w241: Experiments and Causality

Heterogeneous Treatment Effects

David Reiley, David Broockman, D. Alex Hughes
UC Berkeley, School of Information
Updated: 2021-07-21

Introduction to Heterogeneous Treatment Effects

Main Topics

Heterogeneous treatment effects

- Does the same same treatment have different effects on different subjects?
- Use of regression to measure HTEs: Interaction between covariates of interest and treatment variable

Data analysis with non-experimental control

- Multiple comparisons problem: Fishing expeditions can cause overstatement of true statistical significance.

Reading: *Field Experiments*

Please read

- Section 9.0 (the Introduction), and,
- Section 9.1

Motivating Examples of Heterogeneous Treatment Effects

Quiz Question Discussion

- Potential outcomes are hypothetical population parameters.
- In real-world samples, we get to measure only
  - Treatment outcomes for the treatment group
  - Control outcomes for the control group
- We can measure only treatment outcomes or control outcomes for a single person.

Reading: *Field Experiments*, Section 9.3.1

While reading, consider, "Do different groups have different treatment effects?"

Example: Electricity Consumption

Test Groups

1. Below-average; and,
2. Above-average electricity consumers

Treatment

- Social comparison
- How much electricity are your neighbors using?

HTE

- Are people who use a *lot* of energy more responsive than people who aren't using much energy?
- Are people who are using less than their neighbors going to start using more?

Example: Congressional Responsiveness

Test groups

1. Members of Congress from the North; and,
2. Members of Contress from the South

Treatment

- Being informed of a meeting with a donor versus a constituent

HTE

- Responsiveness of different congressmen

Example: eBay Shipping Policies

Test groups

- Buyers of high-priced items
- Buyers of low-priced items

Treatment

- Charging (or not charging) a shipping price on an auction Some measure of value should be predetermined before the experiment

HTE

- Responsiveness to shipping costs on high-value versus low-value items

Example: eBay Seller Reputation

Test Groups

- Buyers of high-priced items
- Buyers of low-priced items

Treatment

- Selling from a high- or low-reputation account

HTE

- Are people differently reactive to reputation when they are buying high-priced, rather than low-priced items?

Example: Donation Matching

Test Groups

- People who live in *Blue States*
- Peolpe who live in *Red States*

Treatment

- Informing individuals that a donation they make to the ACLU will be matched

HTE

- Because the ACLU is perceived to be a "Liberal" non-profit organization, do individuals who live in areas that are more liberal react more strongly to the donor-matching treatment?

Treatment-by-Covariate Interactions

Quiz Review

Answer

- Students whose parents' literacy was above the median

Discussion

- Treatment effect not statistically significantly different between students whose parents have above-median versus below-median literacy
- We need to know standard errors in order to evaluate how much to believe a point estimate

Estimating HTEs

Etimating with two samples

- Split data into two separate samples
- Compute means and standard errors of the treatment effect in each group
  - Do a two-sample test of means.
  - **Group one**: Students whose parents have above-median literacy
  - **Group two**: Students whose parents have below-median literacy

Estimating HTEs (cont'd)

Estimating with regression

- Estimate a regression with dummy variables
  - I indicator for teacher incentive treatment
  - P indicator for parents with above-average literacy
- Include an interaction between $I \times P$

$$Y_i = \beta_0 + \beta_1 I_i + \beta_2 P_i + \beta_3 (I_i \times P_i) + \epsilon_i$$

Inferences

- Test whether there is evidence that the interaction term is different from zero
  - H_0: The treatment effects are no different between the groups
  - H_A: The treatment effects are different between the groups

Reading Clarification

- Always present standard errors with point estimates.
- Always show the number of observations.
- Reports the coefficient on one row, and the standard error in parentheses on the row below
- Includes a reporting of:
  - The number of observations
  - The R^2 and *F-Test* vs. an intercept-only model

Standard Format of Results

```
                    ##
## ========================================
##                   Dependent variable:
##               ---------------------------
##                           Score
## ----------------------------------------
## Treatment               1.01***
##                         (0.03)
##
## Constant                0.81***
##                         (0.19)
##
## ----------------------------------------
## Observations              100
## R2                       0.90
## Adjusted R2              0.89
## F Statistic    839.71*** (df = 1; 98)
## ========================================
## Note:        *p<0.1; **p<0.05; ***p<0.01
```

Reading Clarification (cont'd)

$$\widehat{Y}_i = \beta_0 + \beta_1 I_i + \beta_2 P_i + \beta_3 (I_i \times P_i)$$

- If a test for β_3 rejects the null hypothesis, then treatment effects differ along with levels of the covariate
- Possible to estimate this with regression or randomization inference
  - Gerber and Green use randomization inference,
  - With data that is well-behaved, we can rely on the Central Limit Theorem and use regression

Teacher Incentives: Evidence from India

Example: Teacher Incentives

Muralidharan and Sundararaman (2011)

We present results from a randomized evaluation of a teacher performance pay program implemented across a large representative sample of government-run rural primary schools in the Indian state of Andhra Pradesh. At the end of 2 years of the program, students in incentive schools performed significantly better than those in control schools by 0.27 and 0.17 standard deviations in math and language tests, respectively. We find no evidence of any adverse consequences of the program. The program was highly cost effective, and incentive schools performed significantly better than other randomly chosen schools that received additional schooling inputs of a similar value.

School Enrollments

	Log School Enrollment (1)	School Proximity (8–24) (2)	School Infrastructure (0–6) (3)	Household Affluence (0–7) (4)	Parental Literacy (0–4) (5)	Scheduled Caste/Tribe (6)	Male (7)	Normalized Baseline Score (8)
				Two-Year Effect				
Incentive	−.198	−.019	.28**	.09	.224***	.226***	.233***	.219***
	(.354)	(.199)	(.130)	(.073)	(.054)	(.049)	(.049)	(.047)
Covariate	−.065	−.005	.025	.017	.068***	−.066	.029	.448***
	(.058)	(.010)	(.038)	(.014)	(.015)	(.042)	(.027)	(.024)
Interaction	.083	.018	−.02	.038**	−.003	−.013	−.02	.006
	(.074)	(.014)	(.040)	(.019)	(.019)	(.056)	(.034)	(.031)
Observations	29,760	29,760	29,760	25,231	25,226	29,760	25,881	29,760
R^2	.244	.244	.243	.272	.273	.244	.266	.243
				One-Year Effect				
Incentive	−.36	−.076	.032	.004	.166***	.164***	.157***	.149***
	(.381)	(.161)	(.110)	(.060)	(.047)	(.045)	(.044)	(.042)
Covariate	−.128**	−.016*	−.001	.017	.08***	.007	.016	.502***
	(.061)	(.008)	(.025)	(.013)	(.012)	(.035)	(.020)	(.021)
Interaction	.103	.017	.041	.042**	−.013	−.06	.002	.000
	(.081)	(.011)	(.031)	(.017)	(.016)	(.048)	(.025)	(.026)
Observations	42,145	41,131	41,131	38,545	38,525	42,145	39,540	42,145
R^2	.31	.32	.32	.34	.34	.31	.33	.31

Data Structure: Teacher Gender

Test Score	Male Teacher	Incentive	Interaction
10	1	1	1
12	1	0	0
8	0	1	0
14	0	0	0

Results: Teacher Gender

TABLE 6
HETEROGENOUS TREATMENT EFFECTS
A. HOUSEHOLD AND SCHOOL CHARACTERISTICS

	Log School Enrollment (1)	School Proximity (8–24) (2)	School Infrastructure (0–6) (3)	Household Affluence (0–7) (4)	Parental Literacy (0–4) (5)	Scheduled Caste/Tribe (6)	Male (7)	Normalized Baseline Score (8)
	Two-Year Effect							
Incentive	−.198	−.019	.28**	.09	.224***	.226***	.233***	.219***
	(.354)	(.199)	(.130)	(.073)	(.054)	(.049)	(.049)	(.047)
Covariate	−.065	−.005	.025	.017	.068***	−.066	.029	.448***
	(.058)	(.010)	(.038)	(.014)	(.015)	(.042)	(.027)	(.024)
Interaction	.083	.018	−.02	.038**	−.003	−.013	−.02	.006
	(.074)	(.014)	(.040)	(.019)	(.019)	(.056)	(.034)	(.031)
Observations	29,760	29,760	29,760	25,231	25,226	29,760	25,881	29,760
R^2	.244	.244	.243	.272	.273	.244	.266	.243

Estimating Equation

$$\begin{aligned} T_{ijkm}(Y_n) = \alpha &+ \gamma T_{ijkm}(Y_0) + \delta_1 Incentives_i \\ &+ \delta_2 Characteristic_i \\ &+ \delta_3 (Incentives_i \times Characteristic_i) \\ &+ \beta Z_m + \epsilon_k + \epsilon_{jk} + \epsilon_{ijk} \end{aligned}$$

- T is a test score taken in year Y
- Z_m are covariates that are used to increase efficiency of the estimate
- Subscripts are:
  - i indexes the student, j the grade, and k the school, and m mandal
  - n year of observation

Regressors of interest

- Incentives
- Characteristics

Reading Regression Results

- Each column represents a separate regression estimate
  - **Dependent Variable**: Students' test score after treatment, normalized by the standard deviation across students

TABLE 6
HETEROGENOUS TREATMENT EFFECTS
A. HOUSEHOLD AND SCHOOL CHARACTERISTICS

	Log School Enrollment (1)	School Proximity (8–24) (2)	School Infrastructure (0–6) (3)	Household Affluence (0–7) (4)	Parental Literacy (0–4) (5)	Scheduled Caste/Tribe (6)	Male (7)	Normalized Baseline Score (8)
	Two-Year Effect							
Incentive	−.198	−.019	.28**	.09	.224***	.226***	.233***	.219***
	(.354)	(.199)	(.130)	(.073)	(.054)	(.049)	(.049)	(.047)
Covariate	−.065	−.005	.025	.017	.068***	−.066	.029	.448***
	(.058)	(.010)	(.038)	(.014)	(.015)	(.042)	(.027)	(.024)
Interaction	.083	.018	−.02	.038**	−.003	−.013	−.02	.006
	(.074)	(.014)	(.040)	(.019)	(.019)	(.056)	(.034)	(.031)
Observations	29,760	29,760	29,760	25,231	25,226	29,760	25,881	29,760
R^2	.244	.244	.243	.272	.273	.244	.266	.243

- If a male student started at the 50th percentile, the treatment would have increased him to the 59th percentile.
- No compelling evidence that treatment works differently for male or female identifying students - no HTE

HTEs by Student Characteristics

Quiz Review

- Asterisks help reader pick out statistically significant effects
- Most columns had statistically significant treatment effects...
- ... but no statistically significant HTEs

Household Wealth Interaction Effect

- Only the household affluence interaction term was statistically significant.
- Household affluence variable is difficult to interpret.
  - Interaction coefficient = `0.038 **`
  - Treatment effect increases by 0.038 for each additional point of household-affluence score.
  - *Household-affluence score* ranges from 0–7
  - Assumes all categories have equal treatment-effect benefits -- $1 \rightarrow 2 = 6 \rightarrow 7$

Alternative Specification

Rather than a linear scale, allow multiple dimensions of wealth

- Seven separate dummy variables as covariates.
  - One for owning land
  - One for owning a house
  - One for having running water
  - One for owning a TV
- And then, seven different interaction terms.

One Regression to Rule Them All?

Several specific regressions or one big regression?

- Muralidharan and Sundararaman (2011) Used one column for each covariate
  - Each covariate could have been put into the same regression.
  - But, might be difficult to identify *each* effect, since covariates are not randomly-assigned
  - As a result, many of the covariates might covary with others, and so lead to "unstable" estimates
  - This instability is because one measure is picking up many, related effects

Colinearity isn't a problem with treatment variables

- Because the experiment has randomly assigned treatment, it should be independent from all sets of covariates

Teacher Incentives: Household Affluence

TABLE 6
HETEROGENOUS TREATMENT EFFECTS
A. HOUSEHOLD AND SCHOOL CHARACTERISTICS

	Log School Enrollment (1)	School Proximity (8–24) (2)	School Infrastructure (0–6) (3)	Household Affluence (0–7) (4)	Parental Literacy (0–4) (5)	Scheduled Caste/Tribe (6)	Male (7)	Normalized Baseline Score (8)
	Two-Year Effect							
Incentive	−.198	−.019	.28**	.09	.224***	.226***	.233***	.219***
	(.354)	(.199)	(.130)	(.073)	(.054)	(.049)	(.049)	(.047)
Covariate	−.065	−.005	.025	.017	.068***	−.066	.029	.448***
	(.058)	(.010)	(.038)	(.014)	(.015)	(.042)	(.027)	(.024)
Interaction	.083	.018	−.02	.038**	−.003	−.013	−.02	.006
	(.074)	(.014)	(.040)	(.019)	(.019)	(.056)	(.034)	(.031)
Observations	29,760	29,760	29,760	25,231	25,226	29,760	25,881	29,760
R^2	.244	.244	.243	.272	.273	.244	.266	.243

- Adding household affluence leads the reported effect on `Incentive` to shrink
  - Estimated overall ATE = 3.5 (average affluence) $\times 0.038 = 0.133$
  - (Note, this would only be the case if there was a uniform distribution, or symmetric distribution across the affluence scale)
  - Incentive + estimated overall ATE $0.09 + 0.133 = 0.223$
- In an interaction model, the effect of treatment depends on levels for interacted variables

Teacher Incentives: Parental Literacy

TABLE 6
HETEROGENOUS TREATMENT EFFECTS
A. HOUSEHOLD AND SCHOOL CHARACTERISTICS

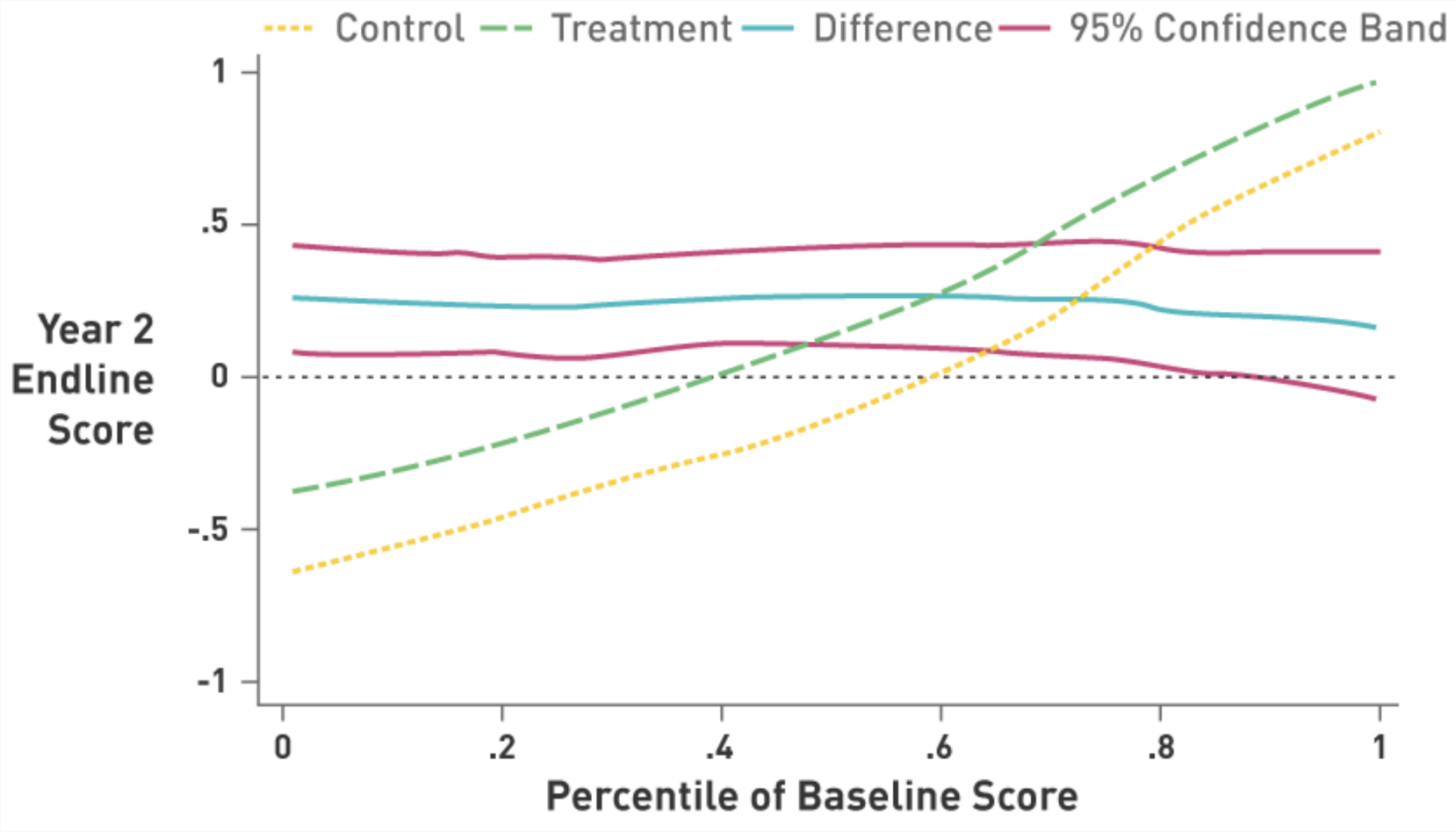
	Log School Enrollment (1)	School Proximity (8–24) (2)	School Infrastructure (0–6) (3)	Household Affluence (0–7) (4)	Parental Literacy (0–4) (5)	Scheduled Caste/Tribe (6)	Male (7)	Normalized Baseline Score (8)
	Two-Year Effect							
Incentive	−.198	−.019	.28**	.09	.224***	.226***	.233***	.219***
	(.354)	(.199)	(.130)	(.073)	(.054)	(.049)	(.049)	(.047)
Covariate	−.065	−.005	.025	.017	.068***	−.066	.029	.448***
	(.058)	(.010)	(.038)	(.014)	(.015)	(.042)	(.027)	(.024)
Interaction	.083	.018	−.02	.038**	−.003	−.013	−.02	.006
	(.074)	(.014)	(.040)	(.019)	(.019)	(.056)	(.034)	(.031)
Observations	29,760	29,760	29,760	25,231	25,226	29,760	25,881	29,760
R^2	.244	.244	.243	.272	.273	.244	.266	.243

- Always be aware of the amount of statistical uncertainty in estimates
  - Confidence interval: -0.003 ± 0.038 per unit of literacy
  - The largest absolute value in that interval is $-0.003 - 0.038 = -0.041$.
- Even using the coefficient estimate with the largest magnitude -0.041 and evaluating it at the largest parental literacy value, 4, the estimated total interaction effect confidence interval is $-0.041 \times 4 = -0.16$
- This is still smaller than the baseline treatment effect, 0.224 for those with totally parents marked as a zero on literacy

Always look at the magnitude of the estimate!

Always look at size of the confidence interval!

Broad based Treatment Effect



HTEs by Teacher Characteristics

Review: Student Characteristic HTE

- Broad-based benefits from teacher incentive program were found.
  - No HTEs by past student test score
  - Household wealth was the only significant HTE interaction.

Teacher Characteristics

Review the table. Answer the quiz questions.

Discussin of Quiz

Disucssion of Quiz

Quiz Review

- Most significant effects (at the 5% level) for covariates of (a) teacher training and (b) years of experience.
  - Displayed in the interaction row
- The interaction terms in columns two and three show the most significant impact.
- Teachers with more experience deliver fewer benefits based on incentives.
- Teachers with more training deliver more benefits based on incentives.
- Understanding the size of the HTEs is important.
- Examine the magnitude of the regressors.

Teacher Training Quiz

Teacher Training Quiz

Teacher training levels:

```
1 = No training
2 = Diploma
3 = Bachelor's
4 = Master's
```

- Starting with a baseline of 0 for the no training value makes coefficient interpretation easier.
- Main treatment effect in this specification is negative. Why?

	(2) Training
Incentive	-0.224
	(0.176)

Teacher Training Effect and Interaction

	(2) Training
Incentive	-0.224
	(0.176)
Covariate	-0.051
	(0.041)
Interaction	0.138^{**}
	(0.061)
Observations	$53,890$
R^2	0.29

- -0.224 is the baseline treatment effect for a teacher with training score of 0.
- Although no teachers have a training score of 0.
- Starting with a baseline of 0 makes coefficient interpretation easier
- Estimate is not statistically different from 0.
- T-ratio is very low.

Estimating Confidence Intervals for HTEs

- Confidence interval for incentive effect (for those with training = 0) in this specification: -0.224 ± 0.35
- Covariate should equal 0 (not 1) for no training.
  - Incentive coefficient would be easier to interpret.
- Few teachers in the sample had literally no training.
- What would be the estimate for a teacher who did have training?

Discussion of Bachelor's Degree Quiz

Discussion of Bachelor's Degree Quiz

Answer: Estimated estimated treatment effect for teachers with bachelor's degrees

$$\begin{aligned} -0.244 + 0.138 \times 3 &= 0.19 \\ \text{Incentive} + \text{Incentive} \times \text{Bachelors} &= \text{Treatment effect} \end{aligned}$$

- Positive treatment effect estimate for those with bachelor's degrees, negative treatment effect for teachers with no training.

Easier Method (for reporting)

- Split up the sample by the levels of training.
- Estimate treatment effect for each group.
- Standard errors on the treatment effect will probably be large.
- Both the incentive and interaction coefficients have large standard errors.

Discussion of Bachelor's Degree Quiz

	(2) Training
Incentive	-0.224
	(0.176)
Covariate	-0.051
	(0.041)
Interaction	0.138^{**}
	(0.061)
Observations	$53,890$
R^2	0.29

- Delta method could be used to compute exact standard errors.
- If the covariate is well defined, it's good enough to:
  - Examine the size of the treatment and interaction coefficients
  - Determine if interaction coefficient is statistically significant
  - Estimate the size of associated standard errors

Conclusion

- Understand point estimates.
- Determine statistical significance of the HTE coefficient.
- Compute the point estimate of the treatment effect for different values of the covariates.

Reading: *Field Experiments*, Section 9.3.2

- Pay attention to the paragraph at the bottom of page 301.
- Subgroup membership is non-experimental in nature

Example of HTEs in a Multifactor Experiment

Reading

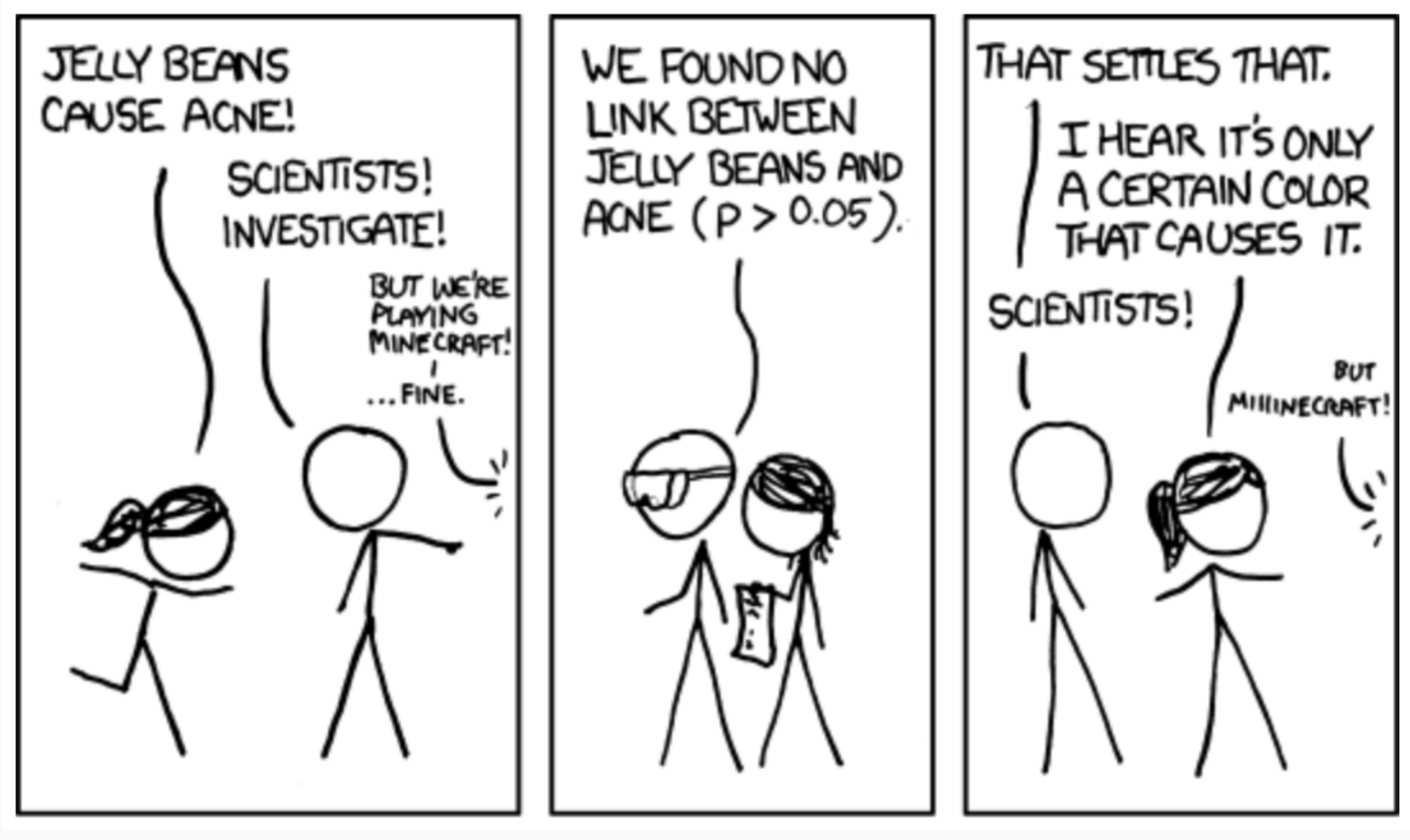
Reading: *Field Experiments* Section 9.4

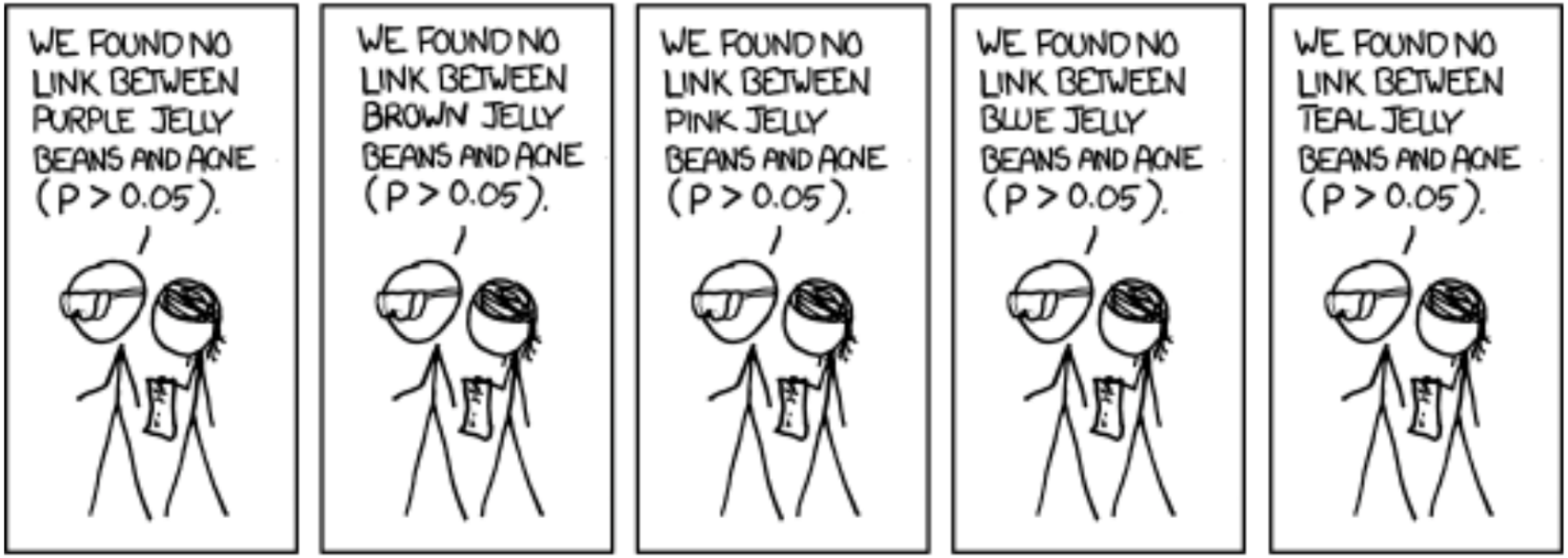
- Two treatment variables, J and G, and one covariate, H
  - J: Name is signed as *Jose* or *Colin*
  - G: Grammar is *bad* or *good*
  - H: Legislator is *Hispanic* or *Non Hispanic*

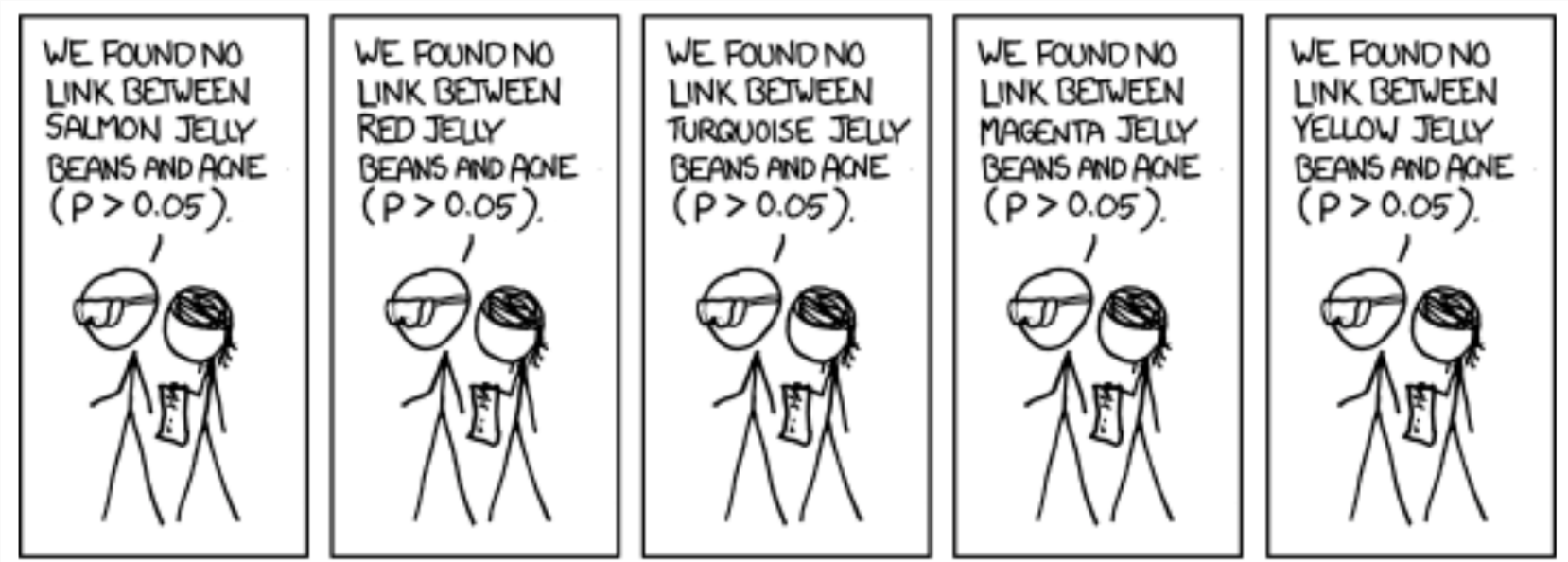
Reading: *Field Experiments* Section 9.5.

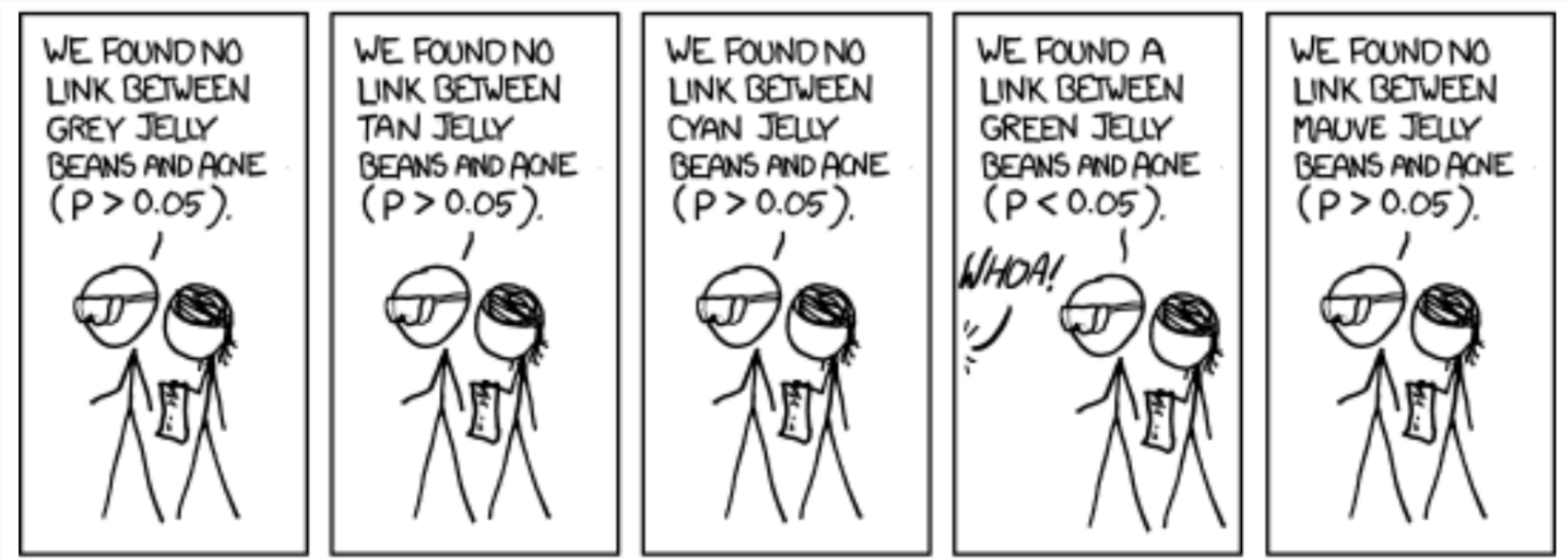
- Study Equation 9.19 and Table 9.2.
- Equation 9.19: regression presentation of data in Table 9.2
- New scenario:
  - Three binary variables
  - Three possible two-way interactions
  - Three-way interaction

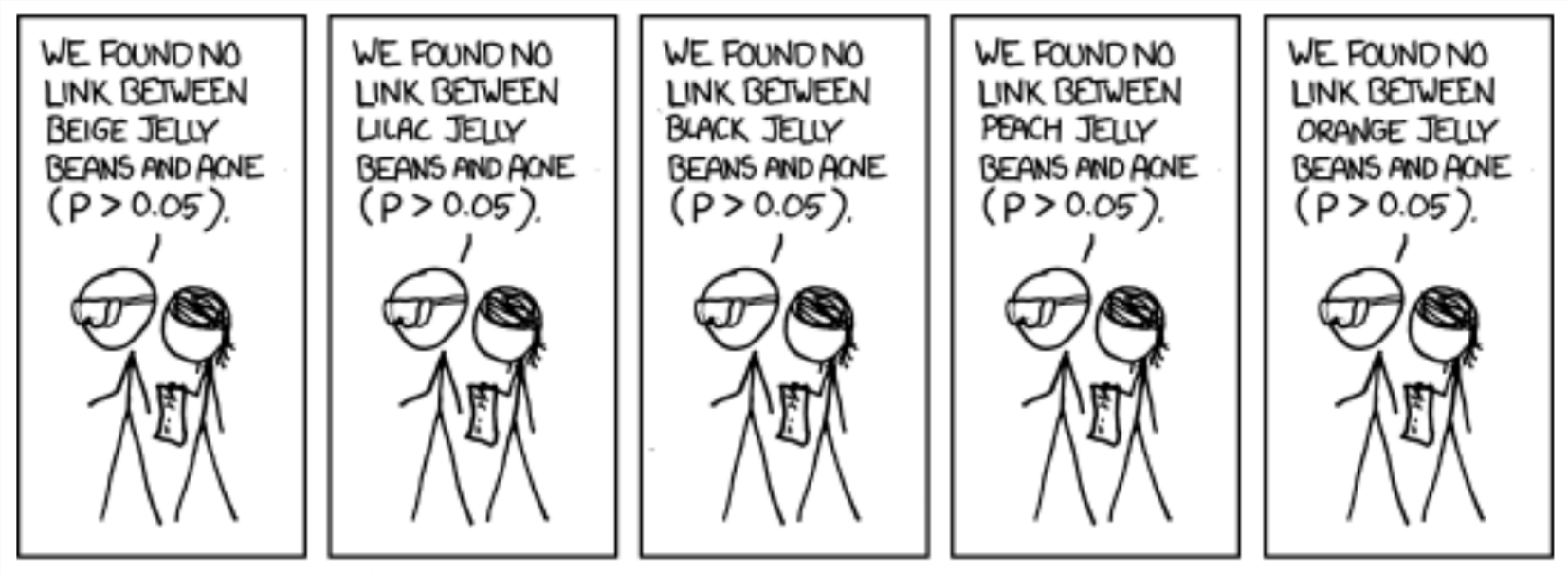
The Multiple-Comparisons Problem


JELLY BEANS CAUSE ACNE!
SCIENTISTS! INVESTIGATE!
BUT WE'RE PLAYING MINECRAFT!
... FINE.
WE FOUND NO LINK BETWEEN JELLY BEANS AND ACNE ($p > 0.05$).
THAT SETTLES THAT.
I HEAR IT'S ONLY A CERTAIN COLOR THAT CAUSES IT.
SCIENTISTS!
BUT MIIIINECRAFT!


WE FOUND NO LINK BETWEEN PURPLE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BROWN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PINK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLUE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TEAL JELLY BEANS AND ACNE (P > 0.05).


WE FOUND NO LINK BETWEEN SALMON JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN RED JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TURQUOISE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN MAGENTA JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN YELLOW JELLY BEANS AND ACNE (P > 0.05).


WE FOUND NO LINK BETWEEN GREY JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN TAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN CYAN JELLY BEANS AND ACNE (P > 0.05).
WE FOUND A LINK BETWEEN GREEN JELLY BEANS AND ACNE (P < 0.05).
WHOA!
WE FOUND NO LINK BETWEEN MAUVE JELLY BEANS AND ACNE (P > 0.05).


WE FOUND NO LINK BETWEEN BEIGE JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN LILAC JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN BLACK JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN PEACH JELLY BEANS AND ACNE (P > 0.05).
WE FOUND NO LINK BETWEEN ORANGE JELLY BEANS AND ACNE (P > 0.05).


NEWS
GREEN JELLY BEANS LINKED TO ACNE!
95% CONFIDENCE

Multiple Comparisons Problem

- More variables means more specification searching is possible.
- Specifications can be changed until the coefficients are suitable.
- All possible covariates can be tried until statistical significance is found.
- Statistical theory assumes we know the correct model.
  - We don't always have the correct model, though.
- Some searching is inevitable.
- An effect isn't necessarily real when one coefficient was significant out of many possibilities tried.

Fishing Expeditions

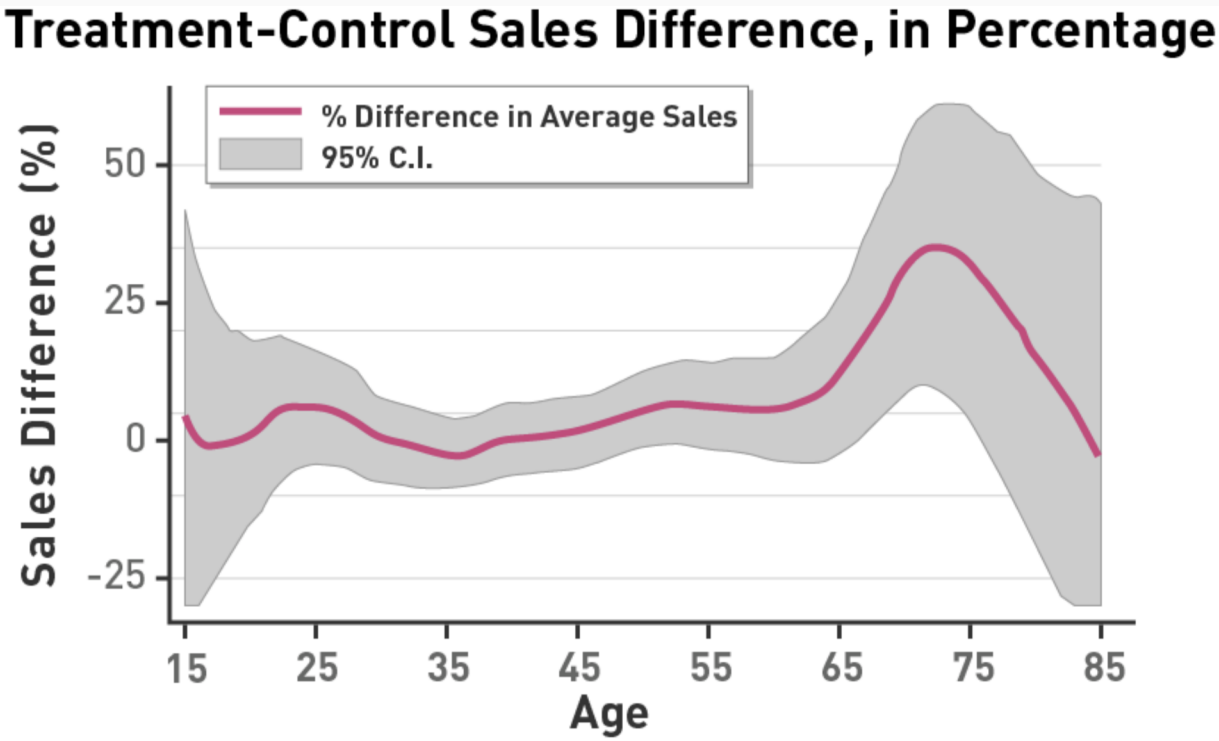
- Trying out multiple hypotheses before picking a favorite.
- Violates assumptions that give valid confidence intervals.
- Analyzing data in different ways is OK.
- But the computed confidence intervals will be too narrow.

Fishing Expeditions: Solutions

- Bonferroni correction (see Box 9.4): helps avoid overstating statistical significance.
- Critical values will be higher than without correcting, i.e. $t_{BF} > 1.96$
- Consider findings to be interesting hypotheses.
- Test them in another experiment to ensure replication

Lewis and Reiley Age Effects

- Found higher sales revenue from older people who watched advertisements
- *Buuuuut...* results didn't replicate in a second experiment.



Heterogeneous treatment effects by age, from Lewis and Reiley (2014)

Teacher Incentives Effect

- Affluent households had statistically significantly larger treatment effects.
  - This was the only significant interaction term out of eight tried.
  - Bonferroni correction may have shown that this effect wasn't significant.
- Muralidharan and Sundararaman examined many covariates.
  - But properly avoided a fishing expedition
- Student ability didn't affect incentive treatment.

Final Thoughts

Quiz Recap

Quiz Recap

- Practiced converting sample means to regression coefficients.
- Examined multifactor experiments.
- Examined HTEs.
- This quiz put together HTEs and multifactor experiments.

Automated Searches for Interactions

Main point:

Fishing expeditions are a problem because of multiple comparisons.

- Keep analysis simple.
- Machine learning can help find heterogeneous treatment effects.
- Performs automated specification searches.
- Multiple comparisons will still be an issue.
- Treat any output as a hypothesis to be confirmed in another experiment.

What to Remember From This Week

Options for reporting heterogeneous treatment effects

- Report separate treatment effects for each subgroup.
- Use regressions where the treatment dummy variable is multiplied by covariates of interest.

Defining covariates so that reading output is easier

Being able to read output of regression models with interaction terms

Testing significance of treatment effects between subgroups

- Can be done with a t-test on one coefficient
- Can also be done with an F-test

What to Remember, continued

Interpreting HTEs

- HTEs explain how different subgroups respond to treatment.
- HTEs don't explain causal effects of reassigning people to new subgroups.

Multiple-comparisons problem

- Examining many different regression models and picking favorites makes us overstate statistical significance.
- Results from fishing expeditions should be treated as hypotheses to be retested.
- Machine learning can help develop testable HTE hypotheses.