups
  - How likely is it that this outcome could have occurred by chance?
  - There are $\binom{4}{3} \times \binom{4}{1} = 16$ possible ways to choose 3 correct cups
  - The p-value associated with this outcome is $16/70 \approx 0.22$

The Lady Tasting Tea: a Hypothesis Test

- The only experimental result that would lead to the rejection of the null hypothesis was correct identification of all 4 treated cups

- In the actual experiment, the null hypothesis was rejected

The Lady Tasting Tea: a Hypothesis Test

- The size of a test is the likelihood of rejecting a true null
- Fisher asserts that tests of size 0.05 are typical
- Alternative experiment: what if we had treated 3 out of 6 cups of tea?
  - There are $\binom{3}{6} = 20$ possible ways to choose 3 of 6 cups
  - Best possible p-value is therefore $\frac{1}{2} = 0.05$ (selecting the one correct combination)
- Alternative experiment: what if we had treated 3 out of 8 cups of tea?
  - There are $\binom{3}{8} = 56$
  - Best possible p-value is therefore $\frac{1}{56} = 0.017$
- Optimal to have equal numbers of treated, untreated cups (see variance formula from last lecture)

The Lady Tasting Tea: a Hypothesis Test

- An alternate experiment: an unknown number of treated cups
  - Combinations of 8 cups are 2^8
  - Under the null, the probability of getting 8 right is 1 in 2^8
  - Probability of getting 7 right is $\frac{8}{256} = 0.03125$
  - Achieve higher power with the same number of trials
  - Possible to reject the hypothesis that the lady tasting tea cannot tell the difference even when her ability to discriminate is imperfect

Randomization inference (RI)

- Elegant precursor to OLS approach to experiments by Neyman (and later by Fisher) in the 1920s
- Randomization creates a **sharp** null hypothesis
- By repeating the exact routine by which the original randomization was conducted many times (e.g., 5,000), we can make inference about variances
  - For each placebo treatment assignments calculate the treatment-control difference
  - The distribution gives the variation in the treatment/control difference given the way the experiment was conducted
  - Place observed treatment/control differential into this distribution
  - If observed difference lies in bottom or top 2.5% of the distribution, reject the two-sided null at 5% level
- Purest form of RI: Calculate every permutation of the treatment assignment
  - Confidence statements become a statement of the relative probability of different counterfactual outcomes occurring

Advantages of RI

- Allows for calculation of tight error bounds without resorting to asymptotics: **Works well in small samples**
- Makes no parametric assumptions about error distributions
- Effective for conducting inference when complex randomization routines were used
- Is it possible that an error or over-restriction in your randomization caused imbalance?
  - Use RI on the randomization that your code could have delivered to understand this
  - Are you imbalanced relative to this universe?
  - Is this universe itself unbalanced?

Some special cases of RI

- How to use RI to test for a treatment effect other than the null of no effect?

Some special cases of RI

- How to use RI to test for a treatment effect other than the null of no effect?
  - If the null is that the treatment-control differential is τ_0
  - Add τ_0 on to every treatment-control differential
  - Test against this RI distribution

Some special cases of RI

- How to use RI to test for a treatment effect other than the null of no effect?
  - If the null is that the treatment-control differential is τ_0
  - Add τ_0 on to every treatment-control differential
  - Test against this RI distribution
- How to conduct the equivalent of a LATE estimation using RI?

Some special cases of RI

- How to use RI to test for a treatment effect other than the null of no effect?
  - If the null is that the treatment-control differential is τ_0
  - Add τ_0 on to every treatment-control differential
  - Test against this RI distribution
- How to conduct the equivalent of a LATE estimation using RI?
  - Maintain the actual cluster-level compliance rate in the re-randomization procedure
  - Inflate each treatment-control comparison by the inverse of the observed compliance differential between the two cells
  - Conduct RI on this distribution

Some special cases of RI

- How to use RI to test for a treatment effect other than the null of no effect?
  - If the null is that the treatment-control differential is τ_0
  - Add τ_0 on to every treatment-control differential
  - Test against this RI distribution
- How to conduct the equivalent of a LATE estimation using RI?
  - Maintain the actual cluster-level compliance rate in the re-randomization procedure
  - Inflate each treatment-control comparison by the inverse of the observed compliance differential between the two cells
  - Conduct RI on this distribution
- How to conduct covariate adjustment with RI?

Some special cases of RI

- How to use RI to test for a treatment effect other than the null of no effect?
  - If the null is that the treatment-control differential is τ_0
  - Add τ_0 on to every treatment-control differential
  - Test against this RI distribution
- How to conduct the equivalent of a LATE estimation using RI?
  - Maintain the actual cluster-level compliance rate in the re-randomization procedure
  - Inflate each treatment-control comparison by the inverse of the observed compliance differential between the two cells
  - Conduct RI on this distribution
- How to conduct covariate adjustment with RI?
  - Use regression or other techniques to adjust for covariates
  - Then conduct RI using the residuals from these regressions

Beyond the basics

Randomization inference

Extensive vs. Intensive Margin Effects

Factorial Designs

Beyond the basics

Randomization inference

Extensive vs. Intensive Margin Effects

Factorial Designs

Extensive vs. Intensive Margin Effects

- Problem akin to Heckman selection: Treatment effects may have 2 components

  1. Extensive margin: A component that works on the likelihood that the outcome is non-zero (or is observed)

  2. Intensive margin: A component that works on the outcome conditional on it being observed and non-zero

- Experiments provide evidence on the reduced-form combination of these effects

- To disentangle, need separate experimental variation to provide identification on the extensive margin (instrument for the selection equation in Heckman terms)

Extensive vs. Intensive Margin Effects: Some examples

1. Attrition may be differential in absolute levels
2. Who attrites (differential selection) across the treatment and control.
3. A treatment may effect both the likelihood that individuals are sexually active as well as the attributes of those with whom they are sexually active
   - Cannot examine effect on partner selection in isolation of being active at all
   - Same problem as attrition statistically, but extensive margin effects are an intrinsic and interesting part of the treatment effect
4. Interest rates can be randomized across different clients
   - Effect on who complies with the treatment **(adverse selection)**
   - Effect how they perform under the treatment **(moral hazard)**
5. Treatment may influence the likelihood that someone enters self-employment, as well as the returns to self-employment if they do
   - "Self-employment income" is a censored variable that confounds the intensive and extensive margin
   - What types of people start new businesses, and how large are the new businesses?

Beyond the basics

Treatment affects probability of observing outcome

$$
\begin{aligned}
Y_i^* &= T_i\beta + X_i\pi_1 + u_i \\
Z_i^* &= T_i\gamma + X_i\pi_2 + \varepsilon_i \\
Y_i &= Y_i^* \mathbb{1}_{Z_i^*>0}
\end{aligned}
$$

- Both Y_i^* and Z_i^* are unobserved
- Even if T is randomized you end up back in a causal inference problem very closely related to the original Heckman formulation of selection bias
- If treatment does not affect observability of the dependent variable you have no problem. You have a correct (L)ATE on those who would anyway have had $Z_i^* > 0$
- If you have an instrument for Z_i^*, you use a Heckman-type estimator

Beyond the basics

Attrition Propensity Weights

- If $Z \perp Y$ then attrition increases standard errors, but will not introduce bias
- But $Z \perp Y$ is unlikely, instead sometimes you can argue $Z \perp Y|X$
- In this case you can use the coefficients from

$$Z_i = X_i\gamma + \varepsilon_i$$

- To predict the propensity not to attrite for all units: $\widehat{Z}_i$
- Run original regression with weights $\frac{1}{\widehat{Z}_i}$
- If likelihood of attrition, conditional on X, is zero, then $\widehat{Z}_i = 1$
- If likelihood of attrition, conditional on X, is zero, then $\widehat{Z}_i = 0.5$
- Attrition analogy to selection on observables in matching estimators

Horowitz and Manski (2000)

- Horowitz and Manski (2000) suggest imputing the missing values with the opposite extremes of the potential values of the dependent variable.

- Can work relatively well if Y is binary, but not otherwise

Zhang, Rubin, and Mealli (2007): "Principal Stratification"

- Problem: Don't observe outcomes unless the unit undertakes some action
- Classify participants into four groups according to whether they would be observed or not in the treatment and in the control
- Divide individual-level counterfactual by decision under (T,C) status
  - (N,N) are never observed (we never observe the variable for them)
  - (O,N) are observed only in the control: assume this group doesn't exist for now
  - (N,O) are observed only if treated (exist in our sample, source of selection problem)
  - (O,O) Observed whether or not treated

Zhang, Rubin, and Mealli (2007): "Principal Stratification"

- The intensive margin effect is only defined for those who would have been observed in the treatment and in the control (O, O)
  - Given that you were going to undertake observable behavior no matter what, how did the treatment change your behavior?
- Difference in means between treated and control confounds effect on (N, O) and (O, O)
- Marginal effect of treatment on (N,N) is undefined
- Marginal effect of treatment on (N,O) is potentially meaningful, but due to selection hard to establish without assumptions (e.g., effect the same as (O, O))

Lee (2009) Bounding

- Lee (2009): Baseline variables related to employment probability could also determine wage rates
- Deals with unequal attrition between treatment and control
- Bounding technique identifies the logical worst case scenarios for how much the difference in the probability of observing the dependent variable can possibly affect the mean of the dependent variable conditional on it being observed

Lee (2009) Bounding

- Rank the outcomes for the control in order of the outcome itself
- Outcome is observed for $p\%$ of the treatment and $q\%$ of the controls
- Assume $p\% > q\%$ (symmetric in the other case). Differential attrition is $(q-p)\%$
- Two estimates by trimming the tails of the distribution of observed outcomes in the control:
  - $\max(E(Y|T=0)$ is the average outcome in the control eliminating the $(q-p)\%$ of lowest outcomes
  - $\min(E(Y|T=0)$ is the average outcome in the control eliminating the $(q-p)\%$ of highest outcomes
- The upper and lower Lee bounds are then given by
  - $E(Y|T=1) - \max(E(Y|T=0)$ (lower bound)
  - $E(Y|T=1) - \min(E(Y|T=0)$ (upper bound)

Lee (2009) Bounding

- Assume you have a treatment that affects sexual behavior
  - 50% of control units are in relationships and 40% of treatment units are
  - Average age in all treatment relationships is 20, and in all control relationships is 18
  - Some of that difference may come from a treatment effect, and some from selection
  - 20% of controls would have selected out of sexual activity if treated
  - How big could the intensive treatment effect on the regularly sexually active be?
  - Take 80% of controls with the oldest partners and 80% with the youngest partners
  - Use these two groups as counterfactuals
  - The marginal effect of the program on the sexualy active group must lie in between these two values (these are the most extreme differences that selection could explain)

Observing Unobservables

- Karlan and Zinman 'Observing Unobservables'
- Degree to which increasing interest rates on loans generates adverse (or advantageous) selection and moral hazard
- Randomize the interest rate offered to potential borrowers
- Differences in the subsequent performance of borrowers depending on this interest rate offer then gives the AS effect
- Once a borrower has come in to take the loan, they then randomize downward from the offer rate the actual rate charged on the loan
- Conditional on the offer rate, this variation in the actual rate isolates MH

Observing Unobservables

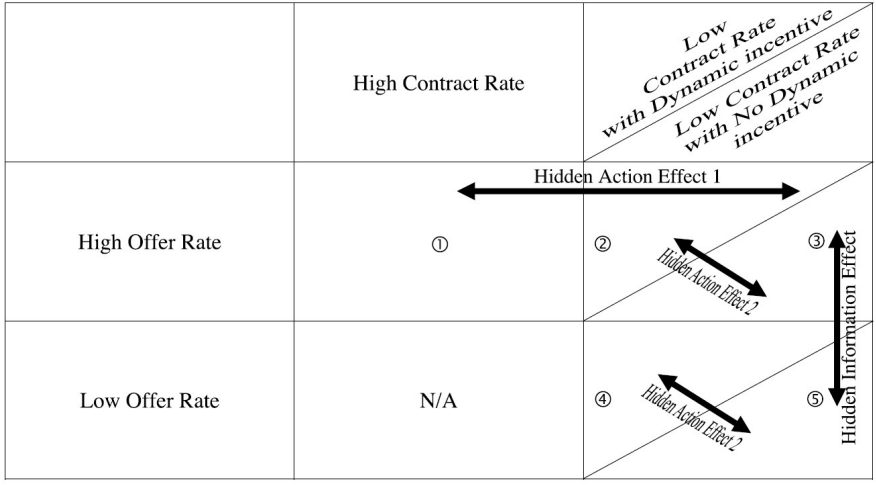


FIGURE 1.—Some basic intuition for our identification strategy.

Beyond the basics

Randomization inference

Extensive vs. Intensive Margin Effects

Factorial Designs

Beyond the basics

Randomization inference

Extensive vs. Intensive Margin Effects

Factorial Designs

Factorial (or cross-cutting) experimental designs

- Widely used in field (and lab) experiments
- Rationale:
  - Budget constraints
  - Interaction effects are often considered a second order issue

Factorial (or cross-cutting) experimental designs

- Widely used in field (and lab) experiments
- Rationale:
  - Budget constraints
  - Interaction effects are often considered a second order issue

Conducting a series of evaluations in the same area allows substantial cost savings. Since data collection is the most costly element of these evaluations, cross-cutting the sample reduces costs dramatically. (Kremer AER P&P, 2003)

Factorial (or cross-cutting) experimental designs

- Widely used in field (and lab) experiments
- Rationale:
  - Budget constraints
  - Interaction effects are often considered a second order issue

Conducting a series of evaluations in the same area allows substantial cost savings. Since data collection is the most costly element of these evaluations, cross-cutting the sample reduces costs dramatically. (Kremer AER P&P, 2003) ...This tactic can be problematic, however, if there are significant interactions between programs. (ibid)

Typical 2×2 design

		T_1	
		No	*Yes*
T_2	*No*	N_1	N_2
	Yes	N_3	N_4

Long model: $$Y = \beta_0 + \beta_1 T_1 + \beta_2 T_2 + \beta_{12} T_1 \times T_2 + \varepsilon \quad (1)$$

Short model: $$Y = \beta_0^s + \beta_1^s T_1 + \beta_2^s T_2 + \varepsilon^s \quad (2)$$

Typical 2×2 design

		T_1	
		No	*Yes*
T_2	*No*	N_1	N_2
	Yes	N_3	N_4

$$\text{Long model:} \quad Y = \beta_0 + \beta_1 T_1 + \beta_2 T_2 + \beta_{12} T_1 \times T_2 + \varepsilon \tag{1}$$

$$\text{Short model:} \quad Y = \beta_0^s + \beta_1^s T_1 + \beta_2^s T_2 + \varepsilon^s \tag{2}$$

Three approaches:

1. Estimate and report (1)

Typical 2×2 design

		T_1	
		No	*Yes*
T_2	*No*	N_1	N_2
	Yes	N_3	N_4

$$\text{Long model:} \quad Y = \beta_0 + \beta_1 T_1 + \beta_2 T_2 + \beta_{12} T_1 \times T_2 + \varepsilon \tag{1}$$

$$\text{Short model:} \quad Y = \beta_0^s + \beta_1^s T_1 + \beta_2^s T_2 + \varepsilon^s \tag{2}$$

Three approaches:

1. Estimate and report (1)
2. Estimate and report (2)

Typical 2×2 design

		T_1	
		No	*Yes*
T_2	*No*	N_1	N_2
	Yes	N_3	N_4

$$\text{Long model:} \quad Y = \beta_0 + \beta_1 T_1 + \beta_2 T_2 + \beta_{12} T_1 \times T_2 + \varepsilon \tag{1}$$

$$\text{Short model:} \quad Y = \beta_0^s + \beta_1^s T_1 + \beta_2^s T_2 + \varepsilon^s \tag{2}$$

Three approaches:

1. Estimate and report (1)
2. Estimate and report (2)
3. Estimate (1); test for $\beta_{12} = 0$ in (1) and estimate and focus on (2) if $\beta_{12} = 0$ is not rejected (model selection)

Regression: long and short

$$
\begin{aligned}
Y &= \beta_0 + \beta_1 T_1 + \beta_2 T_2 + \beta_{12} T_1 \times T_2 + \varepsilon, \\
Y &= \beta_0^s + \beta_1^s T_1 + \beta_2^s T_2 + \varepsilon^s
\end{aligned}
$$

Long model (OLS):

$$
\begin{aligned}
\widehat{\beta_1} &\xrightarrow{p} \beta_1 \quad \text{(main effect of } T_1\text{)} \\
\widehat{\beta_2} &\xrightarrow{p} \beta_2 \quad \text{(main effect of } T_2\text{)} \\
\widehat{\beta_{12}} &\xrightarrow{p} \beta_{12} \quad \text{(interaction effect)}
\end{aligned}
$$

Regression: long and short

$$
\begin{aligned}
Y &= \beta_0 + \beta_1 T_1 + \beta_2 T_2 + \beta_{12} T_1 \times T_2 + \varepsilon, \\
Y &= \beta_0^s + \beta_1^s T_1 + \beta_2^s T_2 + \varepsilon^s
\end{aligned}
$$

Long model (OLS):

$$
\begin{aligned}
\widehat{\beta_1} &\xrightarrow{p} \beta_1 \quad \text{(main effect of } T_1\text{)} \\
\widehat{\beta_2} &\xrightarrow{p} \beta_2 \quad \text{(main effect of } T_2\text{)} \\
\widehat{\beta_{12}} &\xrightarrow{p} \beta_{12} \quad \text{(interaction effect)}
\end{aligned}
$$

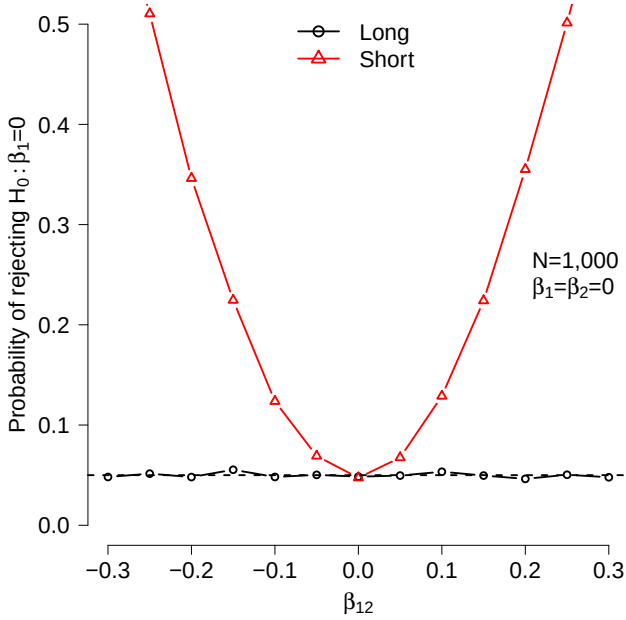
Short model (OLS):

$$
\begin{aligned}
\widehat{\beta_1^s} &\xrightarrow{p} \beta_1^s = \beta_1 + \beta_{12} P(T_2 = 1) \\
\widehat{\beta_2^s} &\xrightarrow{p} \beta_2^s = \beta_2 + \beta_{12} P(T_1 = 1)
\end{aligned}
$$

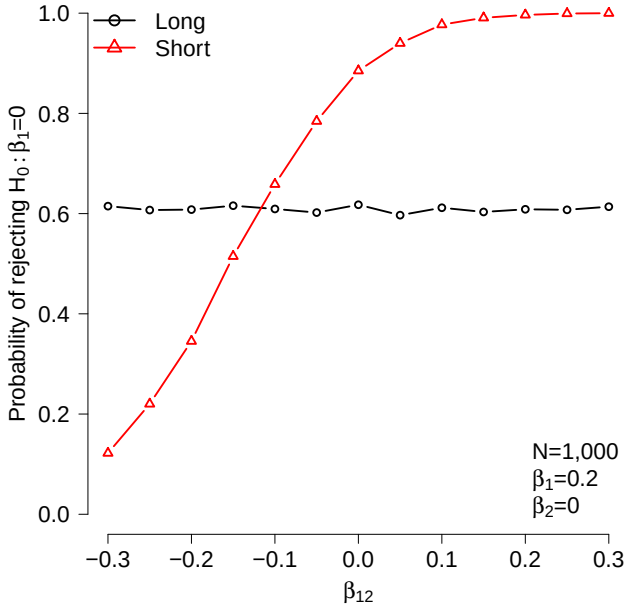
Running example based on a prototypical setting

- 2×2 design with $N_1 = N_2 = N_3 = N_4 = 250$ ($N = 1,000$)
- DGP:
$$Y_i = \beta_1 T_{1i} + \beta_2 T_{2i} + \beta_{12} T_{1i} T_{2i} + \varepsilon_i, \quad \varepsilon_i \sim N(0,1),$$
- T_{1i} and T_{2i} are randomly assigned: $P(T_{1i} = 1) = P(T_{2i} = 1) = 0.5$
- Short model: 90% power to detect an effect of 0.2σ at the 5% level
- Long model: 90% power to detect an effect of 0.29σ at the 5% level
- Monte Carlo simulations to assess the rejection rates of different procedures under the null (size) and the alternative (power)

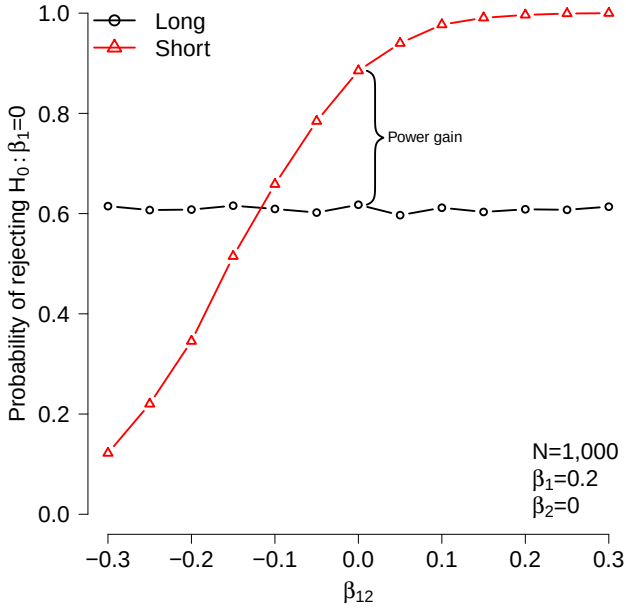
Trade-off: Size distortion if $\beta_{12} \neq 0$



Trade-off: Power gains if $\beta_{12} = 0$



Trade-off: Power gains if $\beta_{12} = 0$



Three problems with the short model

- Affects internal validity of experimental estimates
  - Policy discussions assume a "business as usual" counterfactual
  - Papers written as if estimating main effects (β_1 as opposed to β_1^s)
  - Often, $\beta_1^s = \beta_1 + \beta_{12} P(T_2 = 1)$ is neither of academic nor policy relevance

Three problems with the short model

- Affects internal validity of experimental estimates
  - Policy discussions assume a "business as usual" counterfactual
  - Papers written as if estimating main effects (β_1 as opposed to β_1^s)
  - Often, $\beta_1^s = \beta_1 + \beta_{12} P(T_2 = 1)$ is neither of academic nor policy relevance
  - Problem can be re-cast as "external validity", but not very satisfying

Three problems with the short model

- Affects internal validity of experimental estimates
  - Policy discussions assume a "business as usual" counterfactual
  - Papers written as if estimating main effects (β_1 as opposed to β_1^s)
  - Often, $\beta_1^s = \beta_1 + \beta_{12} P(T_2 = 1)$ is neither of academic nor policy relevance
  - Problem can be re-cast as "external validity", but not very satisfying

Model selection

- In practice, researchers often employ a two-step procedure

  1. Pre-test: estimate the long model and test $H_0 : \beta_{12} = 0$
  2. Post-selection inference:

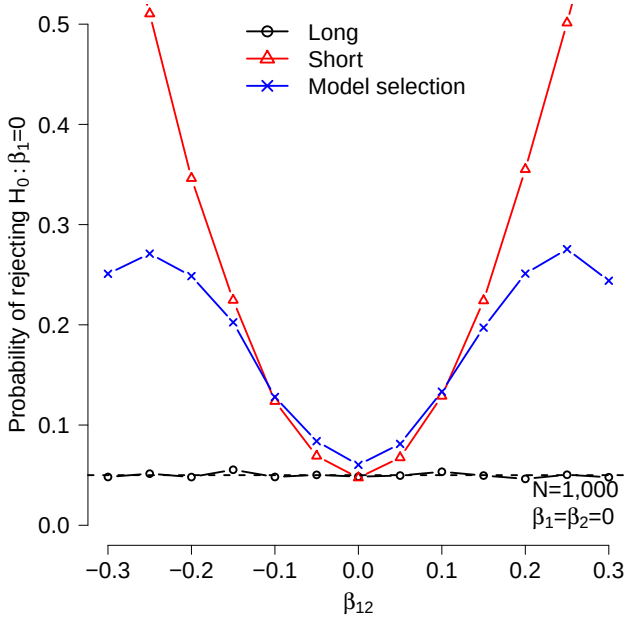
     a. If reject $H_0 : \beta_{12} = 0$, use t-test based on the long model

     b. If cannot reject $H_0 : \beta_{12} = 0$, estimate short model and use t-test based on short model

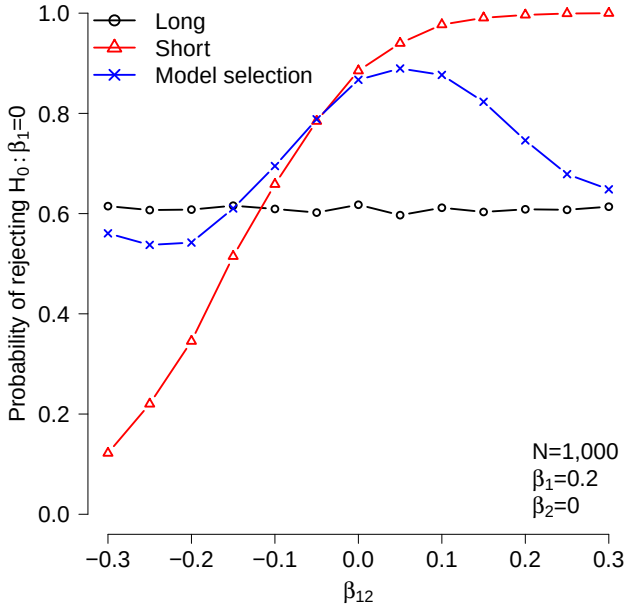
Model selection

- In practice, researchers often employ a two-step procedure
  1. Pre-test: estimate the long model and test $H_0 : \beta_{12} = 0$
  2. Post-selection inference:
     a. If reject $H_0 : \beta_{12} = 0$, use t-test based on the long model
     b. If cannot reject $H_0 : \beta_{12} = 0$, estimate short model and use t-test based on short model
- Does not work
  - True in general (Leeb & Pötscher in several papers)
  - Particularly problematic since most experiments are not powered to detect interactions

Size distortions if $\beta_{12} \neq 0$



Power


Long
Short
Model selection
Probability of rejecting $H_0 : \beta_1=0$
1.0
0.8
0.6
0.4
0.2
0.0
−0.3
−0.2
−0.1
0.0
0.1
0.2
0.3
β_{12}
N=1,000
β_1=0.2
β_2=0

There is no free lunch

- Perceived power gains from the short model come at the cost of false positives in inference on β_1

There is no free lunch

- Perceived power gains from the short model come at the cost of false positives in inference on β_1

- Or come at the cost of a more complicated interpretation of the estimand: $\beta_1^s = \beta_1 + \beta_{12} P(T_2 = 1)$

Recommendations

- Transparency (design and reporting)
  - Always include a figure with the full design
  - (In most cases) report the long model
- Model selection does not work
  - Rationale for the short model should never be that β_{12} is insignificant
- In some cases, the short model and β_1^s may be fine
  - (i) But should be committed to in a pre-analysis plan (with rationale)
  - (ii) Be explicit about interpretation and allow readers to assess whether the other treatments can be interpreted as "background" variables
- For the design of new policy experiments we recommend leaving the interaction cells empty when
  - (i) The "business as usual" counterfactual is especially important
  - (ii) The interactions are not of primary interest